Digital Biology Classes 2012

Datamining and Drug Design

L. Ridgway Scott

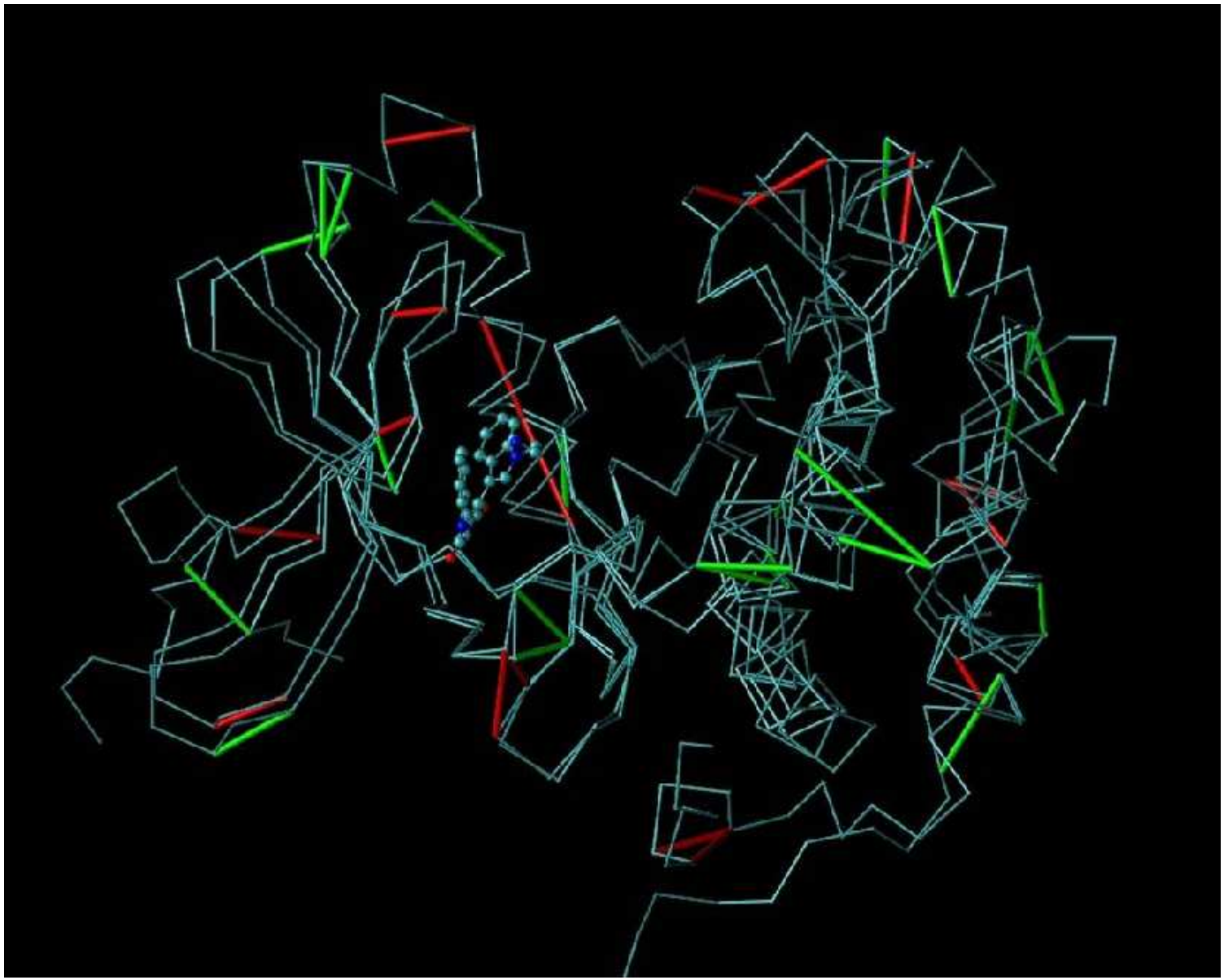
Departments of Computer Science and Mathematics,

Computation Institute, and Institute for Biophysical Dynamics,

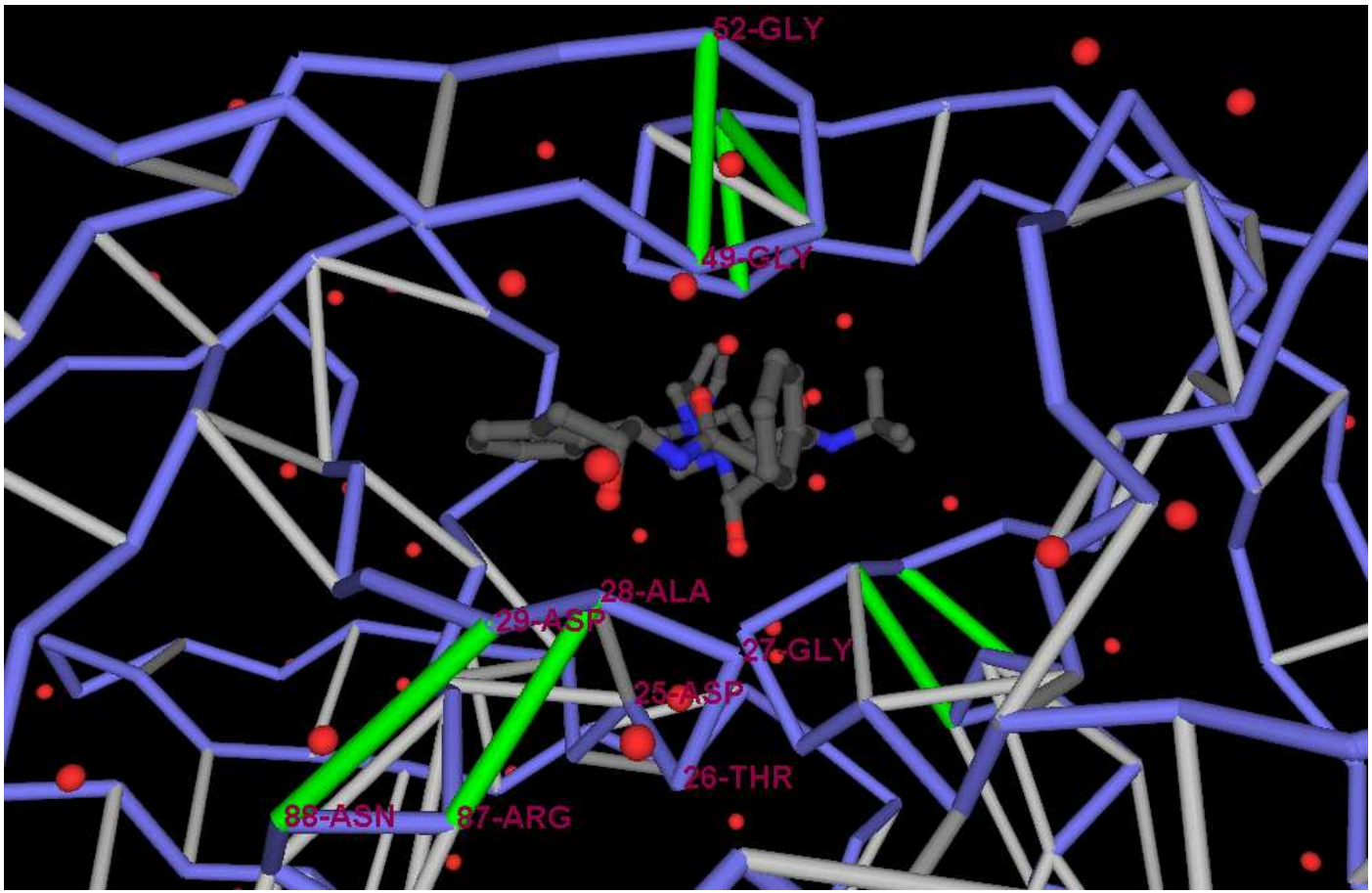
University of Chicago

This talk is based on joint work with Ariel Ferndandez (Univ. of Wisconsin–Madison), Xie Dexuan (Univ. of Wisconsin–Milwaukee), Harold Scheraga (Cornell), and Kristina Rogale Plazonic (Princeton); and at U. Chicago: Steve Berry, Peter Brune, and Chris Fraser.

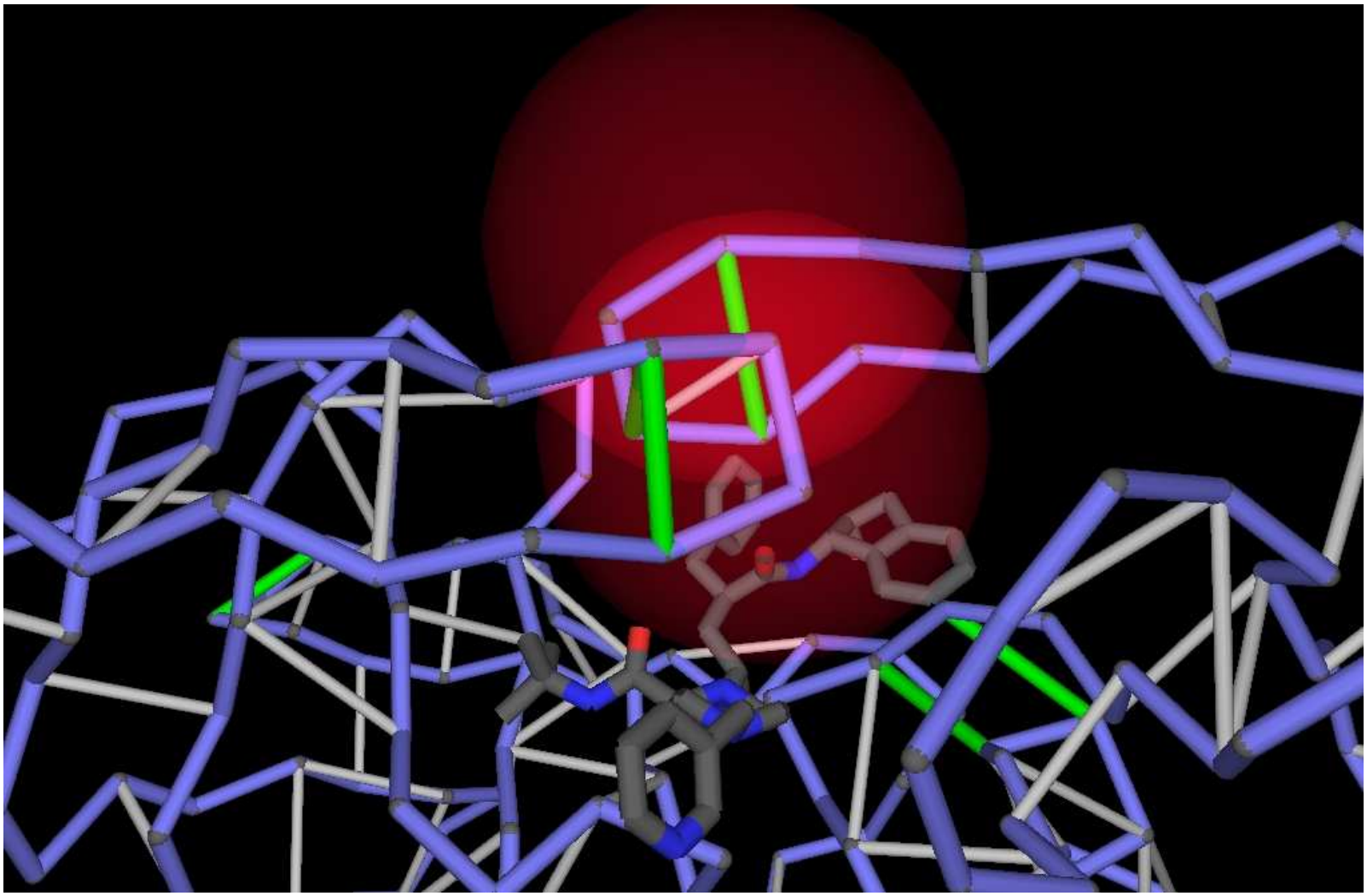
Aligned backbones for two paralog kinases; dehydrons for Chk1 are marked in green and those for Pdk1 are in red [4].



HIV-1 protease with ‘dehydron wrapper’ inhibitor



Detail of the protease cavity, pattern of packing defects, and inhibitor positioned as dehydron wrapper.



Desolvation spheres for flap Gly-49–Gly-52 dehydron containing nonpolar groups of the wrapping inhibitor.

Failure of drug designs in clinical trials [1].

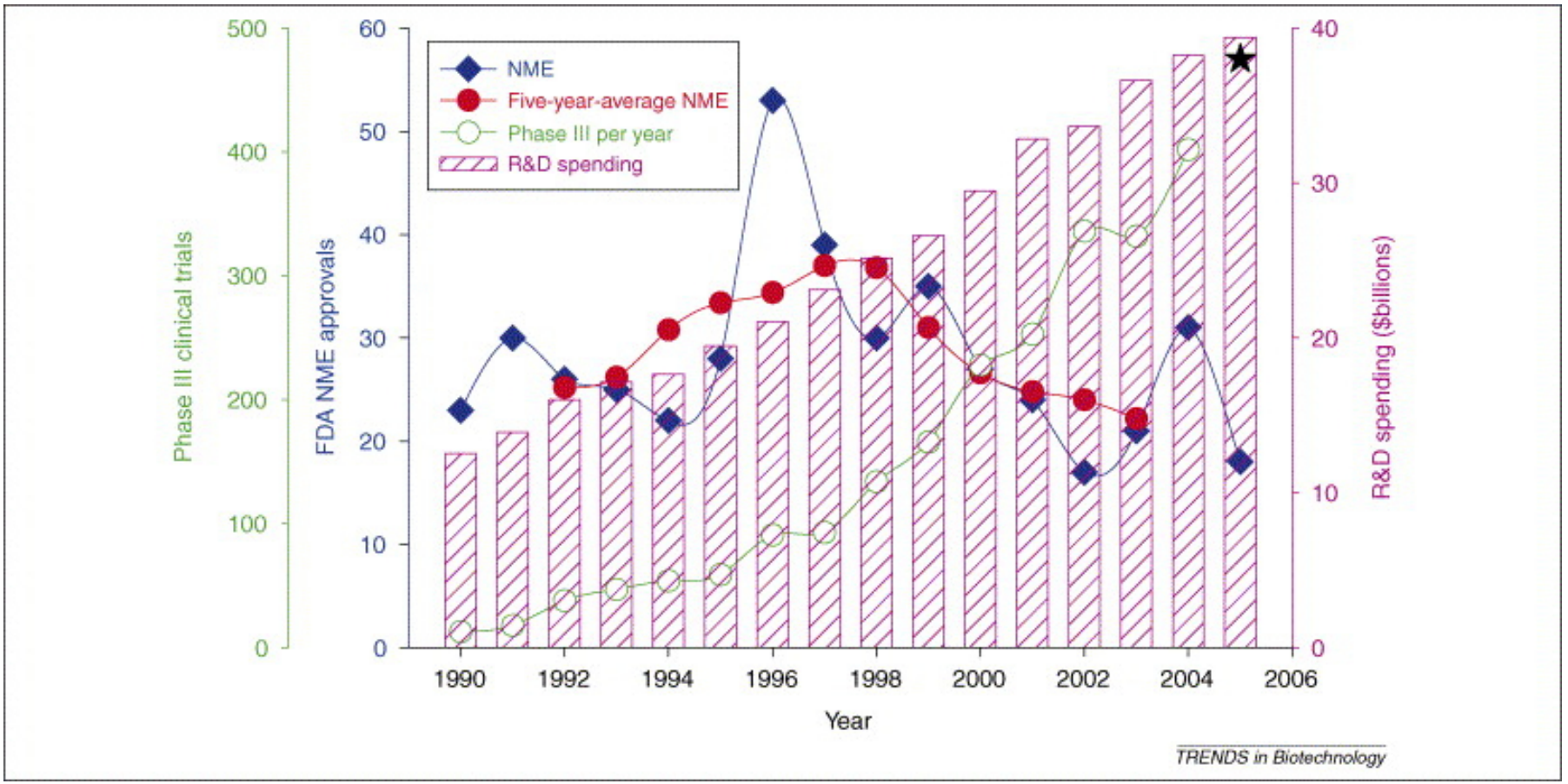
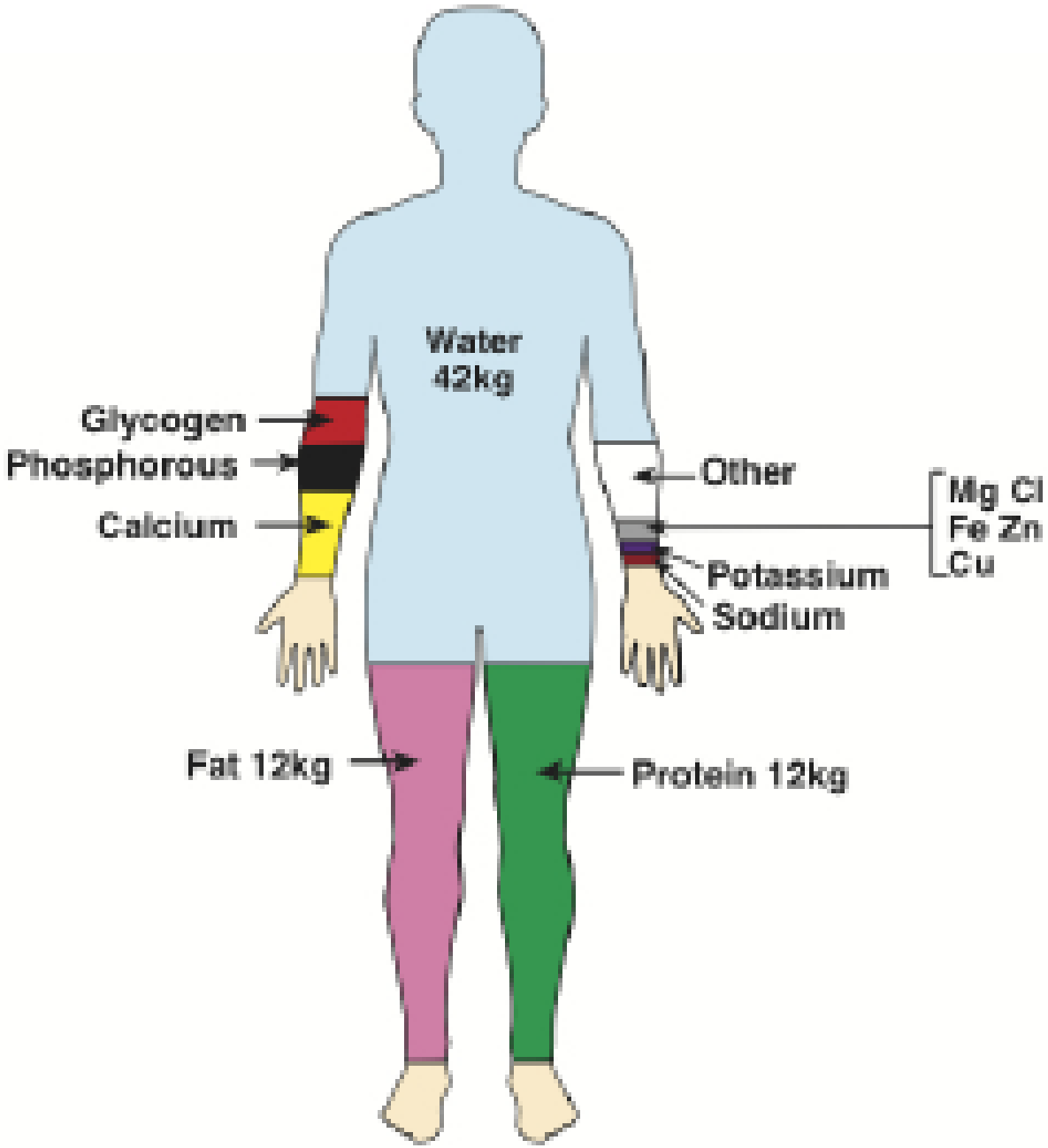


Figure 1: FDA approvals of new molecule entities (NME), numbers of phase III clinical trials completed and pharmaceutical R&D spending from 1990 to 2005. NME approvals reached a peak in the period from 1996 to 1999 and are now dropping in spite of continually increasing R&D spending and phase III clinical trials.


