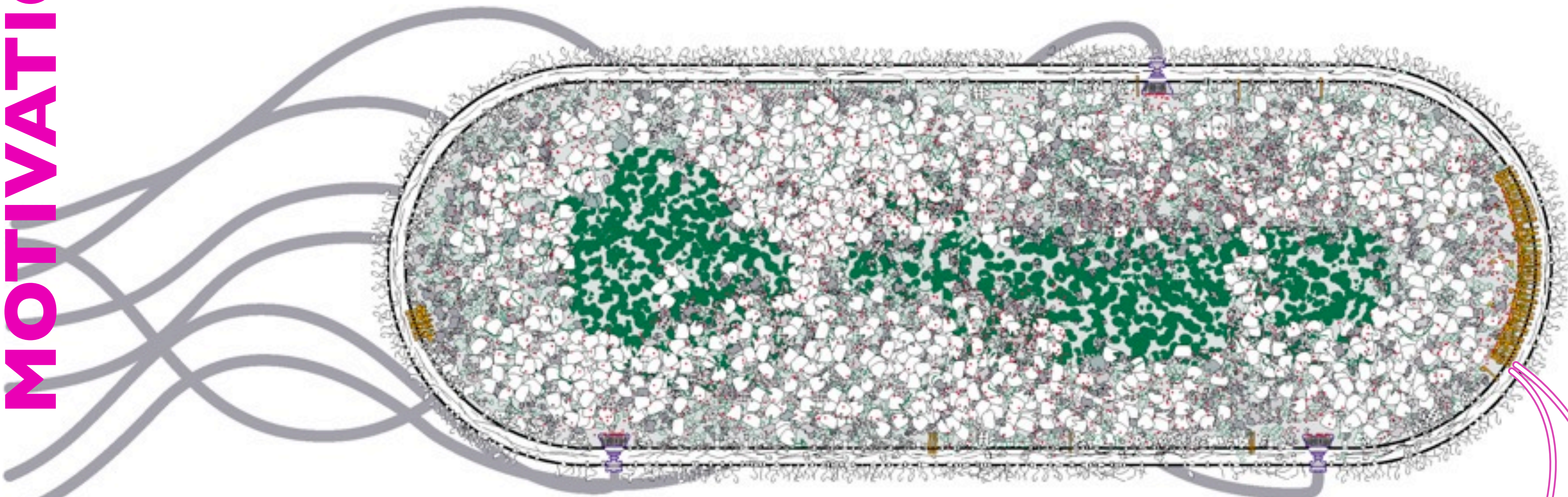
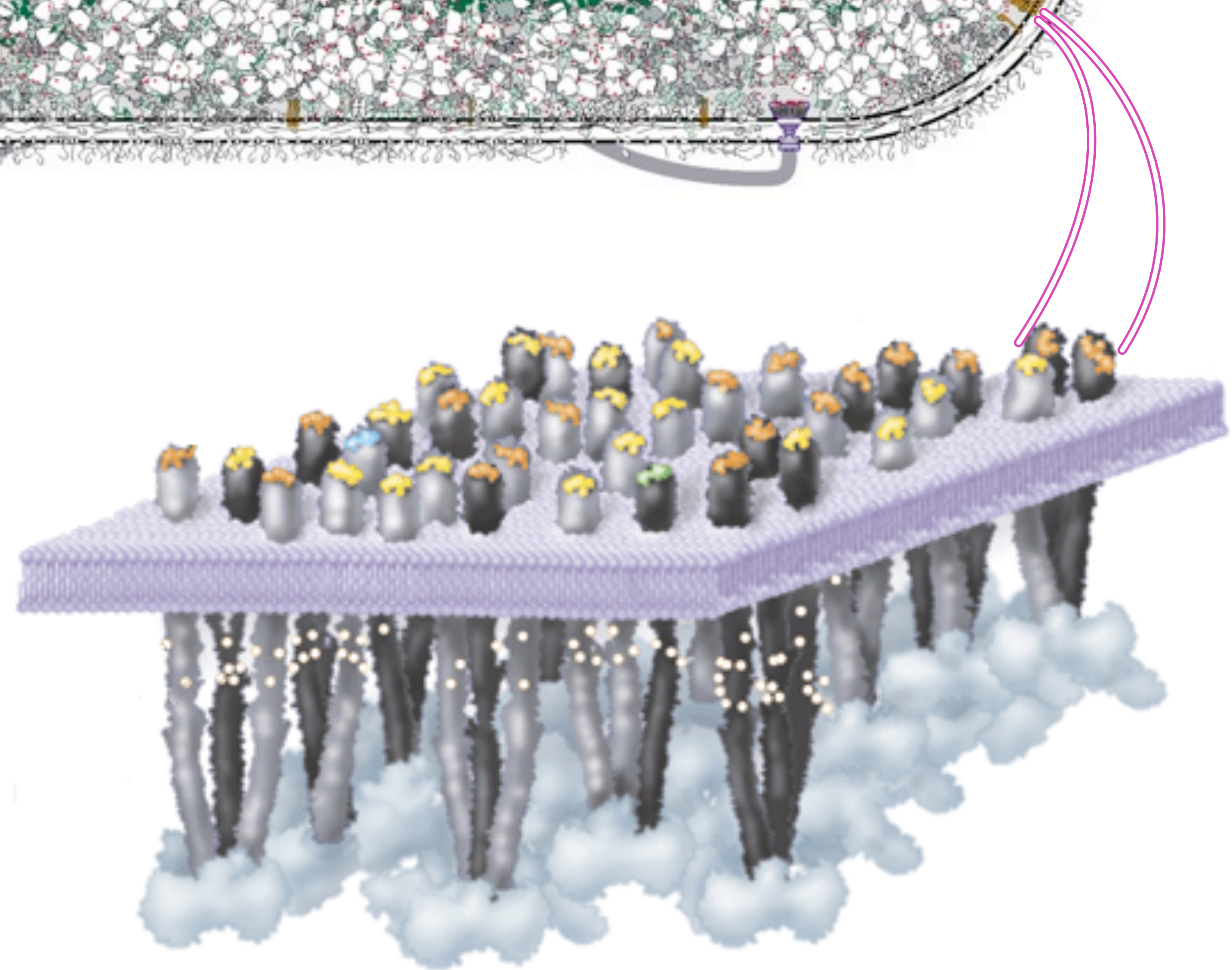
Machine Learning to Predict Protein–Protein Interactions

Sergei Grudinin, 24 Apr 2012

Bacterium

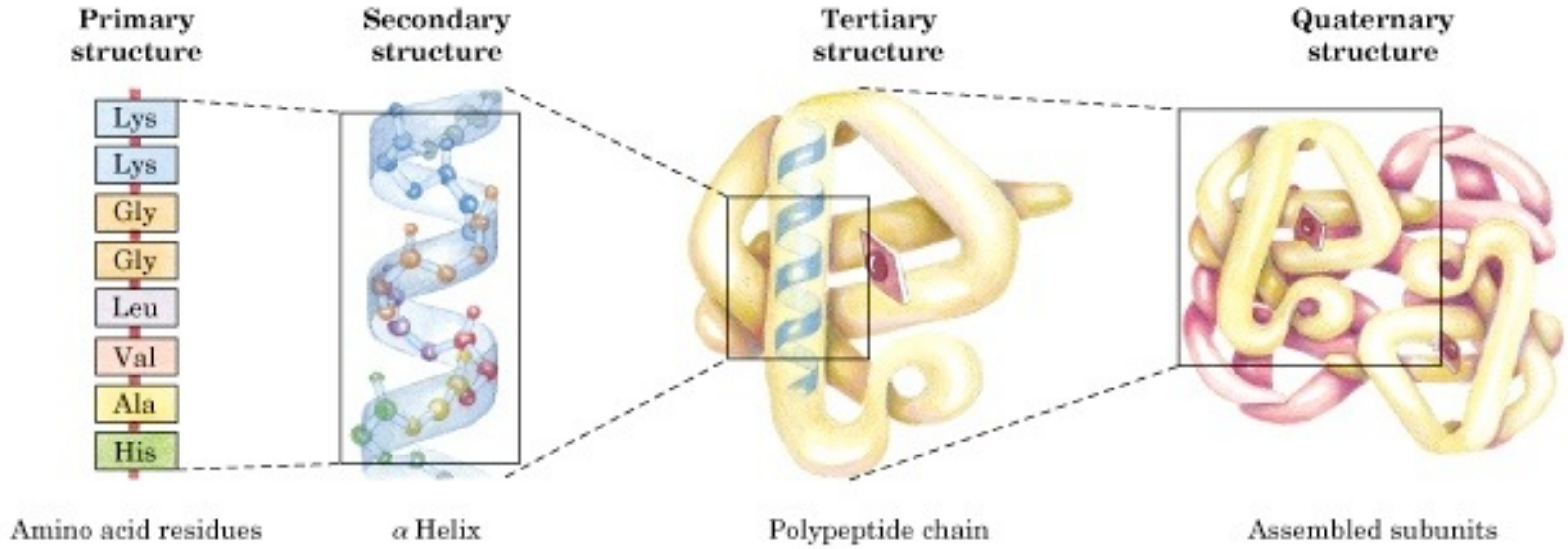


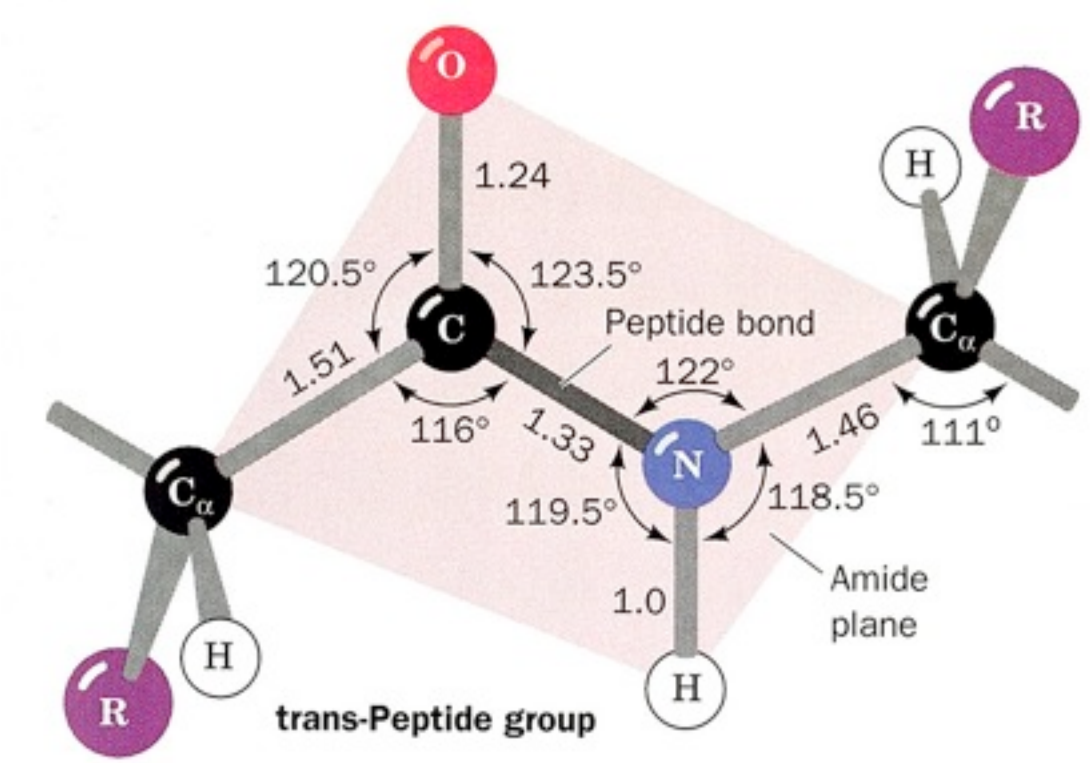
from http://www.pdn.cam.ac.uk/groups/comp-cell,
image is courtesy of Cambridge University

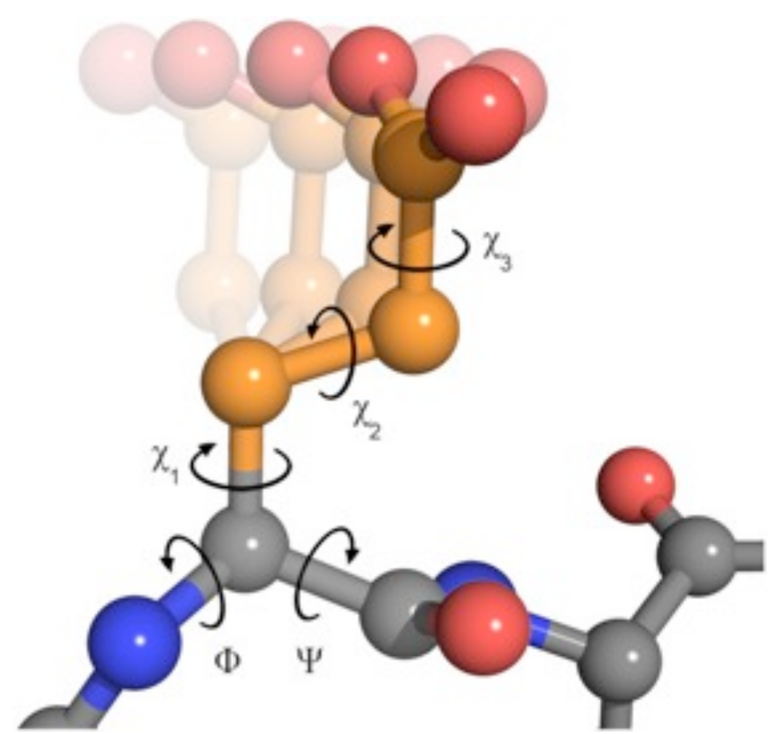


Model of a signalling array.
D. Bray. Science 2003, 299, 1189.

