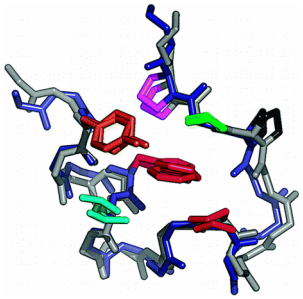
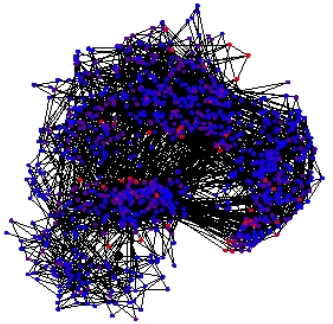
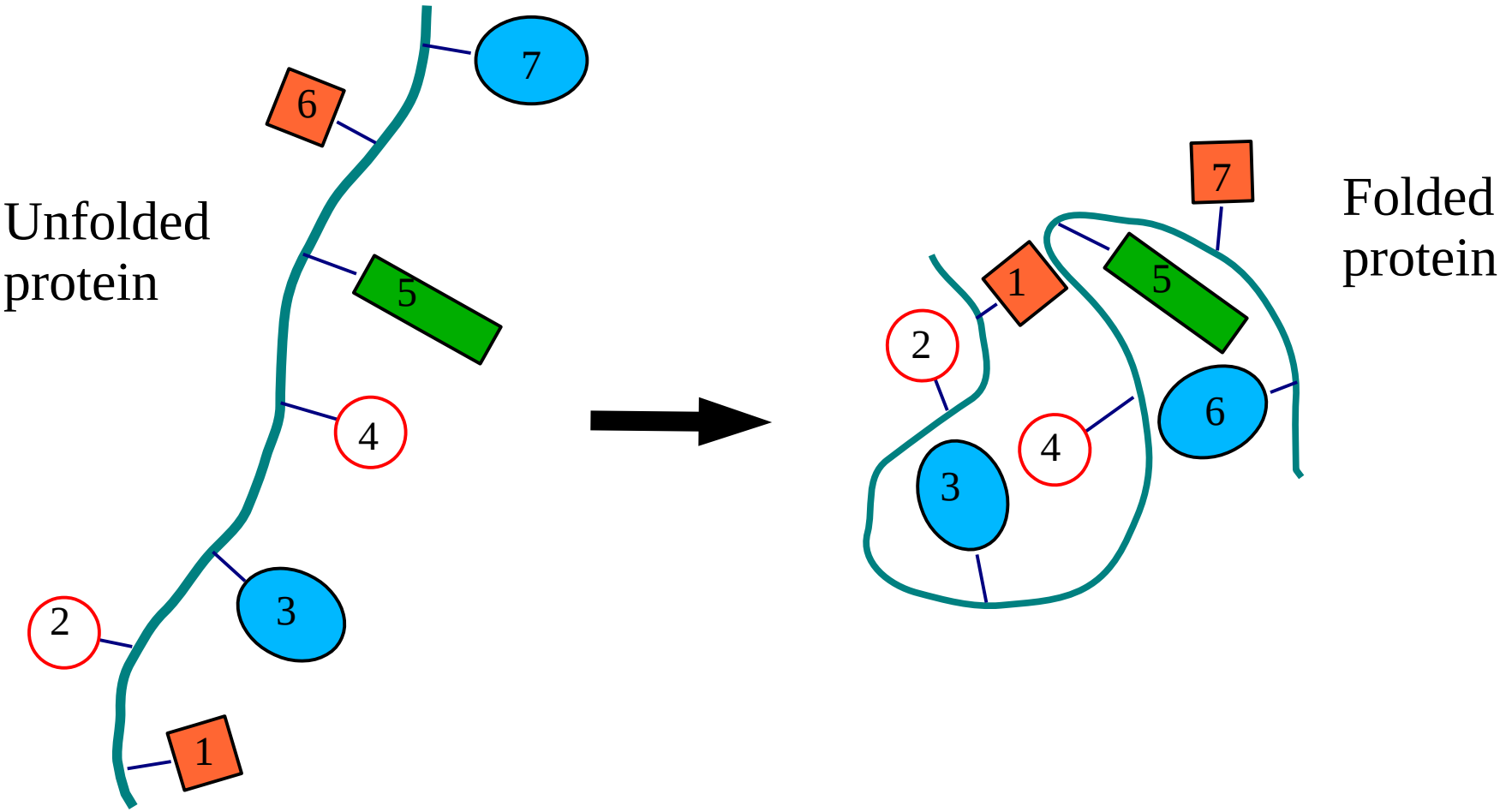




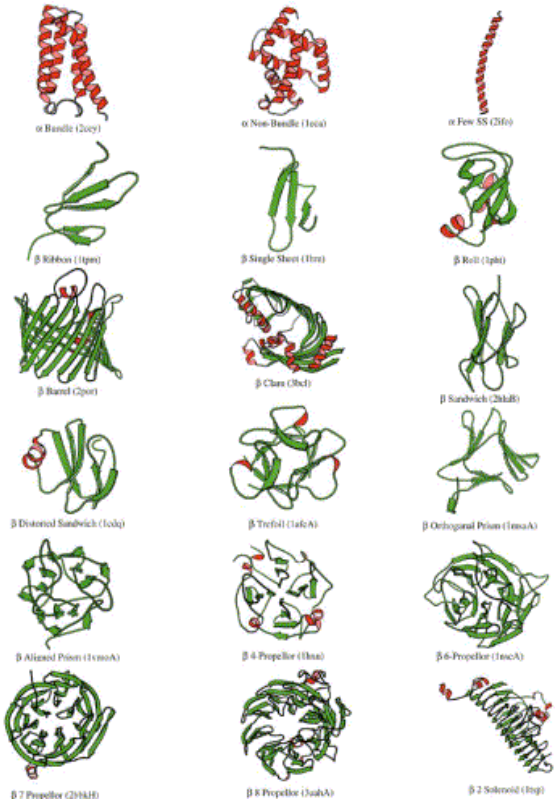
Introduction to computational protein design

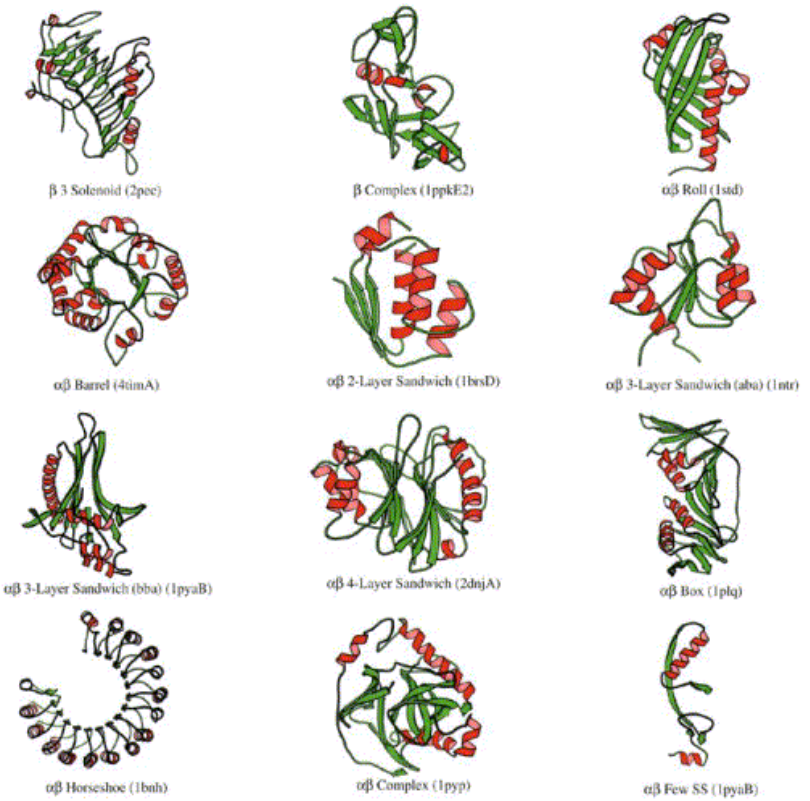
Thomas Simonson
Laboratoire de Biologie Structurale de la Cellule
Ecole Polytechnique, Paris
thomas.simonson@polytechnique.fr
http://biology.polytechnique.fr/biocomputing
https://proteus.polytechnique.fr

The amino acid sequence directs the polypeptide to a particular «native» structure



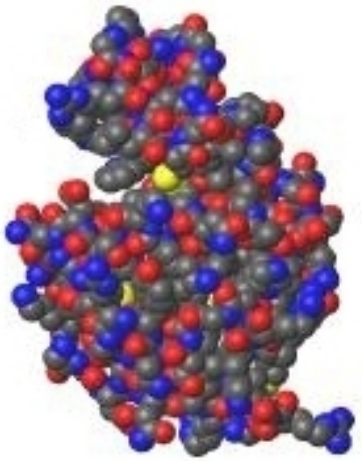
The amino acid sequence directs the polypeptide to a particular «native» structure





The protein folding problem: can we predict protein structures?

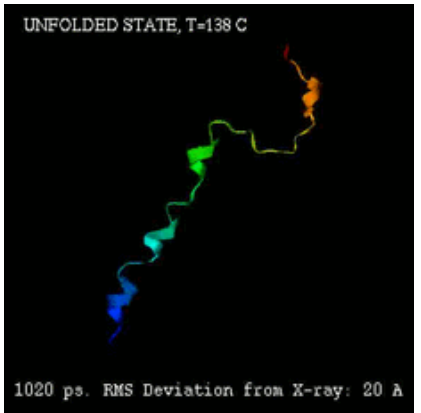
K
L
H
G
G
P
M
L
D
S
D
Q
K
F
W
R
T
P
A
A
L
H
Q
N
E
G
F
T



$N_{states} \sim 10^{n}$

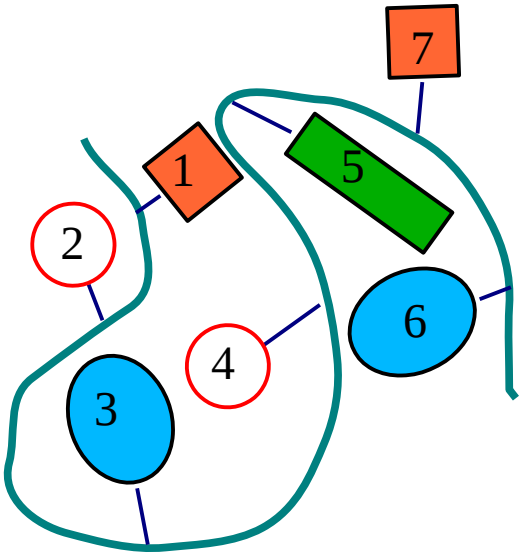
n = 100-300

googols of possibilities...



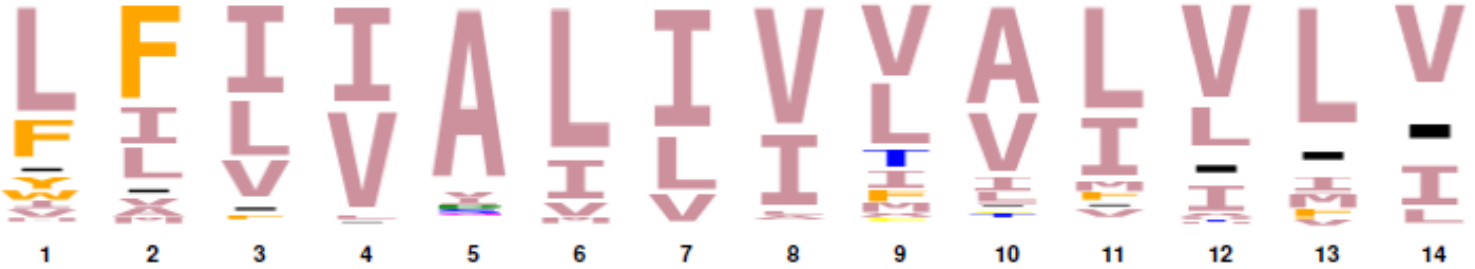
Inverse folding problem: for a given “backbone structure”, or fold, which sidechains “fit”?

The size of the problem is enormous.....
100 amino acids ⇒ 10^{130} sequences



Folded protein

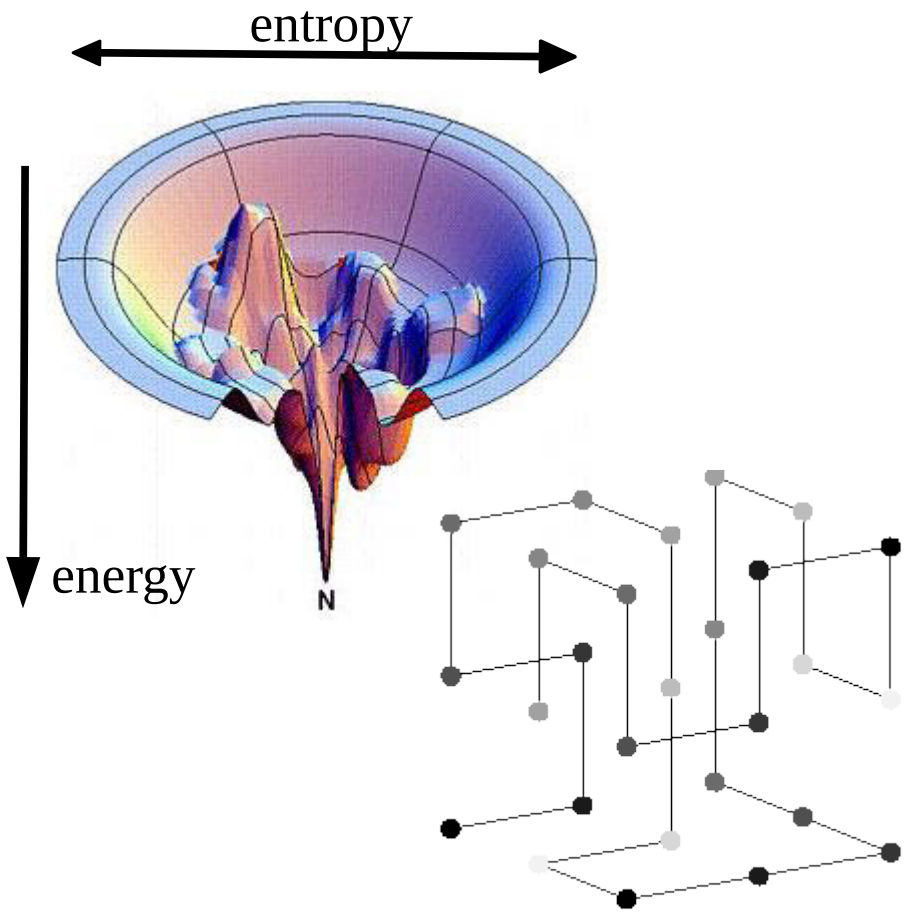
Some natural PDZ sequences (core positions)

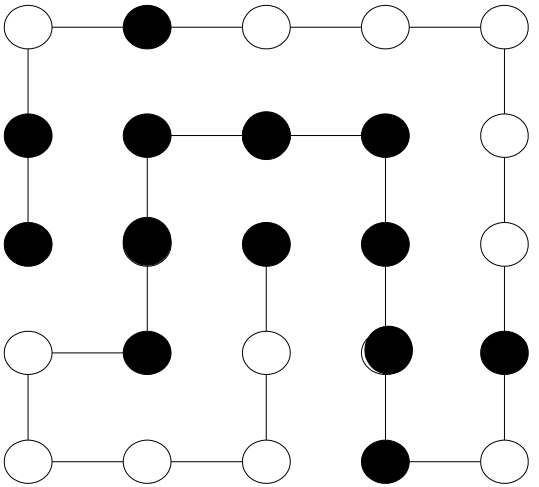


>10,000 natural homologues in UniProt....