Water
42kg
Glycogen
Phosphorous
Calcium
Other
Mg Cl
Fe Zn
Cu
Potassium
Sodium
Fat 12kg
Protein 12kg

Competing effects: why this is so hard

Protein sidechains have large electrostatic gradients

Water is a strong dielectric

Hydrophobic groups modify the water structure

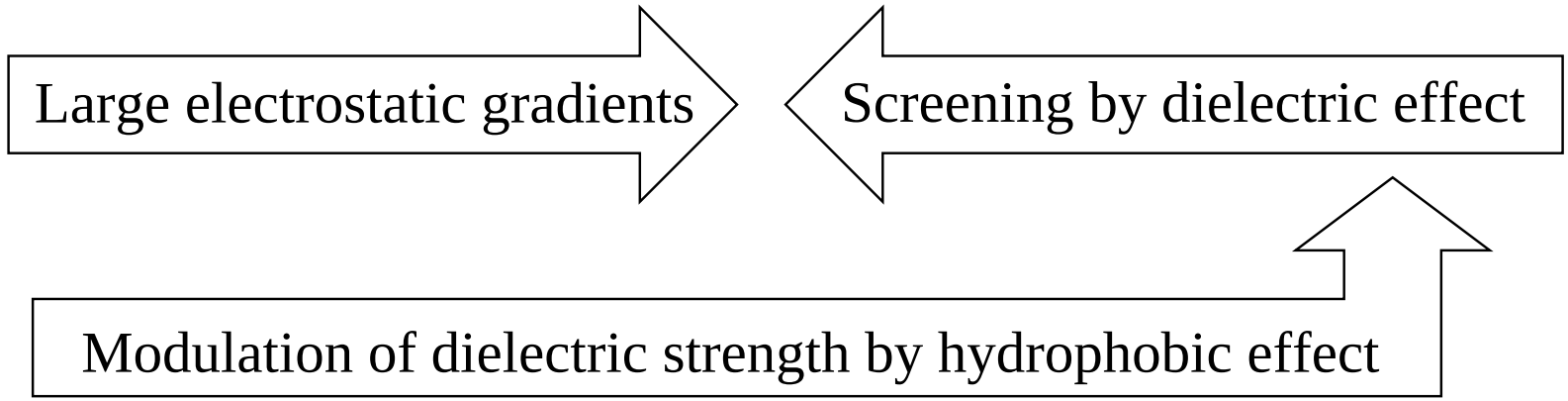


Figure 2: Three competing effects that determine protein behavior. These conspire to weaken interactive forces, making biological relationships more tenuous and amenable to mutation.

1 Protein basics

Proteins are sequences of amino acids which are covalently bonded along a “backbone.”

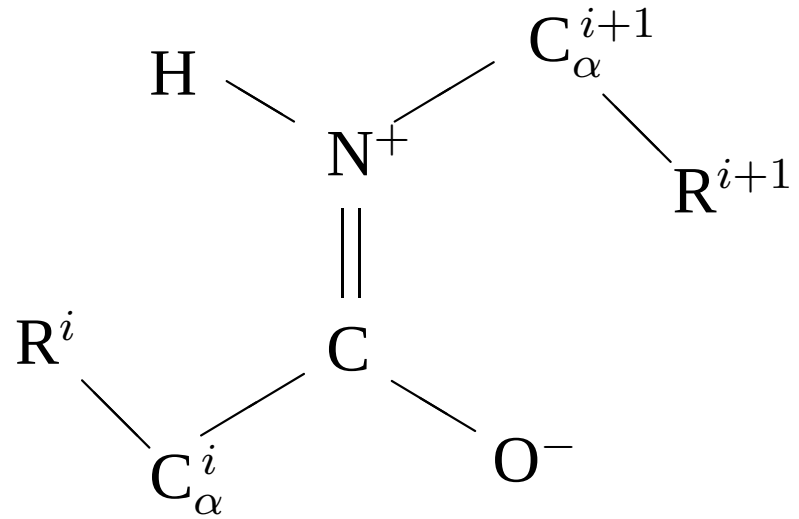
Proteins of biological significance fold into a three-dimensional structure by adding hydrogen bonds between carbonyl and amide groups on the backbone of different amino acids.

In addition, other bonds, such as a salt bridge or a disulfide bond can form between particular amino acids (Cysteine has sulfur atoms in its sidechain).

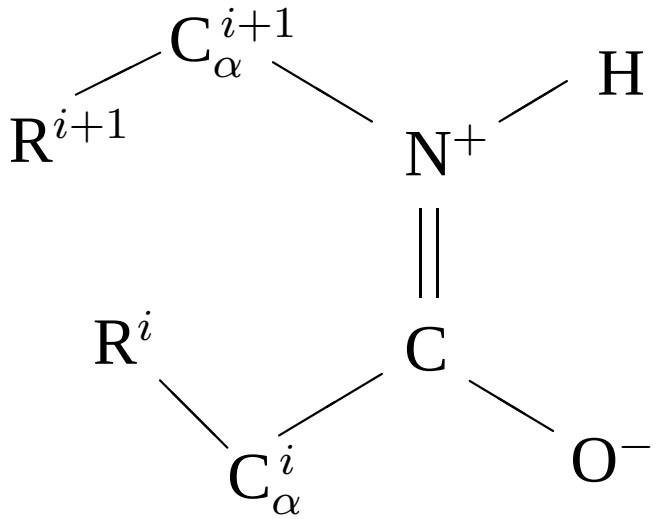
However, the hydrogen bond is the primary mode of structure formation in proteins.

1.1 Chains of amino acid residues

Proteins are chains of amino acid residues whose basic unit is the peptide group.



(a)



(b)

Figure 3: The rigid state of the peptide bond: (a) trans form, (b) cis form. The double bond between the central carbon and nitrogen keeps the peptide bond planar.

1.2 Linear (primary) structure of proteins

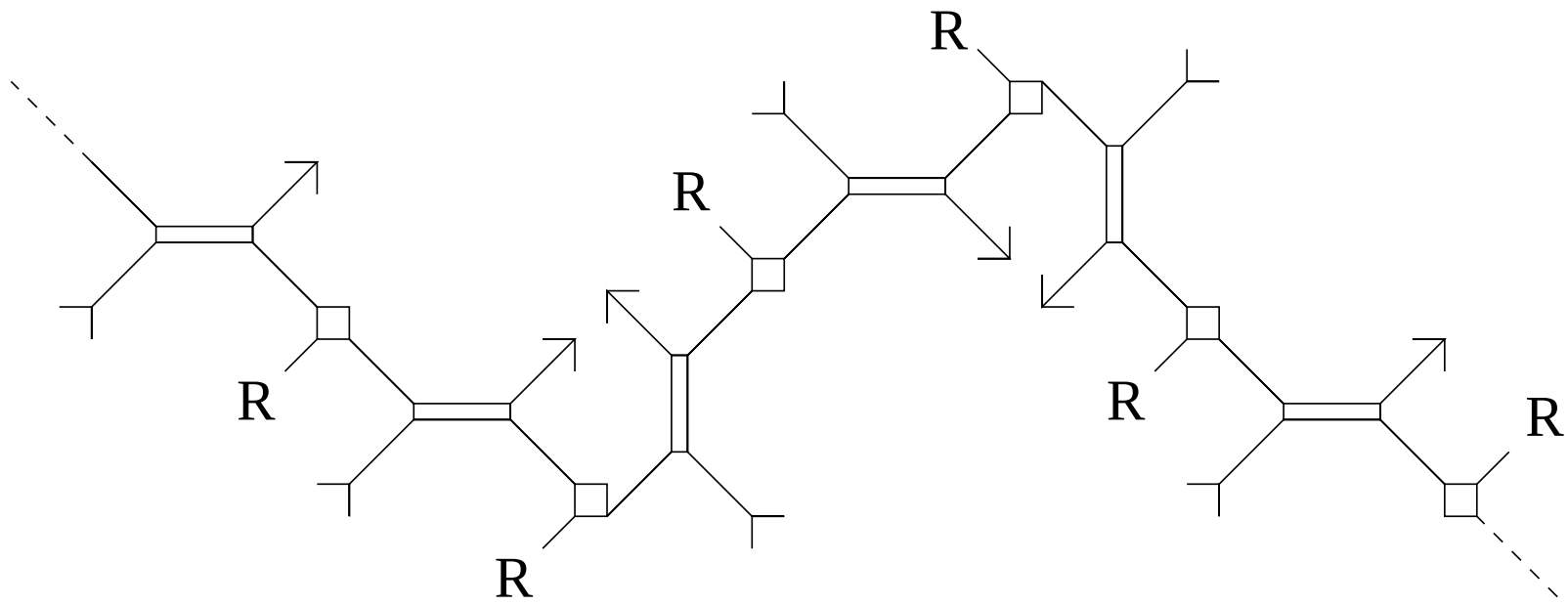
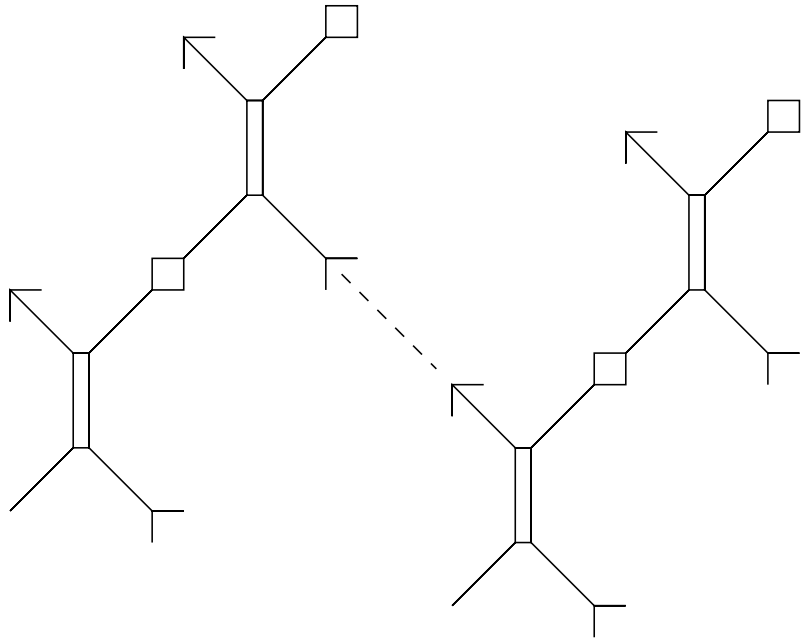


Figure 4: Cartoon of peptide sequence where all peptides are in trans form (cf. Figure 3). Small boxes represent C-alpha carbons, arrow heads represent amide groups NH, arrow tails represent carbonyl groups CO, and thin rectangular boxes are double bond between backbone C and N. The different residues are indicated by R's. The numbering scheme is increasing from left to right, so that the arrow formed by the carbonyl-amide pair points in the direction of increasing residue number. The three-dimensional nature of the protein is left to the imagination.

1.3 Hydrogen bonds and secondary structure

Proteins have a hierarchy of structure, the next being **secondary structure** consisting of two primary types: alpha-helices and beta-sheets (a.k.a., α-helices and β-sheets).

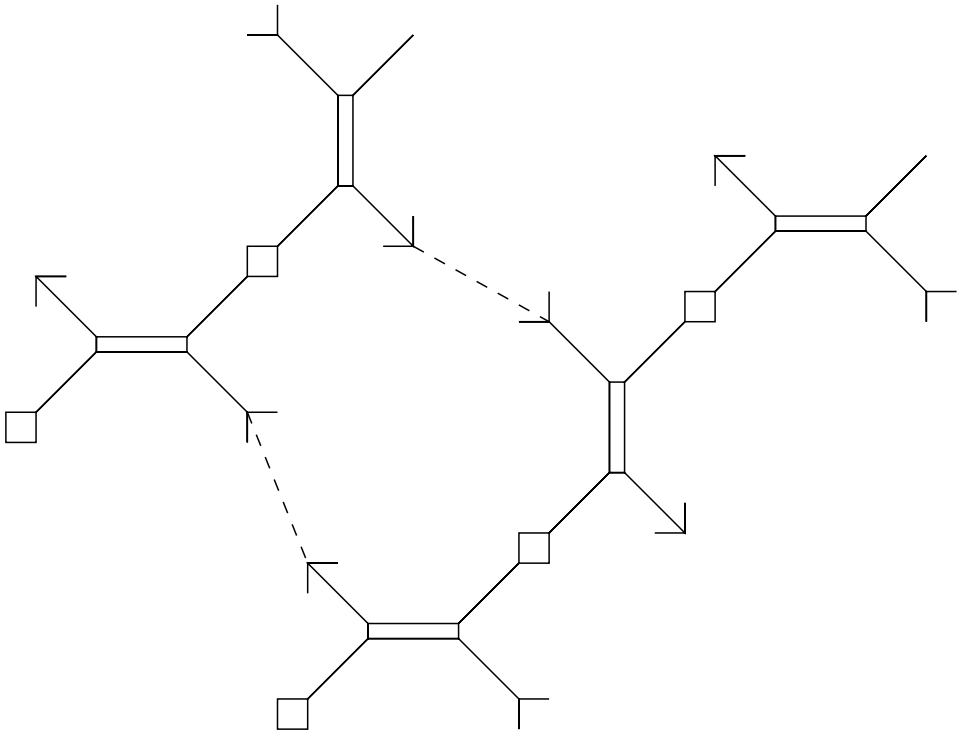
Alpha helices are helical arrangements of the subsequent peptide complexes with a distinctive hydrogen bond arrangement between the amide (NH) and carbonyl (OC) groups in peptides separated by k steps in the sequence, where primarily $k = 4$ but with $k = 3$ and $k = 5$ also occurring less frequently:



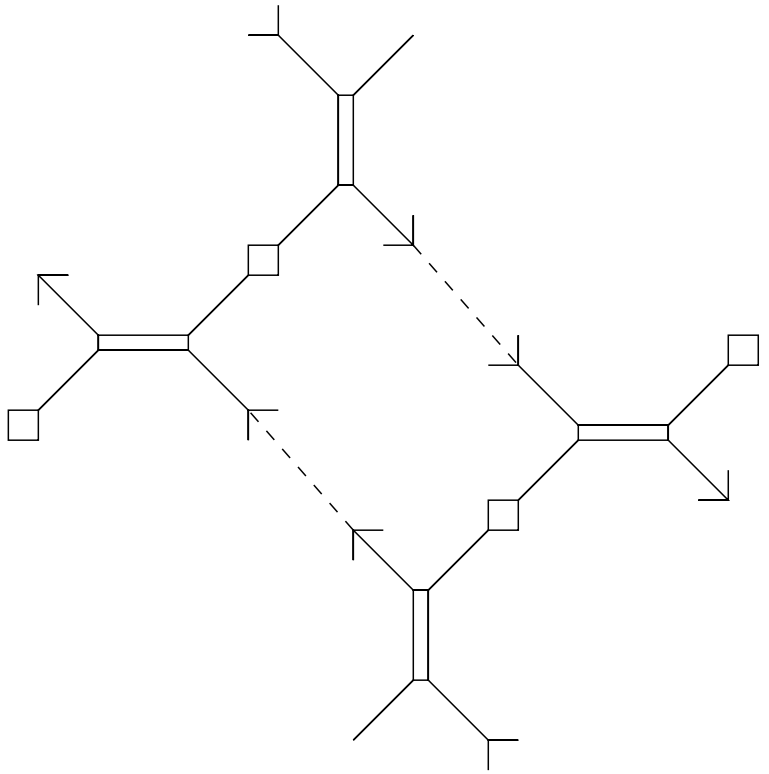
(a)

1.4 Beta sheets

Beta sheets represent different hydrogen bond arrangements: (b) is the anti-parallel arrangement and (c) is the parallel.

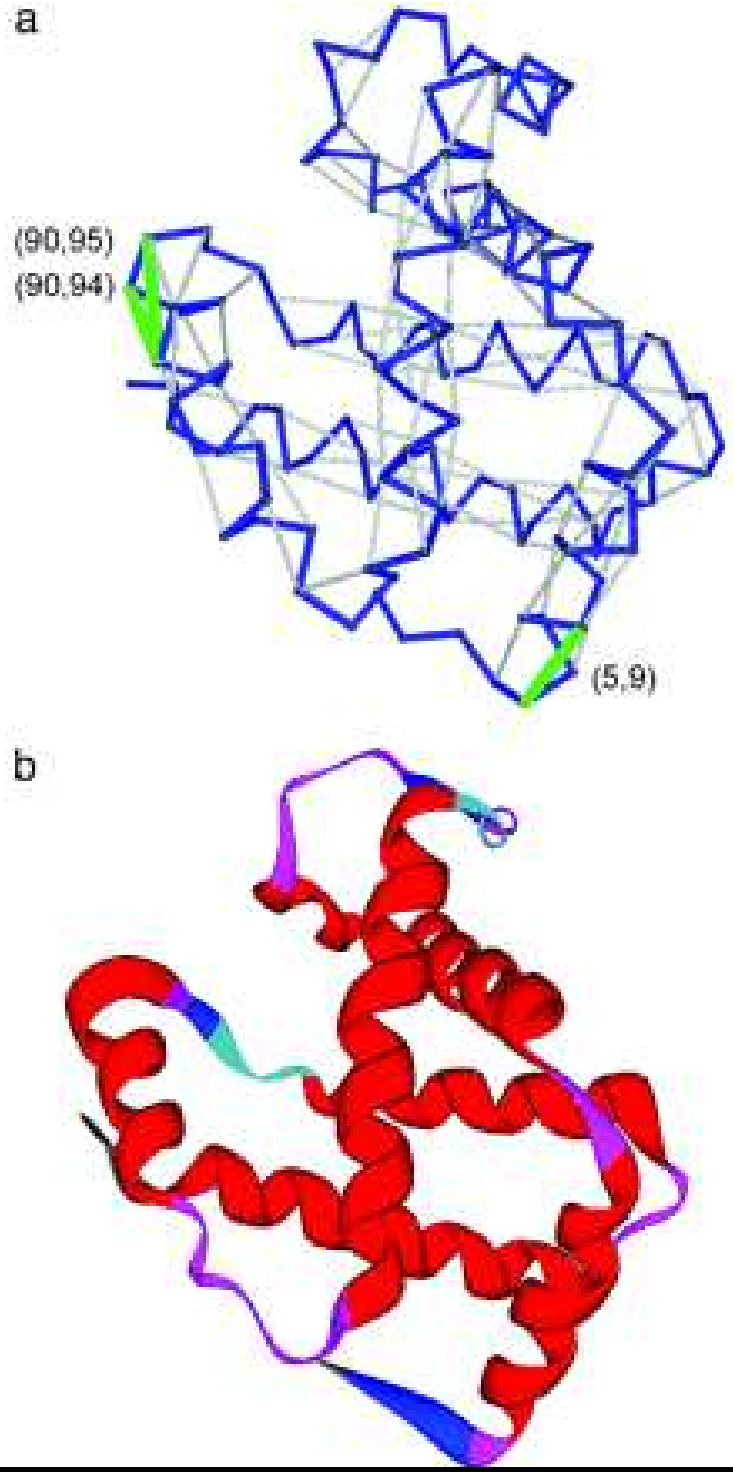


(b)



(c)

Both structures are essentially flat, in contrast to the helical structure in (a).