Protein Structure



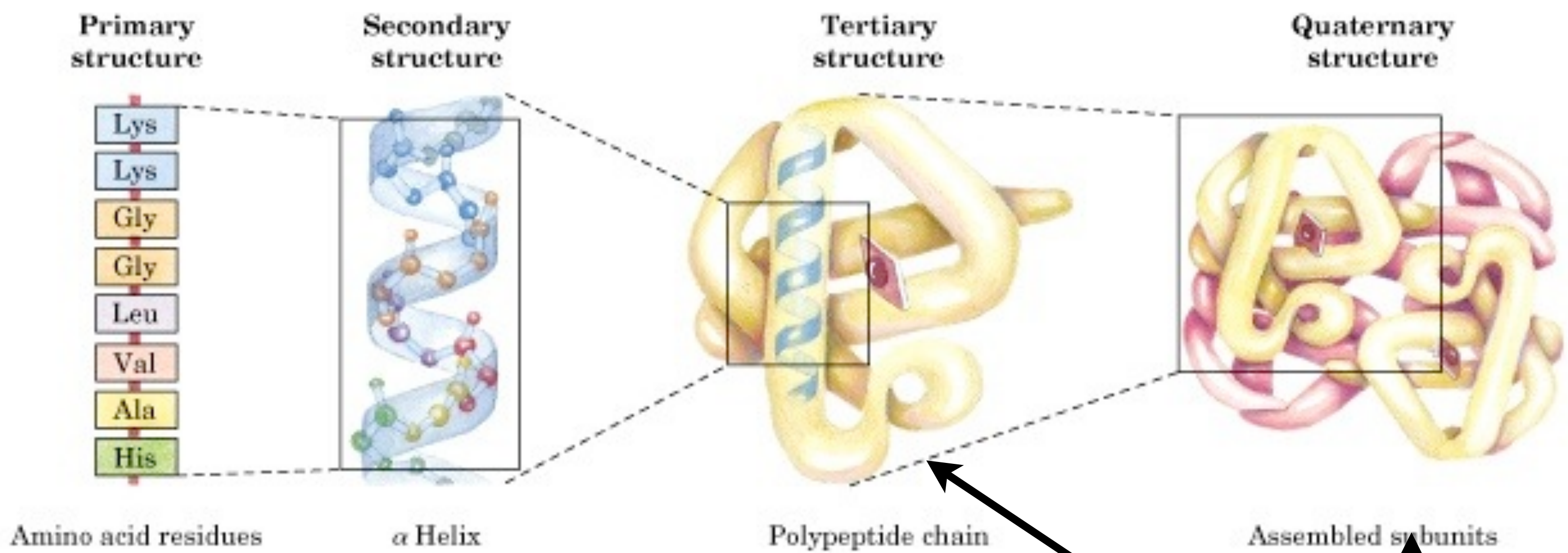




images are courtesy of Uppsala Universitet, Sweden, http://www.uu.se

MOTIVATION

Protein Structure

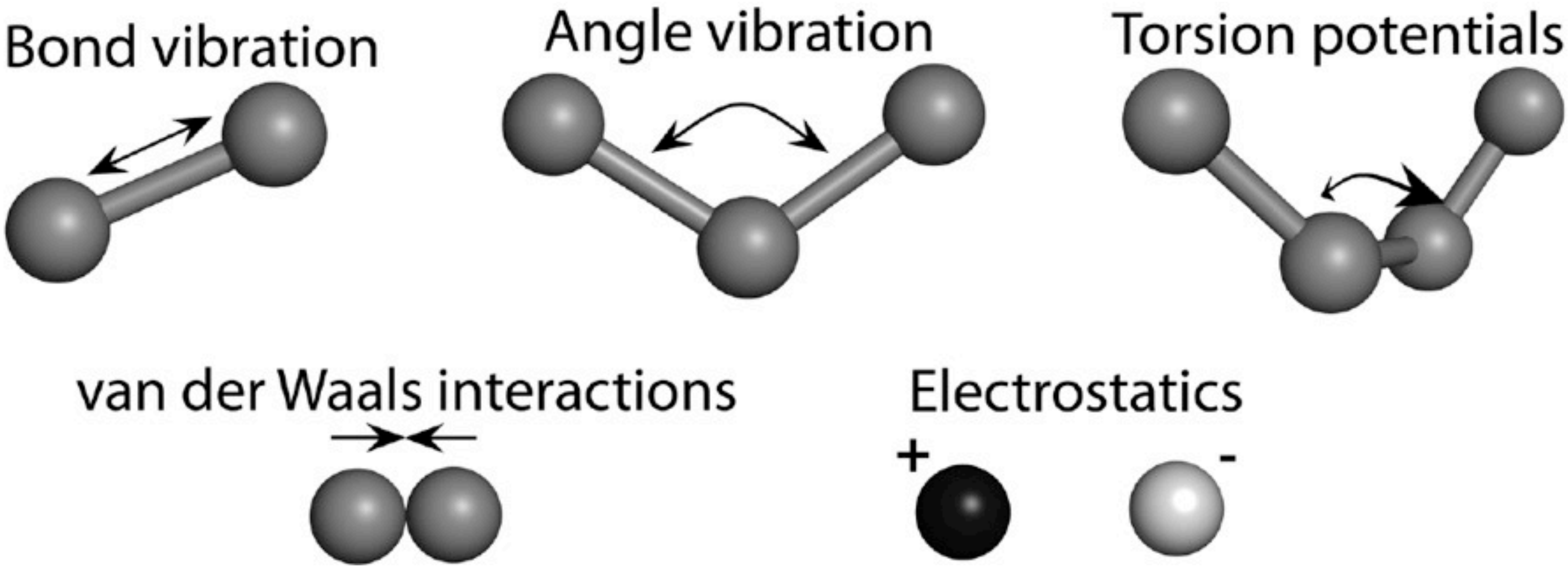


- Protein folding (prediction of protein **tertiary** structure)
  - works very well if there are homologous structures available
  - many web-servers available
  - CASP competitions
- Protein docking (prediction of protein **quaternary** structure)
  - currently much less mature
  - CAPRI competitions

images are courtesy of Uppsala Universitet, Sweden, http://www.uu.se

STANDARD APPROACH

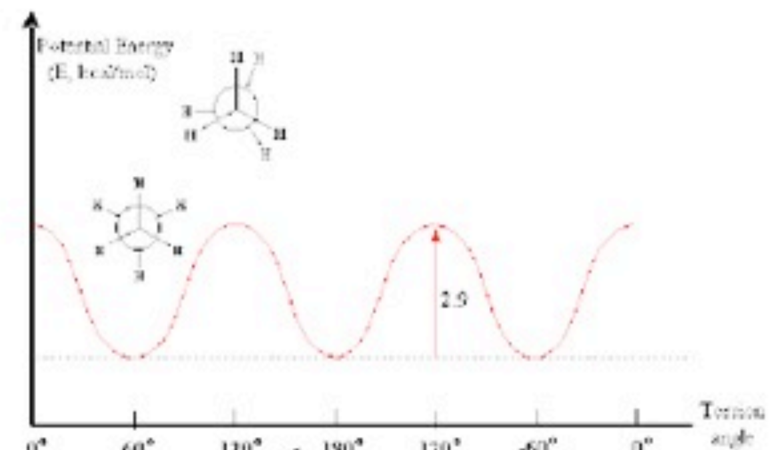
Types of Interactions

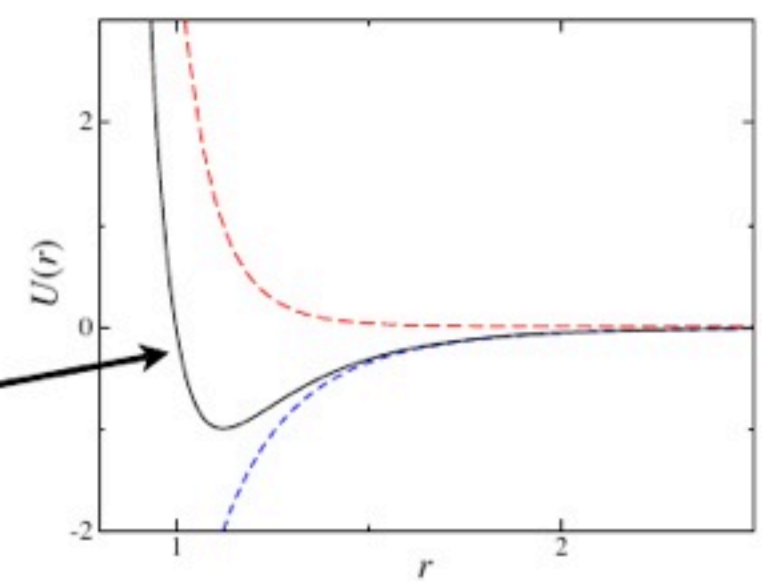


STANDARD APPROACH

Typical Shape of a Force Field

$$U = \sum_{\text{bonds}} \frac{1}{2} k^{b}_{ij} \left(r_{ij} - r^{0}_{ij}\right)^2 + \sum_{\text{angles}} \frac{1}{2} k^{\theta}_{ijk} \left(\theta_{ijk} - \theta^{0}_{ijk}\right)^2 + \sum_{\text{torsions}} \left(\sum_{n} k_{\theta} \left[1 + \cos(n\phi - \phi^{0})\right] \right) + \sum_{\text{impropers}} k_{\xi} \left(\xi_{ijkl} - \xi^{0}_{ijkl}\right)^2 + \sum_{i,j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \sum_{i,j} 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right]$$





Sampling: Bottom–Up approach

STANDARD APPROACH

Remember: all quantities we care about are averages in phase space:

$$\langle F \rangle = \int \mathrm{d}r^N \mathrm{d}p^N F(r^N, p^N) e^{-\beta \mathcal{H}(r^N, p^N)} / Z$$

average, phase space, microscopic value, probability

- Integral over momenta may be evaluated analytically
- The difficult problem is the computation of the average of $F(r^N)$
- Potential looks like

