Построение вероятностных метрических пространств в задачах анализа молекулярных конфигураций

Никита Денисович Уваров

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Кафедра «Интеллектуальные системы»

Научный руководитель: д.ф.-м.н. В.В. Стрижов

Выпускная квалификационная работа магистра

Москва 2020

Мотивация для построения пространства

Задача молекулярного докинга (CASF)

Ранжирование синтезированных молекулярных комплексов (конформаций) по энергетической устойчивости.

Проблема

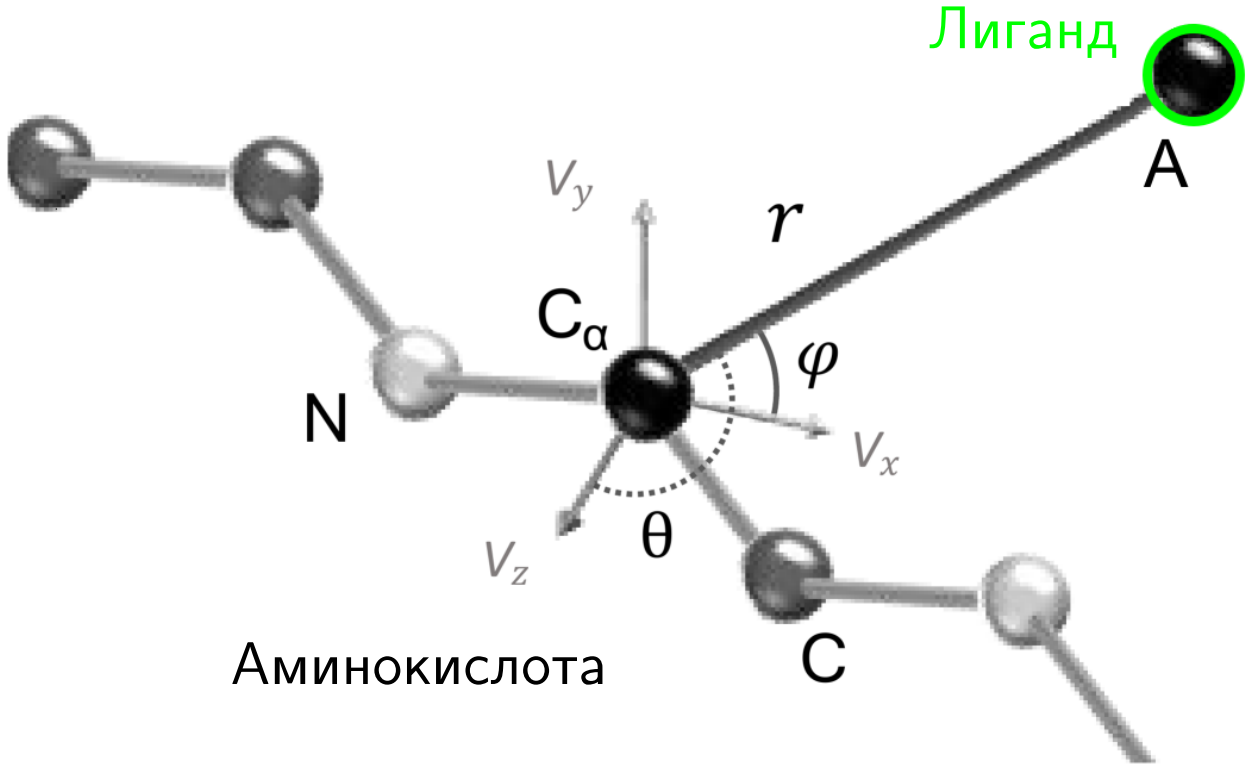
Существующие подходы моделируют физический потенциал взаимодействия с привлечением данных из множества источников и используют избыточно сложные модели.
Предлагается

1. разбить молекулярные комплексы на элементарные взамодействующие пары аминокислота — лиганд,
2. построить модели вероятностных распределений взаимного расположения элементарных пар в $\mathbb{R}^3$,
3. использовать метрические методы в пространстве полученных распределений.

Литература

- **José Ramón López-Blanco, Pablo Chacón, 2019.** KORP: knowledge-based 6D potential for fast protein and loop modeling. Bioinformatics.
- **Maria Kadukova, Sergei Grudinin, 2017.** Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization. Journal of Computer-Aided Molecular Design.
- **Maria Kadukova, Sergei Grudinin, 2018.** Docking of small molecules to farnesoid X receptors using AutoDock Vina with the Convex-PL potential: lessons learned from D3R Grand Challenge 2. Journal of Computer-Aided Molecular Design.