The inverse folding problem:

"Easy" to solve with toy models



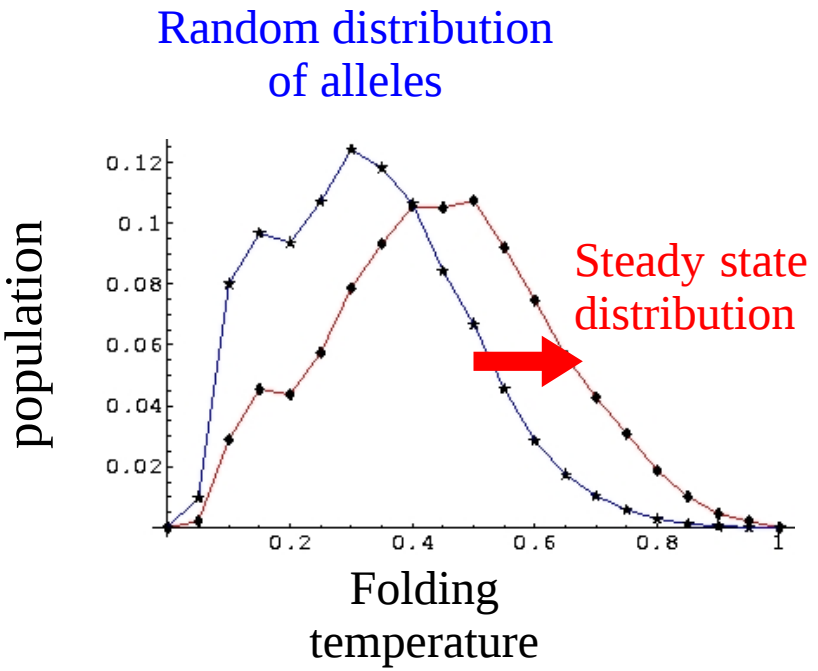


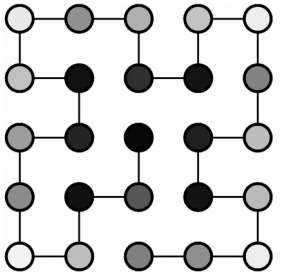
1081 conformations
33.554.432 sequences

- **Designability correlates with fitness, thermophilic adaptation**
- **Of interest for protein engineering and biotechnology**

The inverse folding problem

Easy to solve with toy models of structure and population dynamics





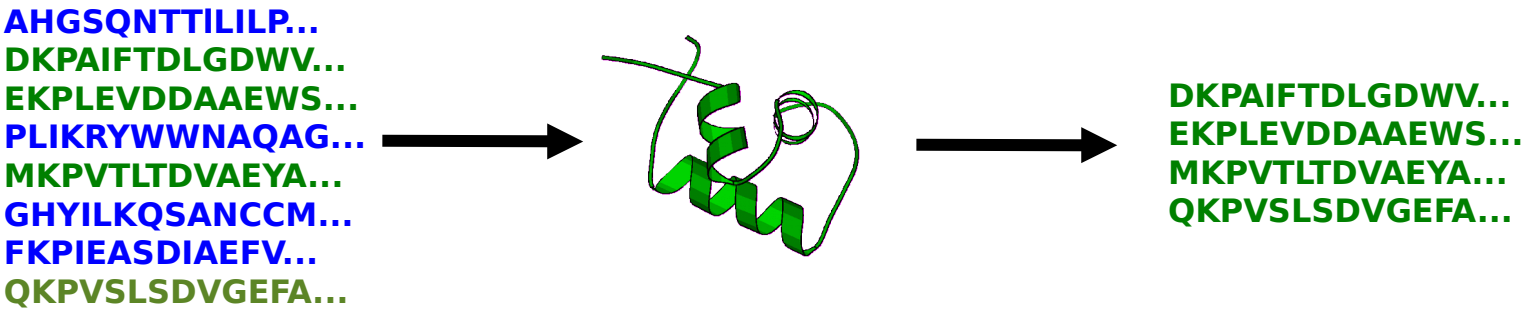
Noirel & TS, J Chem Phys '08

Steady state population enriched in sequences that are thermostable
and tolerant of mutations (designable)

Protein design: search sequence/conformation space for preferred/functional sequences

random mutagenesis + selective pressure

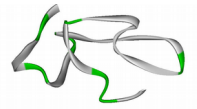
Ponder, Richards (1987); Hellinga, Richards (1994); Mayo (1997) Desjarlais (1998) Koehl, Levitt (1999); **Baker**, Serrano, Wodak, Handel (2000-04); others...



Sequence score: $\Delta G = G_{folded} - G_{unfolded}$

Free energy function

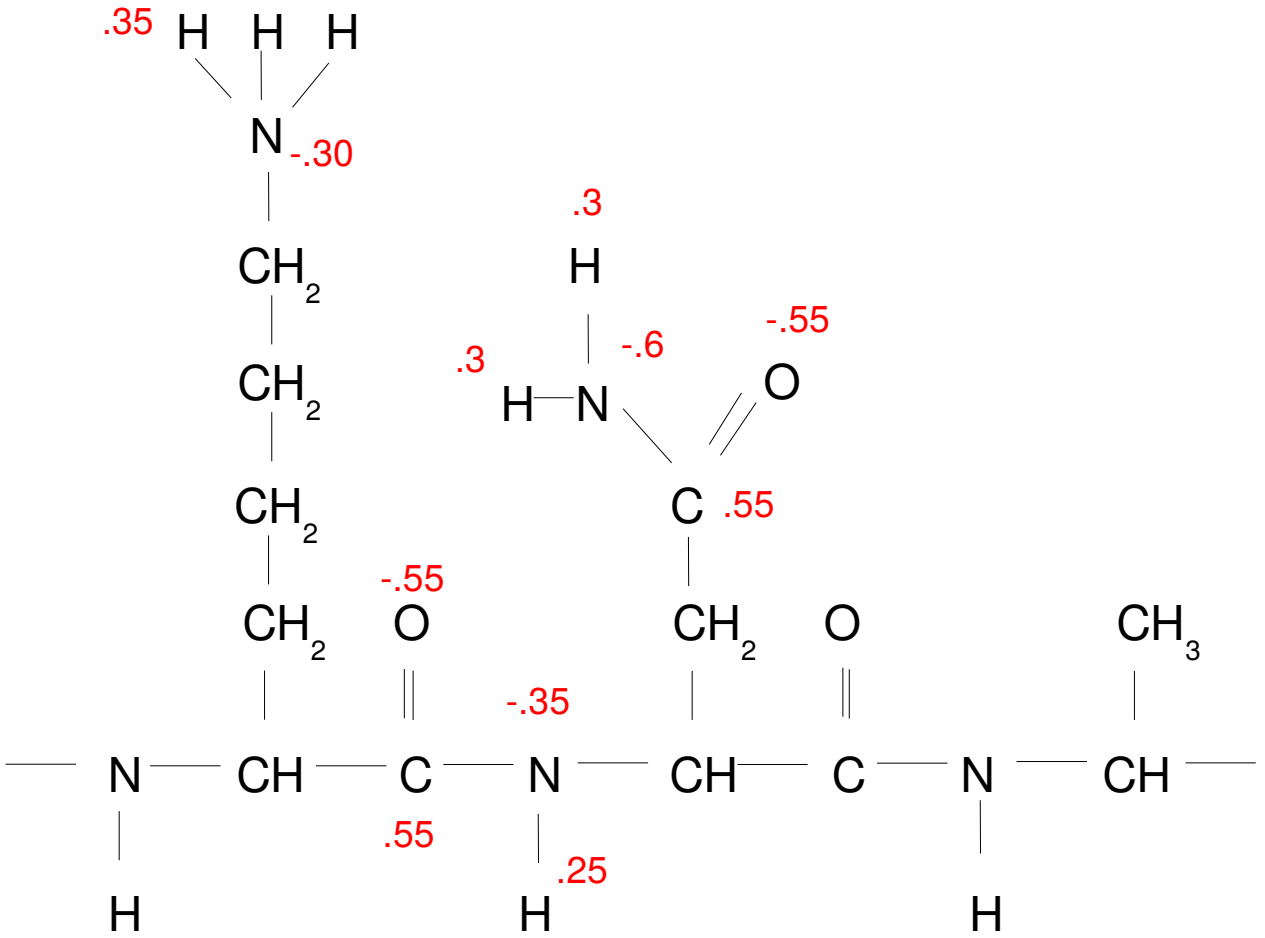
Structural model of both ensembles

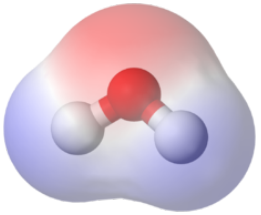


Molecular mechanics energy function

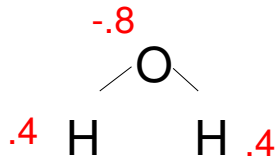
$$U = \Sigma_{liaisons}\, k_b\,(b-b_0)^2 + \Sigma_{angles}\, k_a\,(a-a_0)^2 + \Sigma_{torsions}\, k_t\,[1+\cos(nt-\tau)]$$

$$+ \quad \Sigma_{ij}\,[\, A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^{6} + q_i q_j / r_{ij}\,]$$



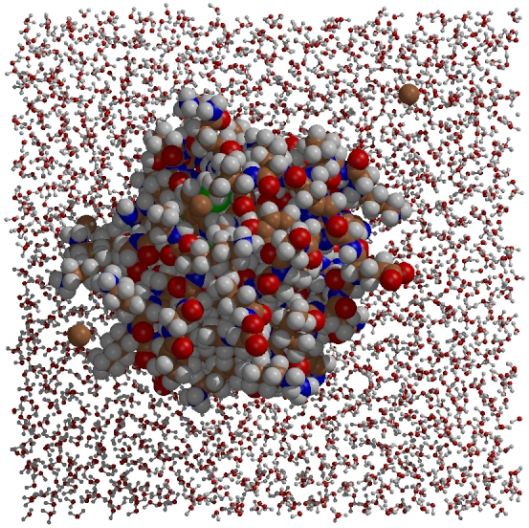


eau

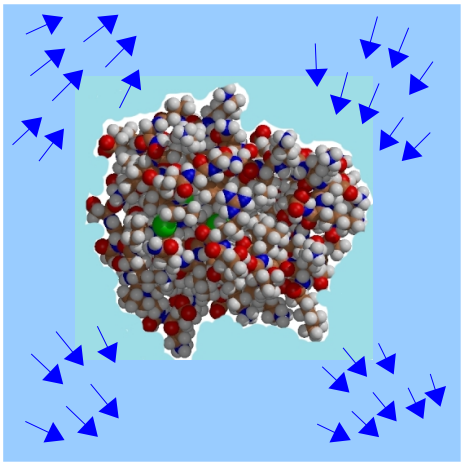


Implicit solvent: polar free energy components

Explicit solvent



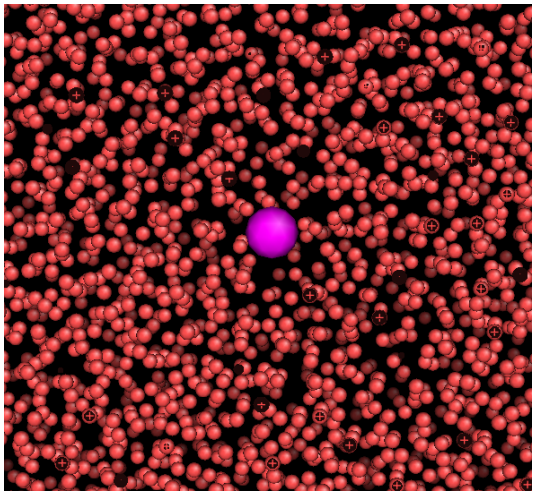
Implicit solvent



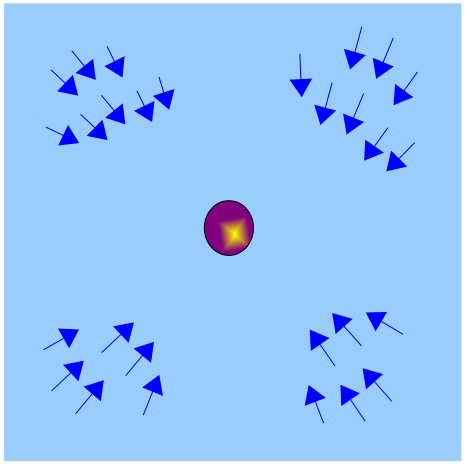
Dielectric continuum

Implicit solvent

“explicit”



“implicit”



Dielectric continuum

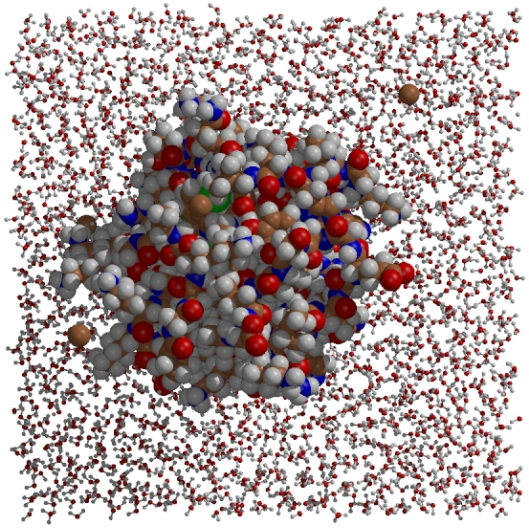
Electrostatic energy of a pair of ions: $U = q\,q' / \varepsilon\, r$

$\varepsilon = 80 =$ dielectric constant of water

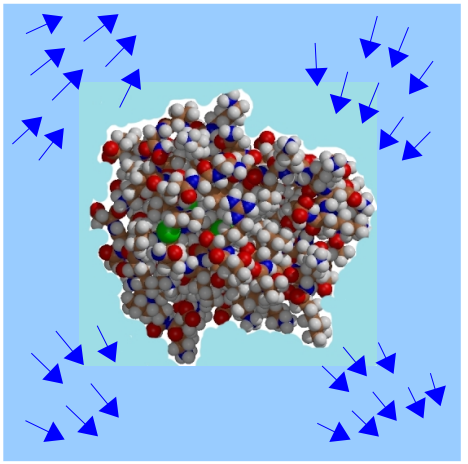
Fine for a few ions in water.....