Dehydrons

in human hemoglobin, From PNAS 100: 6446-6451 (2003) Ariel Fernandez, Jozsef Kardos, L. Ridgway Scott, Yuji Goto, and R. Stephen Berry. Structural defects and the diagnosis of amyloidogenic propensity.

Well-wrapped hydrogen bonds are grey, and dehydrons are green.

The standard ribbon model of “structure” lacks indicators of electronic environment.

Proteins as digital components

Proteins are the essential components of life:

- used to build complexes, e.g., viruses (bricks and mortar)
- involved in signalling (information transmission)
- enzymes essential in catalysis (chemical machines)

In all these cases, protein-ligand interaction is essential.

These interactions are deterministic (always the same).

Proteins function as discrete components not as analog devices.

The hydrophobic effect

Hydrophobic effect crucial in protein-ligand association.

Water is essential to life as we know it, but hostile to proteins.

The role of water in protein biophysics: to modulate electric forces via the dielectric effect.

Hydrophobicity fosters water removal and supports protein-ligand interaction, but it also modulates the dielectric effect.

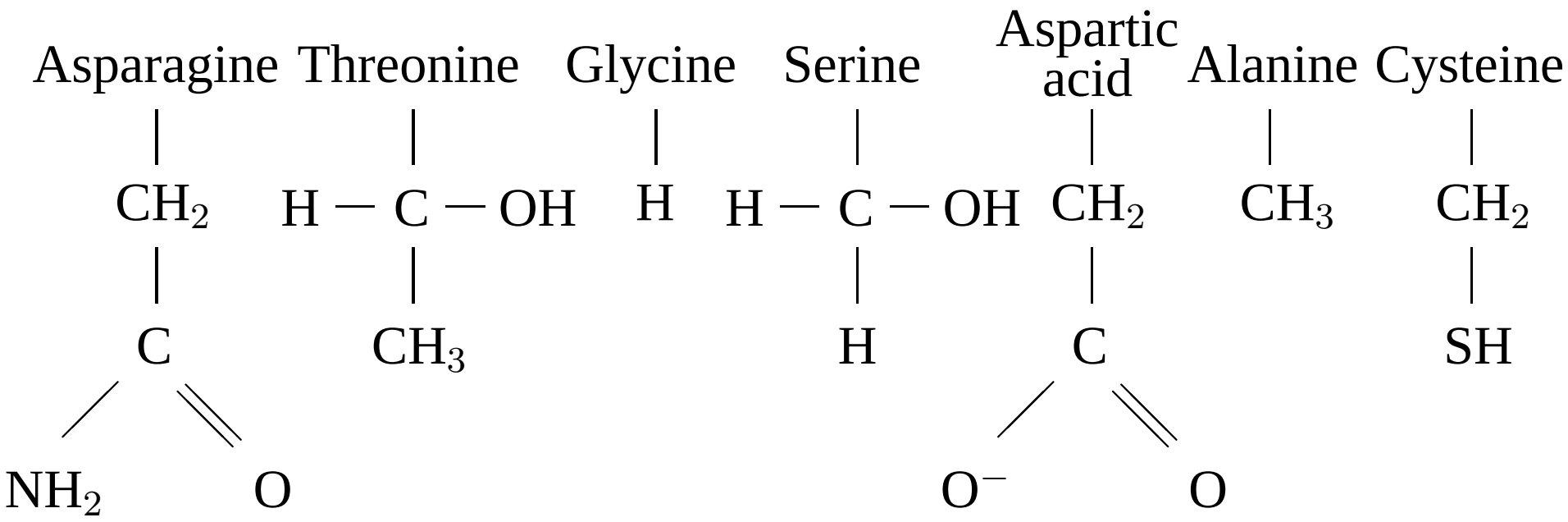
Water is a strong dielectric, and protein sidechains are a complex mix of charged, polar, and hydrophobic parts.

But the hydrophobic effect is non-specific in action.

What makes proteins interact in a repeatable way?

What sidechains are found at interfaces?

By examining interfaces in PDB structures, we can see which residues are most likely to be found at interfaces.



Sidechains most likely to be involved in interactions, ordered from the left (asparagine), are not hydrophobic.

Electronic forces

The only force of significance in biochemistry is the electric force.

— But often modulated by indirection or induction.

In terrestrial biology, water plays a significant role as a dielectric which mediates non-covalent interactions (hydrogen bonds, salt bridges, cation-pi interactions).

But the dielectric effect of water is modulated by hydrophobic components of proteins.

Moreover, a ligand can change the hydrophobic environment upon binding.

In protein-ligand interactions, this makes intramolecular bonds as important as intermolecular interactions.

Our technology

Interaction between physical chemistry and data mining in biophysical data bases.

- Data mining can lead to new results in physical chemistry that are significant in biology.
- Using physical chemistry to look at data provides insights regarding function.

We review some recent results regarding protein-ligand interaction that are based on novel insights about hydrophobic effects.

We show that sidechain configurations modulate dielectric effect.

We discuss how these can be used to understand a novel factor that supports protein-ligand binding.

A quote

from Nature's Robots ....

The exact and definite determination of life phenomena which are common to plants and animals is only one side of the physiological problem of today. The other side is the construction of a mental picture of the constitution of living matter from these general qualities. In this portion of our work we need the aid of physical chemistry.

Jacques Loeb, The biological problems of today: physiology. Science 7, 154-156 (1897).

so our theme is not so new ....

Data mining definition

WHATIS.COM: Data mining is sorting through data to identify patterns and establish relationships.

Data mining parameters include:

- Association - looking for patterns where one event is connected to another event
- Sequence or path analysis - looking for patterns where one event leads to another later event
- Classification - looking for new patterns (May result in a change in the way the data is organized but that's ok)
- Clustering - finding and visually documenting groups of facts not previously known

Conclusion: Data mining involves looking at data.

Data mining lens

If data mining is looking at data then

What type of lens do we use?

- All of these have chemical representations, e.g.,

$$C_{400}H_{620}N_{100}O_{120}P_1S_1$$

- Alphabetic sequences describe much of biology: DNA, RNA, proteins.
- All of these have three-dimensional structure.
- But structure alone does not explain how they function.

Physical chemistry clarifies the picture and allows function to be more easily interpreted.

Sequences can tell a story

Protein sequences

```
aardvarkateataviscicallyacademicianaccelerative
acetylglycineachievementacidimetricallyacridity
actressadamantadhesivenessadministrativelyadmit
afflictiveafterdinneragrypniaaimlessnessairlift
```

and DNA sequences

```
actcatatactagagtacttagacttatactagagcattacttagat
```

can be studied using automatically determined lexicons.

Joint work with John Goldsmith, Terry Clark, Jing Liu.

Sequences can tell a story (a linguistic lens)

Protein sequences

aardvarkateataxisticallyacademicianaccelerative
acetylglycineachievementacidimetrically acridity
actressadamantadhesivenessadministrativelyadmit
afflictiveafterdinneragrypniaaimlessnessairlift

and DNA sequences

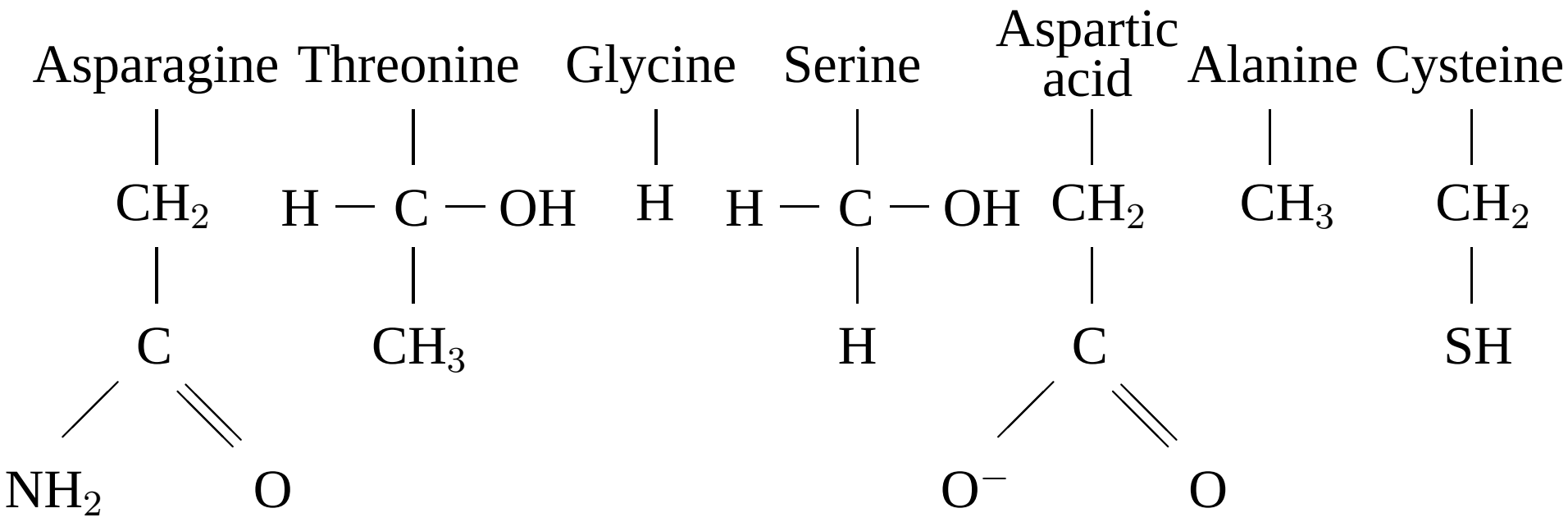
actcatatactagagtacttagacttatactagagcattacttagat

can be studied using automatically determined lexicons.

Joint work with John Goldsmith, Terry Clark, Jing Liu.

What sidechains are found at interfaces?

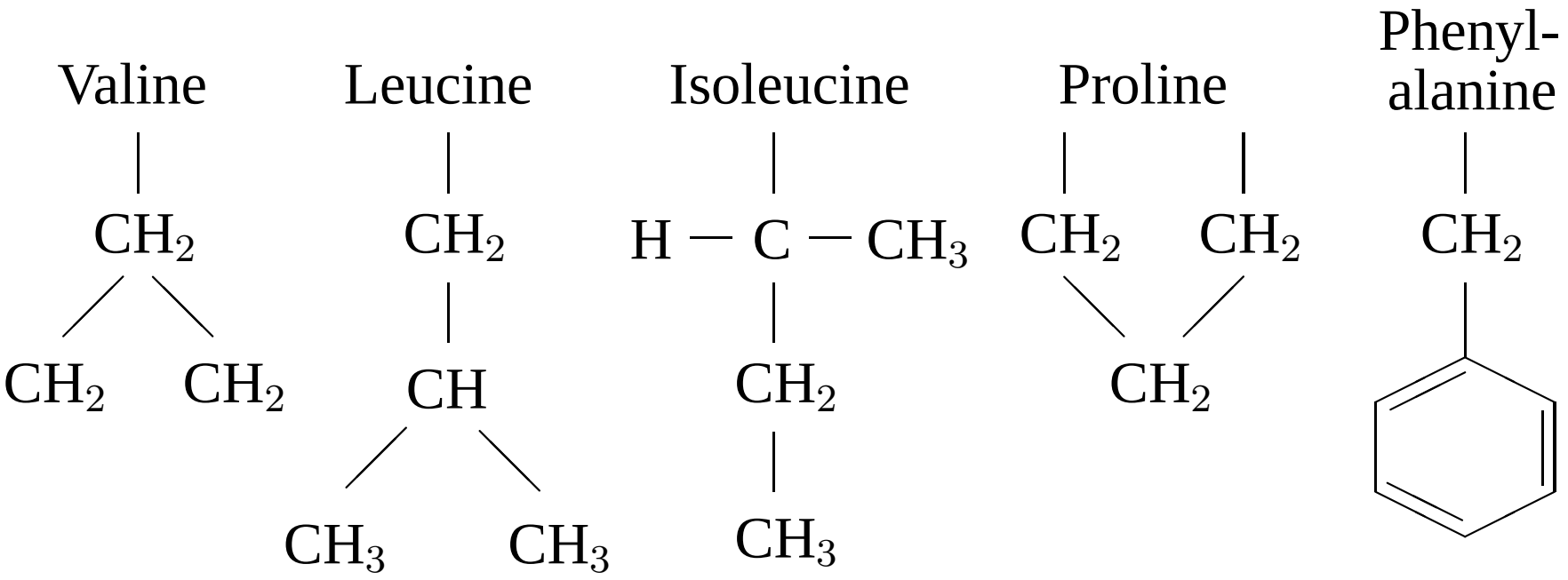
By examining interfaces in PDB structures, we can see which residues are most likely to be found at interfaces.



Sidechains most likely to be involved in interactions, ordered from the left (asparagine), are not hydrophobic.

Amino acid sidechains have different properties

Carbonaceous groups on sidechains are hydrophobic:



Amino acid residues (sidechains only shown) having only carbonaceous groups.

Charges in a dielectric are like lights in a fog.

Wrapping modifies dielectric effect

Hydrophobic (CH_n) groups remove water locally.

This causes a reduction in ε locally.

(Resulting increase in ϕ makes dehydrons sticky.)

This can be quantified and used to predict binding sites.

The placement of hydrophobic groups near an electrostatic bond is called wrapping.

Like putting insulation on an electrical wire.

We can see this effect on a single hydrogen bond.

Unit of hydrophobicity

A single carbonaceous group CH_n can enhance the strength and stability of a hydrogen bond.

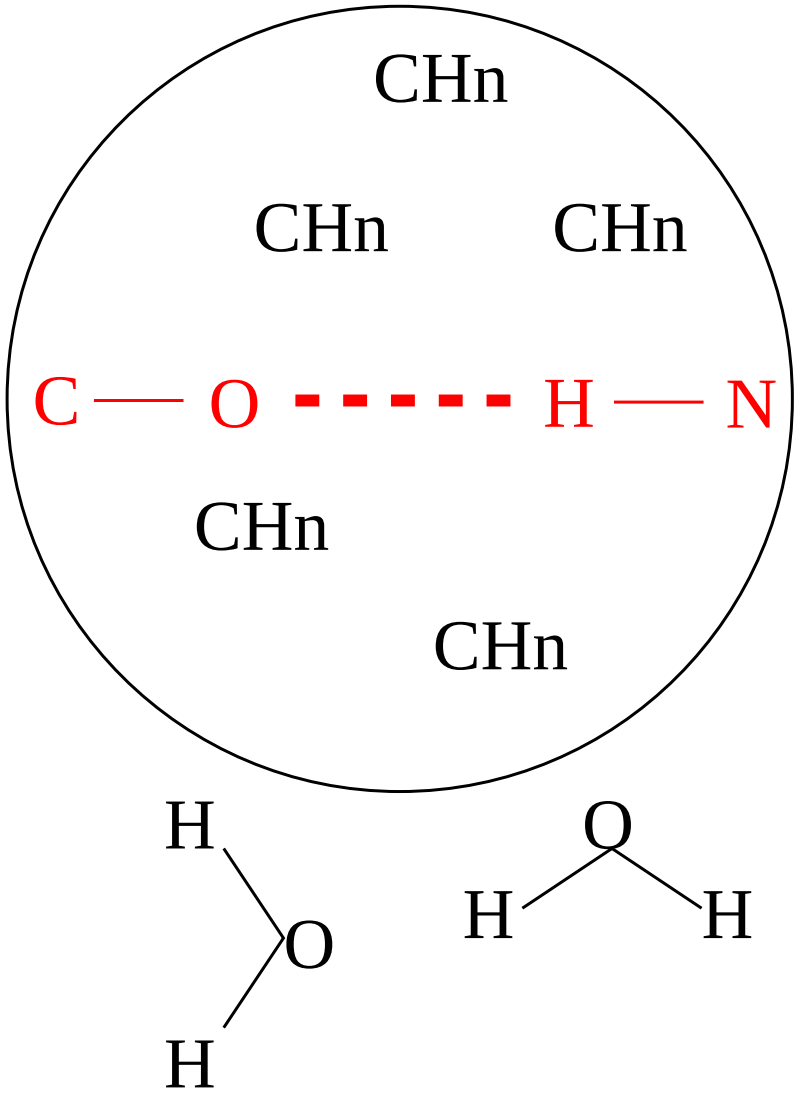
Consider the effect of such a group in

- methyl alchohol versus ethyl alchohol
- ethylene glycol versus propylene glycol
- (deadly versus drinkable)

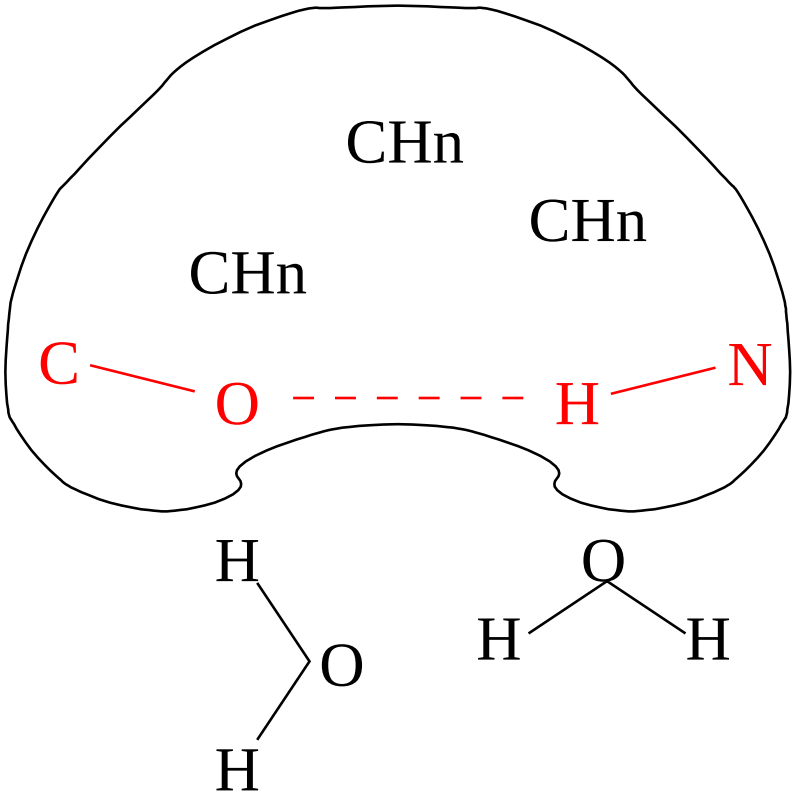
Can we see a molecular-level effect analogous to the change in dielectric permittivity?