Геометрия взаимодействия аминокислоты и лиганда



Аминокислота

Локальная система координат

$$\vec{V}_z = \frac{\overrightarrow{C_\alpha C} + \overrightarrow{C_\alpha N}}{|\overrightarrow{C_\alpha C} + \overrightarrow{C_\alpha N}|}, \quad \vec{V}_y = \frac{V_z \times \overrightarrow{C_\alpha N}}{|\vec{V}_z \times \overrightarrow{C_\alpha N}|}, \quad \vec{V}_x = \vec{V}_y \times \vec{V}_z, \quad \vec{r} = \overrightarrow{C_\alpha A}.$$

Метрическое пространство $\mathcal{M}_\mu$

Объекты пространства

Взаимное расположение аминокислоты $a \in \mathcal{A}$ и лиганда $b \in \mathcal{B}$ определяется вектором $\vec{r} \in \mathbb{R}^3$ или тройкой (r, φ, θ),

$$|\mathcal{A}| = 20, \ |\mathcal{B}| = 40.$$

В пространстве $\mathcal{M}_\mu$ всевозможных вероятностных распределений на $\mathbb{R}^3$ паре $x = (a, b)$ соответствует элемент $p_{a,b} = p_x(r, \varphi, \theta)$.

Метрика

$$d_\mu(x, y) = D_\mu(p_x, p_y) = \int_{\mathbb{R}^3} \mu\left(\frac{p_x(\vec{r})}{p_y(\vec{r})}\right) p_x(\vec{r}) d\vec{r}$$

Метрика: $\mu(t) = \frac{1}{2}|t - 1|$, $\mu(t) = 2(1 - \sqrt{t})$.
Предметрика: $\mu(t) = t \log t$ (KL).

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:

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

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

Семейство моделей $\mathfrak{F}$

1. Гистограмма:

$$p(\vec{r}) = \frac{1}{N}\sum_{t=1}^{N}[\vec{r_t} \in I_{ijk}],\ i_{\vec{r}} = \left\lfloor \frac{r}{dr} \right\rfloor,\ j_{\vec{r}} = \left\lfloor \frac{\varphi}{d\varphi} \right\rfloor,\ k_{\vec{r}} = \left\lfloor \frac{\theta}{d\theta} \right\rfloor.$$

2. Смесь Гауссиан из N компонент:

$$p(\vec{r}) = \sum_{i=1}^{N} w_i \frac{1}{\sqrt{(2\pi)^N |\Sigma_i|}} \exp\left(-\frac{1}{2}(\vec{r}-\vec{\mu_i})^T \Sigma_i^{-1}(\vec{r}-\vec{\mu_i})\right), \sum_{i=1}^{N} w_i = 1.$$

3. Нейросеть:

$$p(\vec{r}) = W_3 \cdot \sigma(W_2 \cdot \sigma(W_1 \cdot \vec{r})), \quad \sigma(\vec{x})_i = \tanh(x_i),$$

$$W_1 \in \mathbf{M}_{50x3}(\mathbb{R}), W_2 \in \mathbf{M}_{50x50}(\mathbb{R}), W_3 \in \mathbf{M}_{1x50}(\mathbb{R}).$$