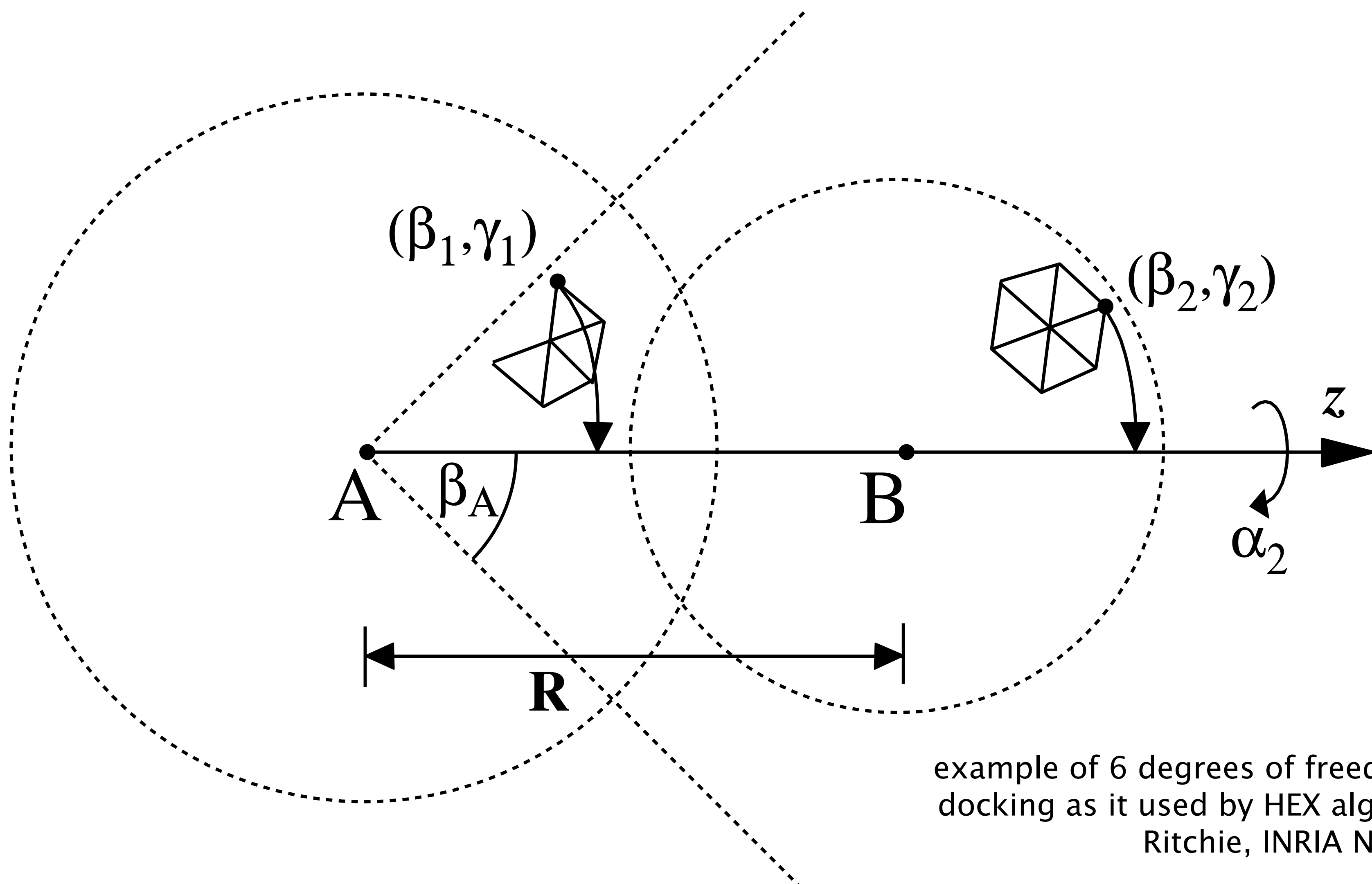
- Typically we use MD or MC

Sampling: Bottom–Up approach

- Binding free energy $W(r^6)$ can be then evaluated as:

$$\exp^{-\beta W(r^6)} \sim \int \mathrm{d}x^{N-6} \exp^{-\beta U(r^N)} / Z$$

- Then, the minimum value of $W(r^6)$ will correspond to the native complex



example of 6 degrees of freedom for rigid body docking as it used by HEX algorithm, from Dave Ritchie, INRIA Nancy

STANDARD APPROACH

Problems

SAMPLING

- Number of degrees of freedom (DOF) in a protein N ~ 10,000.
- We have to include solvation with DOF ~ 100,000.
- Long-range interactions, each atom feels each other atom.
- Extremely computationally expensive. Might take years on a supercomputer.

FF

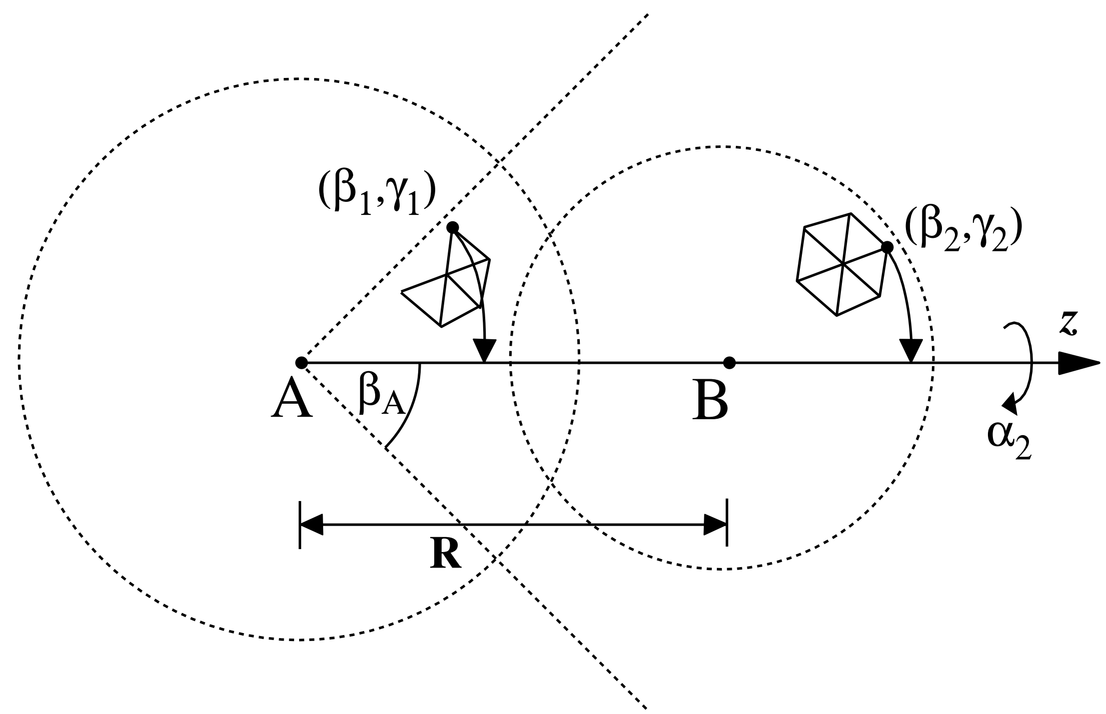
- Standard forcefields are very limited. They do not work for systems with polarization (ion channels) and in reactive centers.
- Forcefields errors accumulate in big systems.
- Forcefields exist only for a limited number of molecules.
- Small molecules must be parametrized separately.

Possible Solution: Rigid-Body Docking

Rigid–Body Docking

Find the minimum of potential function as fast as possible

$$E = \int \phi(\underline{r})\rho(\underline{r})\mathrm{d}V$$



For 2 proteins

$$\begin{aligned}\phi(\underline{r}) &= \phi_A(\underline{r}) + \phi_B(\underline{r}) \\ \rho(\underline{r}) &= \rho_A(\underline{r}) + \rho_B(\underline{r})\end{aligned}$$

Therefore,

$$E = \int (\phi_A(\underline{r})\rho_B(\underline{r}) + \phi_B(\underline{r})\rho_A(\underline{r}))\mathrm{d}V$$

- In the rigid body approximation we have 6 DOFs
- For middle–size proteins we need about 30 points in each direction
- Complexity will be $\sim 10^9$ of such integrals
- Modern algorithms simultaneously treat several such terms

from Dave Ritchie's presentations, INRIA Nancy, http://loria.fr/~ritchied

STANDARD APPROACH

FFT in Cartesian System

$$f_{A_{l,m,n}} = \begin{cases} 1 & : \quad \text{surface of molecule} \\ \rho & : \quad \text{core of molecule} \\ 0 & : \quad \text{open space} \end{cases}$$

$$f_{B_{l,m,n}} = \begin{cases} 1 & : \quad \text{inside molecule} \\ 0 & : \quad \text{open space} \end{cases}$$

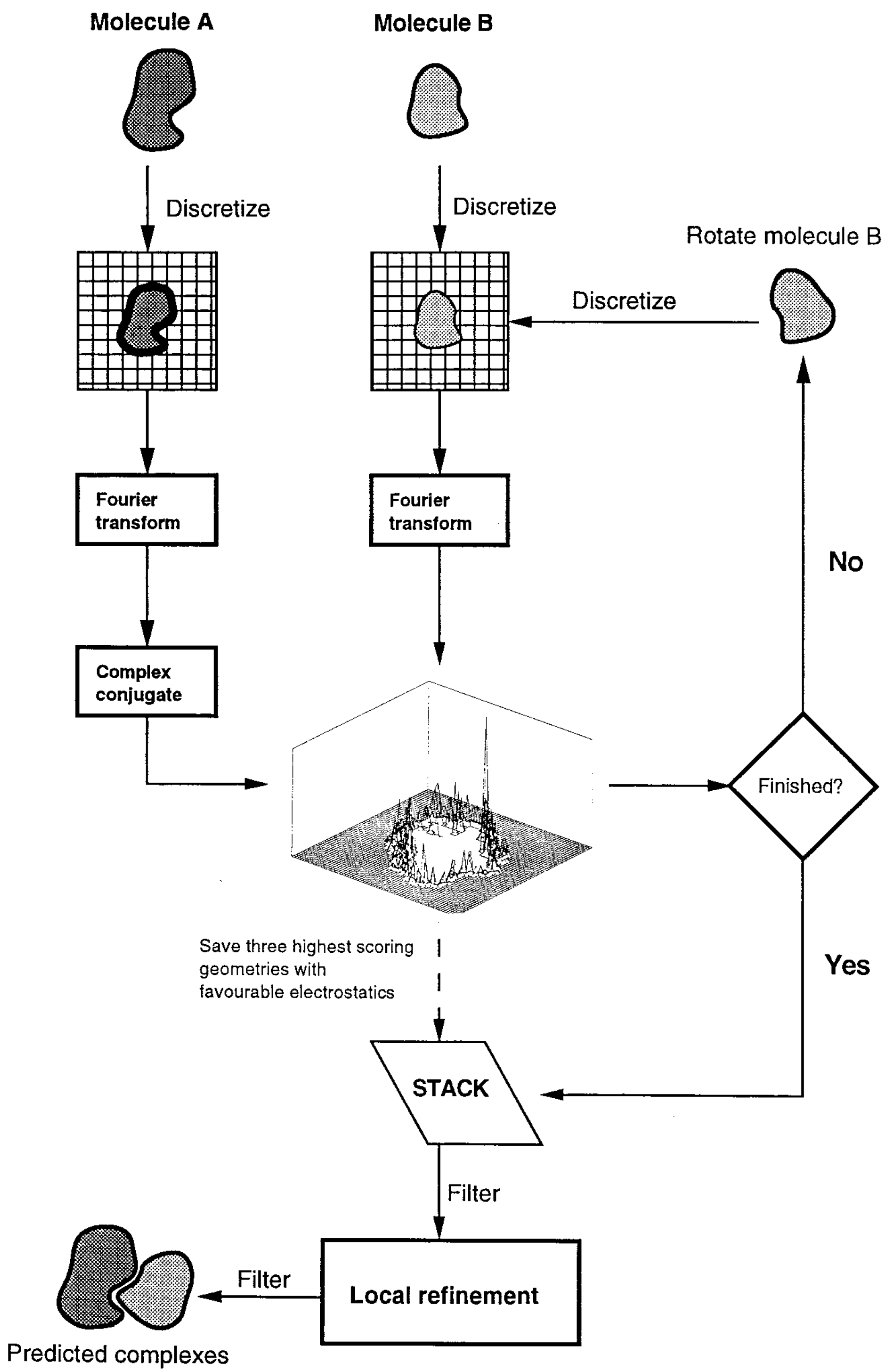
$$f_{C_{\alpha,\beta,\gamma}} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} f_{A_{l,m,n}} \times f_{B_{l+\alpha,m+\beta,n+\gamma}}$$

α, β, γ - shift vectors of A relative to B

N - number of points in each direction

$$F_A = \mathrm{DFT}(f_A)$$
$$F_B = \mathrm{DFT}(f_B)$$
$$F_C = (F_A^*)(F_B)$$
$$f_C = \mathrm{IFT}(F_C)$$