What can a simple model of dielectric modulation predict?

Wrapping protects hydrogen bond from water



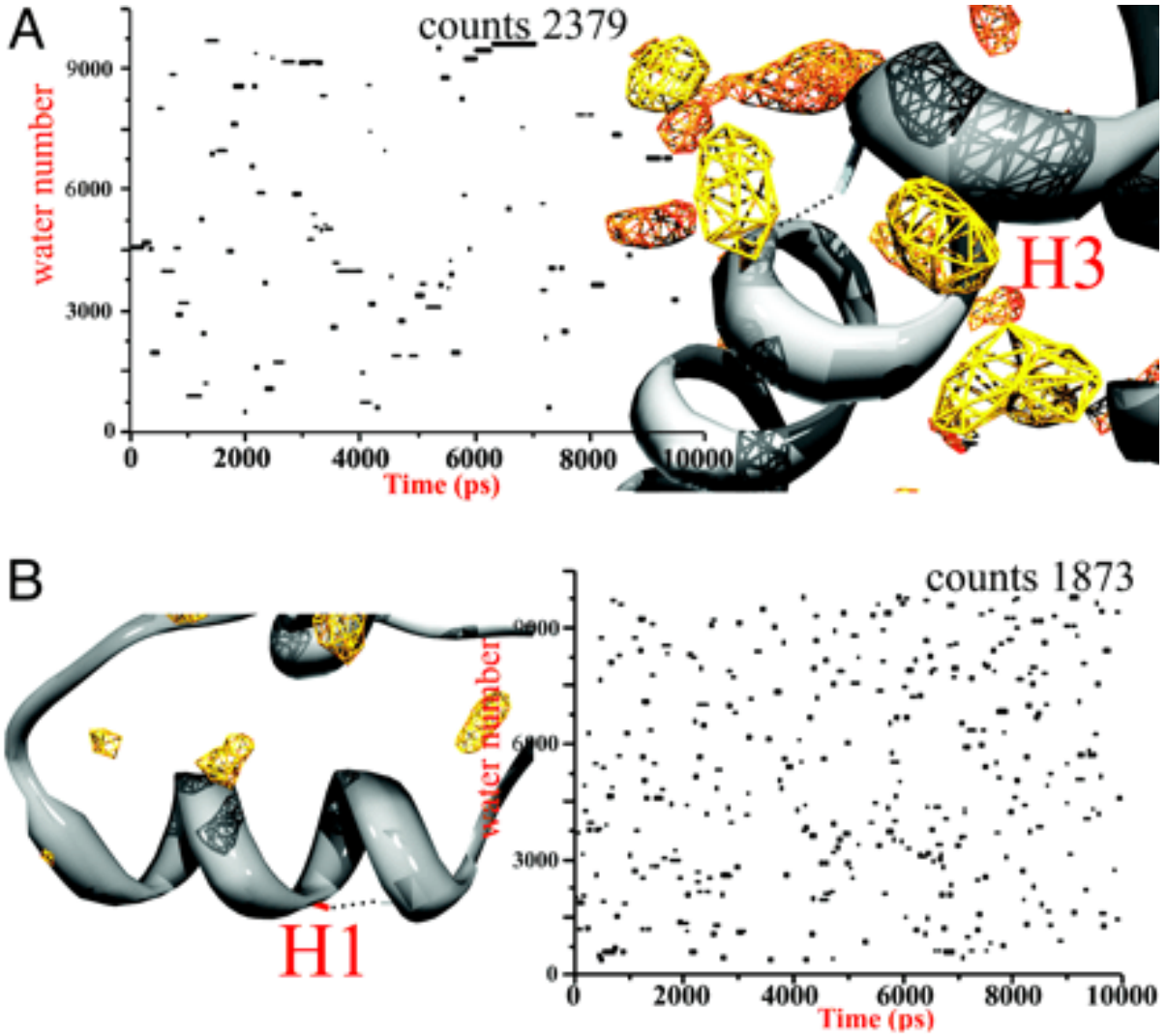
Well wrapped hydrogen bond



Underwrapped hydrogen bond

Extent of wrapping changes nature of hydrogen bond

Hydrogen bonds (B) that are not protected from water do not persist.



From De Simone, et al., PNAS 102 no 21 7535-7540 (2005)

Dynamics of hydrogen bonds and wrapping

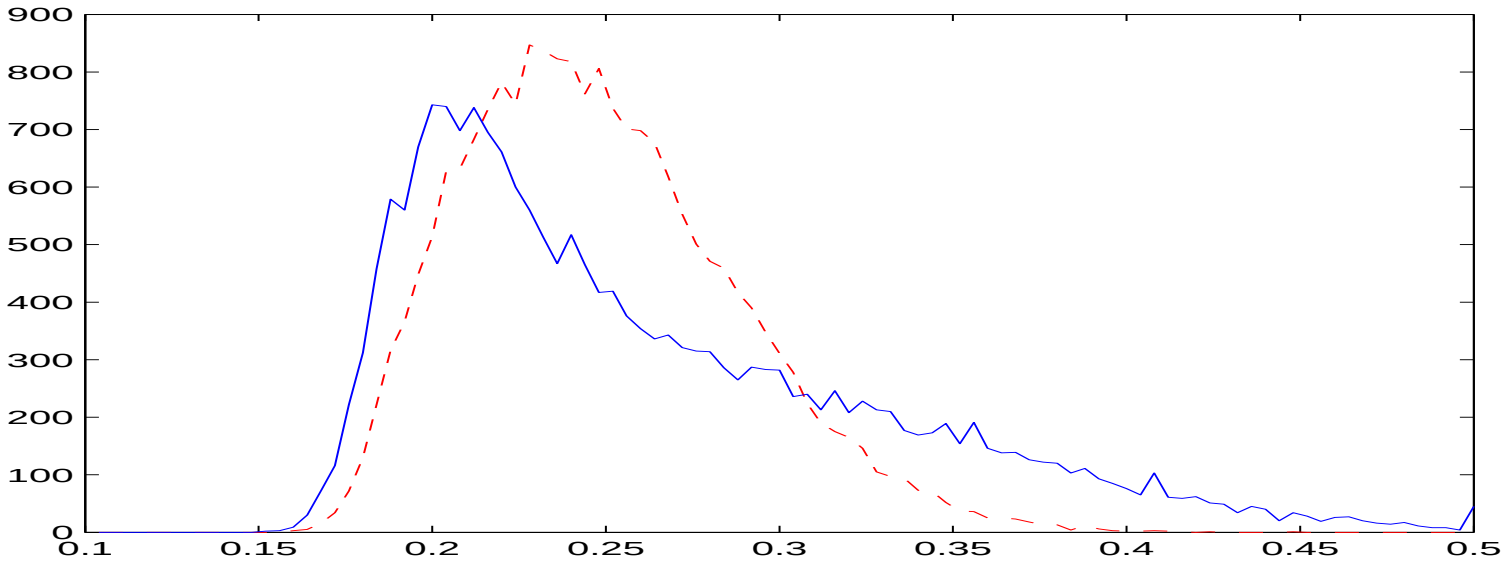
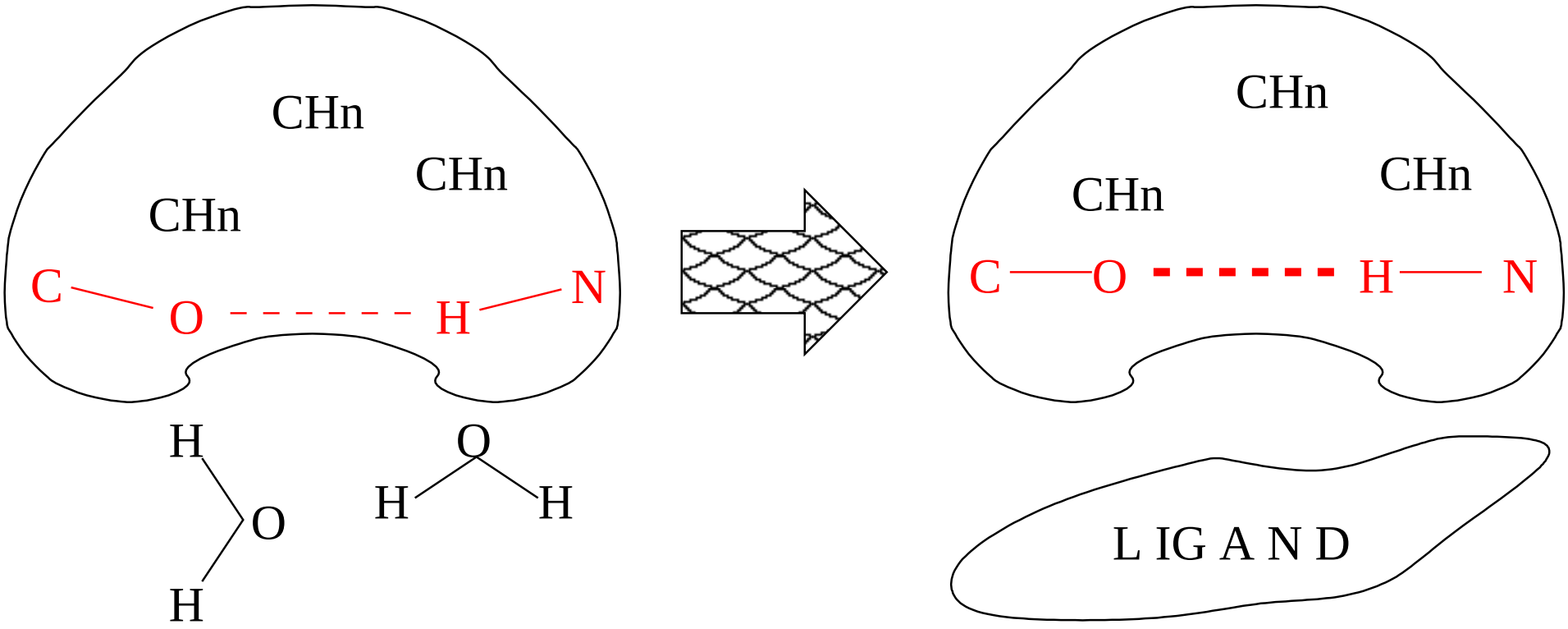


Figure 5: Distribution of bond lengths for two hydrogen bonds formed in a structure of the sheep prion [2]. Horizontal axis measured in nanometers, vertical axis represents numbers of occurrences taken from a simulation with $20,000$ data points with bin widths of 0.1 Ångstrom. Distribution for the well-wrapped hydrogen bond (H3) has smaller mean value but a longer (exponential) tail, whereas distribution for the underwrapped hydrogen bond (H1) has larger mean but Gaussian tail.

Ligand binding removes water



Binding of ligand changes underprotected hydrogen bond (high dielectric) to strong bond (low dielectric)

No intermolecular bonds needed!

Intermolecular bonds are like the power cord on my computer.

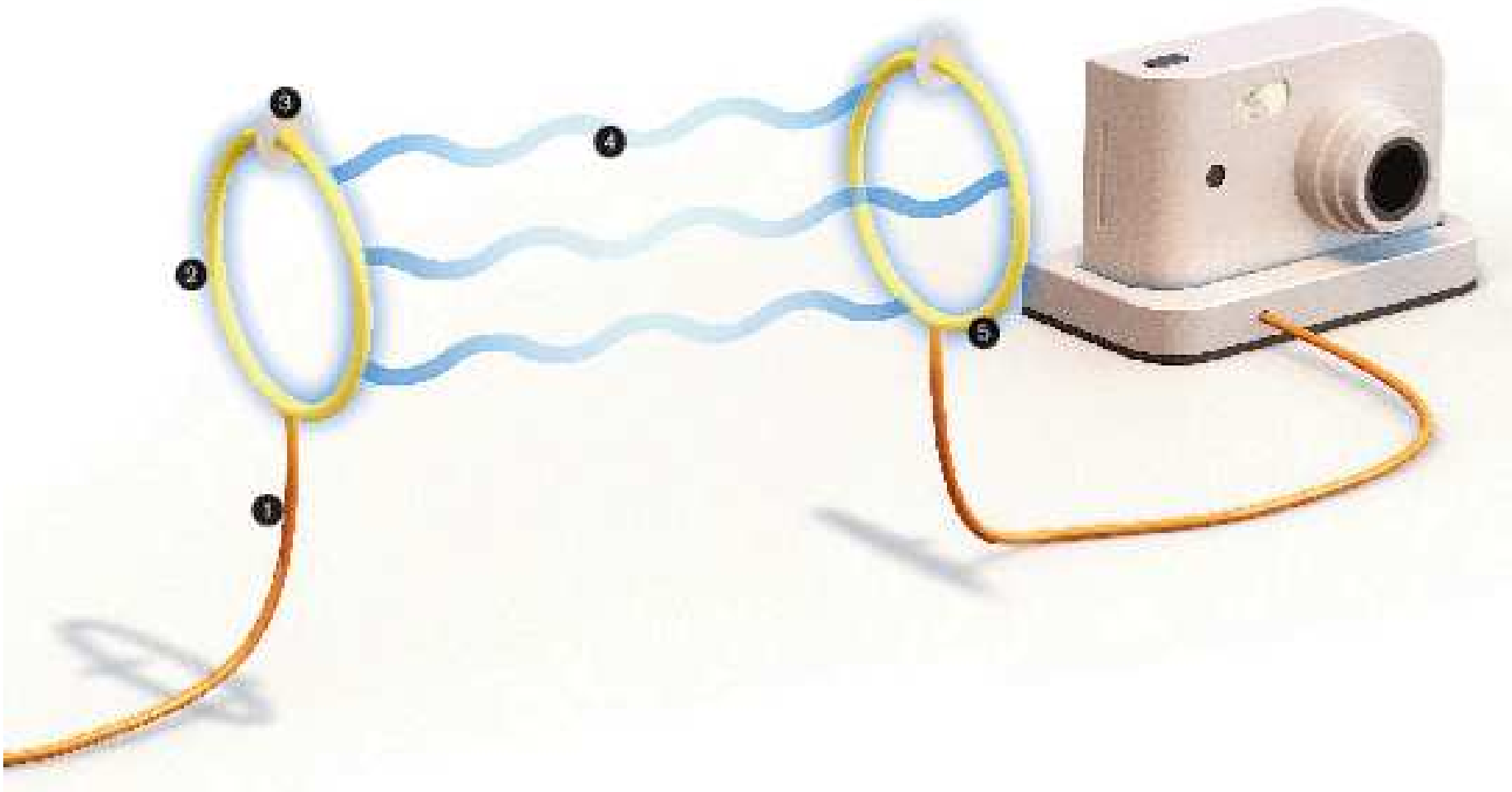
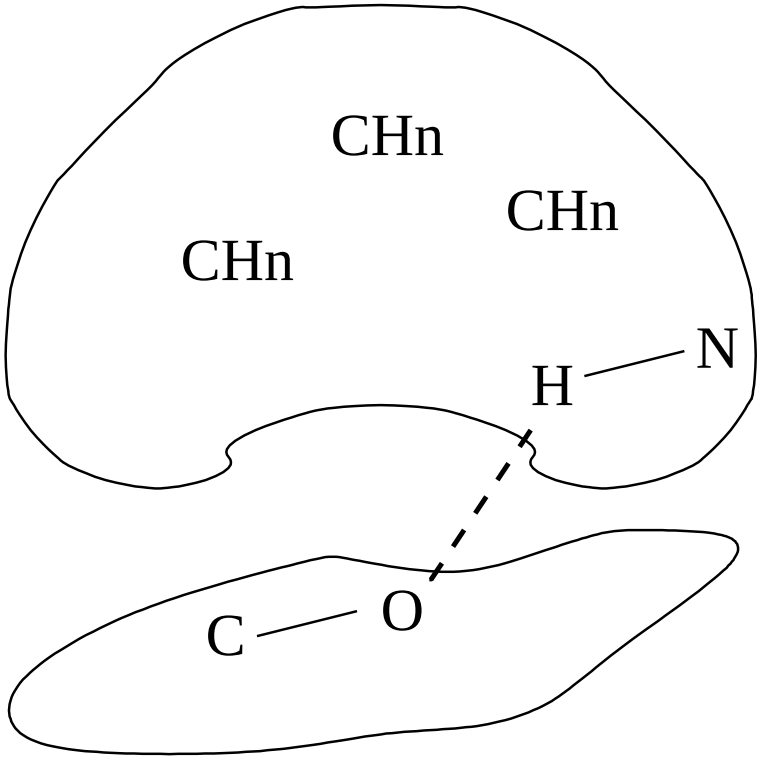
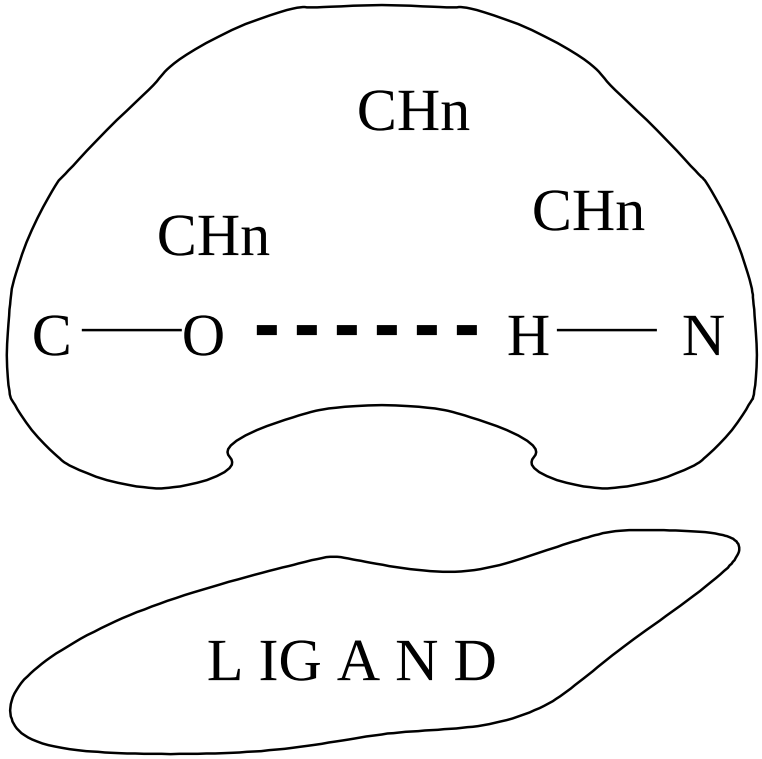


Figure 6: Wireless Charging (from Technology Review).

Intramolecular bonds are like the charger on electric toothbrush.