Вычислительный эксперимент

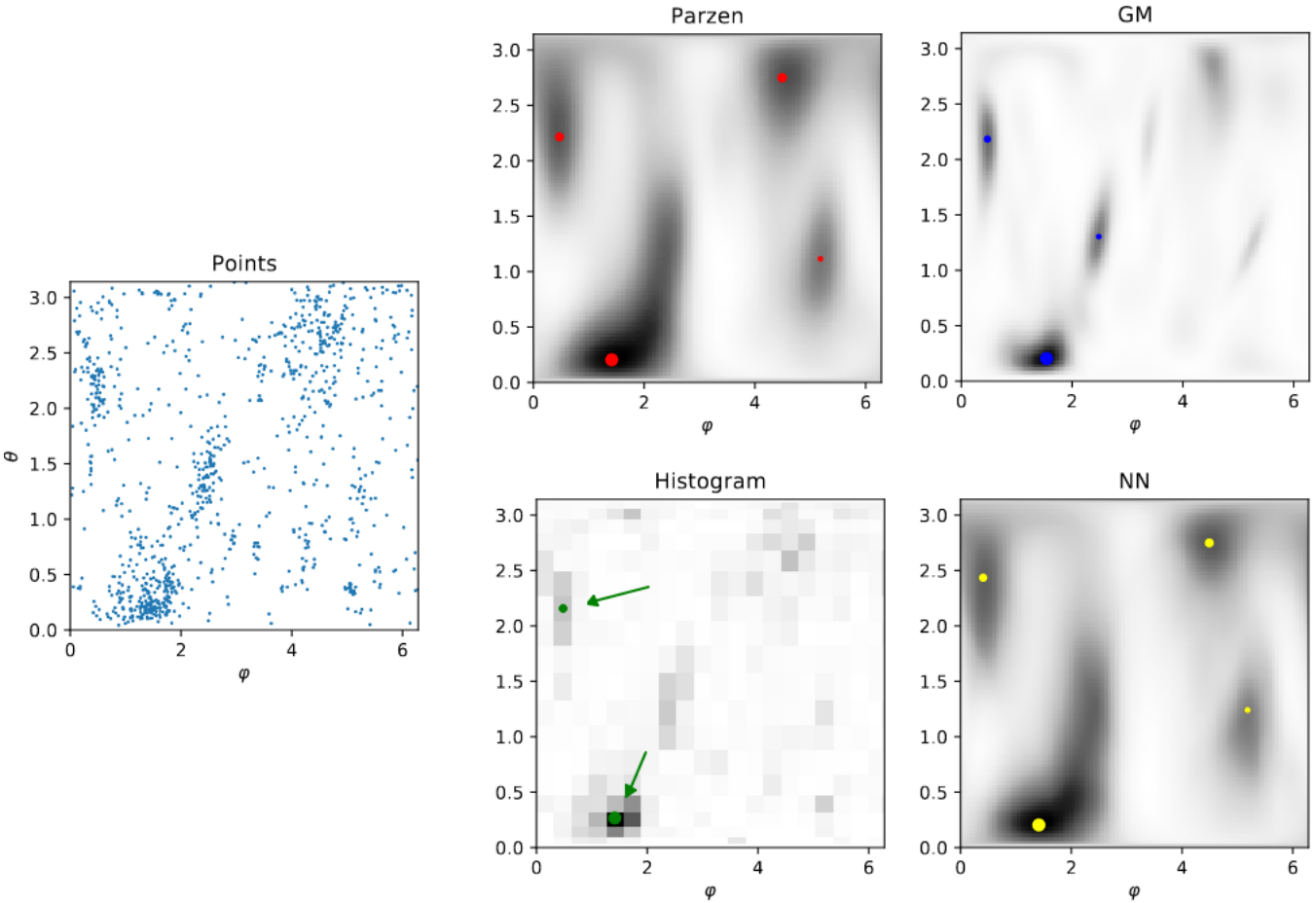
Цель

- Восстановить плотность $p_x(r, \varphi, \theta)$ различными моделями.
- Построить каталог экстремумов плотности $(a, b, (r_{min}, r_{max}], \varphi, \theta)$.
- Построить таблицу средней согласованности распределений $L_1(a, b), L_2(a, b)$.
- Ранжировать аминокислоты по средней согласованности $L_i(a)$ и лиганды по средней согласованности $L_i(b)$.