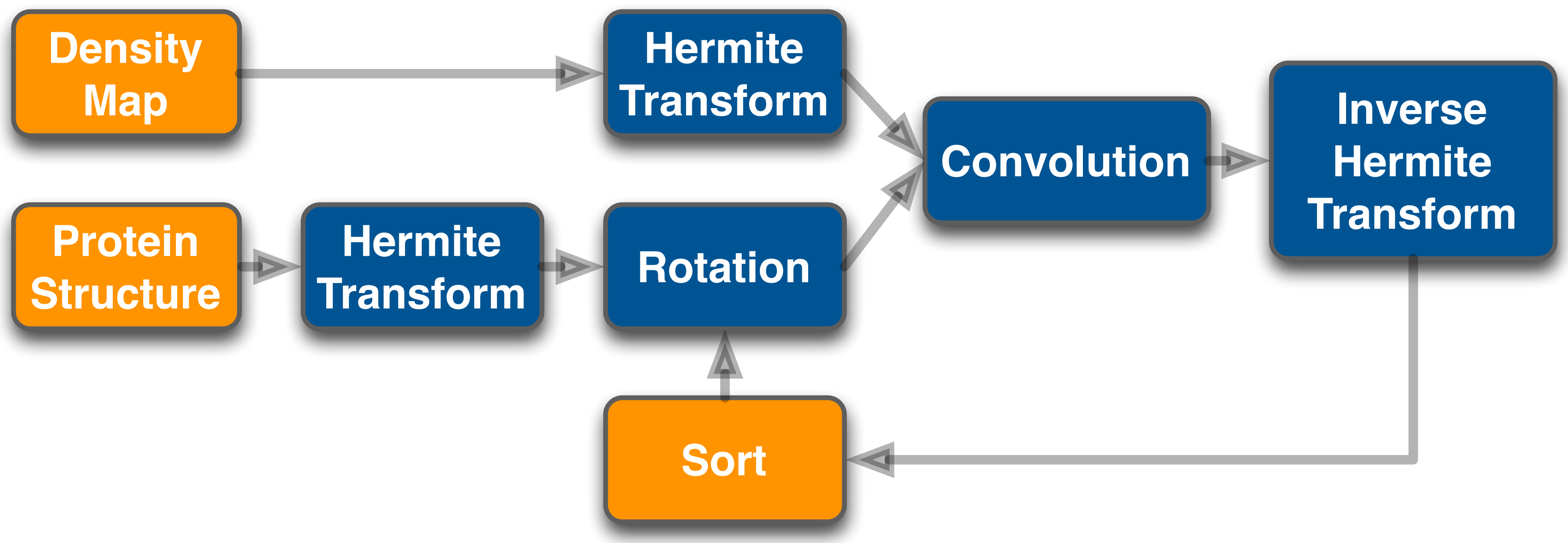
- for each orientation of B we need $O(N^6)$ computations of correlation using the direct method
- or $O(N^3 \log N^3)$ using FFT



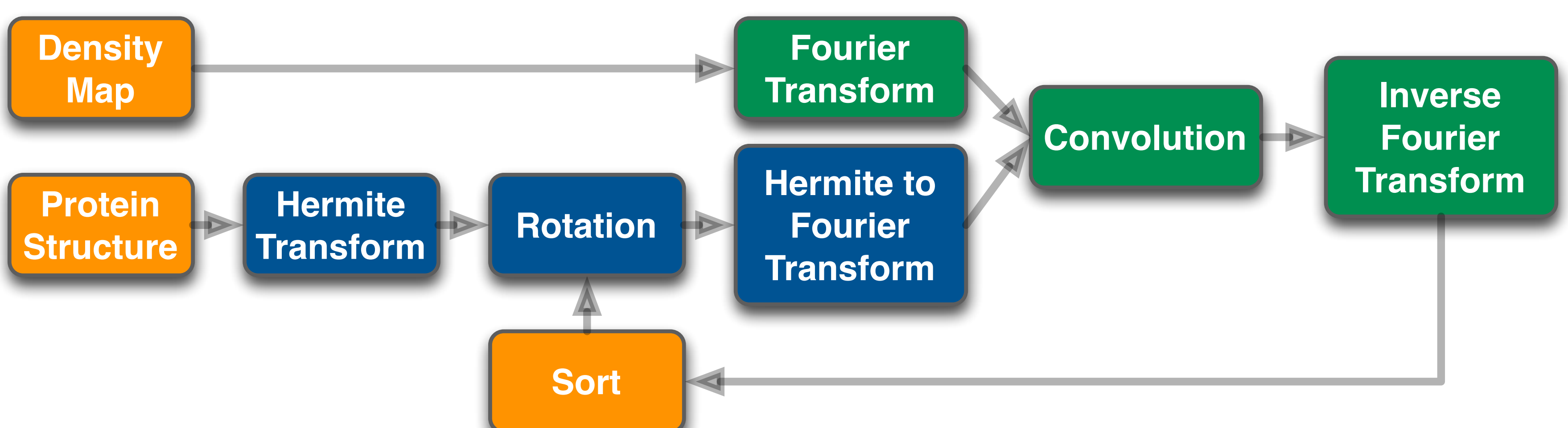
from Henry A. Gabb, Richard M. Jackson and Michael J. E. Sternberg, J. Mol. Biol. (1997) 272, 106-120

Our Approach

- Hermite Space



- Hermite - Fourier Space



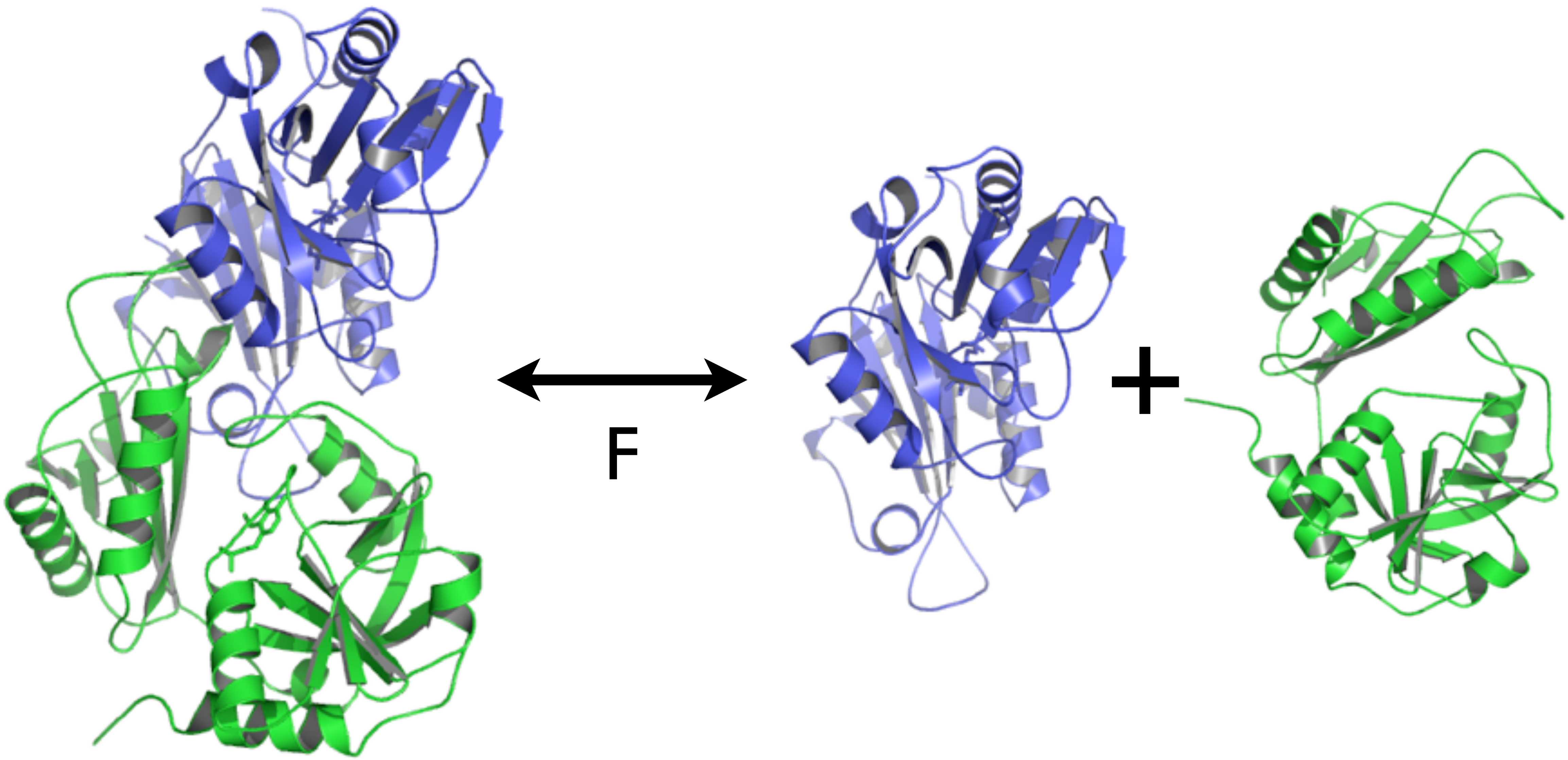
Problems

- Potential function is too simple and in many cases unrealistic.
- 6 DOFs are obviously not sufficient.
- We often start predictions with protein structures in their bound conformations. However, upon binding they adopt different, "unbound" states.

Knowledge-Based Protein Docking: Top-Down Approach

METHOD

Protein Docking



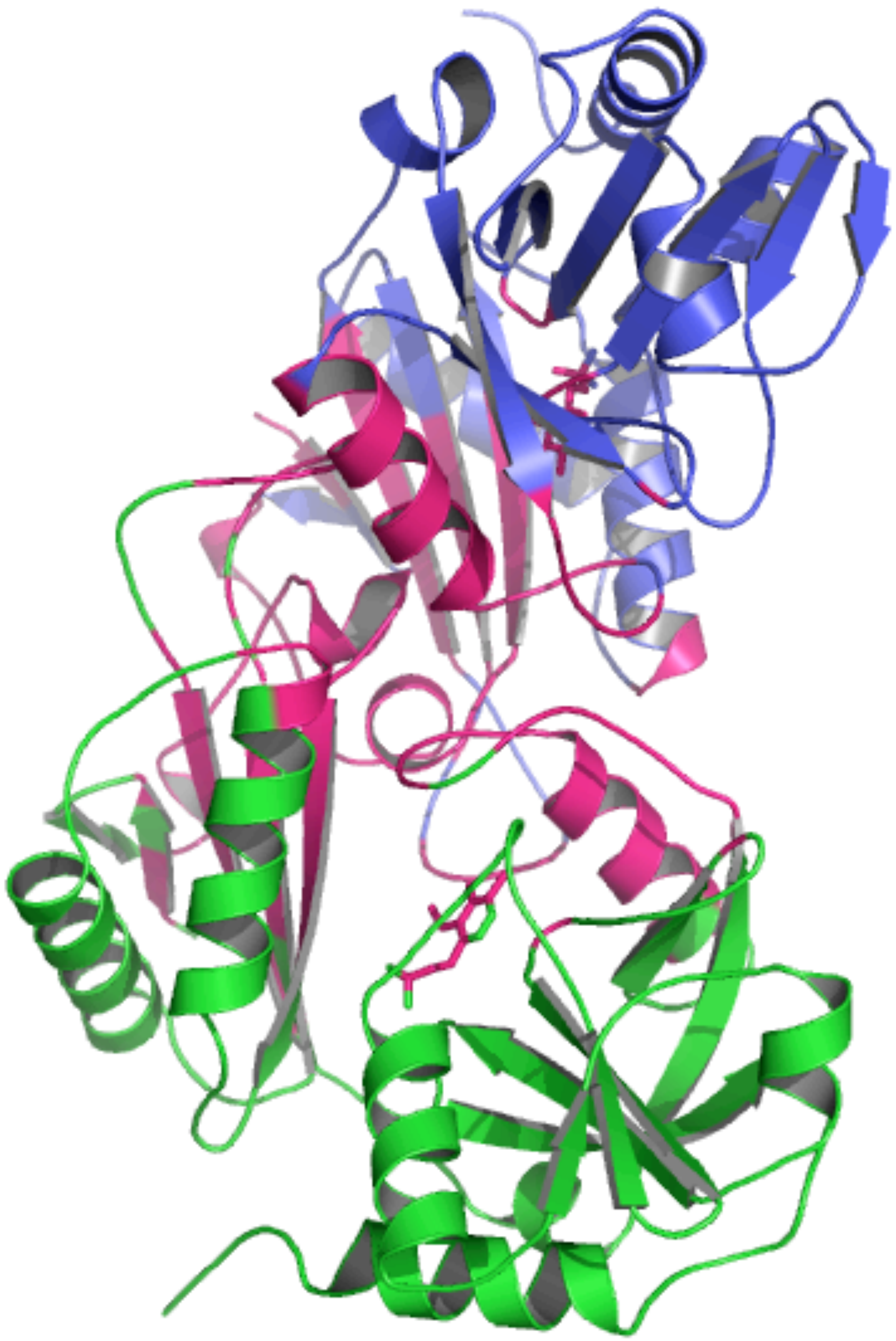
How to find ***Binding Free Energy*** of a protein complex?

