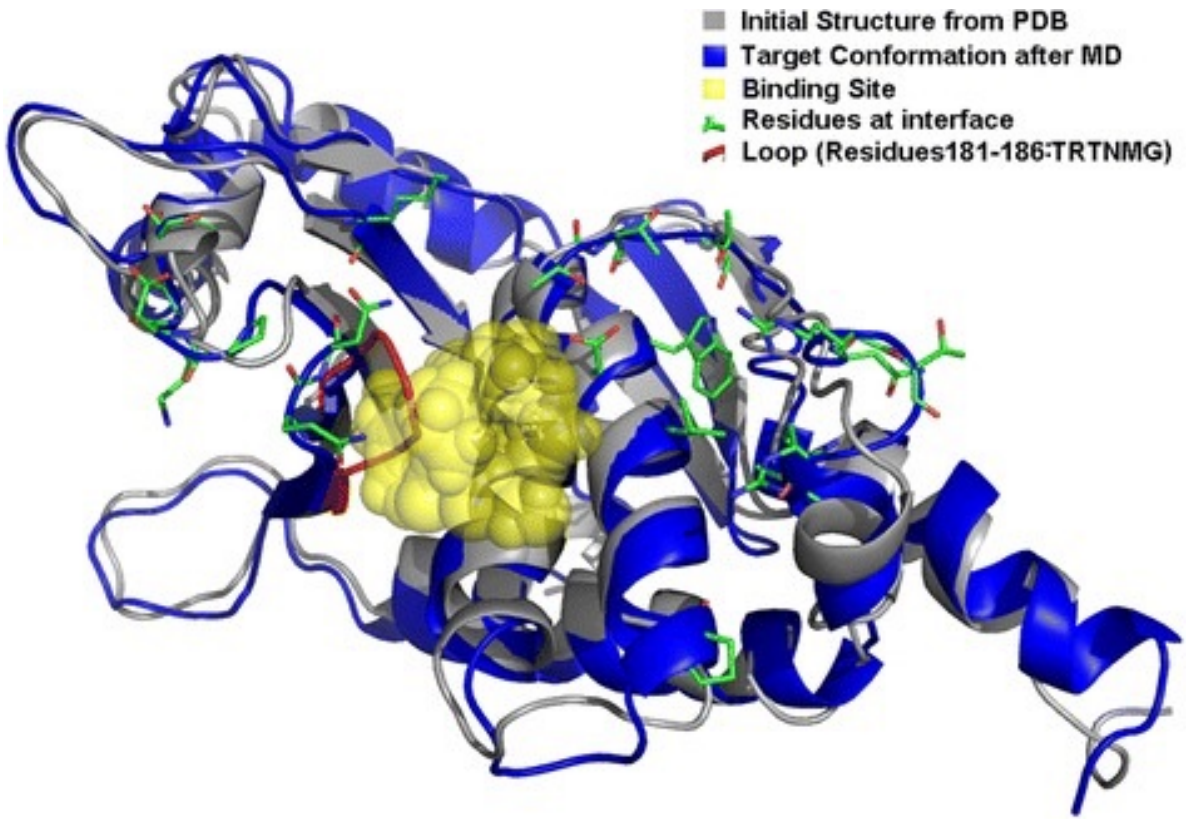
Comparative assessment of scoring functions
(Introduction to the CASF competition project)

Sergei Grudinin, Maria Kadukova, and Vadim Strijov

Moscow Institute of Physics and Technology
Institut national de recherche en informatique et en automatique