Implicit solvent

“explicit”



“implicit”



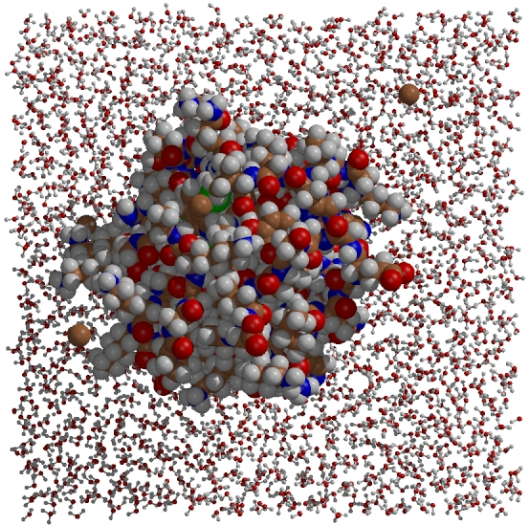
Dielectric continuum

A biomolecule is more complex: heterogeneous system

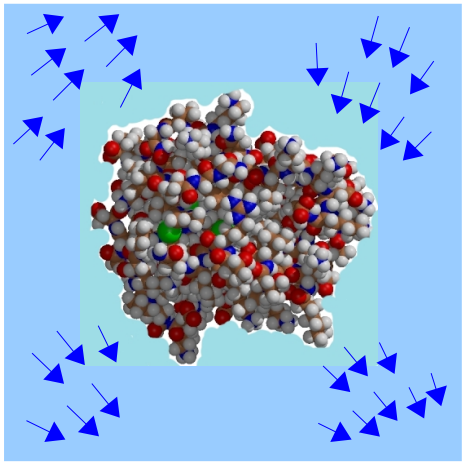
Going further: more sophisticated model

Implicit solvent: "generalized Born" model (GB)

"explicit"



"implicit"

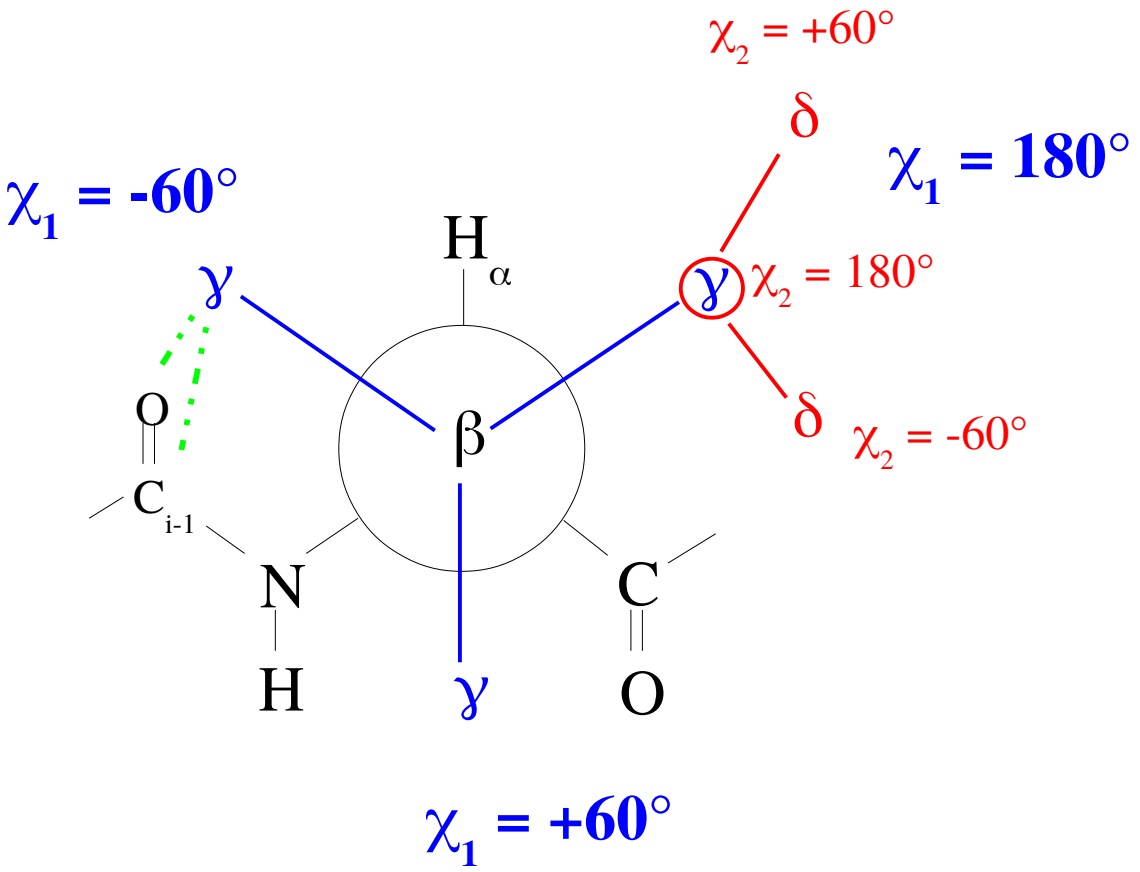


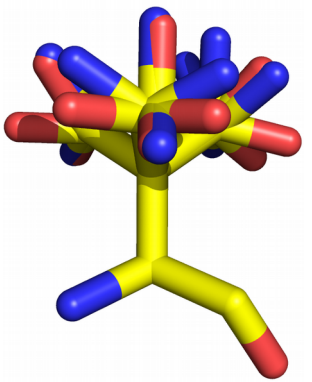
- Buried charges: $E = q\,q' / r$
- Exposed charges: $E = q\,q' / \varepsilon\, r$

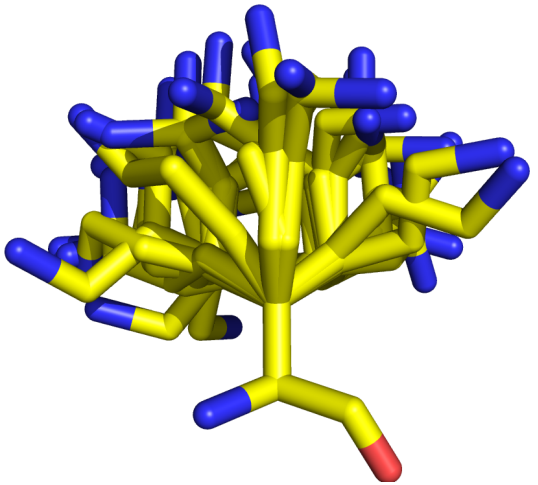
  $\varepsilon = 80$
- Intermediate cases: **interpolation**

Discrete conformational space

Conformational space
=
fixed backbone
+
sidechain
“rotamers”

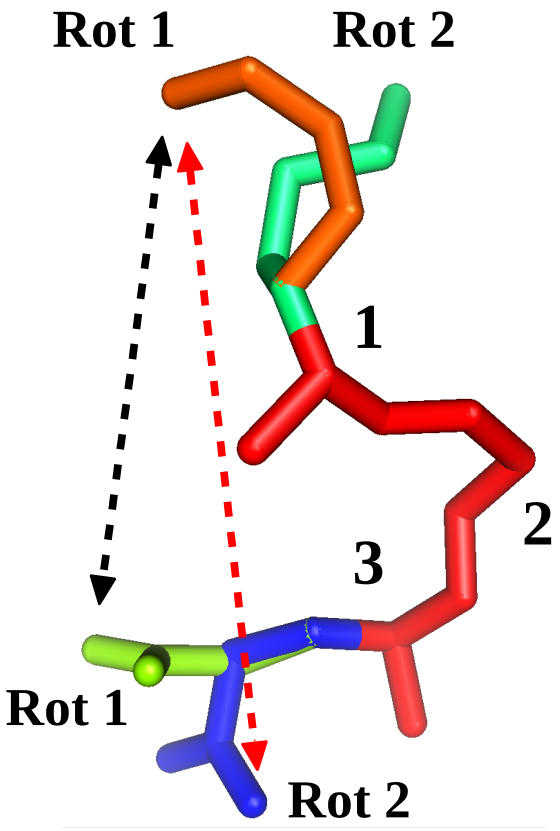






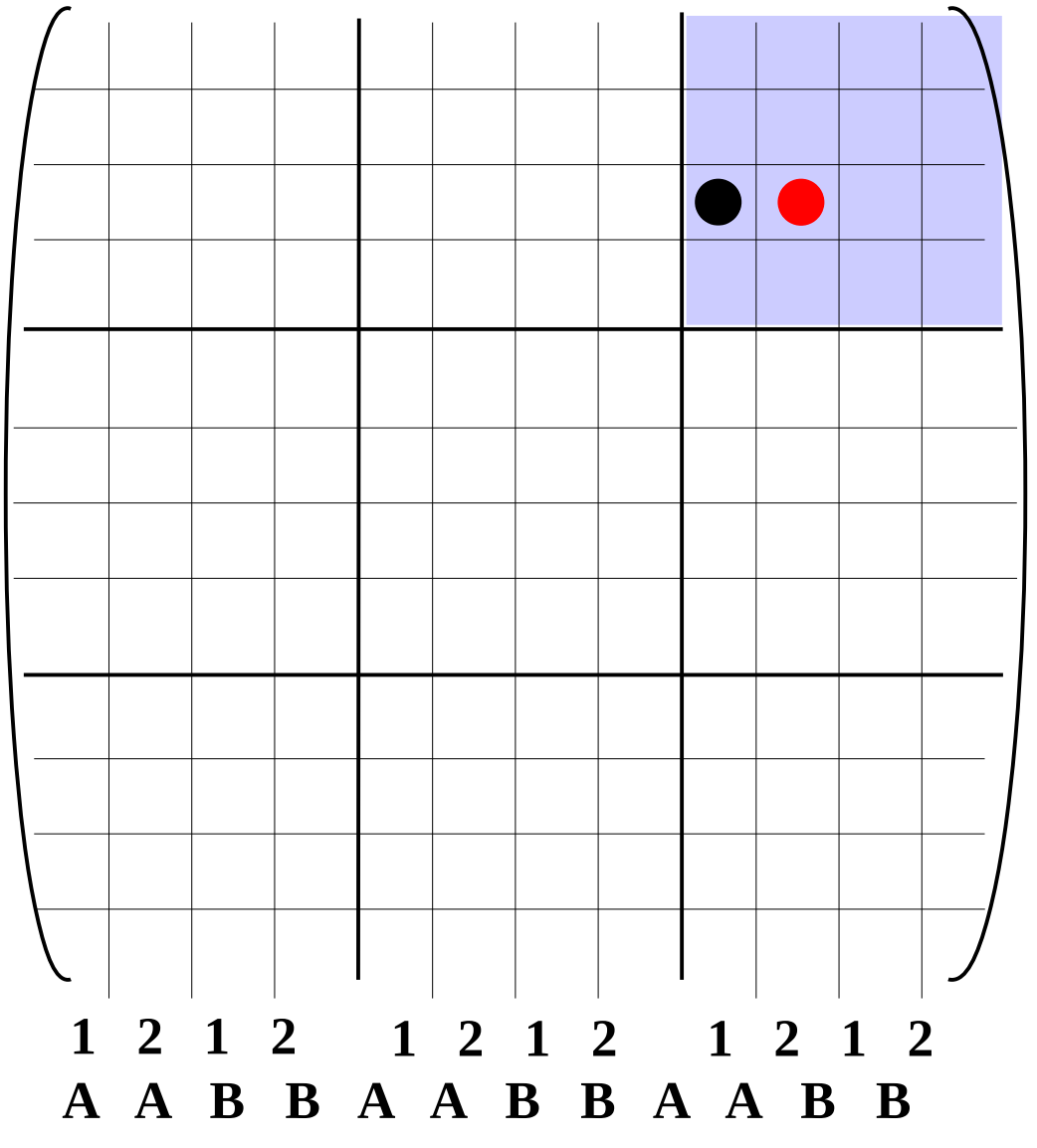
1) Precompute all pair interactions, allowing for all residue types and rotamers (Mayo, 1997)

Structure



Position	Type	Rotamer
1	A	1
	A	2
	B	1
	B	2
2	A	1
	A	2
	B	1
	B	2
3	A	1
	A	2
	B	1
	B	2

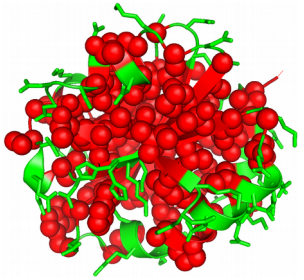
Energy matrix



Protein structure is complex...

Protein structure is simple!

Side chains are dynamic and flexible...



Side chains have just a few conformations !

Unfolded proteins form a molten glubule that is hard to characterize...

Unfolded proteins are just like an extended peptide !

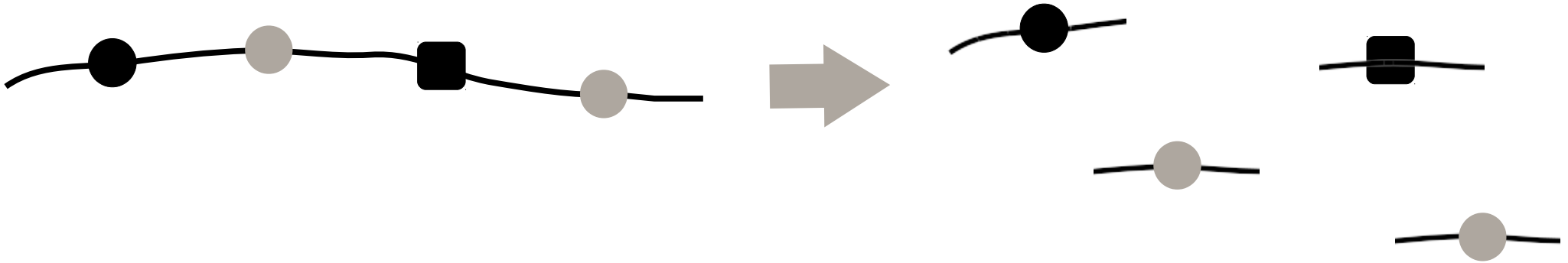
Aqueous solvent has a rich, complex structure...

Solvent is just like a simple dielectric !

The unfolded state : a simple model

extended peptide :

approximated as collection of tripeptides :



$$E_u = \Sigma_i E_u(t_i)$$

• Unfolded energy only depends on composition, through type-dependent amino acid *chemical potentials* ; no structural model.

• The chemical potentials can be computed from an extended peptide model or chosen empirically (see below).

• Longer peptides (5-, 7-mers) don't help (Pokola & Handel, JMB '05)

Maximum likelihood strategy for the unfolded state parameters

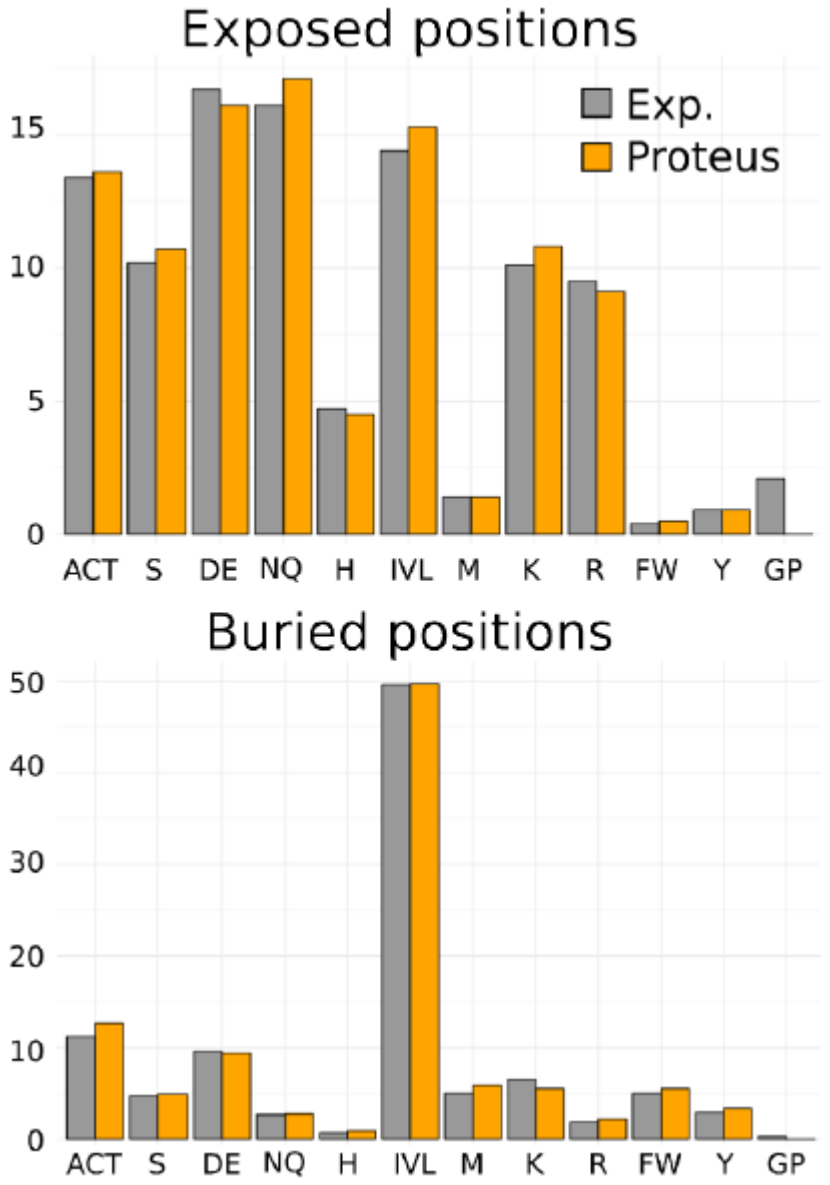
Choose a set of experimental sequences ***S*** = {S, S', ...} as a reference

