# DOCKING PARADIGM IN COMPUTER-AIDED DRUG DISCOVERY

Vladimir Sulimov


DIMONTA

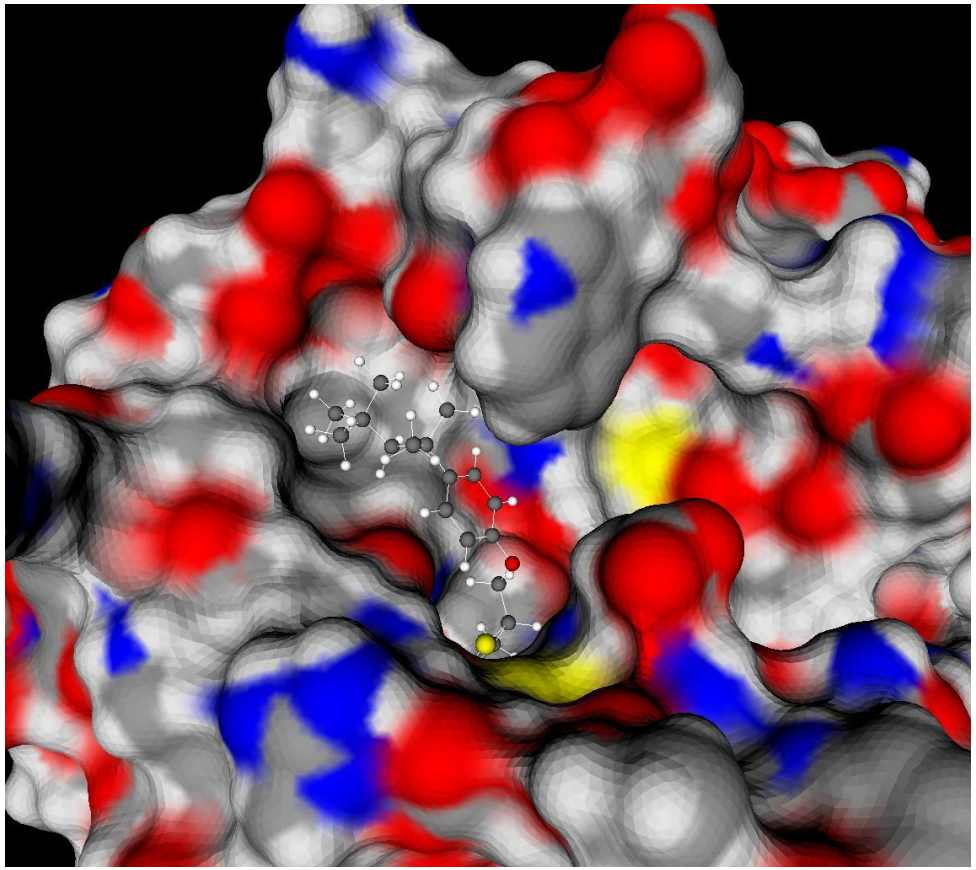

The work is financially supported by the Russian Science Foundation, Agreement no. 21-71-20031

# Docking is a popular software used for the drug development

- **Docking:**
  - Ligand positioning in the target protein
  - Computing the protein-ligand binding energy $\Delta G_{bind}$
- Is it possible to increase docking accuracy?
  - Positioning accuracy – ***satisfactory***
  - Accuracy of the calculations of the protein-ligand binding energy $\Delta G_{bind}$ – ***bad***

**Docking accuracy** → **Drug discovery efficiency**

# Docking paradigm: the ligand binds in the active site of the target protein in close proximity of the global energy minimum of the protein-ligand complex



**Docking is the search for the global minimum of the energy of the protein-ligand complex**

**Reviews:**

- **Sulimov V.B., et al. // Curr. Top. Med. Chem., 2021**
- **Sulimov V.B., et al. // Curr. Med. Chem., 2019**
- **Sulimov A.V. , et al. // Supercomput. Front. Innov., 2019**

# More recent reviews on docking

- D. Bassani et al. *Re-Exploring the ability of common docking programs to correctly reproduce the binding modes of non-covalent inhibitors of SARS-CoV-2 protease Mpro*, Pharmaceuticals, 2022, **15**, 180; comparison of three docking programs, GOLD, Glide, and PLANTS, to reproduce crystallographic positions of known inhibitors of Mpro
- S. Zev et al. *Benchmarking the Ability of Common Docking Programs to Correctly Reproduce and Score Binding Modes in SARS-CoV-2 Protease Mpro*, J. Chem. Inf. Model. 2021, **61**, 2957-2966; comparison of 6 docking programs: Glide, DOCK, AutoDock, AutoDock Vina, FRED, and EnzyDock
- N. Murugan et al. *A Review on Parallel Virtual Screening Softwares for High-Performance Computers*, Pharmaceuticals, 2022, **15**, 63