August 20th, 2020

Protein-ligand docking problem and farmacology

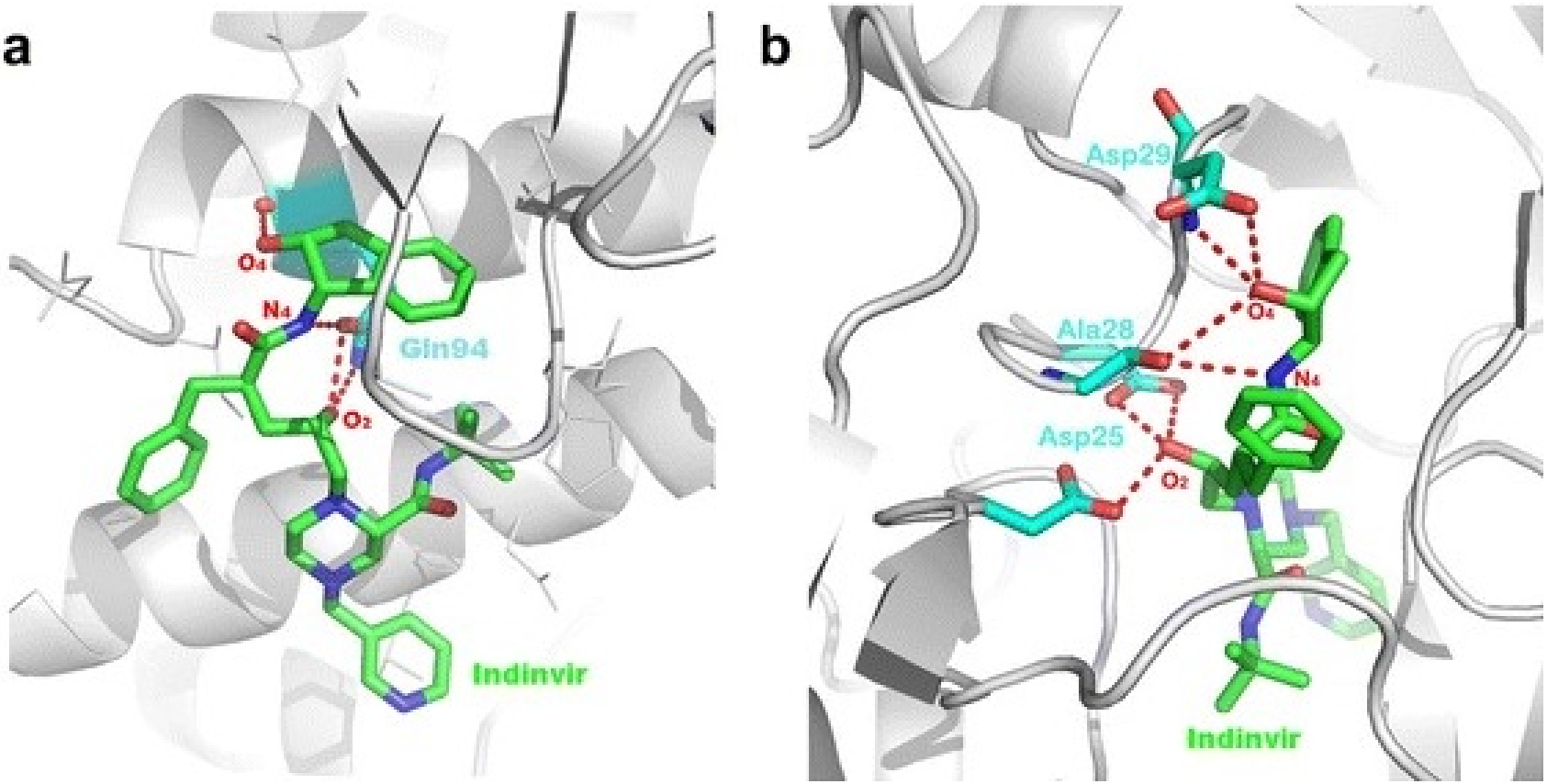


The binding interface of VP24 of Ebola with Karyopherin alpha

---

Drug to target Ebola virus replication and virulence using structural systems pharmacology by Zheng Zhao et al. //BMC Bioinf., 2016

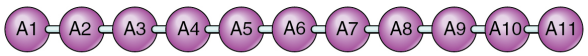
It is important to select a ligand, which fits requirements



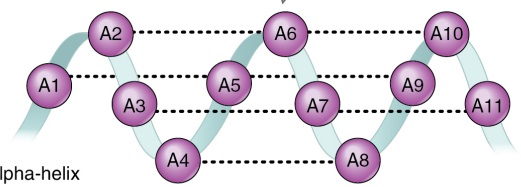
The predicted binding mode of Indinavir in VP24 of Ebola (a) and (b) the binding mode of Indinavir in HIV protease (PDB id 2AVO)

Drug repurposing to target Ebola virus replication and virulence using structural systems pharmacology by Zheng Zhao et al. //BMC Bioinf., 2016

Amino-acids, proteins, and ligands

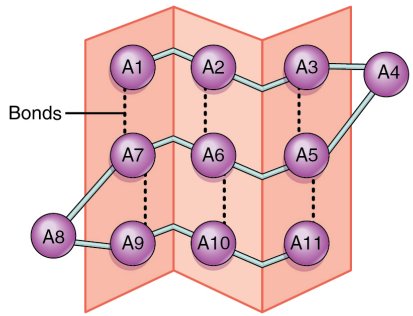


(a) Primary structure — Chain of amino acids

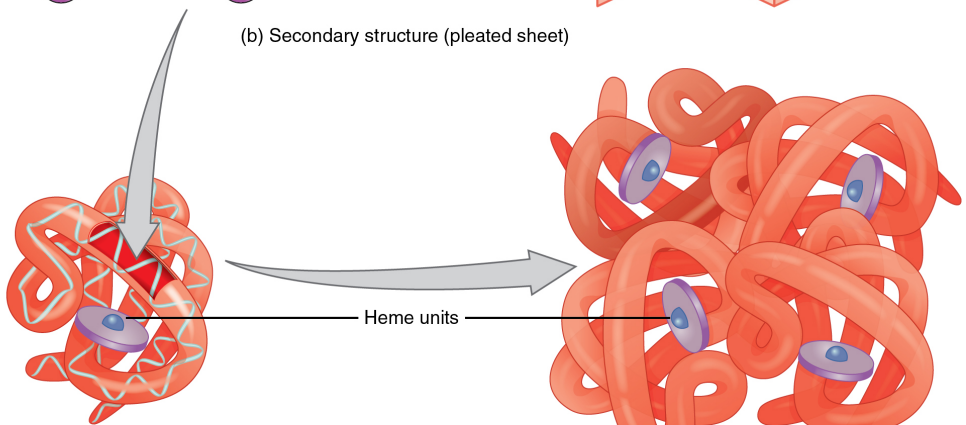


Alpha-helix

OR

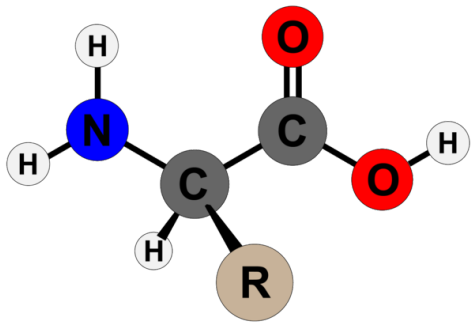


(b) Secondary structure (pleated sheet)