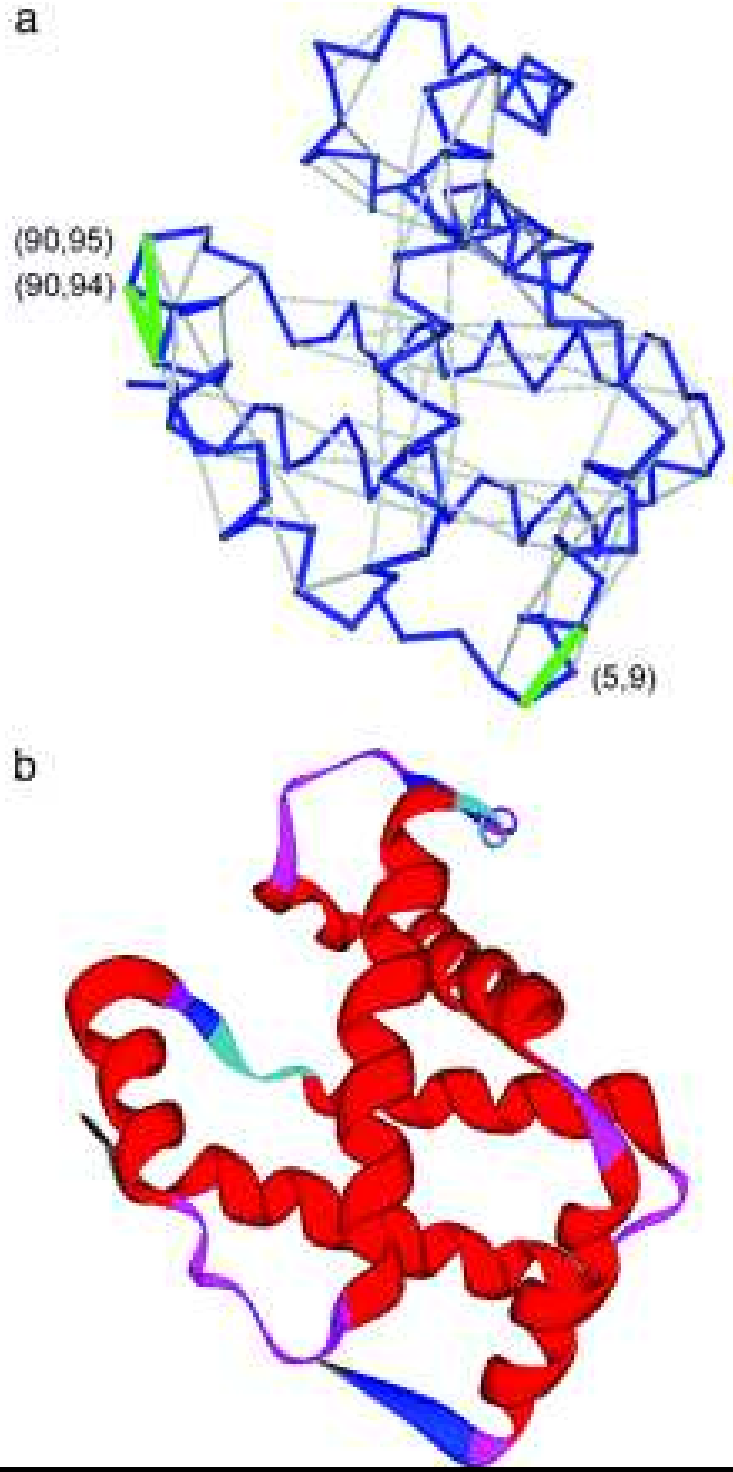
Intermolecular versus intramolecular H bonds





Energetic contribution to binding comparable

but can be better for intramolecular.



Dehydrons

in human hemoglobin, From PNAS 100: 6446-6451 (2003) Ariel Fernandez, Jozsef Kardos, L. Ridgway Scott, Yuji Goto, and R. Stephen Berry. Structural defects and the diagnosis of amyloidogenic propensity.

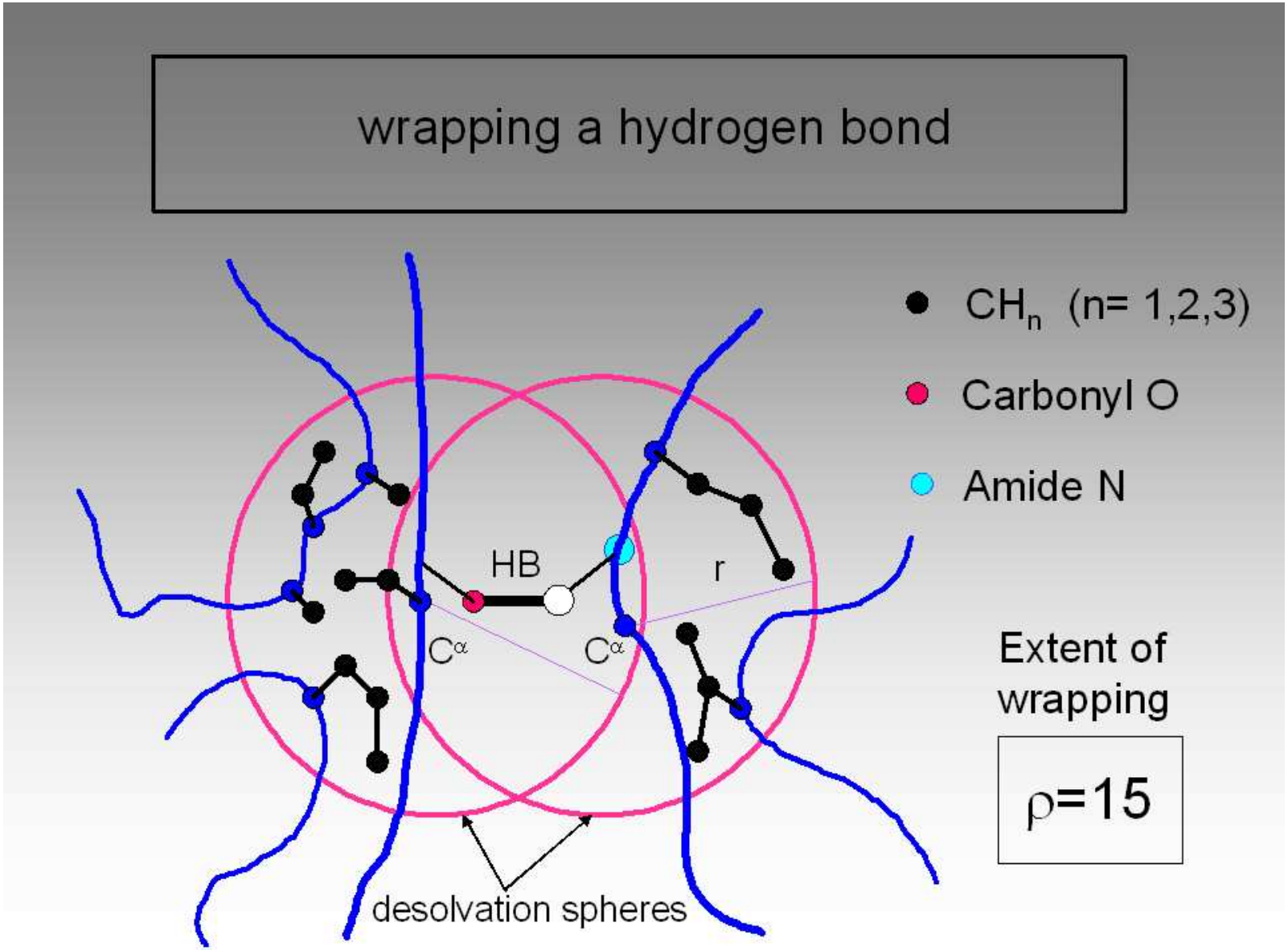
Well-wrapped

hydrogen bonds are

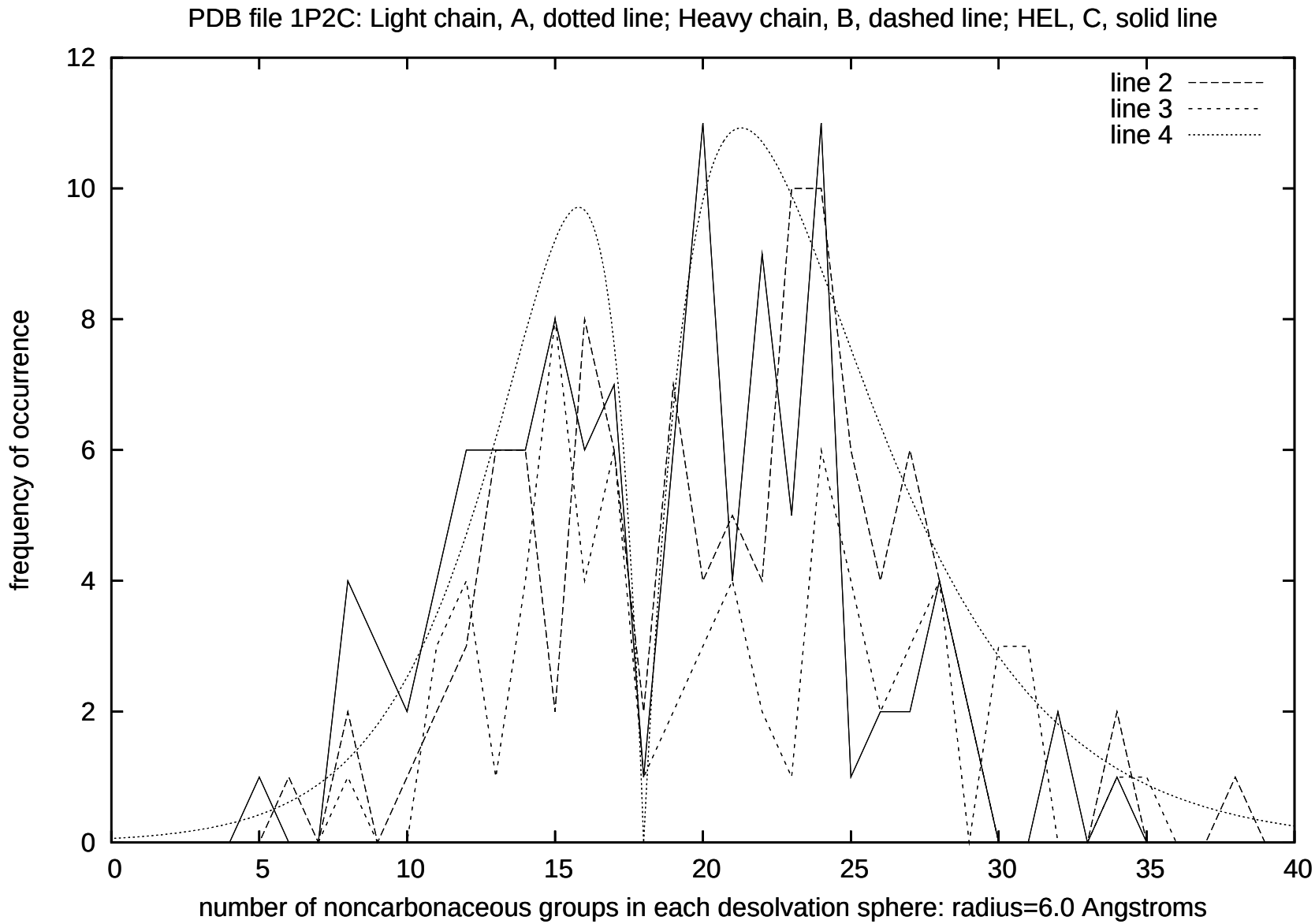
grey, and dehydrons are green.

The standard ribbon model of “structure” lacks indicators of electronic environment.

Wrapping made quantitative by counting carbonaceous groups in the neighborhood of a hydrogen bond.



Distribution of wrapping for an antibody complex.



Stickiness of dehydrons

Attractive force of dehydrons predicted and measured in

Ariel Fernandez and L. Ridgway Scott. Adherence of packing defects in soluble proteins. Phys. Rev. Lett. 2003 91:18102(4)

by considering rates of adhesion to phospholipid (DLPC) bilayer.

Deformation of phospholipid bilayer by dehydrons measured in

Ariel Fernandez and L. Ridgway Scott. Under-wrapped soluble proteins as signals triggering membrane morphology. Journal of Chemical Physics 119(13), 6911-6915 (2003).

Single molecule measurement of dehydronic force in

Ariel Fernandez. Direct nanoscale dehydration of hydrogen bonds. Journal of Physics D: Applied Physics 38, 2928-2932, 2005.

Fine print: careful definition of dehydron requires assessing modification of dielectric enviroment by test hydrophobe. That is, geometry of carbon groups matters, although counting gets it right $\approx$ 90% of the time [3].

Charge-force relationship

Here's the math....

Charges ρ induce an electric field $\mathbf{e} = \nabla\phi$ given by

$$\nabla\cdot(\varepsilon\nabla\phi) = \nabla\cdot(\varepsilon\mathbf{e}) = \rho, \tag{1.1}$$

where ε is the permittivity of the medium. Energy $= \int \rho\phi\, dx$.

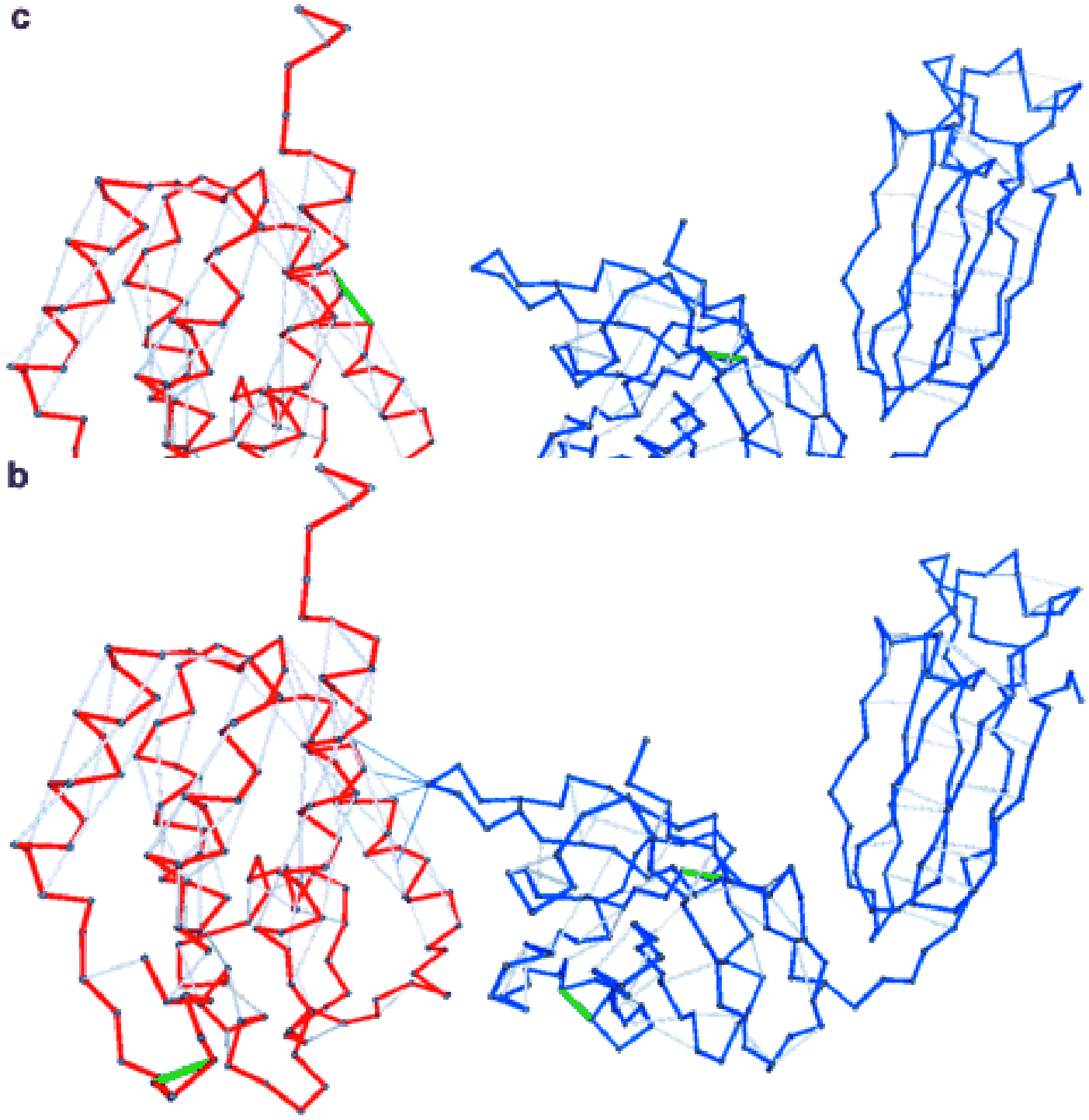
When the medium is a vacuum, ε is the permittivity of free space, ε_0.

In other media (e.g., water) the value of ε is much larger.

The quantity ε measures the strength of the dielectric enviroment.

Water removal decreases the coefficient ε in (1.1), and increases ϕ.

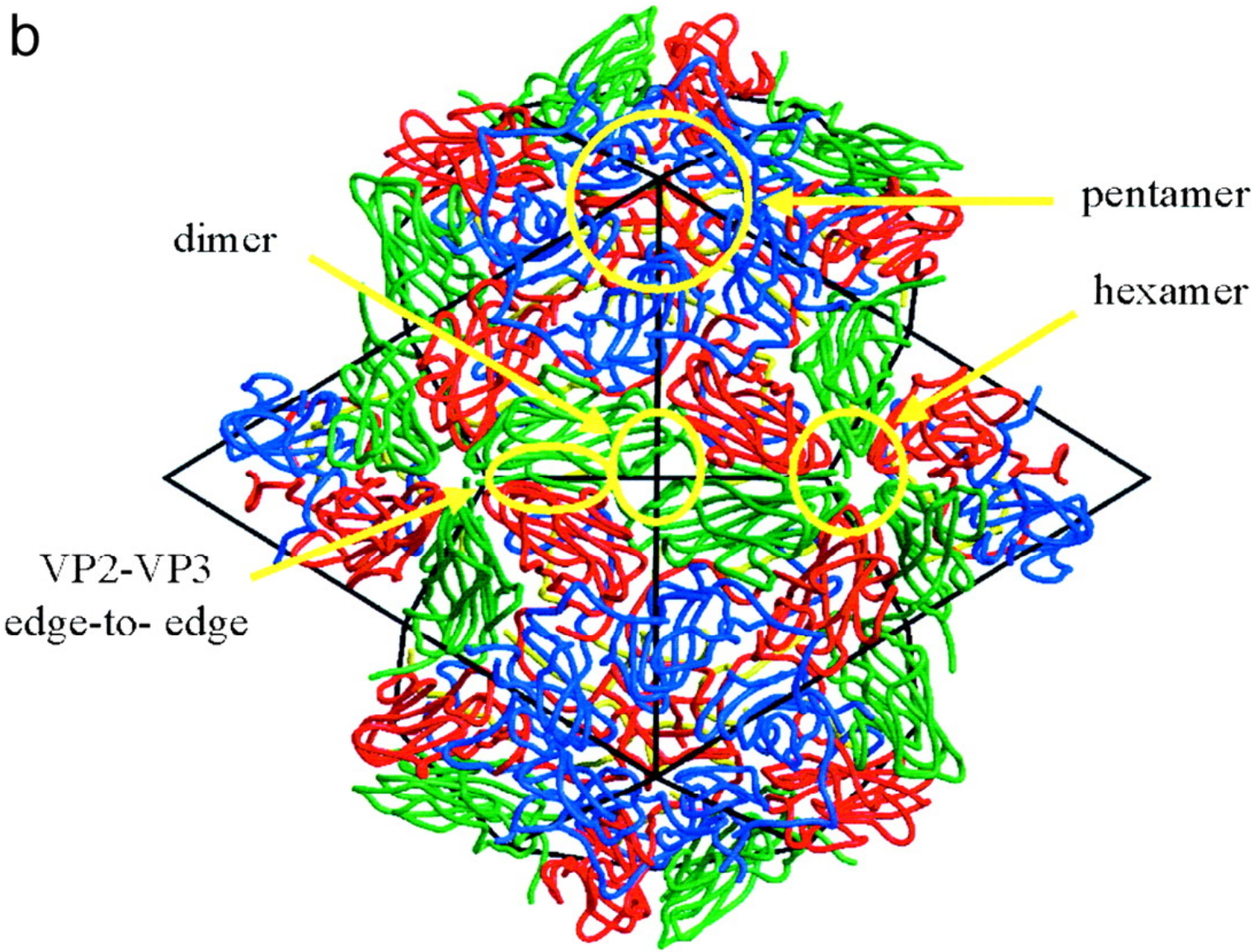
Hydrophilic groups contribute to the right-hand side ρ in (1.1).



