

# Protein Struktur

- Biologen und Chemiker dürfen mit Handys spielen (leise)
- go home, go to sleep
- wake up at slide 39

# Proteins - who cares ?

Most important molecules in life ? Ask the DNA / RNA people

- structural (keratin / hair)
- enzymes (catalysts)
- messengers (hormones)
- regulation (bind to other proteins, DNA, ..)
- industrial – biosensors to washing powder
- receptors
- transporters (O<sub>2</sub>, sugars, fats)
- anti-freeze ...

# Proteins are easy

- data (protein data bank, [www.rcsb.org](http://www.rcsb.org))
  - $1.5 \times 10^5$  files
- literature on function, interactions, structure
- software
  - viewers, molecular dynamics simulators, docking, ..
- nomenclature and rules

# Proteins are not friendly

- one cannot take a sequence and predict structure/function
- data formats are full of surprises
- data contains error and mistakes

# Protein Rules, Physics, Folklore

## Physics / Chemistry

- protein + water = set of interacting atoms
  - can be calculated (not really)

## Rules (not quantified)

- proteins unfold if you heat them (exceptions ?)
- many charged amino acids.. they are soluble
- if they are more than 300 residues, they have more than one domain,
- proteins fold to a unique structure (could you prove this ?)
  - lowest free energy structure

# Protein chemistry

Chemists / biochemists

- sleep, go home
- one tiny surprise at the end of the lectures

Short version

- proteins are sets of building blocks (amino acids, residues, Reste)
- 20 types of residue
- chains of length few to  $10^3$  ( 100 or 200 typical)
- small ones ( $< \approx 50$  residues) are peptides
- they fold up to nice stable structures – why ?

Longer version..

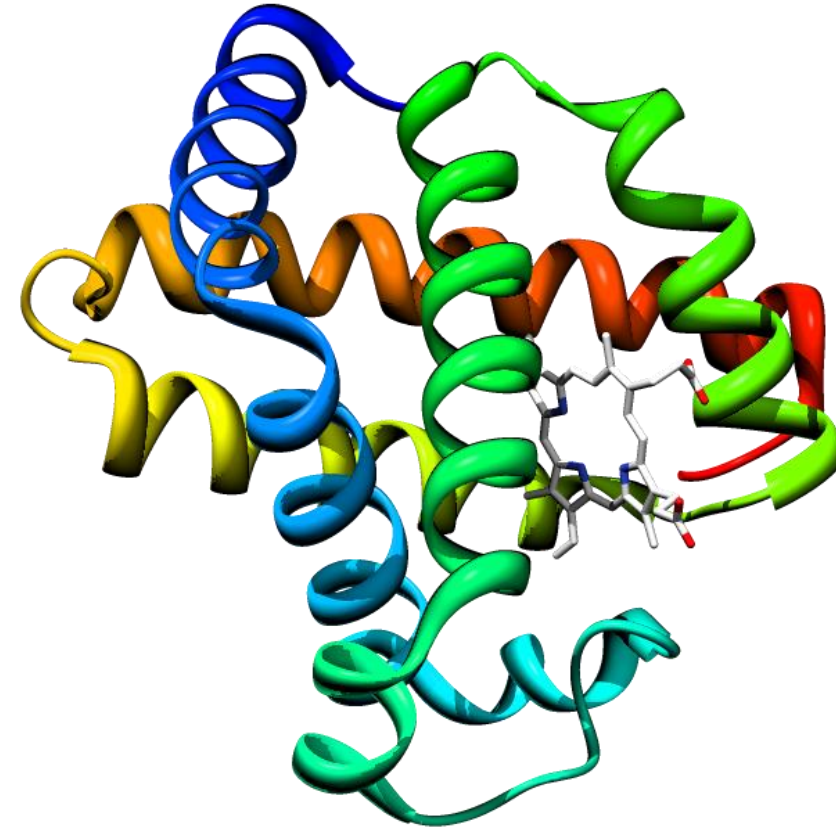
# The Plan

- polymers
- different kinds of sidechain
- structure due to backbone (secondary structure)
- properties of sidechains
- representation

# Sizes

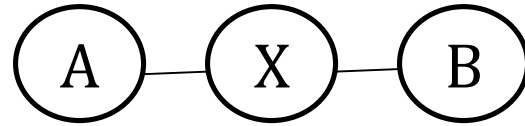
$1 \text{ \AA} = 10^{-10} \text{ m}$  or  $0.1 \text{ nm}$

| structure                        | size                        |
|----------------------------------|-----------------------------|
| bond CH                          | $1 \text{ \AA}$             |
| CC                               | $1.5 \text{ \AA}$           |
| protein radius                   | $10 - 10^2 \text{ \AA}$     |
| $\alpha$ -helix spacing          | $5 \frac{1}{2} \text{ \AA}$ |
| $C_i^\alpha$ to $C_{i+1}^\alpha$ | $3.8 \text{ \AA}$           |

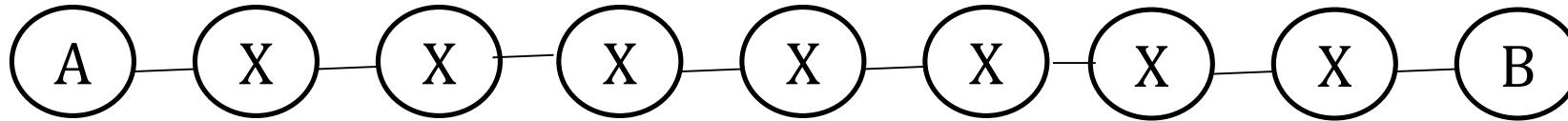


# Proteins are polymers

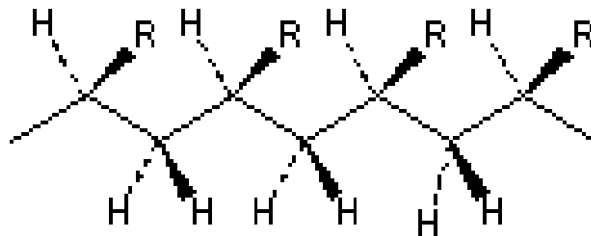
simple polymers



many times gives



example



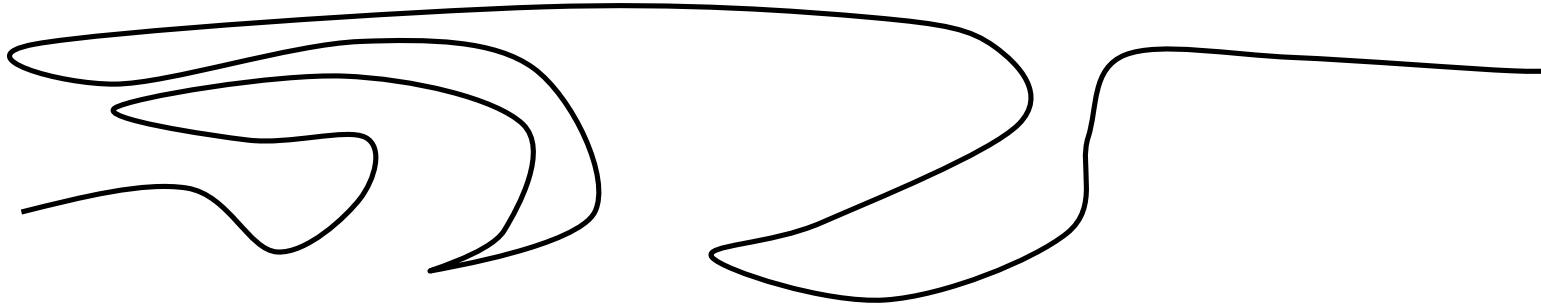
what kind of polymer would this give ?

Do you know what R is ?



# Why are proteins interesting polymers ?

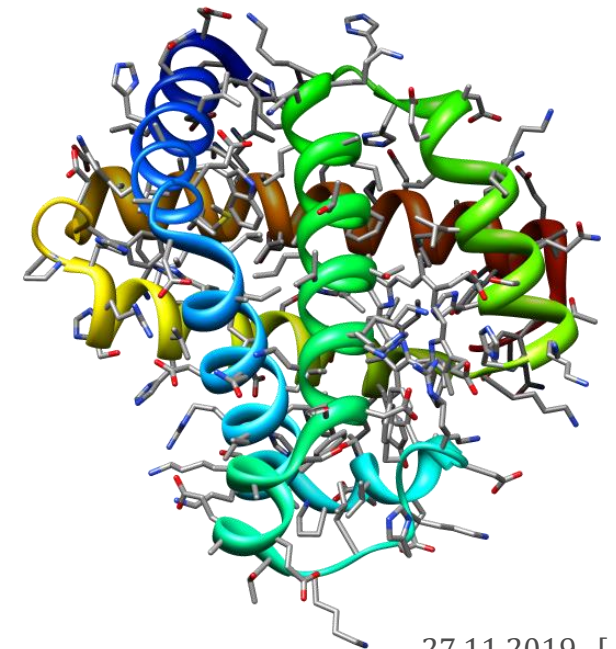
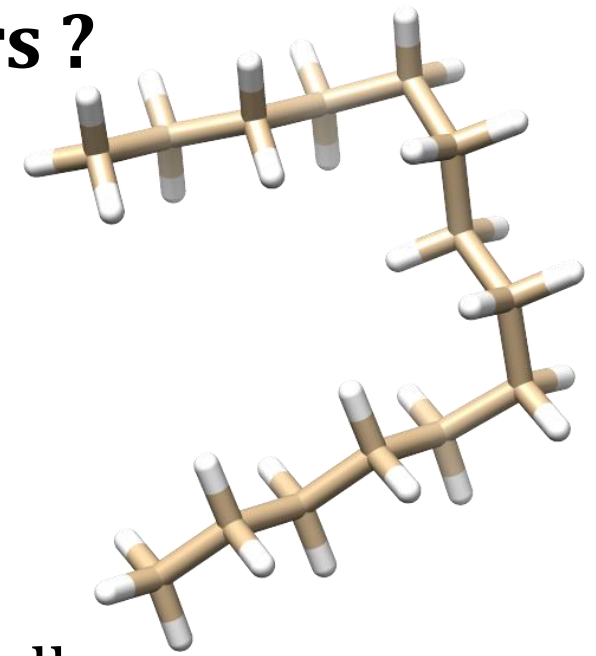
Why do boring polymers not have well-defined structures ?



Each part of polymer wants to interact with all other parts equally

- no structural preferences
- plastic bags, Haushaltsfolie
- no regular structures

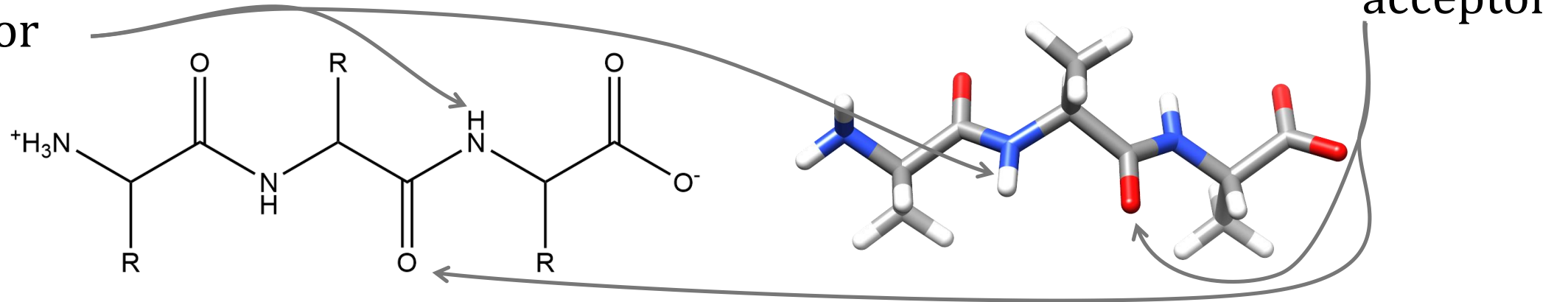
Properties that make proteins different from plastics ..



# Giving proteins character 1

More complicated backbone with H-bond

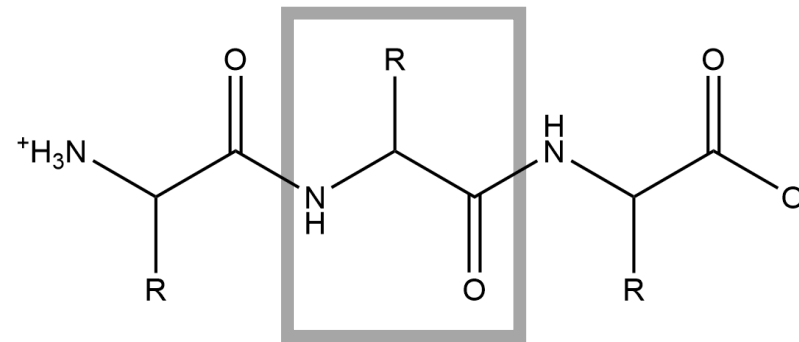
donor



- basis of standard regular structures in proteins (secondary structure)
- repeating polymer unit:

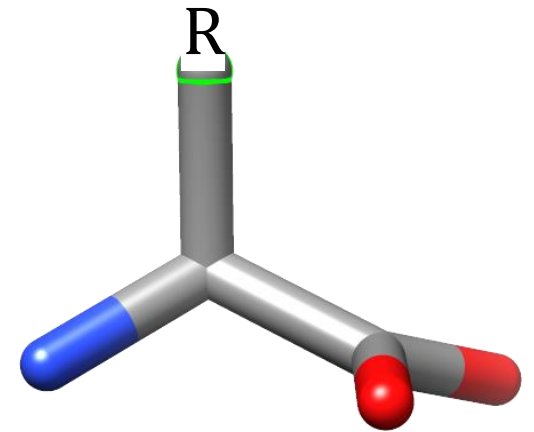
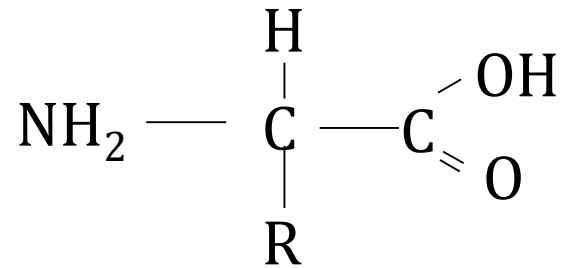
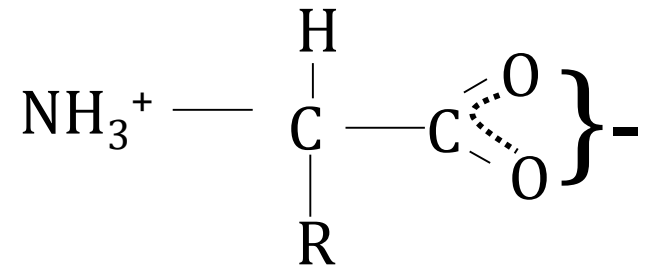
If this was all there was

- all proteins would be the same



# protein chemistry

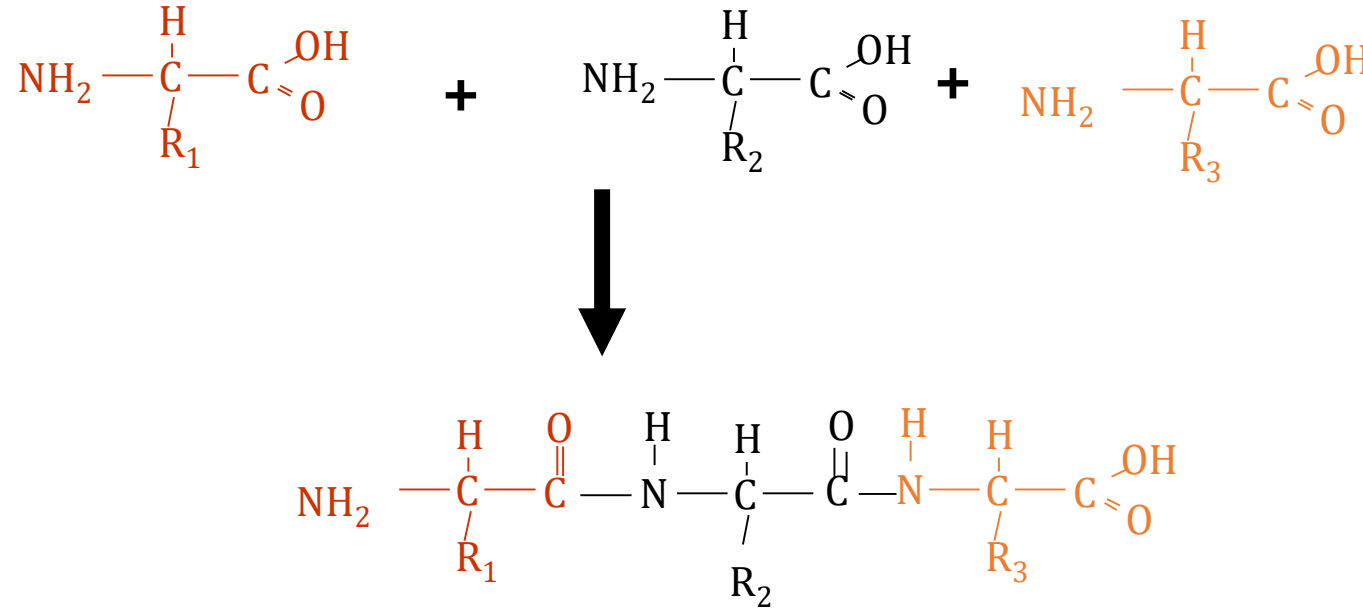
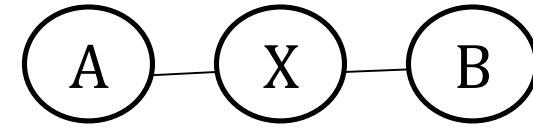
amino acids (monomers) all look like:



How can we construct specific structures ?

