

# Лекция 8. Свойства лигандов. Построение лигандов. QSAR. Докинг

Курс: Структурная Биоинформатика и моделирование лекарств (ВШЭ)

Головин А.В.<sup>1</sup>

<sup>1</sup>МГУ им М.В. Ломоносова, Факультет Биоинженерии и Биоинформатики

Москва, 2017

# Описание молекул

- **Molecular descriptors.** Параметры молекул это некие свойства в количественном описании, которые можно легко рассчитать зная только формулу молекулы.

## **Примеры:**

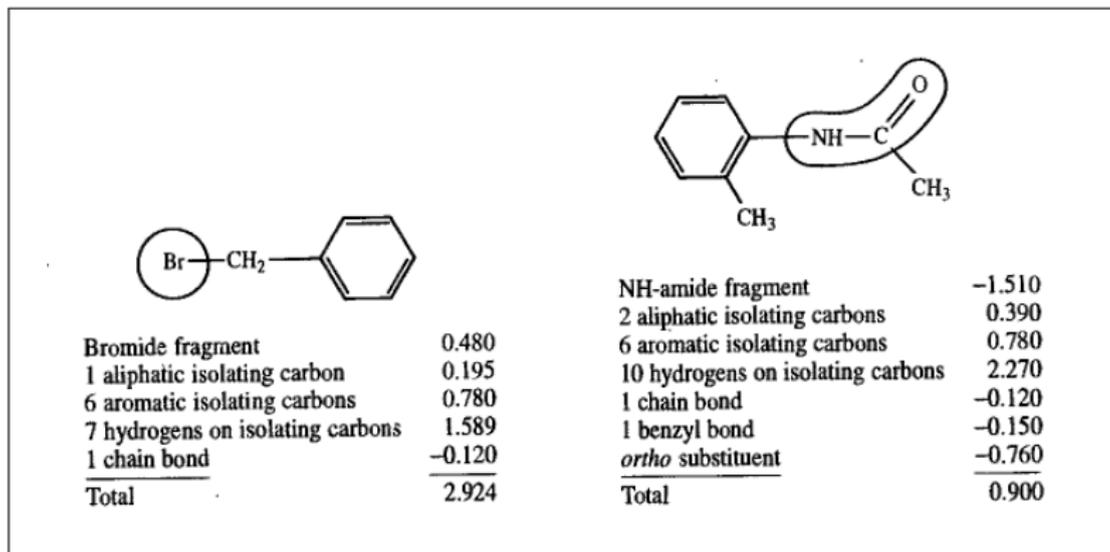
- Молекулярная масса
- Количество атомов
- Распределение вода/октан
- Электротопологические индексы
- Молярная поляризуемость
- Топологические двугранные углы

# Распределение октанол/вода

- $\log P$  где  $P$  степень распределения вещества в системе октанол/вода.
- Для некоторых веществ трудно определить экспериментально.
- Можно использовать метод пертурбации свободной энергии. Есть ряд неудобств.
- Используют фрагментарный подход. Распределение для молекулы равно сумме распределения компонентов с коэффициентами.

# Реализация: CLOGP

- Предложено разбивать молекулу по одинарным связям на фрагменты.



# Молярная рефрактивность

Молярная рефрактивность отражает способность газа преломлять свет, что так же отражает поляризуемость молекулы.

$$MR = \frac{(n^2 - 1)MW}{(n^2 + 1)d}$$

$d$  это плотность

$n$  это некий индекс не сильно изменяющийся для органических соединений.

- На основании этого значения часто судят о размере молекулы и собственно плотности вещества

# Топологические индексы.

- Хол и Кир предложили множество индексов позволяющих судить о молекуле в одной цифре.
- Основа это дельта значения для тяжёлых атомов