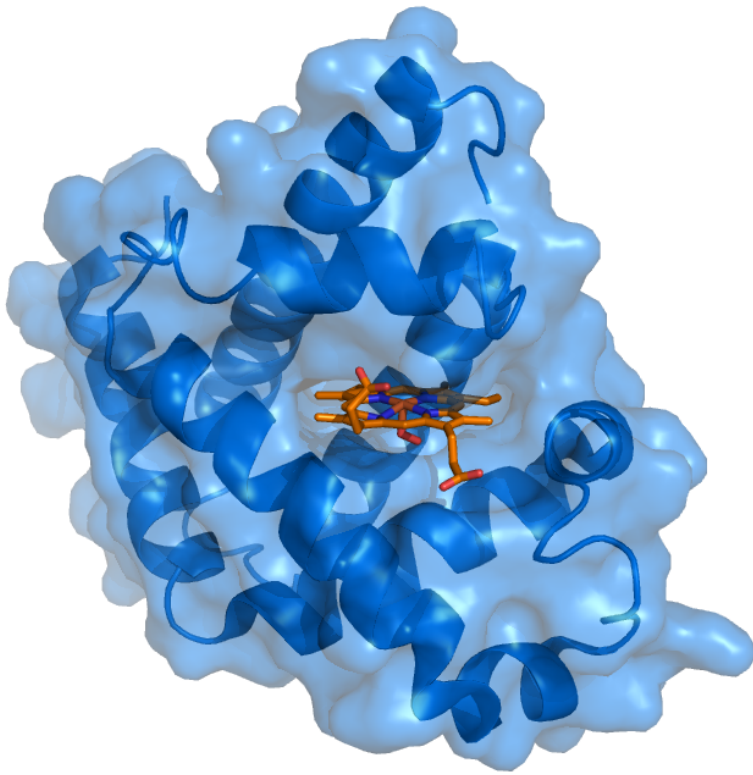


(c) Tertiary structure

(d) Quaternary structure — Hemoglobin (globular protein)

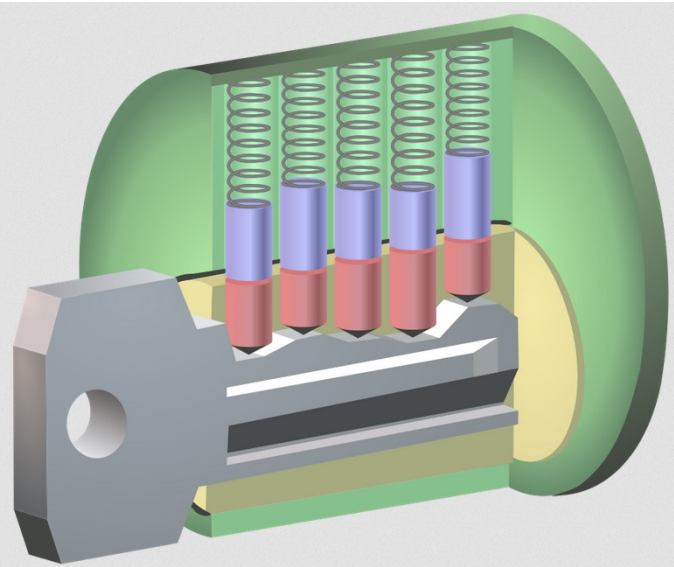


Amino-acid, 20 types



Wikipedia images: Amino acid, Protein, Ligand (biochemistry)