- different kinds of "R" groups

# Putting monomers together



- protein synthesis story (biochemistry lectures)
- peptides and proteins
  - < 30 or 40 residues = peptide
  - > 30 or 40 residues = protein

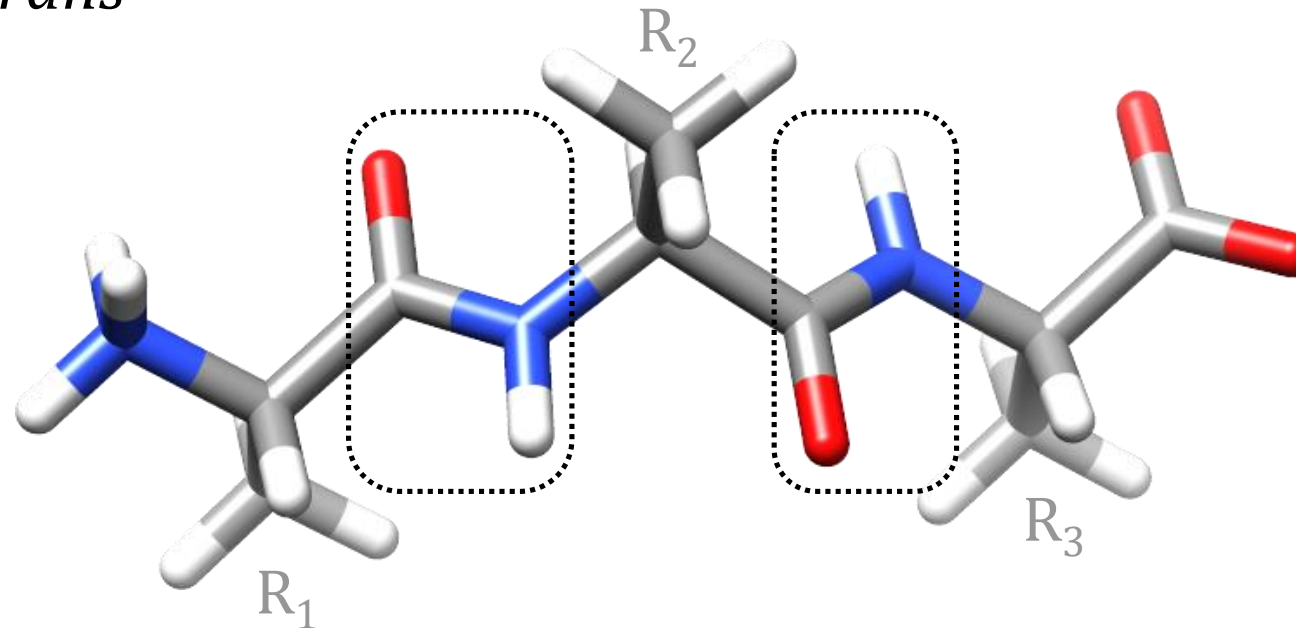
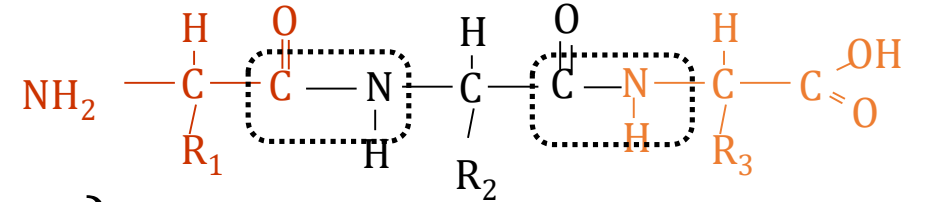
# Backbone peptide bonds

How many backbone angles ?

- 3 ( $\phi$ ,  $\psi$ ,  $\omega$ )

Peptide bond  $\omega$  is planar

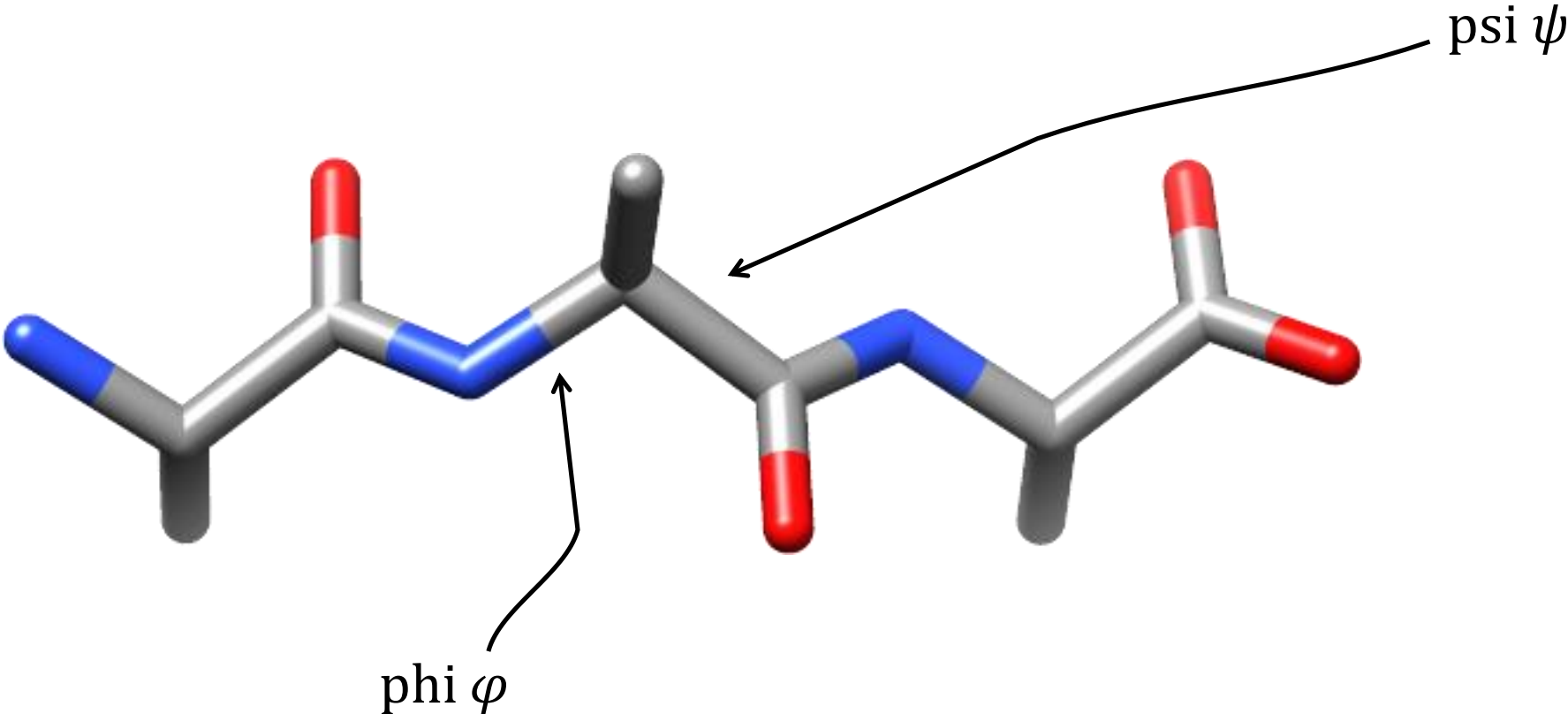
- partial double bond character (resonance forms)
- shorter than other C-N
- nearly always *trans*



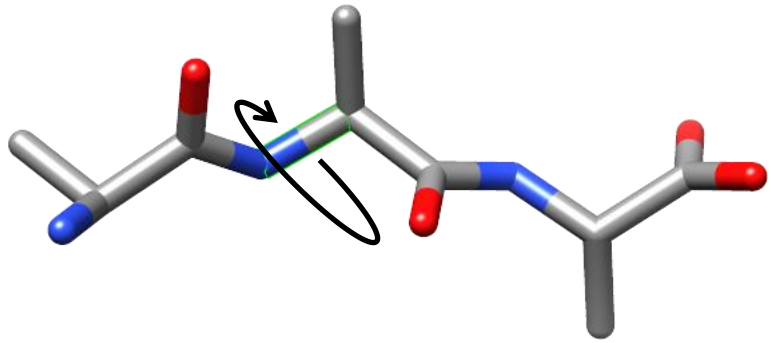
Note:  
usually we  
do not  
draw H  
atoms

# Backbone rotatable angles

Two rotatable angles  $\phi$ ,  $\psi$

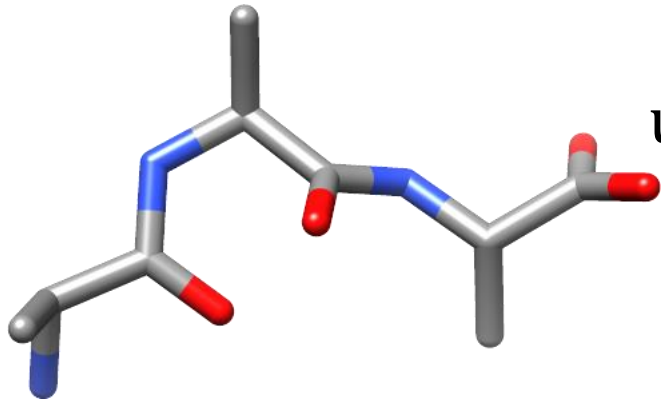
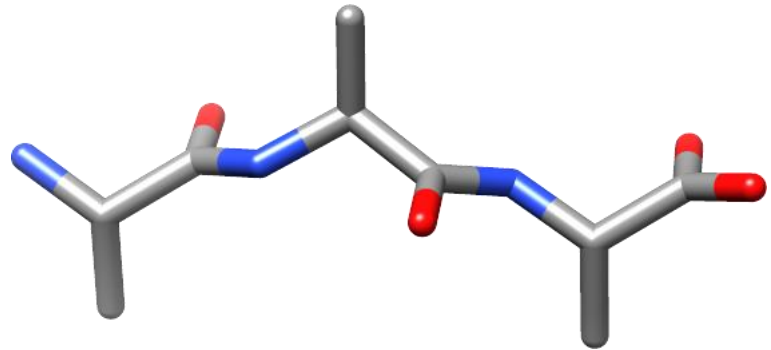


## some $\phi$ rotations



can we rotate freely ?

- no... steric hindrance



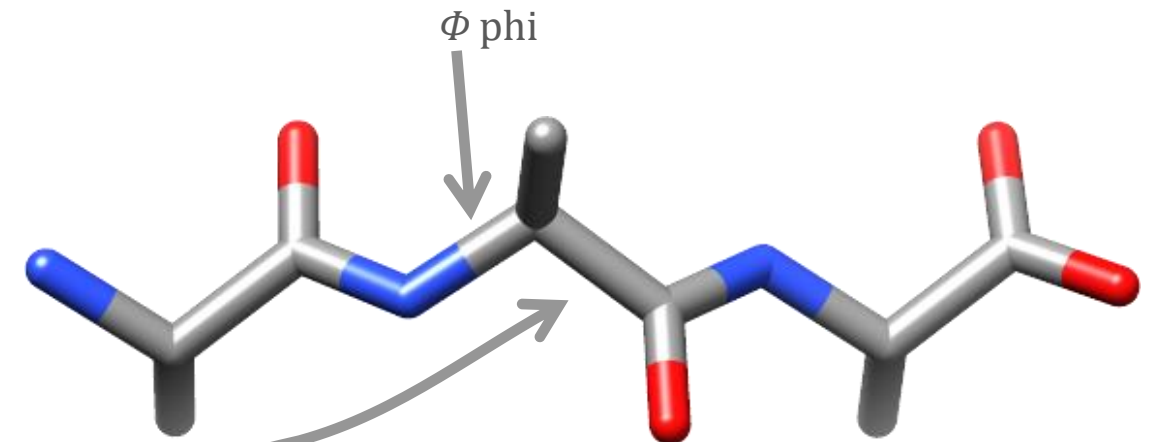
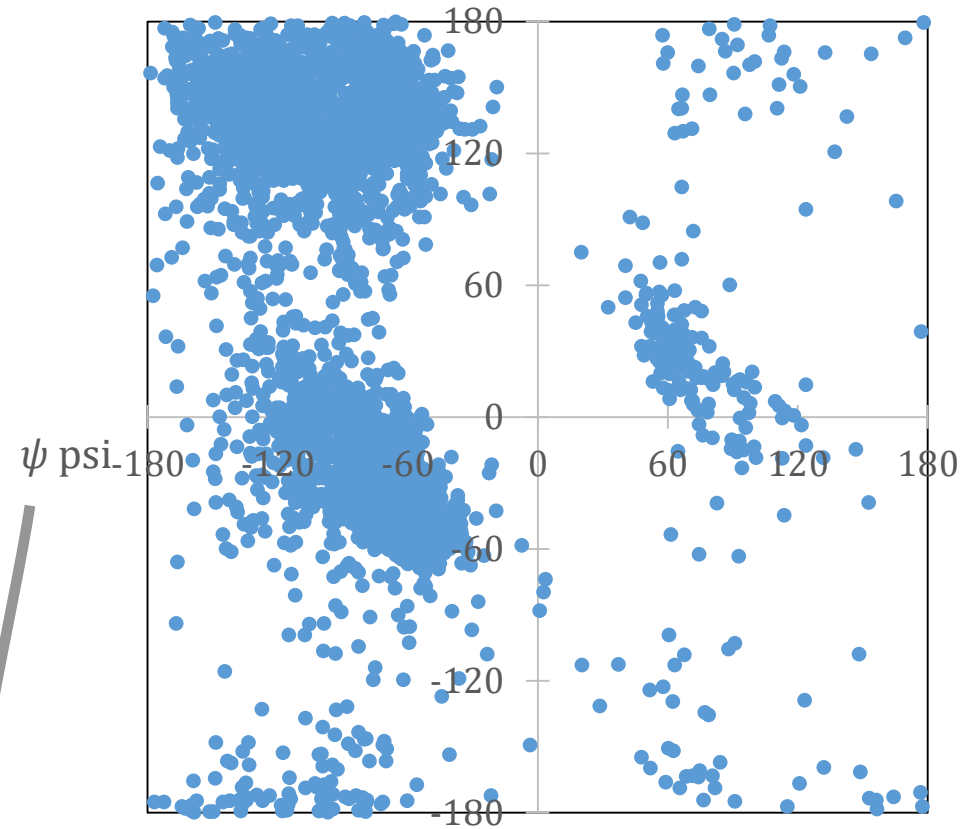
unhappy O atoms – high energy

# ramachandran plot

Can we rotate freely ?

