

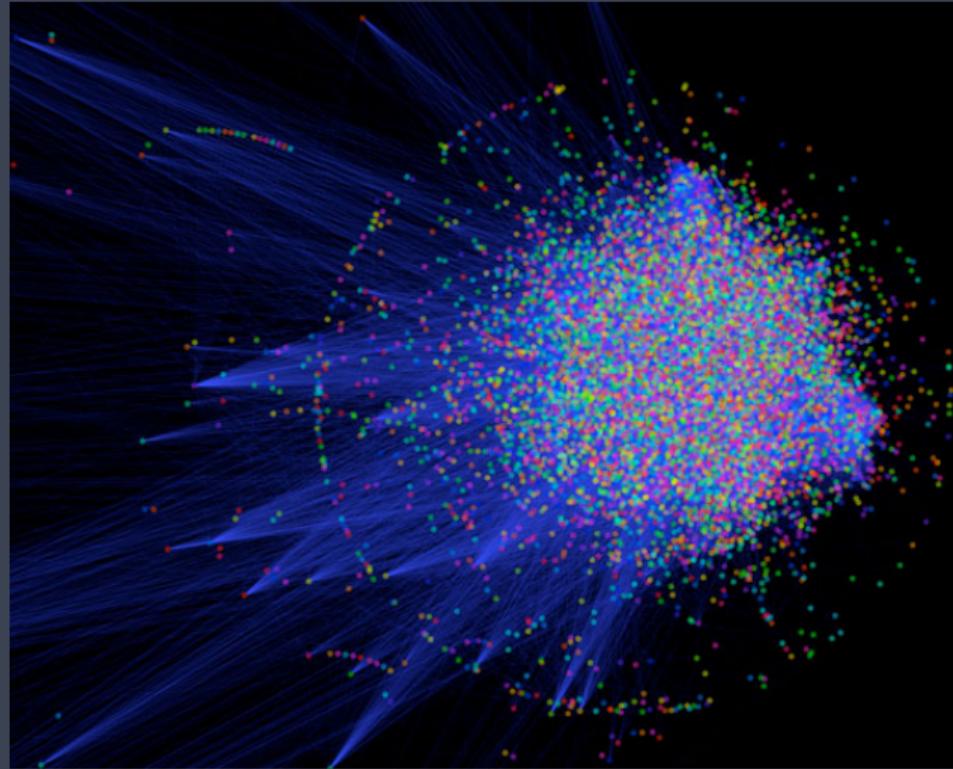
# **Лекция 8. Дизайн белок-белковых взаимодействий**

Курс: Методы машинного обучения в дизайне белков

**by** Головин А.В. <sup>1</sup> ( <sup>1</sup>МГУ им М.В. Ломоносова, Факультет Биоинженерии и  
Биоинформатики )

**on** Москва, 2024

# » Human "Interactome"



## » Особенности белок-белковых контактов

- \* Поверхность интерфейса большая,  $1000\text{-}2000 \text{\AA}^2$
- \* Только 5% остатков дают ключевой вклад в связывание
- \* Экспериментальный поиск затруднен
- \* Широкое разнообразие



## » Способы предсказания белок-белковых взаимодействий

Взаимодействующие белки возможно ко-эволюционируют.

- \* Филогенетический профайлинг.  
Поиск пар белковых семейств среди широкого ряда видов. Появление и исчезновение пар семейств возможно указывает на взаимодействие.
- \* Предсказание на основе подобия филогенетических деревьев.
- \* Методы на основе классификации.
- \* Поиск гомологичных мест контакта.
- \* Ассоциативные методы. Это поиск характеристических последовательностей на основе профилей и мотивов.



## » Способы предсказания белок-белковых взаимодействий

- \* Идентификация структурных паттернов на основе известных структурных данных.  
Построение библиотеки и сканирование по ней.
- \* Методы Байеса для анализа экспериментальных результатов с значимым уровнем шума.
- \* Методы исключения доменных пар.
- \* Моделирование структуры комплекса на основе известной структуры и оценка его качества.
- \* Макромолекулярный докинг.



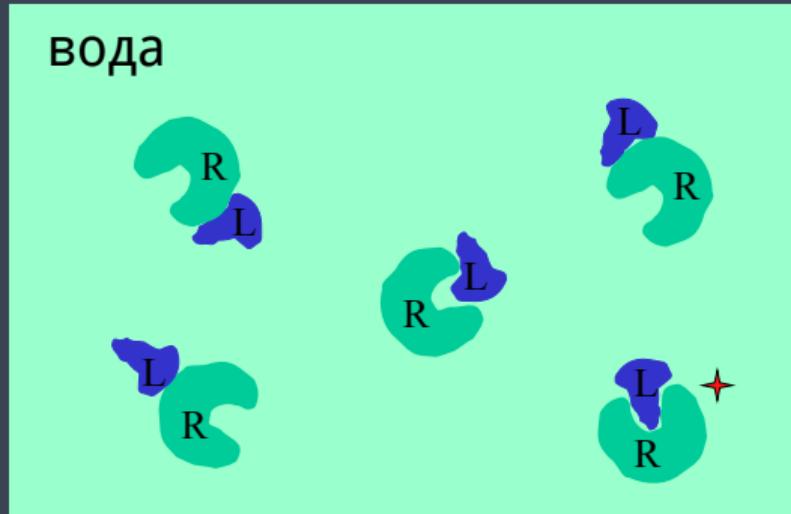
## » Базы данных

- \* **String** - база данных экспериментальных и предсказанных взаимодействий; отличная графика; <http://string-db.org/>
- \* **IntAct** - база данных на основе литературных данных или прямая информация от авторов. <http://www.ebi.ac.uk/intact/>
- \* **iHOP** - Информация слинкованая с другими белками. Построена на основе литературных данных. Представление в виде кусочков текста. <http://www.ihop-net.org/>
- \* **BioGRID** - Источники: литература и результаты high-throughput экспериментов; <http://thebiogrid.org/>
- \* **MIPS** Mammalian Protein-Protein Interaction Database, не работает :). <http://mips.helmholtz-muenchen.de/proj/ppi>

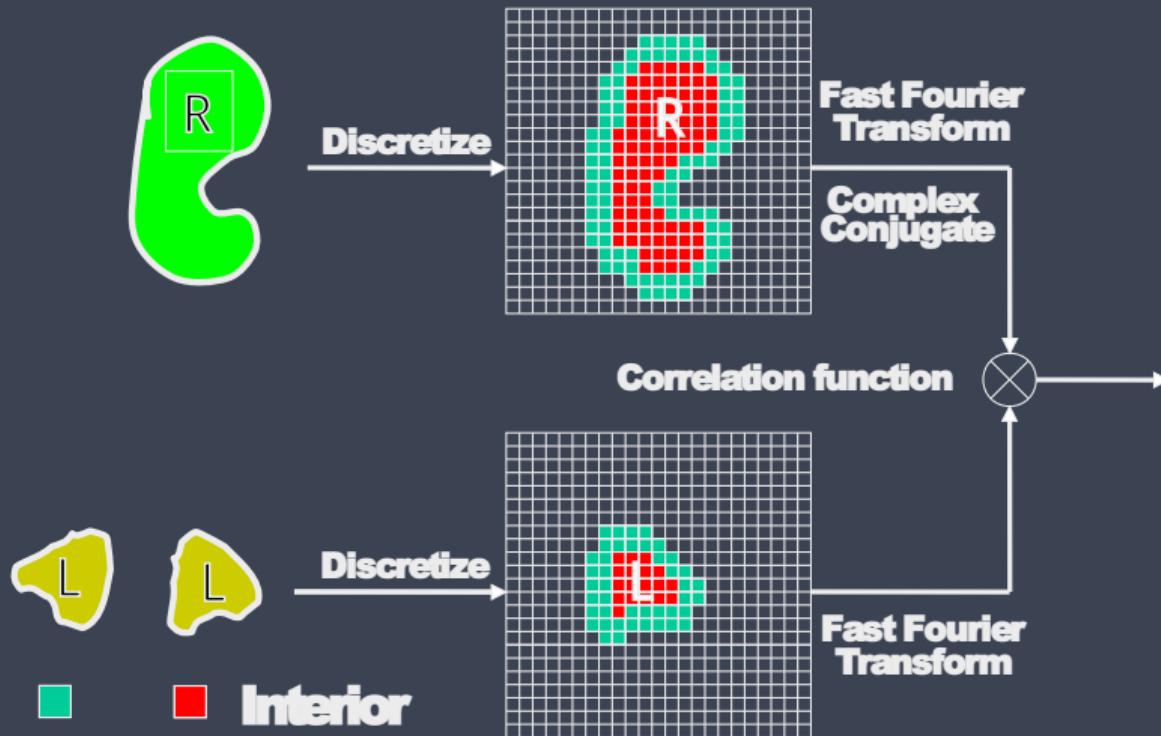


## » Поиск наименьшего $\Delta G$

- \* Суть метода основывается на поиске соответствия поверхностей для достижения максимальной поверхности контакта.
- \* После нахождения возможных конфигураций происходит ранжирование.