Protein-ligand complex and scoring function to dock

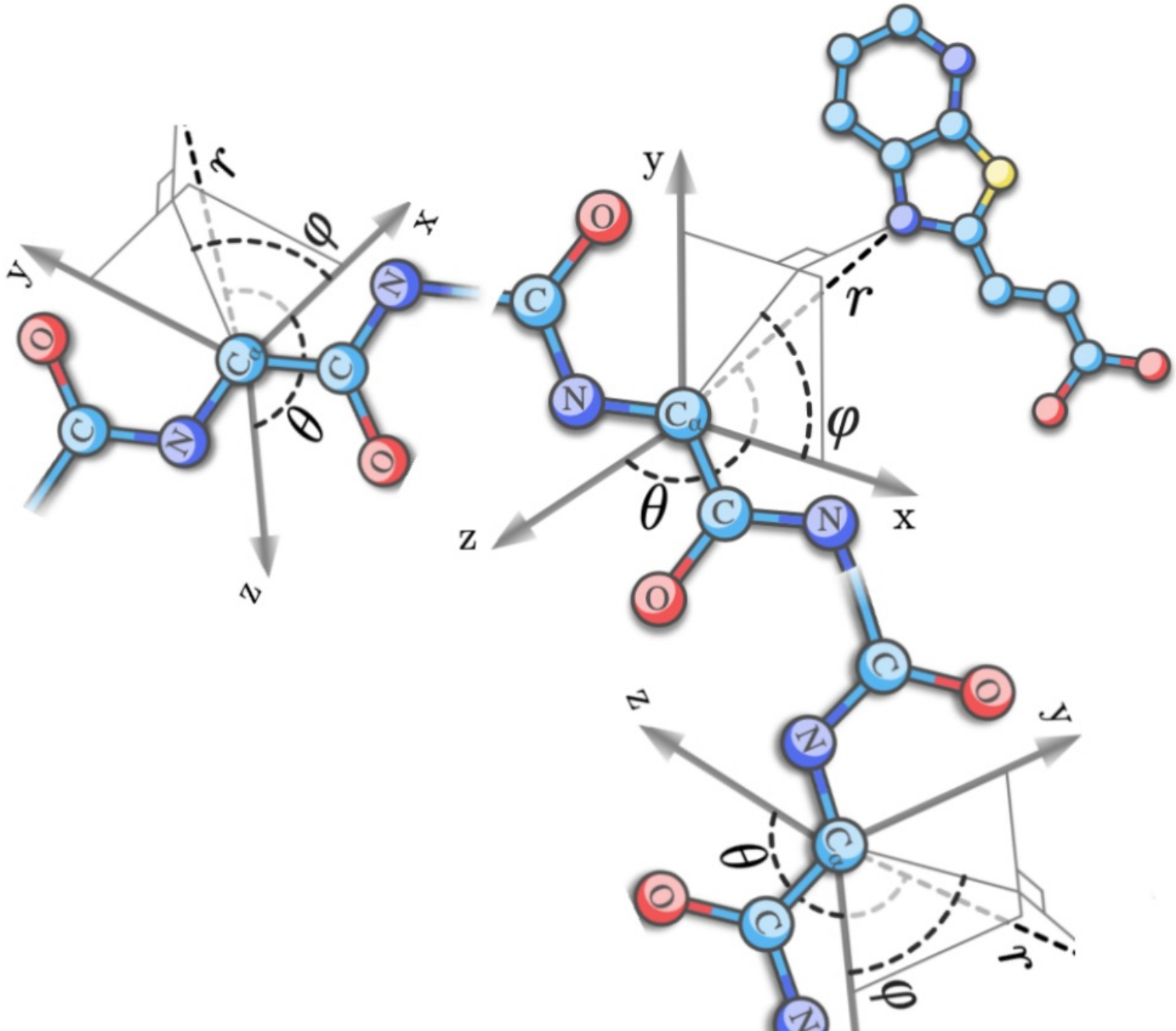


Roles:

- Protein: the lock
- Ligand: the key
- Scoring function: to move along
- Native conformation: to open lock

From: Locksmith Canberra

Docking: a general view



Terminology of CASF

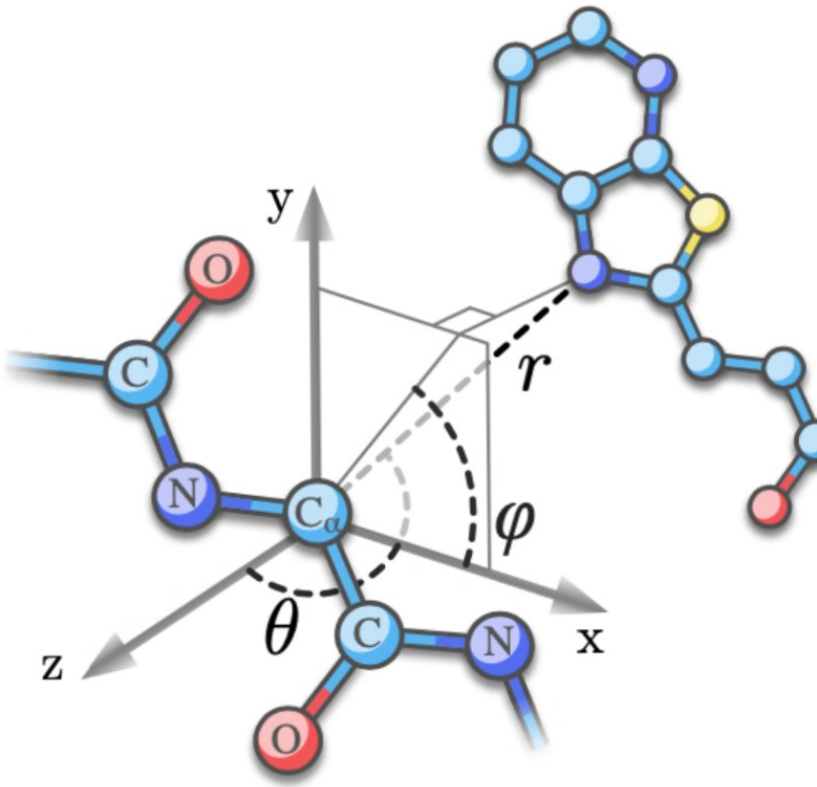
1. **Complex**: a subject to score $C(\text{prt}, \text{lig})$

   supposed to be a stable couple protein-ligand, (an element of metric space).

2. The protein **residue types** corresponds to the 20 standard amino acids, indexed by $\text{aa} \in \{1, \ldots 20\} = \{\text{aa}\}$.

3. The set of 37 ligand **atom types**, indexed by $\text{atm} \in \{1, \ldots 37\} = \{\text{atm}\}$,

   8 carbon types, 12 nitrogen, 7 oxygen, 4 sulfur, 2 phosphorus, and 4 halogens.