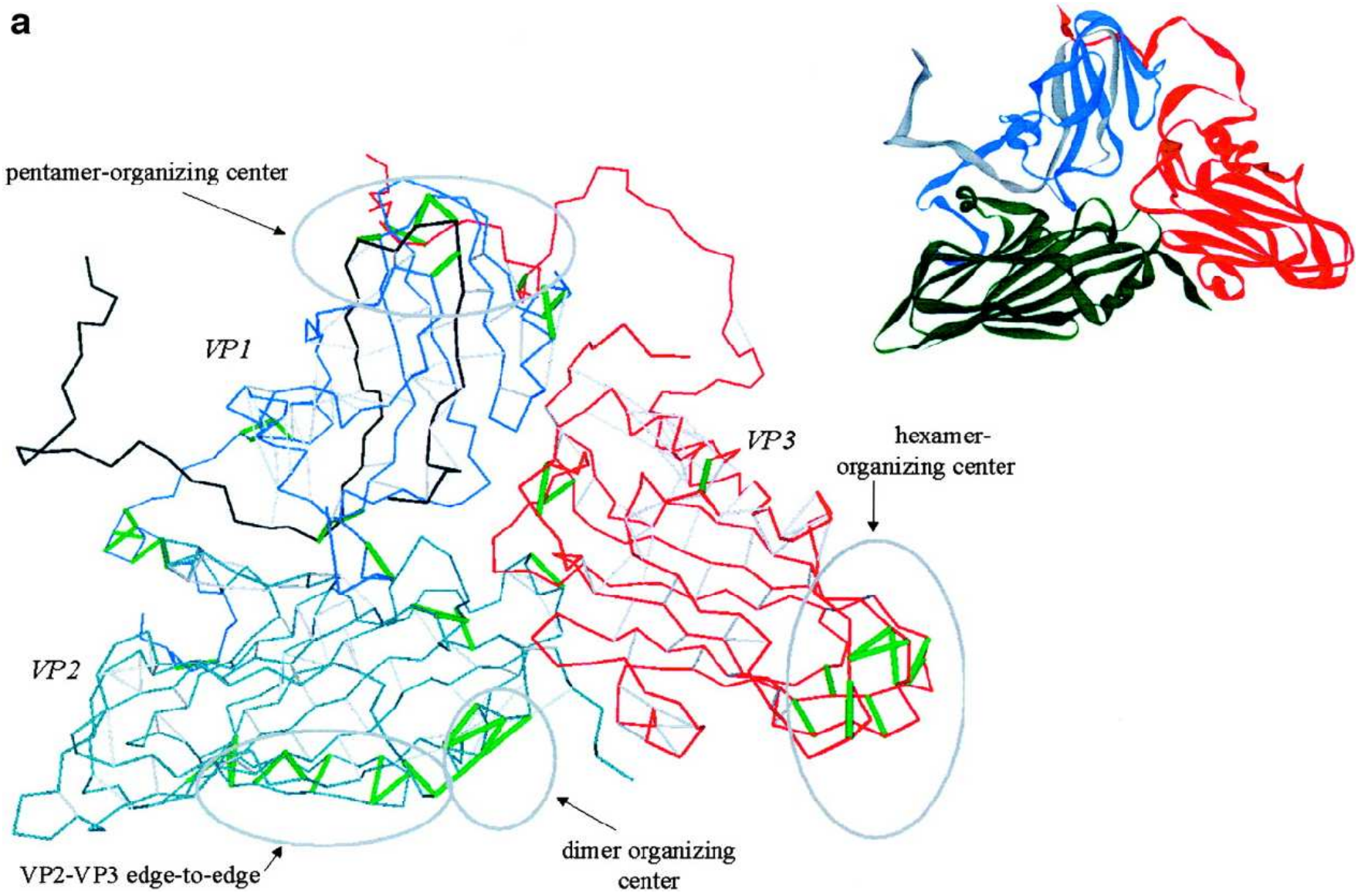
The HIV protease has a dehydron at an antibody binding site.

When the antibody binds at the dehydron, it wraps it with hydrophobic groups.

A model for protein-protein interaction



Foot-and-mouth disease virus assembly from small proteins.



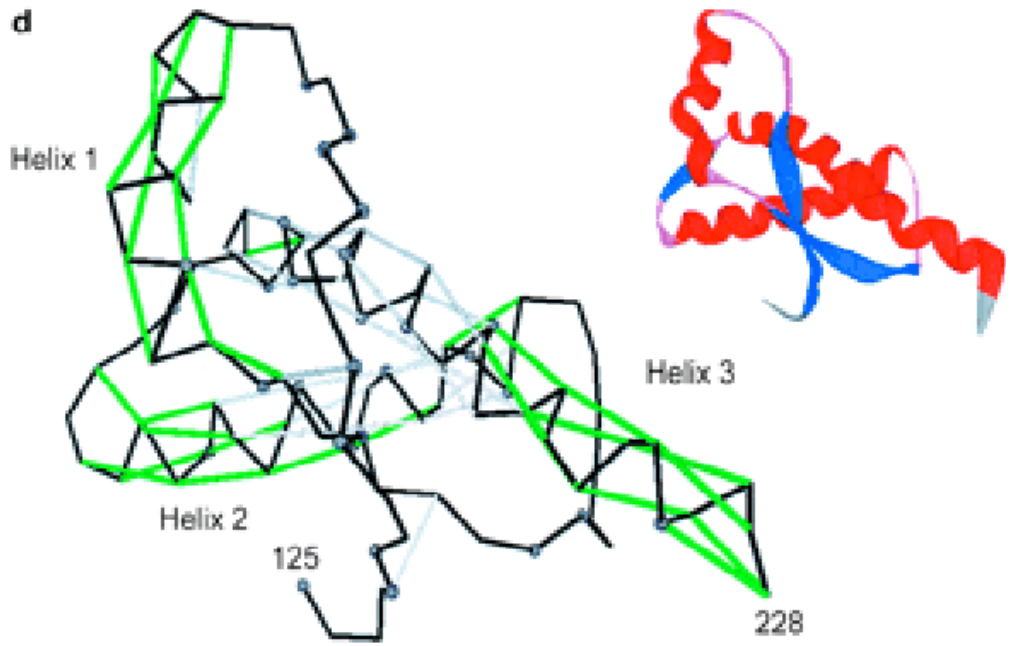
Dehydrons guide binding of component proteins VP1, VP2 and VP3 of foot-and-mouth disease virus.

2 Extreme interaction: amyloid formation

Standard application of bioinformatics: look at distribution tails.

If some is good, more may be better, but too many may be bad.

Too many dehydrons signals trouble: the human prion.



From PNAS 100: 6446-6451 (2003) Ariel Fernandez, Jozsef Kardos, L. Ridgway Scott, Yuji Goto, and R. Stephen Berry. Structural defects and the diagnosis of amyloidogenic propensity.

3 Dehydrons as indicators of protein interactivity

If dehydrons provide mechanism for proteins to interact, then more interactive proteins should have more dehydrons, and vice versa.

We only expect a correlation since there are (presumably) other ways for proteins to interact.

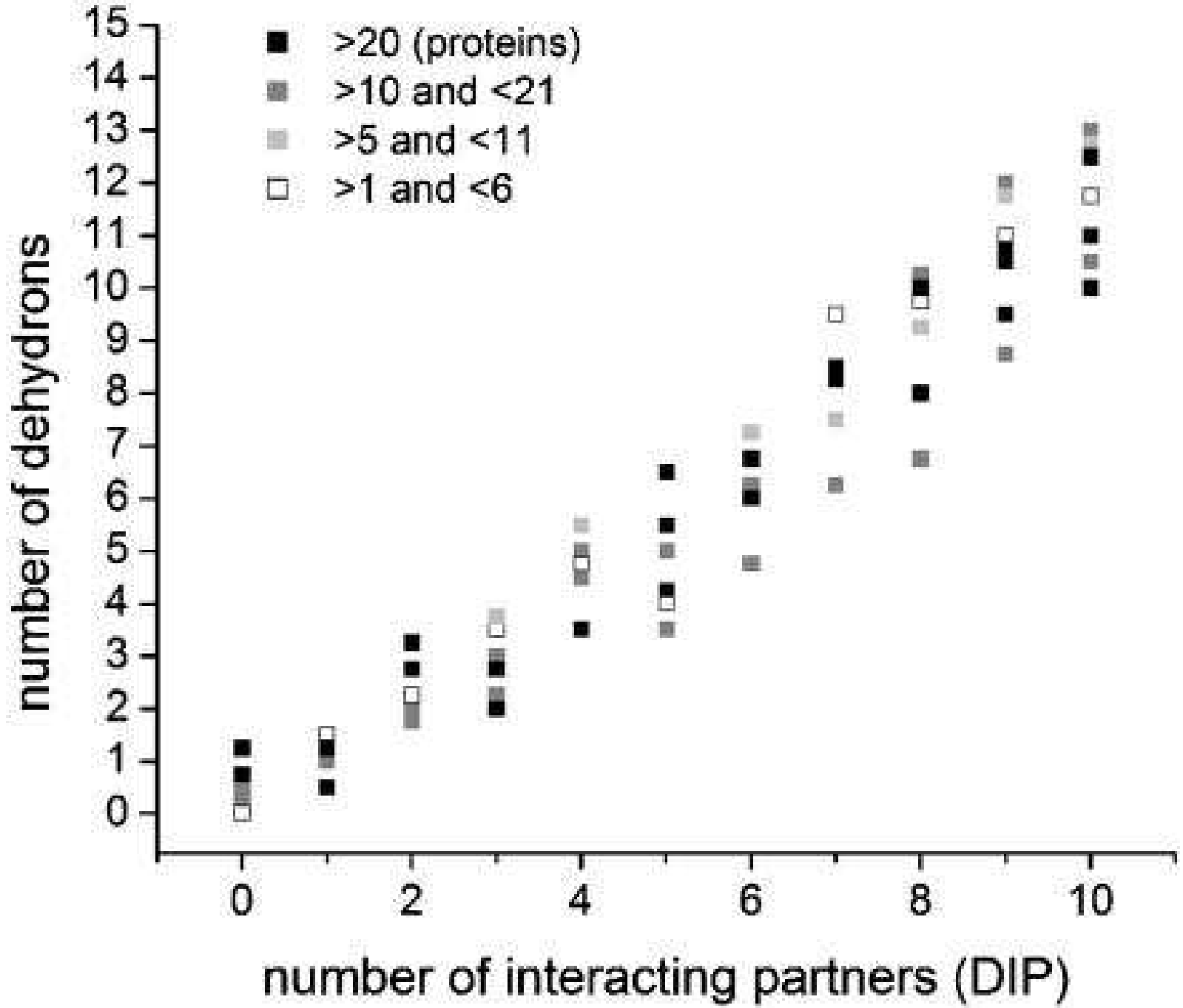
The DIP database collects information about protein interactions, based on individual protein domains: can measure interactivity of different regions of a given protein.

Result: Interactivity of proteins correlates strongly with number of dehydrons.

PNAS 101(9):2823-7 (2004)

The nonconserved wrapping of conserved protein folds reveals a trend toward increasing connectivity in proteomic networks.

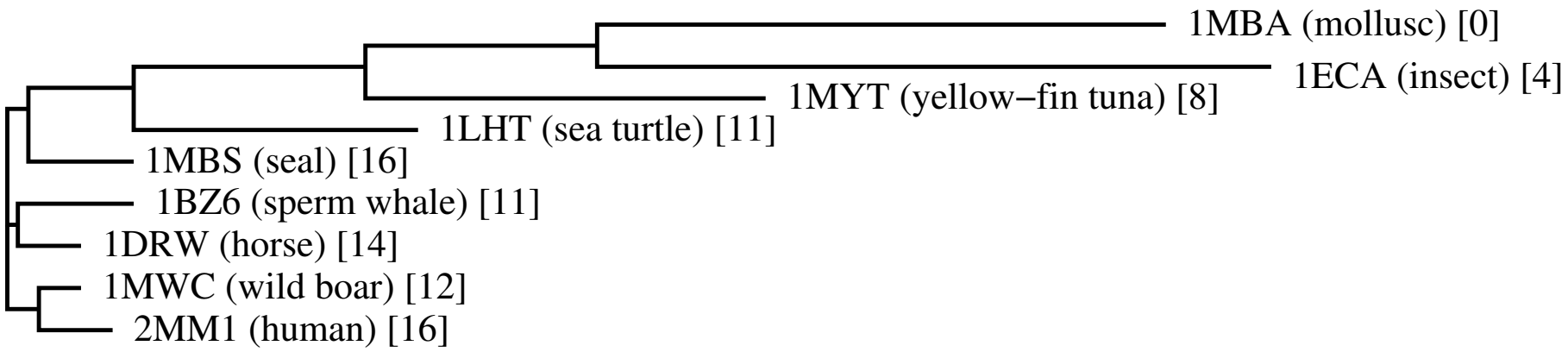
Ariel Fernández, L. R. Scott and R. Steve Berry

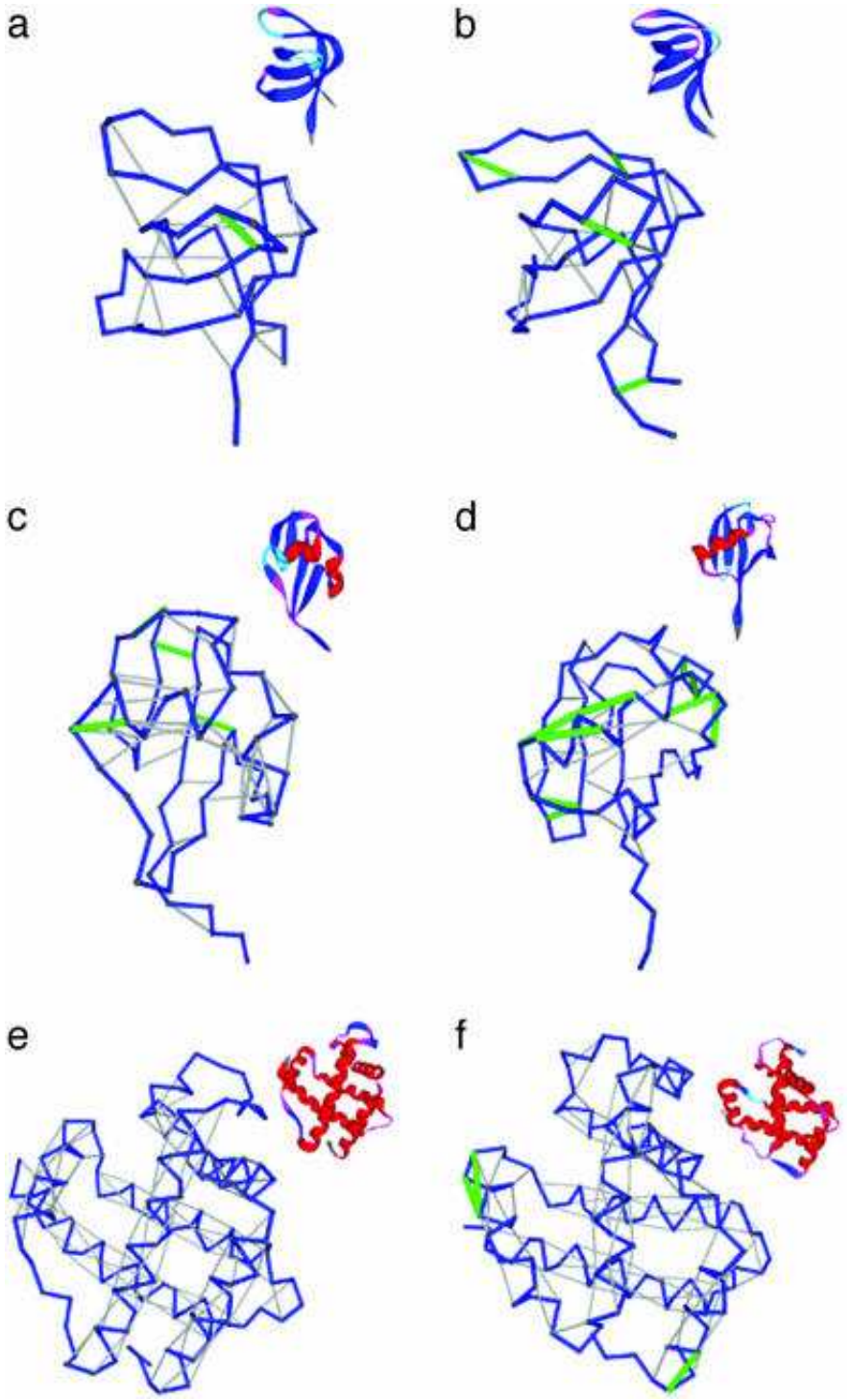

>20 (proteins)
>10 and <21
>5 and <11
>1 and <6
number of dehydrons
number of interacting partners (DIP)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
0
2
4
6
8
10

4 Dehydron variation over different species

Species (common name)	peptides	H bonds	dehydrons
Aplysia limacina (mollusc)	146	106	0
Chironomus thummi thummi (insect)	136	101	3
Thunnus albacares (tuna)	146	110	8
Caretta caretta (sea turtle)	153	110	11
Physeter catodon (whale)	153	113	11
Sus scrofa (pig)	153	113	12
Equus caballus (horse)	152	112	14
Elephas maximus (Asian elephant)	153	115	15
Phoca vitulina (seal)	153	109	16
H. sapiens (human)	146	102	16

Number of dehydrons in Myoglobin of different species





Anecdotal evidence:

the basic
structure is similar, just the
number of dehydrons increases.

SH3 domains are from

nematode C. elegans (a)

H. sapiens (b);

ubiquitin is from
E. coli (c) and H. sapiens (d);

hemoglobin
is from Paramecium
(e). and H. sapiens-subunit (f).

Genetic code

Genetic code minimizes changes of polarity due to single-letter codon mutations, but it facilitates changes in wrapping due to single-letter codon mutations.

First Position	Second Position: u	Second Position: c	Second Position: a	Second Position: g	Third Position
u	uuu, uuc: Phe 7 ||	ucu, ucc, uca, ucg: Ser 0 +−	uau, uac: Tyr 6 +−	ugu, ugc: Cys 0 +−	u, c
	uua, uug: Leu 4 ||		uaa: stop; uag: stop	uga: stop; ugg: Trp 7 +−	a, g
c	cuu, cuc, cua, cug: Leu 4 ||	ccu, ccc, cca, ccg: Pro 2 ||	cau, cac: His 1 +−	cgu, cgc, cga, cgg: Arg 2 ++	u, c, a, g
			caa, cag: Gln 2 +−		
a	auu, auc, aua: Ile 4 ||	acu, acc, aca, acg: Thr 1 +−	aau, aac: Asn 1 +−	agu, agc: Ser 0 +−	u, c
	aug: Met 1 +−		aaa, aag: Lys 3 ++	aga, agg: Arg 2 ++	a, g
g	guu, guc, gua, gug: Val 3 ||	gcu, gcc, gca, gcg: Ala 1 ||	gau, gac: Asp 1 −−	ggu, ggc, gga, ggg: Gly 0 +−	u, c, a, g
			gaa, gag: Glu 2 −−		

First digit after residue name is amount of wrapping. Second indicator is polarity; $||$: nonpolar, $+-$: polar, $--$: negatively charged, $++$: positively charged.