- no... steric hindrance

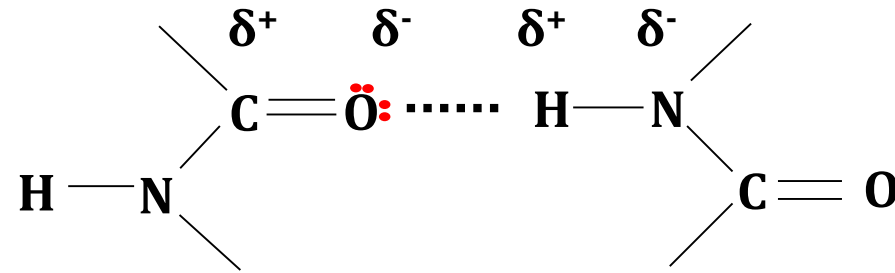
Ramachandran plot  
will reappear very often





# Backbone H bonds

- oxygen is slightly negative
- NH bond is polar



## H-bonds

- can be near or far in sequence
- fairly stable at room temperature

# Secondary structure

Regular structures using information so far

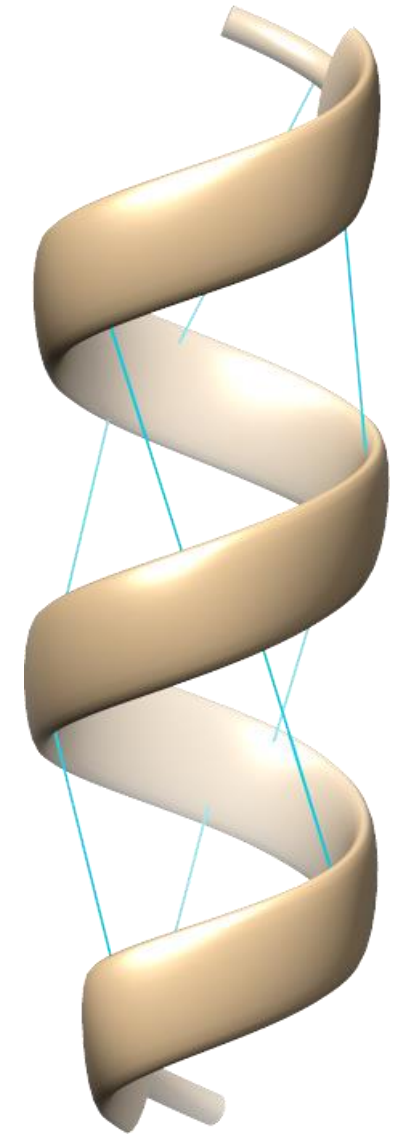
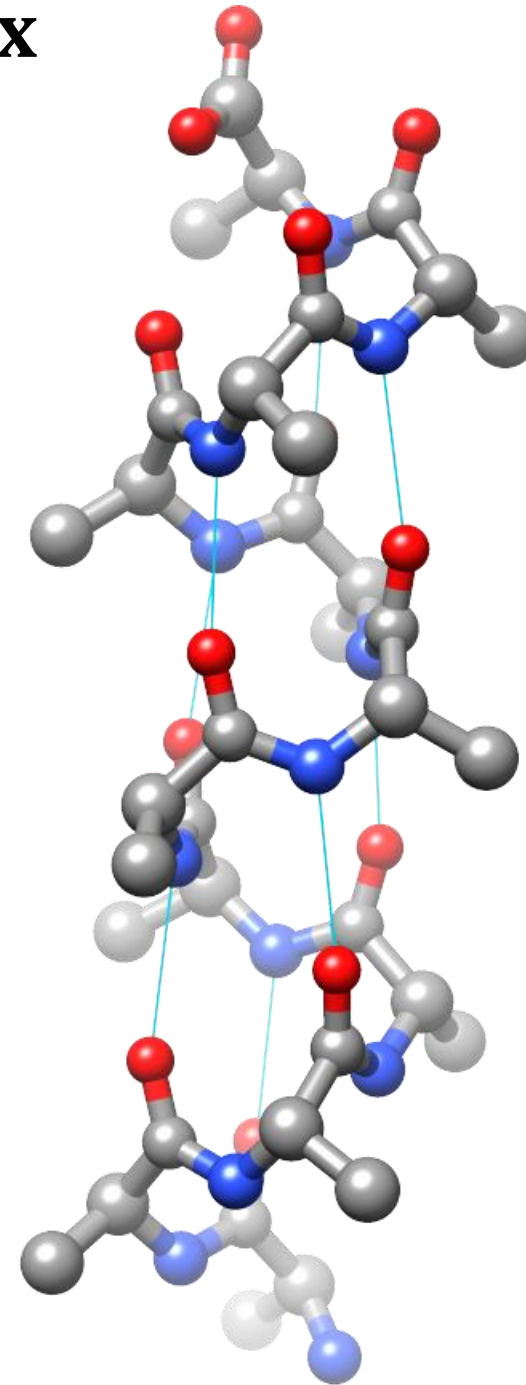
- rotate phi ( $\varphi$ ), psi ( $\psi$ ) angles so as to
  - form H-bonds where possible
  - do not force side chains to hit each other (steric clash)

Two common structures

- $\alpha$ -helix
- $\beta$ -strand / sheet

# $\alpha$ helix

- each CO of residue  $i$  H-bonded to N of  $i+4$
- 3.6 residues per turn
- 2 H-bonds per residue
- side chains well separated



# $\beta$ -sheet

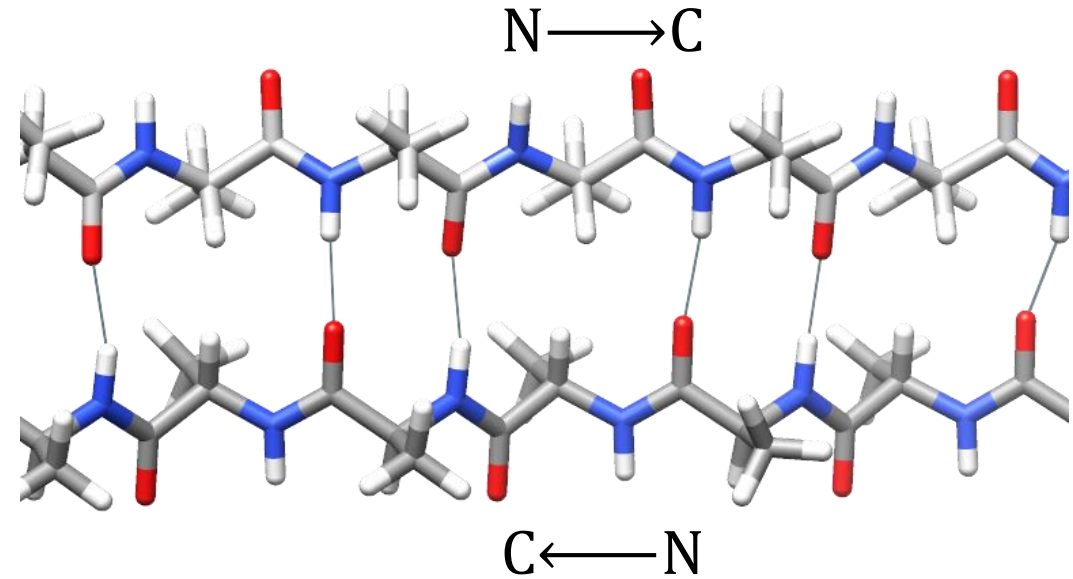
## $\beta$ -strand

- stretch out backbone and make NH and CO groups point out

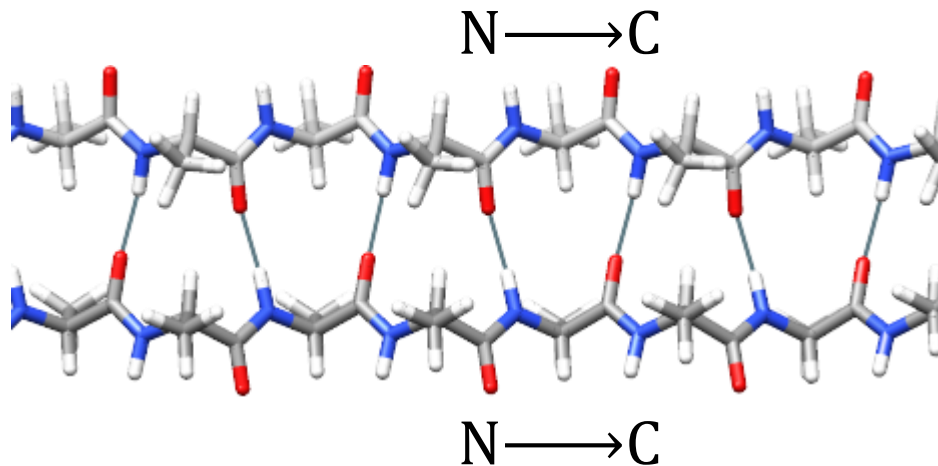
## $\beta$ -sheet

- join these strands together with H-bonds (2 H-bonds/residue)

anti-parallel



or parallel



# After $\alpha$ -helix and $\beta$ -sheet

Do helices and sheets explain everything? No

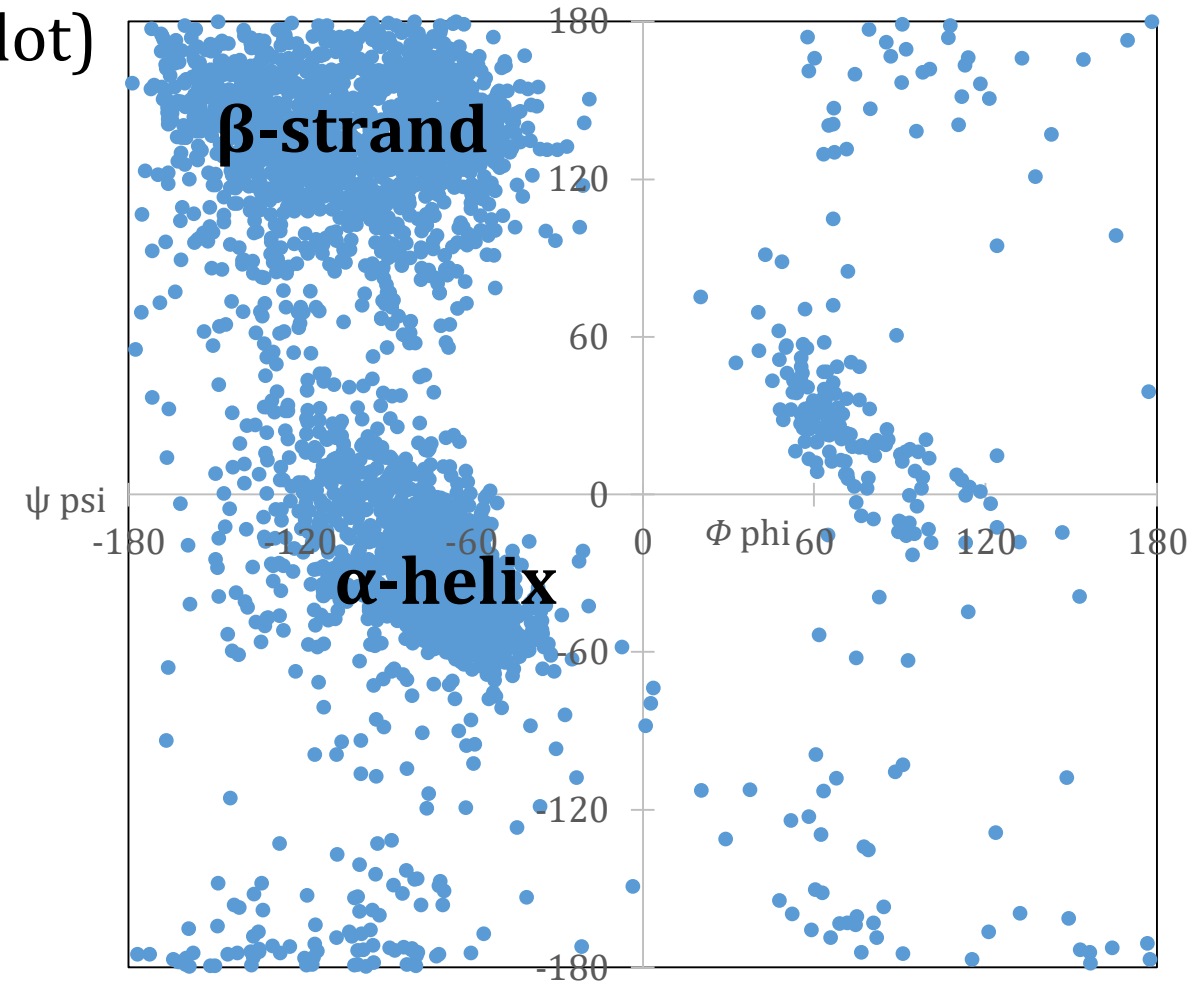
- there is flexibility in the angles (look at plot)
  - geometry is not perfectly defined
- there are local deviations and exceptions

Other common structures

- tighter helices
- some turns

Other structure

- coil, random, not named



# What determines secondary structure ?

So far

- secondary structure pattern of H-bonding

Almost all residues have H-bond acceptor and donor

- almost all could form  $\alpha$ -helix or  $\beta$ -sheet

Difference ?

- sequence of side-chains – overall folding

Why else are sidechains important

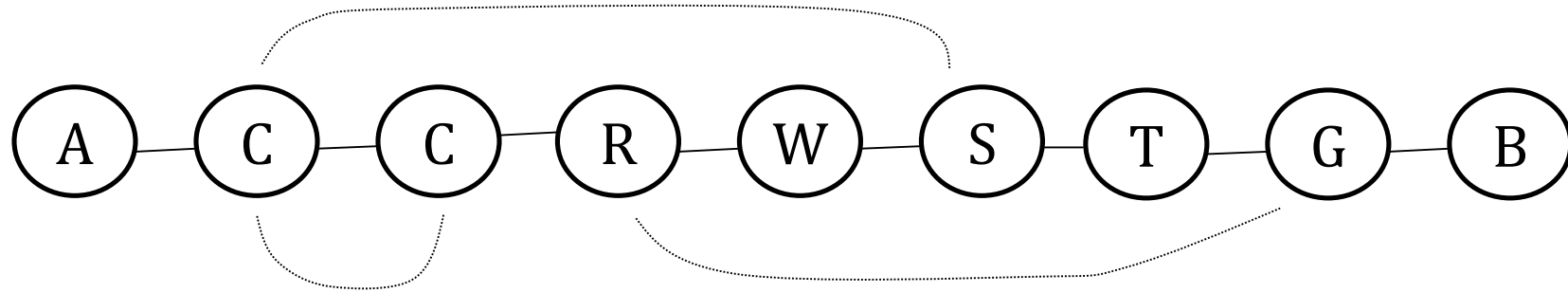
- chemistry of proteins (interactions, catalysis)

Fundamental dogma

- the sequence of sidechains determines the protein shape

# side chain possibilities

- big / small
- charged +, charged -, polar
- hydrophobic (not water soluble), polar
- interactions between sites...



- a CS interaction is different to CC is different to RG ...

# Side chain properties

properties

- big / small
- neutral / polar / charged
- special (...)

example

- phenylalanine side chain looks like benzene (benzin)
  - very insoluble
  - benzene would rather interact with benzene than water
  - what if you have phe-phe-phe... poly-phe ?
    - does not happen in nature (can be made)
    - would be insoluble
    - not like a real peptide
  - phe is a constituent of real proteins – has a role