$$\delta_i = \sigma_i - h_i \quad \delta_i^v = Z_i^v - h_i$$

$\delta$  : количество сигма электронов  $Z$  : количество валентных электронов

**Chi — молекулярные соединения:**

$${}^0\chi = \sum_{atoms} (\sigma_i)^{-1/2}$$

$${}^0\chi^v = \sum_{atoms} (\sigma_i^v)^{-1/2}$$

$${}^1\chi = \sum_{atoms} (\sigma_i \sigma_j)^{-1/2}$$

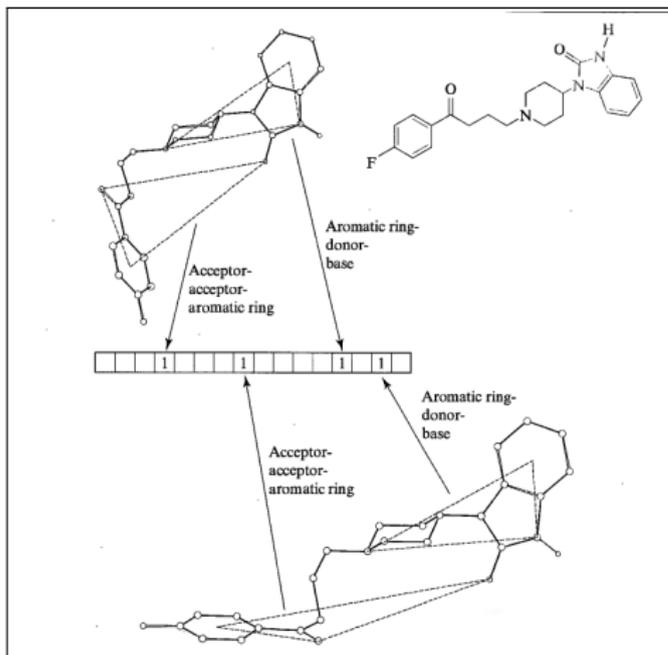
$${}^1\chi^v = \sum_{atoms} (\sigma_i^v \sigma_j^v)^{-1/2}$$

	4	3	2	1	4.828	2.914	1.707
	5	4	1	0	4.992	2.808	1.922
	5	3	2	0	4.992	2.770	2.183
	6	4	0	0	5.155	2.643	2.488
	7	3	0	0	5.207	2.561	2.914

Leach,1996

# Фармакофорные индексы

- Как правило выделяют три точки.
- Определяют расстояния между точками для разных конформеров.



Leach, 1996

# Создание выборки

- Для поиска заготовки для лекарства важно создать хорошую стартовую выборку соединений.
- Используют рассчитанные параметры молекул для оценки качества выборки.
- Надо добиться нормального распределения параметров.
- Избегают высокой корреляции для избежания пере представленности выборки.

# QSAR, количественные соотношения структура/активность

- QSAR — построение математической модели для описания соотношений структура-активность  
Задача предсказать свойства молекулы *in vivo*  
В целом такое соотношение можно записать как:

$$v = f(p)$$

где  $v$ : активность, а  $p$ : свойства из структуры

$$\log \left( \frac{1}{C} \right) = k_1 \log P - k_2 (\log P)^2 + k_3 \sigma + k_4$$

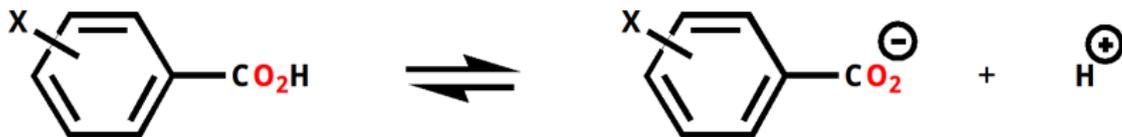
$$\pi = \log \left( \frac{P_X}{P_H} \right); \quad \log \left( \frac{1}{C} \right) = k_1 \pi - k_2 \pi^2 + k_3 \sigma + k_4$$

- На сегодняшний день существует множество модификаций этой простой формулы, которые учитывают множество параметров молекулы

# QSAR, электронный эффект

Константа  $\delta$  отражает влияние заместителя на смещение электронной плотности.