# Docking paradigm

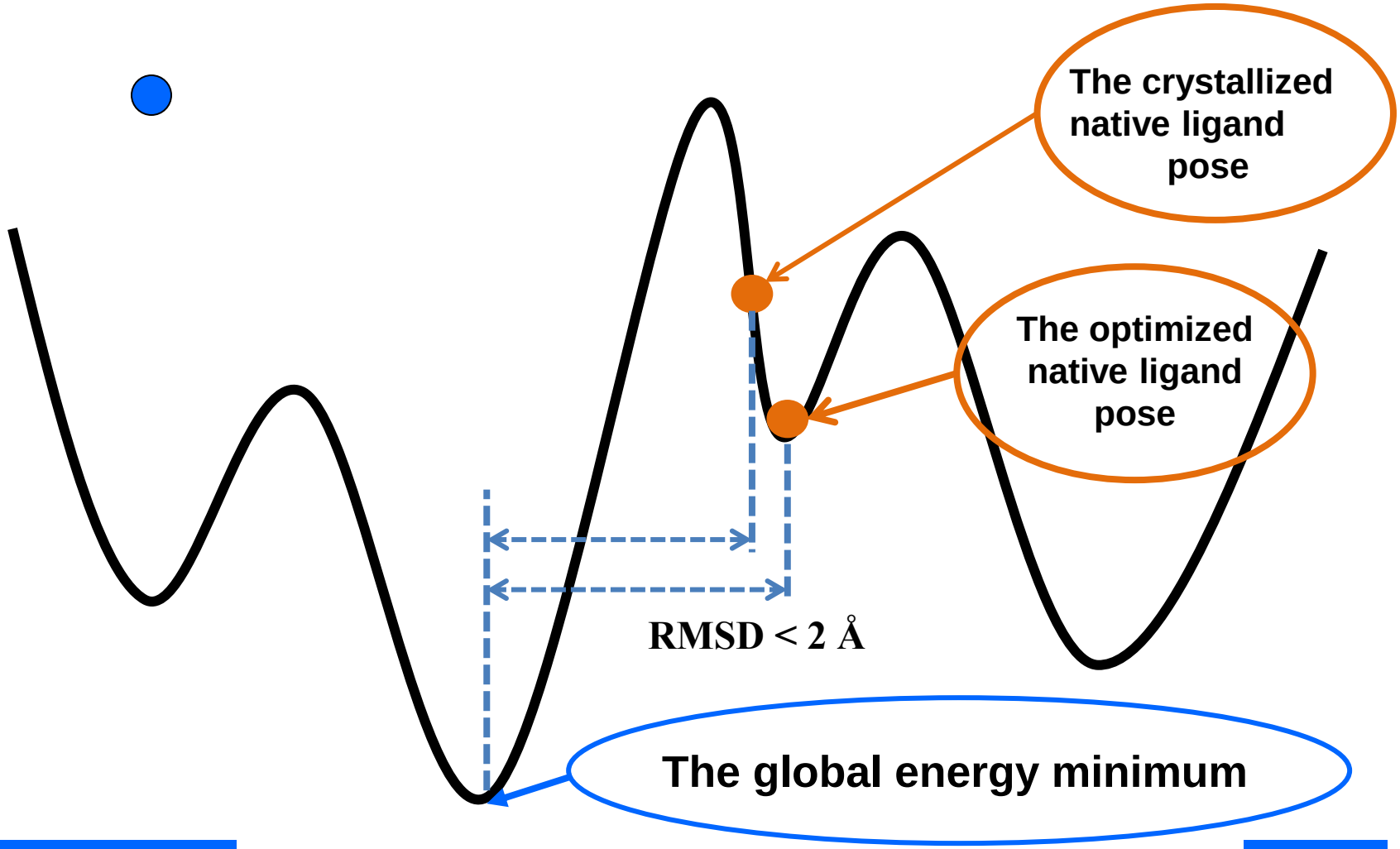
1. **Docking programs are extremely demanded at the initial stage of the drug development pipeline**

**The COVID-19 pandemic has shown a high demand for docking: over 3 years, hundreds of articles have been published on the search for inhibitors of SARS-CoV-2 target-proteins using DOCKING.**

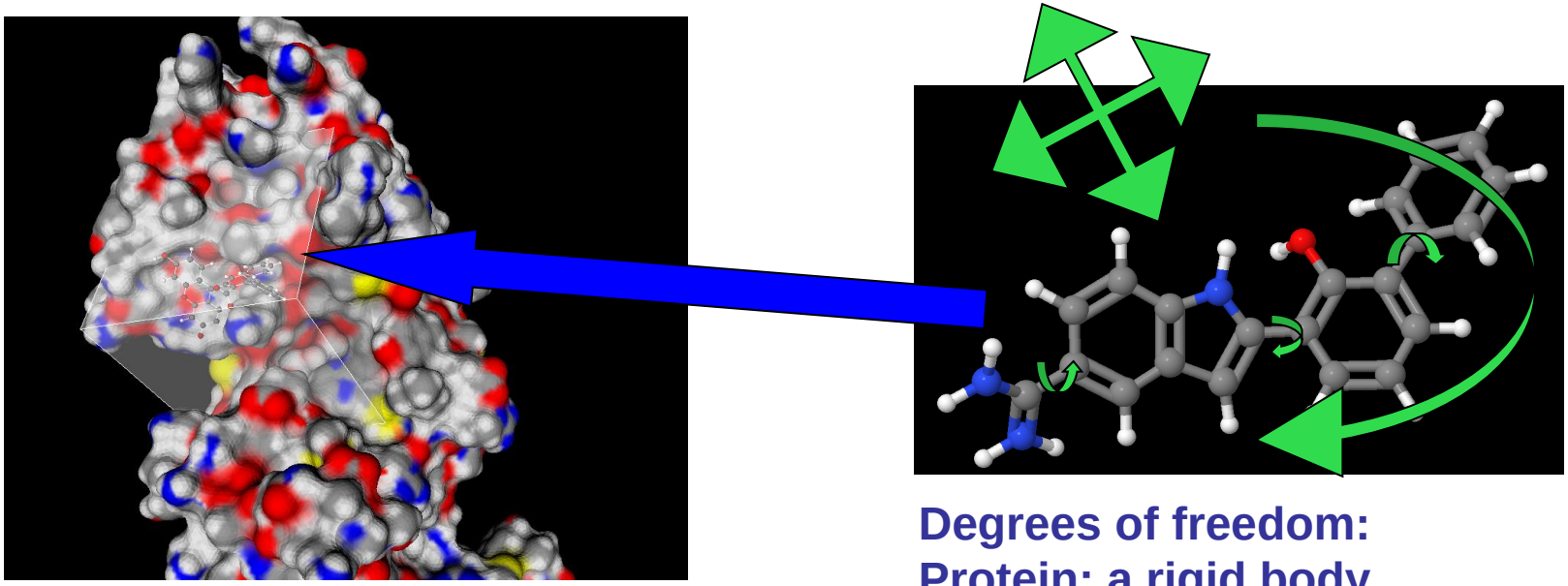
2. **What idea should you use to develop docking programs?**

**The search for the global energy minimum**

# Many local minima on the multi-dimentional protein-ligand energy surface



# The SOL docking program: multi-processor, MPI-based



Adapted for virtual screening of large databases of ligands on the Lomonosov-2 supercomputer of Lomonosov Moscow State University

Degrees of freedom:
Protein: a rigid body
Ligand:
- 3 translations as a whole body
- 3 rotations as a whole body
- 10-15 torsions

**Discovery inhibitors of : thrombin, urokinase (uPA), coagulation factors: Xa, XIa, XIIa, SARS-CoV-2 Mpro, nsp16 (2'-O-Methyltransferase)**

# SOL: Genetic Algorithm of global optimization

Sulimov A.V., et al. J. Chem. Inf. Model. 2013, 53 (8) 1946
Sulimov V.B., et al. J. Turkish Chem. Soc. Sect. A Chem., 2020, 7 (1) 259-276