4. Frame: has a unique basis **B** for any amino acid in a protein;

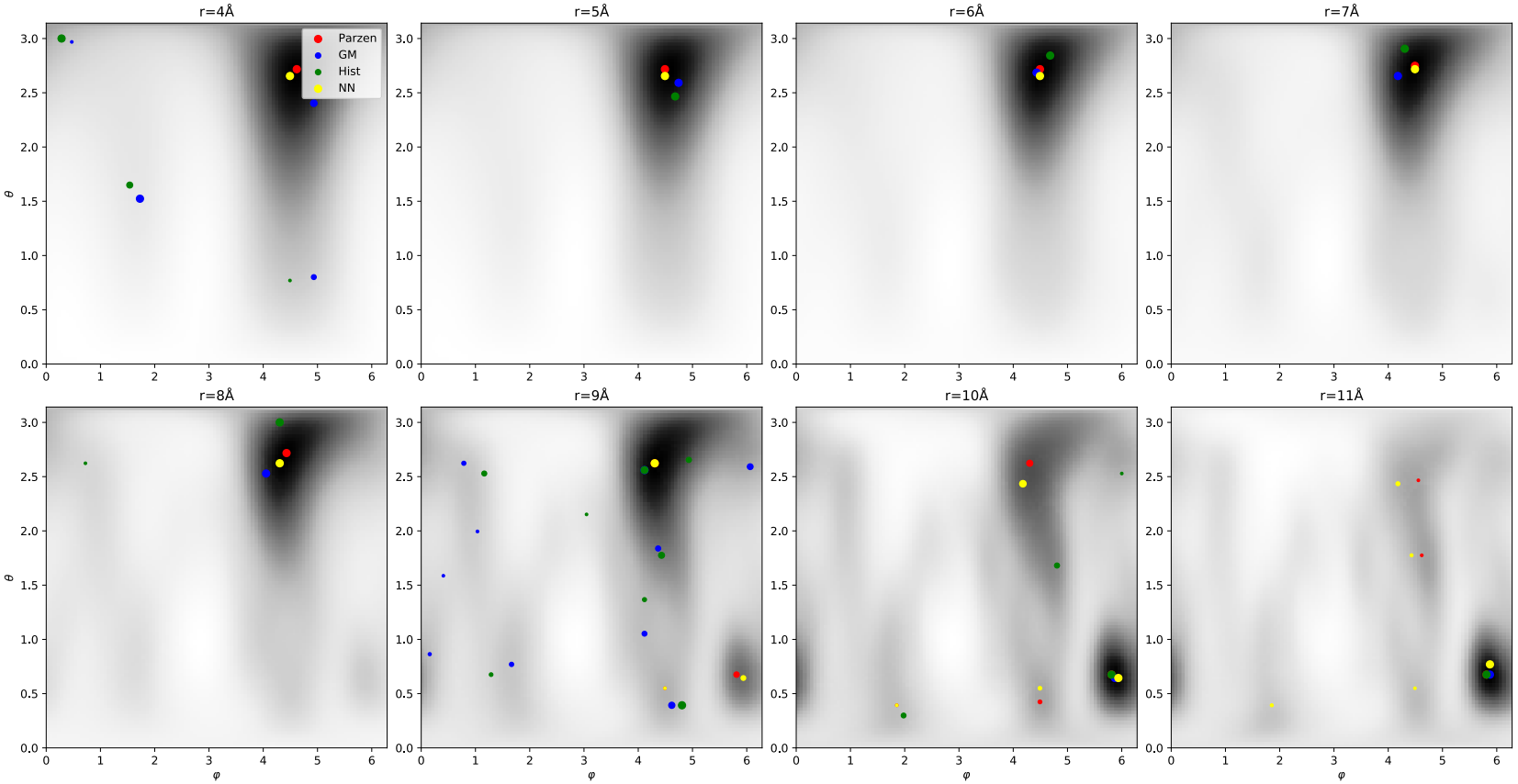
   also a rigid body translation (affine transformation) between any pairs of frames is given.

5. **Pose**: an orientation of a ligand (and so an atom) according to the frame, a triplet $[\theta, \varphi, r]^{\mathsf{T}}$

   in spherical coordinates $\{0, \ldots, 2\pi\}, \{0, \ldots, \pi\}, \{3, \ldots, 20\text{Å}\}$

Каталог экстремумов (1)