Boltzmann probability of sampling a sequence is $p(S) = \exp(-\beta\Delta G)/Z$

ΔG folding free energy, Z normalization constant

Choose unfolded parameters $E_u(t)$ to maximize probability or likelihood ***L*** of ***S***

2 PDZ domains



Maximum likelihood strategy for the unfolded state parameters

Choose a set of experimental sequences $\boldsymbol{S}$ = {S, S', ...} as a reference

Boltzmann probability of sampling a sequence is $p(S) = \exp(-\beta\Delta G)/Z$

ΔG folding free energy, Z normalization constant

Choose unfolded parameters $E_u(t)$ to maximize probability or likelihood $\boldsymbol{L}$ of $\boldsymbol{S}$

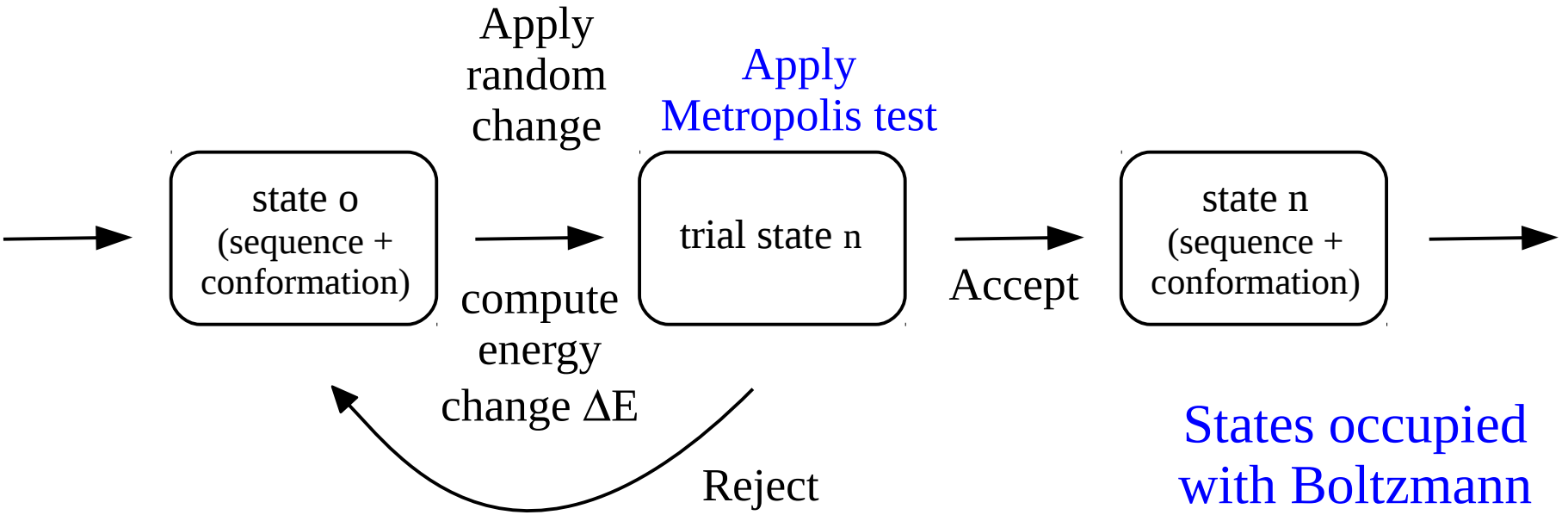
After some algebra, (exercise)

$\partial \boldsymbol{L}/\partial E_u(t) = 0 \iff$ experimental frequency $n_{exp}(t)$ = computational one $n_{comp}(t)$

Grad $\boldsymbol{L}$ along $E_u(t) \propto n_{exp}(t) - n_{comp}(t)$

Search for maximum likelihood parameters using gradient method of your choice

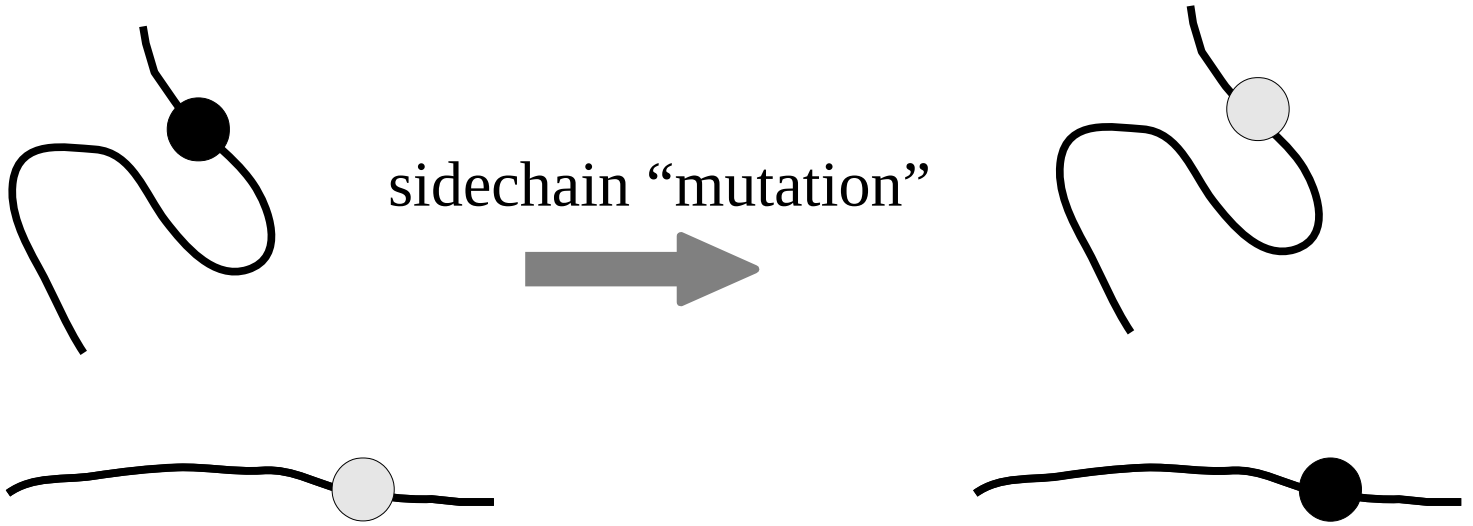
Eg: run MC with $\{E_u(t)\}$; adjust E_u values based on $n_{exp}(t) - n_{comp}(t)$; repeat

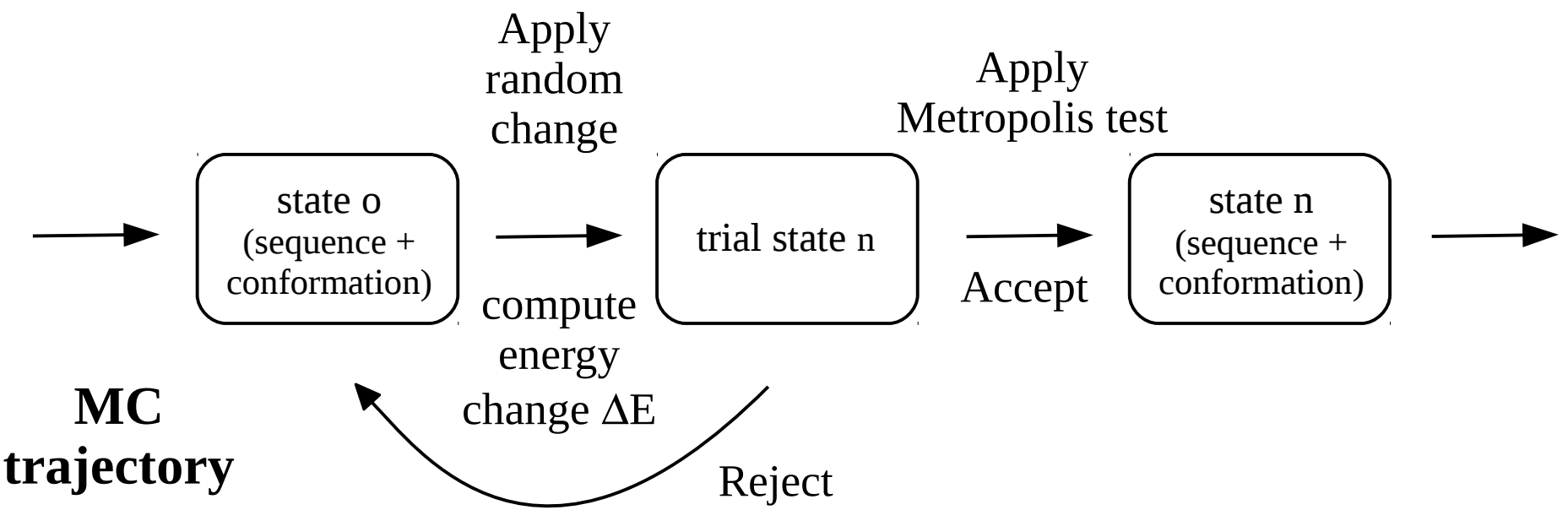


MC trajectory in sequence and structure space

States occupied with Boltzmann probability $\exp(-\beta E)$

MC move in "sequence space"



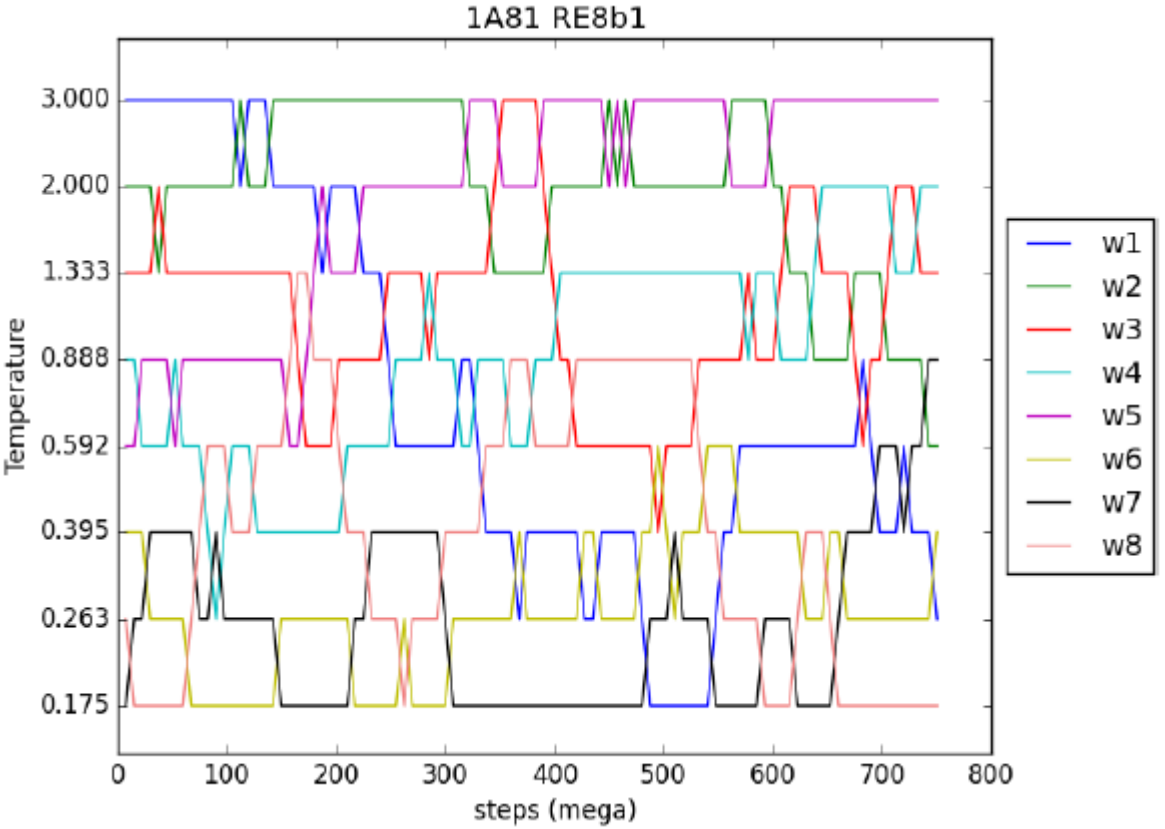


The simulation mimics a collection of sequences, present at equal concentrations, at equilibrium, distributed between folded/unfolded states according to Boltzmann statistics

Replica Exchange Monte Carlo

Multiple simulations or « replicas » are done in // at different temperatures; periodic exchanges of conformations between them are attempted following a Metropolis test: $acc(swap_{ij}) = Min[1, exp[(\beta_i-\beta_j)(\Delta E_i - \Delta E_j)]]$

Thermal energy kT in kcal/mol

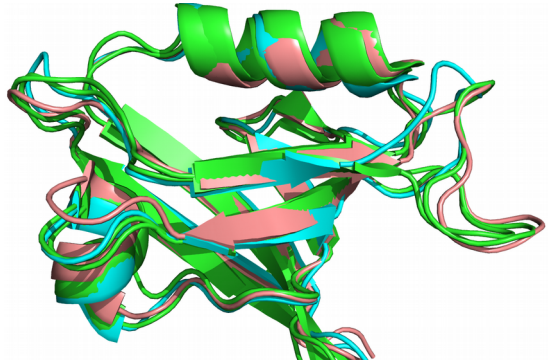


Replicas

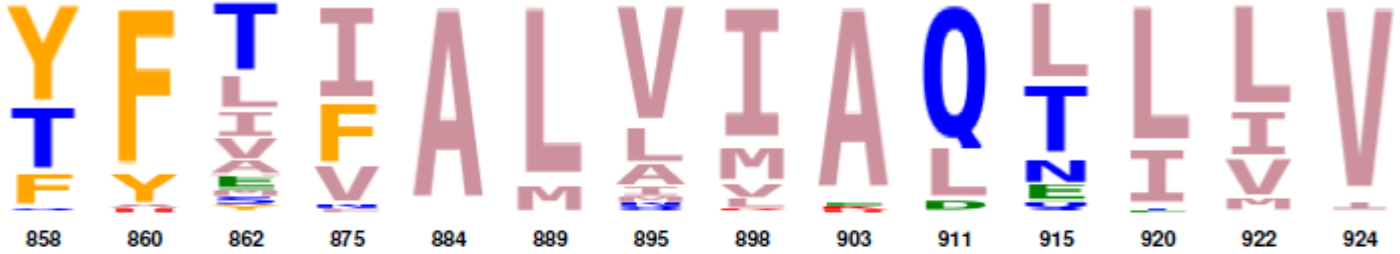
Monte Carlo steps (millions)

Enhanced sampling ; trivial cost thanks to OpenMP parallelization

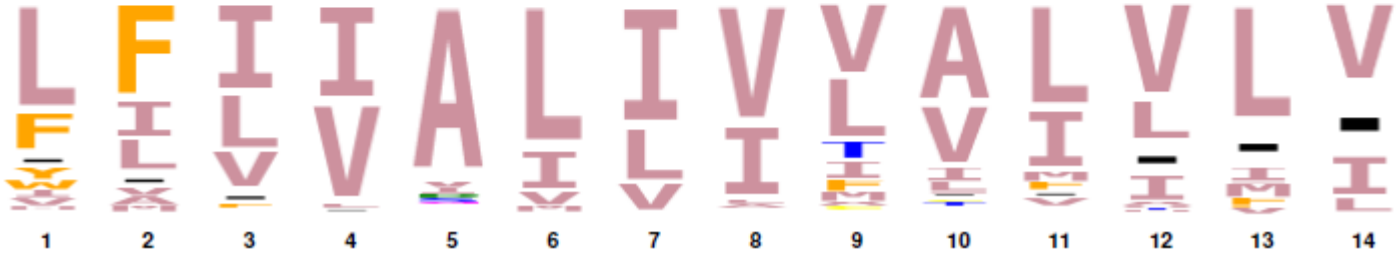
Experimental and designed PDZ sequences are similar (core positions)