- **The rigid protein is represented by a grid of potentials**
- **Potentials of ligand probe atom interactions with the protein:**
  - Coulomb interactions – the MMFF94 force field
  - Van der Waals interactions – the MMFF94 force field
  - Desolvation energy – the Generalized Born model

  **The grid of potentials considerably accelerate docking**
- **The docking region is the cube covering the protein active site:**

  -101 X 101 X 101 points

  - The cube edge = 22 Å, the grid step size = 0.22 Å
- **The global optimization of the target energy function:**

  **ligand grid energy + ligand stress energy**

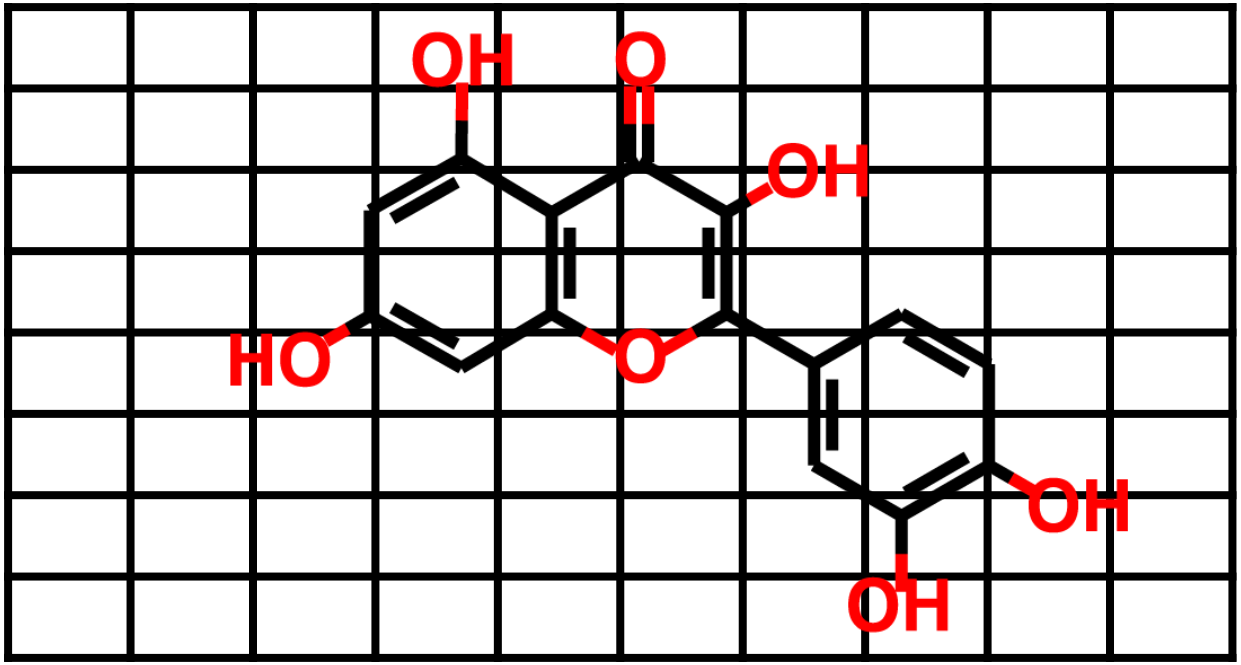
  The ligand stress energy – the MMFF94 force field

# Calculation of ligand energy in the protein field

$$E_{lig-protein} = \sum_{i=1}^{N} E_i$$

N – the number of atoms in the ligand

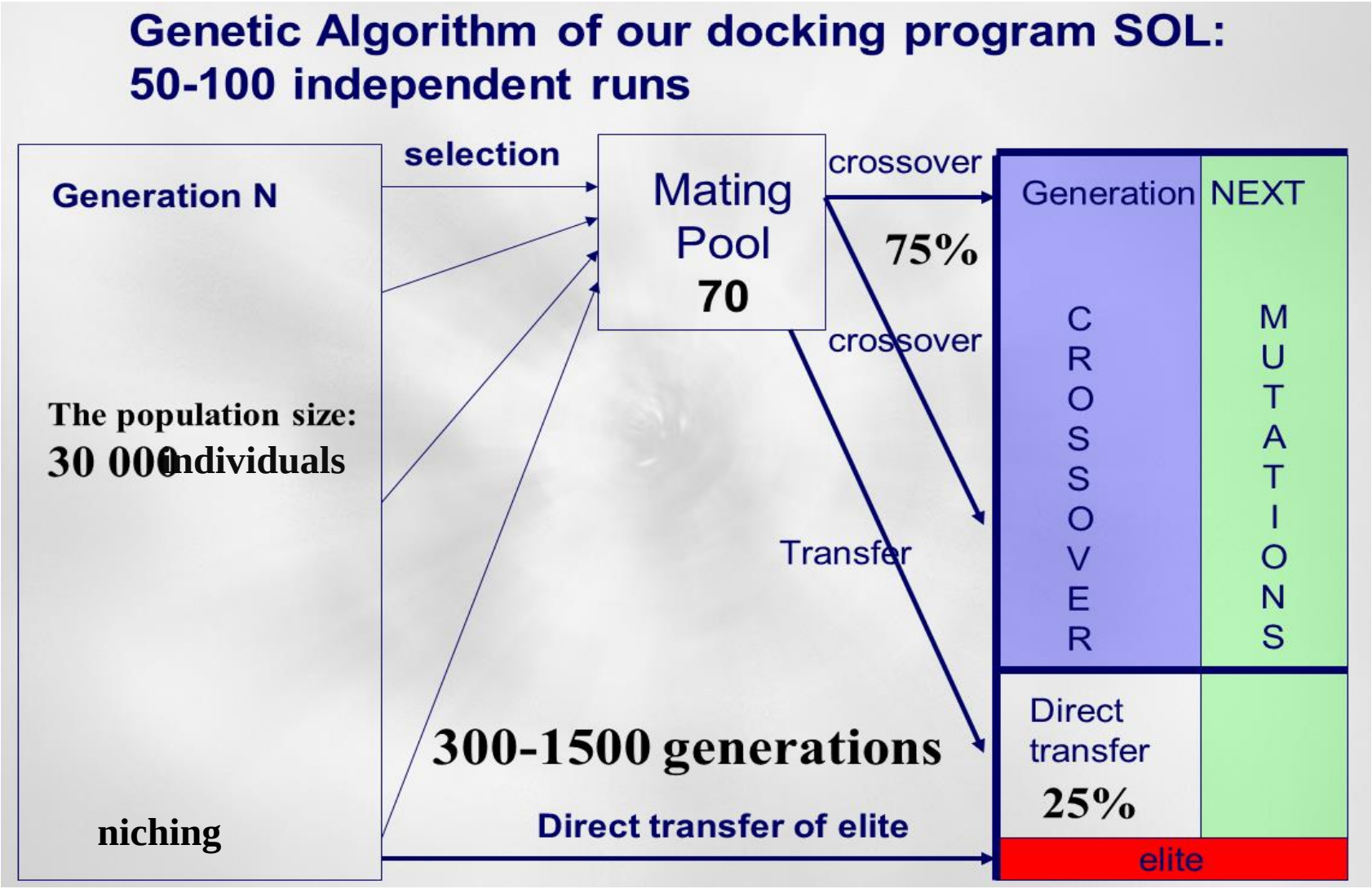
$E_i$ – energy of *i*-th atom of the ligand in the protein field