Эффект можно увидеть при отрыве протона



X=H

$$K_h = K_{diss} = \frac{[PhCO_2^-]}{[PhCO_2H]}$$

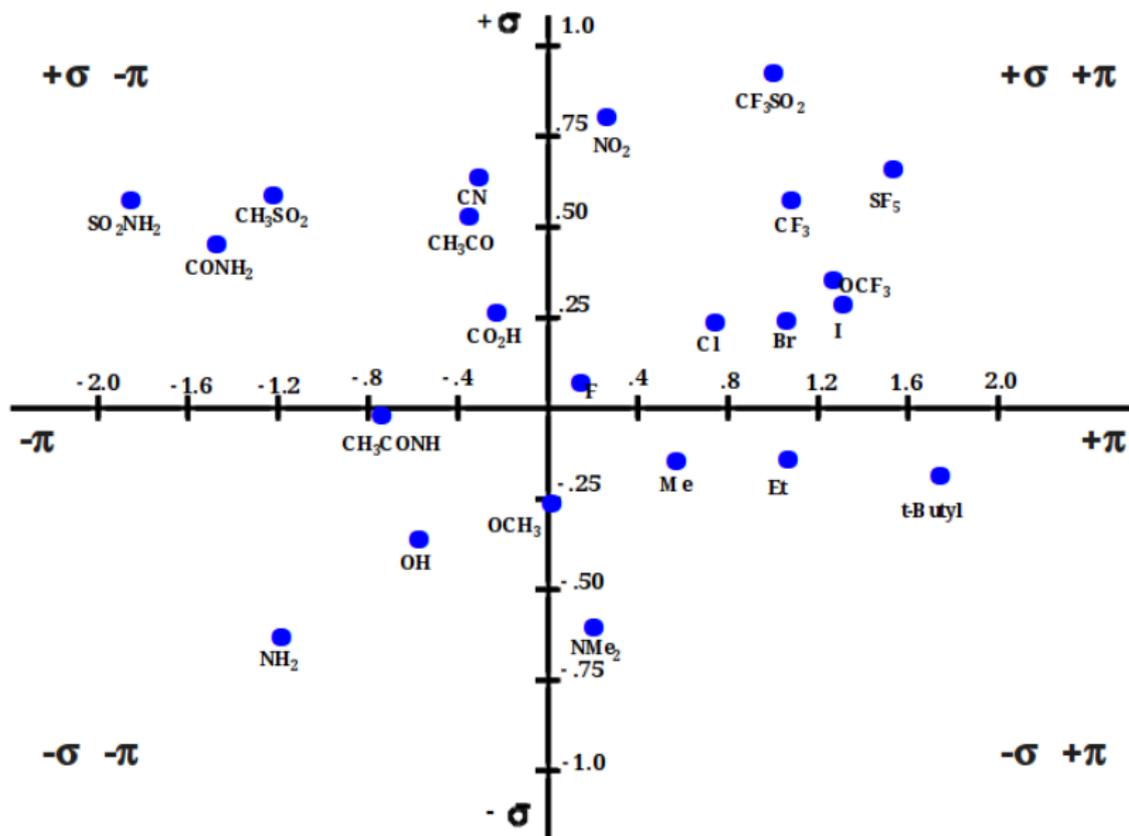
# QSAR, выбор соединений

- Для поиска коэффициентов надо синтезировать набор соединений с разными  $P$ .
- Желательно равномерное распределение выбранных соединений в шкале  $P$ .
- Возможна также вариация соединений при разных  $pH$  и температуре.

# Получение уравнения

- Простейший подход это определение зависимости активности от параметра как линейной регрессии. Тогда определение коэффициентов это просто метод наименьших квадратов.
- Линейная регрессия расширяется до множественной линейной регрессии, где более чем одна независимая переменная.
- Считается, что для статистически значимого определения необходимо не менее 5 соединений на каждый параметр.
- Существуют генетические алгоритмы для поиска коэффициентов регрессии.

# Диаграмма Крейга



# Кросс-валидация

Распространённый способ проверки это выборочное удаление данных из выборки и сравнение результатов.

# Использование QSAR

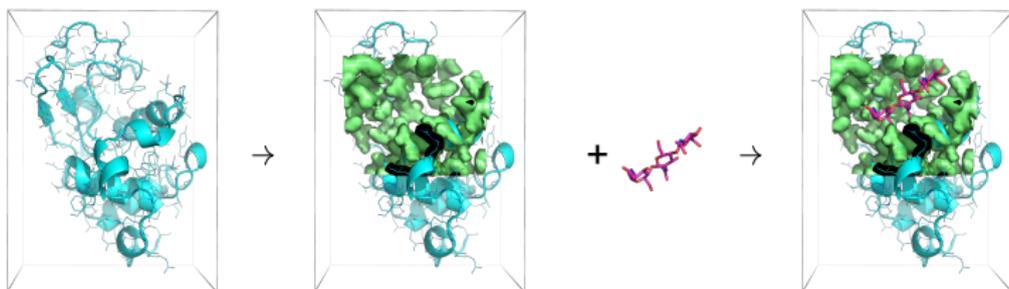
- Главная задача это предсказать оптимальную структуру
- Часто бывает, что QSAR хорошо работает при интерполяции, а не при экстраполяции.
- Бывает необходимо использовать не линейную зависимость, применяют параболическую.
- Существует билинейная модель:

$$\log(1/C) = k_1 \log P - k_2 (\log(\beta P + 1)) + k_3$$

- Отсутствие корреляций для параметра скорее всего означает, что этот параметр не задействован в механизме.

