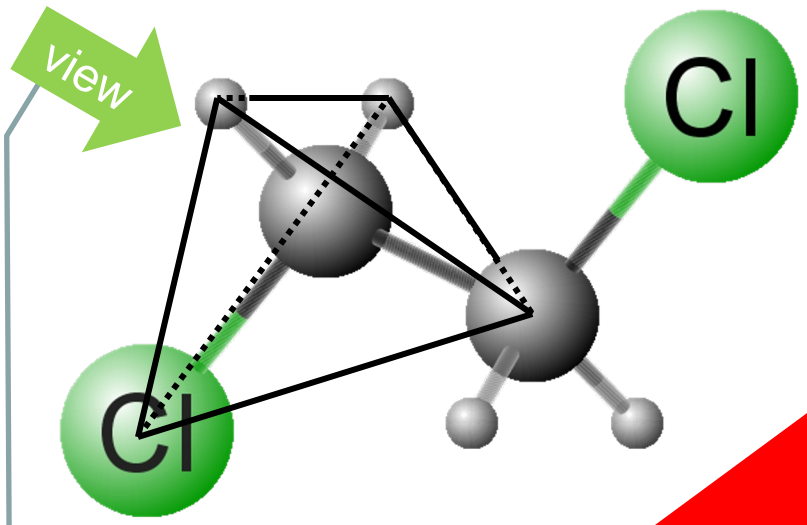
Conformational Sampling

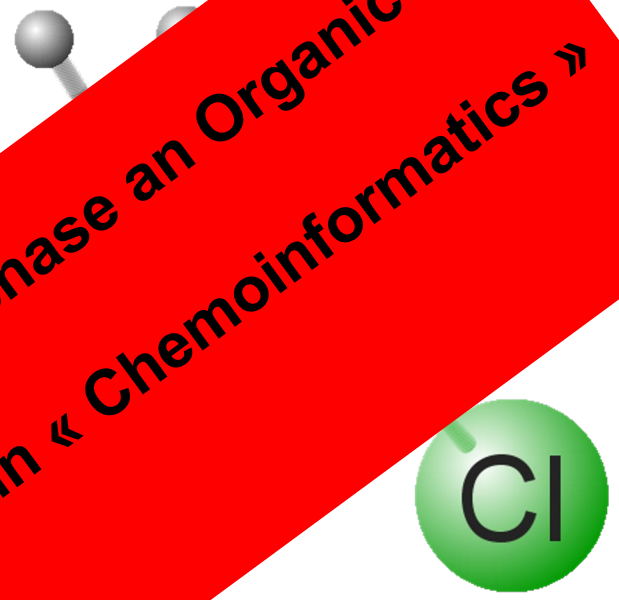
Dragos Horvath

Laboratoire de Chemoinformatique – UMR 7140

dhorvath@unistra.fr

- *Presentation Outline*
  - The Basics: Molecules have Geometries!
    - Intramolecular energy: the Empirical Force Field
  - Sampling Methods: a brief overview
    - Molecular Dynamics: Walking like a molecule
    - Monte Carlo: Molecular Casino
    - Evolutionary Methods: in God/Darwin we trust!
  - Conclusions

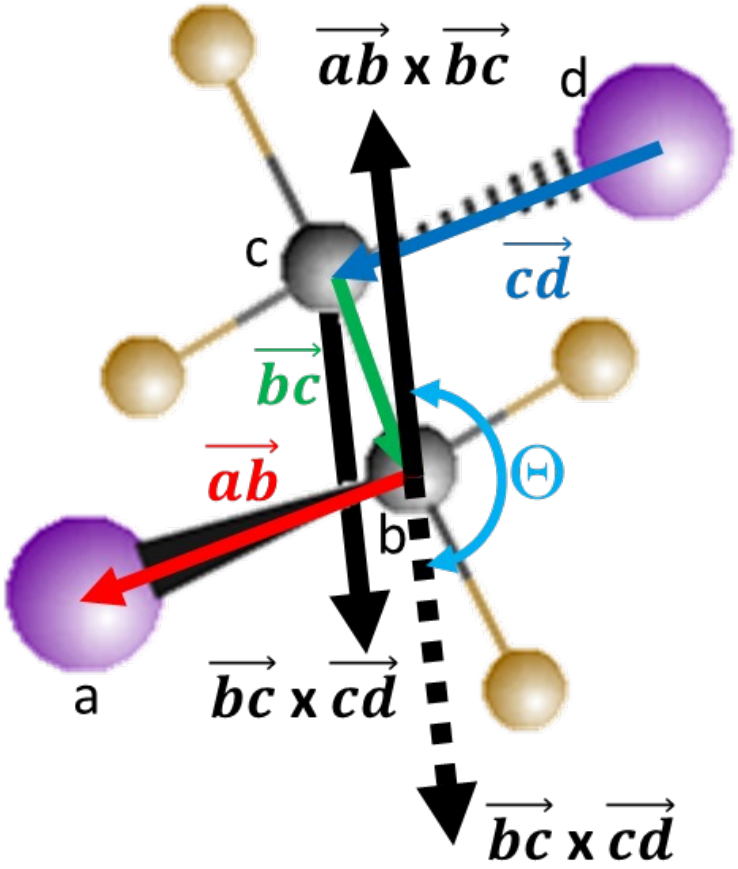


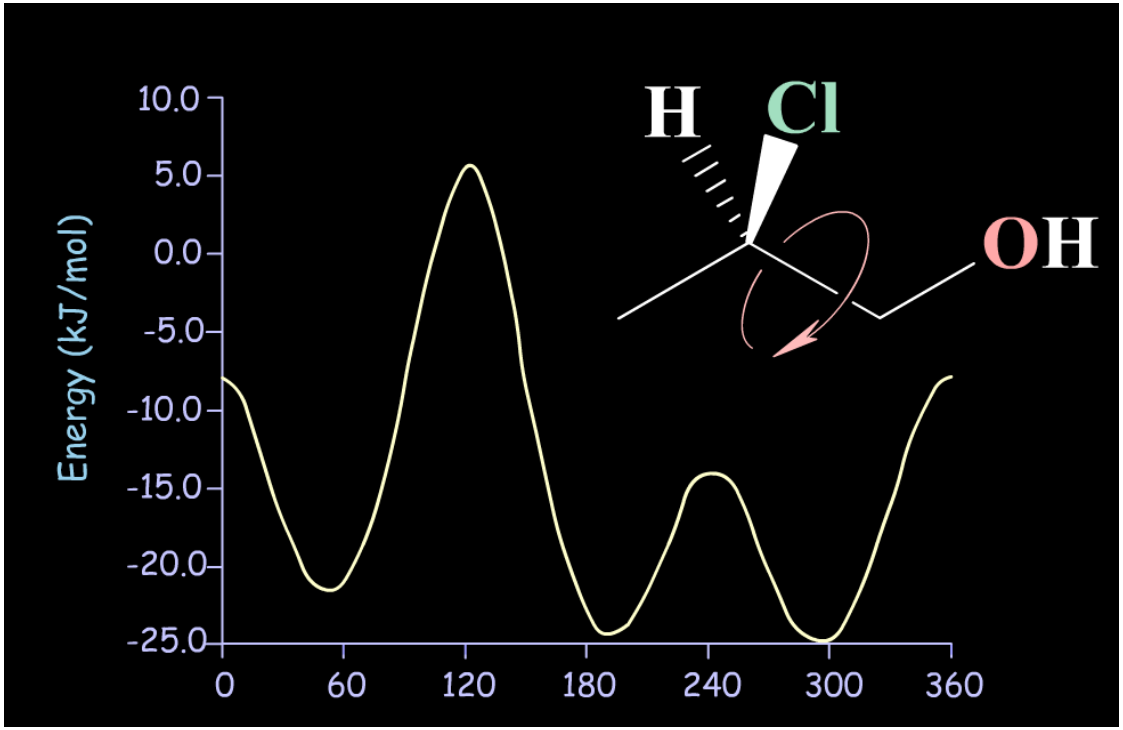


If this slide feels alien to you, urgently purchase an Organic Chemistry Textbook. The prefix « Chemo » in « Chemoinformatics » really matters!

Torsions : the gateway to conformational sampling

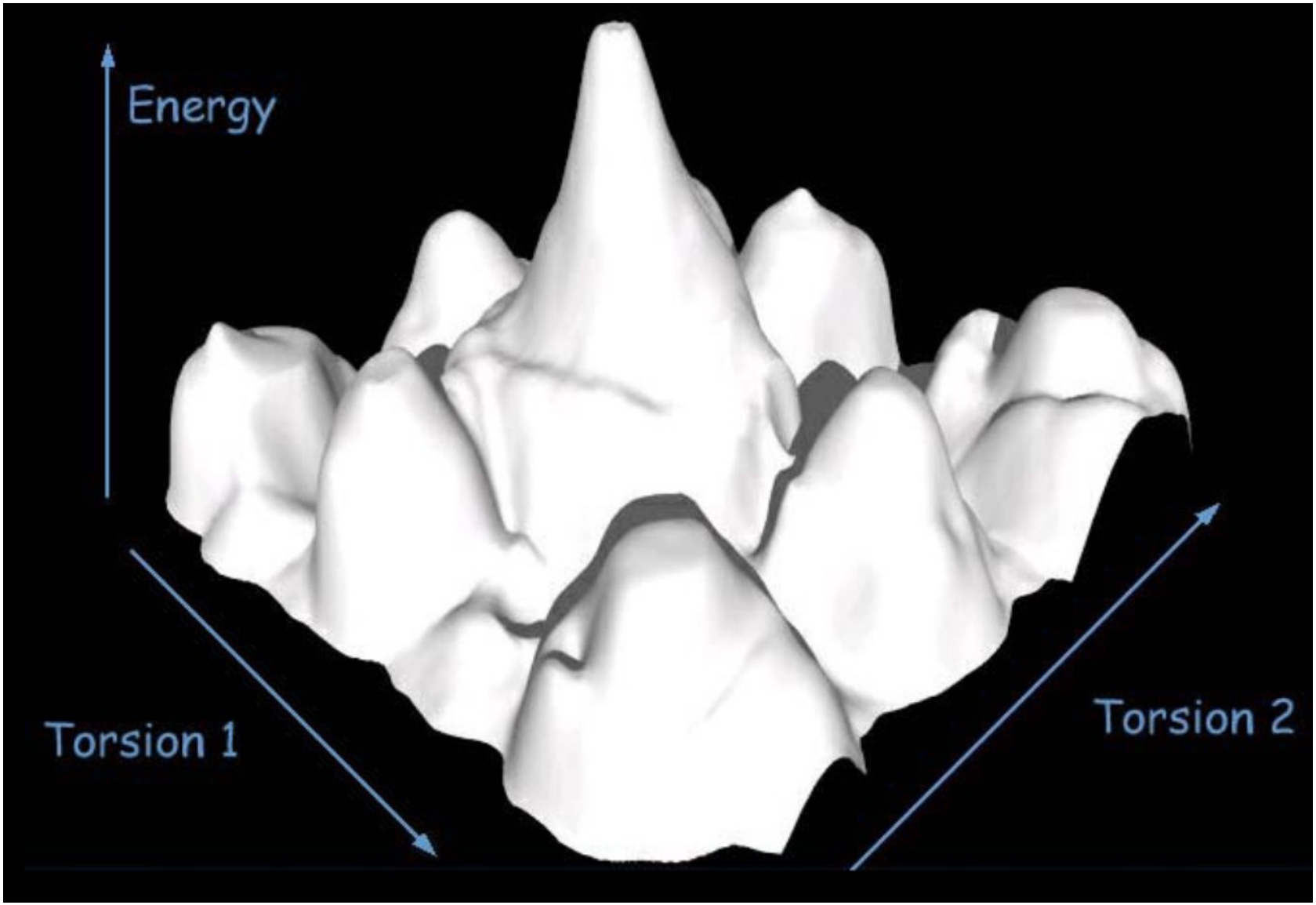
- Rotation around a bond impacts on interatomic distances, thus on energy!