## » Белковый докинг с использованием FFT

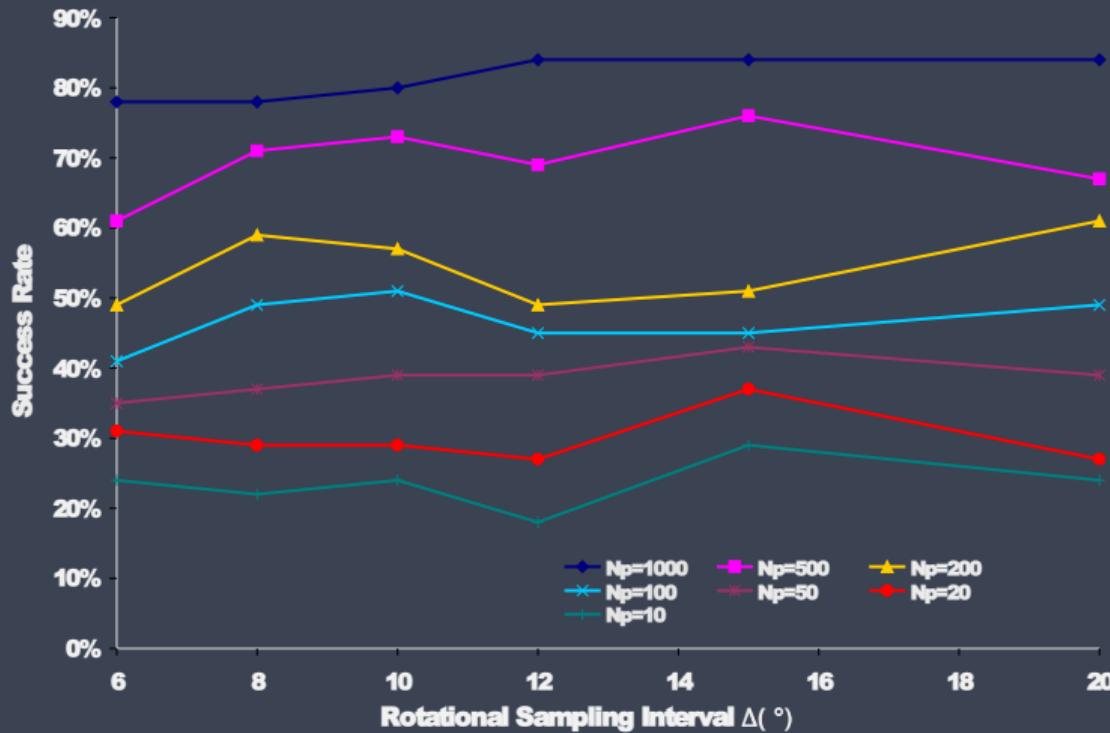


## » Оценка производительности

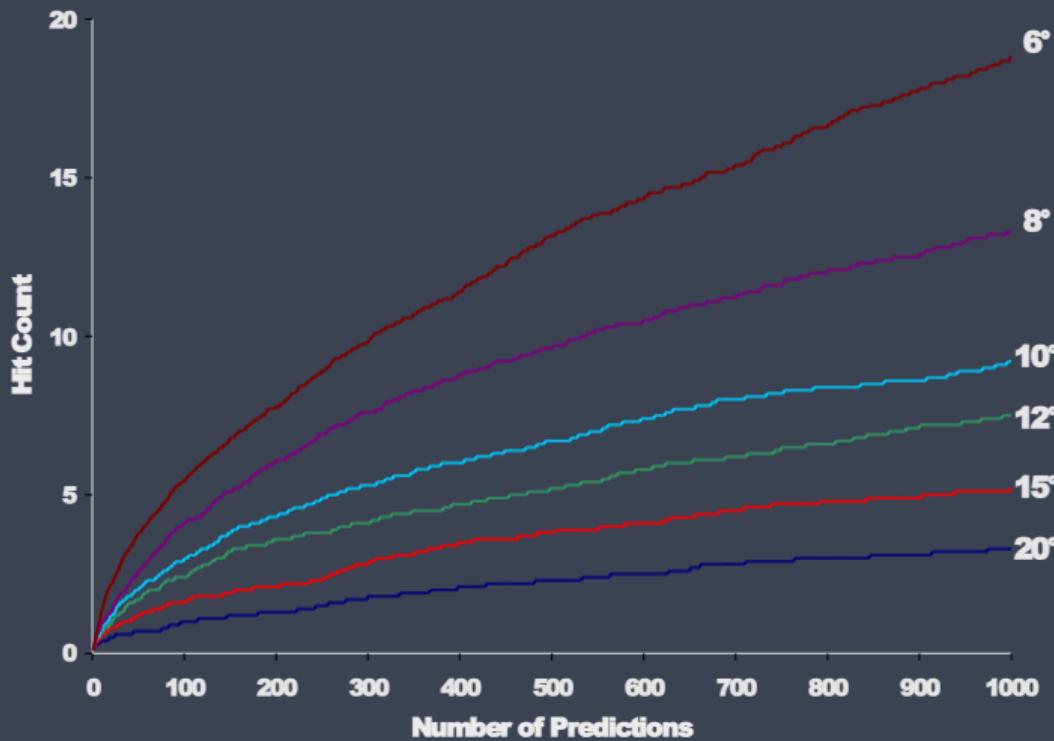
- \* Success Rate: для данного количества предсказаний ( $N_p$ ), это процент структур для которых был найден как минимум один удачный результат
- \* Hit Count: среднее количество хитов при данном значении  $N_p$ .



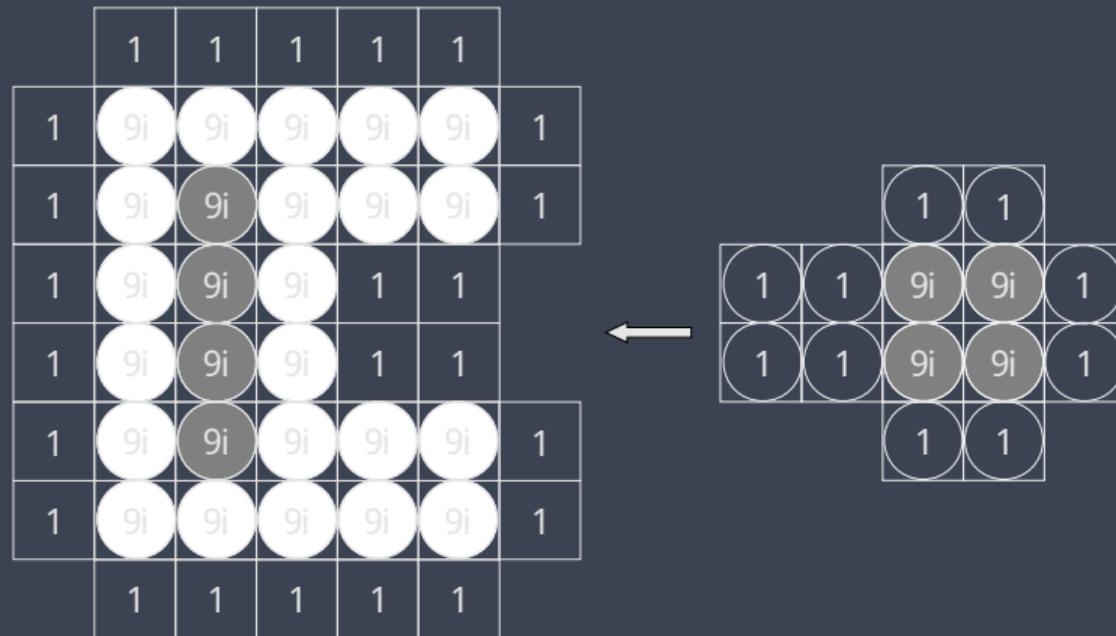
## » Зависимость Success Rate от шага вращения