**If an atom falls between grid nodes, then its energy is calculated as a result of interpolation based on the energy values at the 8 nearest nodes**

**As a result the binary file is obtained containing all interaction potentials of ligand atoms with protein for all types of atoms (C, N, H, S, O...) at all nodes of a 3-dimensional grid covering the entire active center of the target protein**

# Genetic Algorithm of Global Optimization



# SOL: Lomonosov-2 Supercomputer

GA parameters: population size: 30000, number of generations: 1000

1 run of the GA gives 1 solution to the global optimization problem,

50 independent runs of the GA give 50 solutions of the global optimization problem – 50 ligand positions.

Clustering of these positions gives a measure of reliability of docking:
two positions belong to the same cluster if RMSD between them < 1 Å

1 cluster with 50 ligand positions – very reliable result of docking

50 clusters – unreliable result of docking

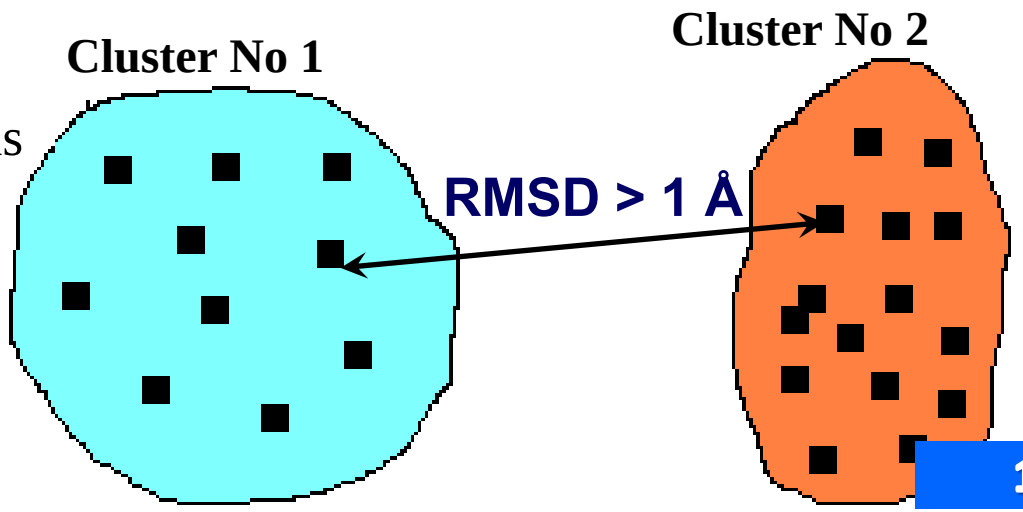
Virtual screening: dozens of thousand up to 1 million ligands
1 ligand per 1 core, docking of 1 ligand in 1 hour
1 ligand per 128 cores: less than 1 minute

High GA parameters: increase the population size up to 6 000 000 results in improvement of the docking reliability; multi-core calculations

# Clustering of 50 independent solutions

- Calculation of the Root Mean Square Deviation (RMSD) between solutions – ligand poses
- When calculating RMSD, the differences in Cartesian coordinates of the same ligand atoms in two positions are taken
- All solutions are divided into groups (clusters): within one cluster RMSD < 1 Å between any two solutions (ligand poses)
- Clusters are numbered by increasing energy of the best ligand pose in the cluster: cluster #1 contains the pose with the lowest energy

If a large percentage of 50 solutions falls into the cluster No. 1, then the global optimization problem – has been solved with high reliability



# Docking score

- The best ligand position corresponds to the global energy minimum of the energy of the protein-ligand system
- This position is used to estimate the free energy of the protein-ligand binding $\Delta G_{bind}$

$$\Delta G_{bind} = \Delta H - T\Delta S \qquad \mathrm{K_d} = e^{\frac{\Delta G_{bind}}{RT}}$$

$\Delta H$- the binding enthalpy, $T\Delta S$- the binding enthalpy

The scoring function (score) is the estimation of $\Delta G_{bind}$

Score = $\sigma E_{protein-ligand} - \mu N_{tor}$, $\sigma, \mu$ – fitting param, $N_{tor}$ - the number of ligand torsions, $E_{protein-ligand}$ - the energy of the ligand in the protein field – grid energy

# Virtual screening

Creating a protein model

Creating 3D structures of chemical compounds

SOL program docking $\sim 10^4$ - $10^6$ ligands

Postprocessing: Quantum Chemistry, Molecular Dynamics

Binding enthalpy calculations by the quantum-chemical PM7 semi-empirical method with the COSMO solvent model

$\Delta H = H_{complex} - H_{protein} - H_{ligand}$: $\sim 10^2 - 10^3$

Experimental *in vitro* verification (≈20 compounds)

# Preparation of ligands and a target-protein for docking

- Using a good quality structures from Protein Data Bank, Resolution < 2.5 Å, no missed atoms and/or residues in the active site of the target-protein
- Addition hydrogen atoms to the target-protein, determination of the residues protonation states – automatic processing
- Preparation 3D-models of ligands using their 2D-models: generation different ligand conformations, different macro-cycles and non-aromatic rings conformations
- Protonation states of ligands