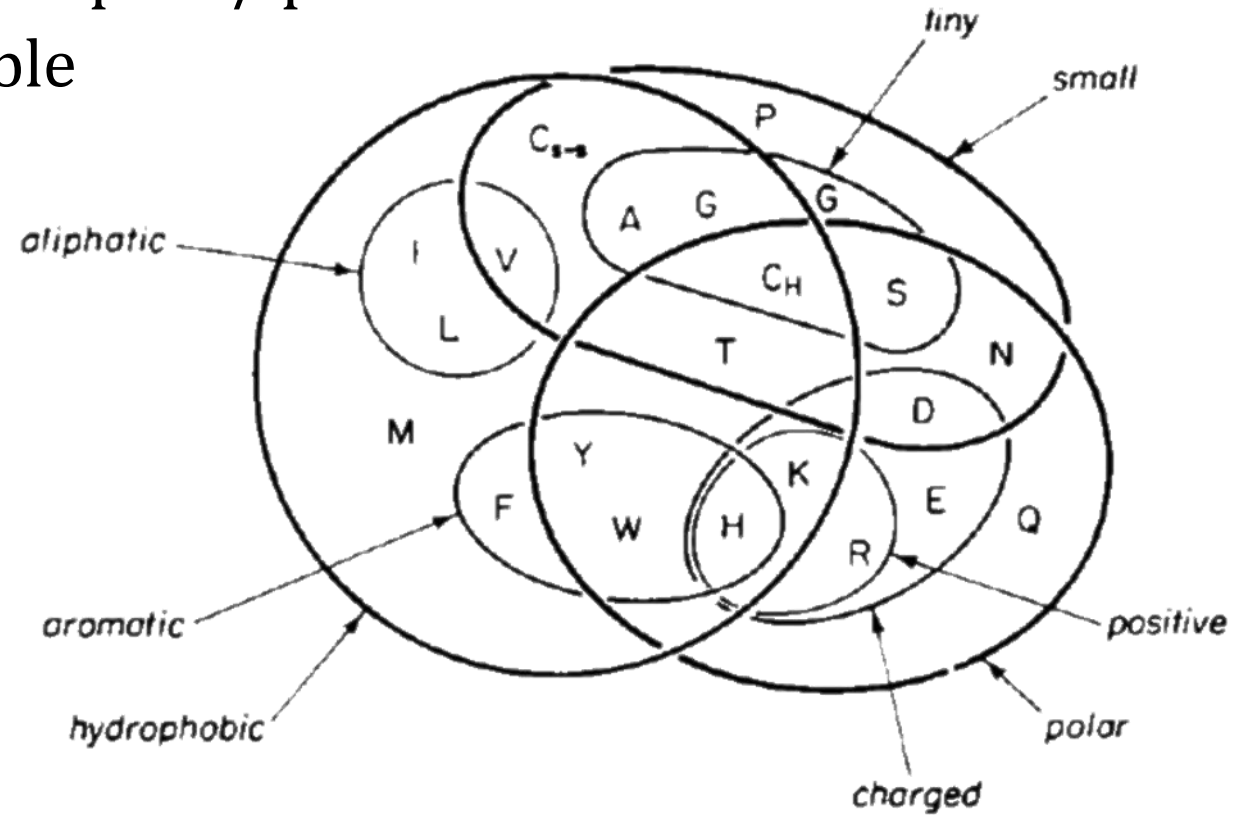




# Properties are not clear cut

You can be big / small, hydrophobic / polar

- combinations are possible

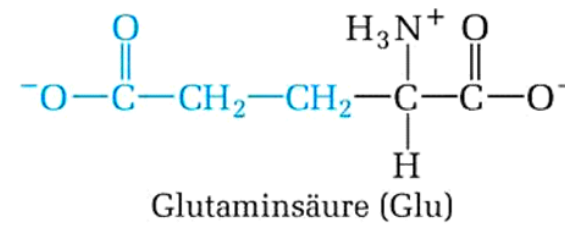
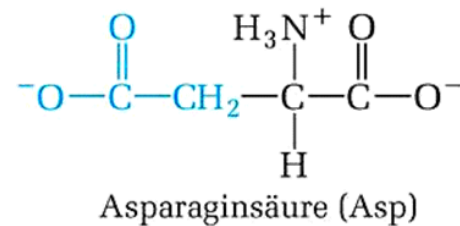
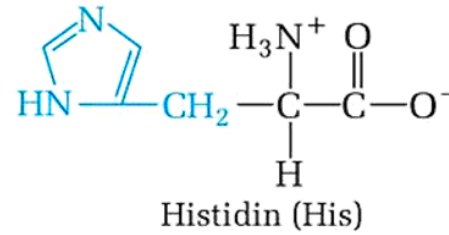
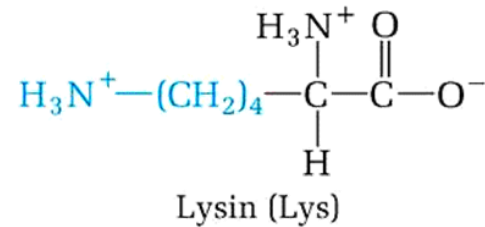
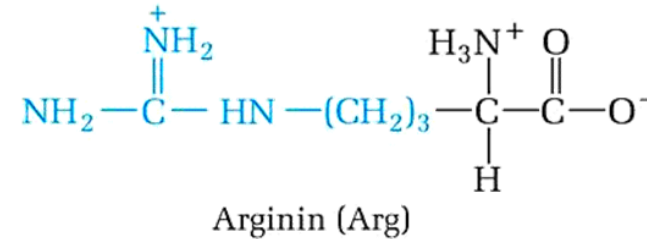
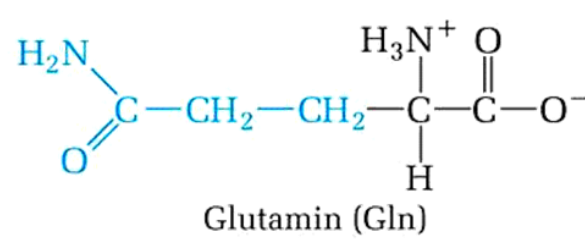
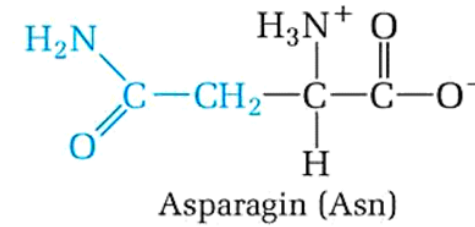
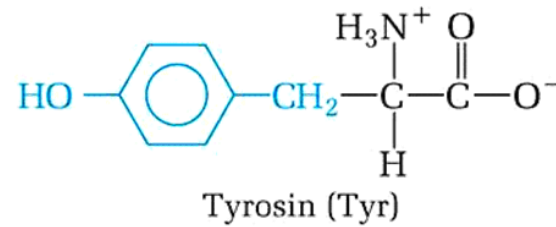
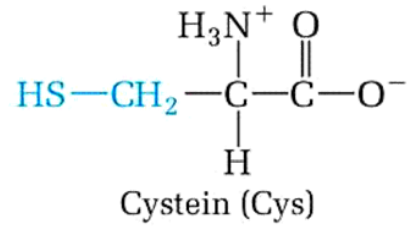
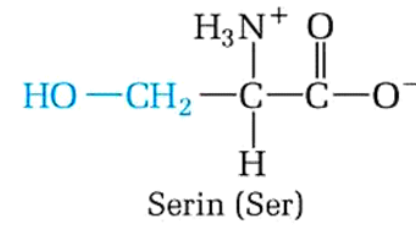
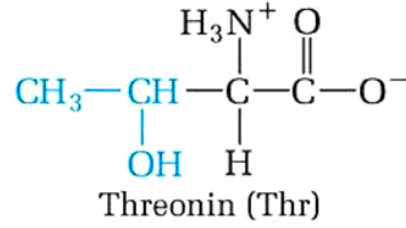
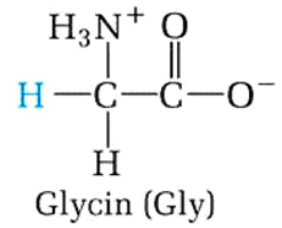


Do not memorise this figure

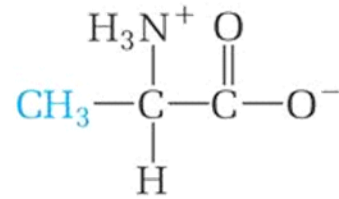
# Sidechain interactions

- ionic (if the sidechains have charge)
- hydrophobic (insoluble sidechains)
- H-bonds (some donors and acceptors)
- repulsive

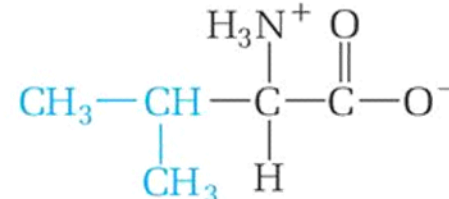
# Summary of amino acids (first dozen)



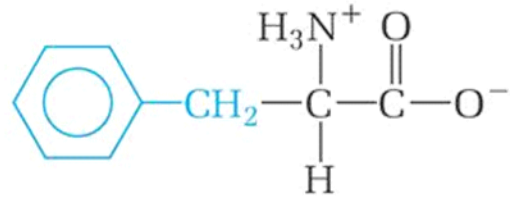
# summary of amino acids (part 2)



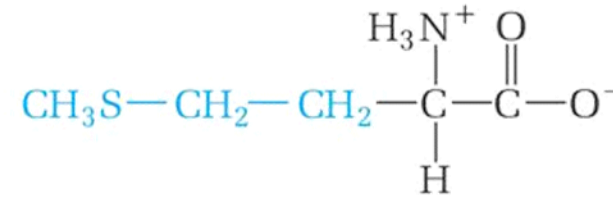
Alanin (Ala)



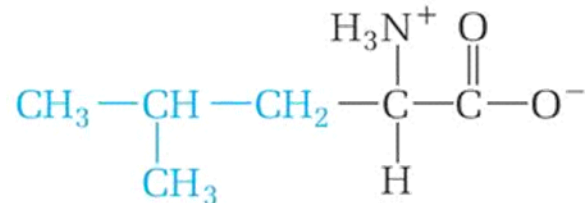
Valin (Val)



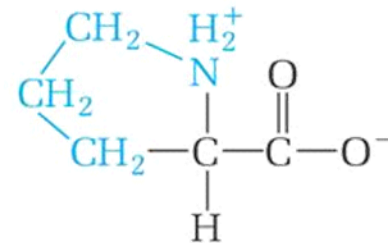
Phenylalanin (Phe)



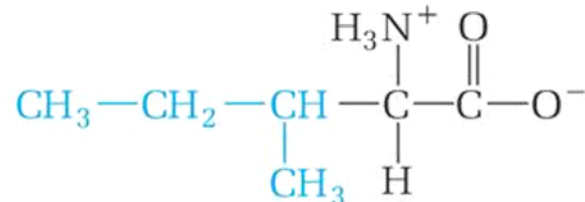
Methionin (Met)



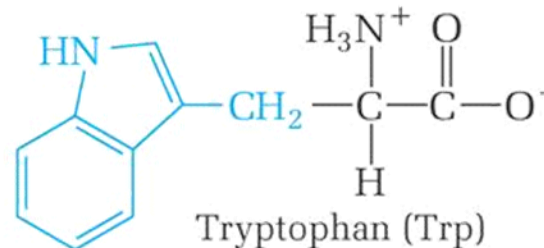
Leucin (Leu)



Prolin (Pro)



Isoleucin (Ile)

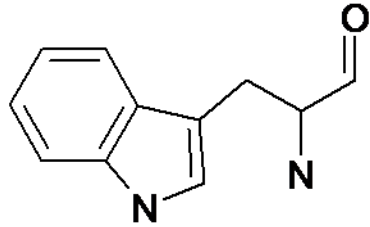


Tryptophan (Trp)

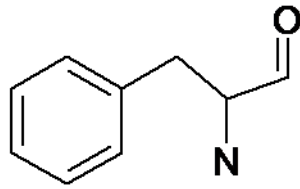
# Amino Acids by property

## aromatic

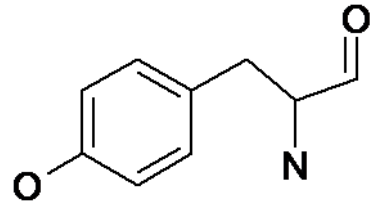
tryptophan



phenylalanine

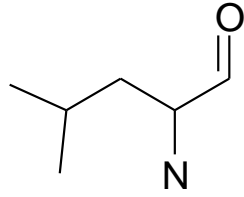


tyrosine

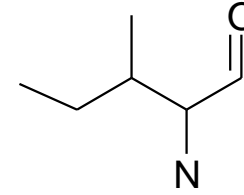


# rather hydrophobic

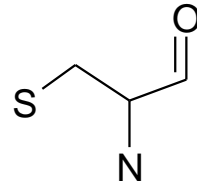
leucine



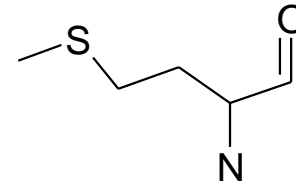
isoleucine



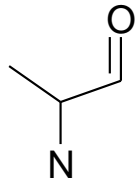
cysteine



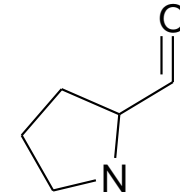
methionine



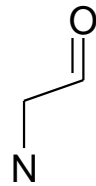
alanine



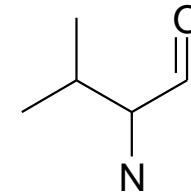
proline



glycine

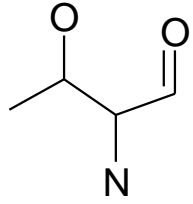


valine

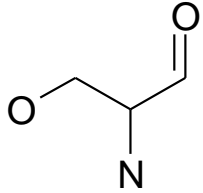


# Polar

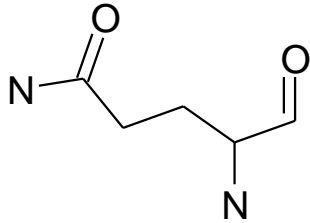
threonine



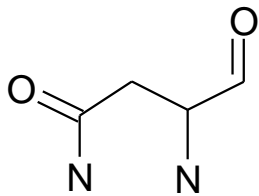
serine



glutamine

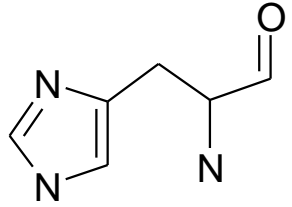


asparagine

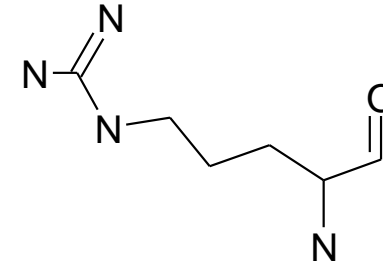


# charged

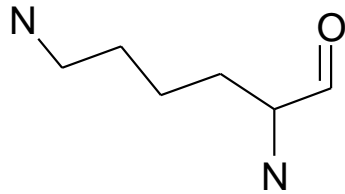
histidine



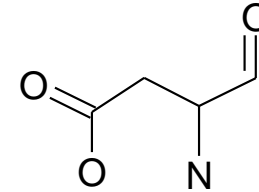
arginine



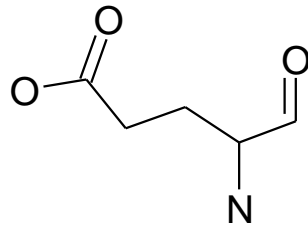
lysine



aspartate



glutamate



Muss ich alle Strukturen für die Klausur wissen ?

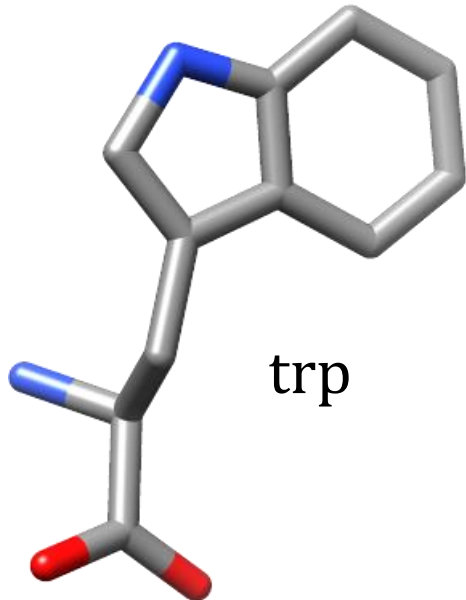


# Hydrophobicity – how serious ?

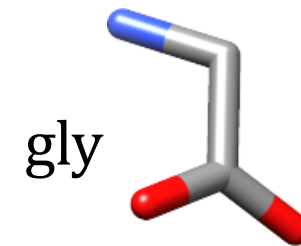
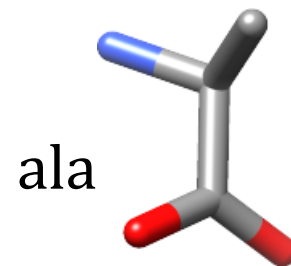
Very serious, but simplified

- the lists above are
  - pH dependent
  - difficult to measure experimentally (some aspects)
- Is there a single definition for hydrophobicity ?

## Other properties - size



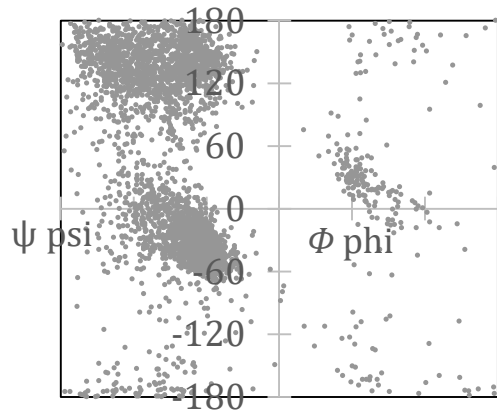
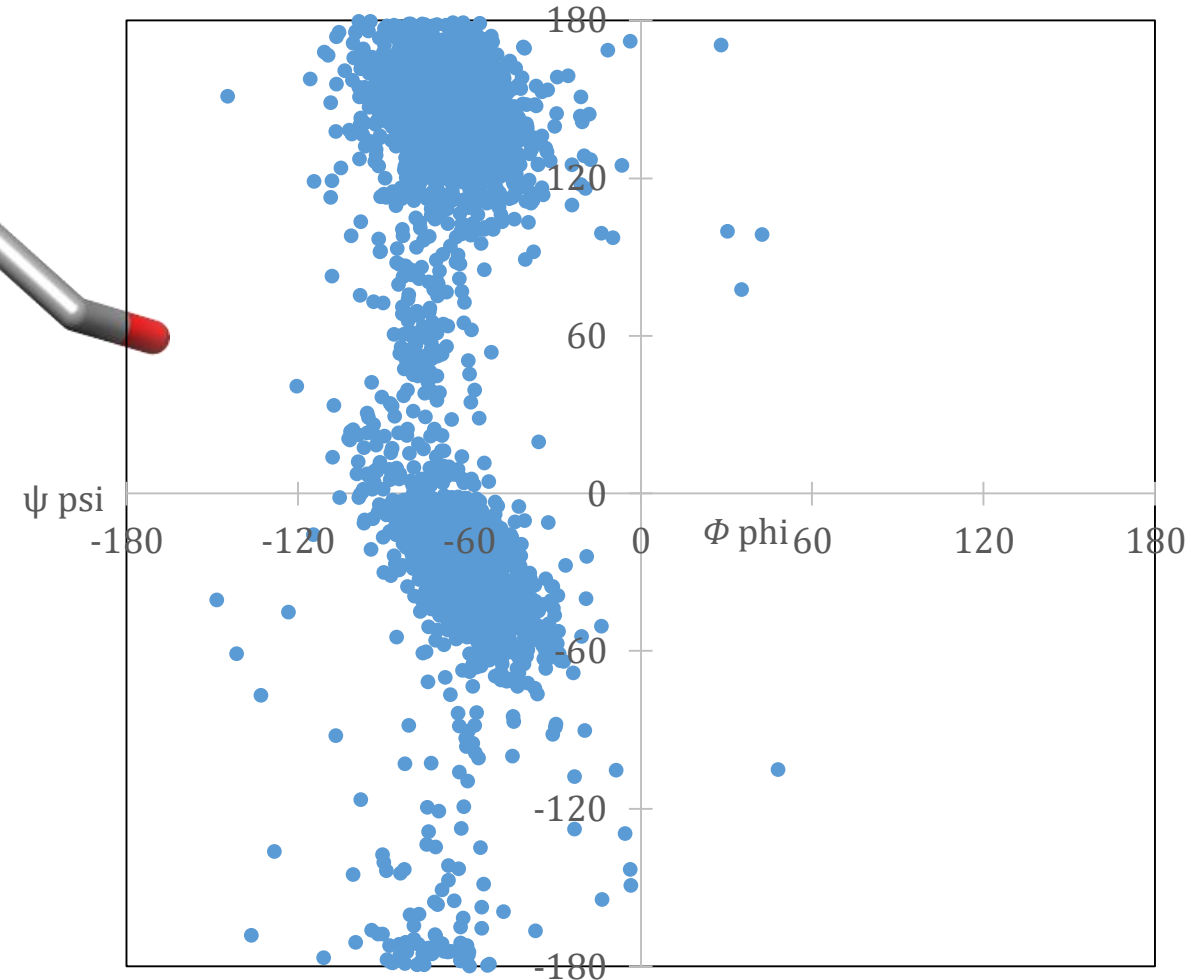
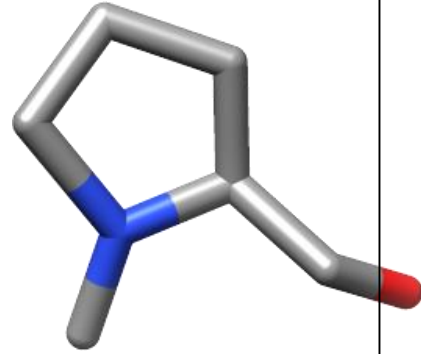
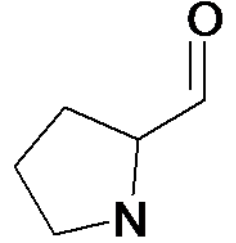
big                      ...                      small  
→