5 Wrapping technology in drug design

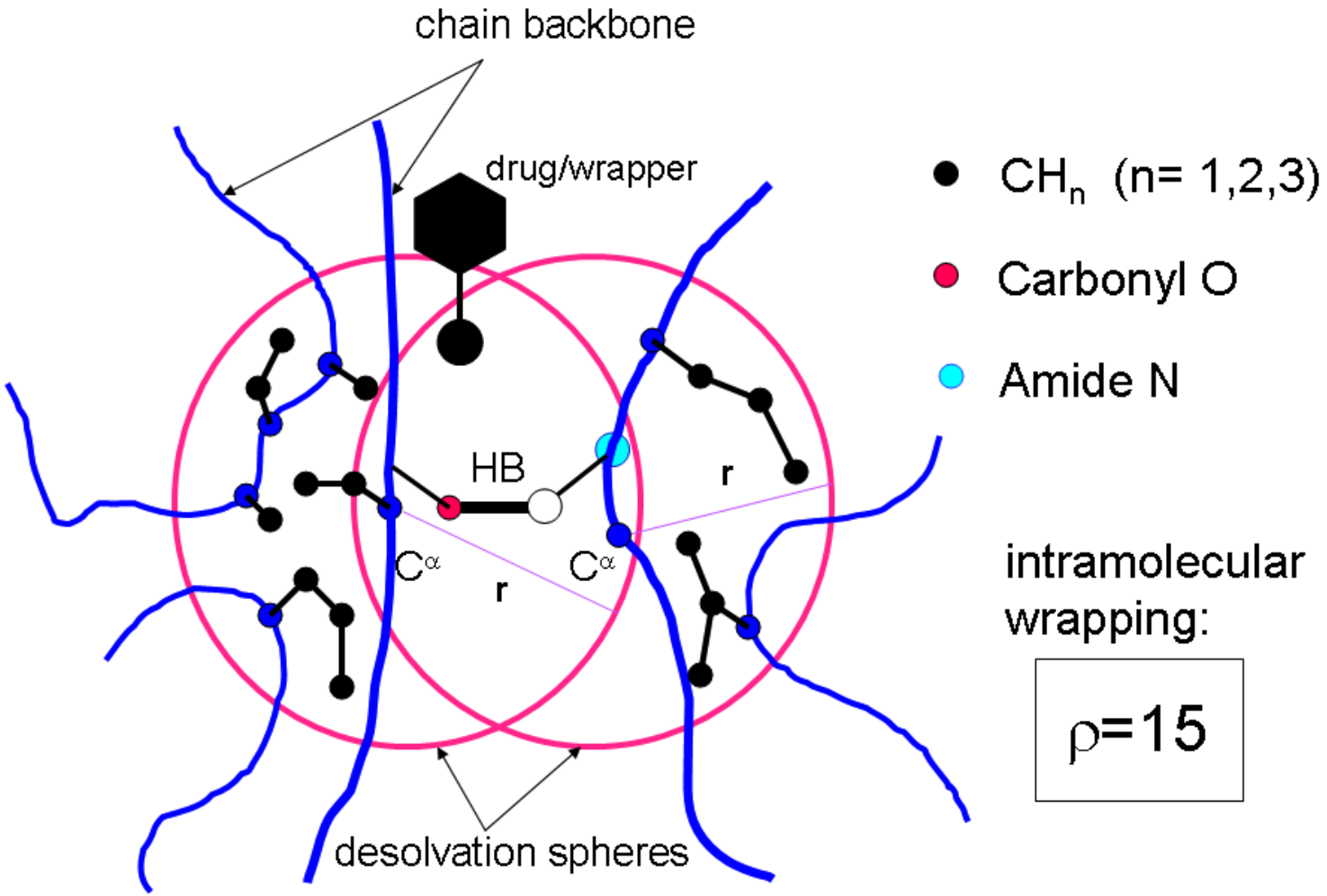
Synopsis of “Modulating drug impact by wrapping target proteins” by Ariel Fernández and L. Ridgway Scott, *Expert Opinion on Drug Discovery* 2007.

Drug ligands often bind to proteins near dehydrons, enhancing their wrapping upon attachment.

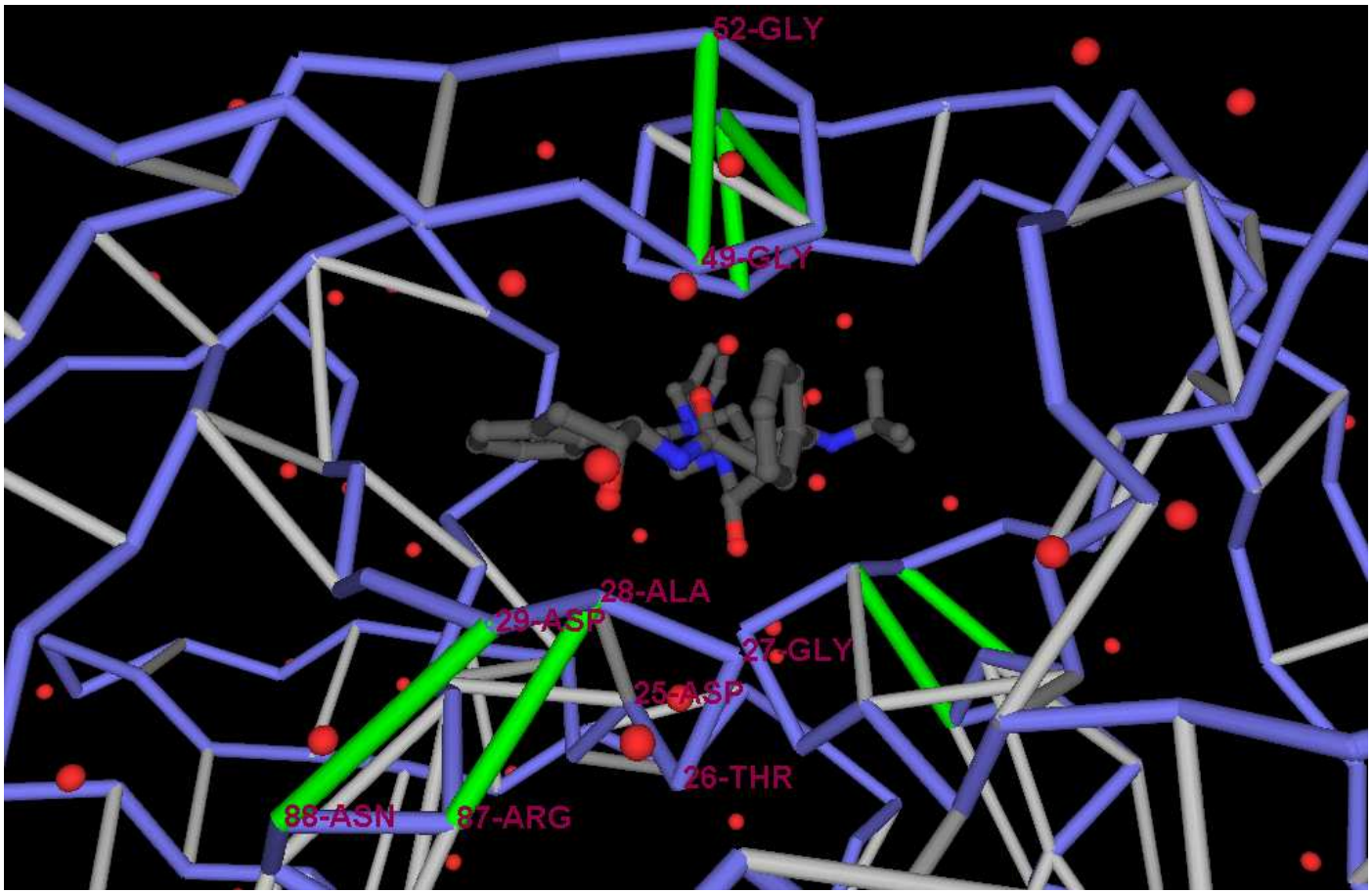
Drug side effects often caused by binding to proteins with structure similar to target.

We can exploit the differences in dehydron patterns in homologous proteins to make drugs more specific.

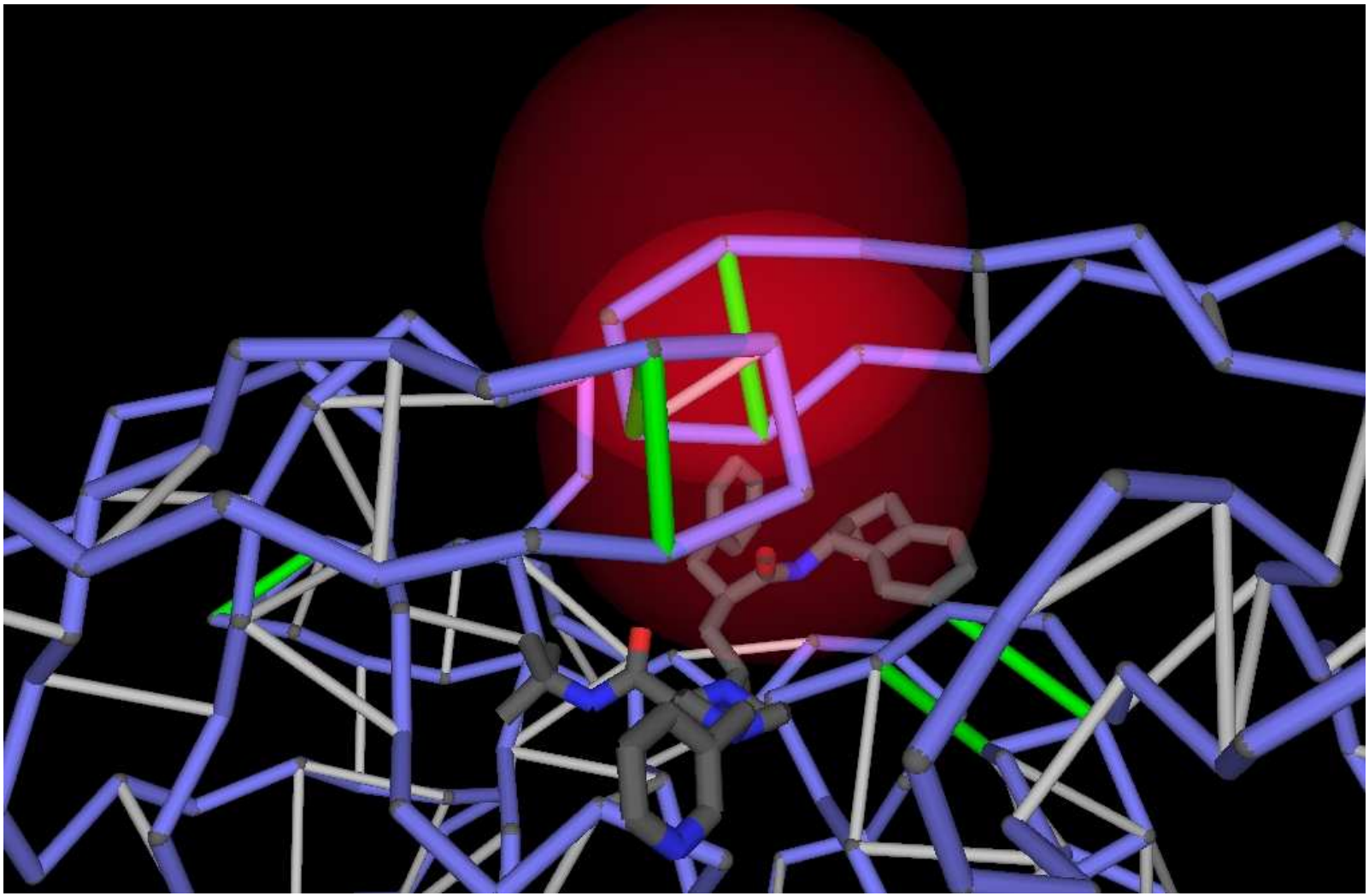
Drug ligand provides additional non-polar carbonaceous group(s) in the desolvation domain, enhancing the wrapping of a hydrogen bond.



HIV-1 protease with ‘dehydron wrapper’ inhibitor



Detail of the protease cavity, pattern of packing defects, and inhibitor positioned as dehydron wrapper.



Desolvation spheres for flap Gly-49–Gly-52 dehydron containing nonpolar groups of the wrapping inhibitor.

Drug specificity

Tyrosine kinases: a family of proteins with very similar structure.

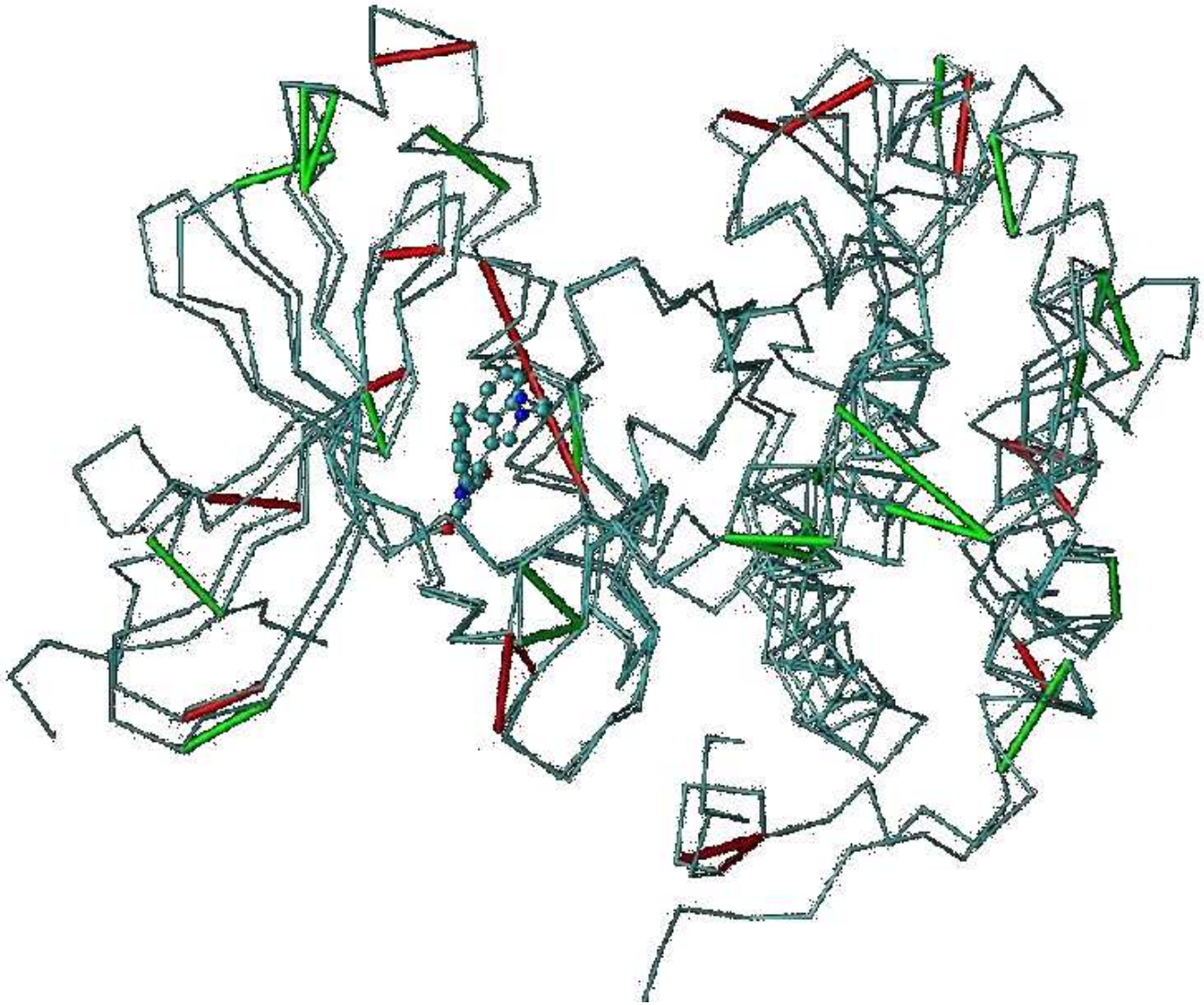
- They are called paralogous because they are similar proteins within a given species.
- These are presumed to have evolved from a common source.
- They are a crucial target of cancer drug therapy.

Gleevec targets particular tyrosine kinases and has been one of the most successful cancer drugs.

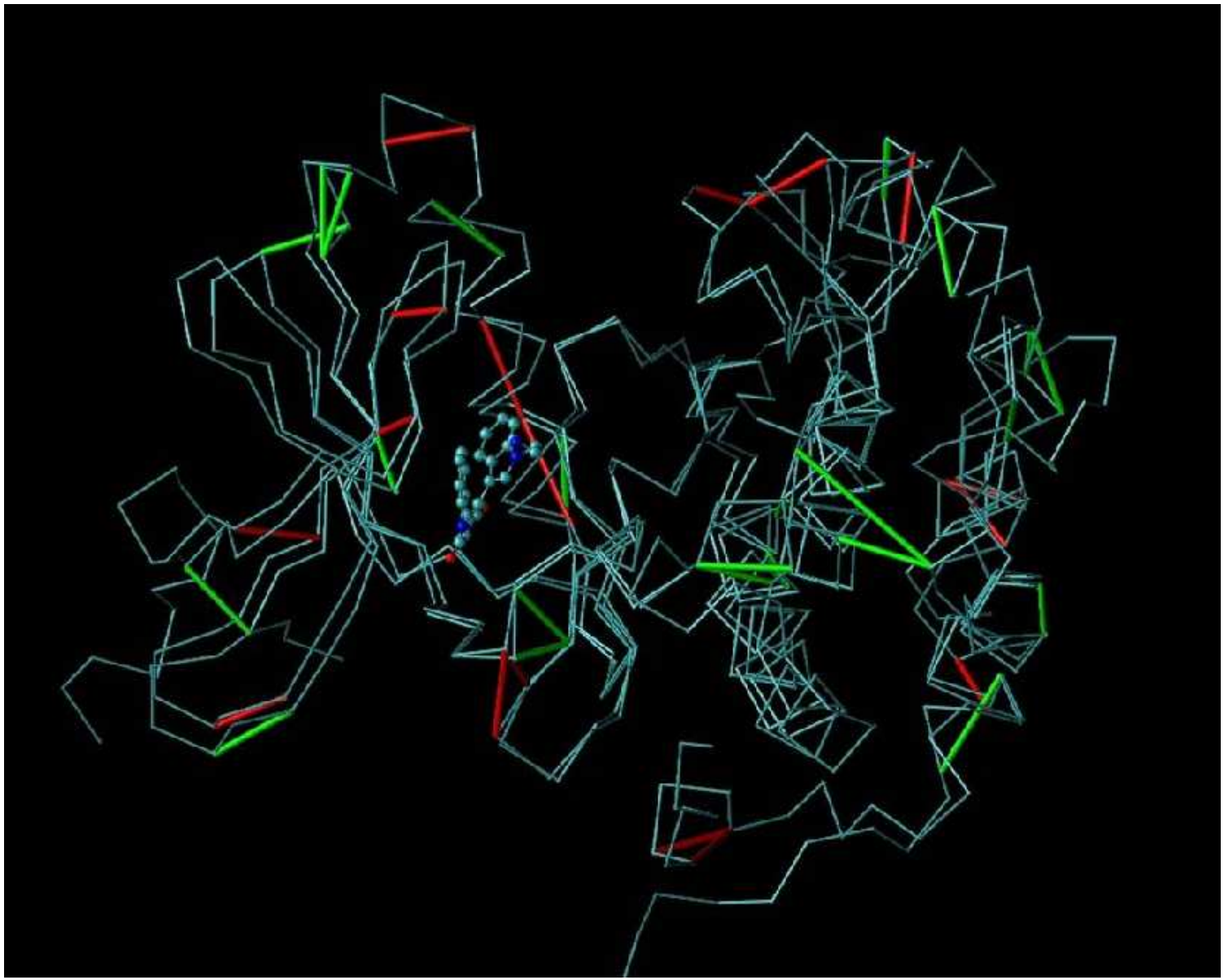
However, it also targets similar proteins and can cause unwanted side effects (it is cardiotoxic).