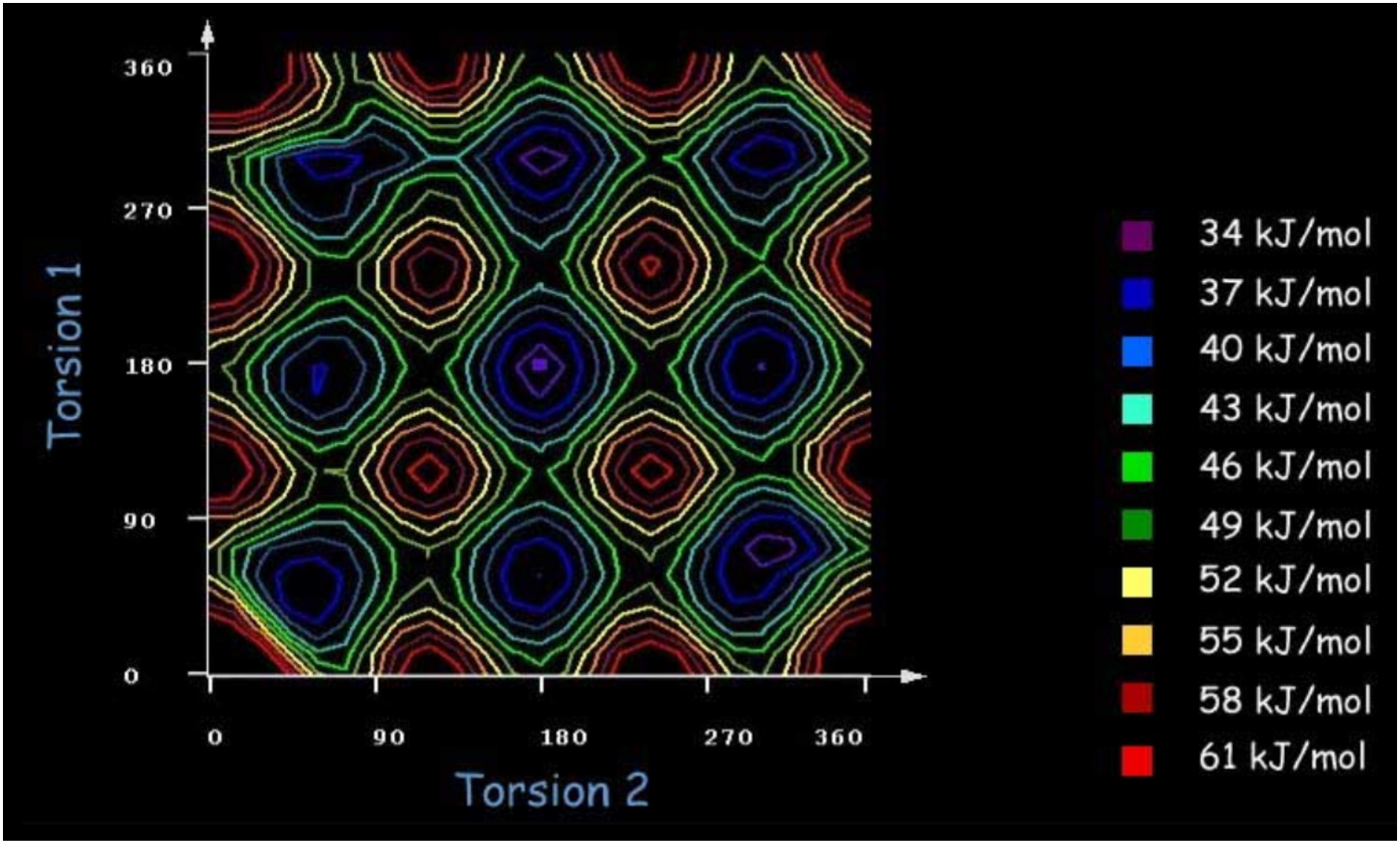
Torsions : the gateway to conformational sampling

- Energy Surface with respect to two torsions....

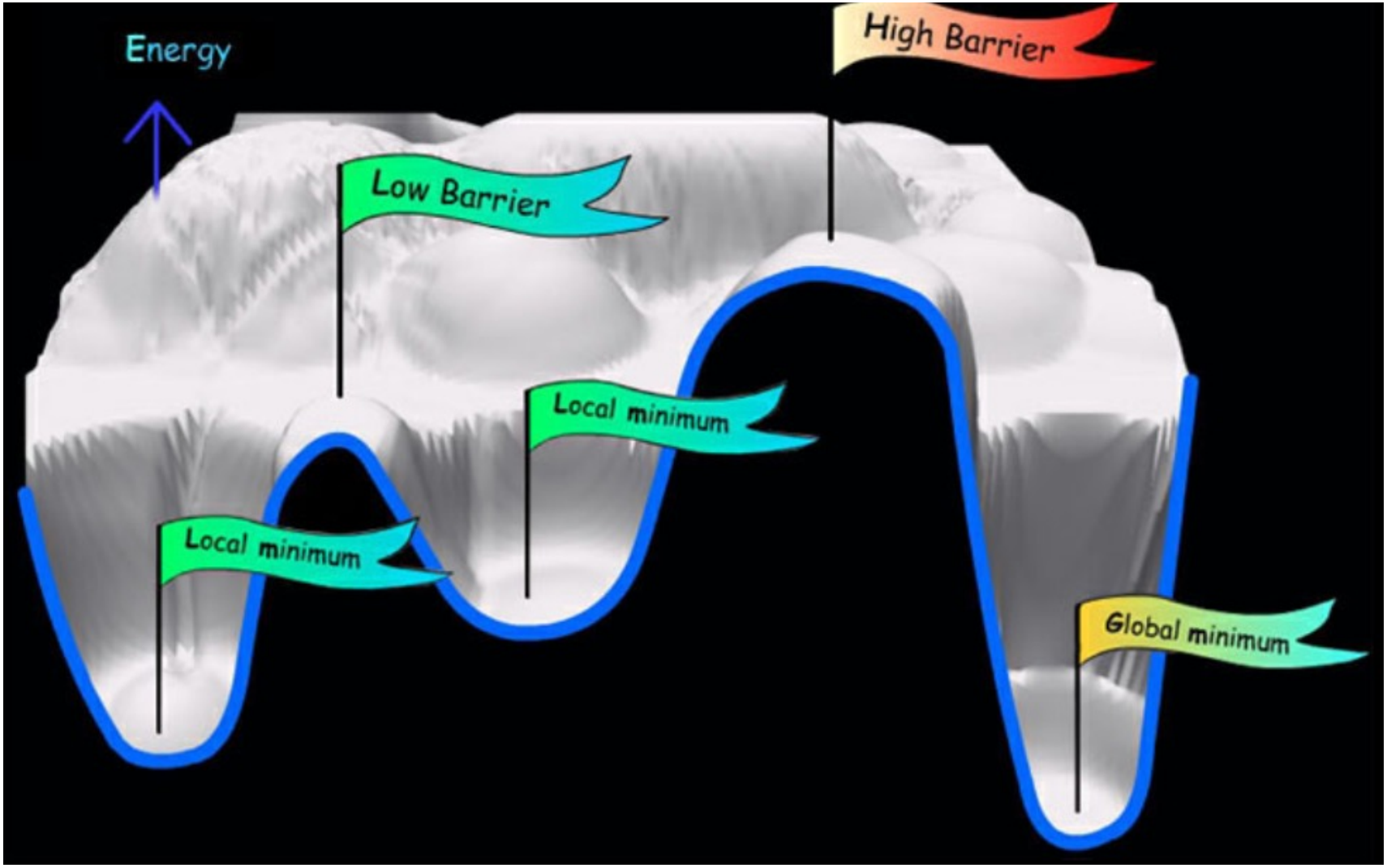


Torsions : the gateway to conformational sampling

- Alternative Contour Plot representation

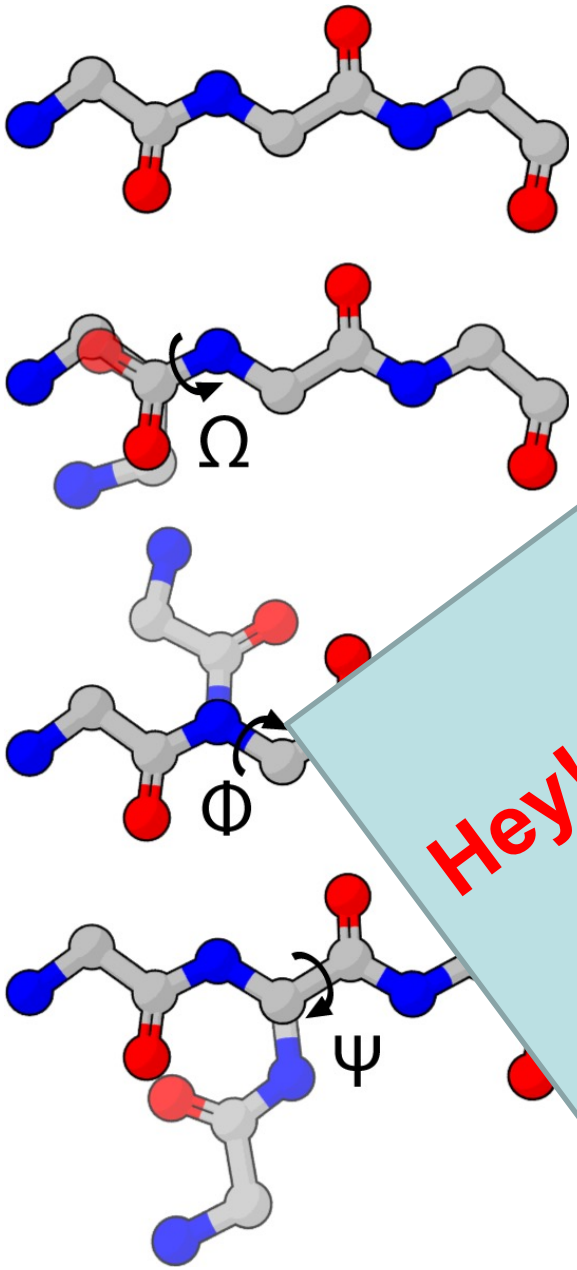


Key points on the energy surface...

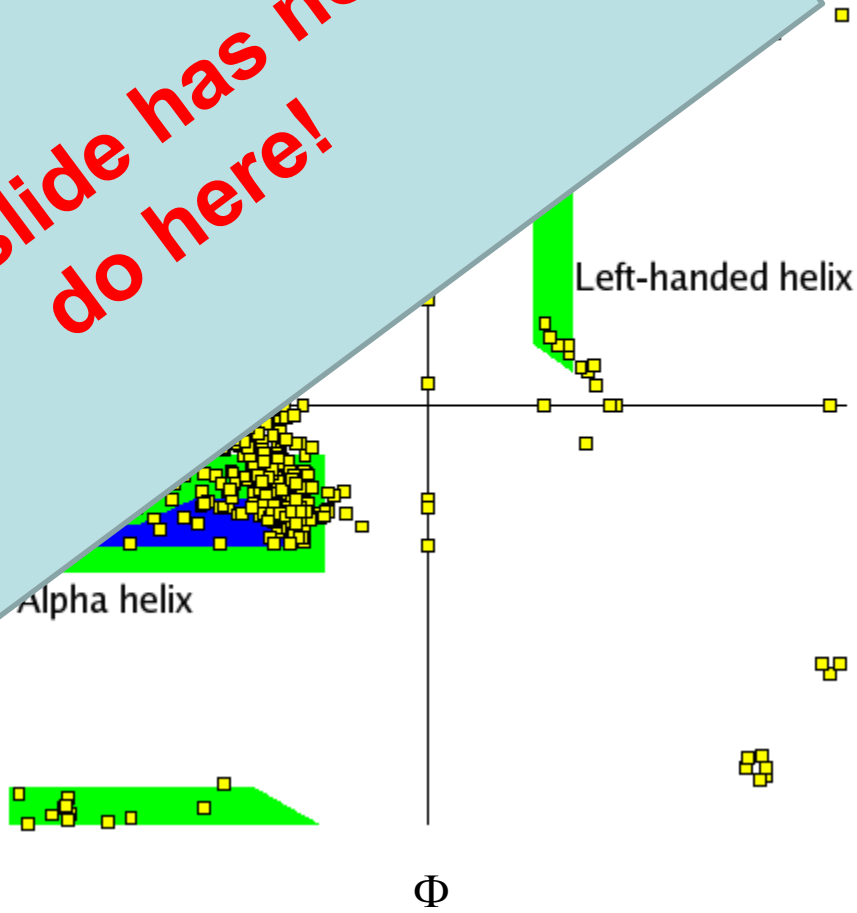


The Ramachandran Plot

http://en.wikipedia.org/wiki/Ramachandran_plot

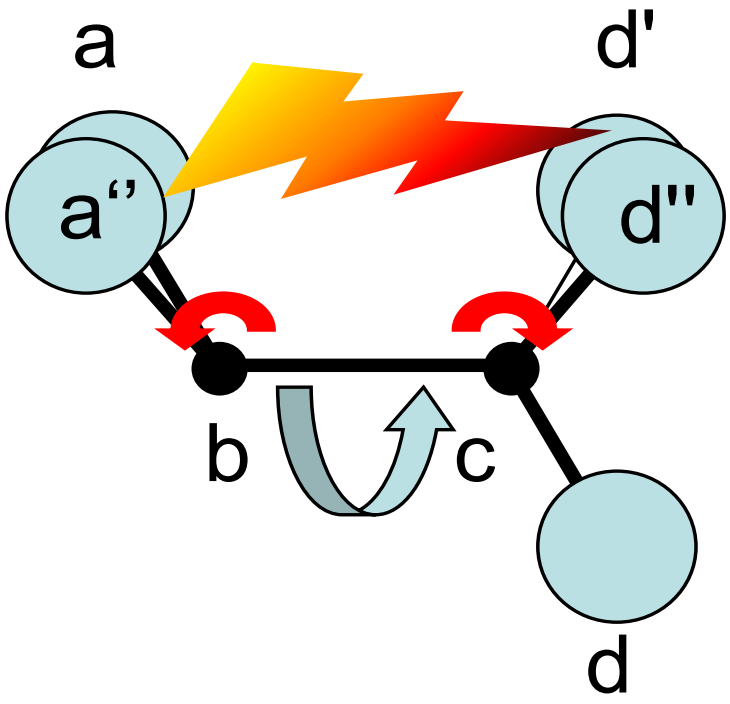


Hey! This slide has nothing to do here!