- have to make several assumptions

METHOD

Assumptions I: Interface

- ***Binding energy*** depends only on the interface between the proteins within a certain *cutoff distance*



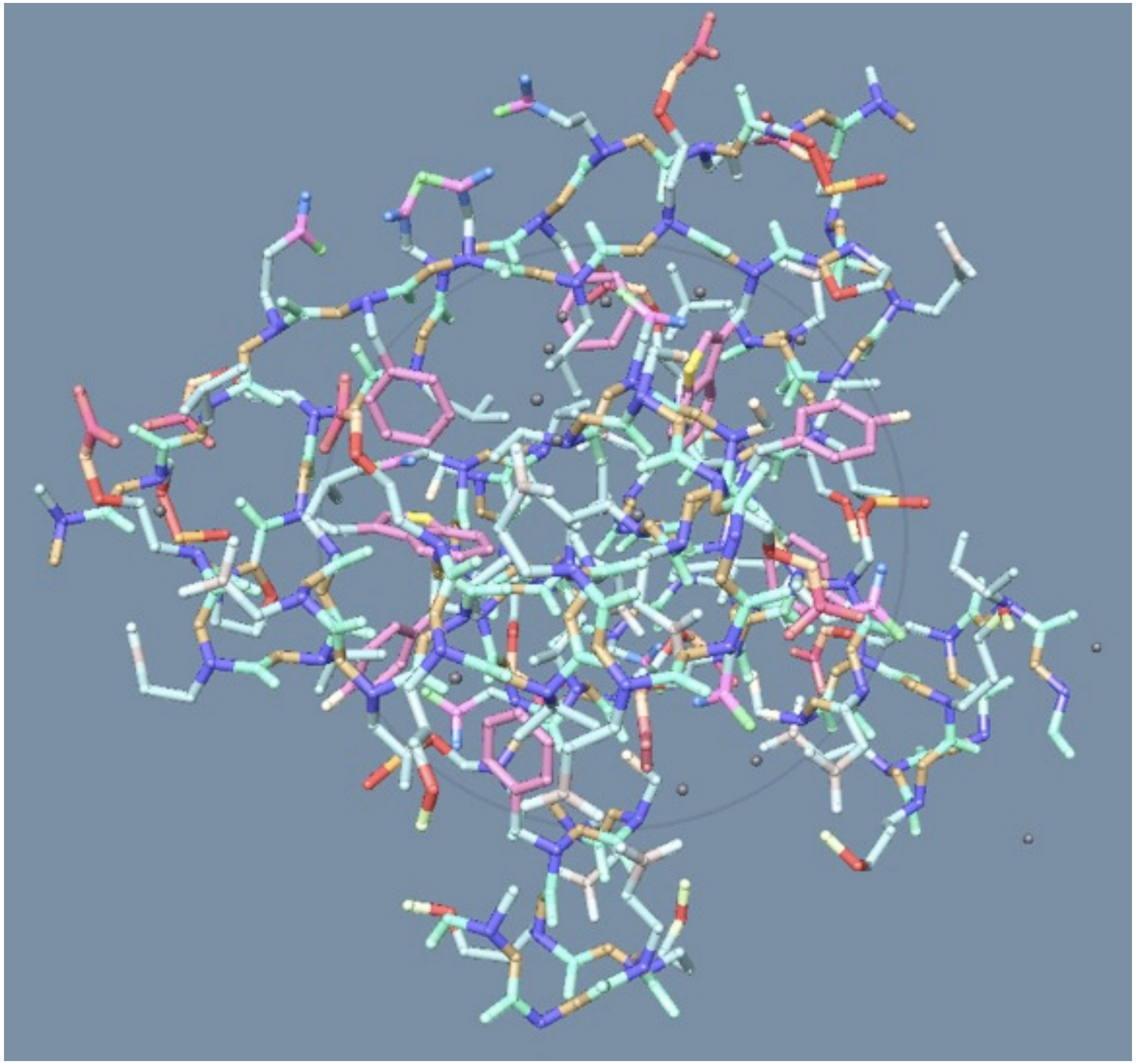
METHOD

Assumptions II: Atom Types

- Protein molecule is represented by a set of M ***discrete interaction sites*** that are located at the sites of the atomic nuclei

- Protein Folding - individual types for all atoms

- Protein Docking - a set of types, about 20



METHOD

Assumptions III:

$$F(n(r)) \equiv F(n_{11}(r), .., n_{kl}(r), .., n_{MM}(r)) = \sum_{k=1}^{M} \sum_{l=k}^{M} \int_{0}^{r_{max}} n_{kl}(r) U_{kl}(r) \ dr$$

- ***Binding energy*** F depends only on the distributions $n_{kl}(r)$ of distances between the interaction sites (the number of site pairs at a certain distance)
- ***Binding energy*** F is a linear functional

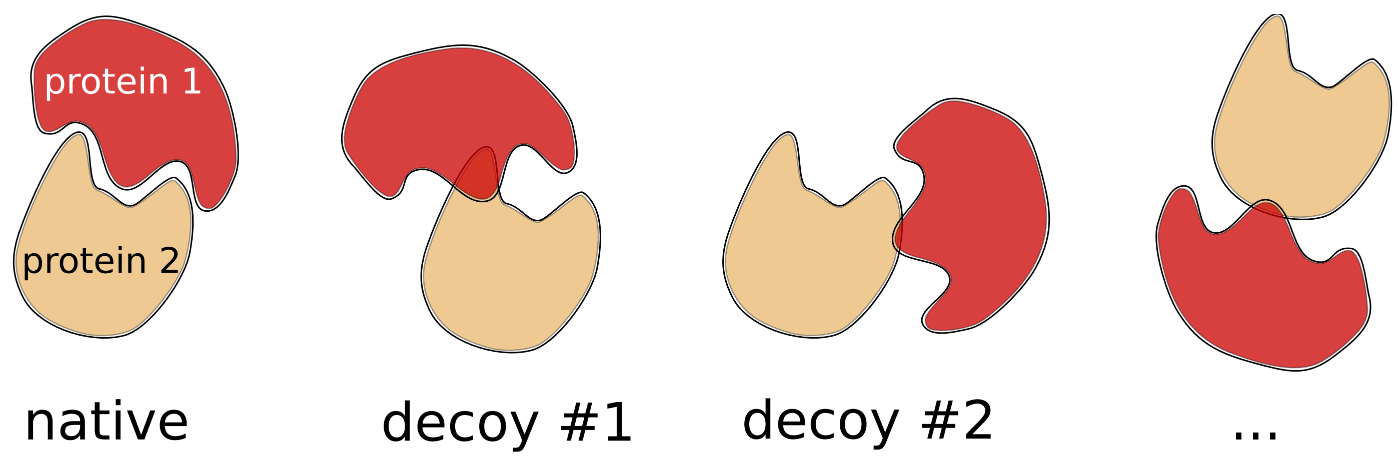
Given a set of $n_{kl}(r)$ and constants $U_{kl}(r)$ we can find the binding free energy $F(n(r))$!

METHOD

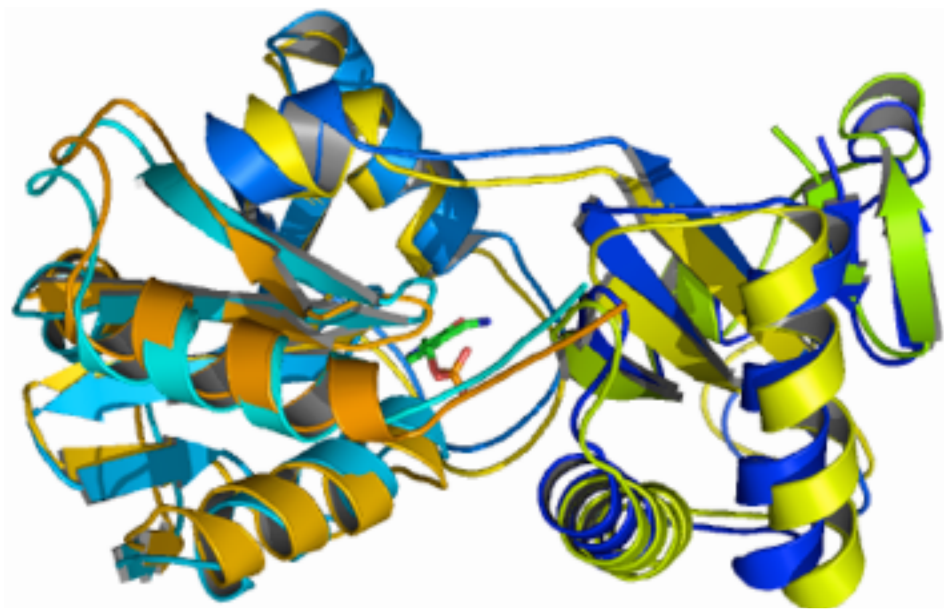
Knowledge-base

- Native: 850 non-homologues complexes from PDB

- Non-native: generated by rolling one over another



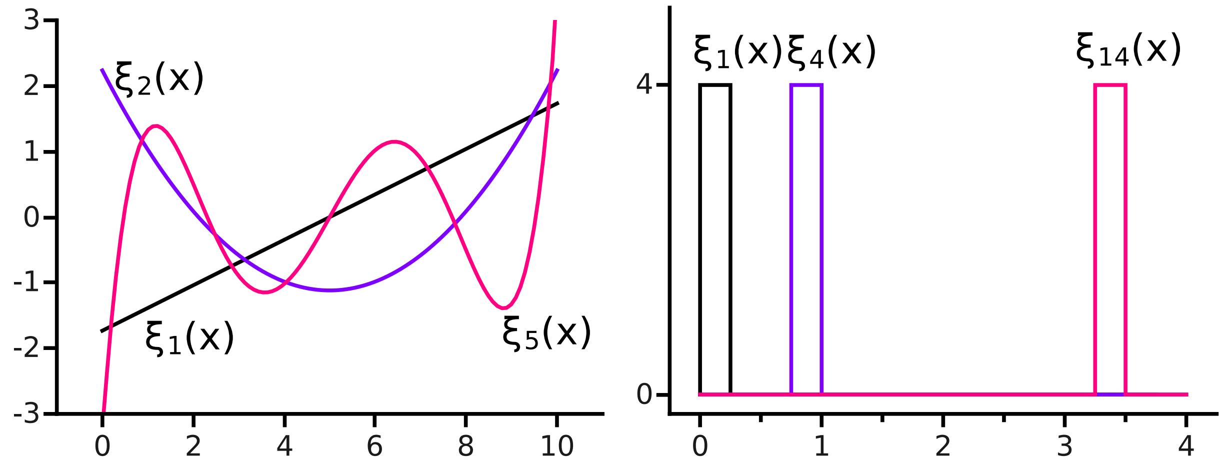
- Non-native: generated using NMA



METHOD

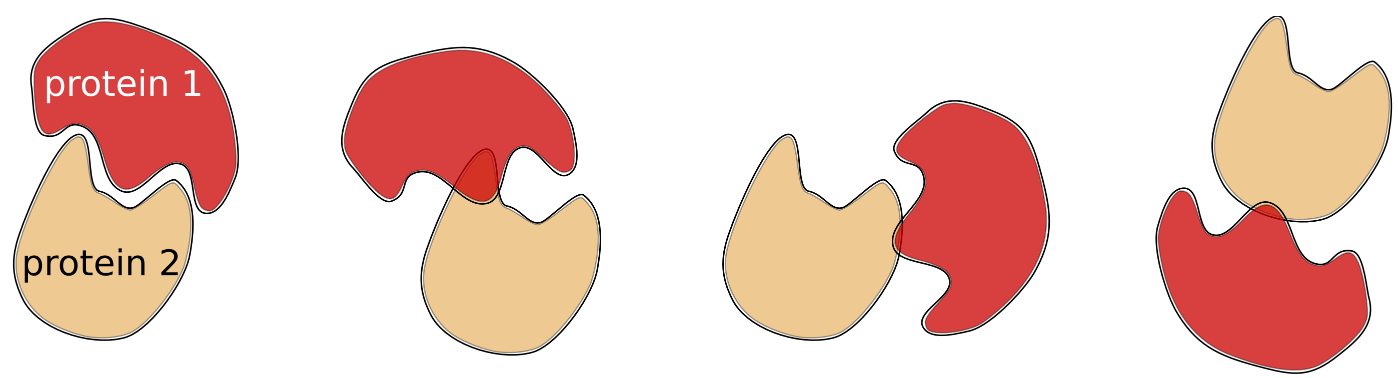
How do we compute $U_{kl}(r)$?