# Докинг белок-лиганд

Метод поиска способа связывания лиганда с белком



В результате мы можем узнать положение лиганда в комплексе с белком и оценить константу связывания

# Положение в сайте связывания

- Сайт связывания — место связывания лиганда
- Геометрия связывания — место связывания, ориентация и конформация лиганда

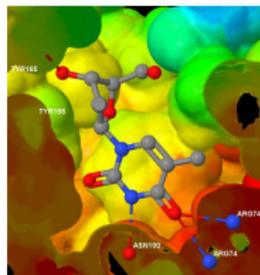
# Использование докинга

## Основные цели докинга:

- Виртуальный поиск лигандов
- Определение геометрии связывания лиганда

## Если мы знаем, как связывается лиганд, то:

- Мы можем узнать, какие части важны для связывания
- Можно предложить изменения для улучшения константы связывания
- Можем избежать ошибок



# Два основных компонента программ для докинга

- Алгоритм поиска
  - Установление места связывания
  - Установление геометрии связывания
- Алгоритм расчёта константы связывания областей с низкой энергией.

# Реализация

## **Сегодня существует много программ для докинга**

- AutoDock, DOCK, e-Hits, FlexX, FRED, Glide, GOLD, LigandFit, QXP, Surflex-Dock...и т.д.
- разные алгоритмы оценки аффинности и разные алгоритмы поиска
- Важно не путать лиганд-белок докинг и белок-белок докинг

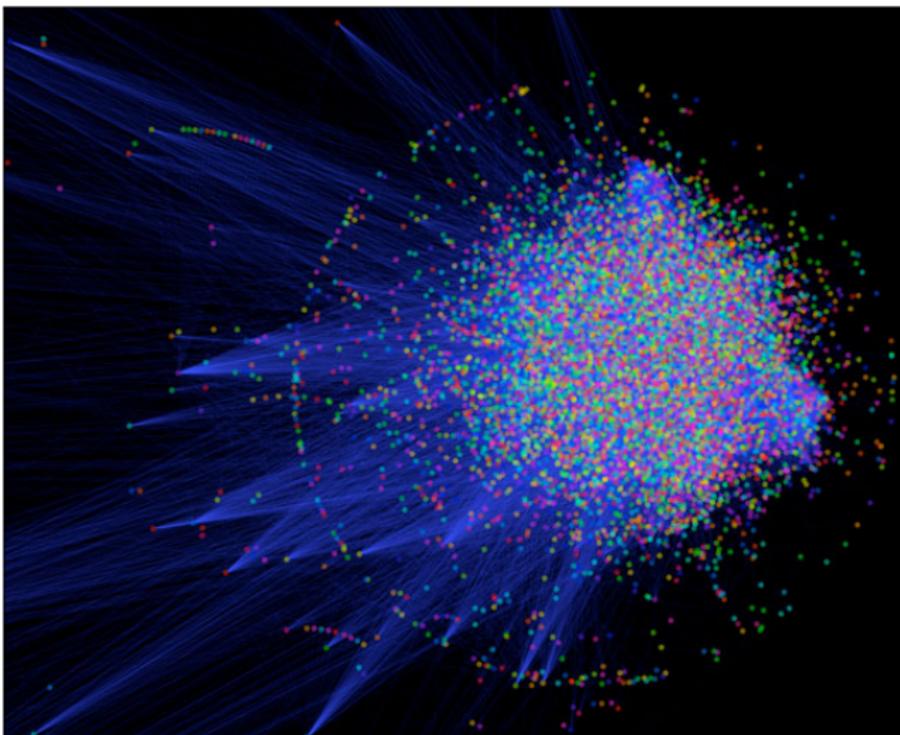
# Практические аспекты

- Часто PDB-структура содержит молекулы воды, почти всегда их надо убрать.
- Надо добавлять протоны к структуре; His?
- Часто в PDB неточно определена ориентация некоторых групп, что сказывается на паттерне водородных связей.
- Протонирование лиганда и его таутомерные формы.