# Other properties - chemistry / geometry

## Proline

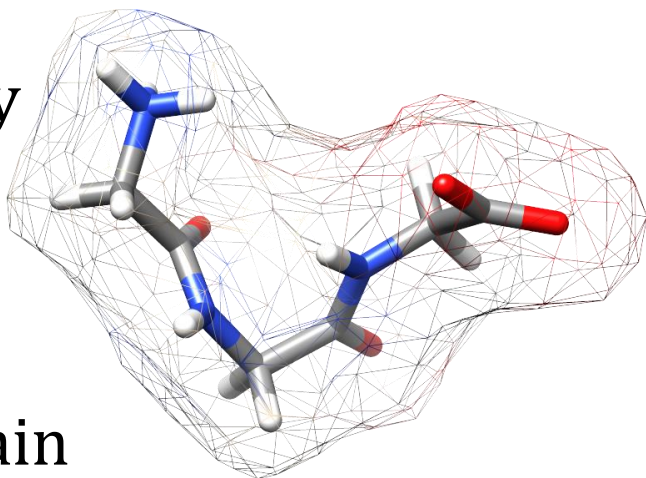
- only one rotatable angle !
- peptide bond sometimes *cis*
- pro ramachandran plot



all  
amino acids

gly

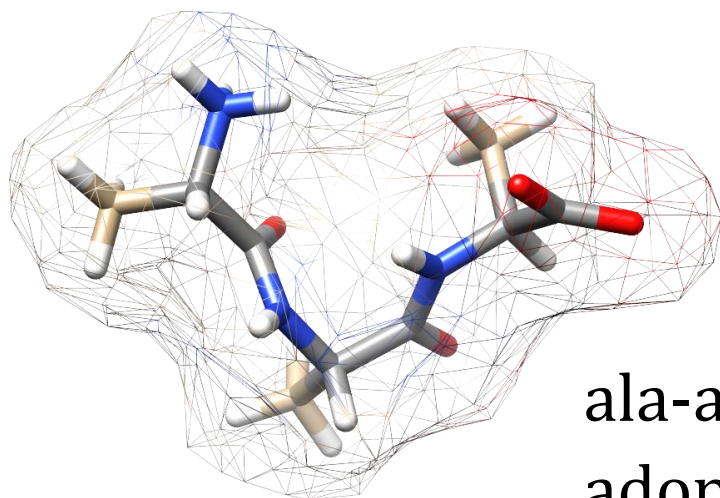
gly-gly-gly



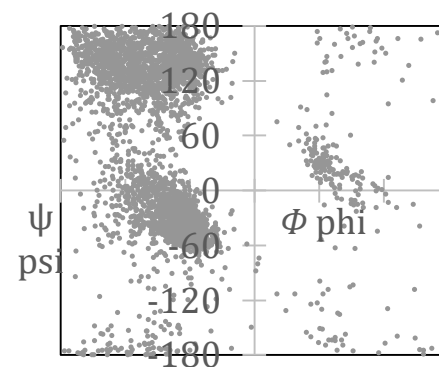
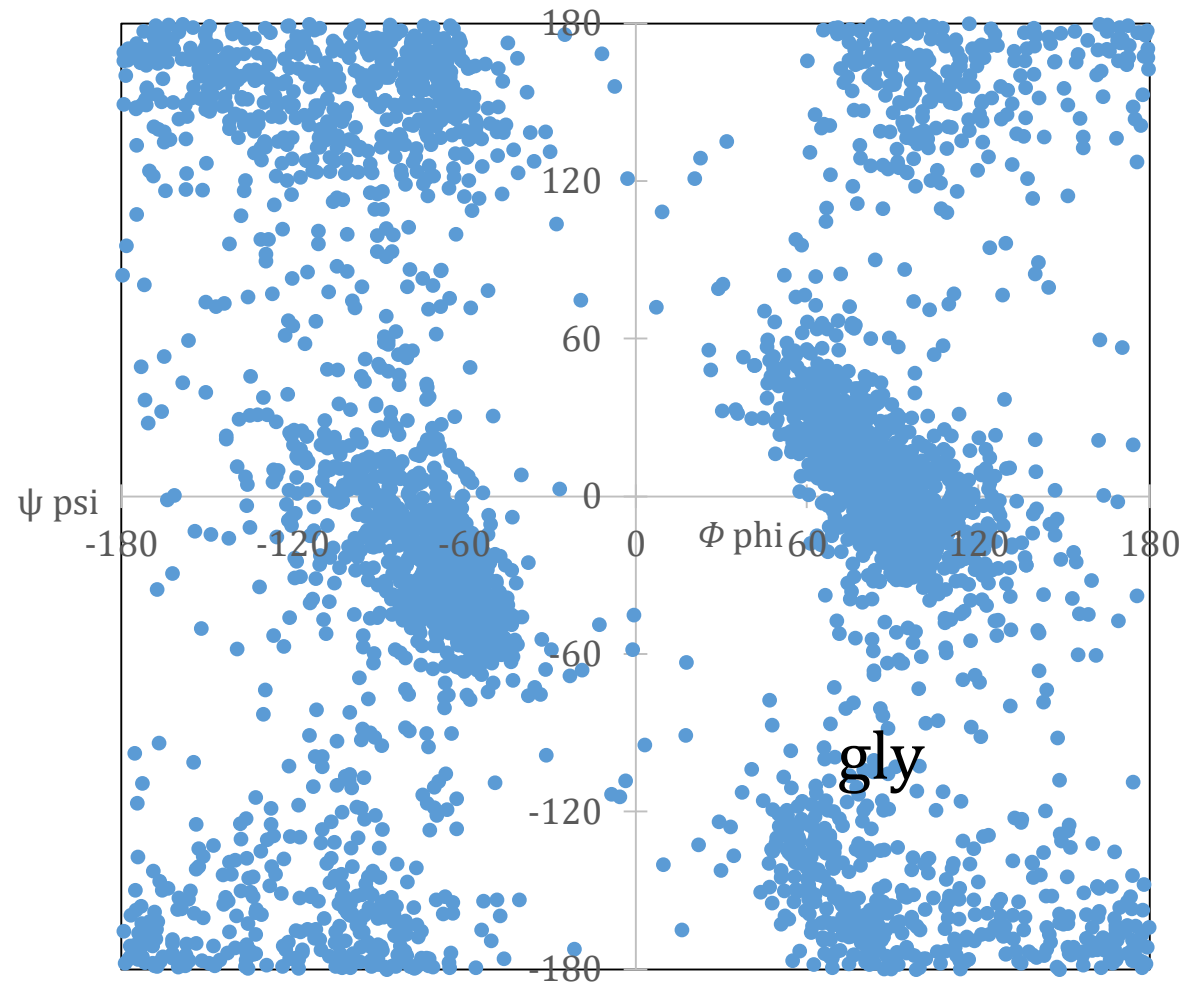
glycine

- no side chain
- can visit forbidden parts of phi-psi map

evolutionary consequences ?



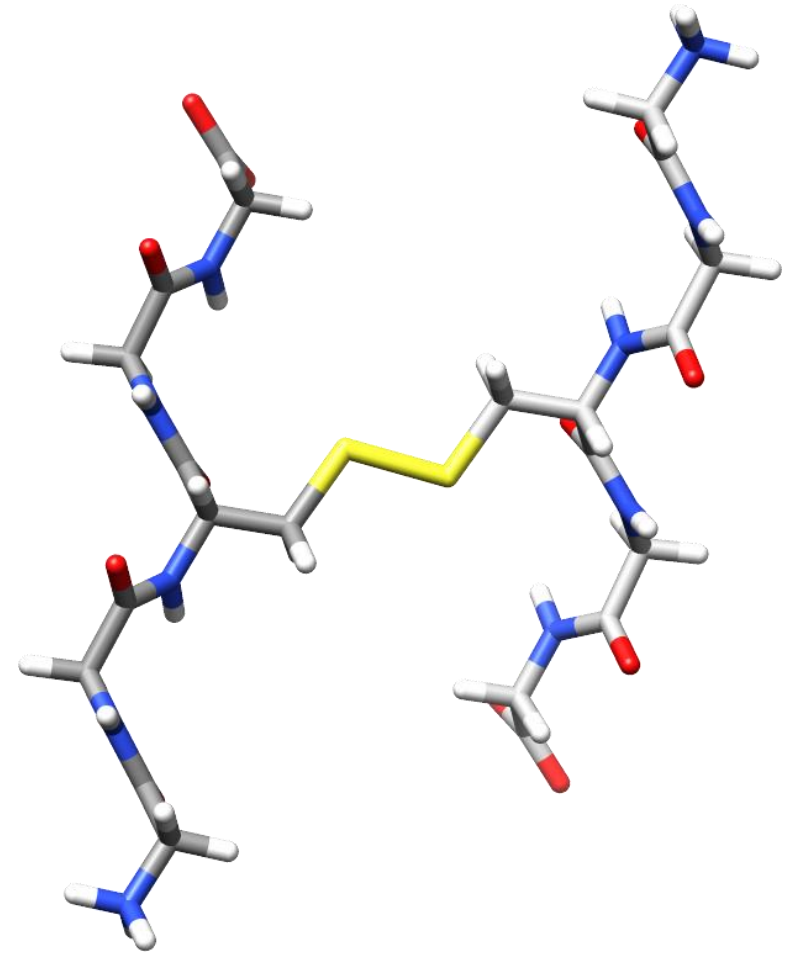
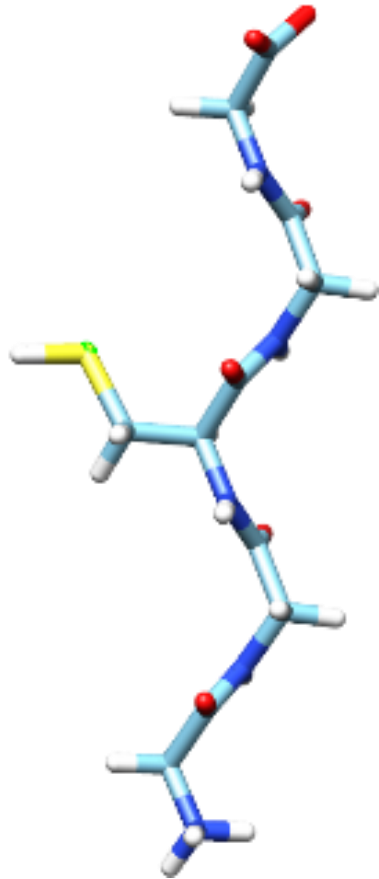
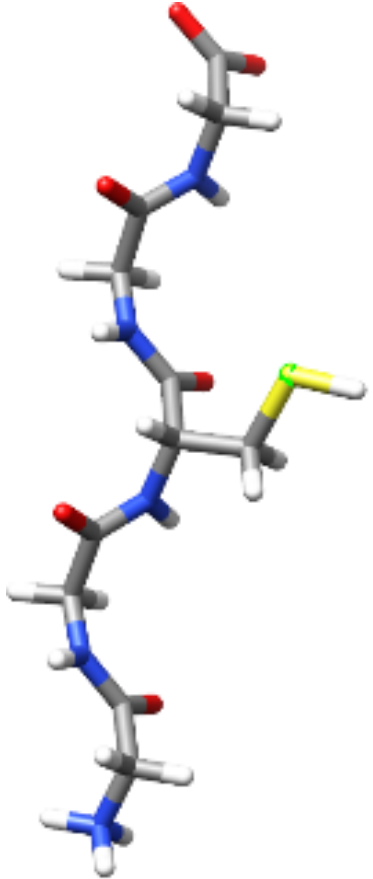
ala-ala-ala cannot  
adopt this conformation



# cys

cys residues can bind to each other

effect on conservation (mutation rates) ?



# Summary so far

- proteins are heteropolymers
- backbone forms  $\alpha$ -helices and  $\beta$ -strands (and more)
  - not sequence specific
- side-chains determine the
  - pattern of secondary structure
  - overall protein shape
- special amino acids
  - cys (forms disulfide bridges)
  - gly (can visit "forbidden" regions of ramachandran plot)
  - pro (no H-bond donor)
- how many sequences can one have ?  $20^n$

# Nomenclature

Some rules are unavoidable

|               |     |   |
|---------------|-----|---|
| Alanine       | Ala | A |
| Cysteine      | Cys | C |
| Aspartic acid | Asp | D |
| Glutamic acid | Glu | E |
| Phenylalanine | Phe | F |
| Glycine       | Gly | G |
| Histidine     | His | H |
| Isoleucine    | Ile | I |
| Lysine        | Lys | K |
| Leucine       | Leu | L |
| Methionine    | Met | M |
| Asparagine    | Asn | N |
| Proline       | Pro | P |
| Glutamine     | Gln | Q |
| Arginine      | Arg | R |
| Serine        | Ser | S |
| Threonine     | Thr | T |
| Valine        | Val | V |
| Tryptophan    | Trp | W |
| Tyrosine      | Tyr | Y |

Always write from N to C terminal  
(convention)

# Definitions, primary, secondary ...

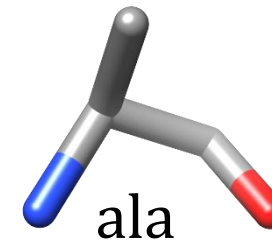
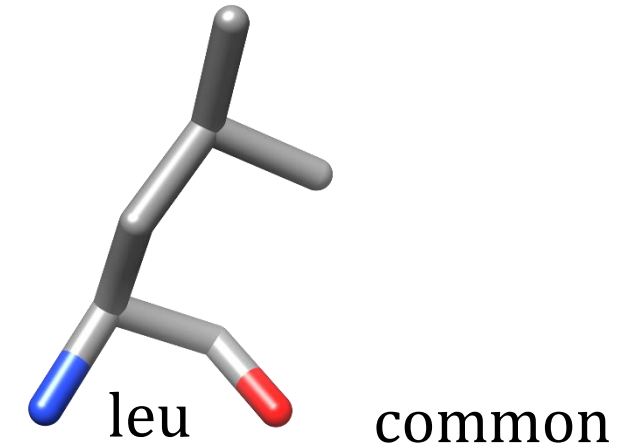
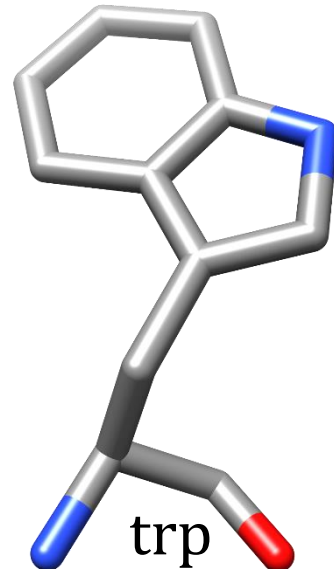
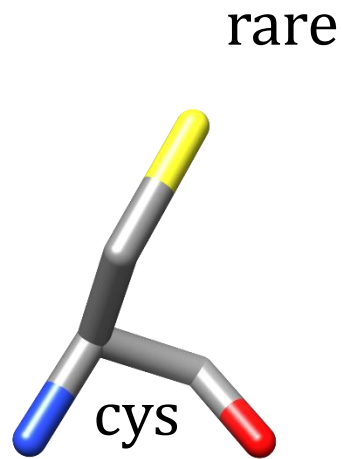
## More definitions

- primary structure
  - sequence of amino acids
    - ACDF (ala cys asp phe...)
- secondary structure
  - $\alpha$ -helix,  $\beta$ -sheet (+ few more)
    - structure defined by local backbone
- tertiary structure
  - how these units fold together
  - coordinates of a protein