Rosetta

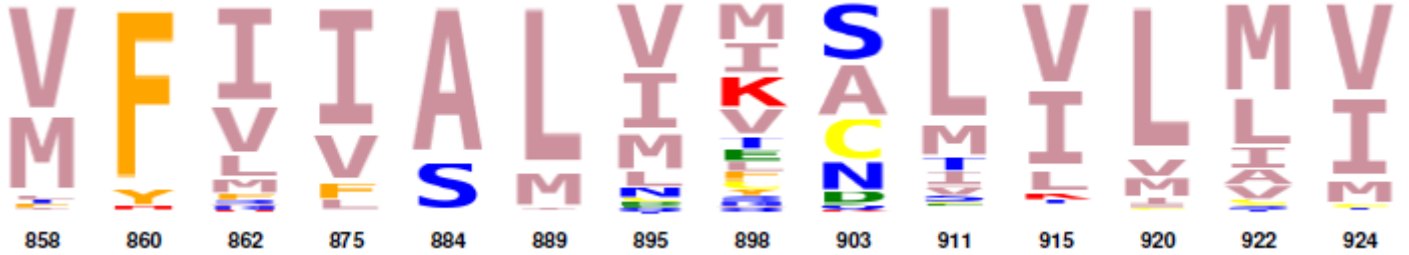


Pfam

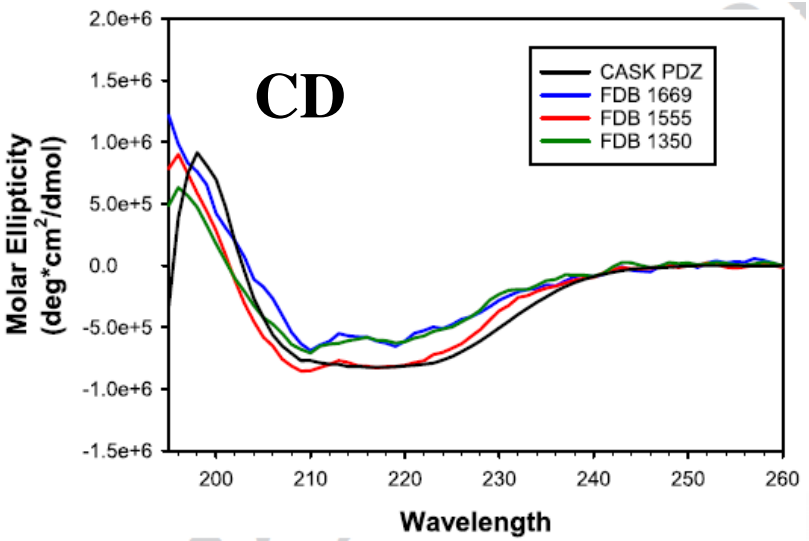


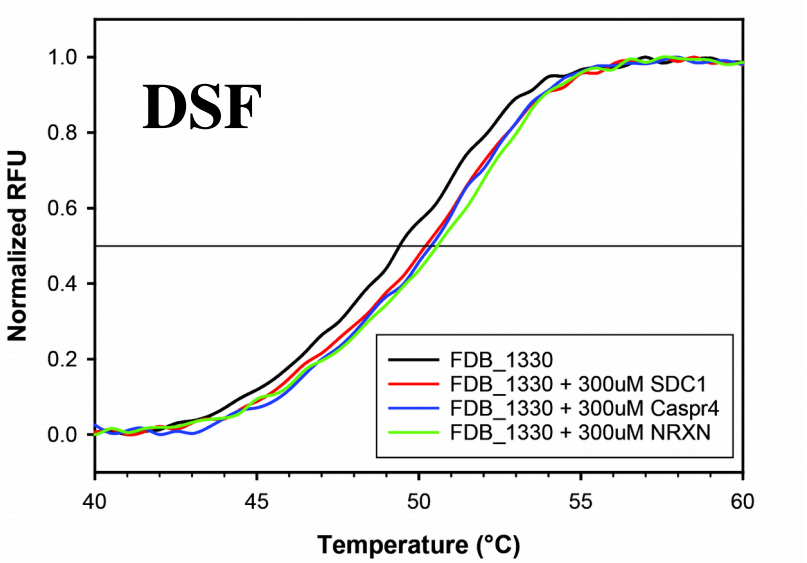
Proteus kT=0.6

Proteus kT=0.9 (kcal/mol)

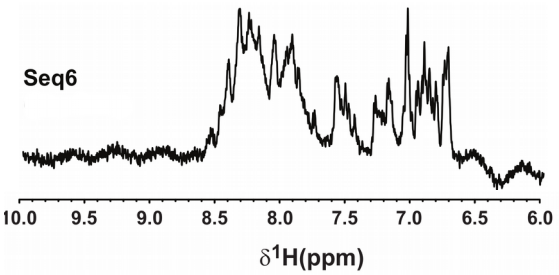


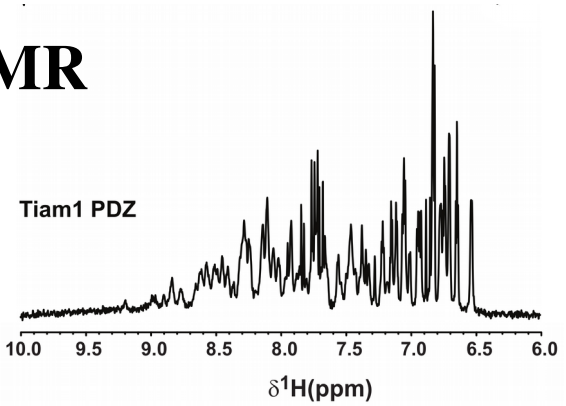
3 out of 3 designed PDZ sequences tested adopt the correct fold and 2 bind the correct ligand

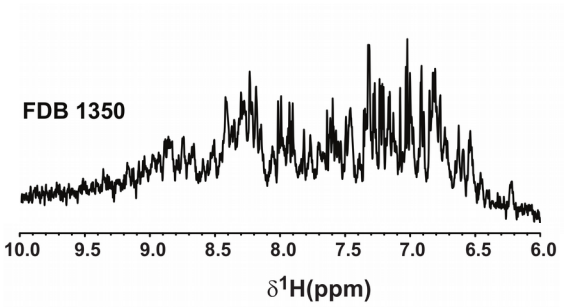


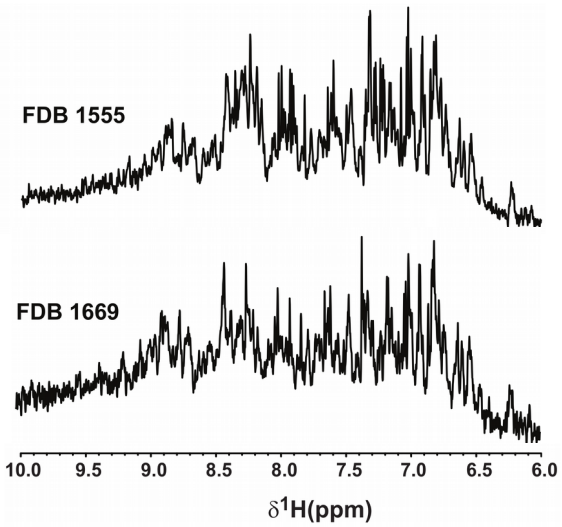


1D-NMR









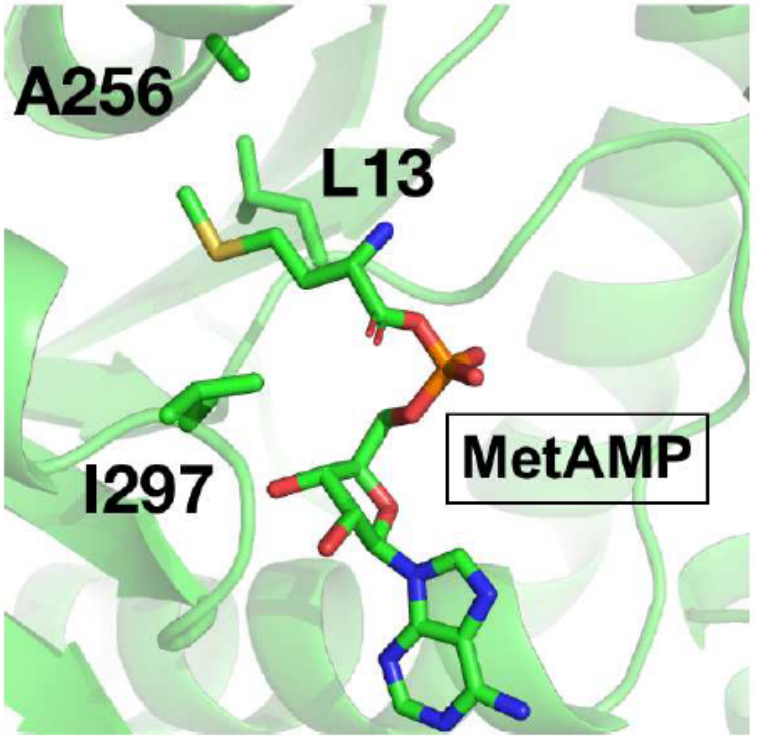
E Fuentes, U Iowa

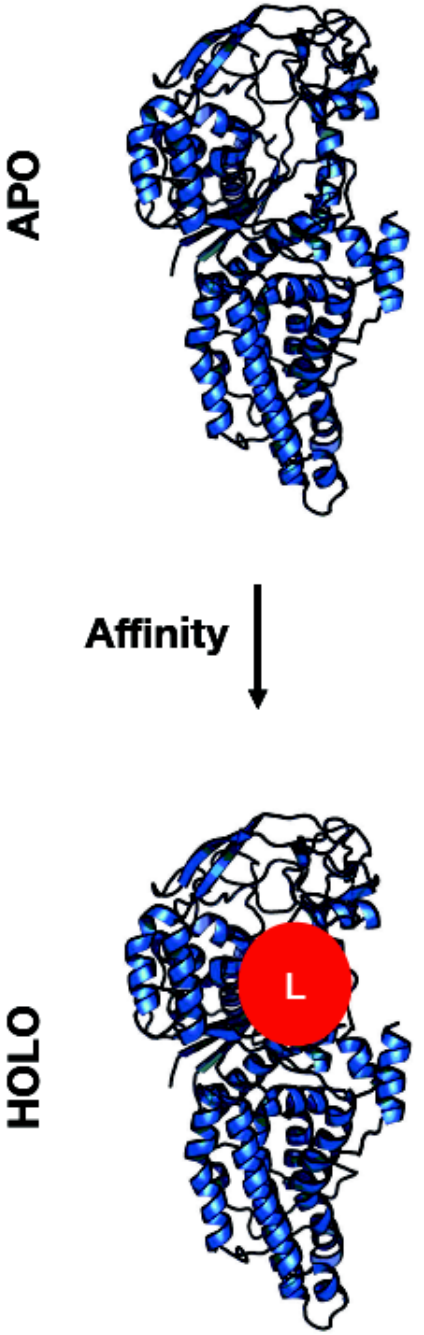
First successful complete protein redesign with a physics-based energy

Opuu et al, Scientific Reports, 2020

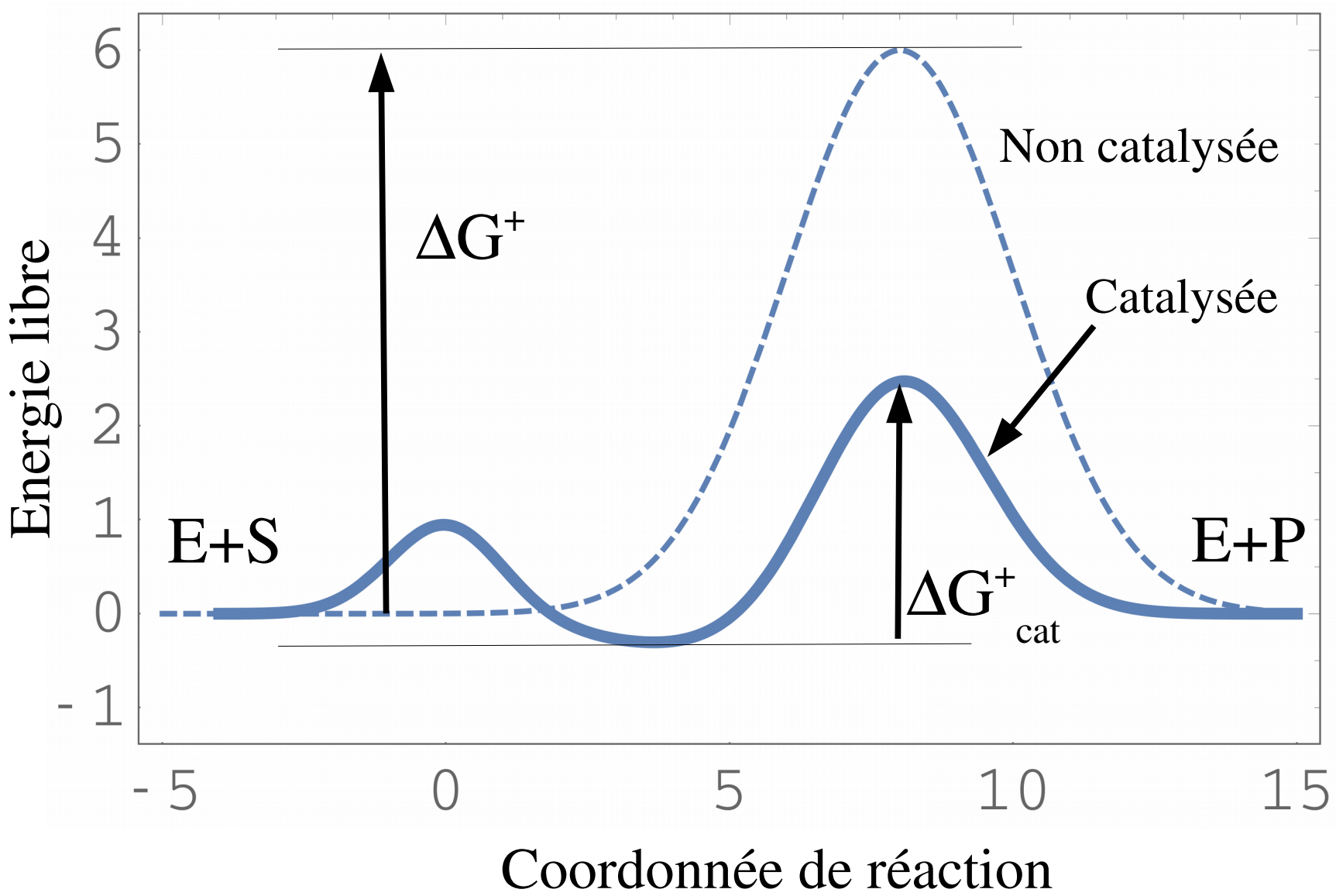
Computational enzyme design

To design new enzymes, or redesign old ones, choose a few amino acids in the active site and explore mutations, in search of activity.