Protein 0, ligand 2

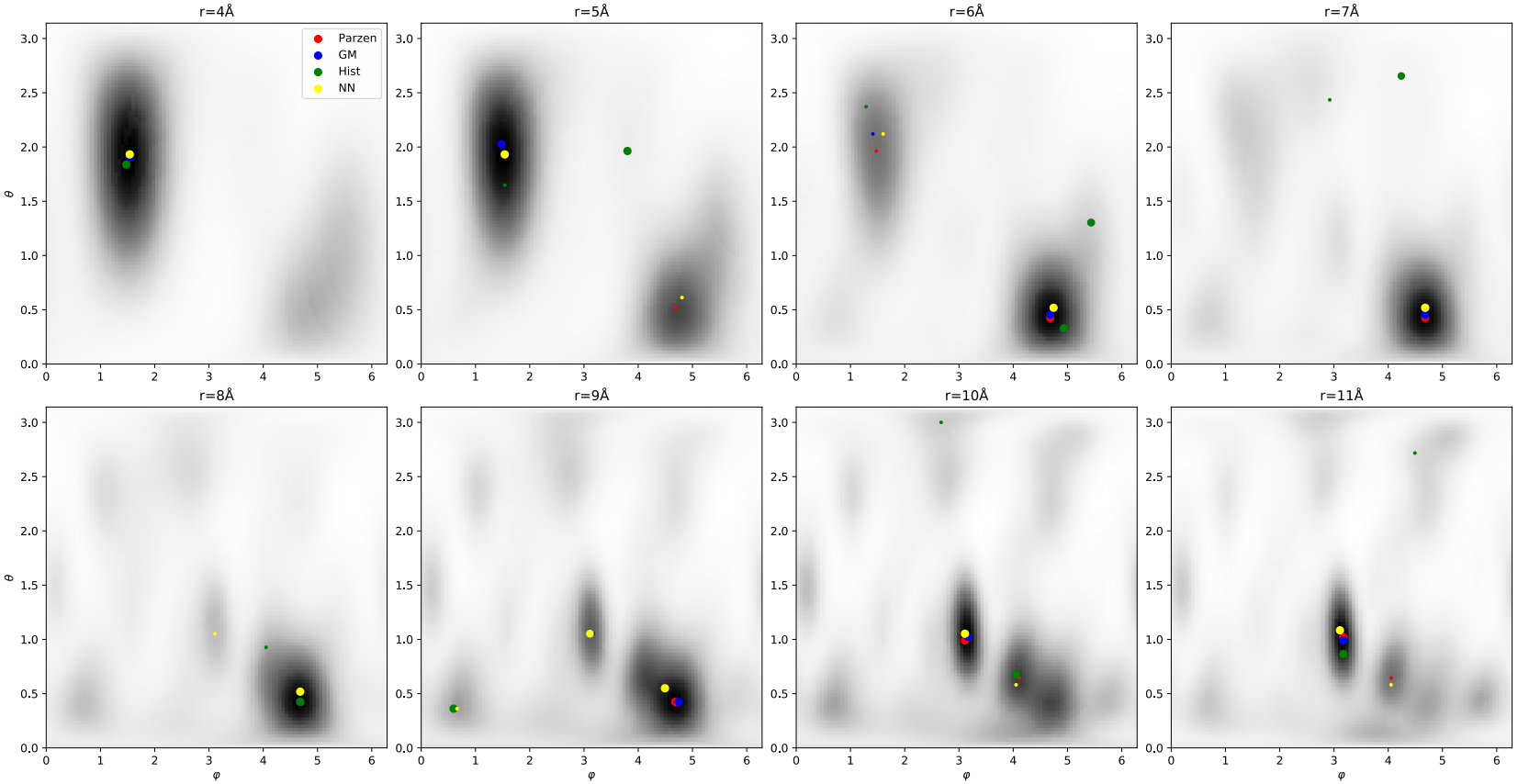


	Parzen	GM	Hist	NN
Parzen				
GM	28.98			
Hist	34.78	23.33		
NN	33.72	54.28	58.24	

MAPE, %

Каталог экстремумов (2)

Protein 10, ligand 23



	Parzen	GM	Hist	NN
Parzen				
GM	30.89			
Hist	60.94	48.88		
NN	39.61	59.71	86.75	

MAPE, %

Восстановление плотности распределения для построения метрического пространства

Модели $\mathfrak{F}$ и оптимизируемые структурные параметры

- Гистограмма — размеры перцентилей $dr, d\varphi, d\theta$.
- Окно Парзена — тип окна, ширина окна.
- Смесь гауссиан — число гауссиан (экстремумов).
- Нейросеть (2 скрытых слоя, 50 нейронов).

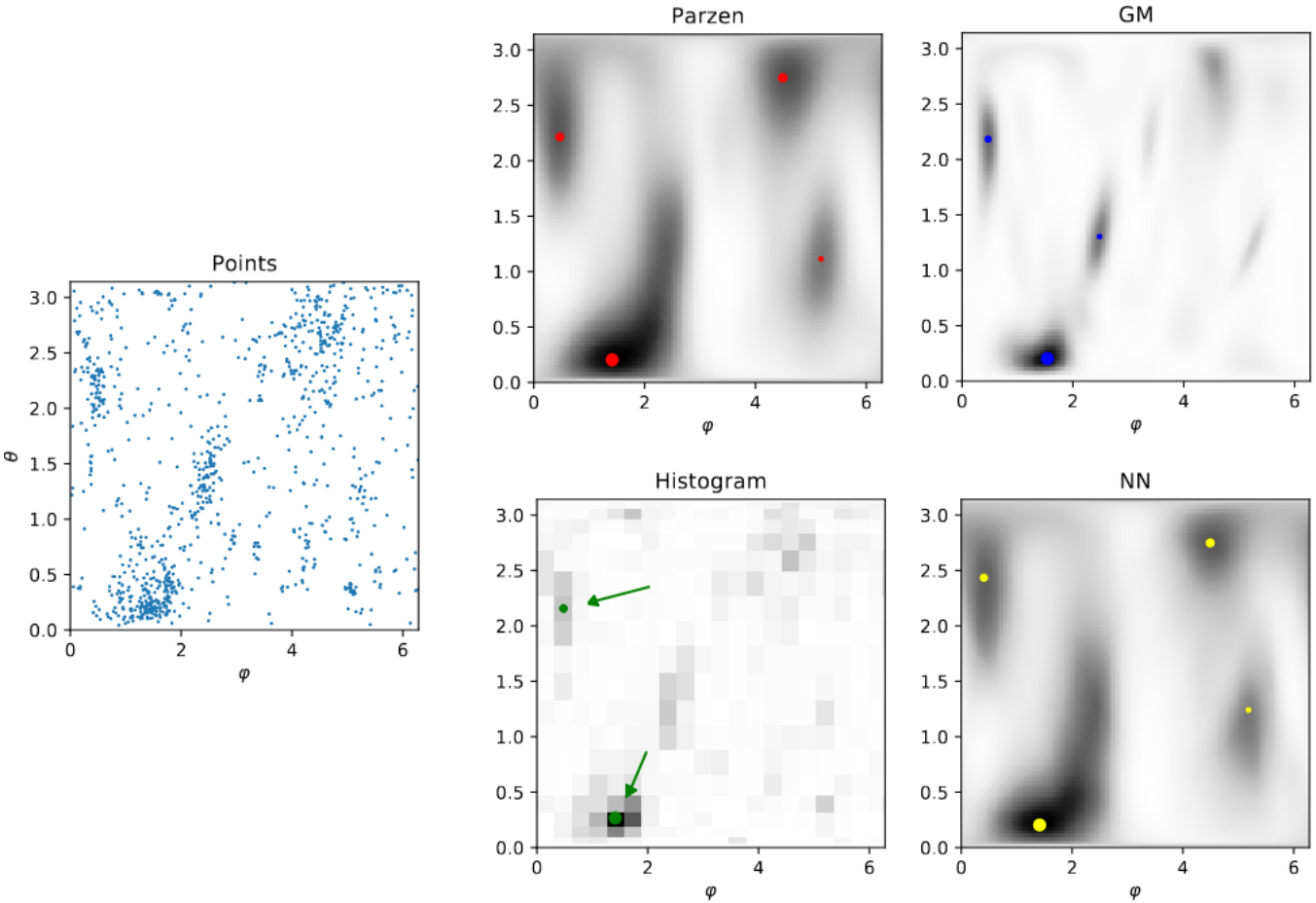
Критерии точности аппроксимации

Расхождение плотности усредняется по всем парам моделей:

$$L_1 = \sum_{1 \leqslant i < j \leqslant 4} \int_{\mathbb{R}^3} (f_i(\vec{r}, \hat{w}_i) - f_j(\vec{r}, \hat{w}_j))^2 d\vec{r},$$