## » Зависимость Hint count от шага вращения



## » Решеточная комплементарность поверхности

 $R_{GSC}$  $L_{GSC}$ 

# » Парная комплементарность поверхности

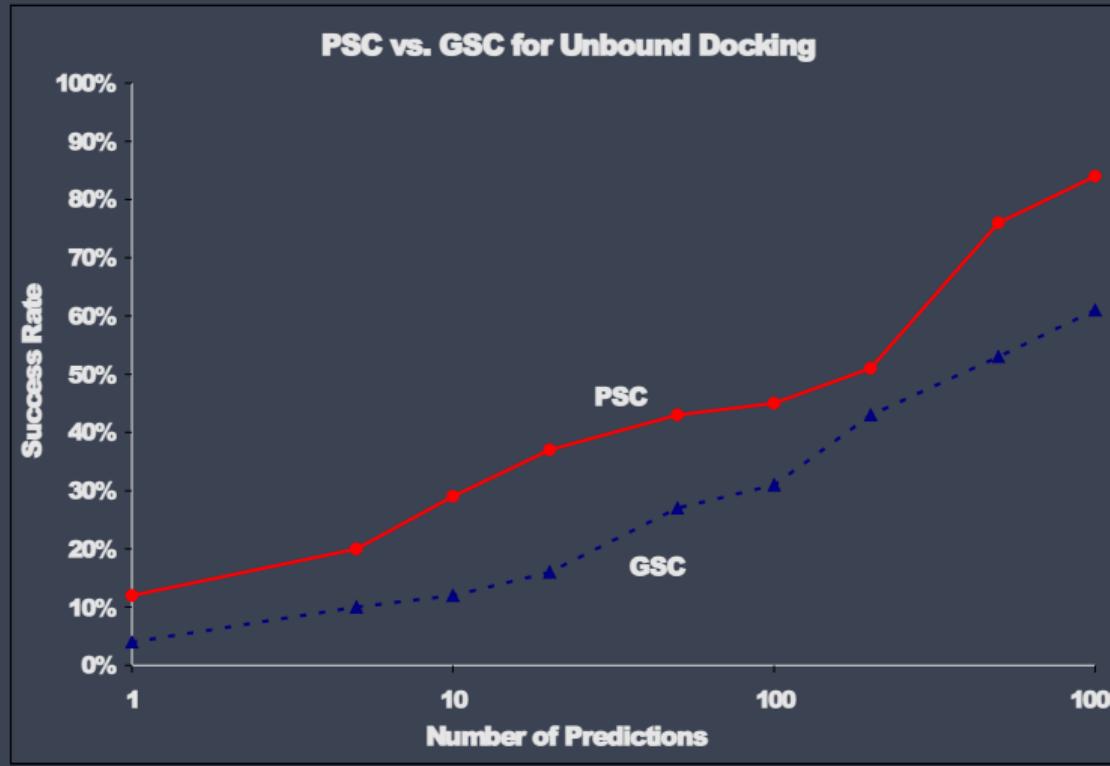
1	2	3	3	3	2	1
2	3i	3i	3i	3i	3i	2
3	3i	9i	3i	3i	3i	2
3	3i	9i	3i	5	2	1
3	3i	9i	3i	5	2	1
3	3i	9i	3i	3i	3i	2
2	3i	3i	3i	3i	3i	2
1	2	3	3	3	2	1



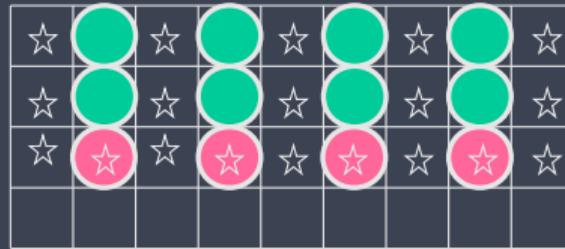
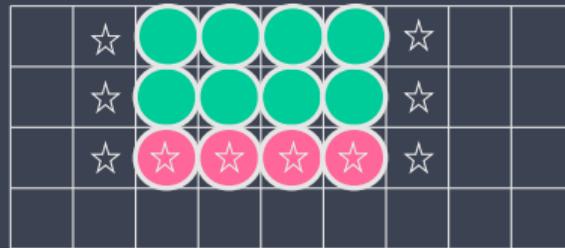
1+3i	1+3i
1+3i	1+3i
1+9i	1+9i
1+9i	1+3i
1+3i	1+3i

**R<sub>PSC</sub>****L<sub>PSC</sub>**

## » PSC vs. GSC и Success Rate



# » Почему так?



# » ФУНКЦИЯ ОЦЕНКИ ЭНЕРГИИ СВЯЗЫВАНИЯ

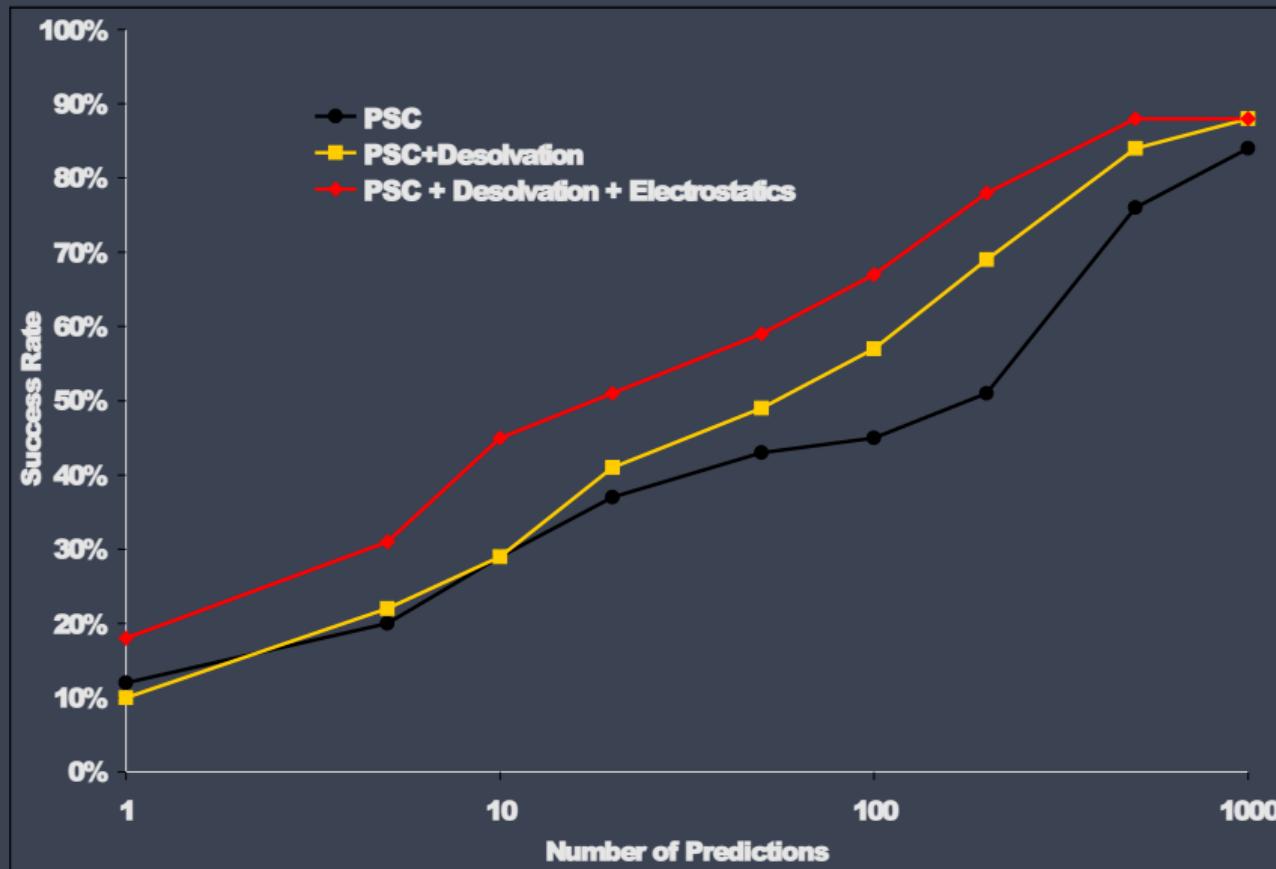
$$\Delta G = \Delta E_{VdW} + \Delta E_{el.} + \Delta G_{desol} + \Delta G_{const}$$

- \*  $\Delta E_{VdW}$  - это комплементарность поверхности
- \*  $\Delta G_{desol}$  - это гидрофобика
- \*  $\Delta E_{el.}$  - это электростатика
- \*  $\Delta G_{const}$  - это изменение вращательной и прочих энтропий.