Данные

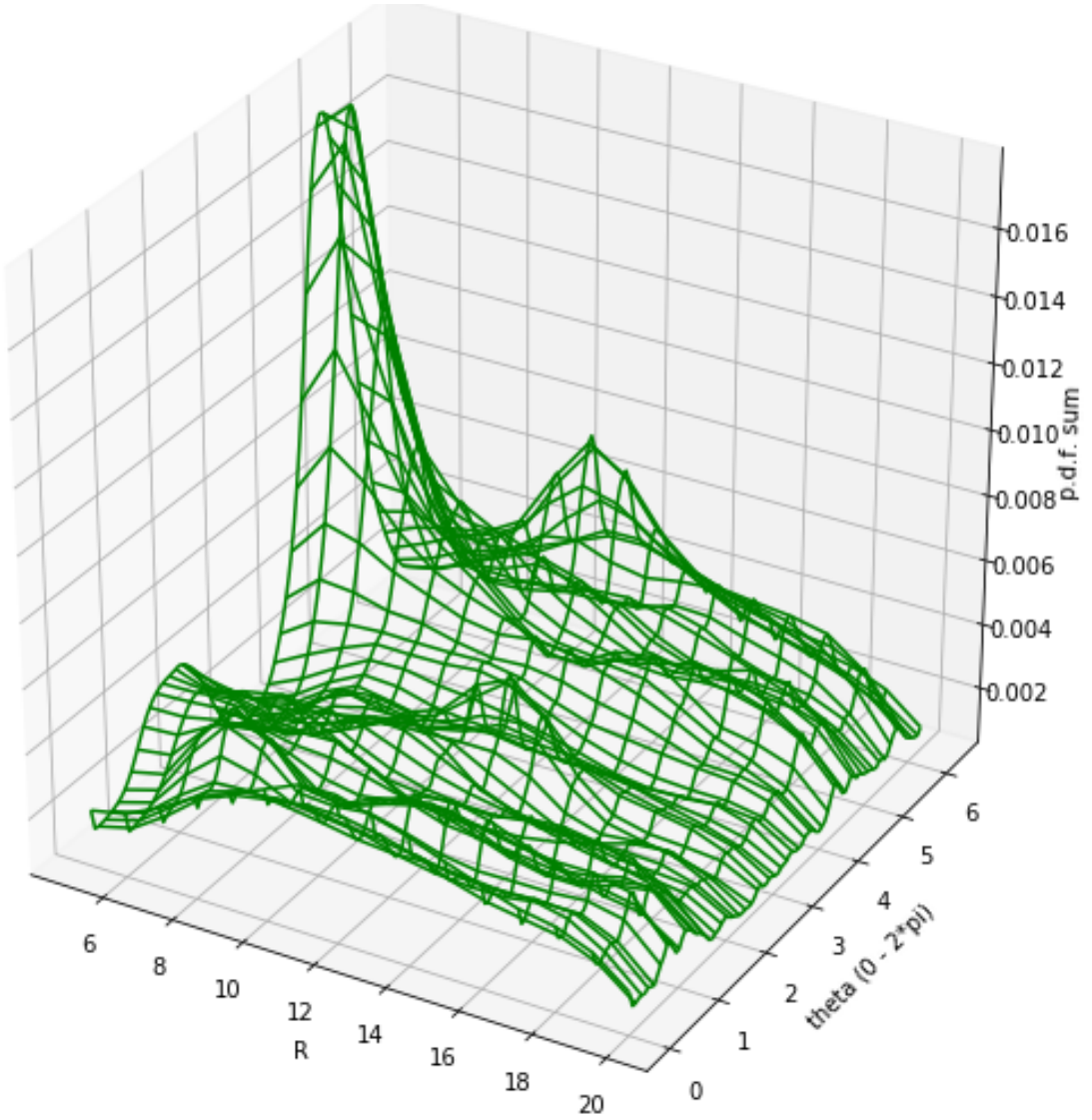
Наблюдения пространственного расположения 40 белков $\times$ 20 лигандов, $5 \cdot 10^7$ записей $(a, b, r, \varphi, \theta)$. $r_{min} = 4$Å, $r_{max} = 20$Å.