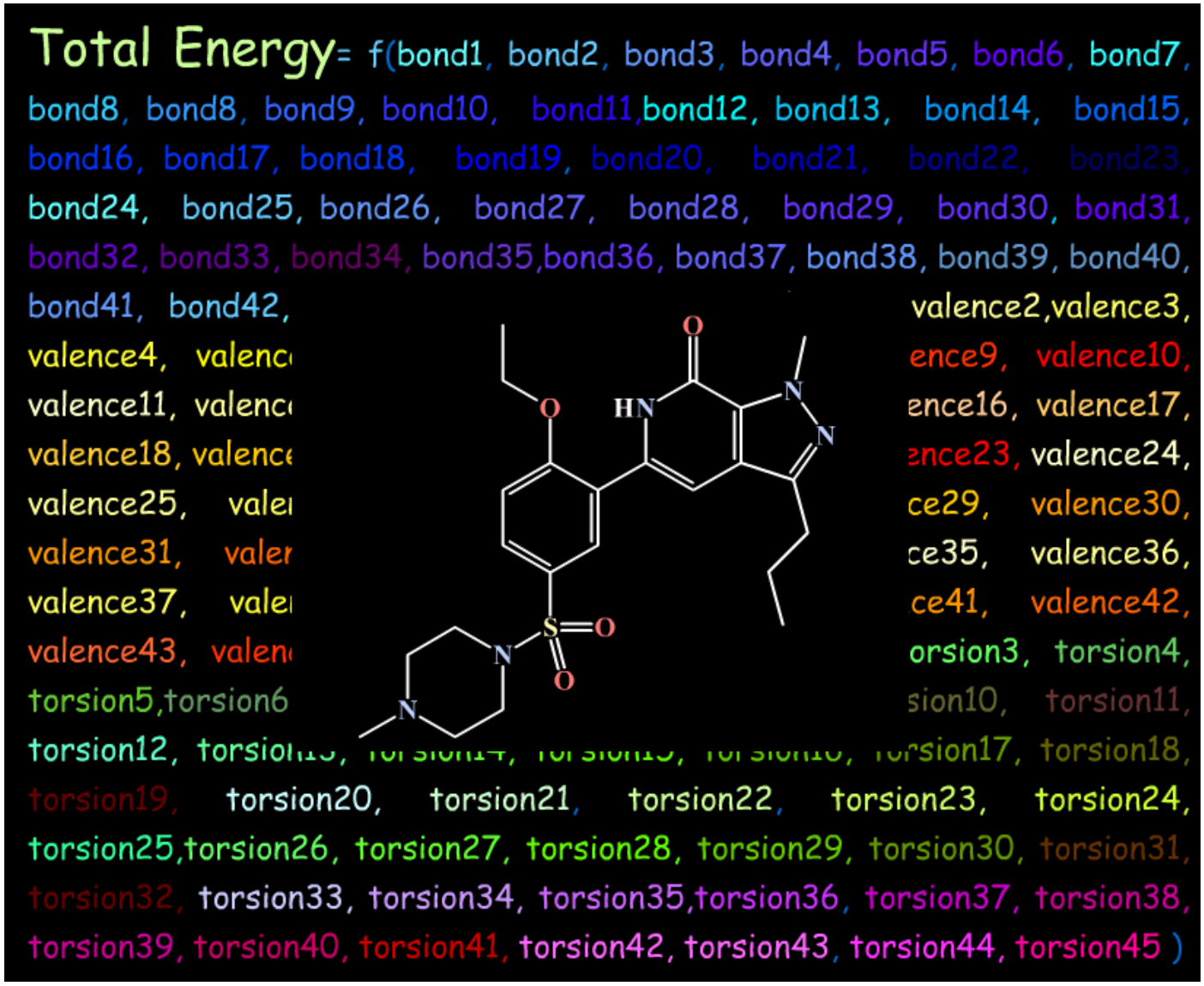


Valence angle bending helps release some strain…

- Bond length oscillations mainly affect spectral properties

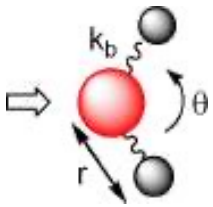


Energy: function of internal coordinates, which depend on Cartesian *x,y,z*



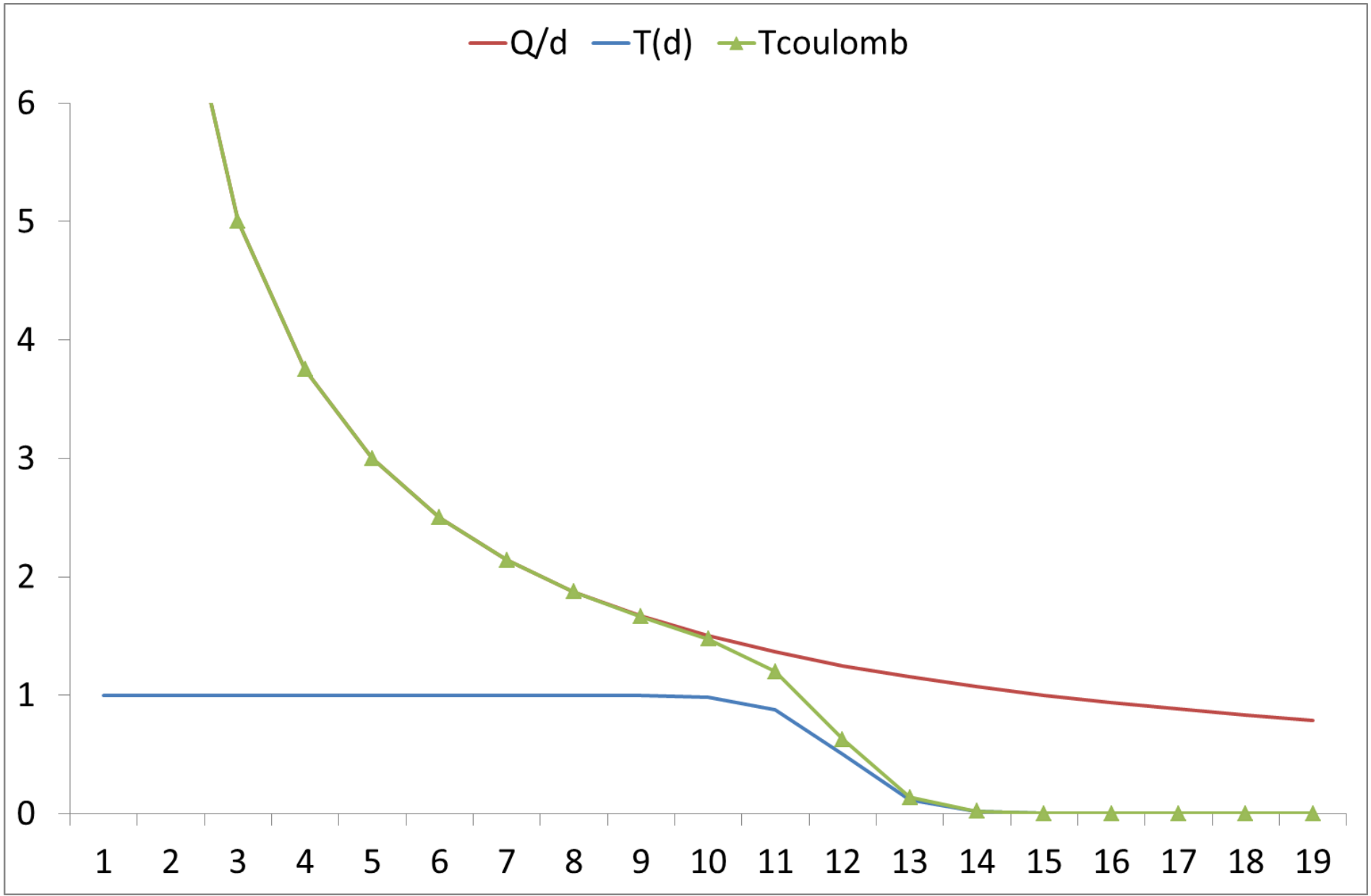
Potential energy calculation is based on the Empirical Force Field (FF) approach

- Quantum chemical calculations are too time-consuming: atoms and their interactions are approximated as "classical" objects
- Atoms need to be "parameterized" in function of their chemical environment
- Covalent terms:



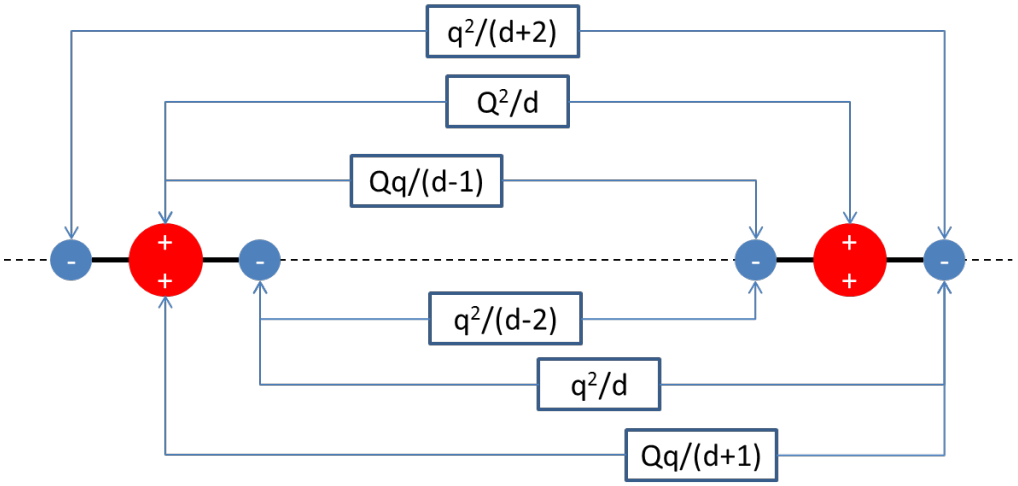
  - bonds are modeled as harmonic springs. The energy required to stretch or compress a bond by Δb with respect to its natural length b is expressed as $K_b \Delta b^2$
  - Valence angle bending modeled by harmonic potential $K_\phi \Delta \phi^2$
- Atoms that are not directly bonded or do not form an angle interact "through space" by means of non-bonded interactions.
  - Electrostatics interactions – based on partial charges
  - Continuum Solvent models
  - Van der Waals interactions
- Terms that should not be, but they're needed to make it work!
  - Torsional potentials, cross-terms, *etc.*

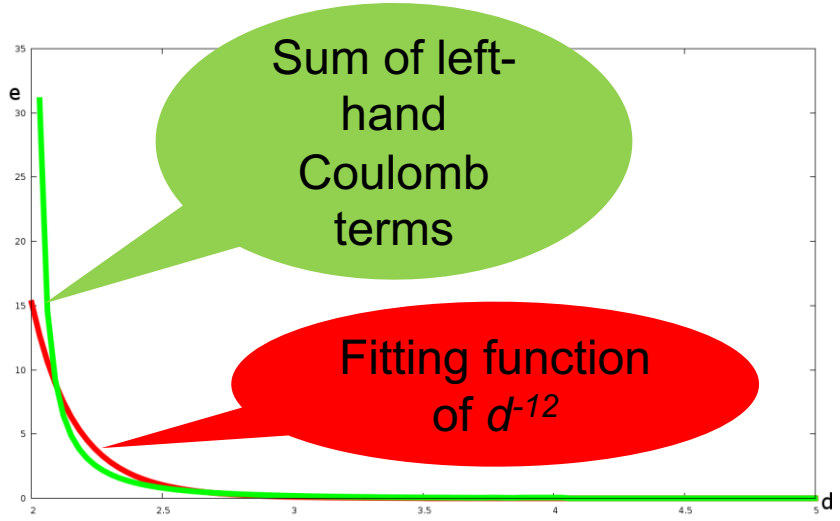
Non-bonded interactions: (1) - Coulomb



Non-bonded interactions: (2) – van der Waals terms

- These are "cryptoelectrostatic" interactions:
  - a repulsive term acting at low distances, when electron spheres of non-bonded atoms overlap (d^{-12})
  - An attractive term in d^{-6} representing London dispersion terms (fluctuation-induced dipole-dipole interactions)