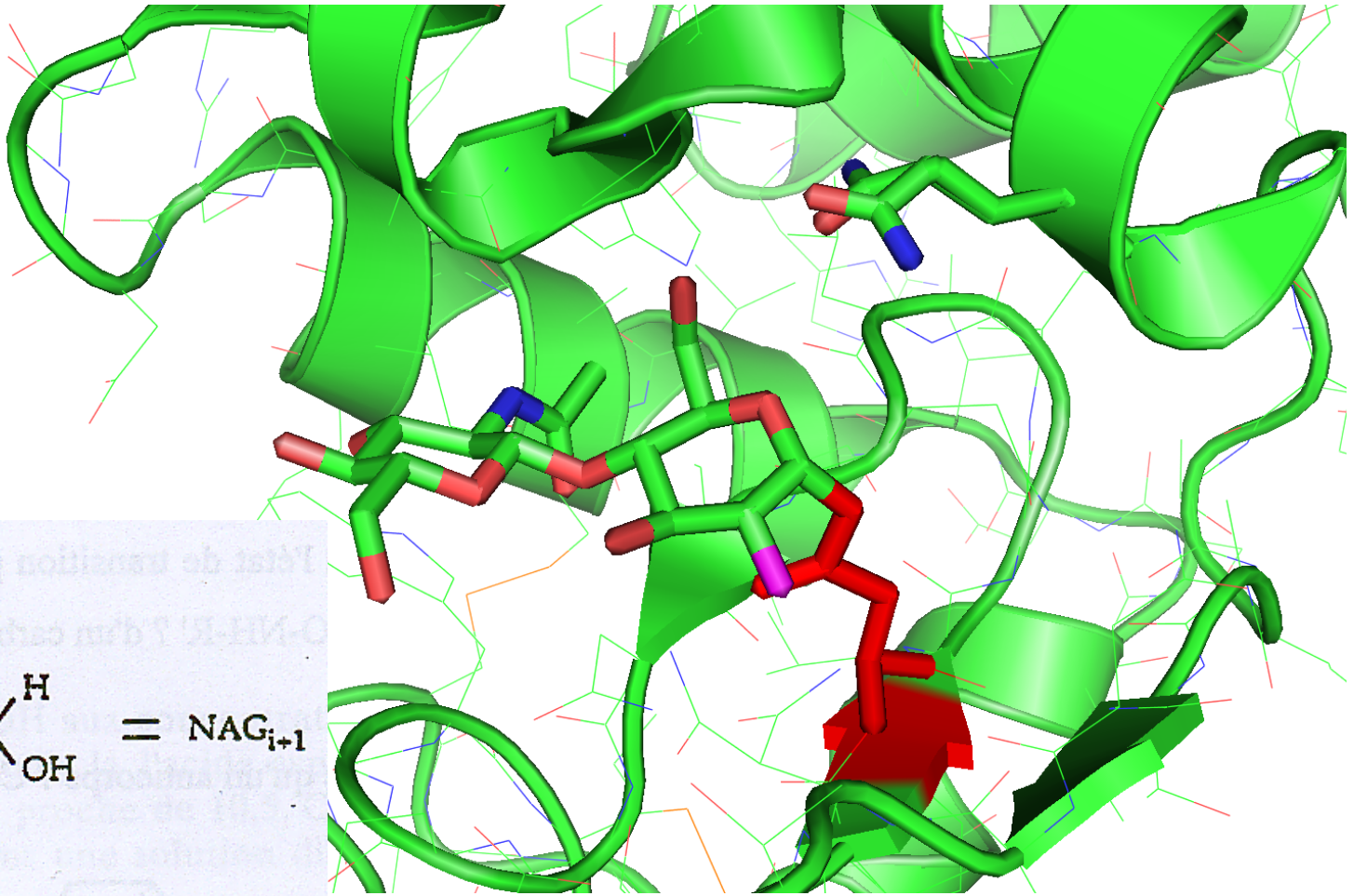


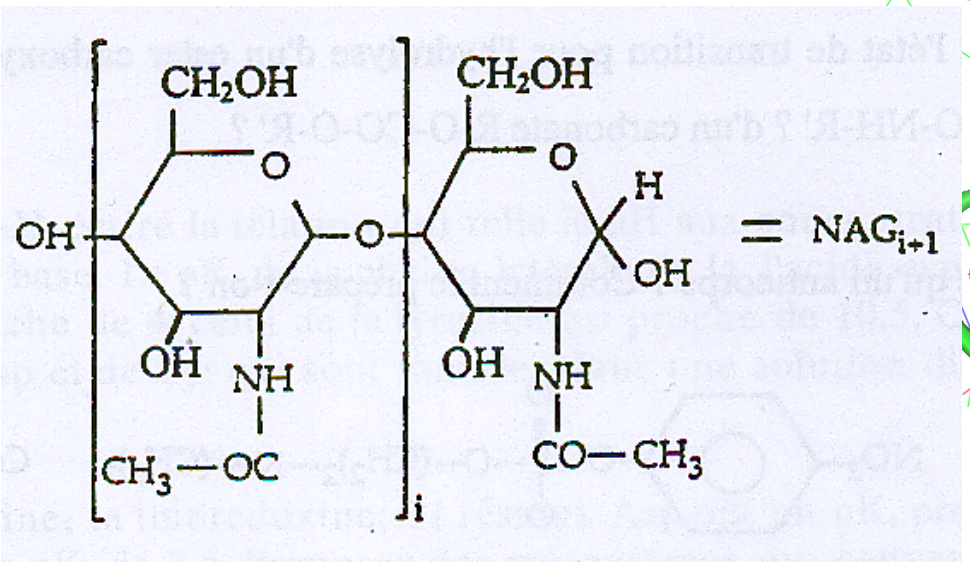


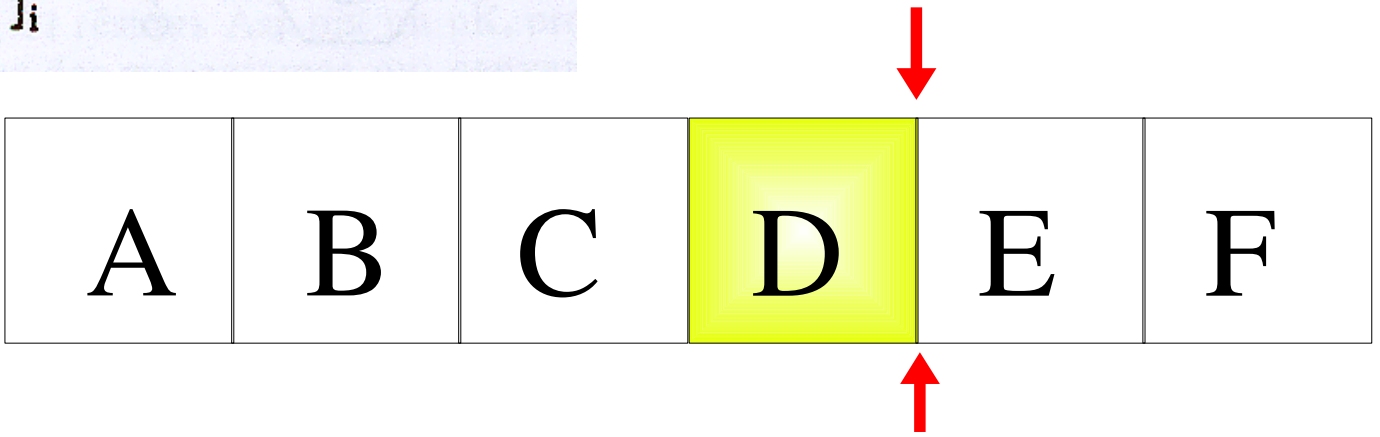
Design criteria : protein stability, substrate binding, low reaction barrier
Novel substrate, new reaction.


Energie libre
6
5
4
3
2
1
0
- 1
- 5
0
5
10
15
Coordonnée de réaction
Non catalysée
Catalysée
ΔG+
ΔG+ cat
E+S
E+P

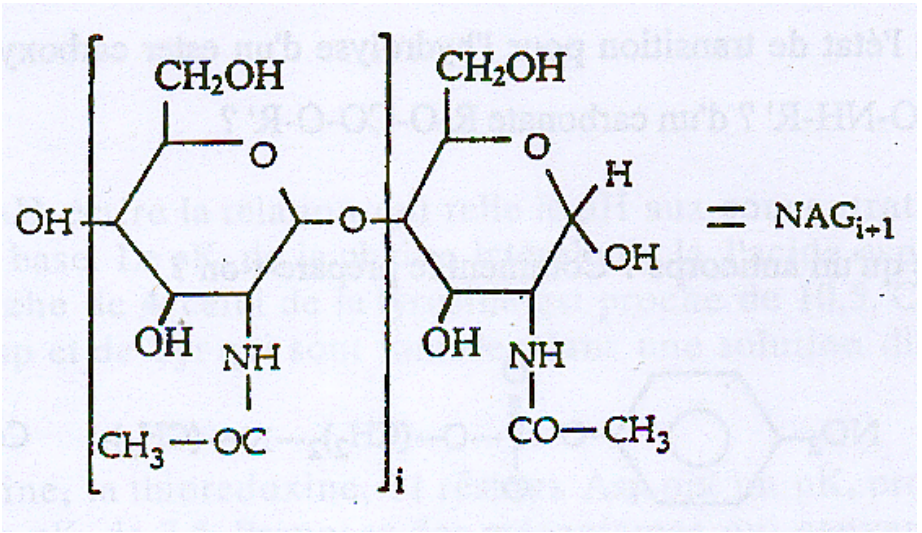
Lysozyme active site







N-acetyl-glucosamine (NAG)