- Expand $U_{kl}(r)$ and $n_{kl}(r)$ in an orthogonal basis



Legendre   Rectangular

- Compute distance distributions $n_{kl}(r)$ for native and nonnative structures



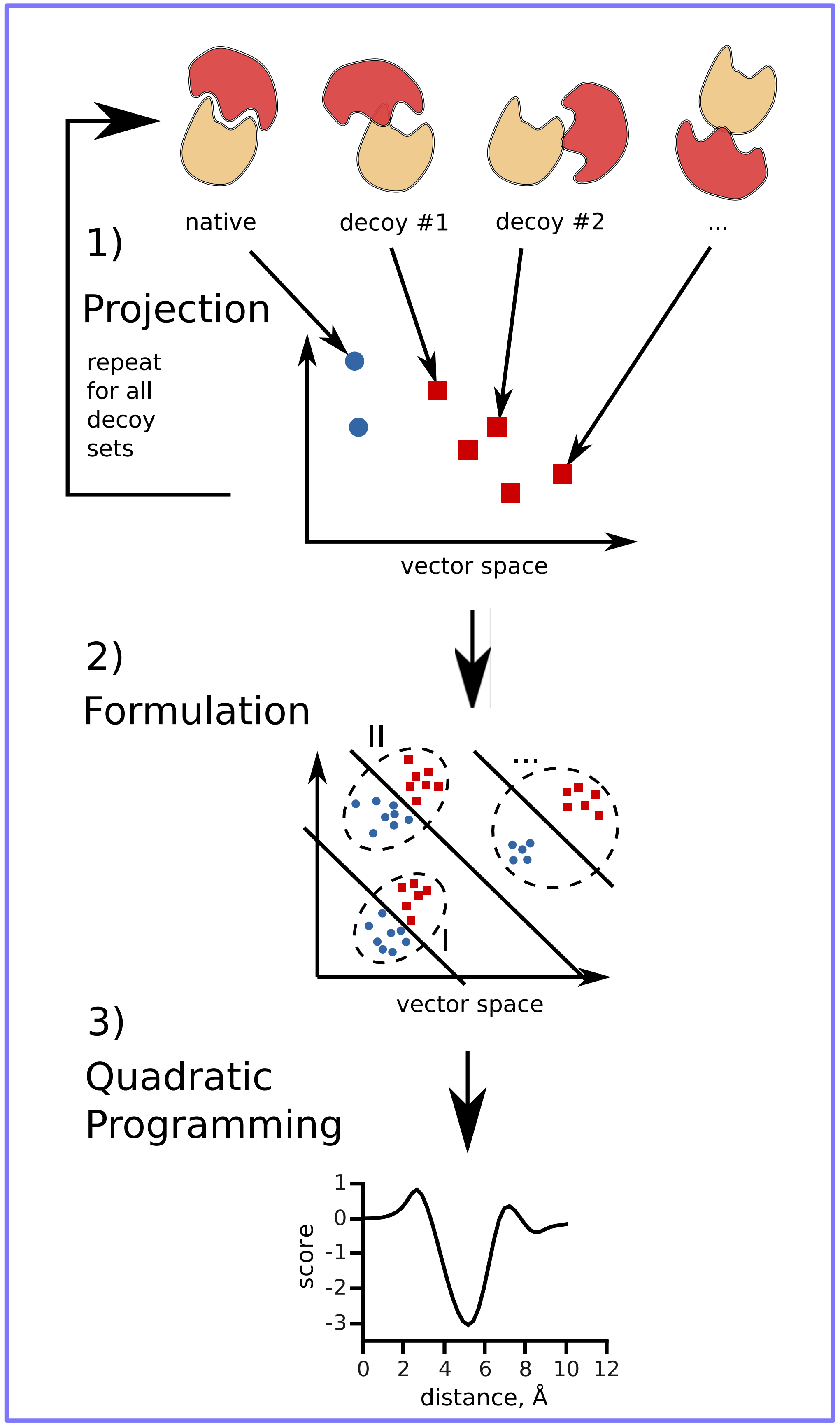
native   decoy #1   decoy #2   ...

- Find energy expansion coefficients $\mathbf{w}$ by solving convex quadratic problem with about 10^5 - 10^6 linear constraints

$$
\begin{aligned}
&minimize: && \frac{\mathbf{w}\cdot\mathbf{w}}{2} + \sum_{i=0}^{m} C_{ij}\xi_{ij} \\
&subject\ \ to: && y_{ij}\left[\mathbf{w}\cdot\mathbf{x}_{ij} + b\right] - 1 + \xi_{ij} \geq 0 \\
& && \xi_{ij} \geq 0
\end{aligned}
$$

METHOD

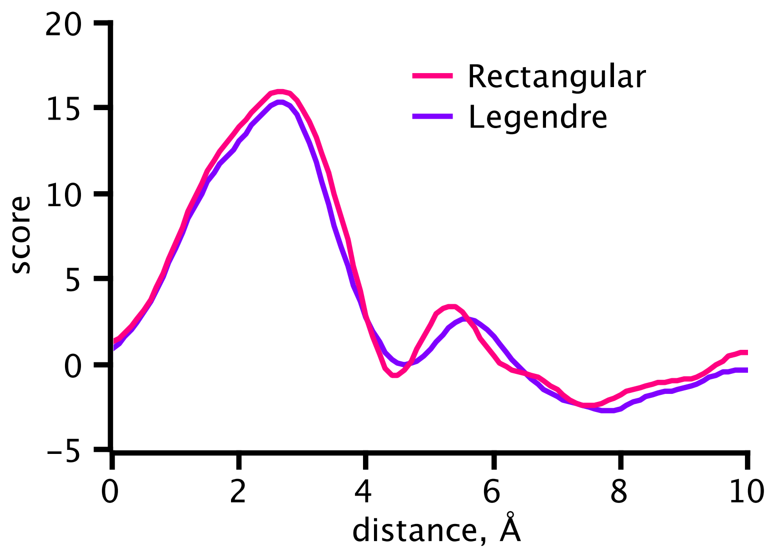
Algorithm



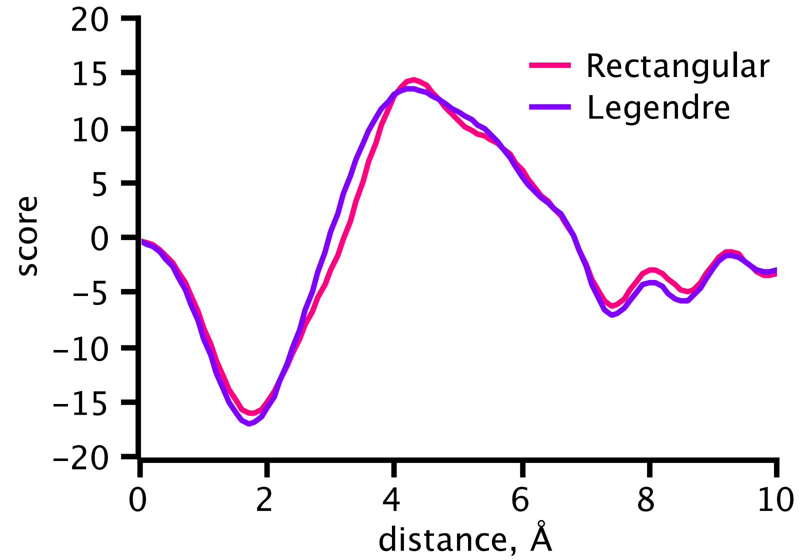
RESULTS

Potentials $U_{kl}(r)$

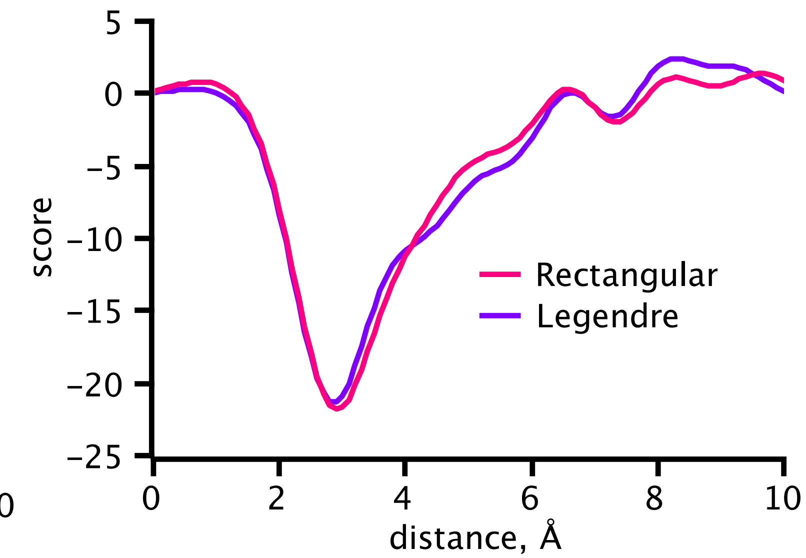
aliphatic carbons – C_a carbons



amide nitrogens – oxygens

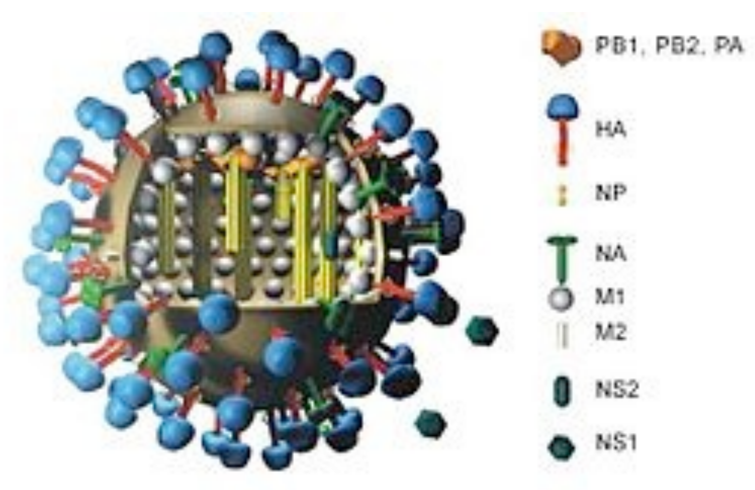


N+ - O-

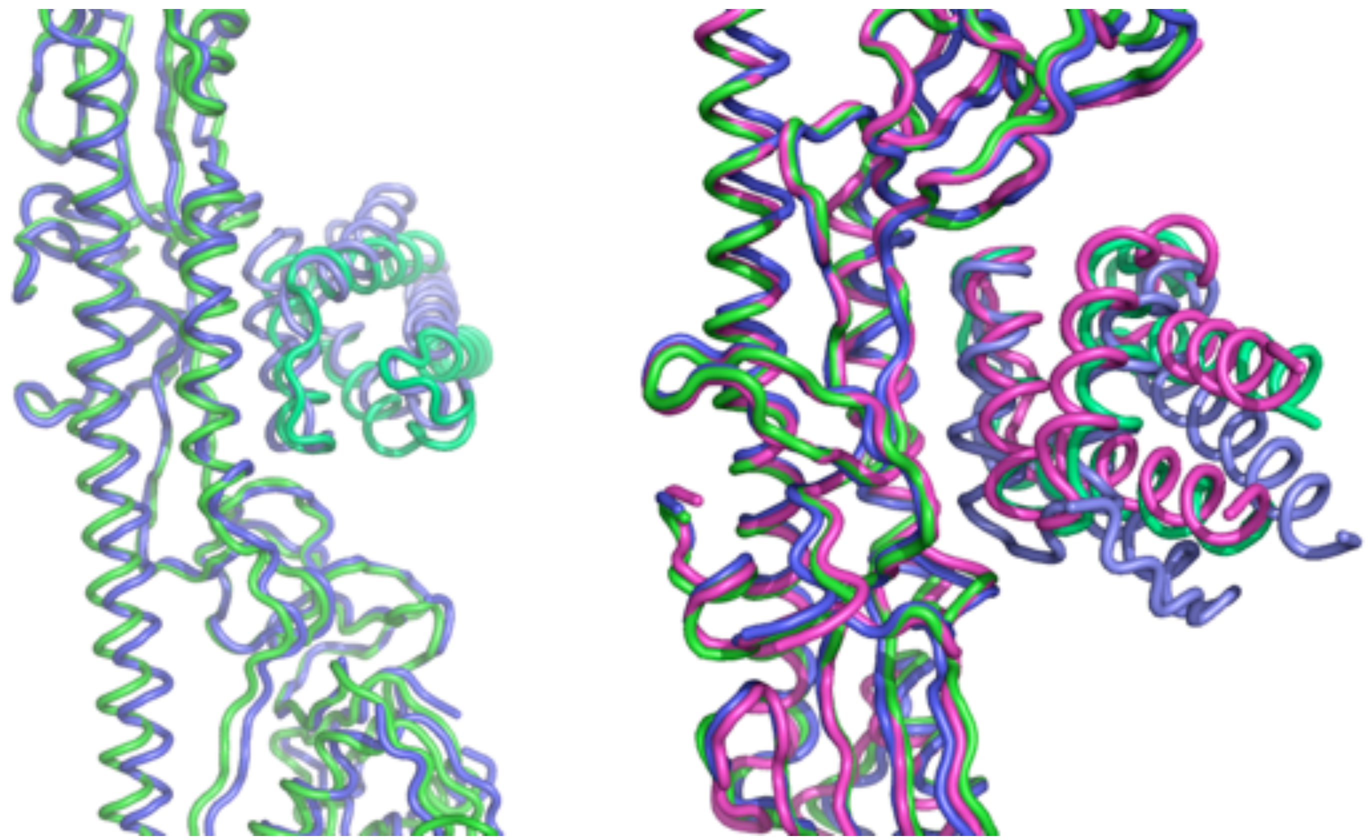


RESULTS

CAPRI Blind Predictions



Influenza virus with hemagglutinin protein trimers (HA) on the surface of the viral capsid