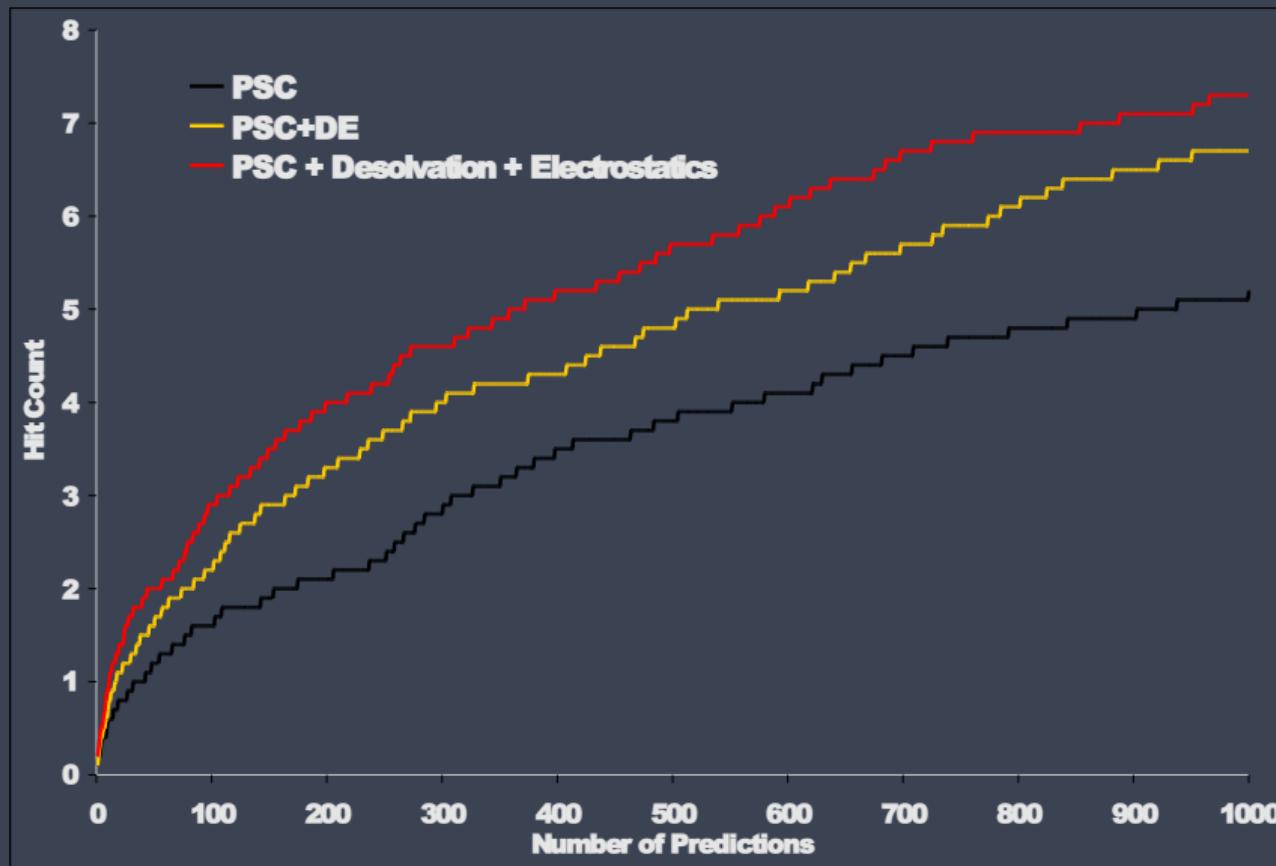
$$\Delta G_{desol} = \sum_i \sum_j N_{ij} \Delta G_{ij}$$



## » Влияние на Success rate



## » Влияние на Hit Count



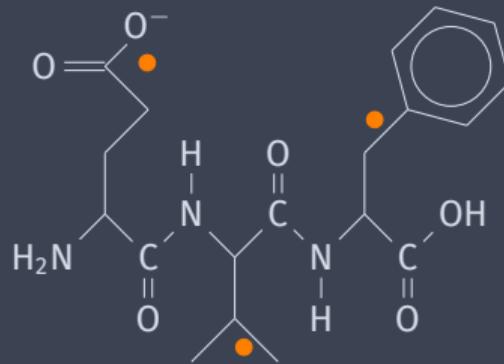
## » Rosettadock Алгоритм



# » Поиск с низким разрешением

Особенности метода:

- \* Поиск методом Монте-Карло.
- \* Вращение и смещение белка как жесткого тела.
- \* Остаток белка представляется как атомы остова и средний центроид представляет боковой радикал.
- \* Процедура старается воспроизвести физическую диффузию.



## » Уточнение с высоким разрешением

- \* Из библиотеки ротамеров добавляются полноатомные боковые цепи
- \* Используется полноценная оценка энергии (ММ)
- \* Монте-Карло + оптимизация геометрии
- \* Циклическое использование оптимизации положения как твердого тела и полноатомная оптимизация положения боковых радикалов

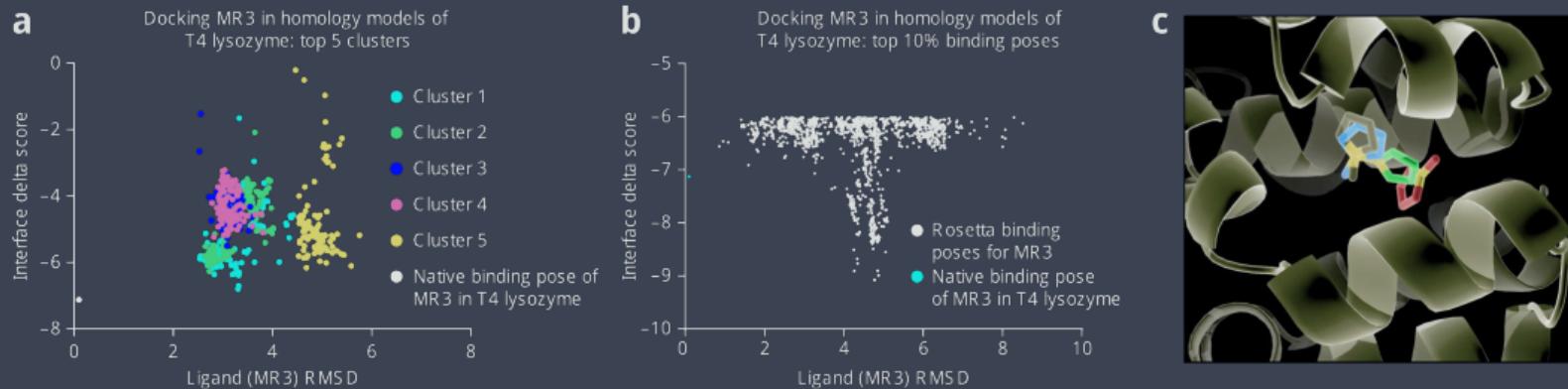


## » Rosettadock Алгоритм



# » Пример докинга лигнада

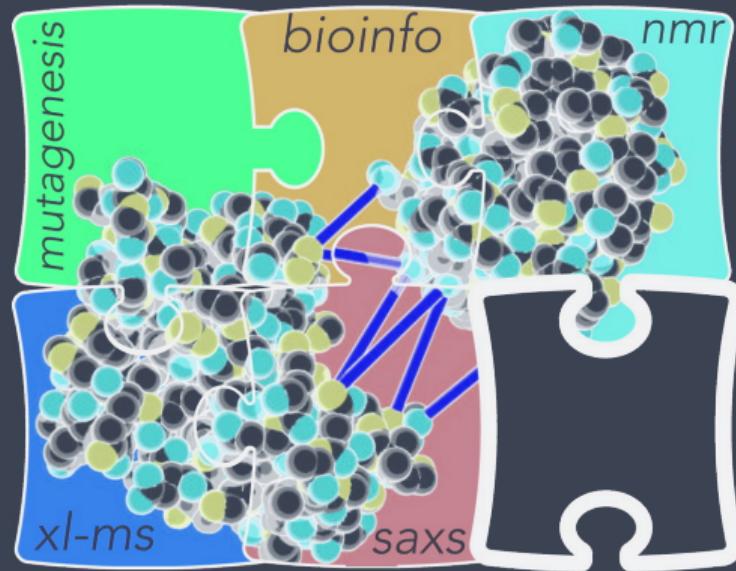
## protocol



**Figure 5** | Docking MR3 into comparative models of T4 lysozyme. The MR3 ligand was docked into the ten lowest-energy comparative models of T4 lysozyme, as detailed in Steps 17–22 of the protocol. (a) 10,000 binding modes were clustered by RMSD using applications available in the bcl::Commons. The largest five clusters are shown, with the interface\_delta score plotted against the RMSD to the native ligand-binding mode (shown in black). Generally, the largest clusters are also those with the lowest RMSD to the native binding mode. (b) The RMSD between 10,000 binding modes and the native binding mode (shown in red) was computed. The top ten percent of models by interface\_delta score are shown here. Sub-angstrom binding modes are within the top ten percent of models, but Rosetta also identifies an alternative lower-energy binding mode within the site. (c) The lowest RMSD binding mode (orange) is closer to the native binding mode (gray) compared with the lowest-energy binding mode of the largest cluster (magenta) and the lowest-energy binding mode overall (cyan).