# Rigid|Flexible докинг

- Rigid: лиганд не имеет внутренних степеней свободы, т.е. вращение вокруг связей запрещено.
- Flexible: предполагает учёт вращения вокруг связей лиганда.
- Часто белок рассматривается как жёсткое тело

# Human "Interactome"



# 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.

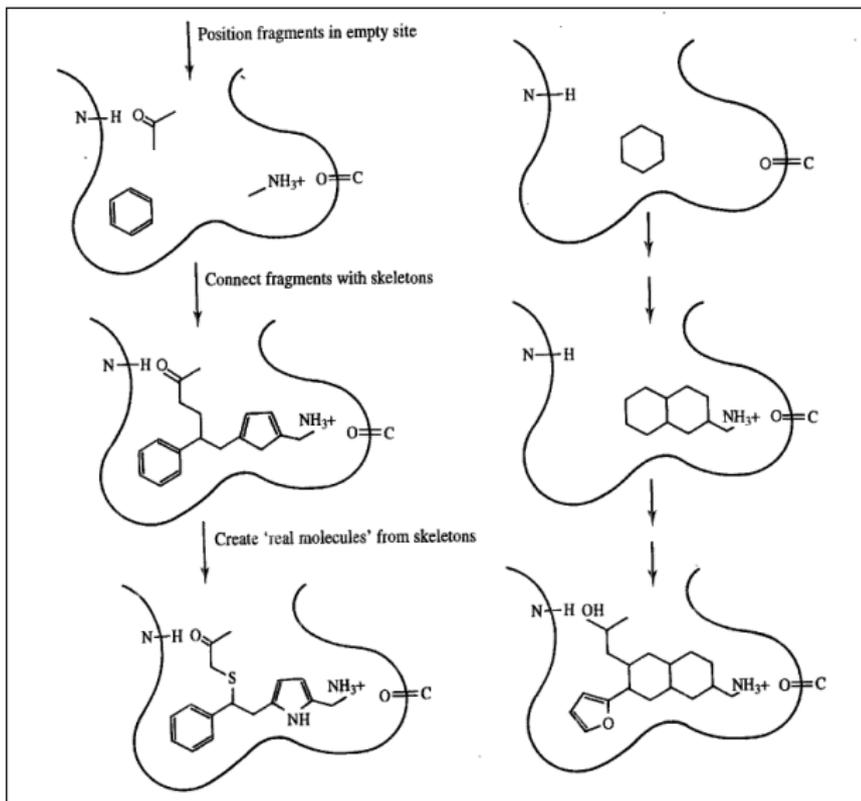
# Базы данных

- **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 слинкованная 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>

# Фрагментарное построение лиганда

- Сканирование по базам данных удобно, так как позволяет тут же проверить молекулу.
- Сканирование не предполагает создание лиганда *de novo*.
- Суть фрагментарного построения лиганда состоит в поиске мест в активном центре белка, где связываются небольшие фрагменты молекул
- Соединяя фрагменты связями при сохранении места связывания, добиваются высоких констант.