# distributions of residue types

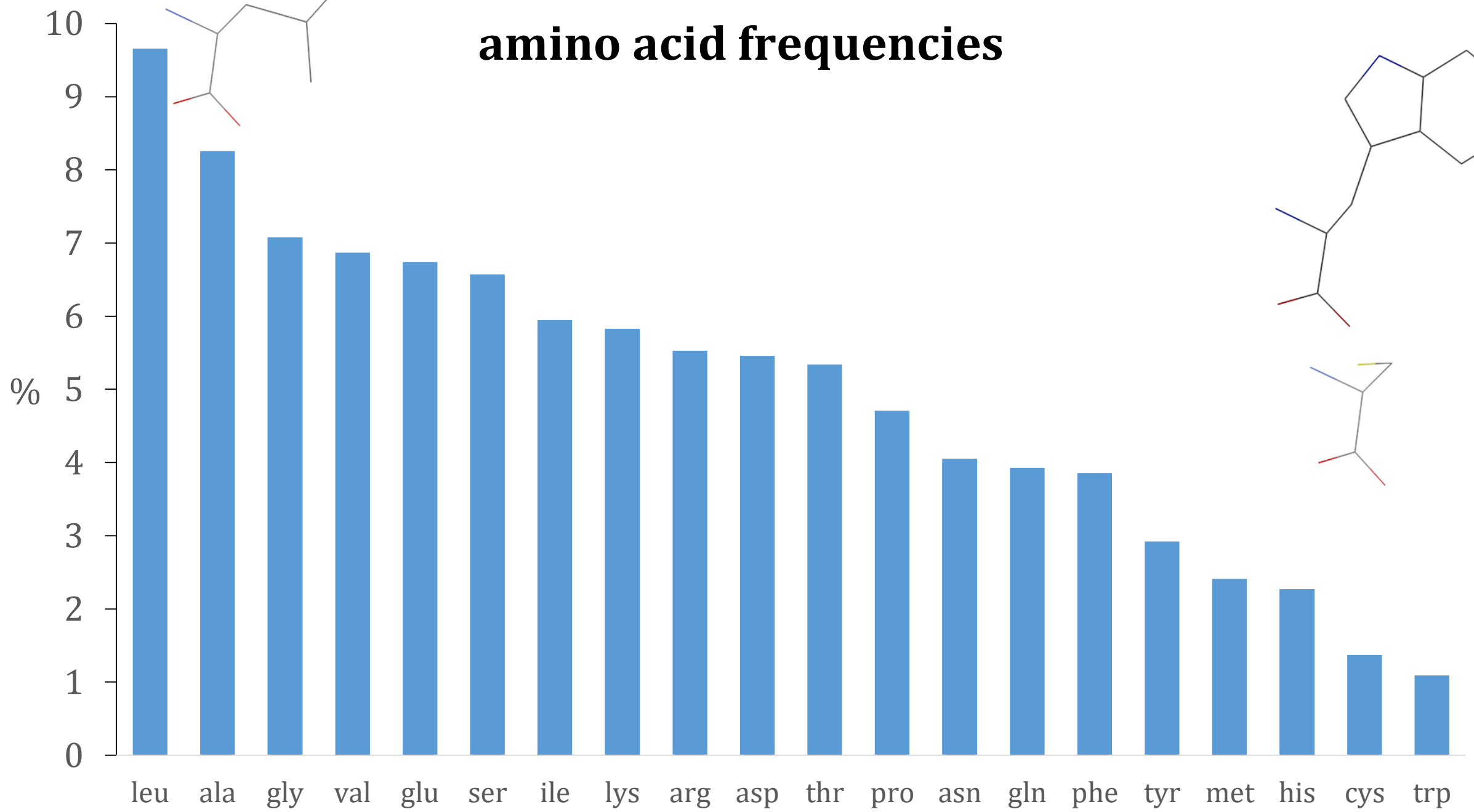
Surprise coming

- 20 amino acid types – are they all equally common ?
- Are you made of  $1/20 = 5\%$  of ala, leu, cys, ... ?





# amino acid frequencies



# What would Darwin say ?

Why ?

- so much ala, leu
- so little trp, cys, his, met

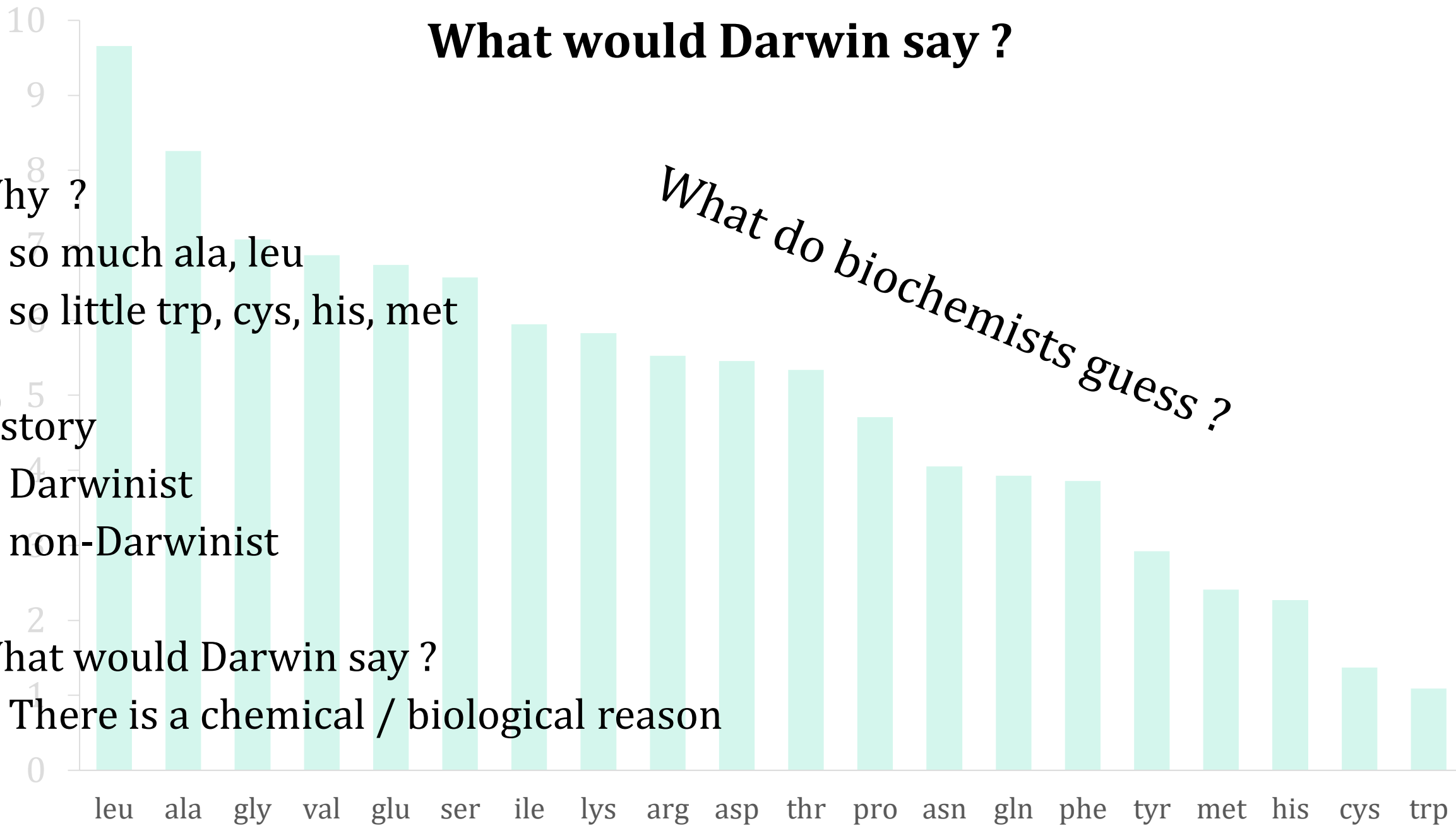
A story

- Darwinist
- non-Darwinist

What would Darwin say ?

- There is a chemical / biological reason

*What do biochemists guess ?*



# Think Darwinist

## Empirical fact

- trp, cys, met are rare in proteins

## Interpretation / explanation

- too much trp is bad for you / expensive / dangerous

## Possibilities

- metabolic cost issues
  - does it cost energy to make trp ? cys with its sulfur ?
- protein structure – lots of chemical differences between amino acid types
  - if you put lots of trp / cys / met in a protein
    - does it not fold ? Does it become unstable ?
- is free trp toxic ?

# Common amino acids

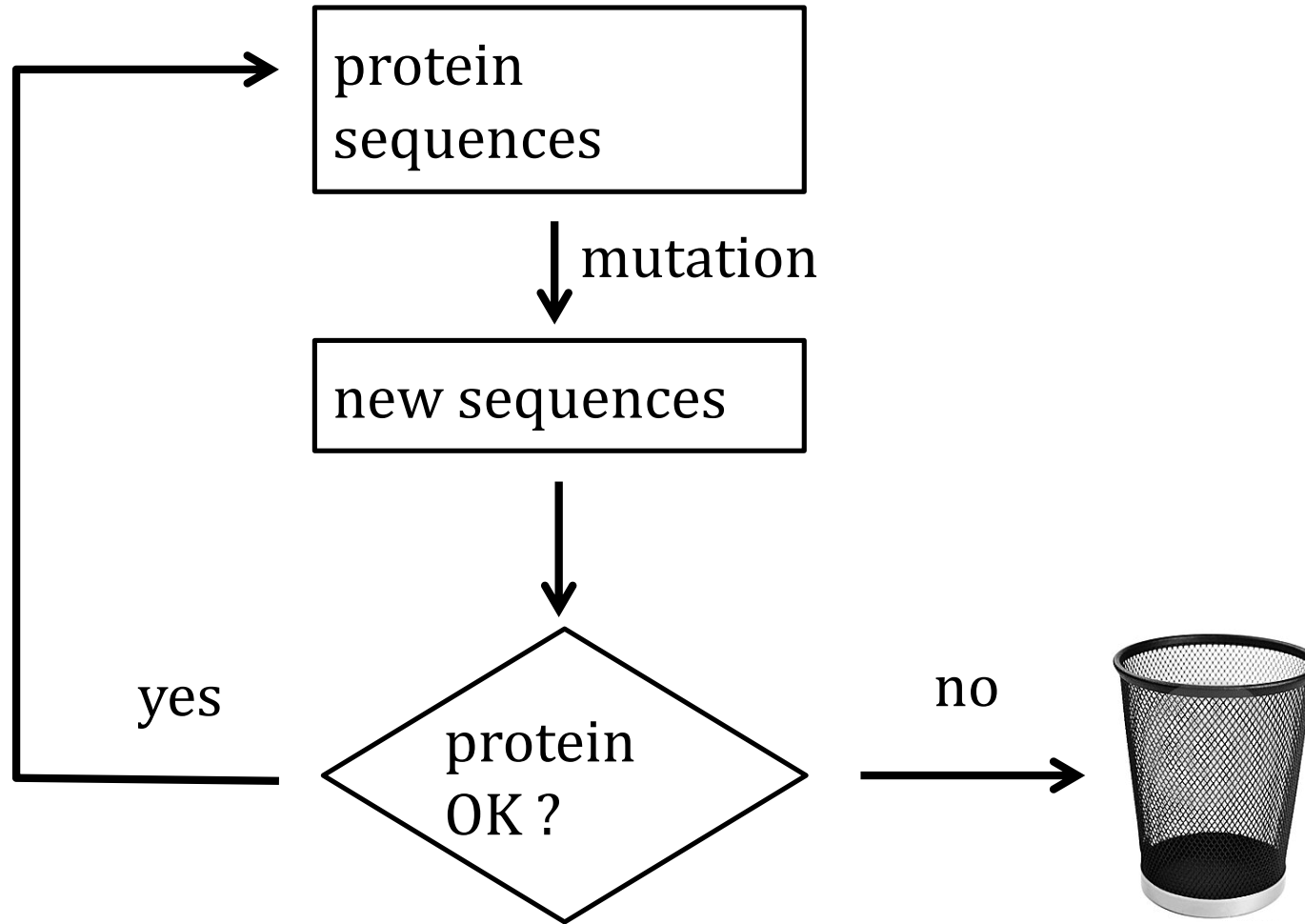
Leu and ala

- cheap to synthesise ?
- do you get them as by-products from other biochemistry ?
- what is their advantage in protein structure ?
  - stability ? rigidity ? flexibility ?

Forget Darwin – think neutral evolution

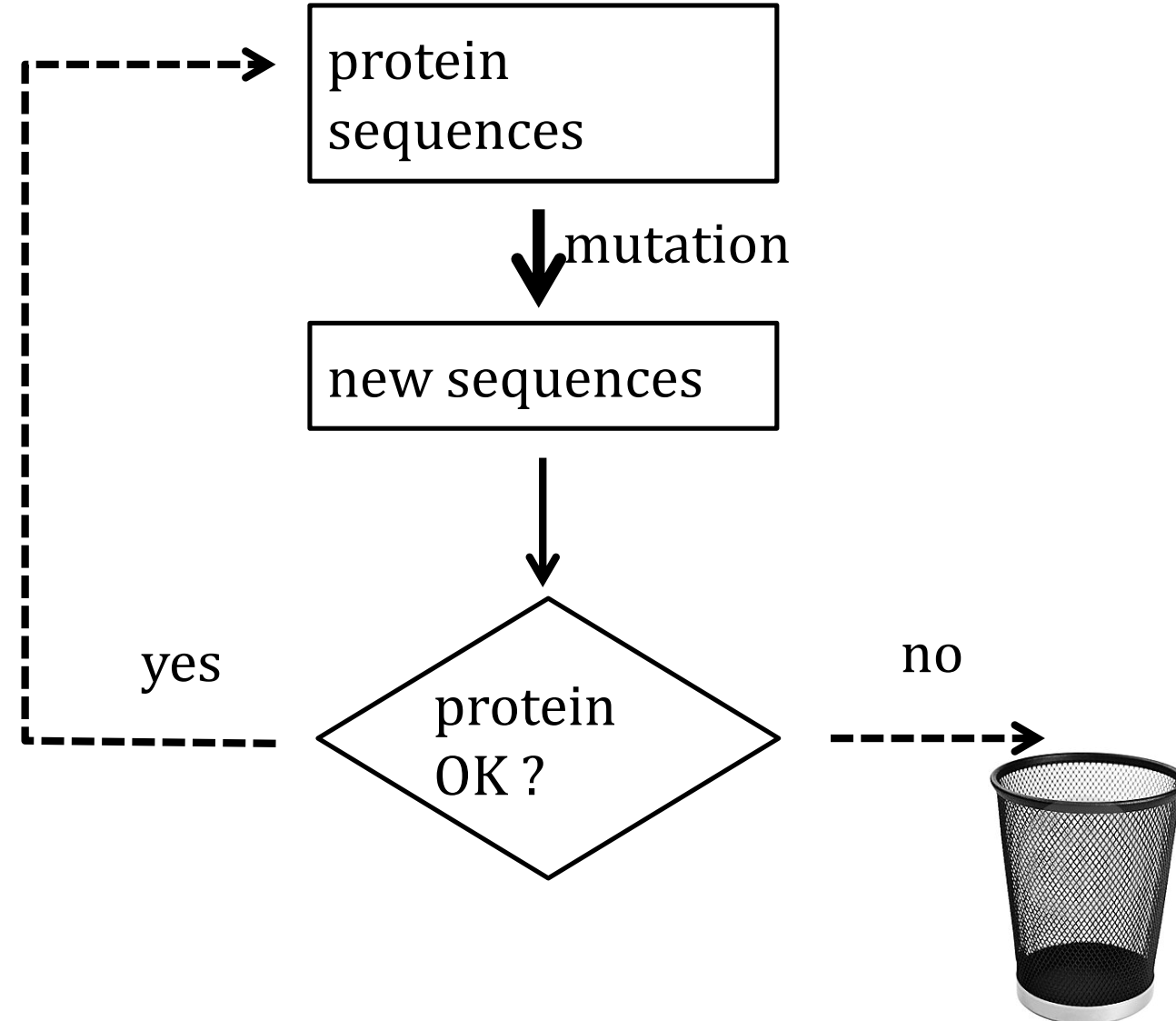
- what do we mean by Darwinism ?

# Very Darwinist



# Think neutralist

- OK/not OK step (selection) less important
- What determines the sequences you see ?
  - "mutation" step
- mutation step looks very simple
  - not really
- consider the meaning and biases



# Codon bias

- look at the most rare amino acids...