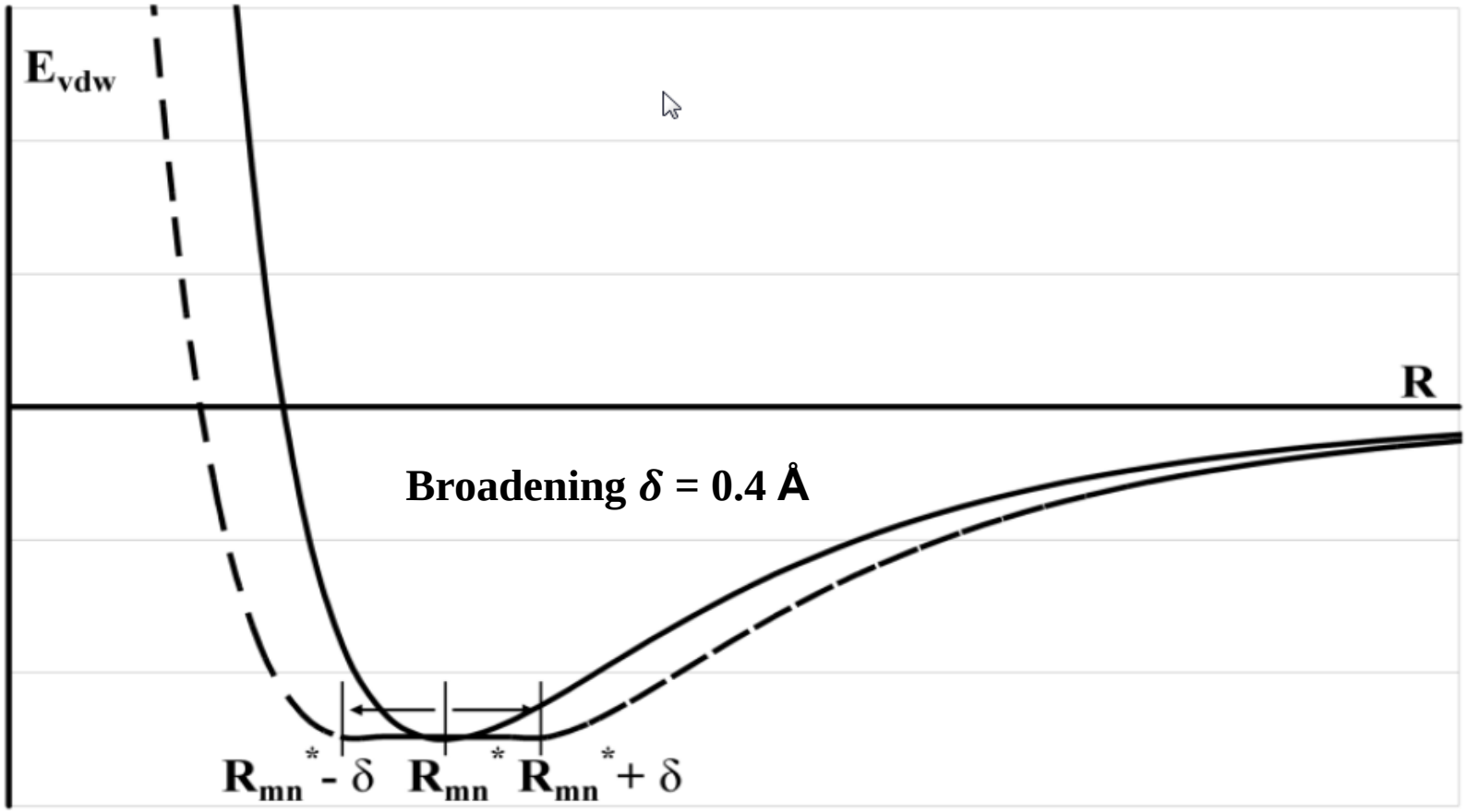
High quality of full atomic models of the target-protein and ligands play a key role for high quality of docking

# Most popular docking programs

- **AutoDock Vina** – one of the most popular docking programs
- **AutoDock** – The Scripps Research Institute – **USA**
- **DOCK** – one of the oldest docking programs - University of California, San Francisco, **USA**
- **GOLD** – The Cambridge Crystallographic Data Centre - **UK**
- **ICM** - Molsoft, LLC, San Diego, CA, **USA**
- **FlexX** – BioSolveIT - **Germany**
- **GLIDE** (Schrodinger, Inc.) – one of the most advanced programs using its own OPLS force field, expensive; in general, it is not adapted for supercomputing.
- Each docking program has its own peculiarities of work due to an individual combination of models and approximations
- **Supercomputer docking: Faster and Larger:**
- **HSP-DOCK**, **BUDE**, **VinaLC**, **VirtualFlow** (**AutoDock Vina** etc.): automatic preparation of ligands and analysis of the docking results

# Protein flexibility in docking

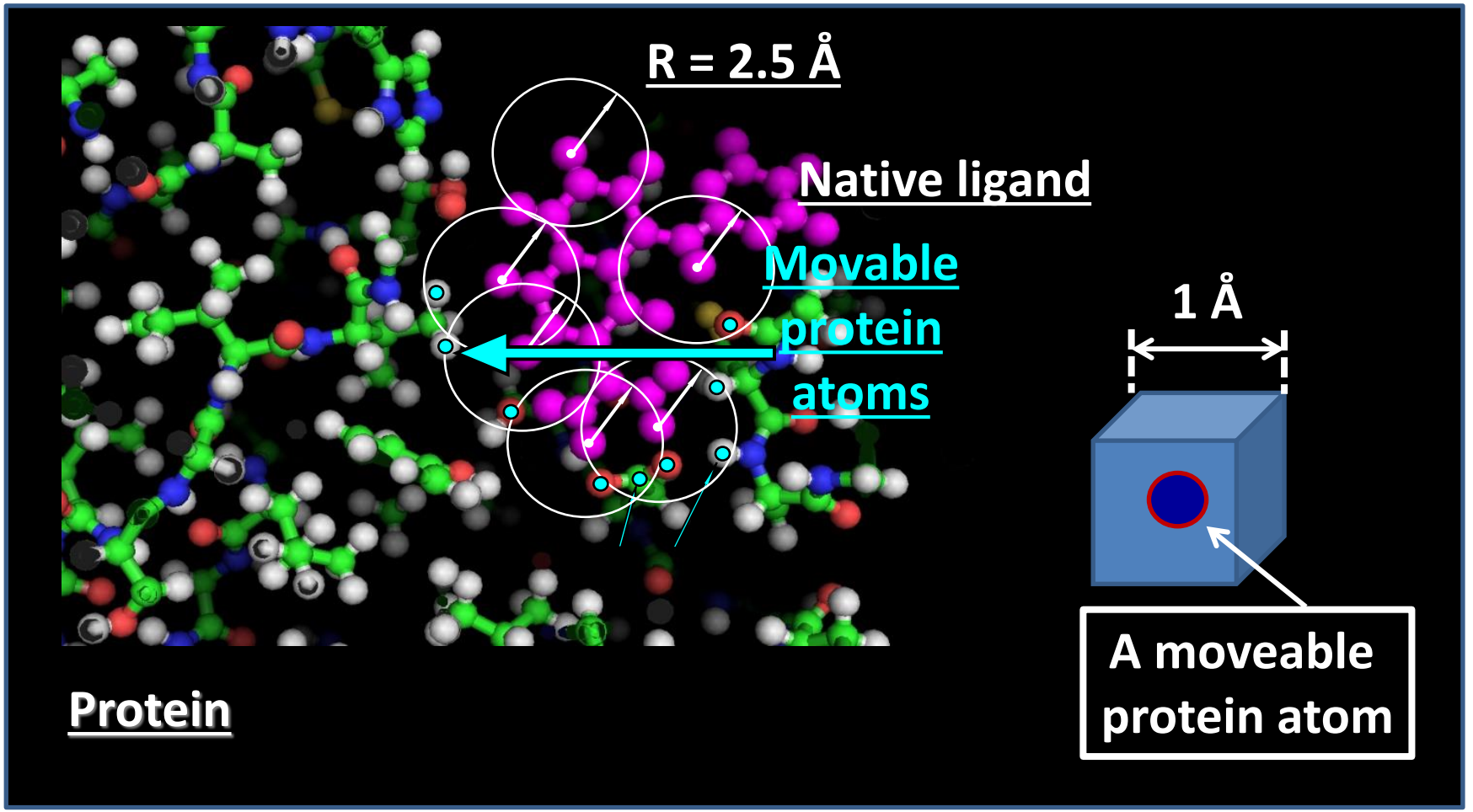
Параметр δ называется уширением и задается в файле *par-file* входных параметров модуля SOLGRID.