Prediction of the complex of HA with the designed protein HB36

X-Ray Us Baker's group

RESULTS

Validation

- Training set of 850 complexes is predicted with 100% accuracy
- Top 1 predictions on Standard Benchmarks (1000 complexes of different qualities, contact side chains rebuilt)
  - Rosetta Unbound 83%
  - Rosetta Bound 89%

Rosetta Unbound Benchmark

PDB	Quality	Rank	PDB	Quality	Rank
1A0O	3	1	1MAH	2	1
1ACB	2	1	1MDA	2	2
1AHW	3	1	1MEL	2	1
1ATN	1	1	1MLC	3	1
1AVW	2	1	1NCA	1	1
1AVZ	-	>10	1NMB	1	1
1BQL	3	1	1PPE	1	1
1BRC	2	1	1QFU	1	1
1BRS	3	1	1SPB	1	1
1BTH	3	1	1STF	1	1
1BVK	3	1	1TAB	2	1
1CGI	3	3	1TGS	3	1
1CHO	1	1	1UDI	2	1
1CSE	2	1	1UGH	2	1
1DFJ	2	1	1WEJ	3	2
1DQJ	3	1	1WQ1	2	1
1EFU	-	>10	2BTF	1	1
1EO8	3	1	2JEL	2	1
1FBI	3	1	2KAI	3	7
1FIN	-	>10	2PCC	3	1
1FQ1	3	4	2PTC	2	1
1FSS	2	1	2SIC	1	1
1GLA	2	1	2SNI	2	1
1GOT	3	1	2TEC	1	1
1IAI	2	1	2VIR	2	1
1IGC	3	1	3HHR	3	1
1JHL	3	4	4HTC	2	1
Top1	ITScore **59.3%**		RosettaDock **66.7%**	Us **83.3%**	

CAPRI Assessment, 2010–2012

http://web.mit.edu/sheny/capri.html

Rank	Group	T46	T47 (Water-mediated interactions)	T48	T48 (Trimer)	T49	T49 (Trimer)	T50	T51.1	T51.2	T51.3	T52 (Not assessed)	T53	T54	Summary: #Targets / *** + ** + *
1	Bonvin	*	**		*		*	**	*				**		7 / 3 ** + 4 *
2	Shen		*	**	**	**	**	*					**	*	6 / 3 ** + 3 *
3	Bates		**	*		*		*		*			*		6 / 1 ** + 5 *
4	Vajda		**		**		*	**					***		5 / 1 *** + 3 ** + 1 *
5	Eisenstein		**		**	*	*	**					*		5 / 3 ** + 2 *
6	Fernandez-Recio		*		*		*	**					**		5 / 2 ** + 3 *
6	Zacharias		***		*		*	*					*		5 / 1 *** + 4 *
8	Vakser		**	*	*	*	*	*						*	5 / 1 ** + 4 *
9	ClusPro				**		*	**					**		4 / 3 ** + 1 *
9	Zou		***	**	*	*	*	*							4 / 1 *** + 1 ** + 2 *
11	Nakamura		***						*				*	*	4 / 1 *** + 3 *
12	Weng		*			*	*	*					**		4 / 1 ** + 3 *
13	**Grudinin**	**–**	******	**–**	**–**	**–**	**–**	******					*****		**3 / 2 ** + 1 ***
14	HADDOCK	*	**				*								3 / 1 ** + 2 *
14	PIE/DOCK				*		*	**							3 / 1 ** + 2 *
14	Wolfson		*	*	**	*	*								3 / 1 ** + 2 *
17	Zhou		*	*	*		*								3 / 3 *
18	Seok		**										**		2 / 2 **
19	Elber				*			**							2 / 1 ** + 1 *
19	Fernandez-Fuentes							**					*		2 / 1 ** + 1 *
19	Gray		**										*		2 / 1 ** + 1 *
22	SwarmDock												*	*	2 / 2 *
23	Camacho							**							1 / 1 **
23	Cui			*	**										1 / 1 **
23	LZerD												**		1 / 1 **
23	Ritchie		**												1 / 1 **
23	Ten Eyck												**		1 / 1 **
23	Wang		**												1 / 1 **
29	Luethy							*							1 / 1 *
29	Pal							*							1 / 1 *
29	Poupon												*		1 / 1 *
29	SurFit												*		1 / 1 *
29	Zhang												*		1 / 1 *
34	About 24 Others														0 / 0 *

Problems

- Protein flexibility must be taken into account

⟶ Collective motions with Normal Modes

- Sidechain flexibility must be taken into account

⟶ Rotamers optimization

Predicting Positions of Water Around a Protein

METHOD

Assumptions:

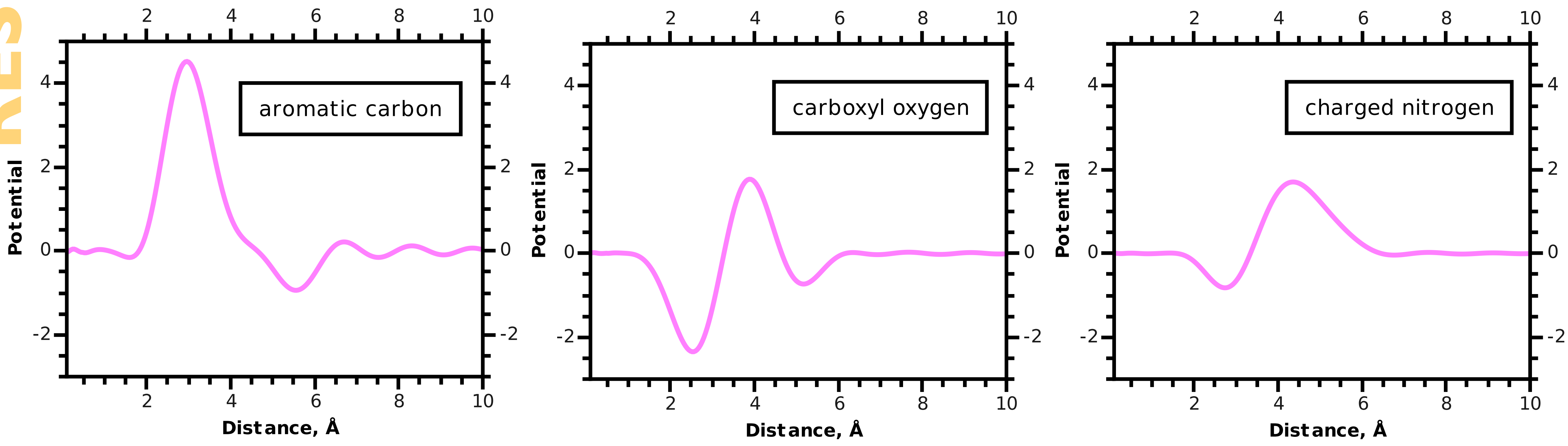
$$F(n(r)) \equiv F(n^1(r), .., n^M(r)) = \sum_{k=1}^{M} \int_0^{r_{max}} n^k(r) U^k(r)\, dr,$$

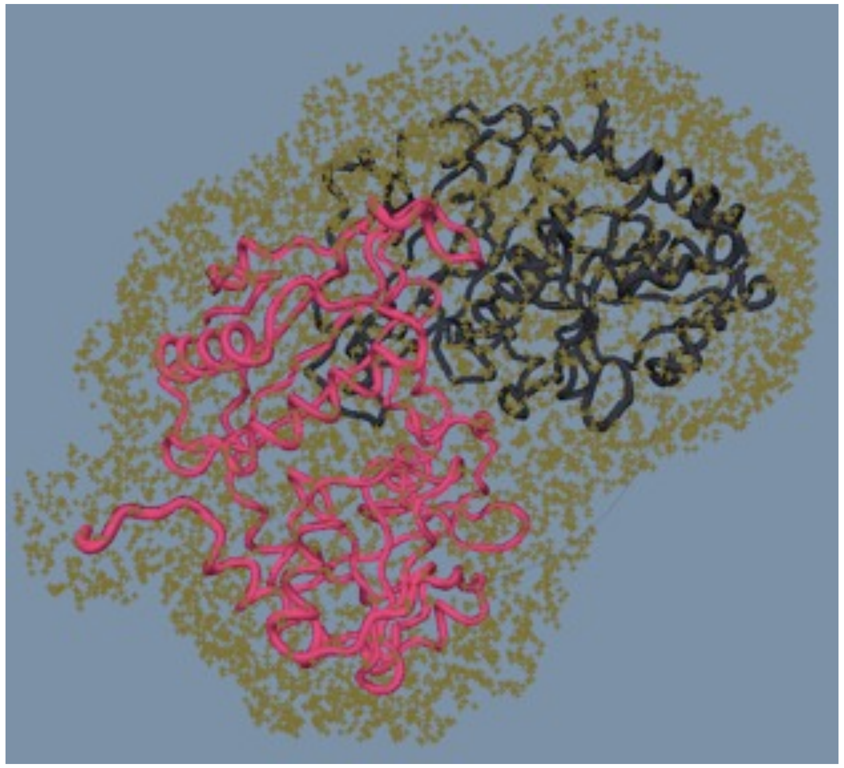
- ***Solvation free energy*** F depends only on the n^k distributions of distances between the interaction sites (the number of site pairs at a certain distance)
- ***Solvation free energy*** F is a linear functional

Given a set of $n^k(r)$ and constants $U^k(r)$ we can find the solvation free energy $F(n(r))$!