---

|     |                              |
|-----|------------------------------|
| ser | UCU, UCA, UCC, UCG, AGU, AGC |
|-----|------------------------------|

|     |                              |
|-----|------------------------------|
| leu | CUU, CUA, CUC, CUG, UUA, UUG |
|-----|------------------------------|

...

|     |          |
|-----|----------|
| his | CAU, CAC |
|-----|----------|

|     |     |
|-----|-----|
| met | AUG |
|-----|-----|

|     |     |
|-----|-----|
| trp | UGG |
|-----|-----|

---

- number of codons not quite everything

- some bases are more common than others

$$\begin{aligned}
 p(\text{his}) &= p(\text{C}) \cdot p(\text{A}) \cdot p(\text{C}) + p(\text{C}) \cdot p(\text{A}) \cdot p(\text{C}) \\
 &= 0.22 \cdot 0.30 \cdot 0.22 + 0.22 \cdot 0.30 \cdot 0.22 \\
 &\approx 0.03
 \end{aligned}$$

---

|   |      |
|---|------|
| U | 22 % |
|---|------|

|   |      |
|---|------|
| A | 30 % |
|---|------|

|   |      |
|---|------|
| C | 22 % |
|---|------|

|   |      |
|---|------|
| G | 26 % |
|---|------|

---

- does this predict the probability of all amino acids ?
- if yes, there is no selection for amino acids

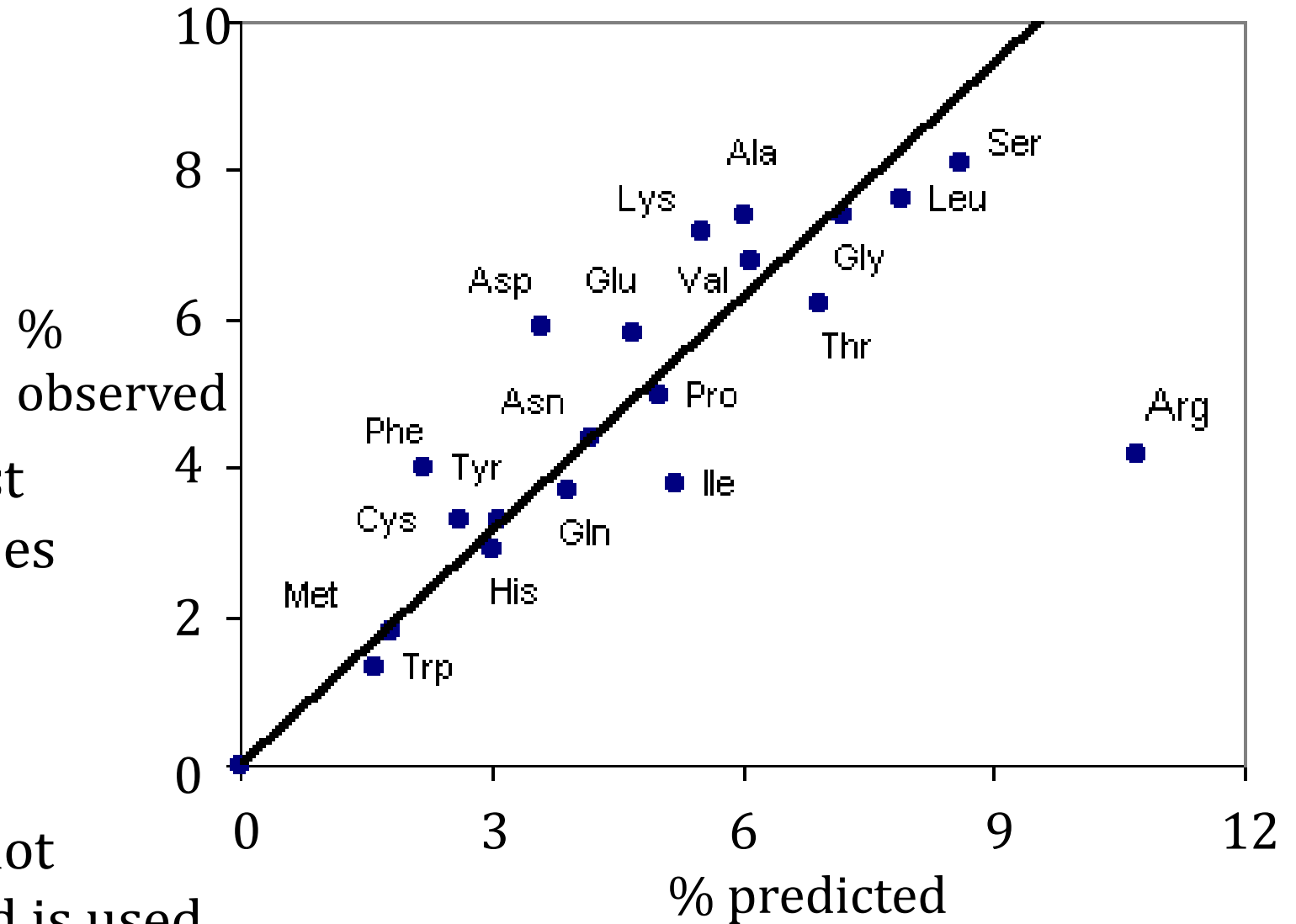
# How relevant is Darwinism ?

One outlier (arg)

Little evidence of Darwinist selection in amino acid types

Logical consequence

- there are many sites in proteins where it does not matter which amino acid is used





# Forget Darwinism and selection of amino acids ?

No

- arg example
- lots of mutation data
  - for an enzyme
    - most mutations are a bit bad, some do not matter
- Do not be a pure Darwinist
- do not interpret everything you see in terms of fitness

# Representation

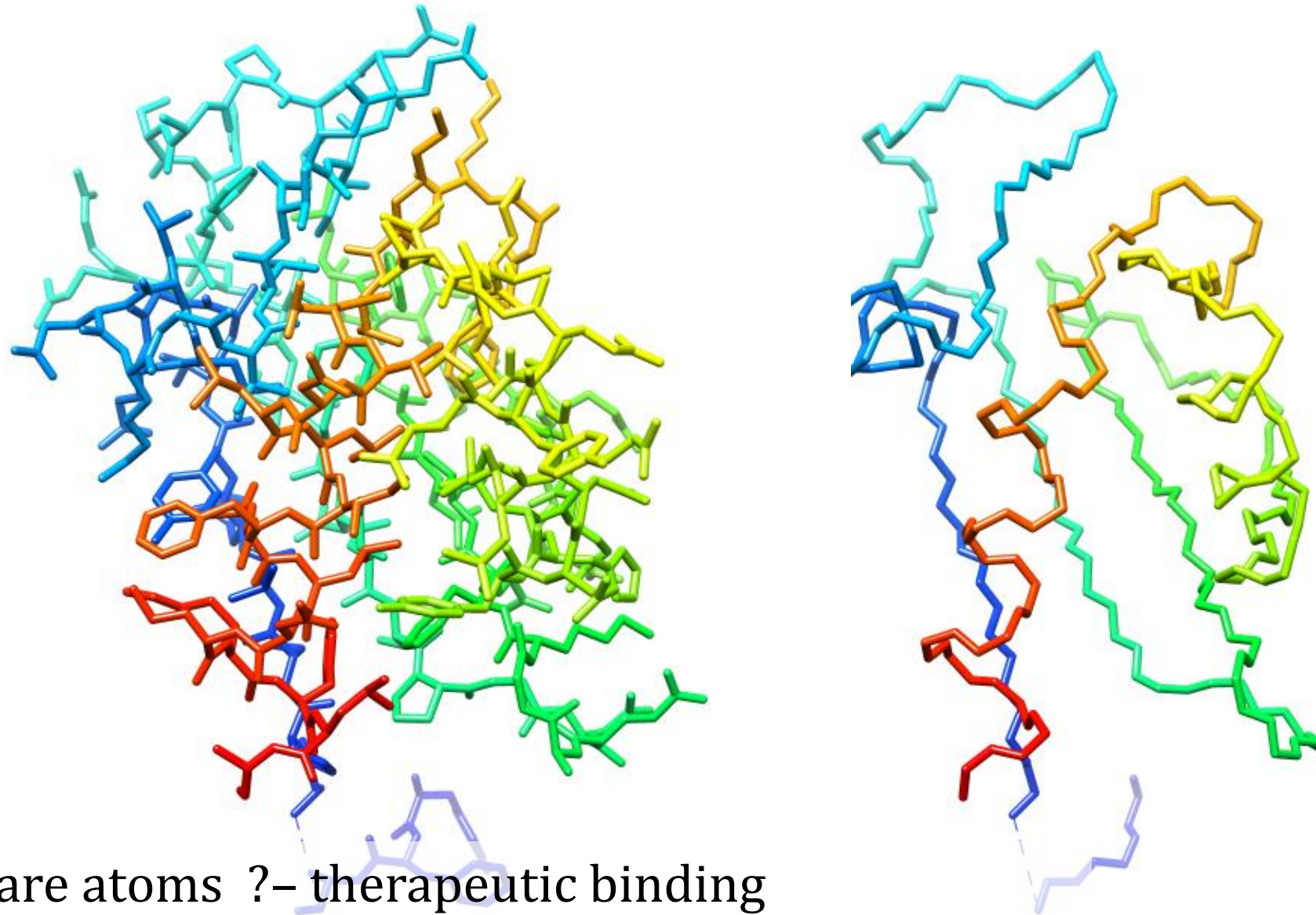
Ultimately, our representation of a structure...

|      |    |     |     |   |        |        |         |      |       |      |     |
|------|----|-----|-----|---|--------|--------|---------|------|-------|------|-----|
| ATOM | 1  | N   | ARG | 1 | 31.758 | 13.358 | -13.673 | 1.00 | 18.79 | 1BPI | 137 |
| ATOM | 2  | CA  | ARG | 1 | 31.718 | 13.292 | -12.188 | 1.00 | 14.26 | 1BPI | 138 |
| ATOM | 3  | C   | ARG | 1 | 33.154 | 13.224 | -11.664 | 1.00 | 18.25 | 1BPI | 139 |
| ATOM | 4  | O   | ARG | 1 | 33.996 | 12.441 | -12.225 | 1.00 | 20.10 | 1BPI | 140 |
| ATOM | 5  | CB  | ARG | 1 | 30.886 | 12.103 | -11.724 | 1.00 | 16.74 | 1BPI | 141 |
| ATOM | 6  | CG  | ARG | 1 | 29.594 | 11.968 | -12.534 | 1.00 | 15.96 | 1BPI | 142 |
| ATOM | 7  | CD  | ARG | 1 | 28.700 | 13.182 | -12.299 | 1.00 | 15.45 | 1BPI | 143 |
| ATOM | 8  | NE  | ARG | 1 | 27.267 | 12.895 | -12.546 | 1.00 | 12.82 | 1BPI | 144 |
| ATOM | 9  | CZ  | ARG | 1 | 26.661 | 13.087 | -13.727 | 1.00 | 17.38 | 1BPI | 145 |
| ATOM | 10 | NH1 | ARG | 1 | 27.370 | 13.558 | -14.735 | 1.00 | 18.38 | 1BPI | 146 |
| ATOM | 11 | NH2 | ARG | 1 | 25.367 | 12.797 | -13.838 | 1.00 | 25.73 | 1BPI | 147 |
| ATOM | 12 | N   | PRO | 2 | 33.800 | 13.936 | -10.586 | 1.00 | 17.07 | 1BPI | 148 |
| ATOM | 13 | CA  | PRO | 2 | 34.976 | 13.367 | -9.840  | 1.00 | 14.99 | 1BPI | 149 |
| ATOM | 14 | C   | PRO | 2 | 34.960 | 11.922 | -9.660  | 1.00 | 13.11 | 1BPI | 150 |
| ATOM | 15 | O   | PRO | 2 | 33.962 | 11.306 | -9.391  | 1.00 | 10.57 | 1BPI | 151 |
| ATOM | 16 | CB  | PRO | 2 | 34.922 | 14.145 | -8.523  | 1.00 | 15.81 | 1BPI | 152 |
| ATOM | 17 | CG  | PRO | 2 | 34.058 | 15.391 | -8.737  | 1.00 | 18.91 | 1BPI | 153 |
| ATOM | 18 | CD  | PRO | 2 | 33.371 | 15.273 | -10.096 | 1.00 | 19.41 | 1BPI | 154 |
| ATOM | 19 | N   | ASP | 3 | 33.000 | 8.730  | -9.707  | 1.00 | 8.73  | 1BPI | 155 |

x, y, z  
coordinates

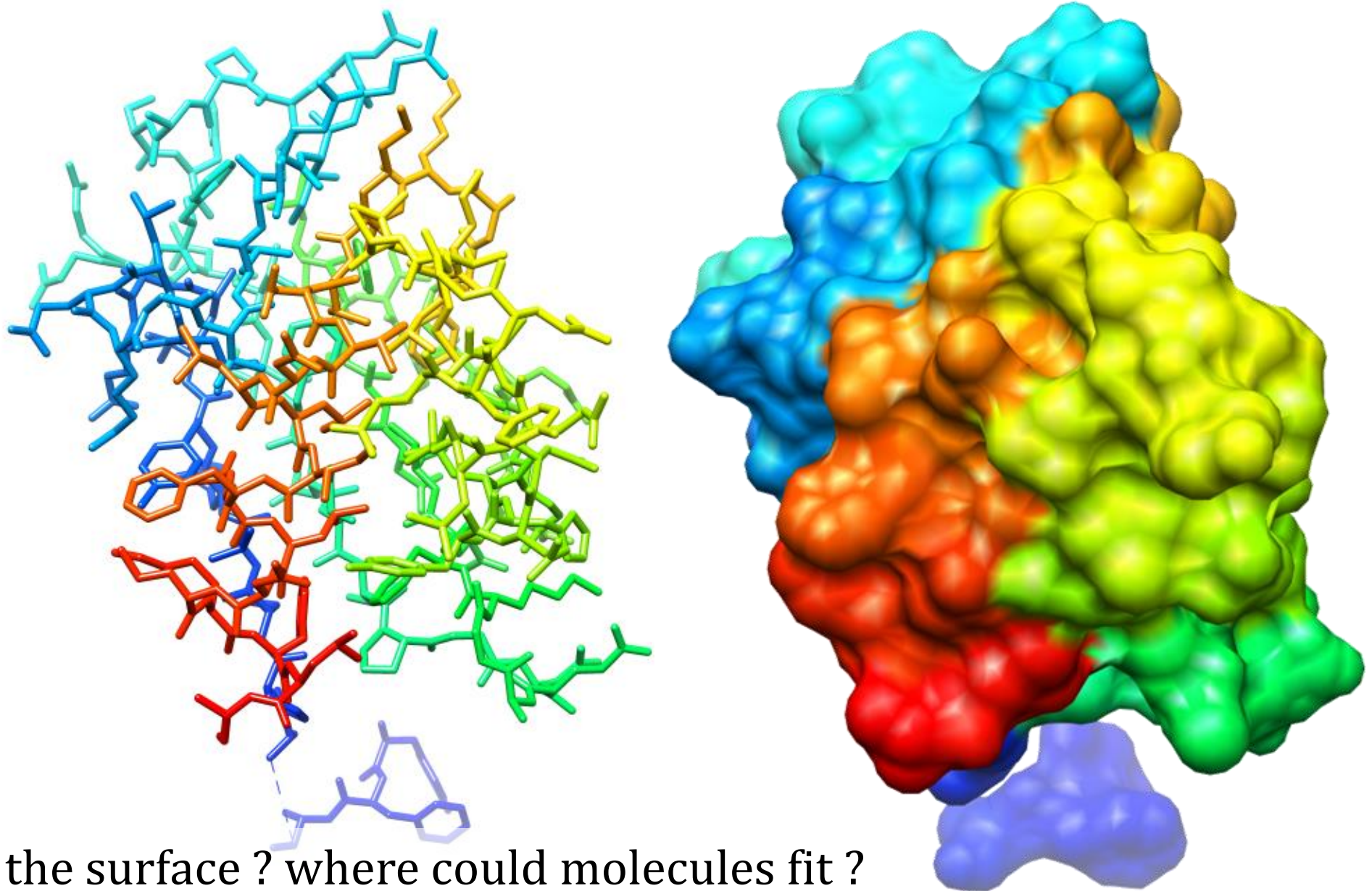
Drawing the structure ?

# Representations



- where are atoms ? – therapeutic binding
- which residues could be involved in interactions ?