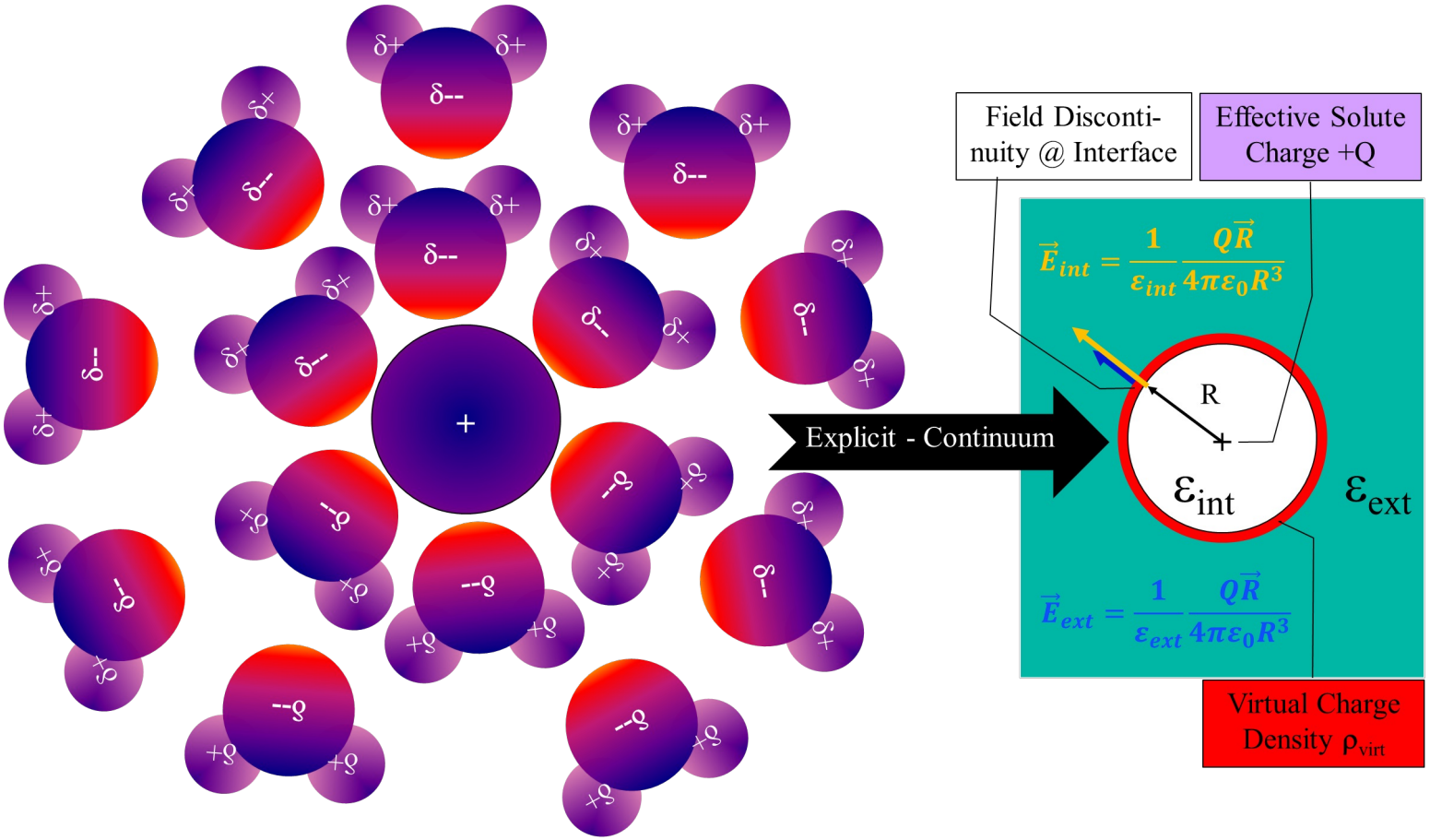




$$E_{VdW} = \frac{\sqrt{A_i A_j}}{d_{ij}^{12}} - \frac{\sqrt{B_i B_j}}{d_{ij}^{6}}$$

Solvent effects…

- Explicit solvent molecules not only increase system size, but must be simulated in all their possible orientations with respect to the solute…
- Continuum solvent models assume dielectric effects to accurately capture this averaging over all possible solvent states…



Continuum solvent models may be as complex as you please…

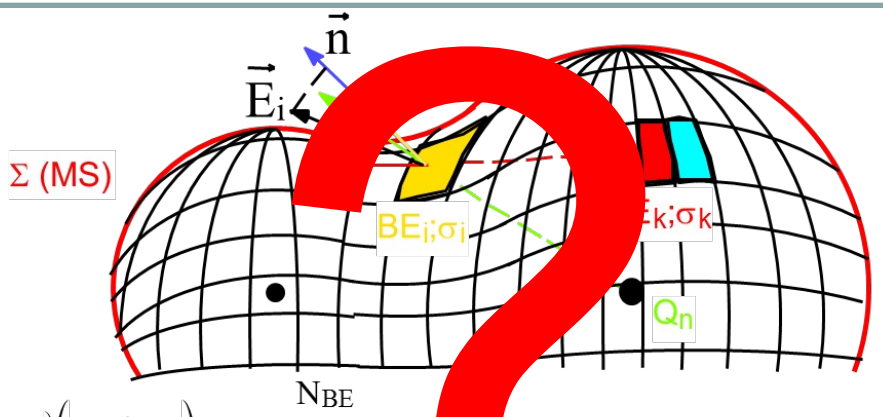
Empirical Terms

$$F_{Solv} = F_{Pol} + F_{Hphob} = k_{solv}\frac{Q_i^2 V_j + Q_j^2 V_i}{d_{ij}^4} - k_h \delta^{hphob}(i,j)$$

Generalized Born Methods α: Generalized Born Radii

$$F_{pol} = \left(\frac{1}{\varepsilon_{ext}} - \frac{1}{\varepsilon_{int}}\right)\sum_{i,j=1}^{N_{atoms}} \frac{Q_i Q_j}{8\pi\varepsilon_0\sqrt{d_{ij}^2 + \alpha_i\alpha_j e^{\left(-d_{ij}^2/4\alpha_i\alpha_j\right)}}}$$

Explicit Solving of Poisson – Boltzmann Equation



$$\frac{\sigma_i}{\varepsilon_0} = \left(\vec{E}_i.\vec{n}\right)\left(1-\frac{\varepsilon_{ext}}{\varepsilon_{int}}\right) \Rightarrow \sum_k a_{ik}\sigma_k = b_i$$

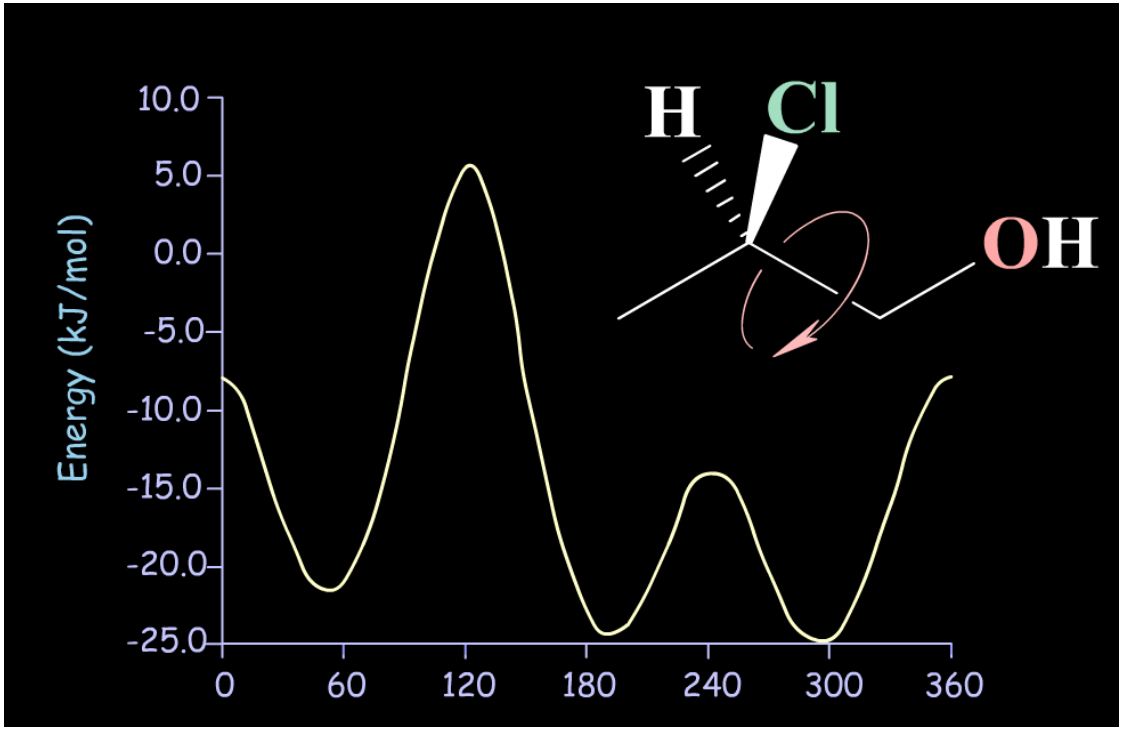
$$\int_{(\Sigma)} \sigma dS \sum_k^{N_{BE}} \sigma_k.\delta S_k = Q_{(\Sigma)}\left(\frac{1}{\varepsilon_{ext}} - \frac{1}{\varepsilon_{int}}\right)$$

$$\Rightarrow F^{pol} = \sum_{k=1}^{N_{atoms}}\left[\frac{Q_k}{8\pi\varepsilon_0}.\sum_{p=1}^{N_{BE}}\frac{\sigma_p\delta S_p}{r_{kp}}\right]$$

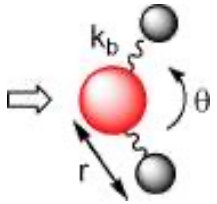
Torsional terms – what for?

$$E_{tors} = k_t\left[1+\cos(3\Theta-\Theta_0)\right]$$

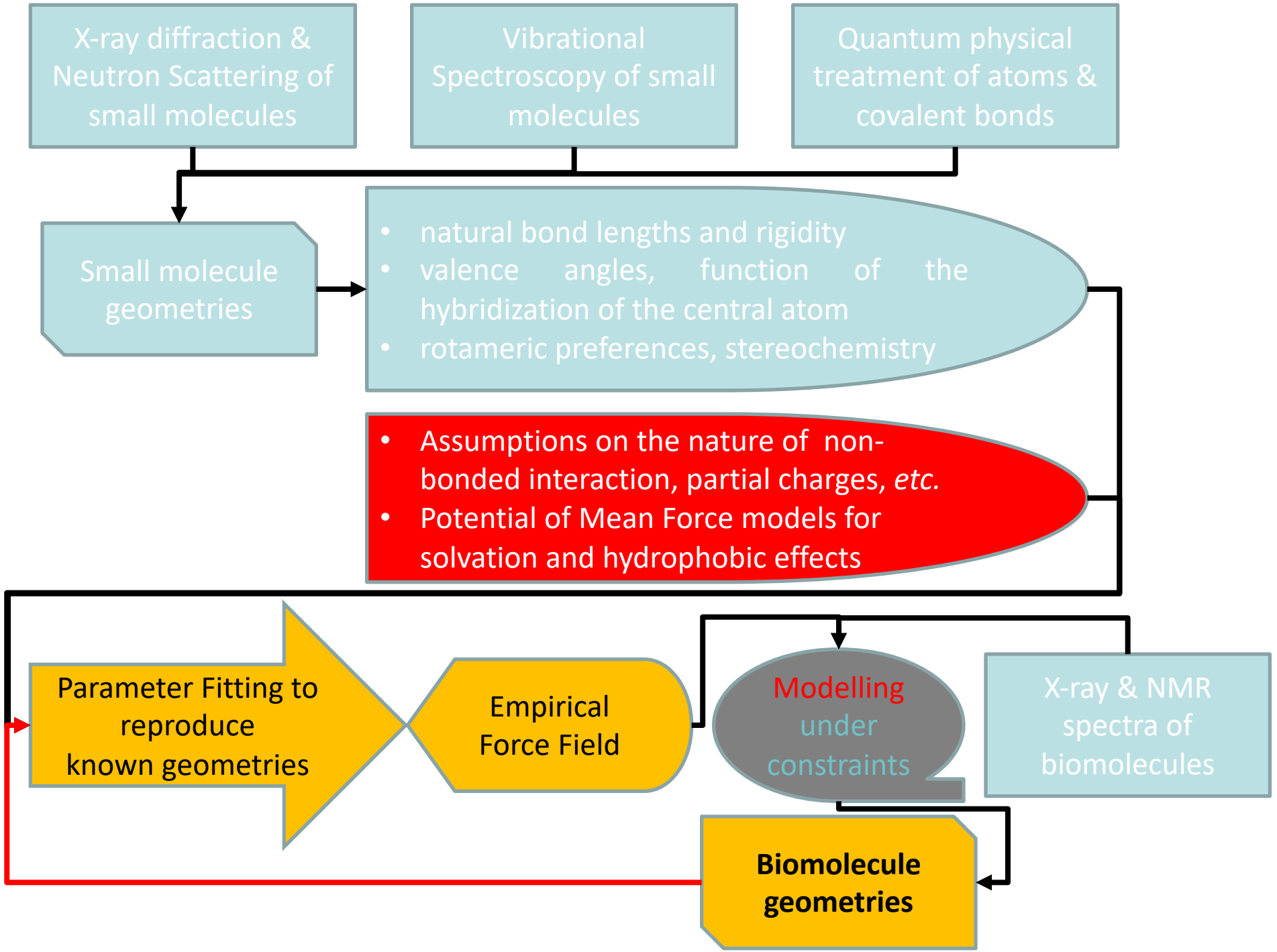
- They must be introduced in order to correct for incompatibilities between long-range and short-range van der Waals contributions (*N.Allinger, MM2*)
- A set of FF parameters must be *consistent* – do not expect individual terms to have any physical meaning (*K. Rasmussen*)



Setting up a Force Field – where do all those parameters come from ?



- Few are directly issued from experimental observations:
  - bond & angle deformation constants relate to IR vibration frequencies
  - van der Waals parameters can be measured… for ideal gas atoms.
- Atomic partial charges from electronegativity equilibration, molecular orbital “collapsing”.
- Most are fitted, making sure that force field simulations reproduce:
  - experimentally determined geometries & interconformational barriers
  - Quantum-chemically determined potential energy landscape


X-ray diffraction & Neutron Scattering of small molecules
Vibrational Spectroscopy of small molecules
Quantum physical treatment of atoms & covalent bonds
Small molecule geometries
• natural bond lengths and rigidity
• valence angles, function of the hybridization of the central atom
• rotameric preferences, stereochemistry
• Assumptions on the nature of non-bonded interaction, partial charges, etc.
• Potential of Mean Force models for solvation and hydrophobic effects
Parameter Fitting to reproduce known geometries
Empirical Force Field
Modelling under constraints
X-ray & NMR spectra of biomolecules
Biomolecule geometries

Coarse-grain (CG) models for huge molecules….