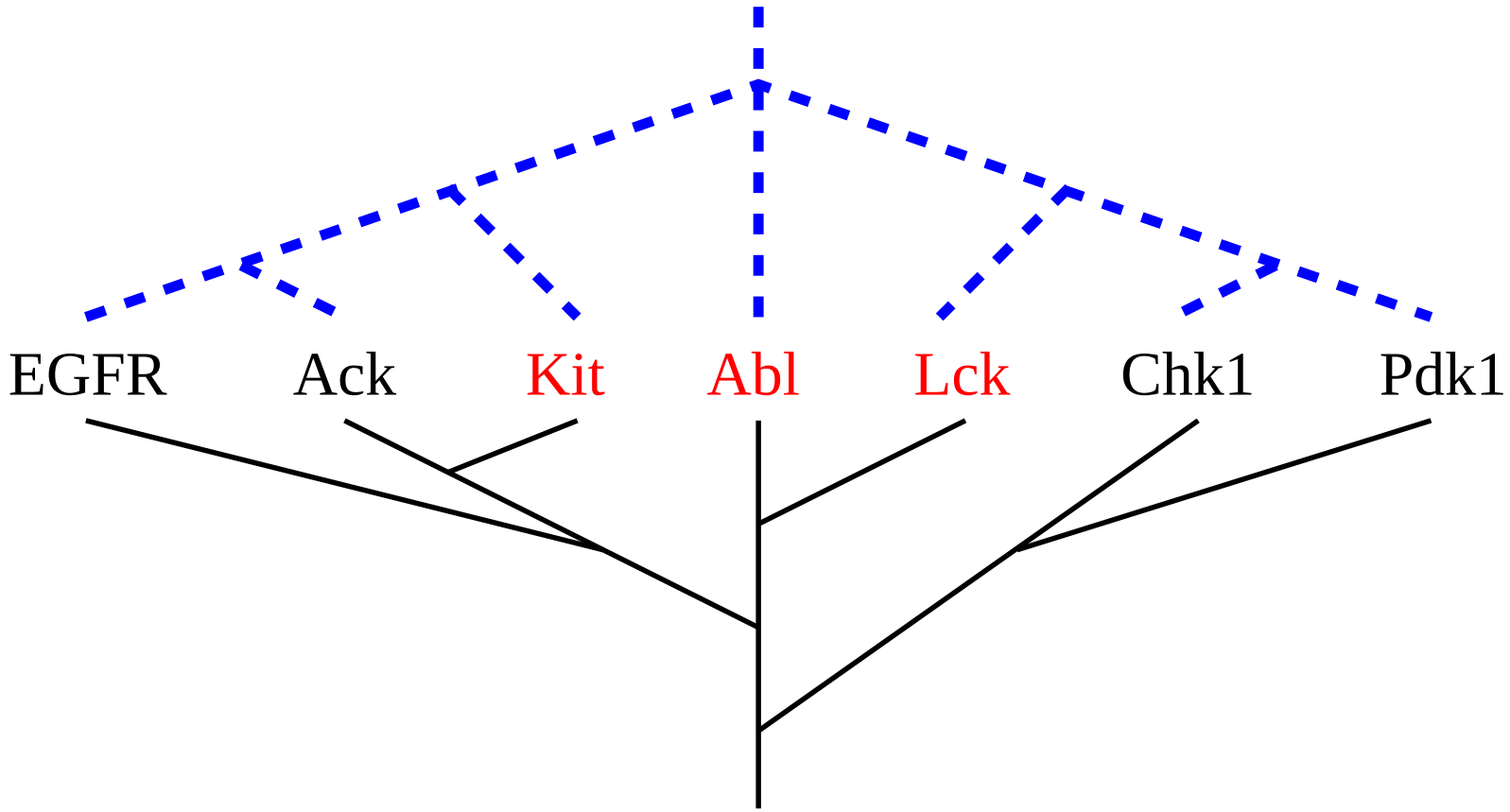
Differences between the dehydron patters in similar proteins can be used to differentiate them and guide the re-design of drug ligands.

Aligned backbones for two paralog kinases; dehydrons for Chk1 are marked in green and those for Pdk1 are in red.

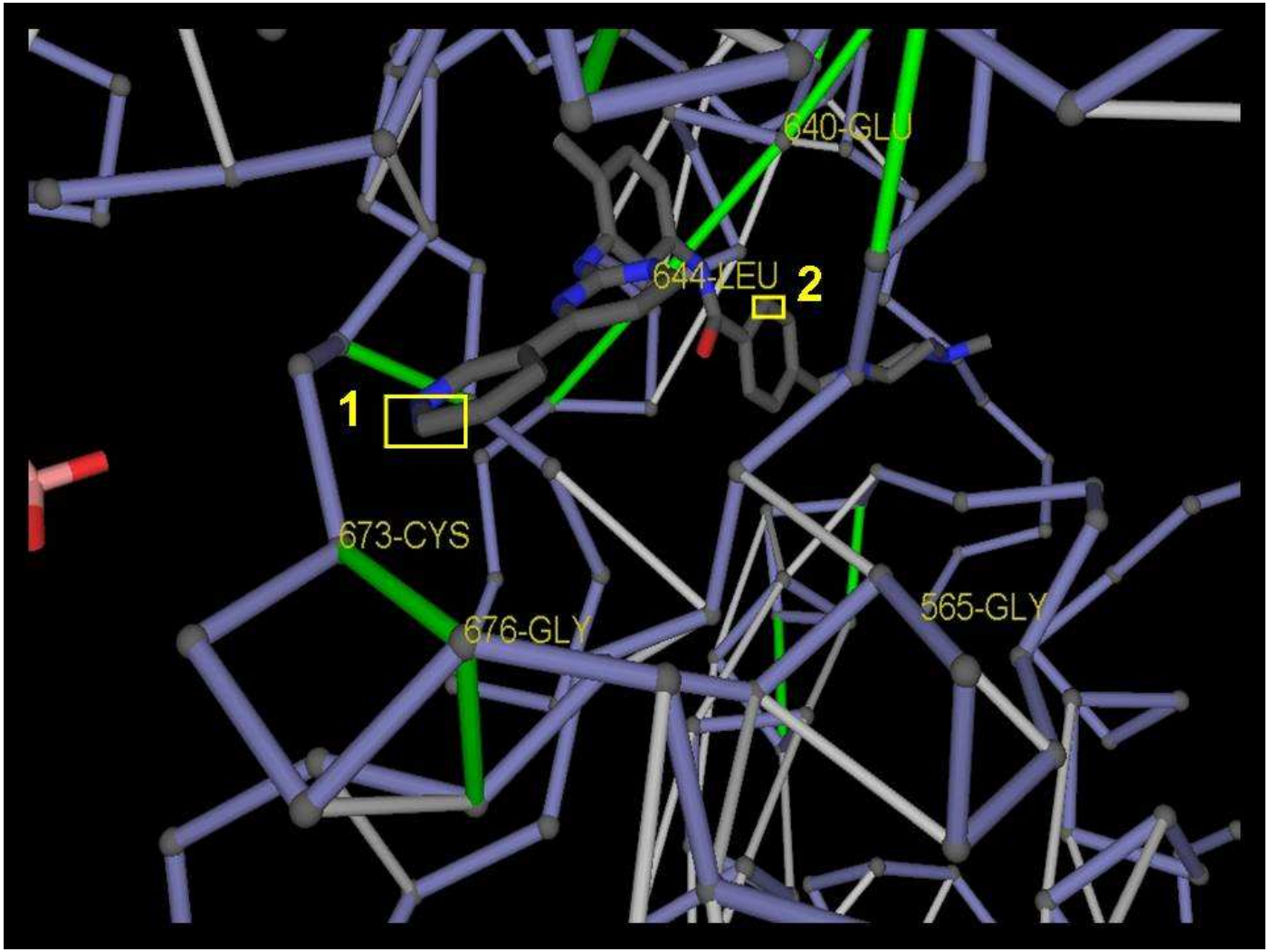


Aligned backbones for two paralog kinases; dehydrons for Chk1 are marked in green and those for Pdk1 are in red.

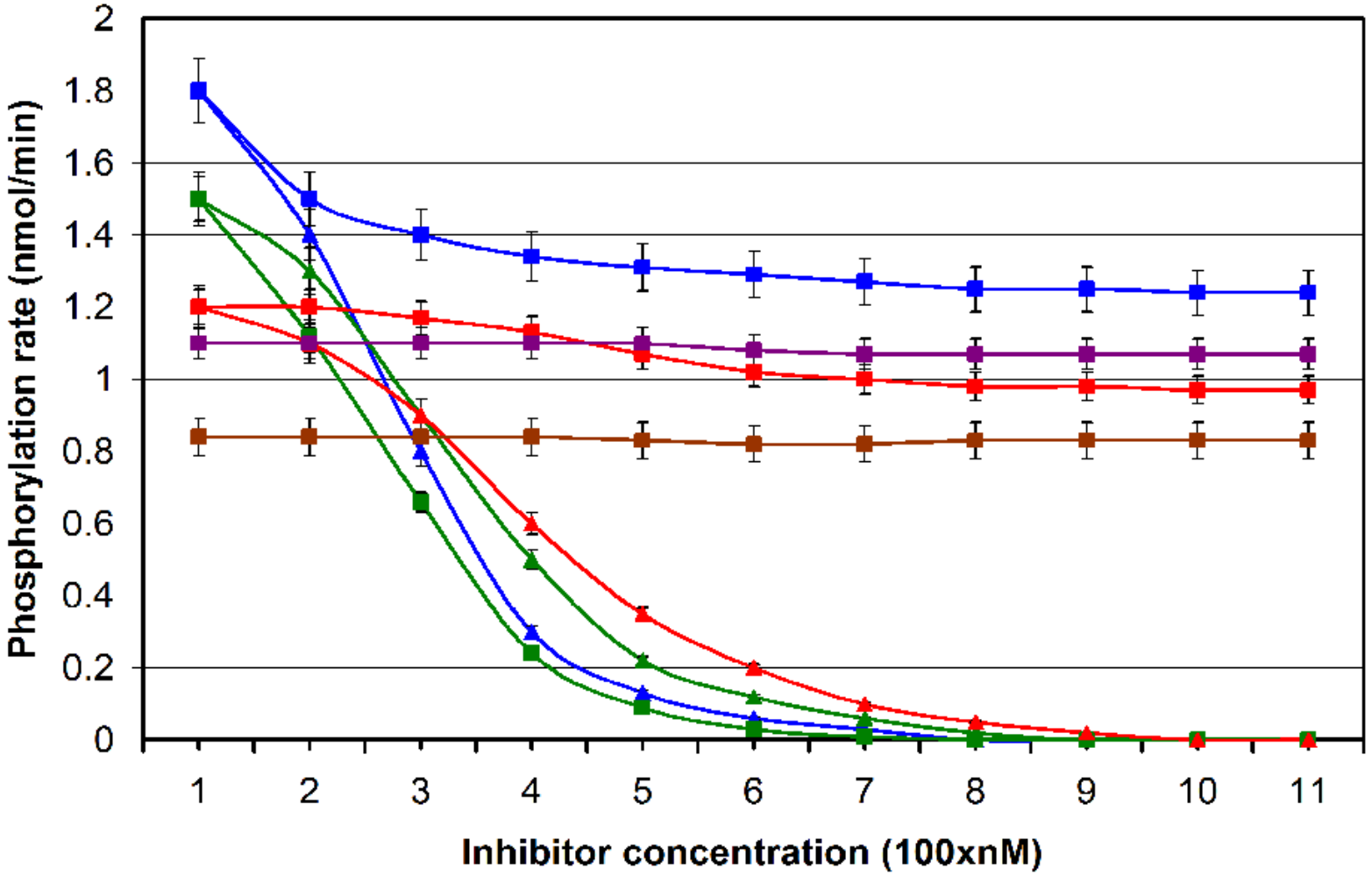




Packing similarity tree (PST, bottom in black) for the seven structurally aligned paralogs of Bcr-Abl. The PST restricted to the alignments of the Gleevec wrapped region in Bcr-Abl is shown (top) with blue dashed lines. The paralogs in red have the most similar packing in the region that aligns with the Gleevec wrapped region in Bcr-Abl and are also primary targets of this inhibitor.

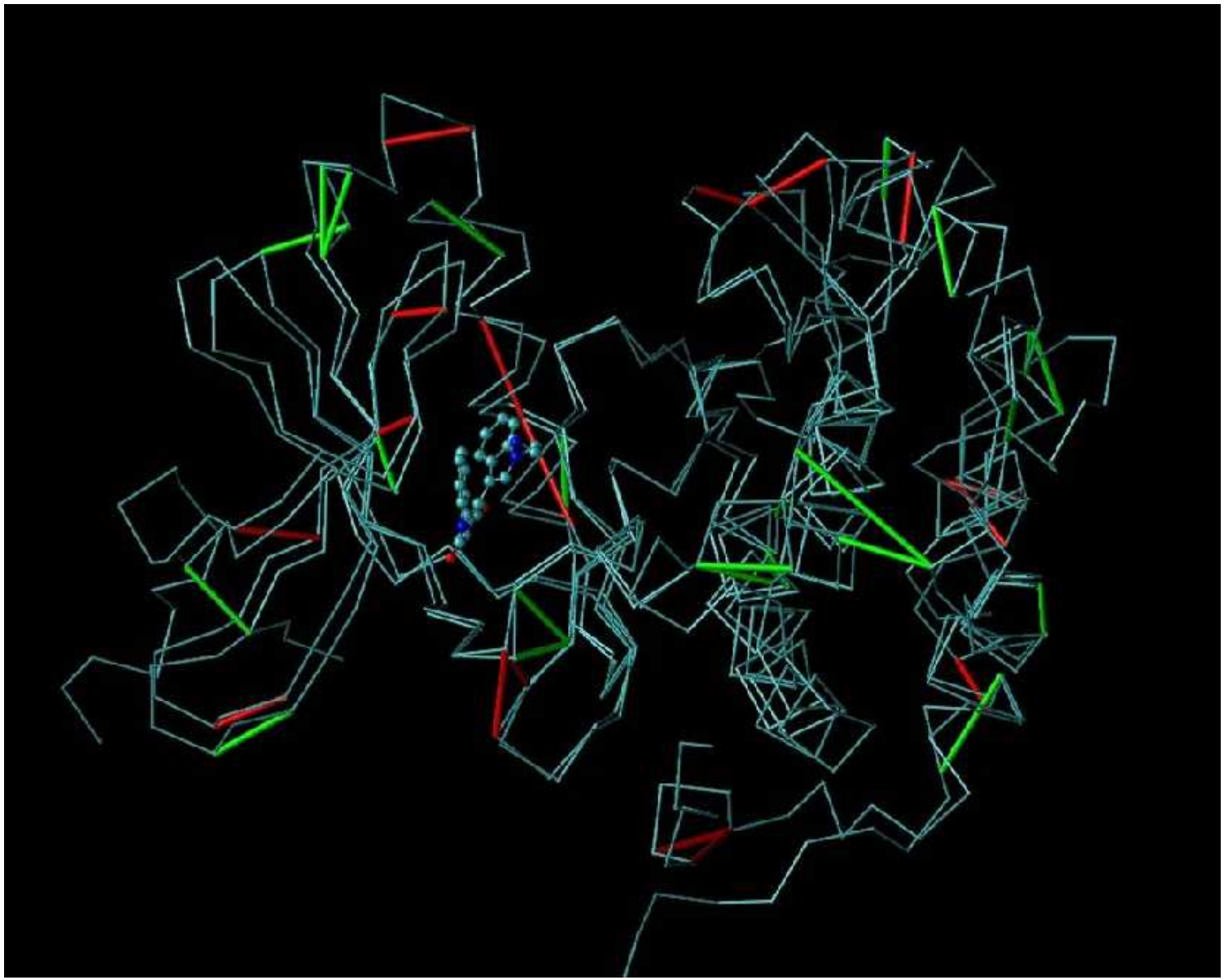


Dehydron Cys673-Gly676 in C-Kit is not conserved in its paralogs Bcr-Abl, Lck, Chk1 and Pdk1. By methylating Gleevec at the para position (1), the inhibitor becomes a selective wrapper of the packing defect in C-Kit.



Phosphorylation rates from spectrophotometric assay on the five kinases Bcr-Abl (blue), C-Kit (green), Lck (red), Chk1 (purple), and Pdk1 (brown) with Gleevec (triangles) and modified Gleevec methylated at positions (1) and (2) (squares). Notice the selective and enhanced inhibition of C-Kit.

Aligned backbones for two paralog kinases; dehydrons for Chk1 are marked in green and those for Pdk1 are in red [4].



6 Remaining challenges

Modeling hydrogen placement

- Hydrogens not resolved by imaging techniques, e.g., in Histidine sidechain

Role of ionic solvents

- How do ions affect local dielectric behavior?

Increasing entropy versus decreasing enthalpy

$$\Delta G = \Delta H - T\Delta S \tag{6.2}$$

References

[1] Yugong Cheng, Tanguy LeGall, Christopher J. Oldfield, James P. Mueller, Ya-Yue J. Van, Pedro Romero, Marc S. Cortese, Vladimir N. Uversky, and A. Keith Dunker. Rational drug design via intrinsically disordered protein. *Trends in Biotechnology*, 24(10):435 – 442, 2006.

[2] Alfonso De Simone, Guy G. Dodson, Chandra S. Verma, Adriana Zagari, and Franca Fraternali. Prion and water: Tight and dynamical hydration sites have a key role in structural stability. *Proceedings of the National Academy of Sciences, USA*, 102:7535–7540, 2005.

[3] Ariel Fernández and L. Ridgway Scott. Dehydron: a structurally encoded signal for protein interaction. *Biophysical Journal*, 85:1914–1928, 2003.

[4] Ariel Fernández and L. Ridgway Scott. Modulating drug impact by wrapping target proteins. *Expert Opinion on Drug Discovery*, 2:249–259, 2007.