## » HADDOCK



**ADDOCK**  
High-Ambiguity Driven Docking



## » История

- \* Первые работы в 2001 году
- \* Суть задачи: для двух данных последовательностей предсказать взаимодействие
- \* Типы представлений: состав, доменный состав, мотивы, профили гидрофобности, геномные особенности, филогенетические особенности



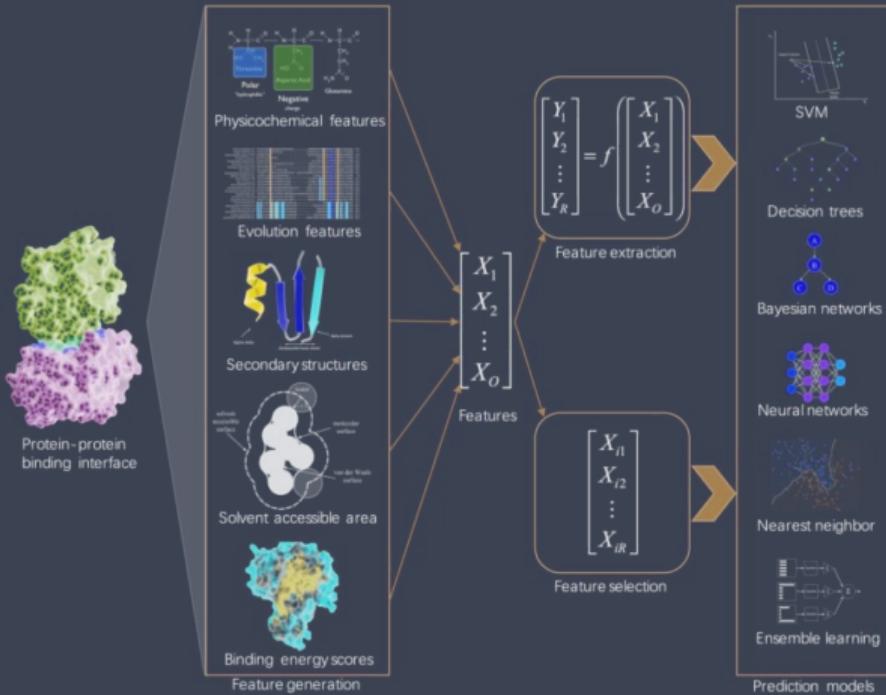
## » Типы подходов

- \* Обучение с учителем: NN, Баевские методы, SVM, RF,
- \* Кластеризация

\*Наборы представлений, не могут полностью охватить динамические и сложные явления, которые могут однозначно идентифицировать истинные IPP



# » Поиск "hot spots"



10.3390/molecules23102535

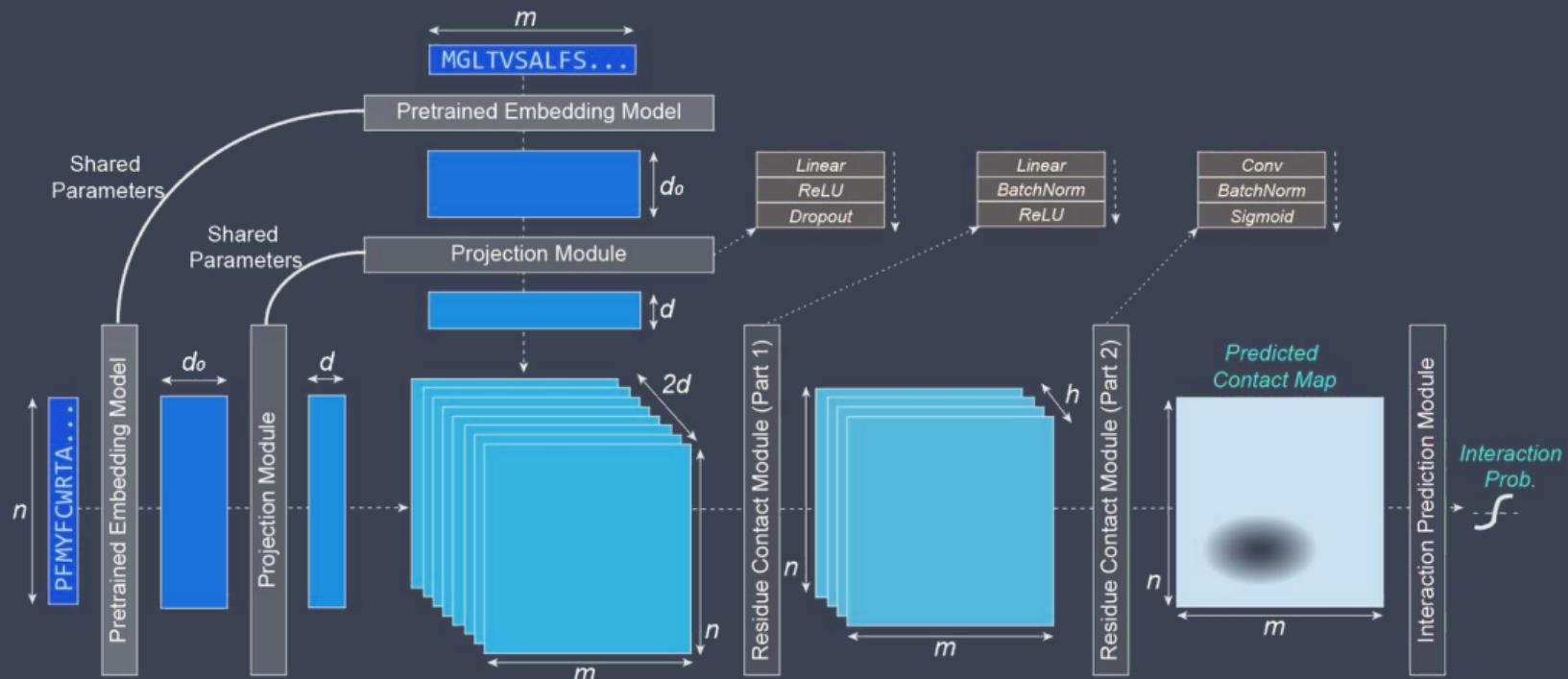


## » Достижения

- \* Существенный прогресс, но есть еще "вызовы"
- \* ddG из экспериментов не унифицированно
- \* Малое наполнение данными
- \* Часто случается переобучение
- \* Предикторы 'hot spots' используют последовательность и структурную информацию для представления, но 3D не используется полностью
- \* \*Перспективным считается интеграция физических методов (докинг, МД) и ML