# Фрагментарное построение лиганда

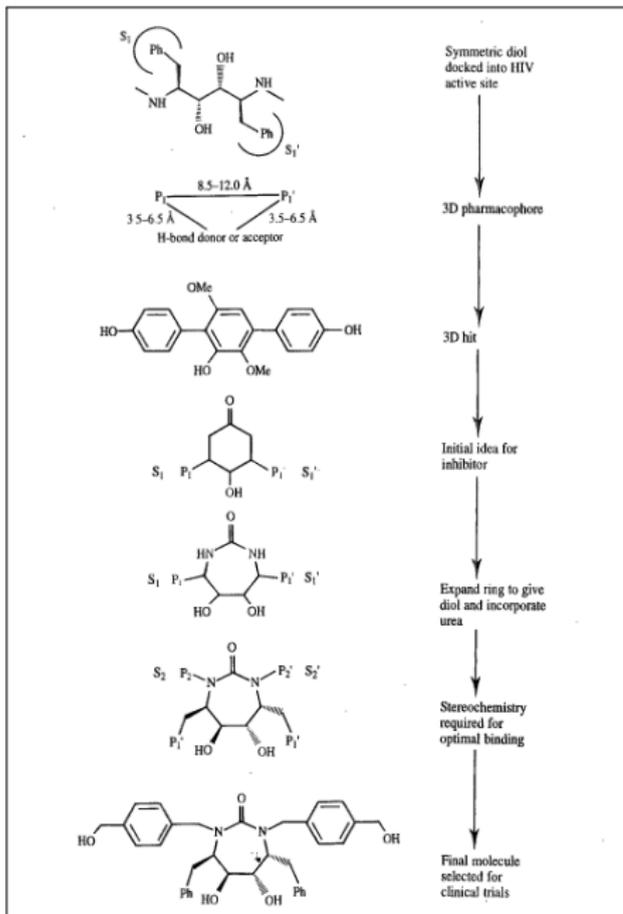


# Фрагментарное построение лиганда, реализация

- GRID аналог докинга.
- MCSS сайт наполняется фрагментами и с помощью EM вычленяется место где фрагмент наиболее предпочтителен. Взаимодействия между фрагментами не учитываются.
- LUDI использует информацию из банка PDB для задания фрагментов образующих водородные связи и т.д.
- Необязательно всё моделировать, можно использовать PCA и YMP для определения места связывания фрагмента.

# Реализация ковалентного связывания фрагментов

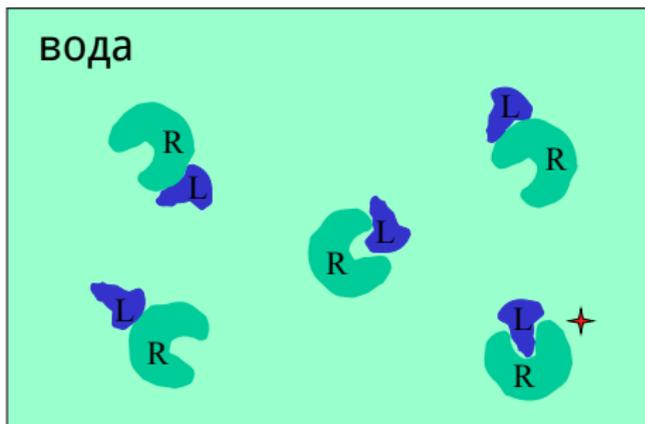
- Если у Вас есть два и более фрагмента, то можно искать способ их соединения по базам данных.
- Реализовано в CAVEAT.
- Можно строить автоматически строить скелеты. Главный критерий это сохранение взаимного положения фрагментов.
- Переход от скелета к молекуле сложен, так как надо реализовать возможность синтеза молекулы.