# Protein flexibility in docking

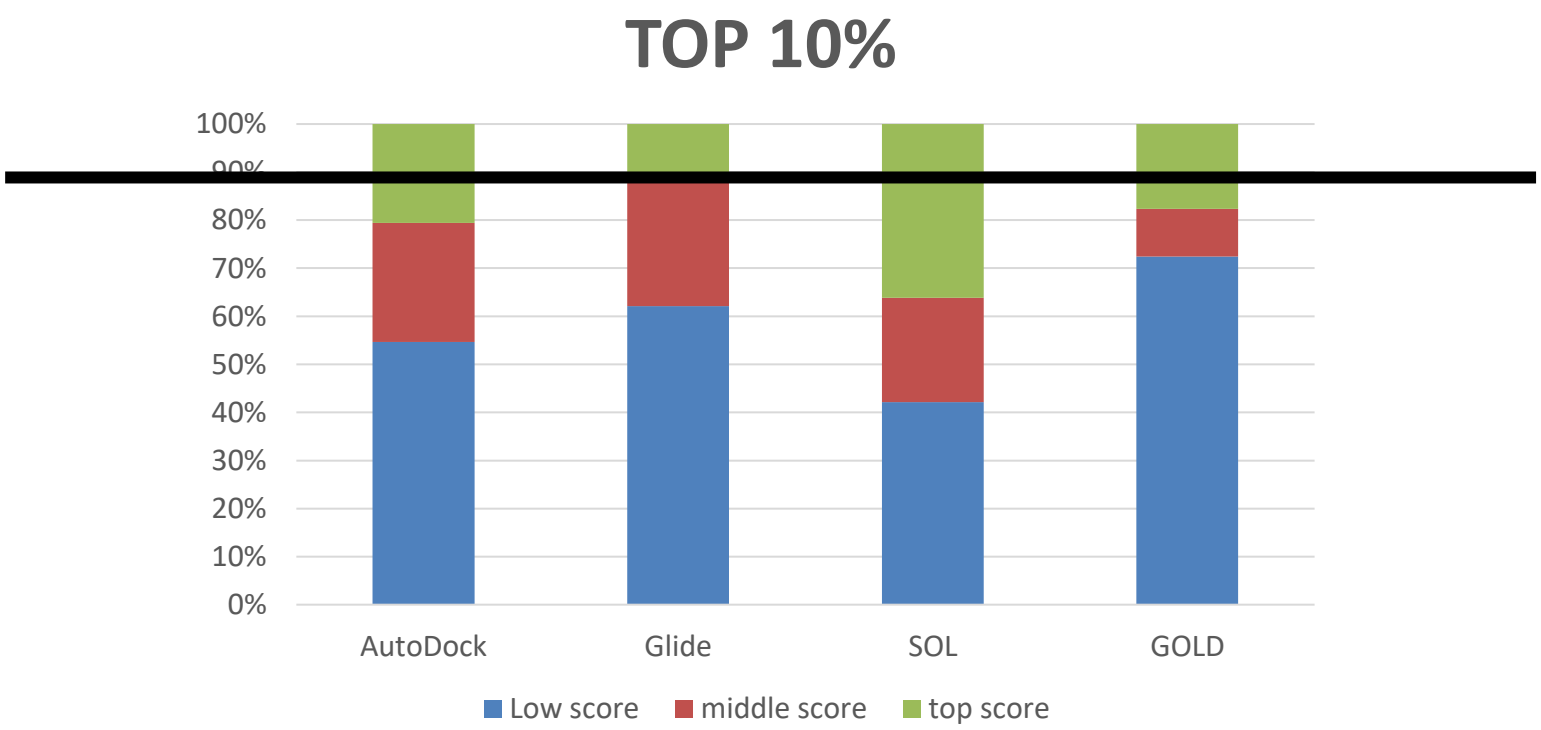
- In the docking, specified side chains of the protein are separated and processed explicitly together with the ligand. AutoDock4 and AutoDock Vina: 10 degrees of freedom of a ligand + 6 degrees of freedom of two side chains
- Ensemble docking: generation of several target-protein conformation followed by independent docking to each rigid conformation. AutoDock, ICM, FlexE, FlipDock, SurflexDock, Glide
- Selecting mobile protein atoms, which are moved in a restricted space simultaneously with a ligand in the docking process