Lysozyme problem

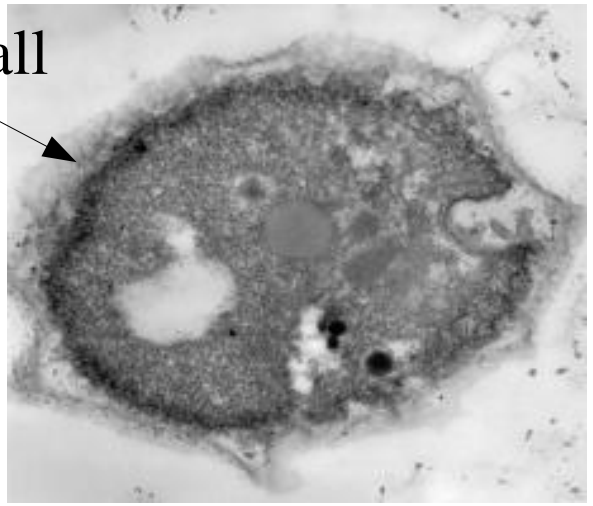
Starch, glycogen

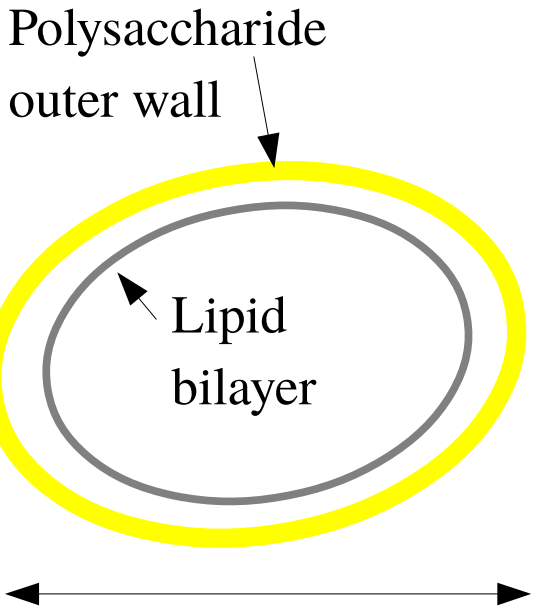
energy storage

cellulose, chitin

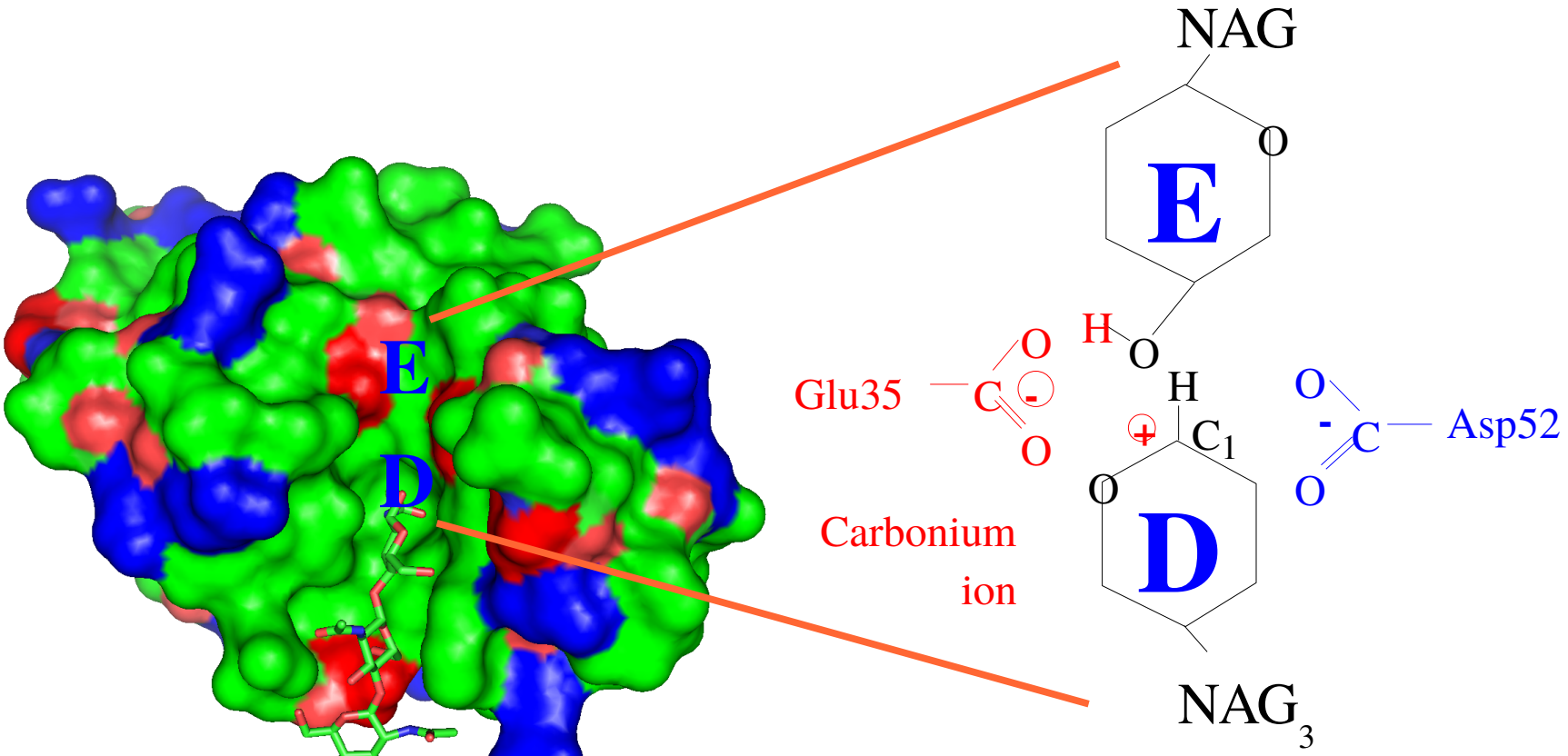
structure

Cell wall



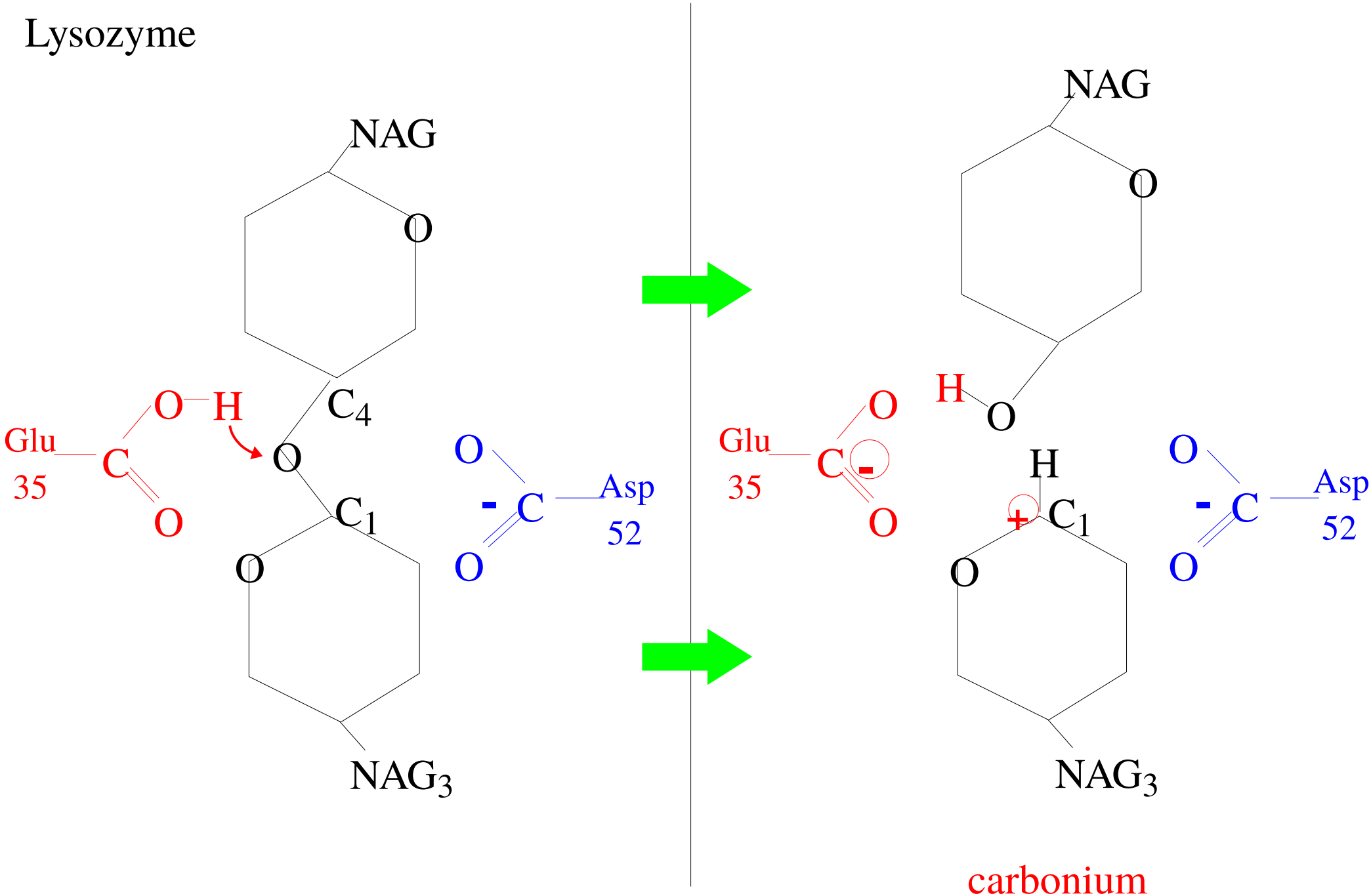


Transition state binding by lysozyme



Lysozyme active site

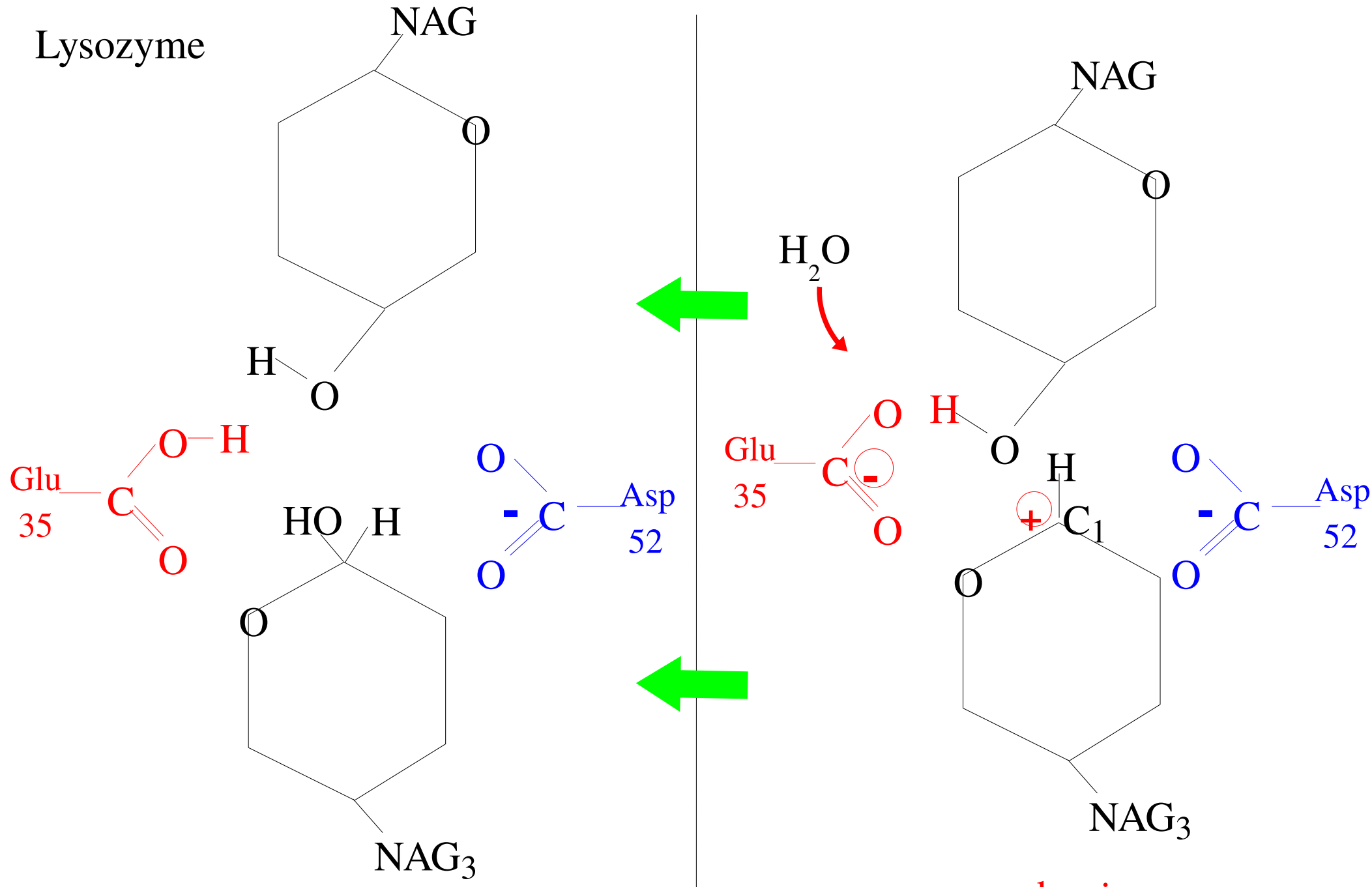
Lysozyme



carbonium ion

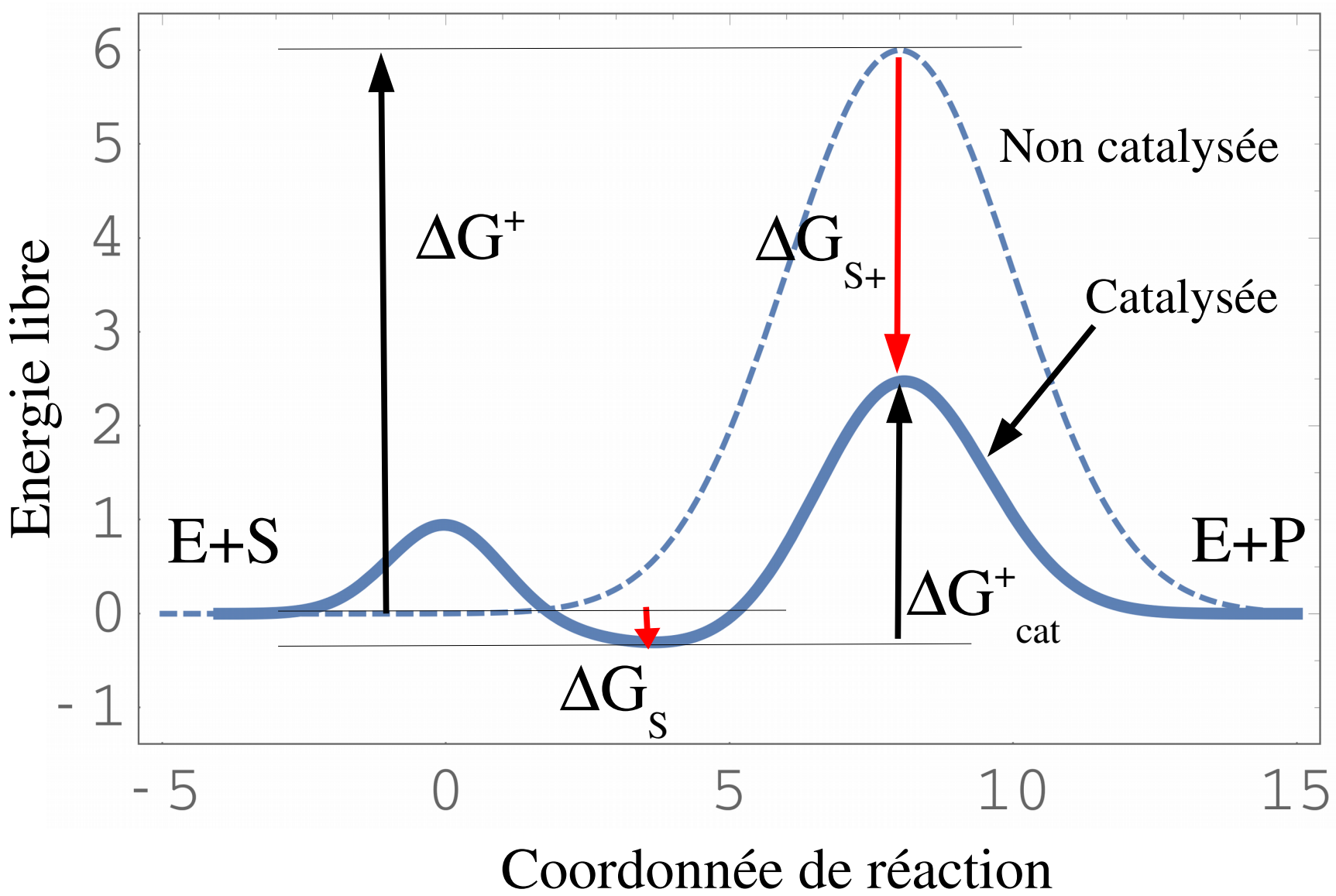
1st step along the probable reaction pathway

Lysozyme

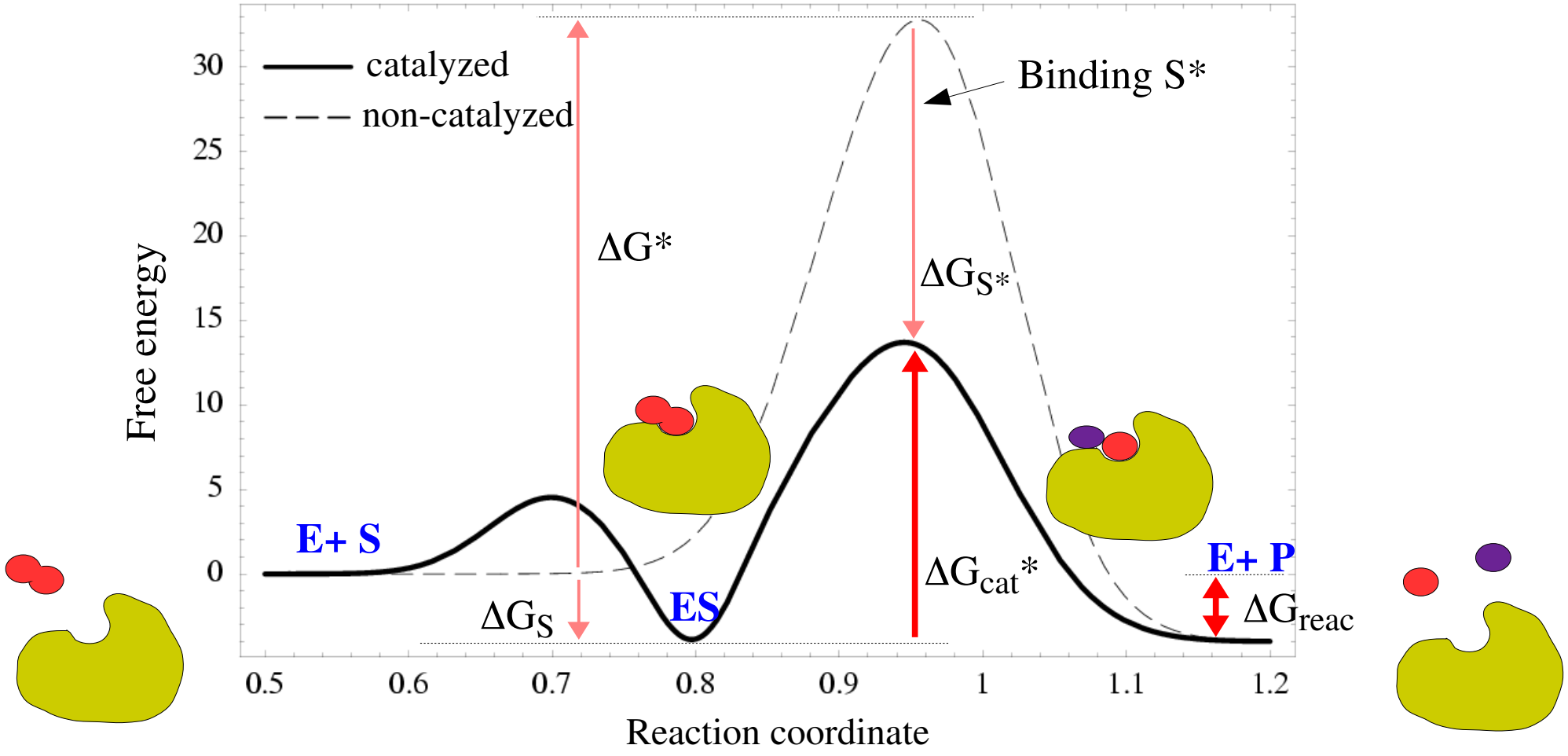


carbonium ion

2nd step along the probable reaction pathway


Energie libre
6
5
4
3
2
1
0
- 1
- 5
0
5
10
15
Coordonnée de réaction
E+S
E+P
ΔG+
ΔGS+
ΔGS
ΔG+cat
Non catalysée
Catalysée

Reaction coordinate and reaction pathway of an enzyme.
Activation barrier for the reactions with and without the enzyme.