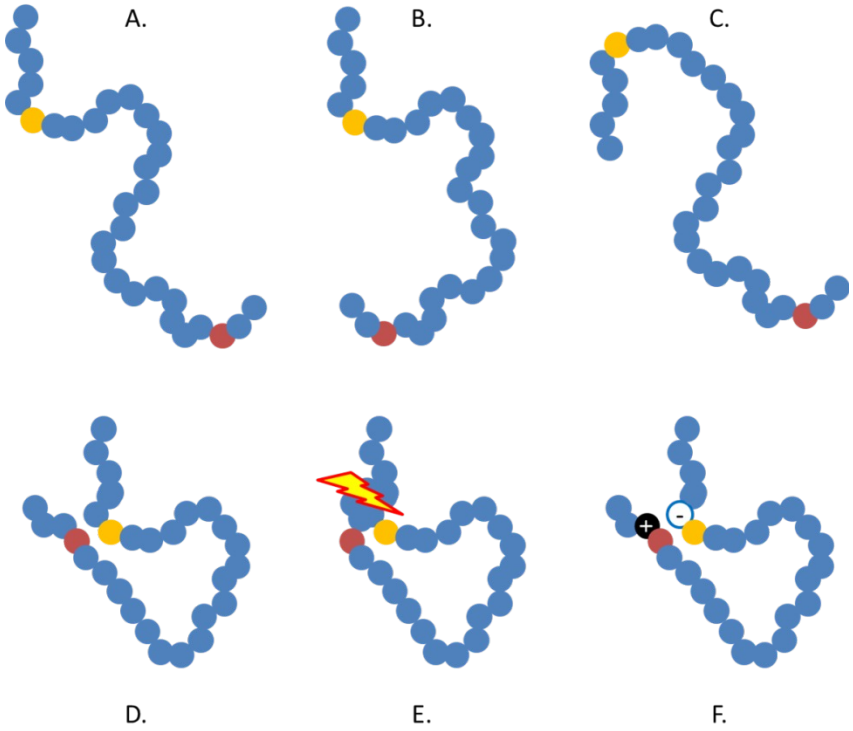
- Replace standard fragments by some generic ‘object’ (bead):
    - What is the functional form of bead-bead interaction? Such “Potential of Mean Force” (PMF) should represent the average contact strength at inter-bead distance *d*.

$$PMF_{CG}(\text{d}) = -k_B T \ln Z_d =$$

$$f\{E(\text{}), E(\text{}), E(\text{}), \ldots\}$$

The back-engineering alternative: Knowledge-based potentials

- If, in a structural data base, the red-yellow inter-particle distance is, on the average, *shorter* than expected (over the average of all possible chain arrangements – geometries like “D” seen more often than “normal”), then one may conclude that they **attract** each other



- *O(d)* = observed frequency @ *d*
- *B(d)* = baseline frequency @ *d*

$$\Rightarrow \exp\left[-\frac{E(d)}{k_b T}\right] = \frac{O(d)}{B(d)}$$

The Question: Why care for conformational sampling?

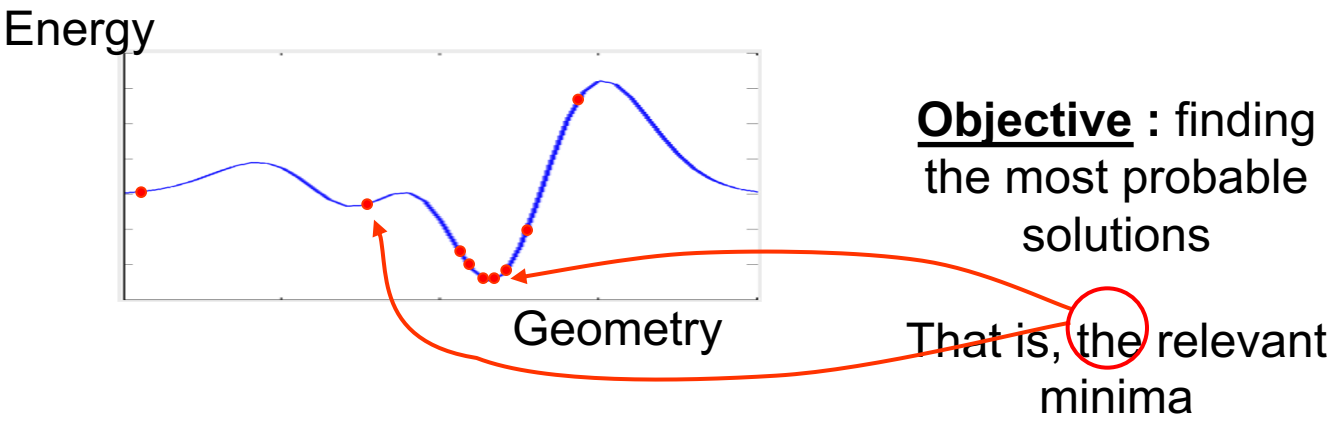
– Because experimental properties of a molecule are given by the Boltzmann Average of properties of populated geometries

Boltzmann's probability distribution:

$$P(geometry\ of\ energy\ E) \sim \exp\left\{-\frac{E}{k_B T}\right\}$$

Boltzmann Averaging:

$$ObservedProperty = \sum_{\substack{possible \\ geometries}} P(geometry) \times Property(geometry)$$

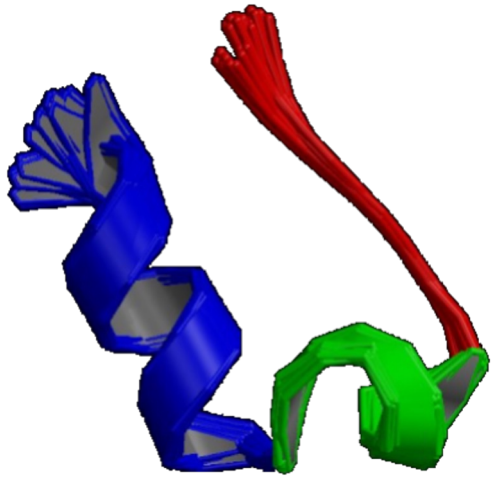


By the way… what *is* a "Conformer"?

- A conformer is the set of geometries within the neighborhood of a local or global minimum energy geometry.
  - A geometry is a *point* in coordinate Phase Space, $\vec{X} = (x_1, y_1, z_1, x_2, y_2, ..., x_N, y_N, z_N)$
  - A conformer is a Phase Space Domain *C* containing many similar geometries
- Therefore, a geometry is characterized by its energy, a conformer by its *partition function Z* and *free energy F*

$$Z(C) = \sum_{geometries \in C} \exp\left\{\frac{-E(geometry)}{k_B T}\right\}$$

$$F(C) = -k_B T \ln Z(C)$$



- At equal minimum depth *min E* over all geometries, the *broader energy well* (with more distinct low-energy geometries) will correspond to the experimentally more often seen conformer.

$$P(C) \sim \exp\left\{-\frac{F(C)}{k_B T}\right\}$$

ONE

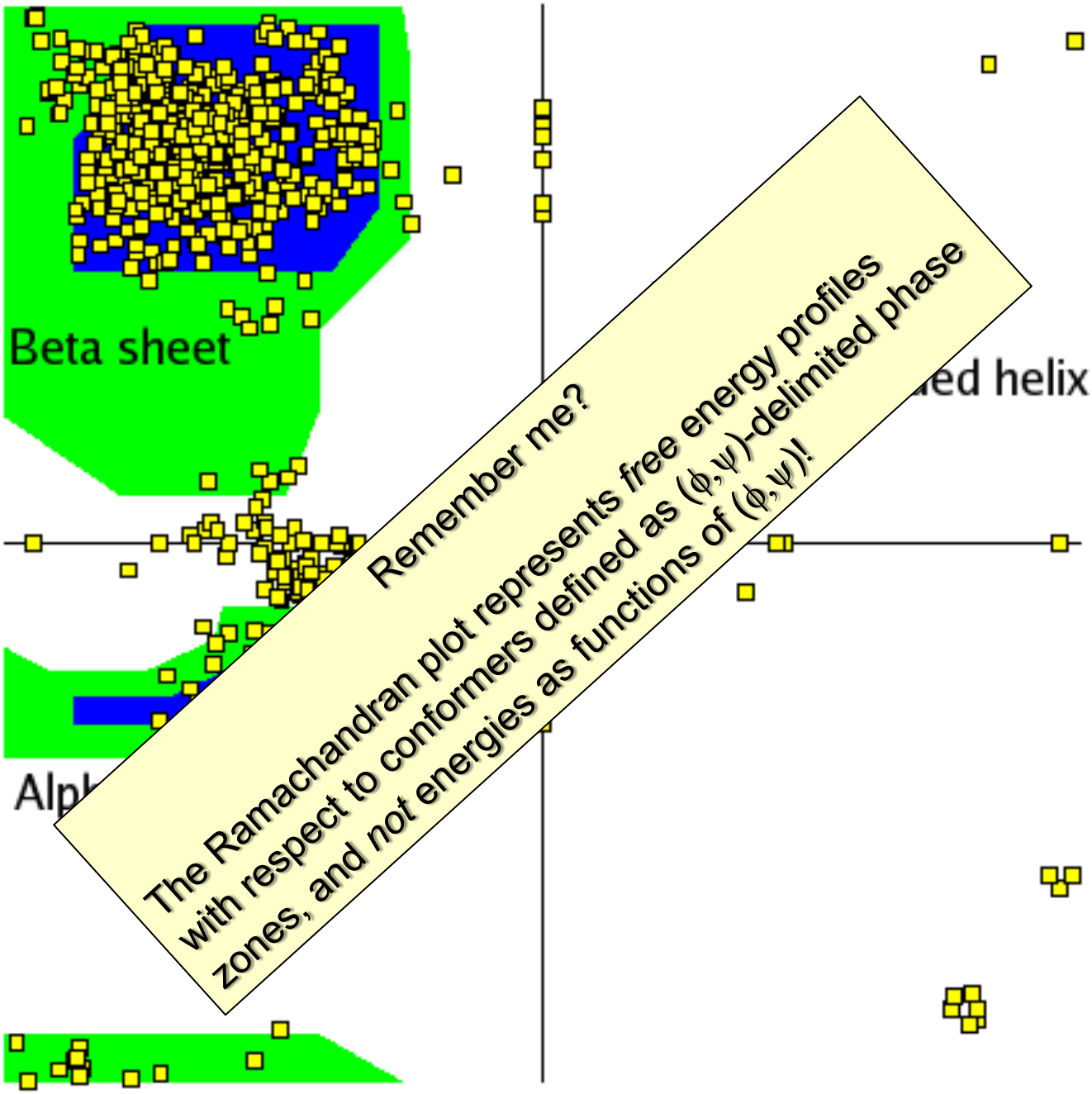
THERMODYNAMICS AND STATISTICAL MECHANICS