# Choice of the protein moveable atoms

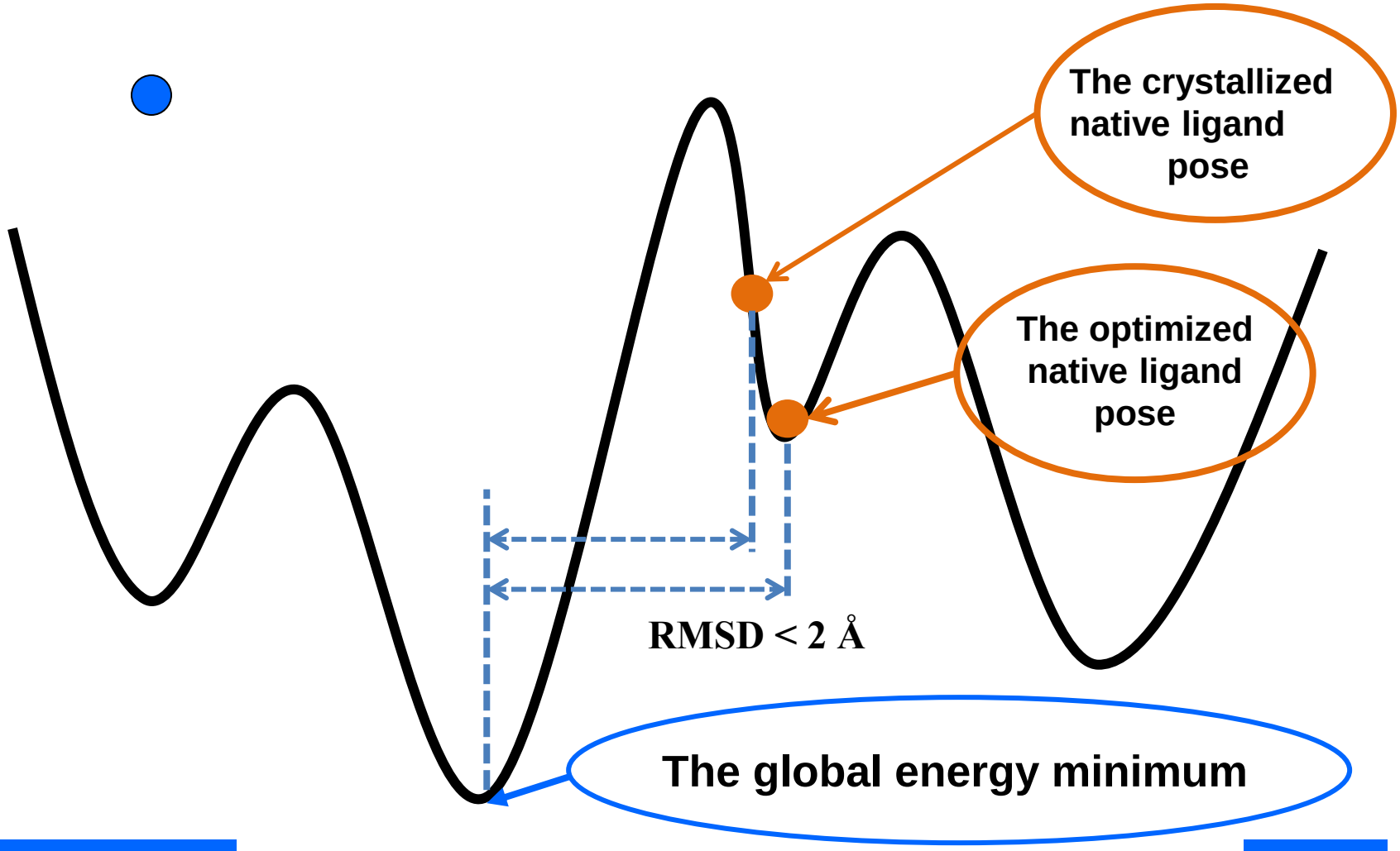


# Consensus docking

- Docking a given library of ligands into a target-protein using several different docking programs
- Selection of only those ligands which are in the top 10% of the results of most docking programs



# To find all low energy minima on the energy surface



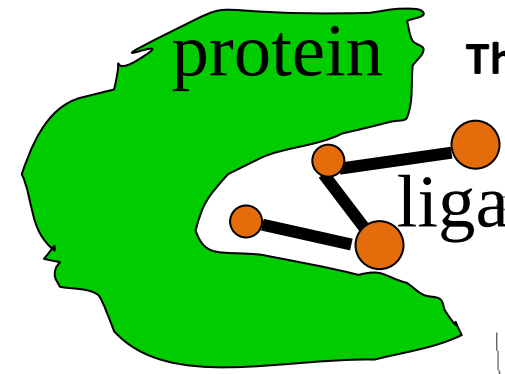
# FLM docking program – Find Local Minima

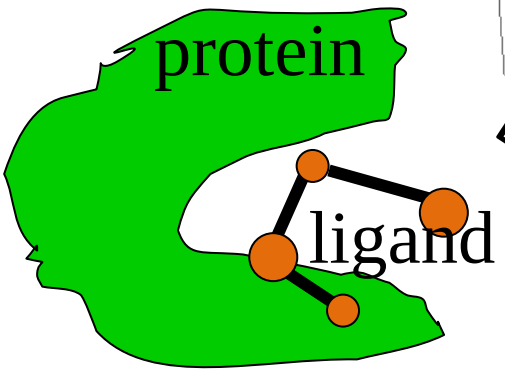
- FLM does not use any preliminary calculated energy grid
- Rigid protein
- Local energy optimization: the variation of positions of all ligand atoms from a random initial ligand pose
- Vacuum or implicit solvent models PCM or Generalized Born
- The MMFF94 Force Field, no simplification, no fitting parameters
- **Exhaustive search for the low energy minima spectrum**
- Parallel calculations 1 complex: **8191 cores** several hours on the Lomonosov supercomputer ≈20 000 CPU·hours
- FLM can be used: verification of global optimization algorithms and for comparison of different energy functions in docking

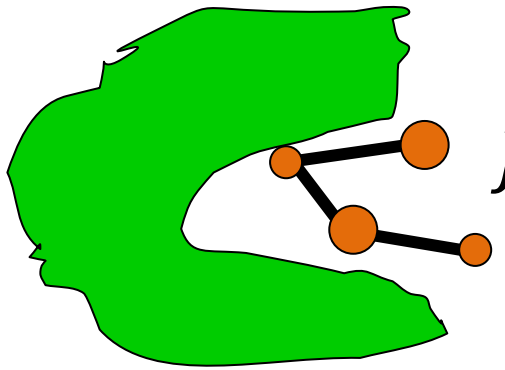
# The SOL-P supercomputer program: docking with the Tensor Trains global optimization