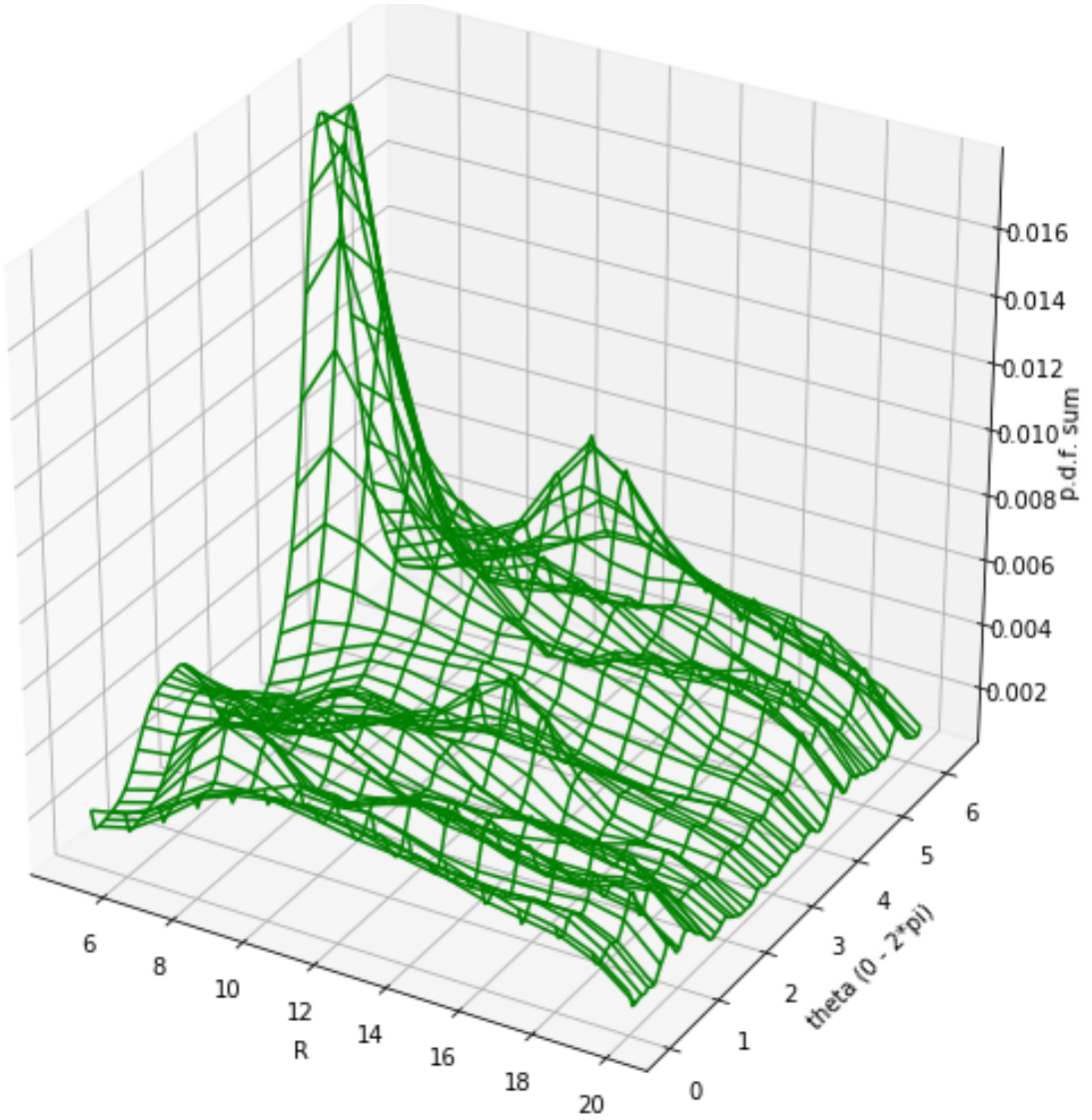
$$L_2 = \sum_{1 \leqslant i < j \leqslant 4} \int_{|f_i(\vec{r}, \hat{w}_i) - f_j(\vec{r}, \hat{w}_j))| > \varepsilon} d\vec{r}.$$

Восстановленная плотность $f(\varphi, \theta, r)$ при $r = 5$Å



Для элемента $x = (4, 0)$. Стрелками отмечены согласованные экстремумы, попадающие в каталог.

Экстремумы плотности $f(r, \theta)$



Положение экстремумов на сфере меняется в зависимости от расстояния незначительно, что обуславливает структуру каталога.