RESULTS

Potentials $U^k(r)$





On The Way

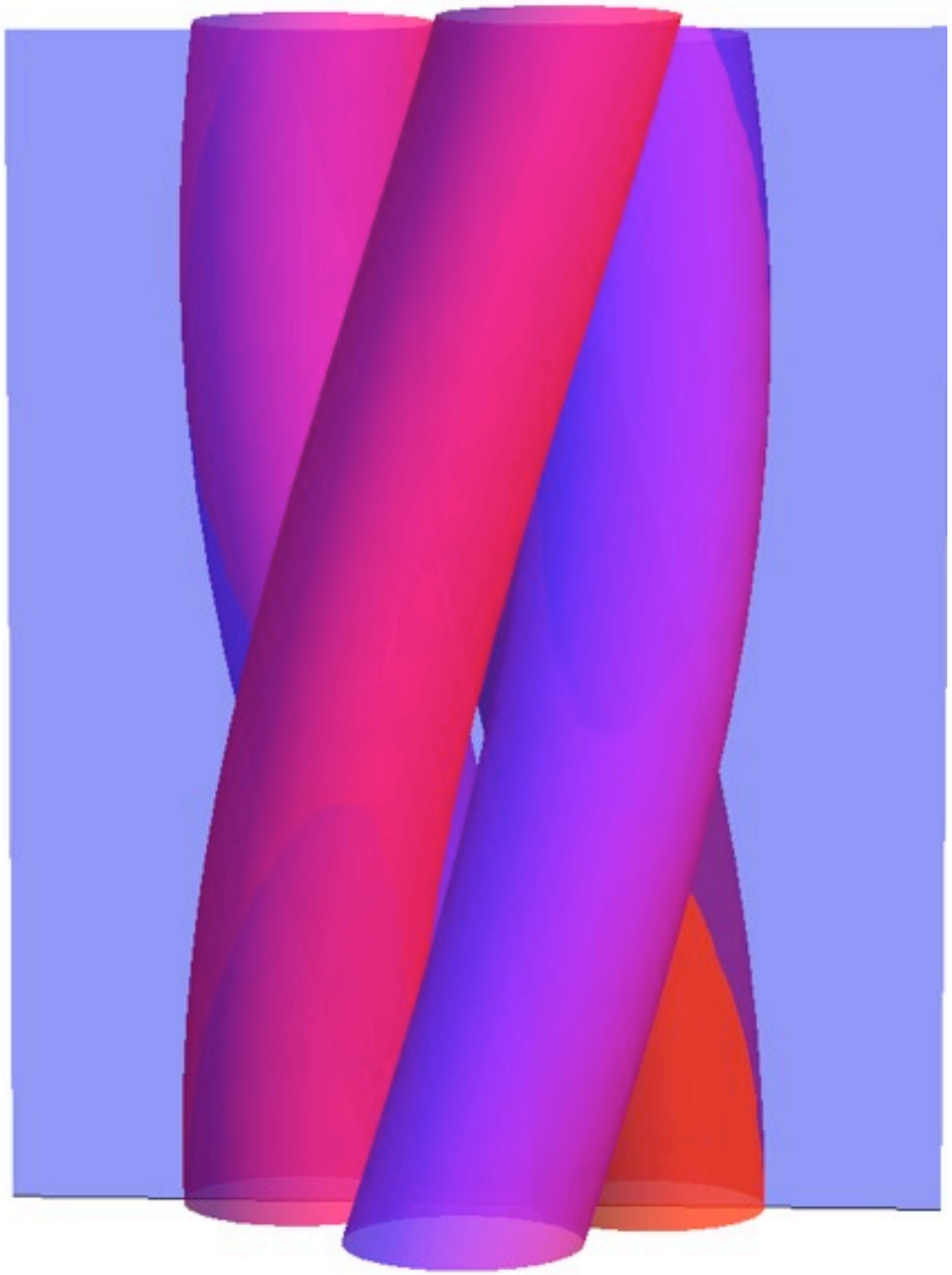
Minimization with a KB-Potential

Set without native structures	Top1 (q = 1,2,3)	Top10 (q = 1,2)	Top1Q1*	Top10Q1*
Before minimization	422 (52.88%)	502 (62.90%)	351 (77.31%)	417 (91.85%)
After minimization	652 (81.70%)	679 (85.09%)	611 (95.17%)	639 (99.53%)

Set without native and near-native structures	Top1 (q = 1,2,3)	Top10 (q = 1,2)	Top1Q1*	Top10Q1*
Before minimization	248 (31.07%)	311 (38.97%)	171 (76.00%)	204 (90.67%)
After minimization	563 (70.55%)	593 (74.31%)	504 (95.64%)	525 (99.62%)

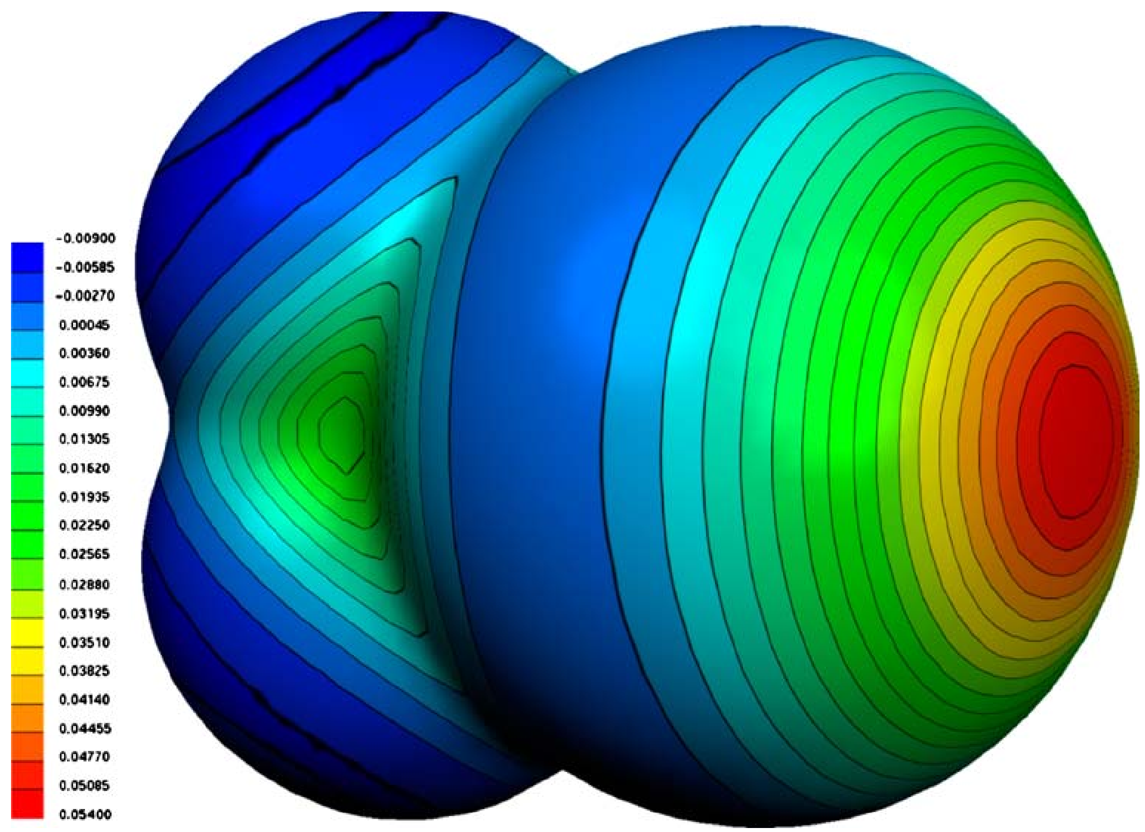
Flexibility

- Combination with internal-angle mechanics
- CG orientation-dependent potential
- QP optimization



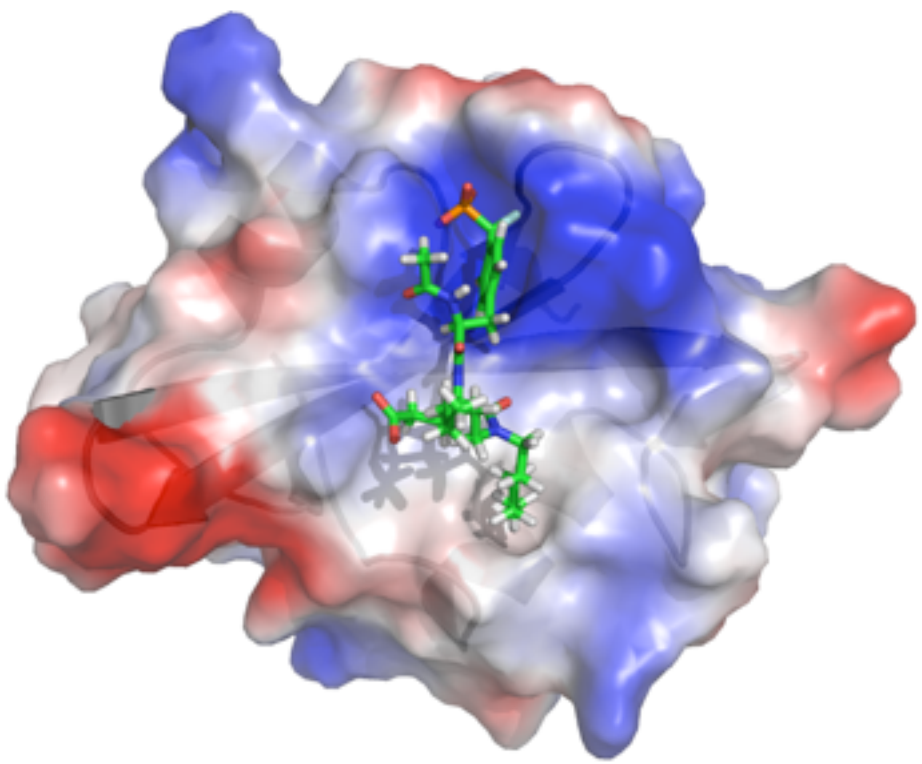
Angular-dependent KB-potential

- Halogen-bonds
- Hydrogen-bonds
- Aromatic interactions
- Accepted for the INRIA International Internship Program



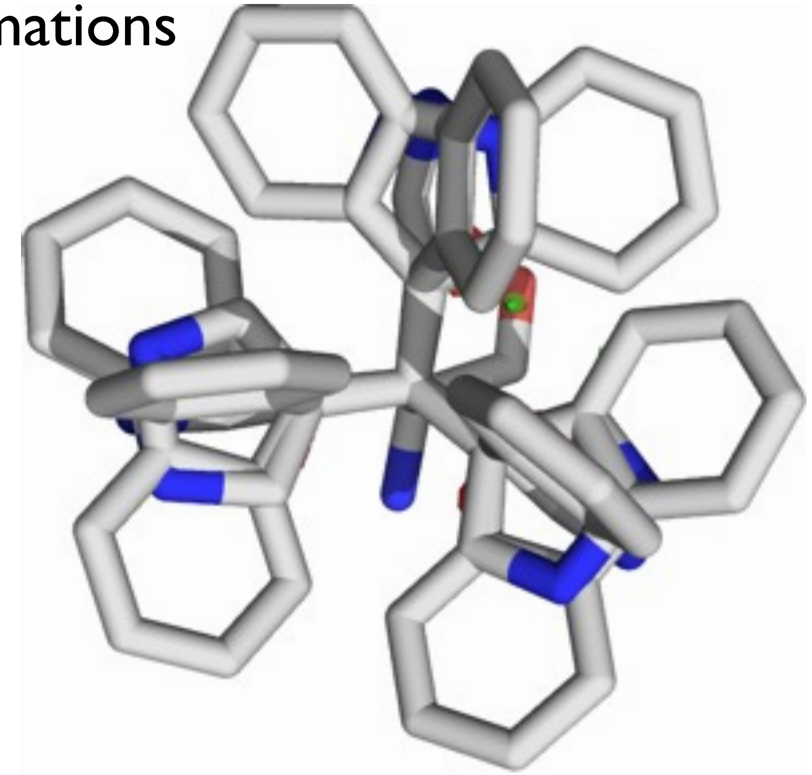
Protein-Ligand Interactions

- Pairwise-additive KB function
- ~ 50 atom types
- QP optimization
- Bachelor project of a MIPT student
- Currently tested



Open Problems

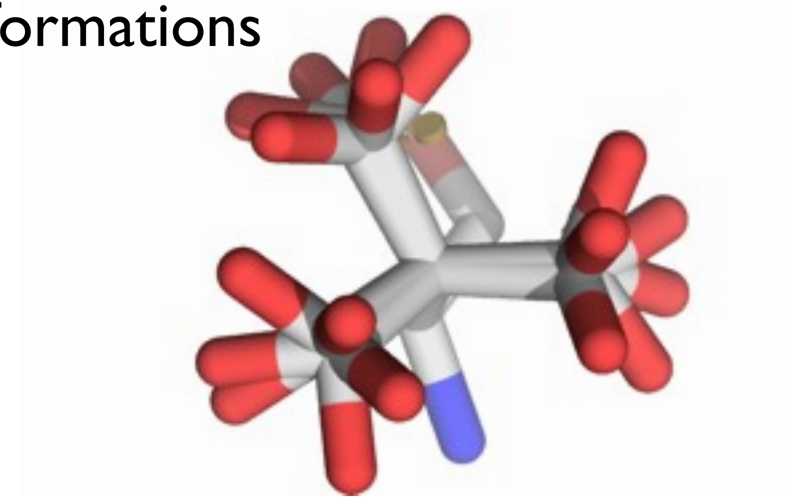
A, 6 conformations



- We have A with N_1 conformations and B with N_2 conformations
- All states of A and all states of B are accessible
- Then, partition function is given by

$$Z = \sum_k \exp^{-\mathbf{x_k} \circ \mathbf{w}}$$

B, 9 conformations



- And Helmholtz free energy is

$$F = -\log Z = -\log \sum_k \exp^{-\mathbf{x_k} \circ \mathbf{w}}$$

- So, optimization problem is:

$$\log \sum_k \exp^{-\mathbf{x_k} \circ \mathbf{w}} > \log \sum_k \exp^{-\mathbf{x'_k} \circ \mathbf{w}}$$