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



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

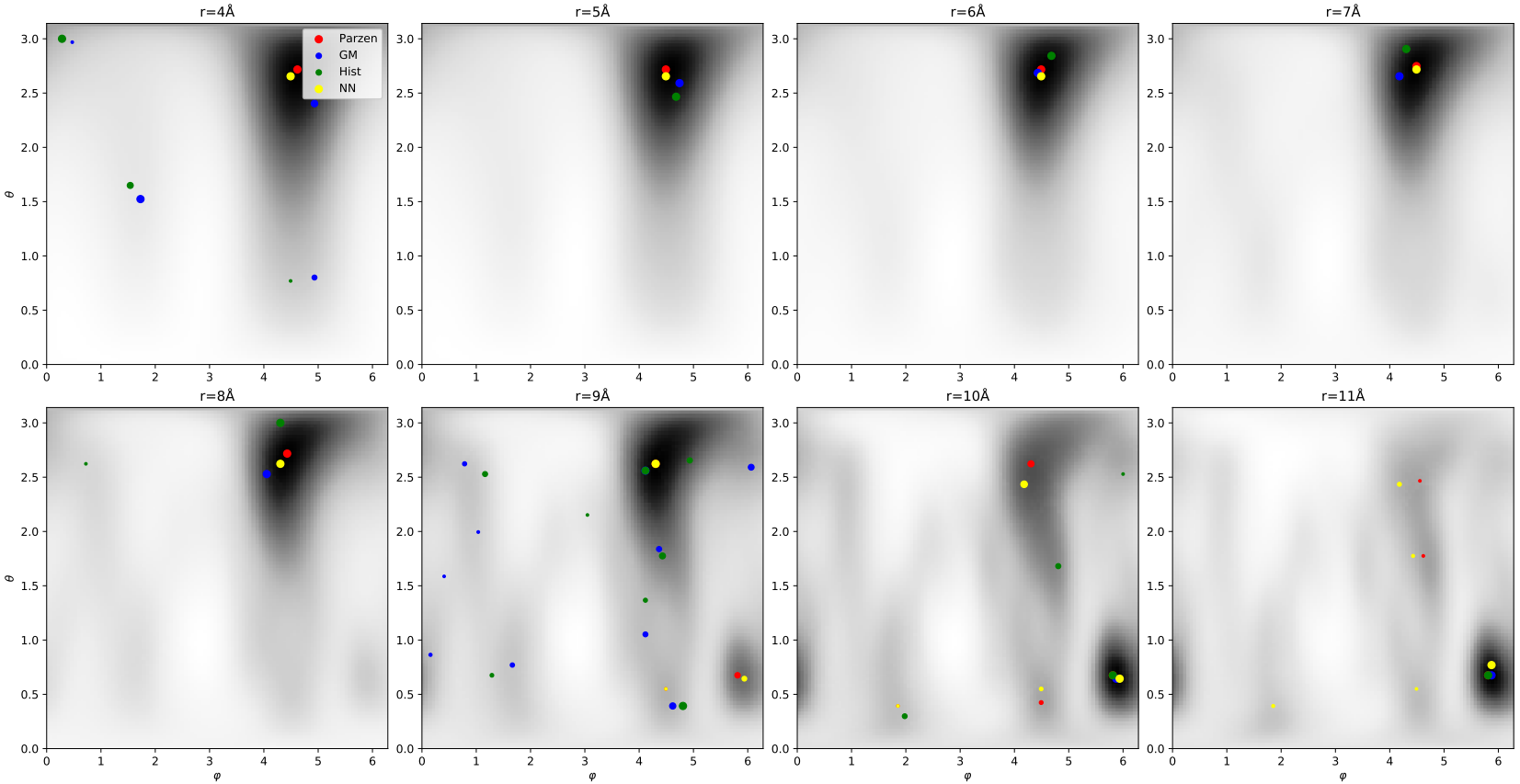
Экстремумы плотности $f(r, \theta)$



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

Каталог экстремумов (1)