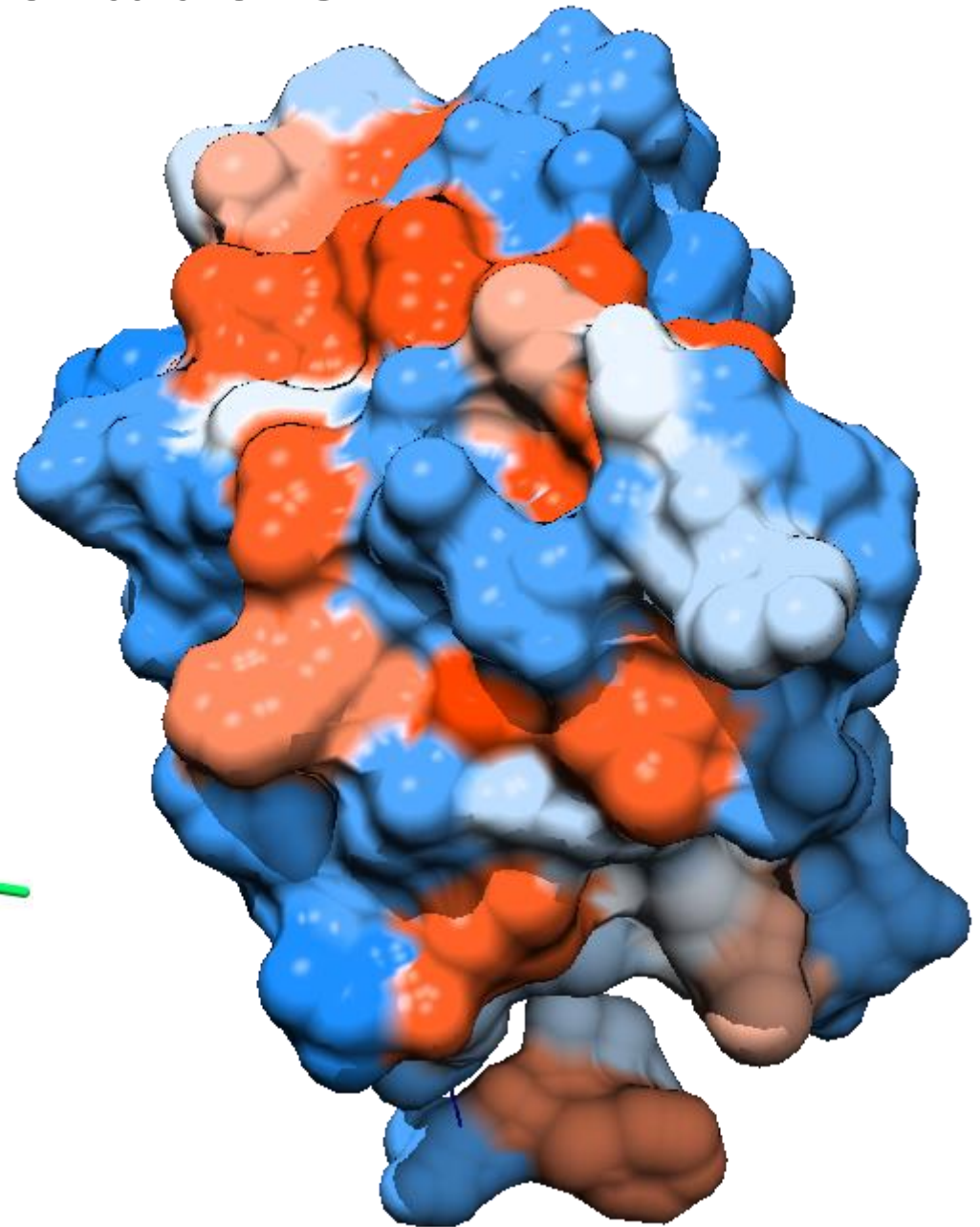
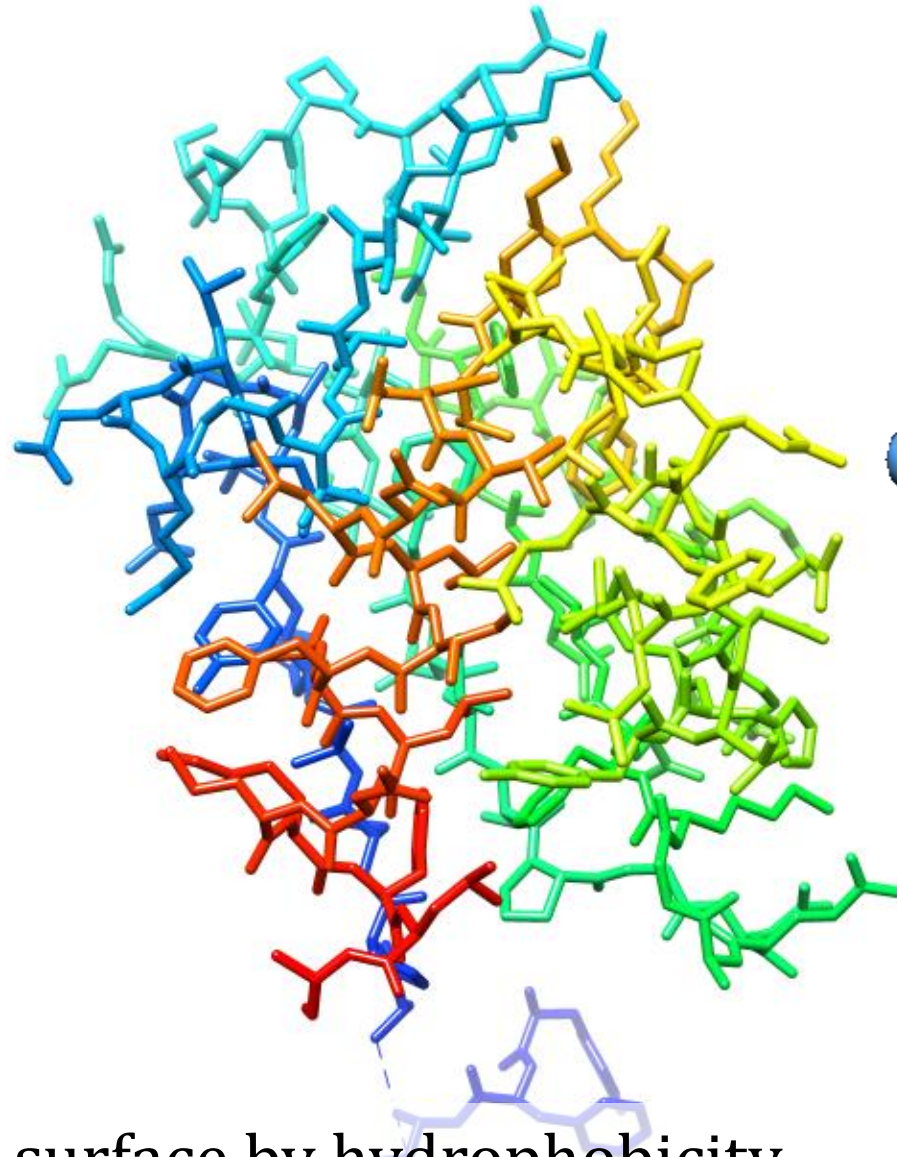
# Representations



What is the surface ? where could molecules fit ?

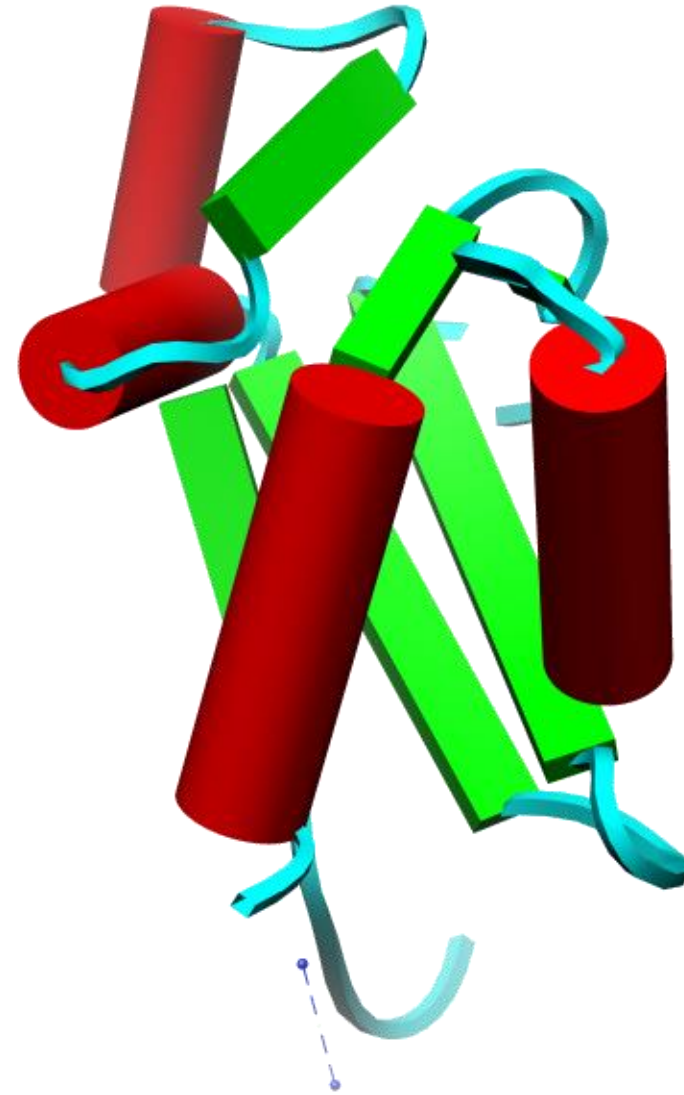
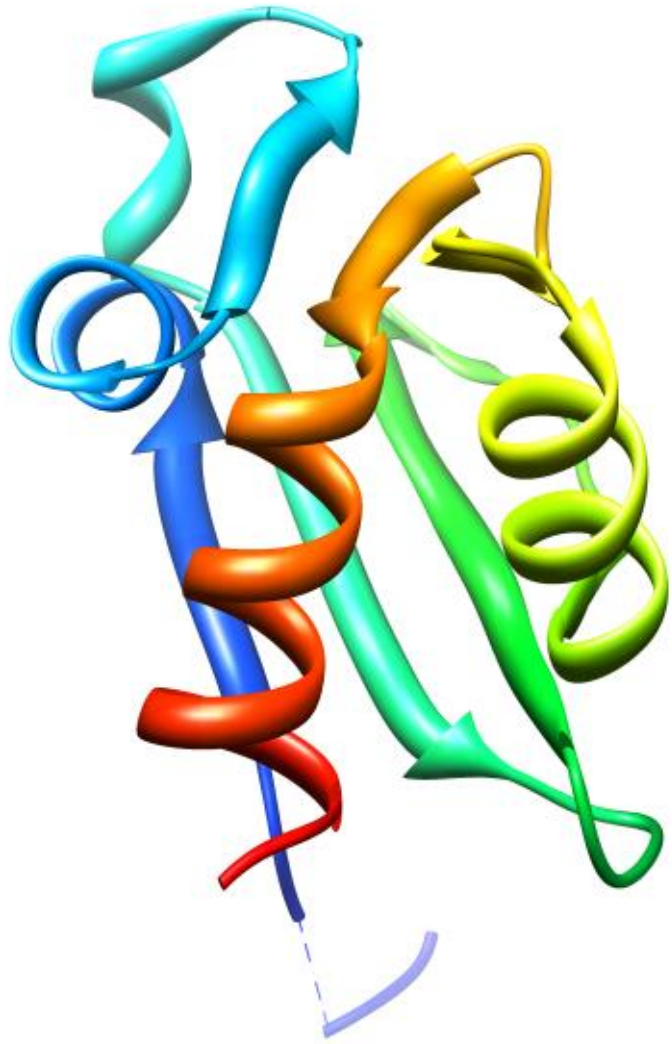


# Representations



Colour surface by hydrophobicity

# Representations



Highlight / emphasise regular structures

# Why does structure matter ?

- what residues can I change and preserve function ?
- what is the reaction mechanism of an enzyme ?
- what small molecules would bind and block the enzyme ?
- is this protein the same shape as some other of known function ?

# Where do structures come from ?

- X-ray crystallography
- NMR
- + a bit of small angle X-ray scattering, electron diffraction, neutron diffraction...

# resolution, precision, accuracy

Coordinates 27.370 13.558 -14.735

- what do they mean ?

Random errors

- non-systematic / noise / uncertainty
- should be scattered around correct point

X-ray crystallography has model for data

- uncertainty (probability)
- resolution (experimental)
  - $< 1 \text{ \AA}$  (unusually good)
  - $> 5 \text{ \AA}$  (bad, but examples..)

3LJ5 Full Length Bacteriophage P22 Portal Protein

3M0C X-ray Crystal Structure of PCSK9 in Complex with the LDL receptor



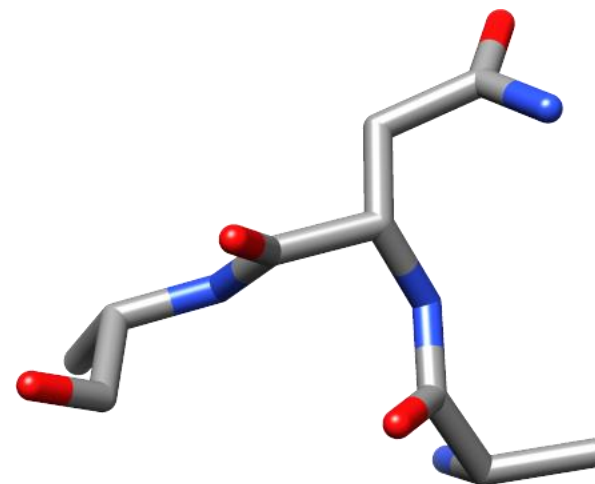
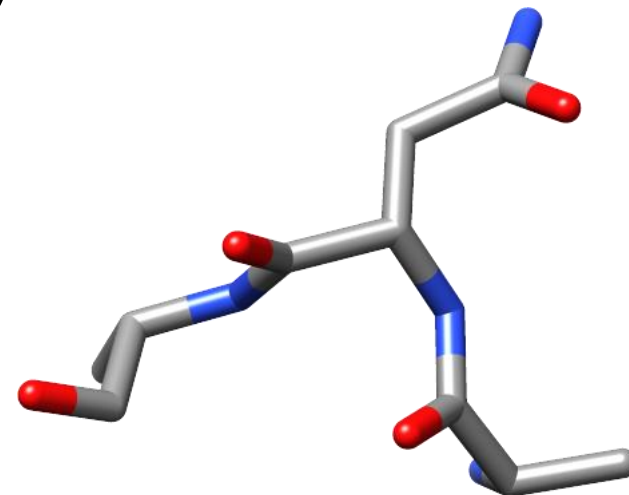
# X-ray crystallography

## Non-systematic errors

- small problems: (O and N look the same)
- few huge problems
- newer structures are better

## Proteins are not static

- overall motion
- local motion



# NMR structures

Different philosophy to X-ray

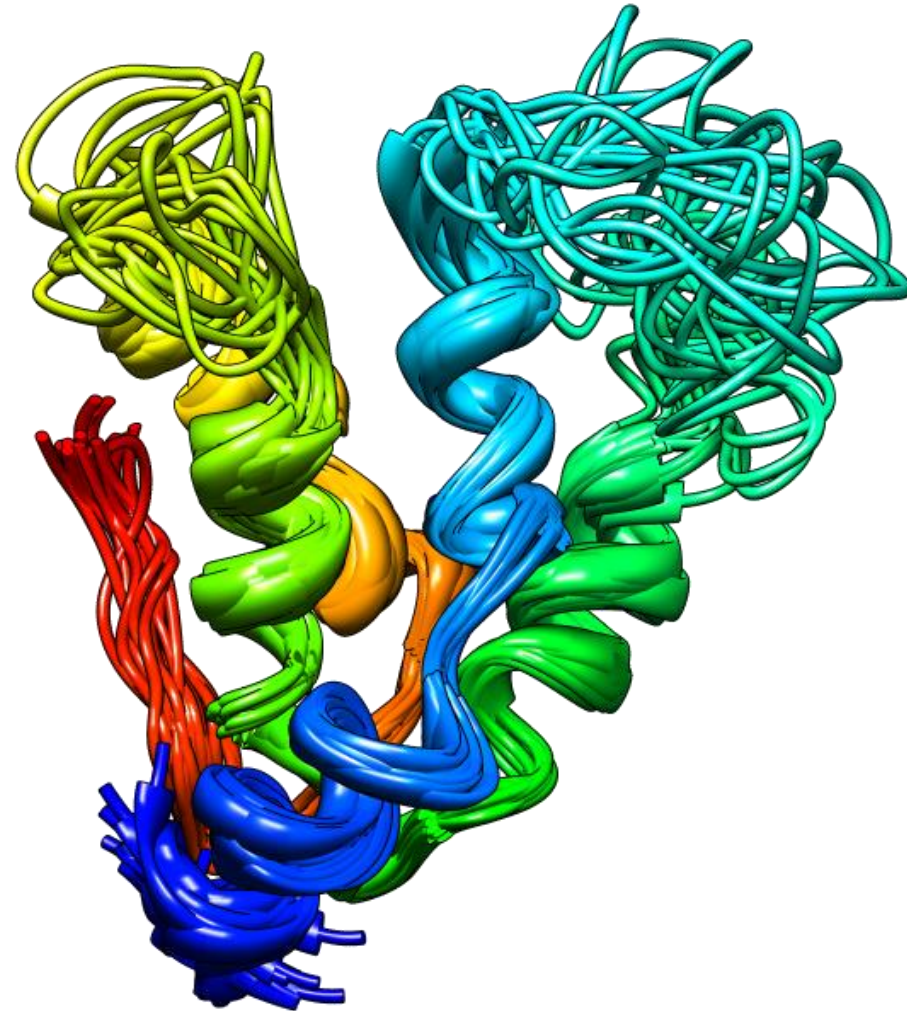
- lots of little internal distances
- do not quite define structure

Generate 50 or  $10^2$  solutions

- look at scatter of solutions

As with X-ray

- some parts are well defined
- some not



# Summarise and stop

- roles of proteins
- heteropolymers – 20 types of amino acid / residue
- geometry – avoiding atomic clashes, forming H bonds
  - leads to regular secondary structure
- chemistry of amino acids very different to another
- unique structure for a sequence reflects these differences
- representations of structures
- structures in PDB are experimental – have errors

# some questions

- (Asp)<sub>100</sub>
  - is it soluble ? Is it acidic / basic ?
  - would it form a compact regular structure ?
- Why does a protein fold to a specific structure, but polyethylene not ?
- Glycine can reach parts of the angle space which other residues cannot. What are the evolutionary consequences ?
- How big is sequence space ? How much has been tried by evolution ?
- if you have a protein of poly-trp, would it form a specific structure ? How would it behave in solution ?
- for length  $n$ , do all / many / few of the  $n^{20}$  sequences form specific structures ?
- how would a Darwinist explain the uneven distribution of amino acid usage ?
- why would you want to represent a protein by its surface ?
- why might you draw it as a series of helices and strands ?
- what is the biggest chain in the protein data bank ? Examples
  - fatty acid synthase  $> 2 \times 10^3$  residues/chain
  - dynein heavy chain motor domain  $> 4 \times 10^3$  residues/chain