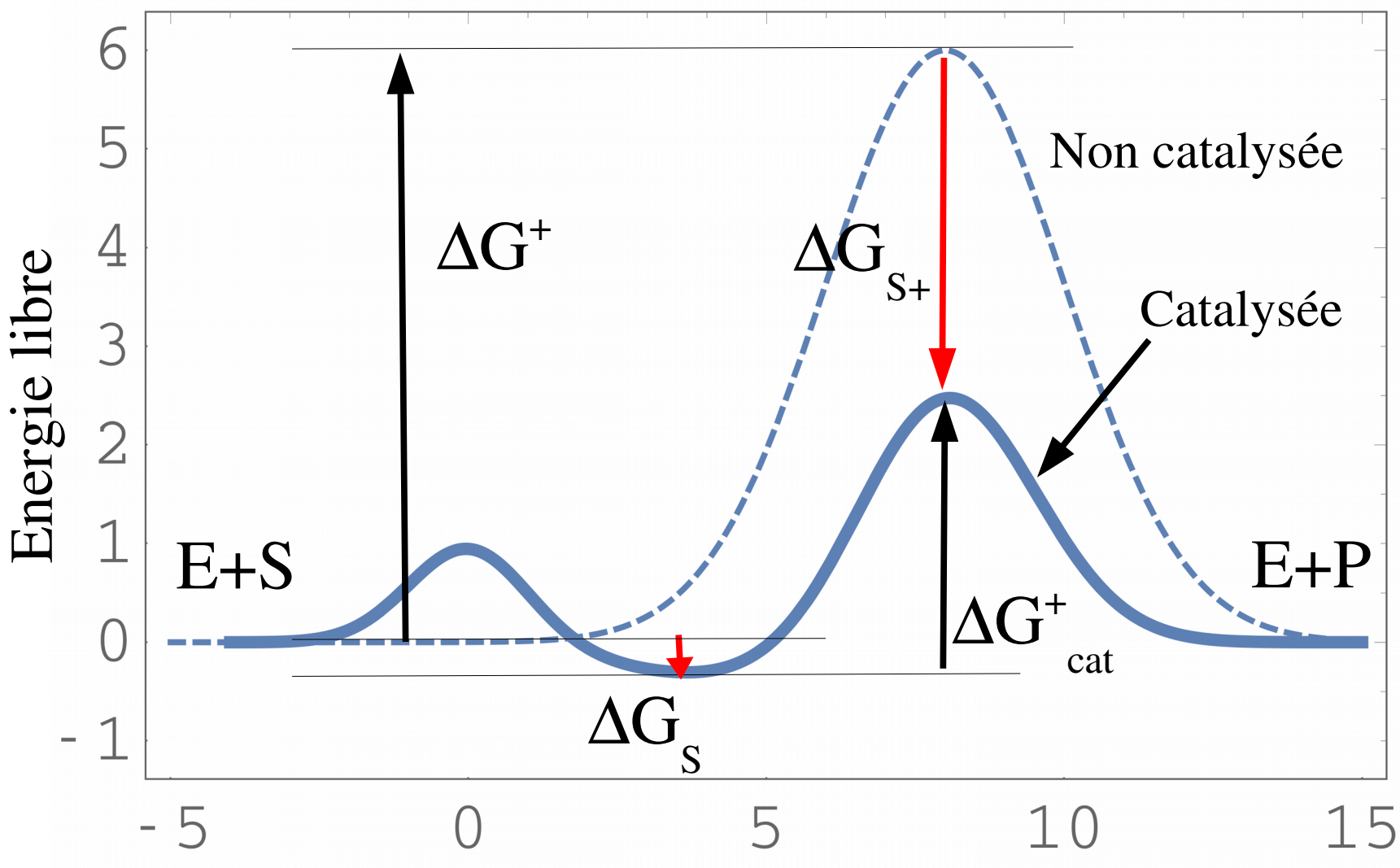


$$\Delta G_{cat}^{*} - \Delta G^{*} = \Delta G_{S*} - \Delta G_S$$

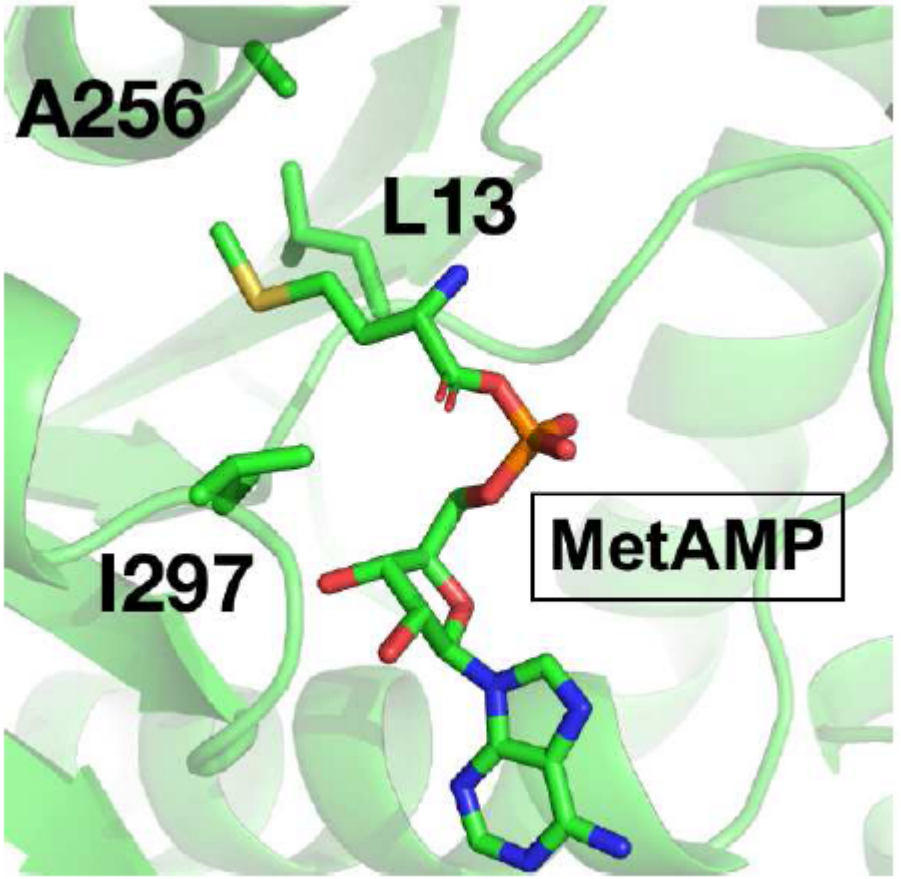
$$\Delta G^{+} - \Delta G_{S} = \Delta G^{+}_{cat} - \Delta G_{S+}$$

$$\Delta G^{+} - \Delta G^{+}_{cat} = \Delta G_{S} - \Delta G_{S+}$$

Pouvoir catalytique $\Leftrightarrow$ Liaison de S^{+} > liaison de S



Computational enzyme design



Substrate complex

Design criterion : transition state binding.

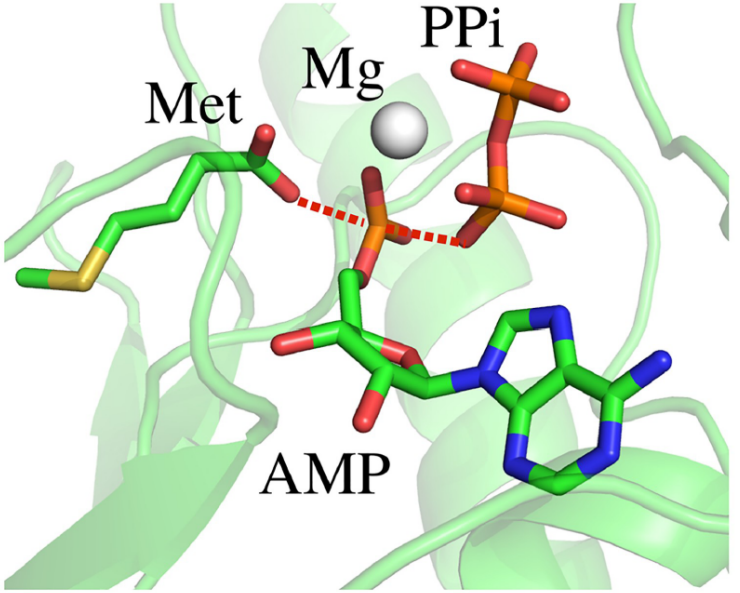
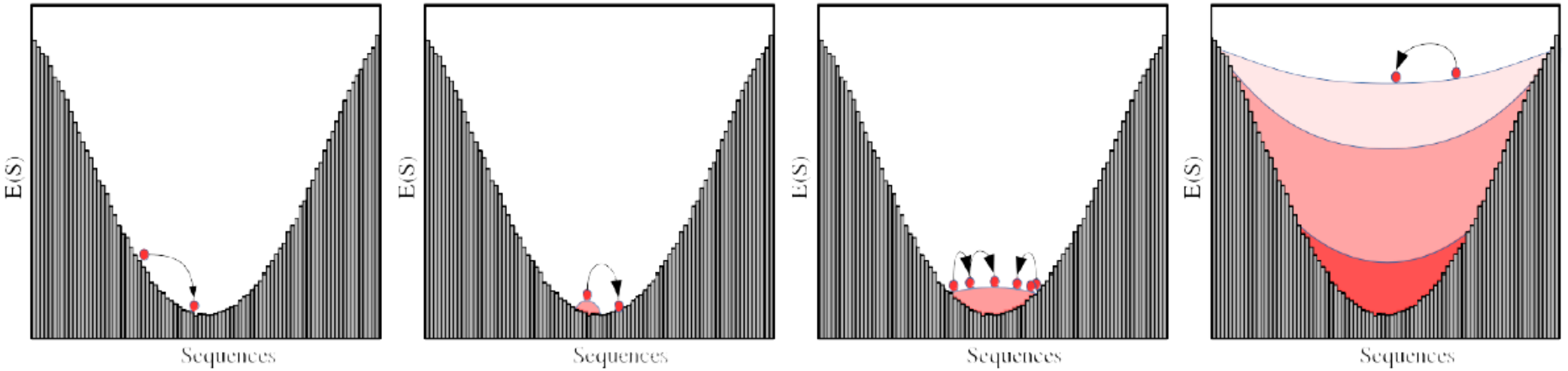


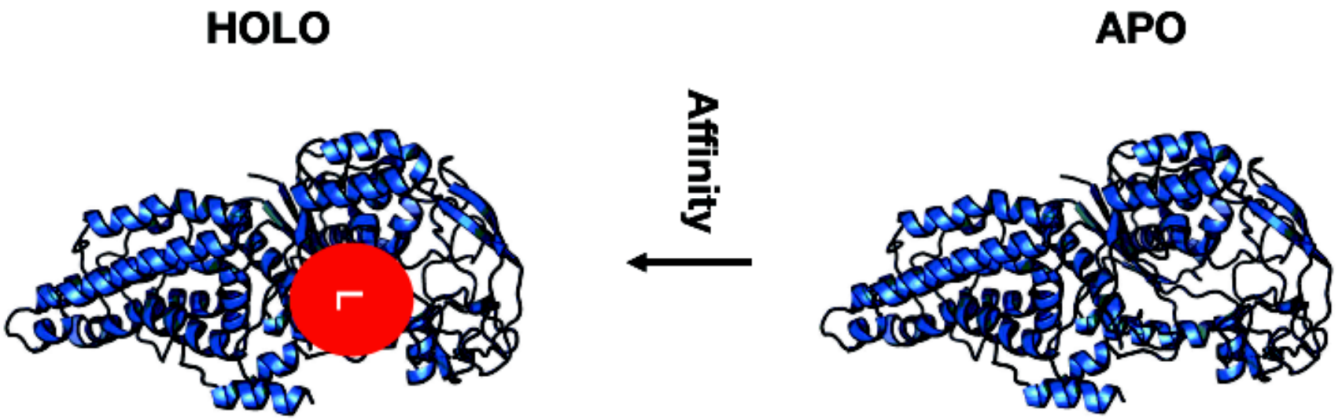
Fig 1. MetRS transition state for MetAMP formation. Closeup of the ligands.

Transition state complex

Step 1 : Adaptive landscape flattening of the unbound state

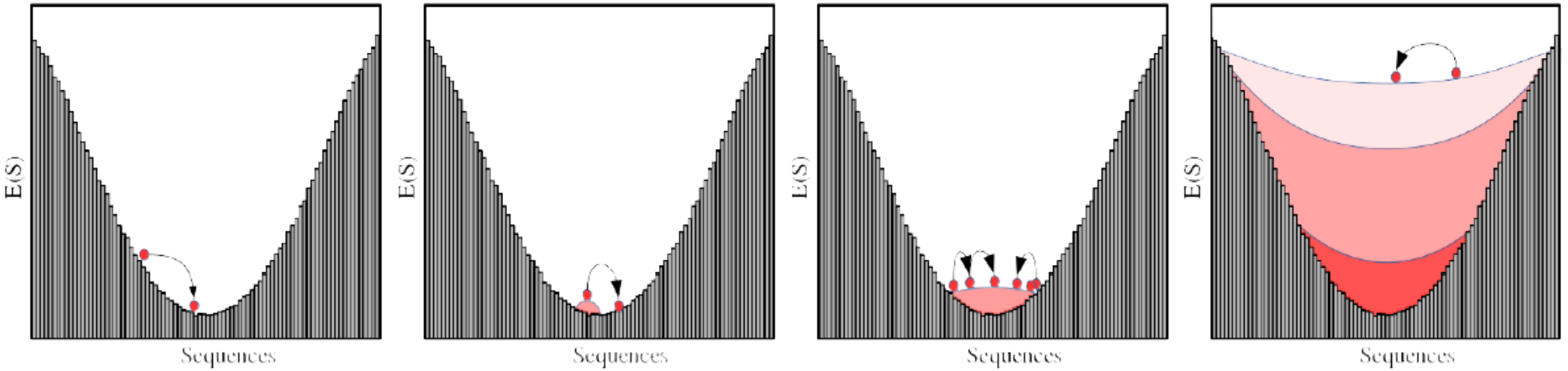
Simulate unbound state ; when a sequence S is sampled, add a small penalty to the energy function. After a while, the energy surface has been flattened.





Step 1 : Adaptive landscape flattening of the unbound state

Simulate unbound state ; when a sequence S is sampled, add a small penalty to the energy function. After a while, the energy surface has been flattened. The penalty function is $-E_{apo}(S)$.



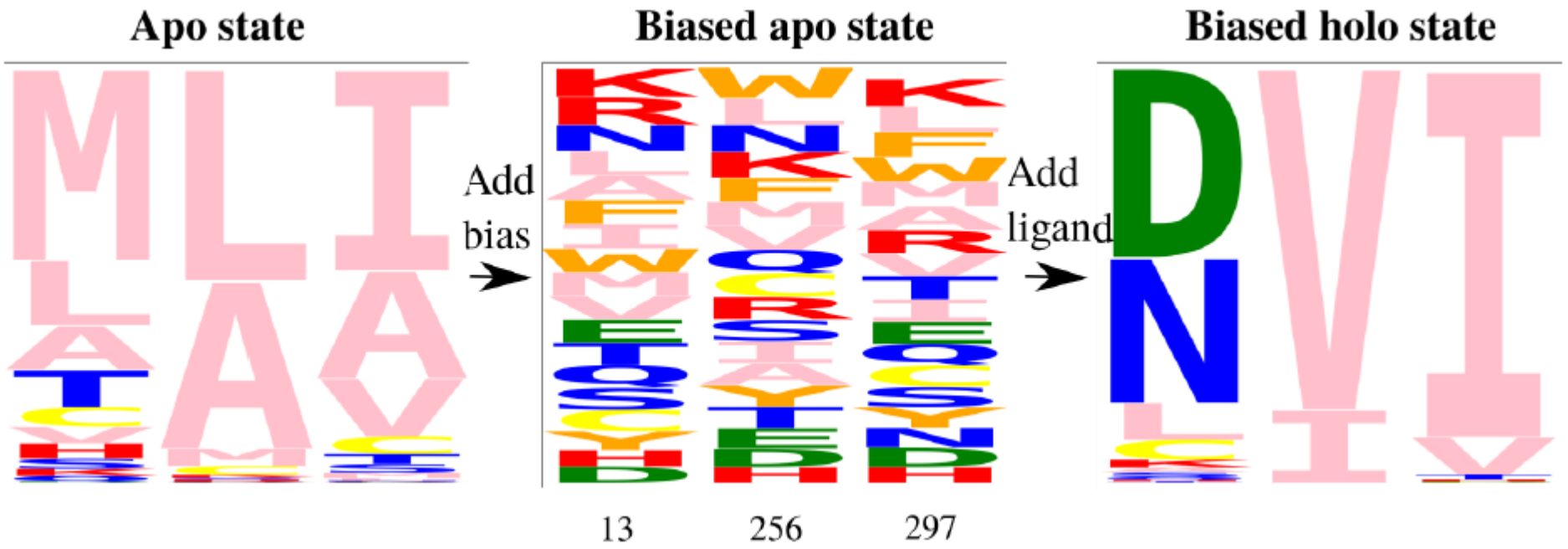
Step 2 : simulate the bound state, including the penalty function

The simulation is controlled by $E_{holo}(S) - E_{apo}(S)$, the binding energy

Step 1 : Adaptive landscape flattening of the unbound state

Step 2 : simulate the bound state, including the penalty function

The simulation is controlled by $E_{holo}(S) - E_{apo}(S)$,the binding energy



Methionyl-tRNA synthetase redesigned to accept nonnatural amino acids as substrates

A tool for synthetic biology and biotechnology

Inverse folding problem

Monte Carlo simulation, biased simulation techniques

Likelihood maximization = unsupervised machine learning of unfolded parameters

Adaptive landscape flattening : another unsupervised learning technique

An example targeting genetic code expansion (underway)

Enzyme catalysis !