Protein 0, ligand 2

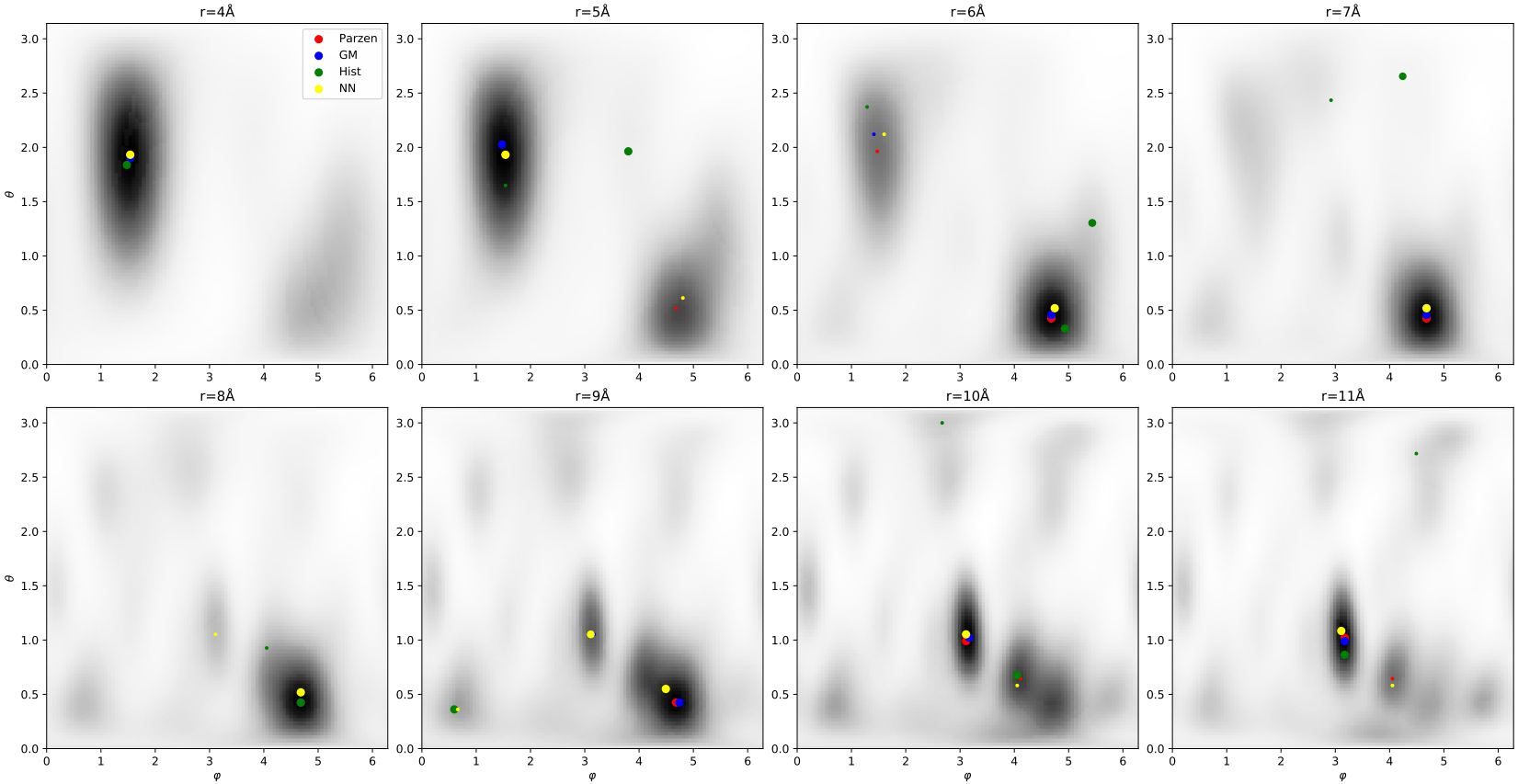


	Parzen	GM	Hist	NN
Parzen				
GM	28.98			
Hist	34.78	23.33		
NN	33.72	54.28	58.24	

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, %

Применение для задачи ранжирования CASF

Постановка

Данные: набор конформаций $\{z_i\}_{i=1}^{\ell}$, $z_i = \{(a, b, r, \varphi, \theta)\}_{j=1}^{\ell_i}$, для обучающей выборки известны истинные энергии E_i.
Критерий качества:

$$P@K = \frac{1}{K} \sum_{i:\text{Rank}(E_i)\leqslant K} [\text{Rank}(\widetilde{E}_i) \leqslant K], K = 5,$$

$\text{Rank}(\widetilde{E}_i)$ - позиция $\widetilde{E}_i$ в порядке по возрастанию $\widetilde{E}_i$.

Оценка устойчивости синтетической конформации

1. Восстанавливаем для пар (a, b) взаимодействующих элементов конформации z_i распределения $\widetilde{p}_{a,b}(z_i) \in M_\mu$.
2. Используем в качестве оценки устойчивости

$$\widetilde{E}_i = \sum_{a\in\mathcal{A}, b\in\mathcal{B}} D_\mu(\widetilde{p}_{a,b}(z_i), p_{a,b}).$$

Результаты

- Плотность вероятности взаимодействия протеин-лиганд восстановлена различными моделями (окно Парзена, гистограмма, смесь гауссиан, нейросеть).
- Построен каталог экстремумов плотности и введена метрика согласованности экстремумов.
- Отвергнута гипотеза о равномерности плотности на больших расстояниях.
- На основе полученных распределений построено метрическое пространство для использования в задаче молекулярного докинга.

Публикация ВАК

Н.Д. Уваров, М.П. Кузнецов, А.С. Малькова, К.В. Рудаков, В.В. Стрижов. Выбор суперпозиции моделей при прогнозировании грузовых железнодорожных перевозок // Вестник Моск. ун-та. Сер.15. Вычисл. матем. и киберн., 2018.