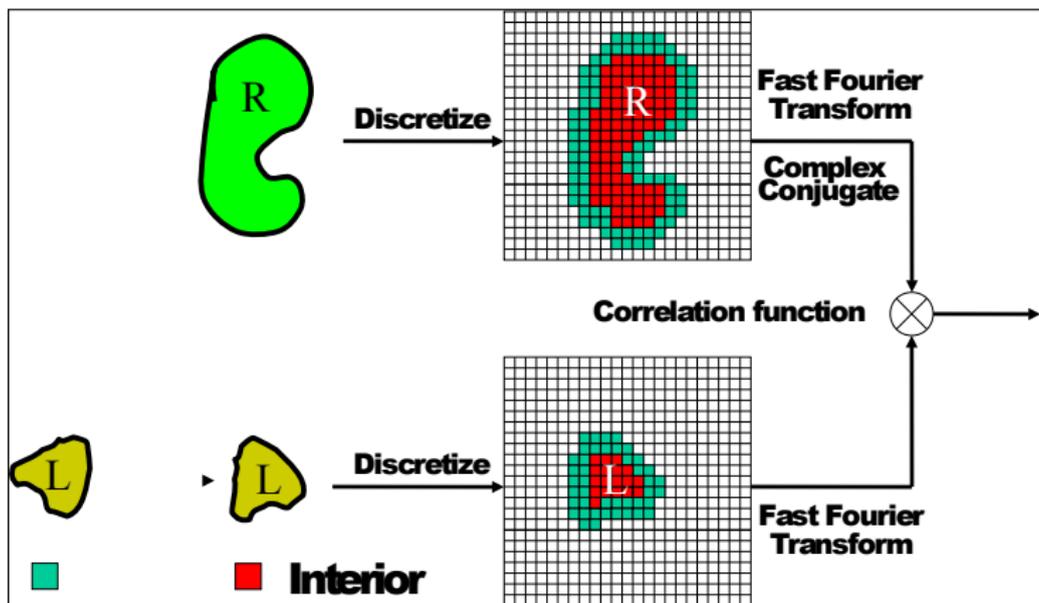
Leach, 1996

# Поиск наименьшего $\Delta G$

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



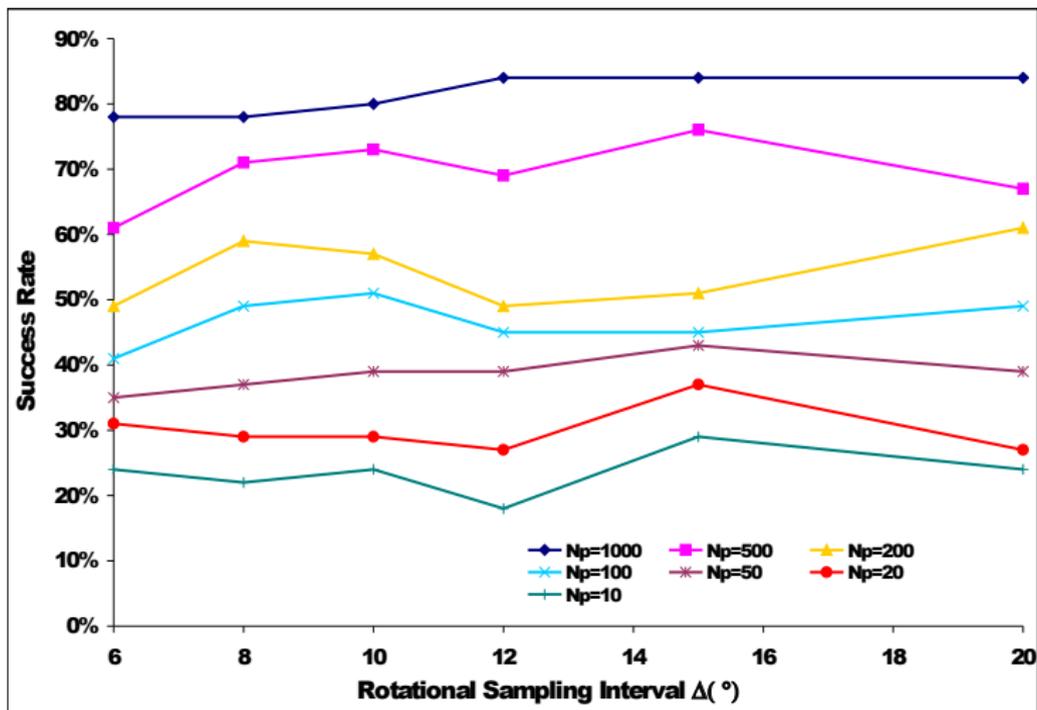
## Белковый докинг с использованием FFT



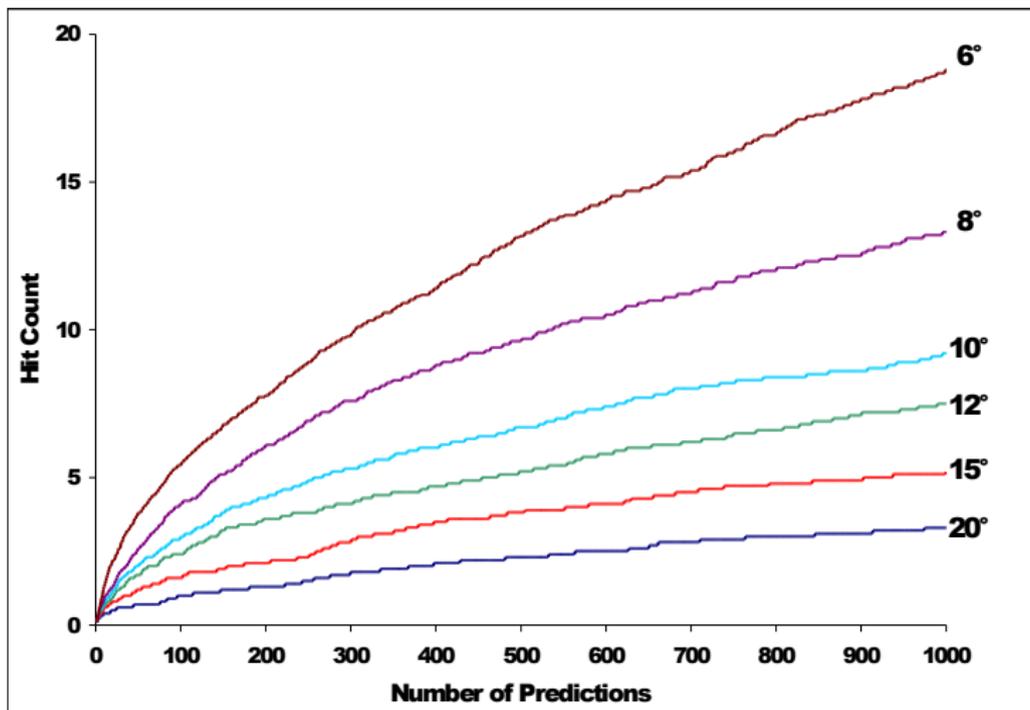