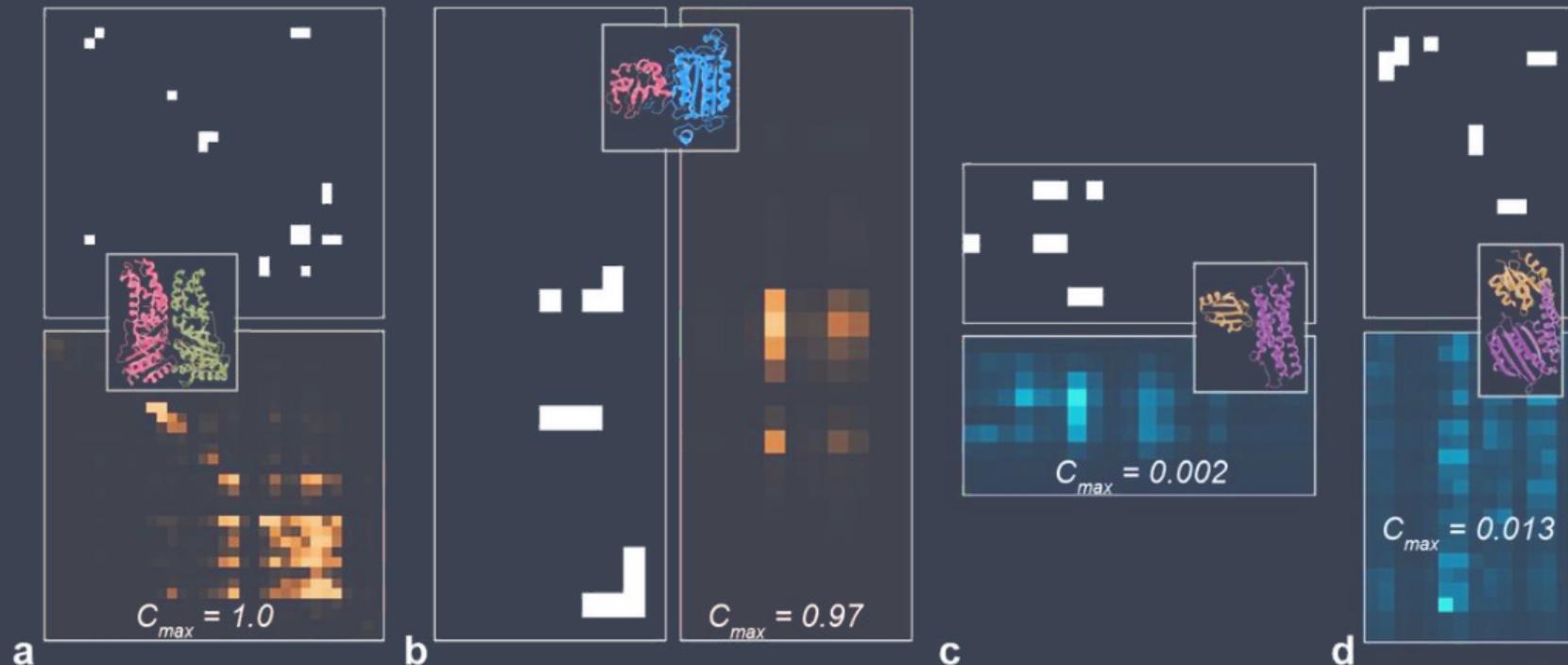
# » D-SCRIPT



10.1016/j.cels.2021.08.010



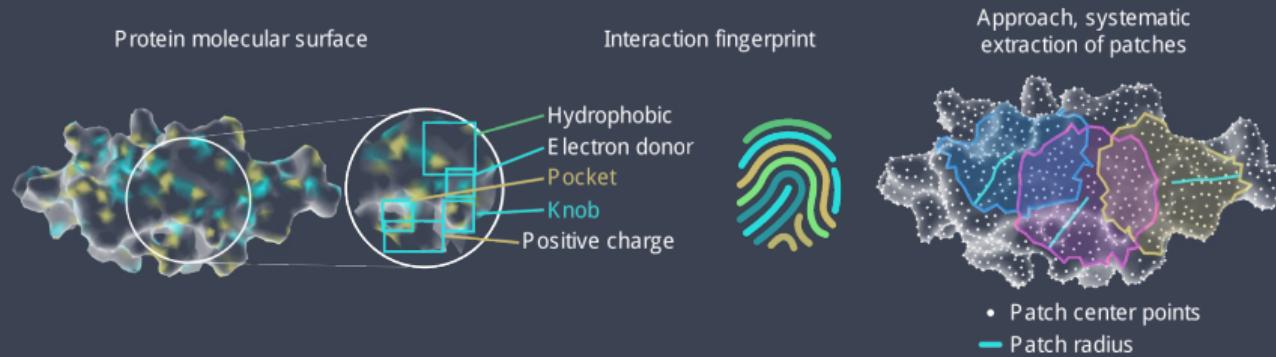
# » D-SCRIPT, результат



\*When D-SCRIPT correctly predicts an interaction, its contact maps are significantly similar to the ground truth.



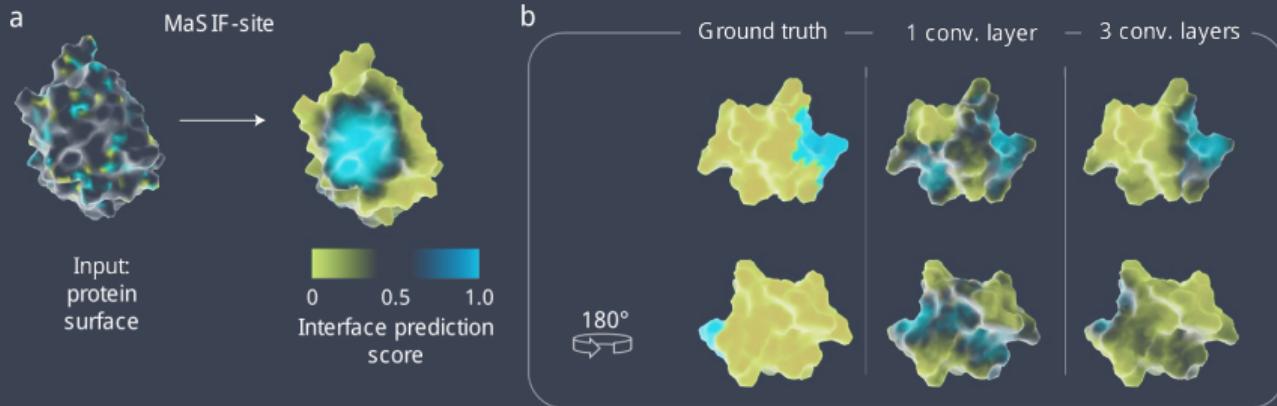
## » MASIF



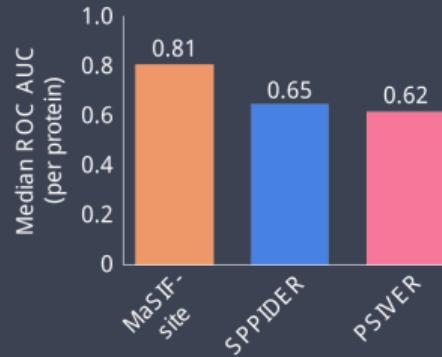
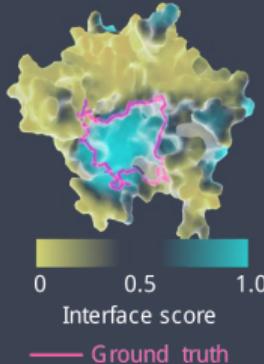
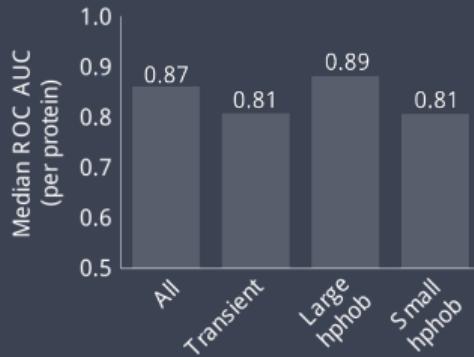
MaSIF-site: классификатор, на входе поверхность белка, на выходе прогнозируемая оценка для каждой вершины поверхности на вероятность участия в PPI