1.1 INTRODUCTION: THERMODYNAMICS AND STATISTICAL MECHANICS OF THE PERFECT GAS

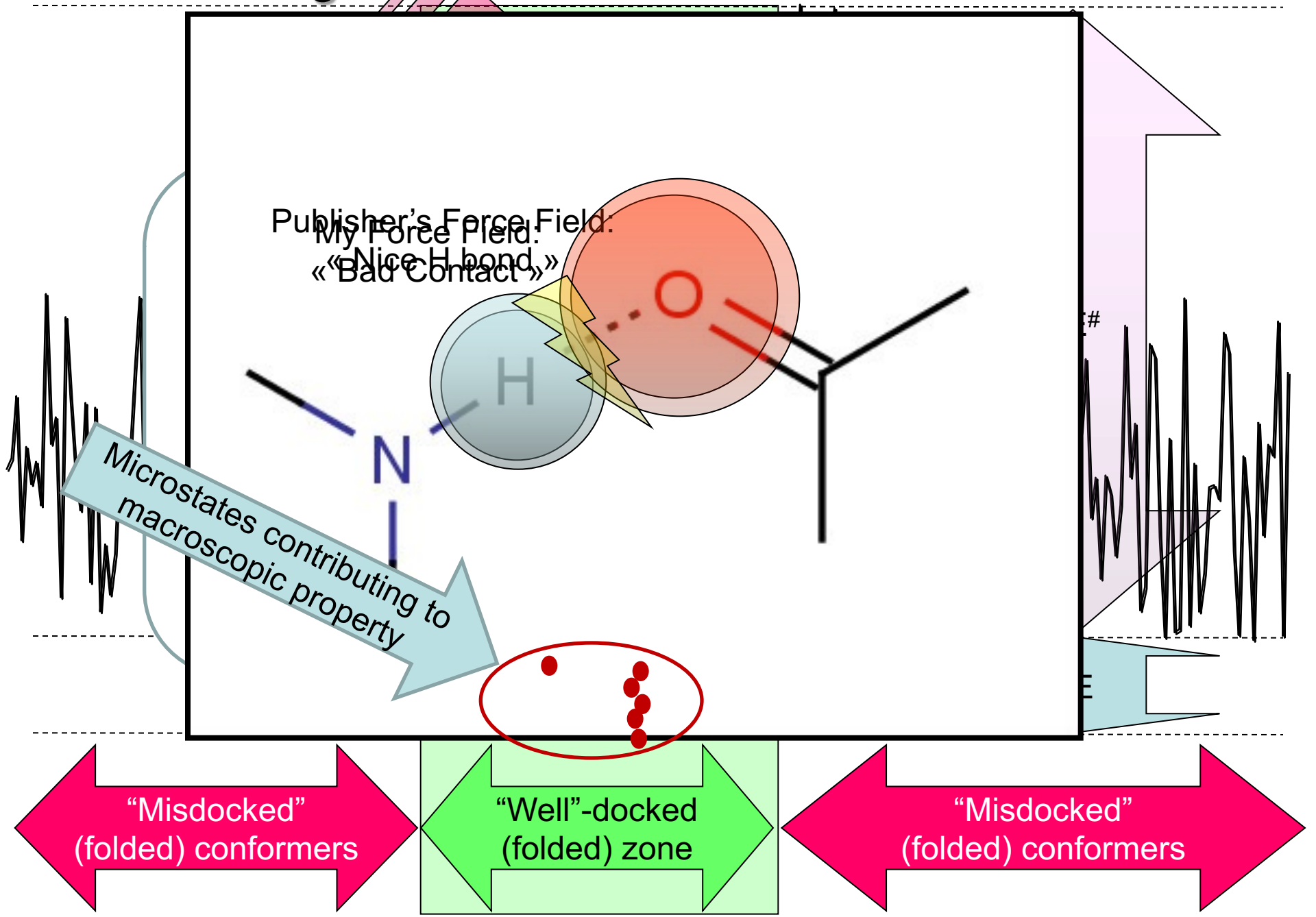
Ludwig Boltzmann, who spent much of his life studying statistical mechanics, died in 1906, by his own hand. Paul Ehrenfest, carrying on the work, died similarly in 1933. Now it is our turn to study statistical mechanics.

Perhaps it will be wise to approach the subject cautiously. We will begin by considering the simplest meaningful example, the perfect gas, in order to get the central concepts sorted out. In Chap. 2 we will return to complete the solution of that problem, and the results will provide the foundation of much of the rest of the book.

The quantum mechanical solution for the energy levels of a particle in a box (with periodic boundary conditions) is


Beta sheet
ed helix
Alp
Remember me?
The Ramachandran plot represents *free* energy profiles with respect to conformers defined as (ϕ,ψ)-delimited phase zones, and *not* energies as functions of (ϕ,ψ)!

The Challenge…



- *Presentation Outline*
  - The Basics: Molecules have Geometries!
    - Intramolecular energy: the Empirical Force Field
  - Sampling Methods: a brief overview
    - Molecular Dynamics: Walking like a molecule
    - Monte Carlo: Molecular Casino
    - Evolutionary Methods: in God/Darwin we trust!
  - Conclusions

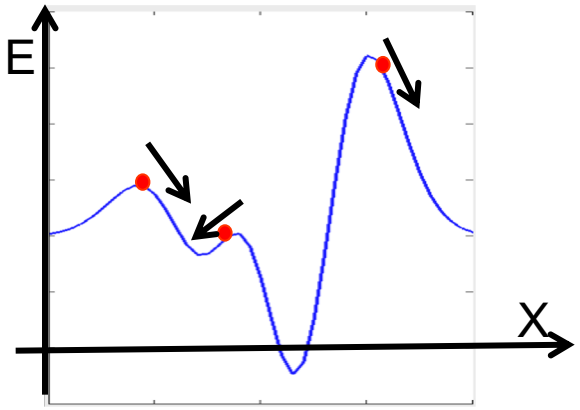
Energy Minimization is only [the easy] part of the problem

- Given a starting geometry, deterministic algorithms allow the discovery of the adjacent local minimum
- Descent methods follow the local gradient

$$molecular\ geometry\ \vec{X} = \left(x_1, y_1, z_1, x_2, y_2, \ldots, x_N, y_N, z_N\right)$$

$$\nabla E(\vec{X}) = \left(\frac{\partial E}{\partial x_1}, \frac{\partial E}{\partial y_1}, \frac{\partial E}{\partial z_1}, \frac{\partial E}{\partial x_2}, \frac{\partial E}{\partial y_2}, \ldots, \frac{\partial E}{\partial x_N}, \frac{\partial E}{\partial y_N}, \frac{\partial E}{\partial z_N}\right)$$

$$iteratively: \quad \vec{X}^{new} = \vec{X}^{curr} - s\nabla E(\vec{X}^{curr})$$



Molecular Dynamics: walk the energy surface like the molecule does… following Newtonian dynamics

- Assume random atom velocities according to Maxwell's distribution at temperature T
- Calculate forces on atoms $\vec{F} = -\nabla E(\vec{X}) = -\left(\frac{\partial E}{\partial x_1}, \frac{\partial E}{\partial y_1}, \frac{\partial E}{\partial z_1}, \frac{\partial E}{\partial x_2}, \frac{\partial E}{\partial y_2}, \ldots, \frac{\partial E}{\partial x_N}, \frac{\partial E}{\partial y_N}, \frac{\partial E}{\partial z_N}\right)$
- Recalculate velocities at future time point: $v(t+dt) = v(t) + F/m \times dt$
- Recalculate coordinates at future time point:

$$x(t+dt) = x(t) + \frac{v(t) + v(t+dt)}{2} \times dt$$

- Continue for the relevant time scale of the phenomenon to simulate… all while recalling that *dt* must be small enough to ensure that F≈constant during *dt.*
  - Bond stretching forces are the most rapidly varying – they impose ***dt*=10^{-15} s**

But… a molecule is never alone: need to simulate the its energy exchange with the rest of the Universe…

- "Microcanonical" (purely Newtonian) dynamics conserves total energy…
- Langevin dynamics emulates both the intrinsic "viscosity" of the environment (friction term slowing down fast atoms, controlled by γ) and occasional energy gains due to stochastic collisions with environment atoms (controlled by η)

$$\mathbf{F}_i = m_i \frac{\mathrm{d}^2\mathbf{x}_i}{\mathrm{d}t^2} = -\nabla E(\mathbf{x}_i) - \gamma m_i \frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t} + \eta\sqrt{m_i T_0}\delta,\ \delta \sim \mathcal{N}(0,1)$$

- Berendsen's empirical "coupling to a thermal bath" basically does the same, with a different formalism.

$$\mathbf{F}_i = m_i \frac{\mathrm{d}^2\mathbf{x}_i}{\mathrm{d}t^2} = -\nabla E(\mathbf{x}_i) - \frac{1}{\tau}\left(1 - \frac{T_0}{T}\right) m_i \frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t}$$

- Nose-Hoover's formalism is formally proven to be correct from a statistical physics standpoint… but quite tricky to use!

Now, a molecule is not interested in *effectively* exploring its conformational space…

- So, it must be pushed towards more efficacious sampling, by creatively "tampering" with Newton's equations…
  - Tabu or Memory dynamics, a.k.a "Poling": forces the trajectory to become a self-avoiding walk, by means of bias potential terms that are large and positive for geometries similar to already visited ones, and zero outside the already explored phase space domain!
  - Potential Smoothing approaches: Artificially lower the high energy barriers of the energy landscape, without modifying the low energy zones – thus ensure for faster transitions…
  - Adaptive bias: add an extraneous force coupled to some global geometric descriptor (say: gyration radius), thus compelling the algorithm to generate both unfolded, linear and folded, globular geometries
  - Umbrella sampling: force the MD trajectory to follow a predefined "reaction path"… if you have a clue on how to define it!

The M timum
by Pla

– T

– R tion in
C

– A mall
st

– R
g

•

•

– L

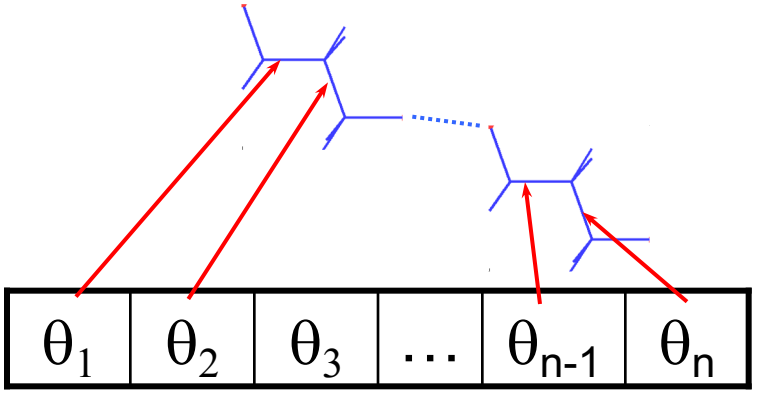
criterion,

Genetic Algorithms

- Applying a Darwinian Evolution Scenario to a population of vectors (“chromosomes”) encoding the solution to a problem
- Solution Quality is the “Fitness” score, and the fittest survive…

Data representation :

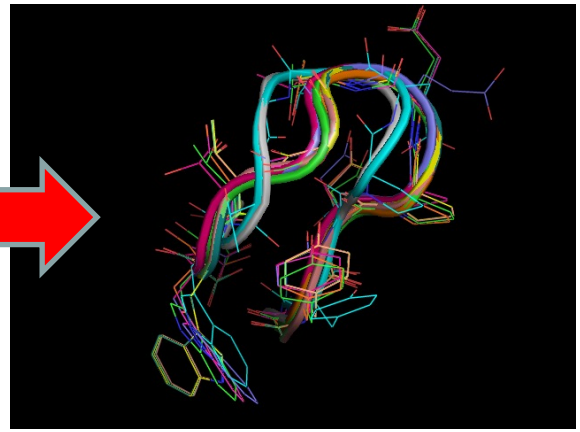
« individual »
or
« chromosome »
=
list of its torsional angles



Population of individuals :

θ^1_1	θ^1_2	θ^1_3	…	…	…	…	…	…	…	θ^1_n
θ^2_1	θ^2_2	θ^2_3	…	…	…	…	…	…	…	θ^2_n
θ^3_1	θ^3_2	θ^3_3	…	…	…	…	…	…	…	θ^3_n
θ^4_1	θ^4_2	θ^4_3	…	…	…	…	…	…	…	θ^4_n