1. The MMFF94 force field
2. There is no a grid of protein-ligand interaction potentials
3. No simplifications
4. No fitting parameters
5. Multi-processor performance: several hundred computing cores
6. The continuous energy of the protein-ligand complex is transformed into a multi-dimension tensor with a very fine grid
7. The modern tensor analysis methods are applied to the search of the largest in module element of the tensor

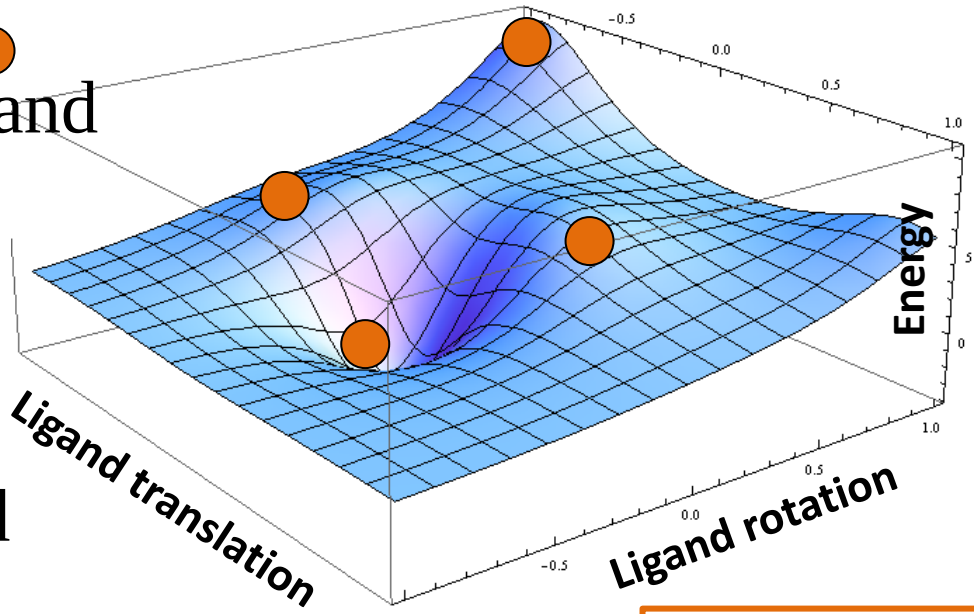
# Transformation of the continuous function into the multi-dimensional tensor

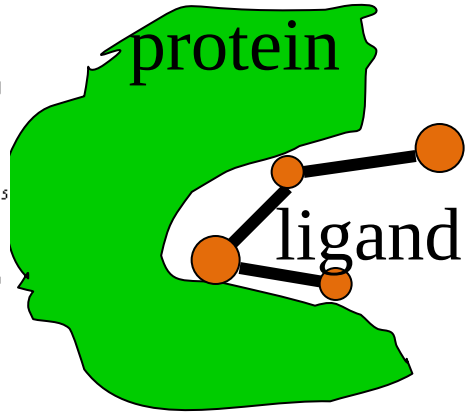


**The energy discretization grid : $n = 2^{16}$ – along one dimension**




**The search of the minimal (maximal) element of the multi-dimensional tensor**

$$f(x_1, x_2, \ldots, x_d) \Rightarrow A(i_1, i_2, \ldots i_d)$$

**The search of the global minimum (maximum) of the continuous function**

$i$=1, 2, 3,…$10^{16}$

# Tensor Train Decomposition

- Multidimensional array (tensor) $A \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ can be decomposed in the form:

$$A(i_1, \ldots, i_d) \approx$$

$$\approx \sum_{\alpha_1=1,\ldots,\alpha_{d-1}=1}^{r_1,\ldots,r_d} G_1(i_1, \alpha_1) G_2(\alpha_1, i_2, \alpha_2) \ldots G_{d-1}(\alpha_{d-2}, i_{d-1}, \alpha_{d-1}) G_d(\alpha_{d-1}, i_d)$$

- $r_1, \ldots, r_{d-1}$ are called **TT**-ranks of the tensor
- 3–dimensional tensors $G_i \in \mathbb{R}^{r_{i-1} \times n_i \times r_i}$ are called cores (carriages) of the tensor train
- **If TT-ranks are small TT-decomposition is useful**

# The SOL-P supercomputer program: docking with the Tensor Trains global optimization

1. **For a rigid protein SOL-P docks faster than FLM:**
   - **SOL-P needs 100 CPU*hours**
   - **FLM needs 10 000 CPU*hours**

2. **SOL-P docks ligands with up to 25 torsions**

   **to be compared with max 10-15 torsions SOL,**

   **max 10 torsions Glide, GOLD and ICM docking programs**

3. **SOL-P can dock flexible ligands with several dozen moveable protein atoms: up to 157 degrees of freedom,**

Sulimov A. V., et al. SAR QSAR Environ. Res., 2019, Vol. 30, No. 10, P. 733–749.

# Quantum Quasi-docking

- Using the FLM docking program with MMFF94 and the PCM solvent model lowest energy minima of test complexes are found: **8192 minima for each complex**
- Each energy minimum is re-calculated with **a quantum-chemical semiempirical methods PM7 with a solvent model COSMO**

**PM7 – the new quantum-chemical semiempirical method:**
- **Improved dispersion interactions**
- **Improved Hydrogen Bonds description**

J. J. P. Stewart *J. of Molecular Modeling*, 2013, vol. 19, 1–32

- The global minimum of the PM7/COSMO energy is determined
- **A ligand pose corresponds to this global energy minimum**
- **PM7/COSMO demonstrates much better accuracy than force fields**

A. Sulimov, et al. Nanomaterials, 2022, 12, 274.

# Conclusions

- Docking is the only tool for searching for inhibitors at the initial stage of drug developmentSelected compounds were tested in vitro
- Docking has proven to be extremely in demand during the COVID-19 pandemic
- All necessary conditions are available to improve docking accuracy
- The most important goal is to create a quantum docking program that will use quantum chemistry methods
- Currently, docking efficiency is lagely determined by the post-processing method used
- The choice of ligand database plays a critical role in the success of virtual screening using docking

# Thank you for attention!

… Surely every medicine is an innovation; and he that will not apply new remedies, must expect new evils …

*Francis Bacon*
(1561-1626)
OF INNOVATIONS

**The work is financially supported by the Russian Science Foundation, Agreement no. 21-71-20031**