Переход к угловым переменным в распределении Кента

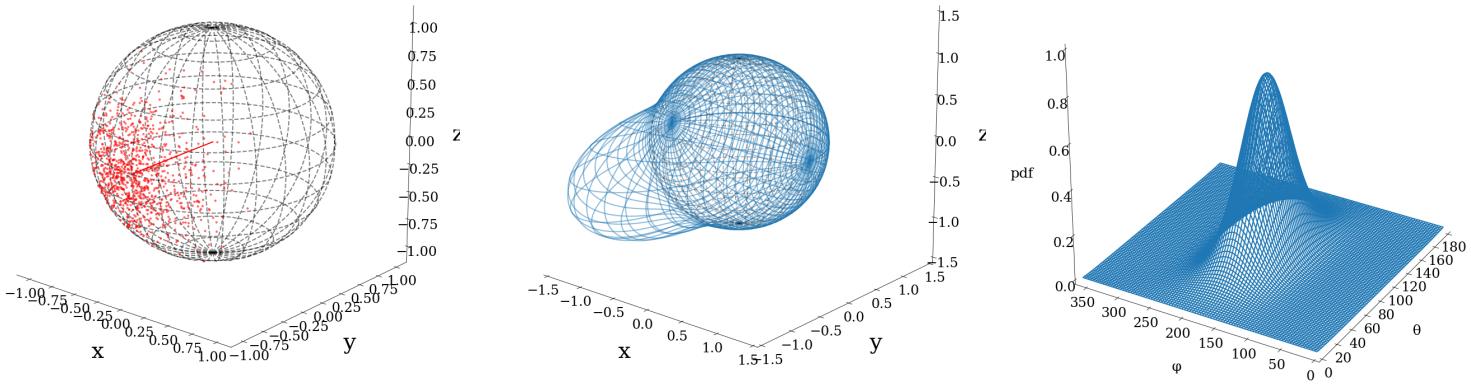
От компонент единичного вектора $[x_1, x_2, x_3]^\intercal$ перейдём к сферическим углам $\theta \in [0, \pi]$, $\varphi \in [0, 2\pi]$:

$$x_1 = \cos\theta, \;\; x_2 = \sin\theta\cos\varphi, \;\; x_3 = \sin\theta\sin\varphi,$$

В таком случае плотность распределения обозначим

$$\mathcal{K}(\theta, \varphi | \kappa, \beta, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3) \text{ или кратко } \mathcal{K}(\theta, \varphi | \boldsymbol{\nu})$$

.



Иллюстрации плотности распределения типичного представителя семейства распределений Кента

Instances of f-divergence

Divergence	Corresponding *f(t)*
KL-divergence	$t \log t$
reverse KL-divergence	$-\log t$
squared Hellinger distance	$(\sqrt{t}-1)^2,\ 2(1-\sqrt{t})$
Total variation distance	$\frac{1}{2}\lvert t-1\rvert$
Pearson χ^2-divergence	$(t-1)^2,\ t^2-1,\ t^2-t$
Neyman χ^2-divergence (reverse Pearson)	$\frac{1}{t}-1,\ \frac{1}{t}-t$
α-divergence	$\begin{cases} \frac{4}{1-\alpha^2}\left(1-t^{(1+\alpha)/2}\right), & \text{if } \alpha \neq \pm 1, \\ t\ln t, & \text{if } \alpha = 1, \\ -\ln t, & \text{if } \alpha = -1 \end{cases}$
Jensen-Shannon Divergence	$(t+1)\log\left(\frac{2}{t+1}\right)+t\log t$
α-divergence (other designation)	$\begin{cases} \frac{t^\alpha - t}{\alpha(\alpha-1)}, & \text{if } \alpha \neq 0,\ \alpha \neq 1, \\ t\ln t, & \text{if } \alpha = 1, \\ -\ln t, & \text{if } \alpha = 0 \end{cases}$

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x)\log\left(\frac{p(x)}{q(x)}\right)\,dx$$

Scoring function: the present state of research

There given P native protein-ligand complexes $C_{i0}, i \in \{1, \ldots, P\}$. For each native (from database) configuration of the complex there exists D non-native configurations (decoys), $j \in \{1, \ldots, D\}$. For each complex C_{ij} we have $D + 1$ conformations, 1 native and D non-native. We have to select a scoring function E, such that

$$E(C_{i0}) < E(C_{ij}), \text{for alternatives } i \in \{1, \ldots, P\}, j \in \{1, \ldots, D\}.$$

Scoring function: a function to represent the integral binding energy; from a ligand native, non-native (a pose to be ranked).

$$E_{ij} = -RT \ln \frac{P^{\text{obs}}_{\text{aa,lig}}(\theta, \varphi, r) + z}{P^{\text{ref}}(\theta, \varphi, r) + z}, \qquad E = \sum_k E_k(\theta, \varphi, r).$$

A weighted variant:

$$E = \sum_i \sum_j r_i c_j E_{ij}(\theta, \varphi, r).$$

Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization by

Tests of the competition CASF-16

Docking power

ability of a scoring function to predict the native or the best near-native docking pose among a set of computer-generated configurations.

Scoring power

correlation of scoring function predictions with the experimental binding affinity data.

Ranking power

ability of a scoring function to correctly rank a set of known ligands for a target protein. In CASF-2016, where five known ligands are available for each target protein, it is measured by Spearman's correlation coefficient.

Screening power

ability of a scoring function to identify true binders for a target protein among a set of small molecules.

Previous results with code

1. For probabilistic models of protein-ligand docking by Nikita Uvarov:
   thesis, slides, code (the project to start with)
2. Mixture of spherical distributions by Sviatoslav Panchenko:
   thesis, slides, code
3. Approximation the empirical distributions with neural networks by Natalia Varenik:
   thesis, slides, code

References

1. CASF Project description page at m1p.org
2. Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization by Maria Kadukova and Sergei Grudinin, 2017
3. Predicting Binding Poses and Affinities in the CSAR 2013–2014 Docking Exercises Using the Knowledge-Based Convex-PL Potential by Sergei Grudinin et al, 2016
4. Comparative Assessment of Scoring Functions: The CASF-2016 Update by Minyi Su et al., 2018
5. Docking rigid macrocycles using Convex-PL, AutoDock Vina, and RDKit in the D3R Grand Challenge 4 Maria Kadukova et al., 2020
6. Basic solution is a linear combination; it is described in draft

Docking: a residue, its frame and an atom