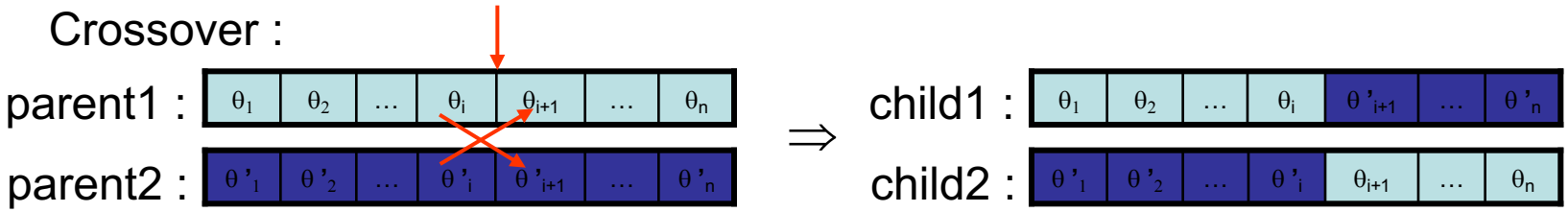
…



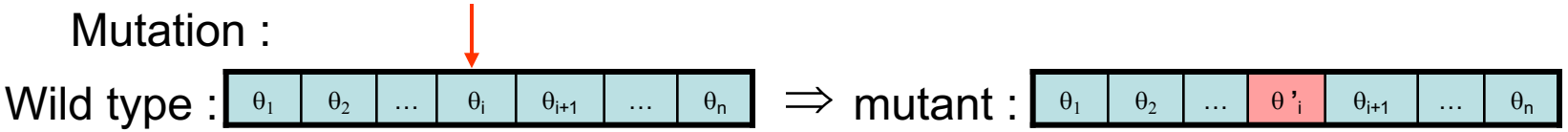
Adults only

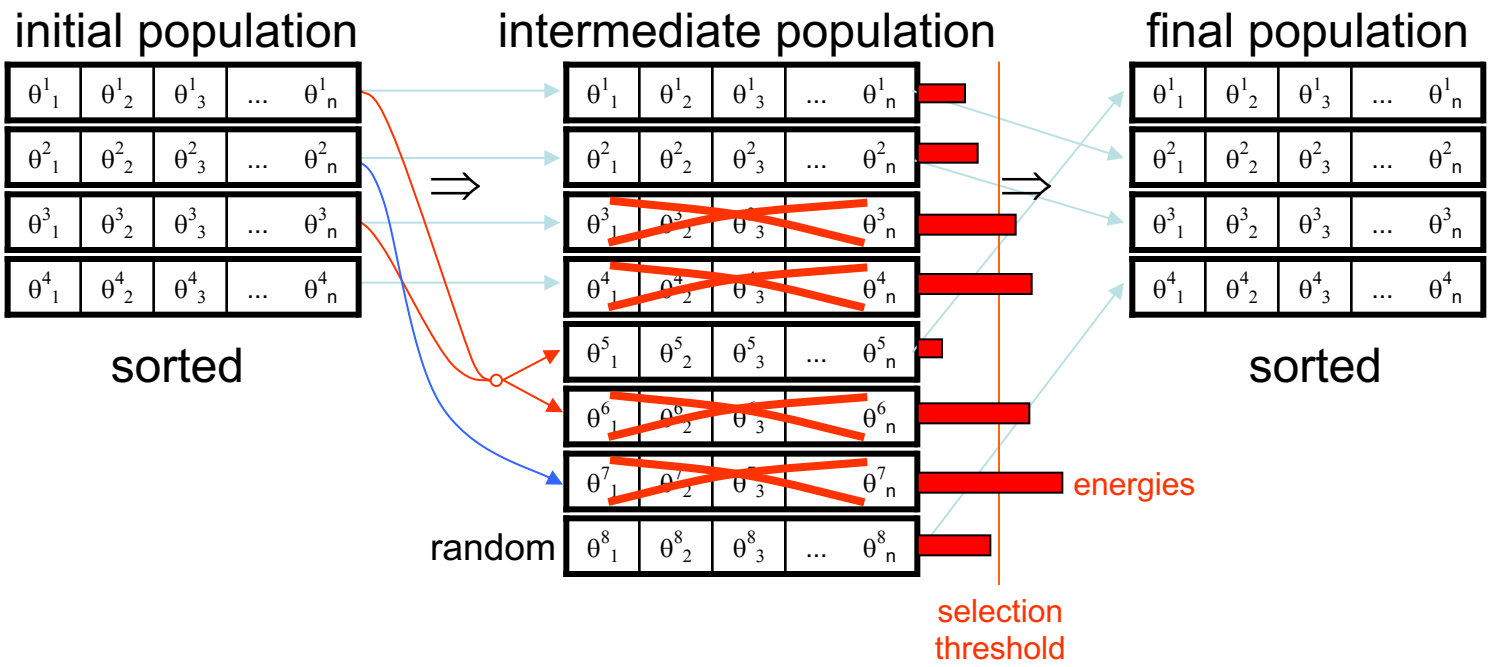
Generation of new offspring :

Crossover :



Mutation :





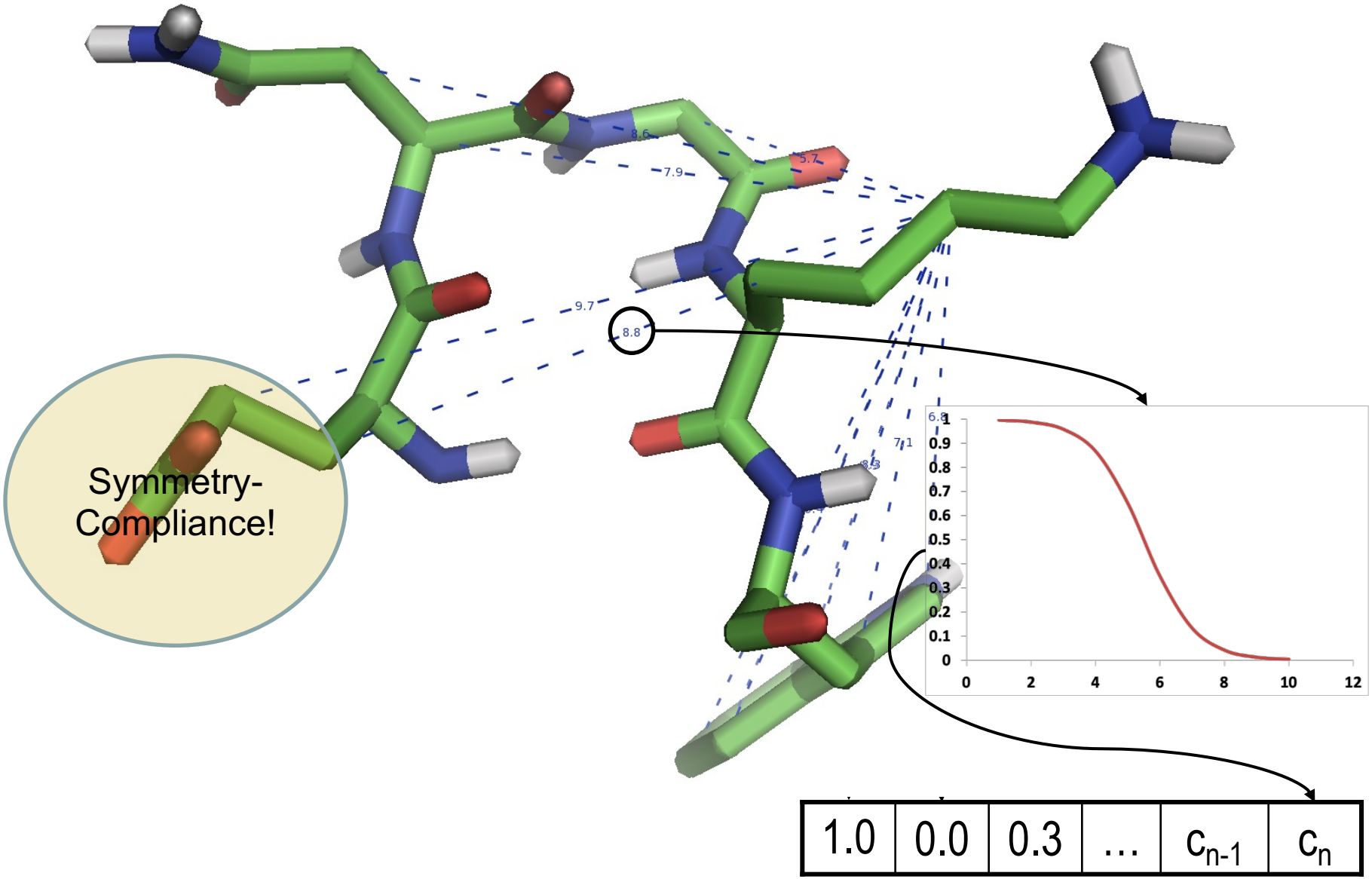
Evolution of the average fitness,
Evolution of the fitness of the best

$\Rightarrow$ **the algorithm converges**

Population Diversity Control is a Key Issue

- Multiple 'Island' models – parallel simulations occasionally swapping solutions
- Discarding of redundant chromosomes (requires a metric defining how similar two encoded solutions are!)

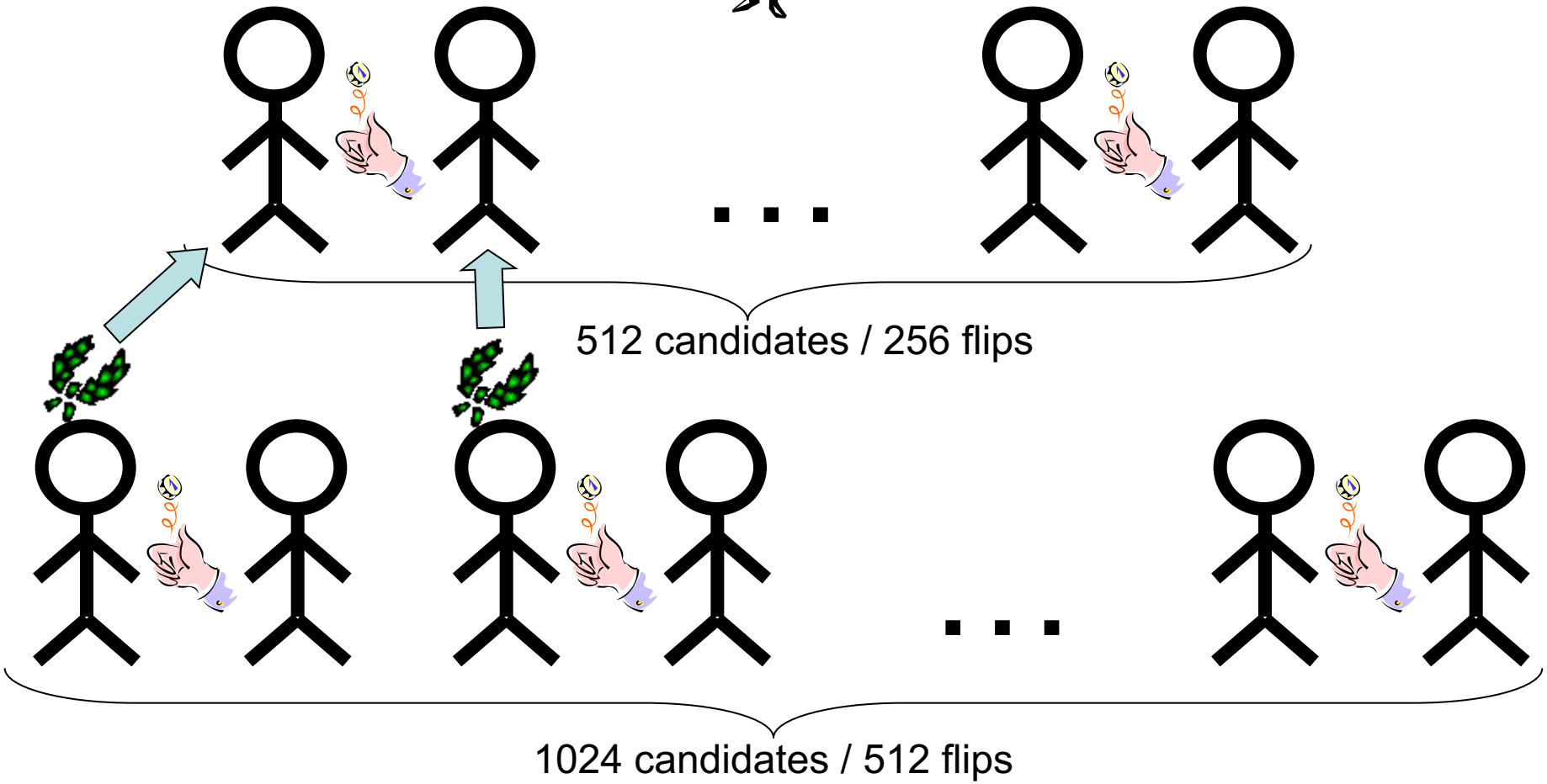
Interaction Fingerprints to check for redundancy



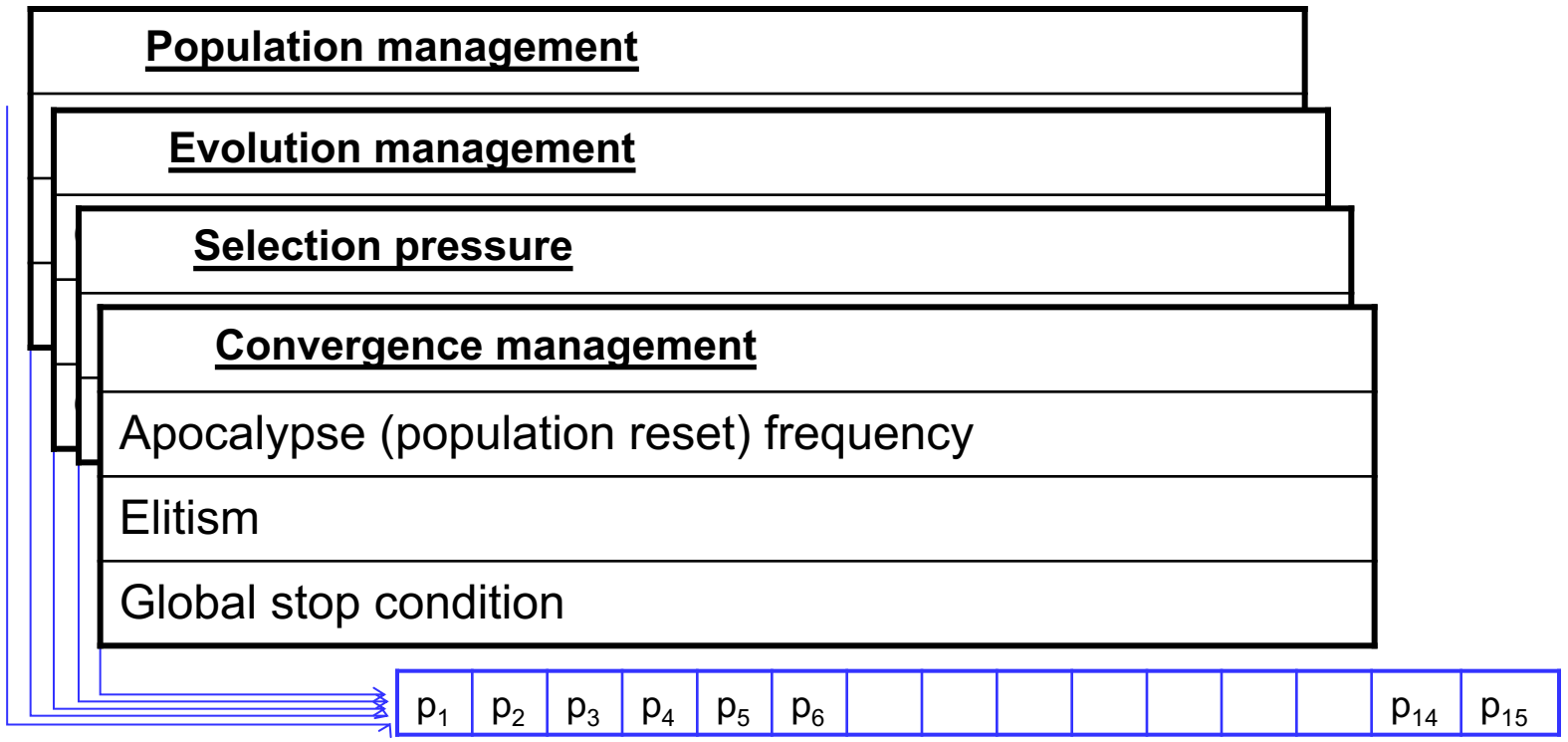
Genetic Algorithms: Chance, Selection & the Coin Flipper's bet!