# » MASIF, предсказание участков



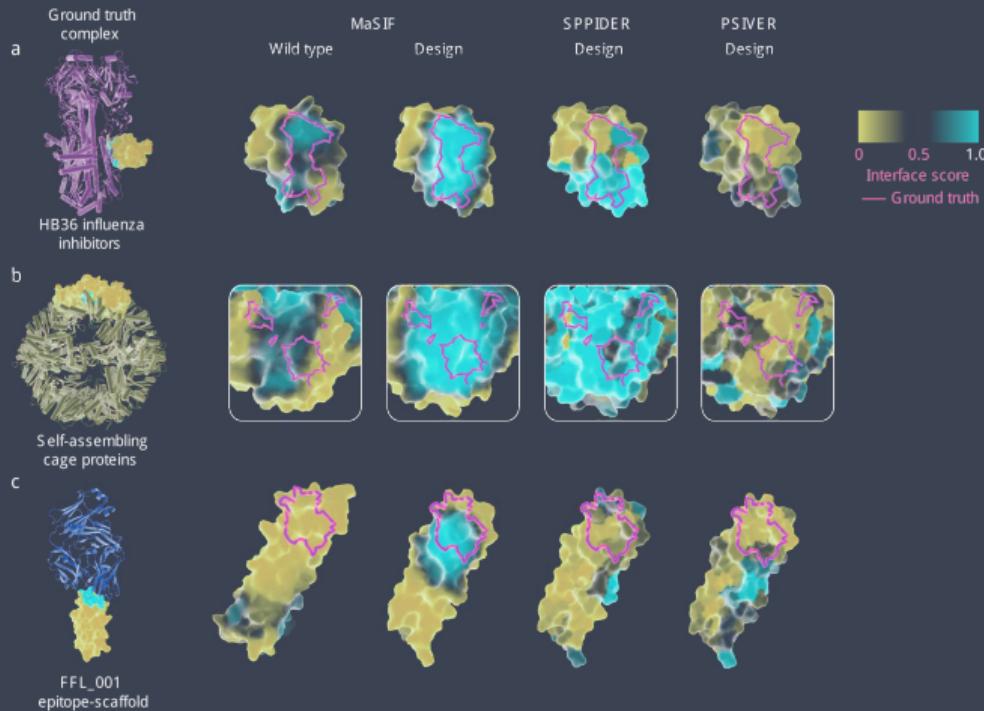
# » MASIF, сравнение



- \* PSIVER - Naïve Bayes classifier (NBC) and a kernel density estimation method (KDE)
- \* SPIDER - SAS метрика + MSA,



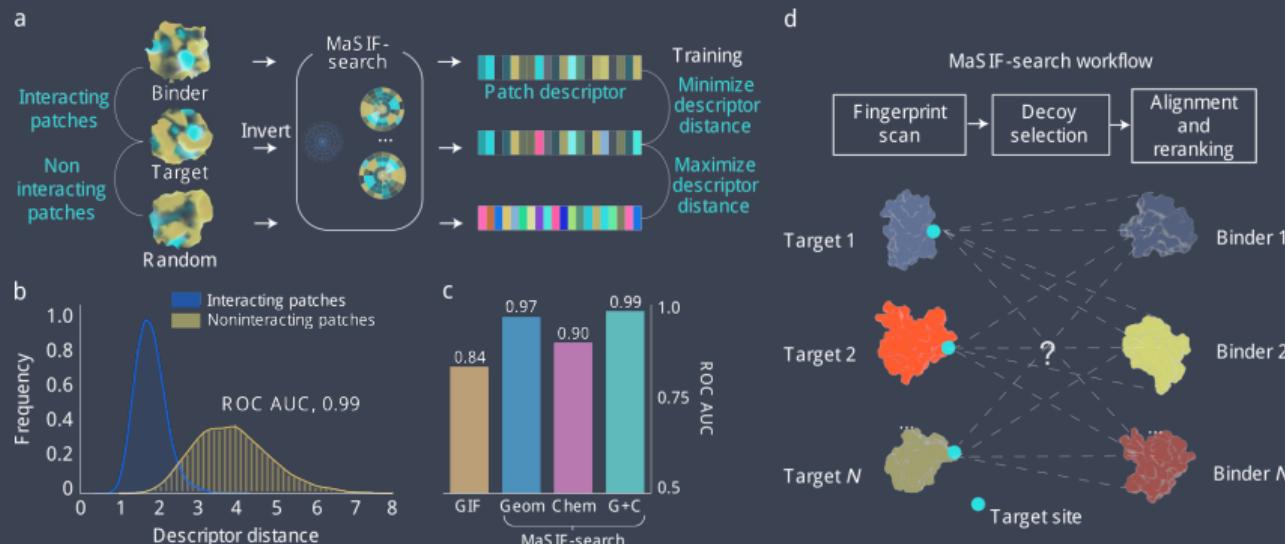
# » MASIF, предсказание для "новых" белков



\*MaSIF-site clearly labels the interfaces of the designs



# » Ultrafast scanning with MaSIF



\*MaSIF-search inverts the numerical features of one protein partner (multiplied by -1), with the exception of hydrophathy.



# » "FATALITY" or not?

**Table 1 | Results for large-scale docking benchmark for PatchDock, MaSIF-search (with multiple numbers of decoys), ZDock and ZDock+ ZRank2 on bound (holo) complexes**

Method	Number of solved complexes in the top			time (min)
	100	10	1	
MaSIF-search decoys = 100	37	36	30	4
MaSIF-search decoys = 2,000	67	56	43	39
PatchDock	43	32	21	2,743
ZDock	58	36	18	134,934
ZDock+ ZRank2 decoys = 200,000	77	63	45	159,902

No. of solved complexes in the top, number of target–binder complexes within 5 Å iRMSD found in the top 100, top ten or top one (for holo cases) or top 1,000, top 100 and top ten (for apo cases). Time (min), CPU time in minutes for each program, which excludes precomputation time for MaSIF-search.

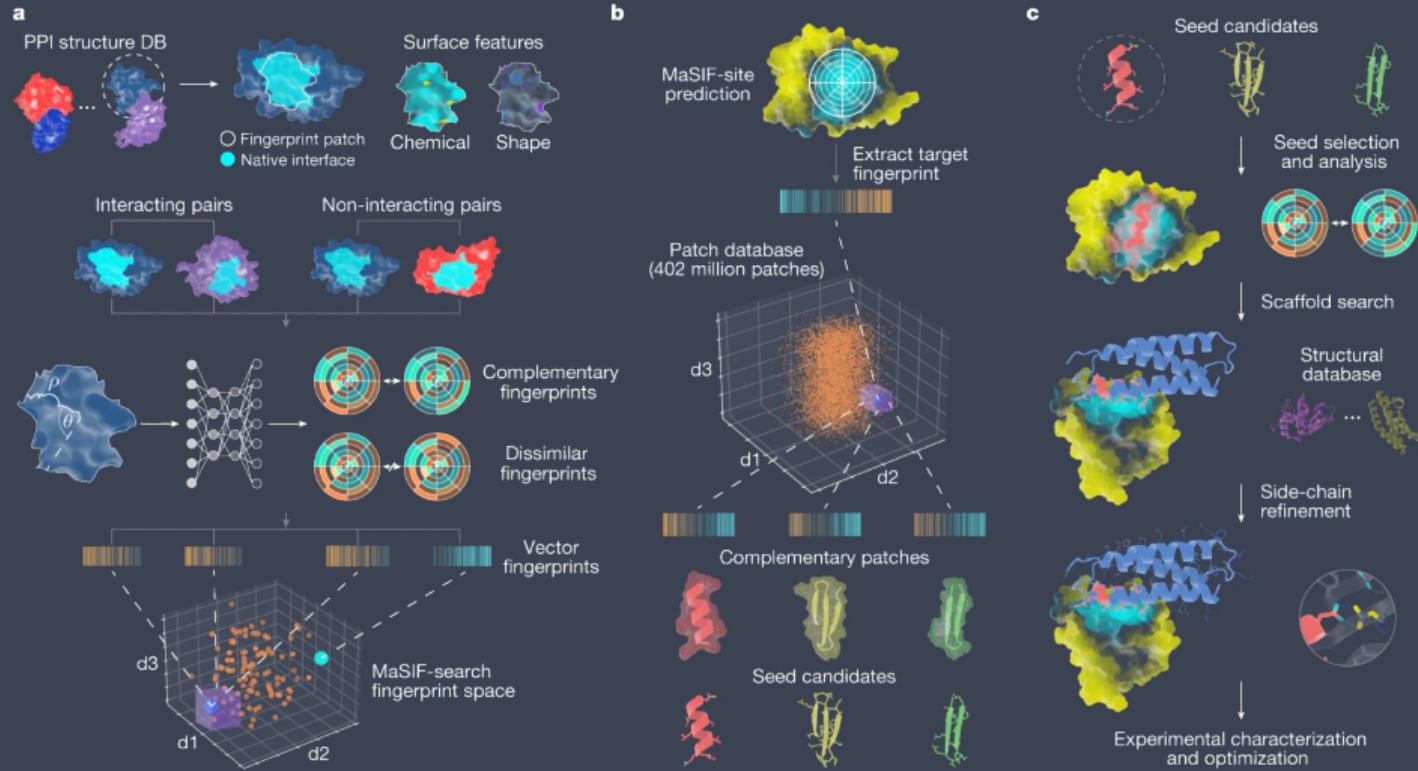
**Table 2 | Results for large-scale docking benchmark for PatchDock, MaSIF-search (with multiple numbers of decoys), ZDock and ZDock+ ZRank2 on unbound (apo) complexes**