- Any problem admitting a *vector* as a solution may be coded by a "chromosome" and left in the hands of Darwin…**or God??**
- I bet (1M€) I can find a person who won a coin-flipping challenge 10 times in a row, at his/her first attempt!!
  - In order to fulfill my promise, I need a total of 1024 coin flips to happen,
    - 1024/10=102 pretendents, each with a chance of $(1/2)^{10}$ to score 10 successive winning coin flips: ~90% chance to loose 1M€!
    - If you read "Darwin's Dangerous Idea" by D.C.Dennett, you are not allowed to bet !!

Selection is the Key!

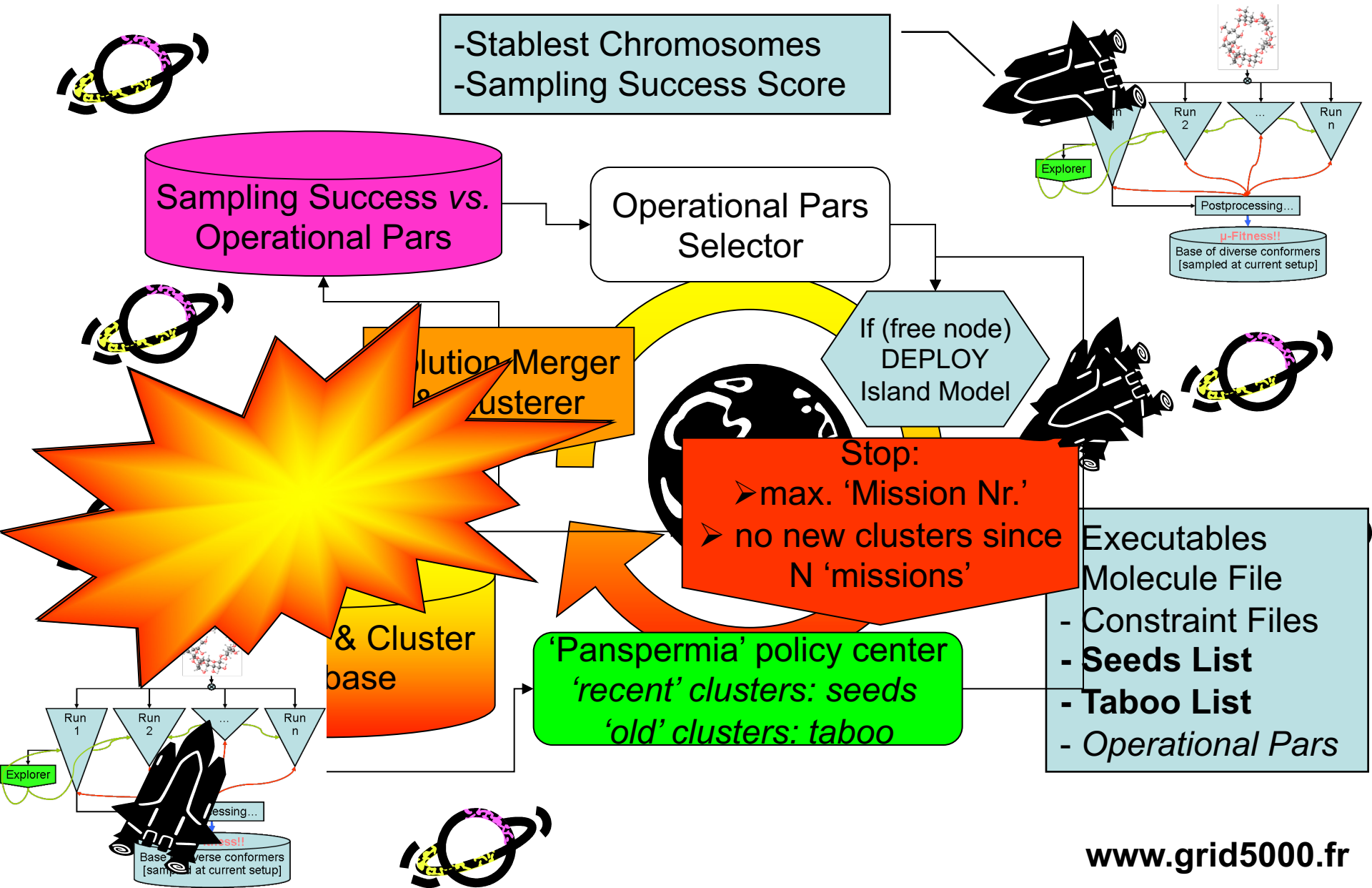


Search for Optimal Sampling Setups in the Strategy Parameter Space…



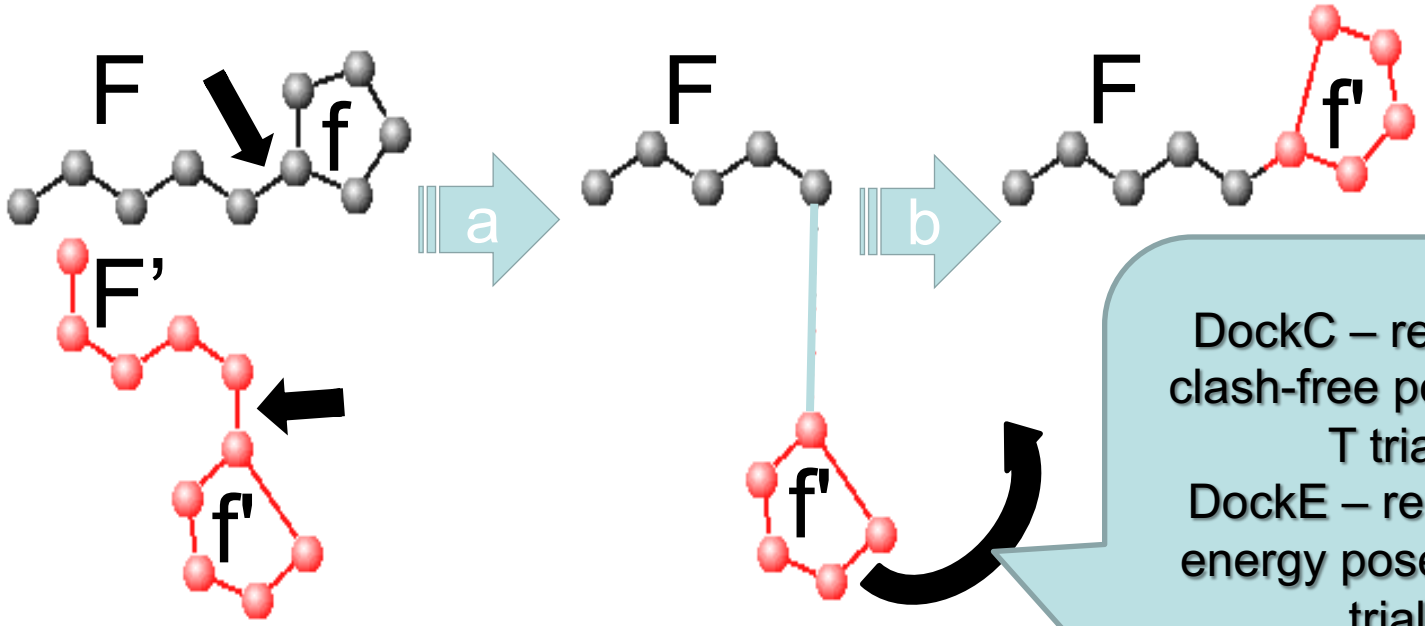
$$SamplingSuccessScore = f(p_1, p_2, ..., p_n)$$

GRID 5000-based ‘Planetary’ Model



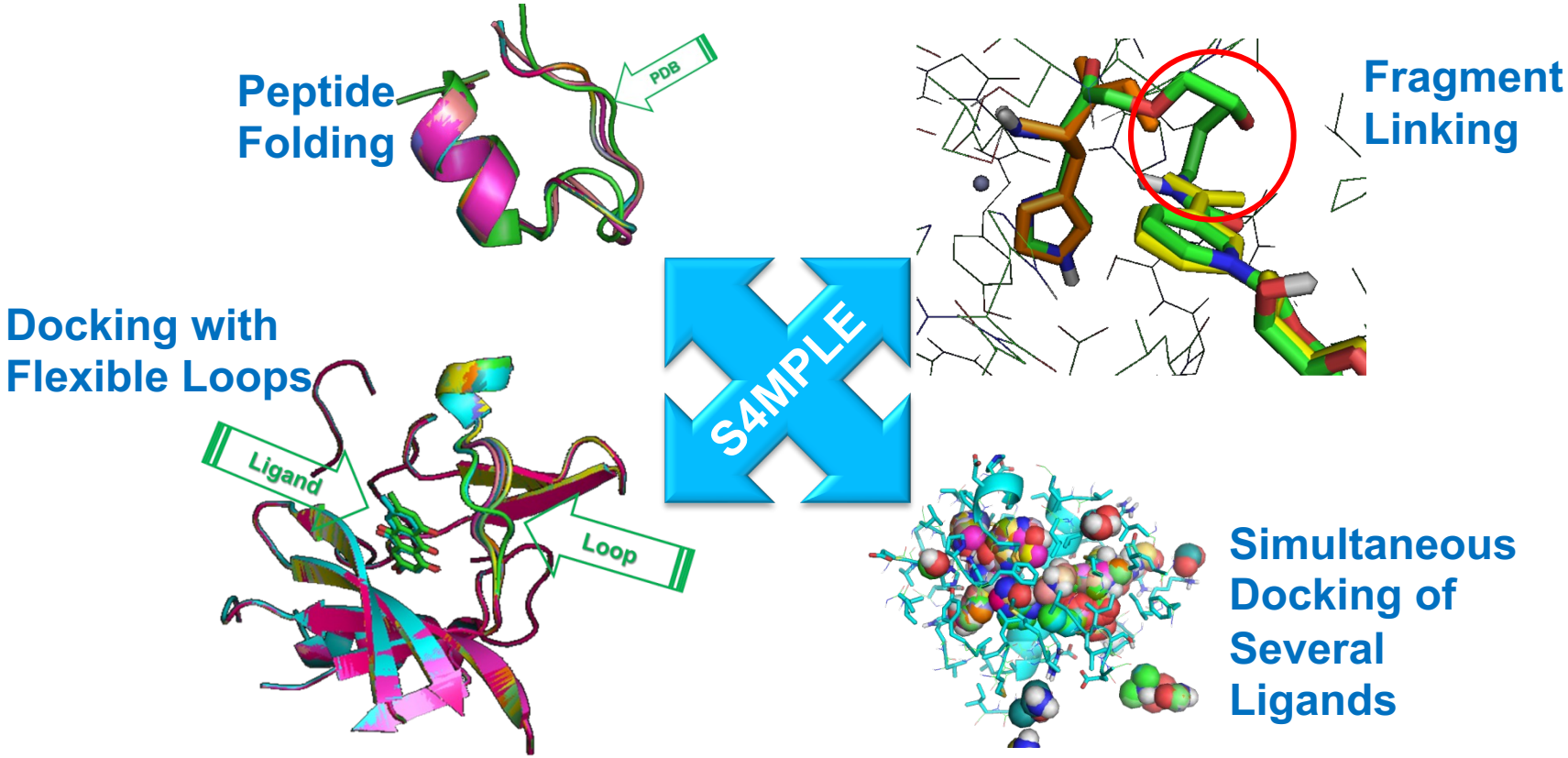
www.grid5000.fr

Essential: you do not NEED to stick to torsions: Genetic Operators in Fully Flexible Mode (S4MPLE)



- If fragments are not bound, a putative favorable contact is used instead of the bond!

Evolutionary Computing allows a unified approach to conformational sampling…



- *Presentation Outline*
  - The Basics: Molecules have Geometries!
    - Intramolecular energy: the Empirical Force Field
  - Sampling Methods: a brief overview
    - Molecular Dynamics: Walking like a molecule
    - Monte Carlo: Molecular Casino
    - Evolutionary Methods: in God/Darwin we trust!
  - Conclusions

Conclusions

- Conformational Sampling is the Key Element for Understanding of Molecular Behavior
- It may range from very simple to extremely difficult, to impossible
- **If you don't do it well, better don't do it at all**: empirical methods based on molecular topology only may be more accurate than 3D models based on wrong – or too few – conformations
- Two main sources of errors: A.) wrong calculated energy-geometry landscape (poor Force Field parameterization) and B.) – insufficient sampling!
- Docking is just a specific case of conformational sampling, involving at least two molecules: a binding "site" and one or more "ligands"
- You will often hear that the knowledge of the "bioactive" conformer is paramount to understand binding. This is necessary, but sometimes not sufficient.
- Entropic effects perversely insist on being important, in spite our inability to properly estimate them!