Method	Number of solved complexes in the top			time (min)
	1,000	100	10	
MaSIF-search decoys = 2,000	17	7	2	16
PatchDock	11	4	1	560
ZDOCK	17	13	5	13,174
ZDock+ ZRank2 decoys = 80,000	23	12	5	16,866

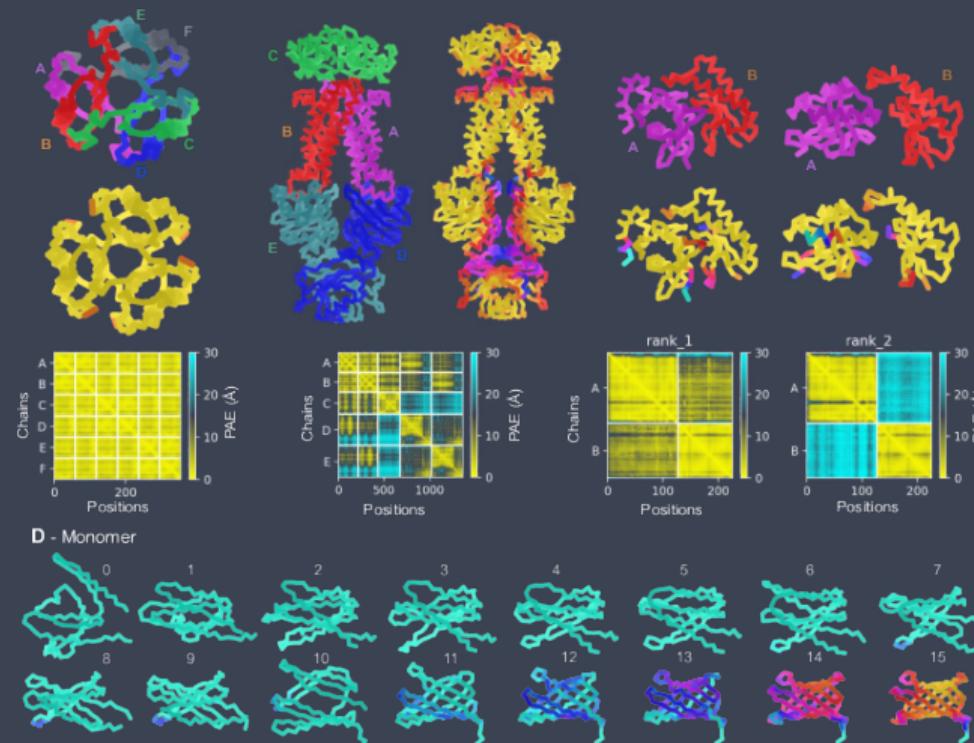
\*Moreover, all these methods could benefit from sequence evolutionary data to improve their predictive capabilities.



# » Masif-seed



# » ColabFold



[10.1101/2021.08.15.456425v1.full.pdf](https://doi.org/10.1101/2021.08.15.456425v1.full.pdf)



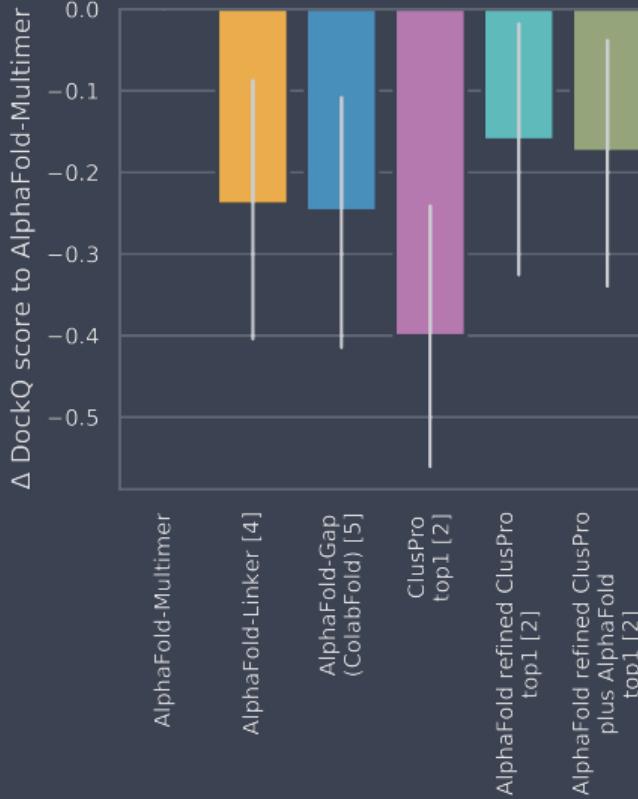
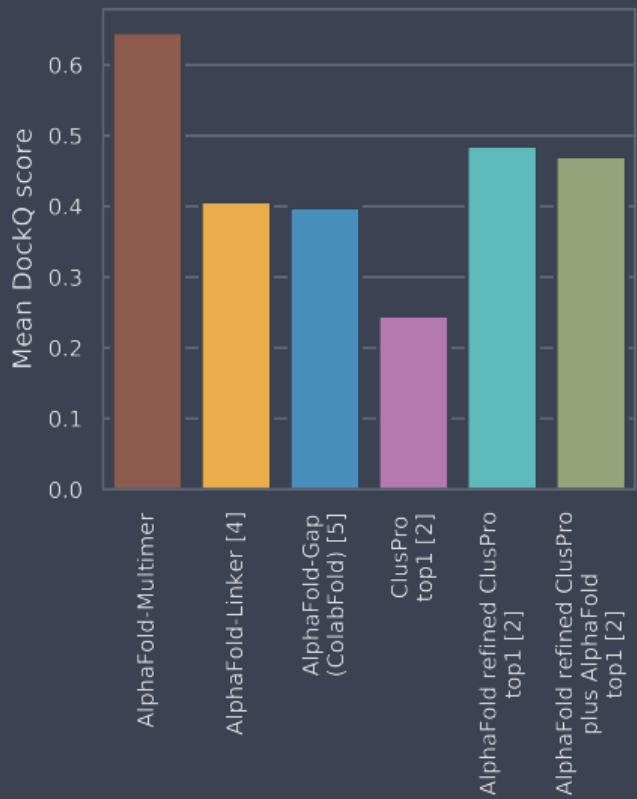
## » AlphaFold-Multimer

- \* Модификации функции loss, чтобы учесть симметрию перестановок для идентичных цепей
- \* Совмещение двух MSA для индивидуальных цепей для утилизации генетической информации об контакте
- \* Новый способ выборки набора остатков для обучения
- \* Разные мелкие оптимизации

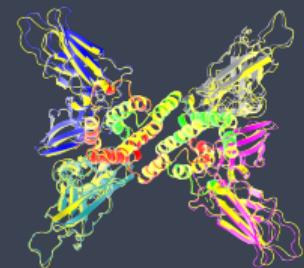
10.1101/2021.10.04.463034v1



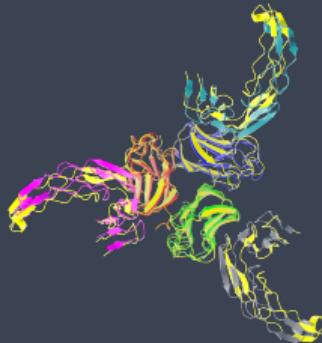
# » AlphaFold-Multimer



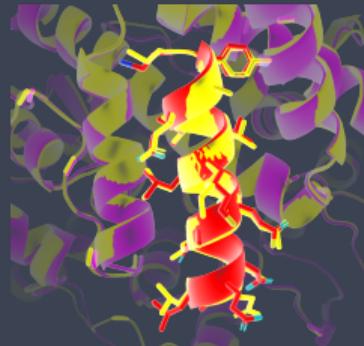
# » AlphaFold-Multimer



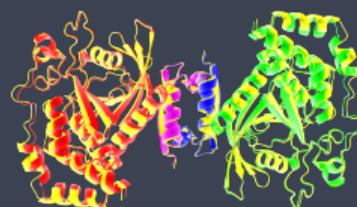
(a) A2B2C2 heteromer  
TM-score = 98.0,  $N_{\text{res}} = 1,246$ , PDB ID = 6E3K



(b) A3B3 heteromer  
TM-score = 89.3,  $N_{\text{res}} = 795$ , PDB ID = 7KHD



(c) Protein-peptide complex  
TM-score = 96.0, DockQ = 0.948,  
 $N_{\text{res}} = 385$ , PDB ID = 6JMT



(d) A2B2 heteromer  
TM-score = 98.3,  $N_{\text{res}} = 716$ , PDB ID = 6IWD



## » AlphaFold-Multimer