# Оценка производительности

- **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  $N_p$ .

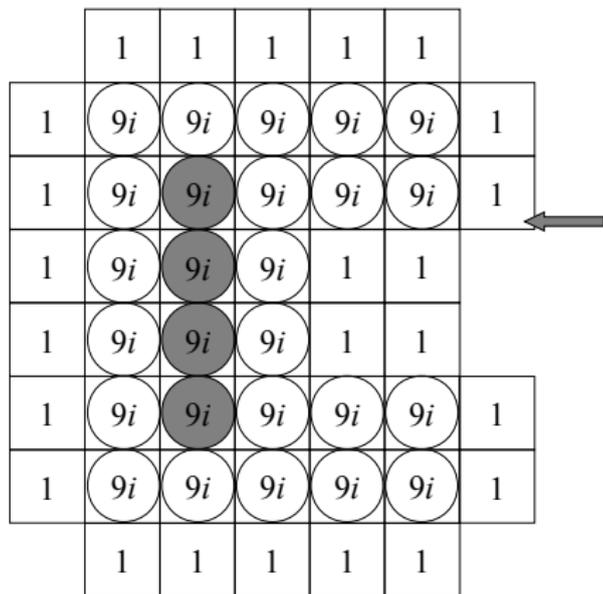
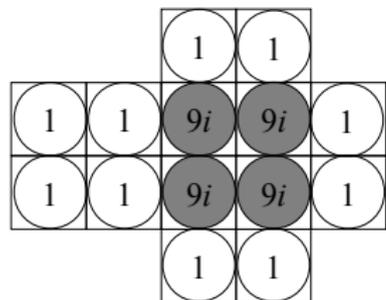
## Зависимость Success Rate от шага вращения



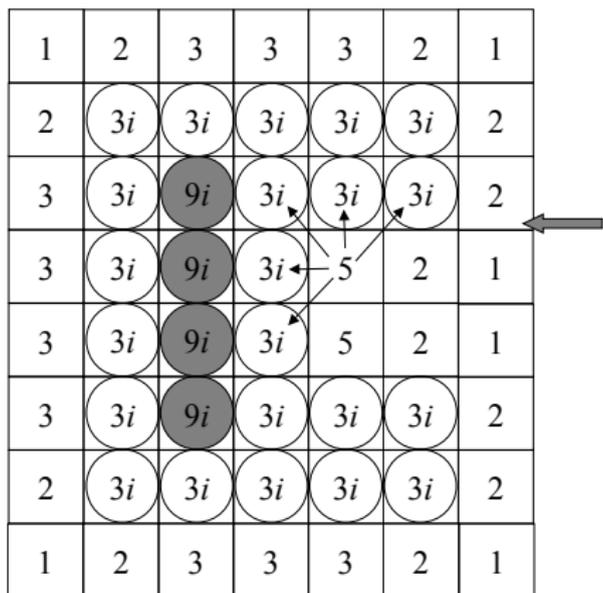
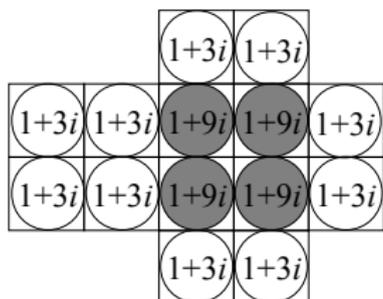
## Зависимость Hit count от шага вращения



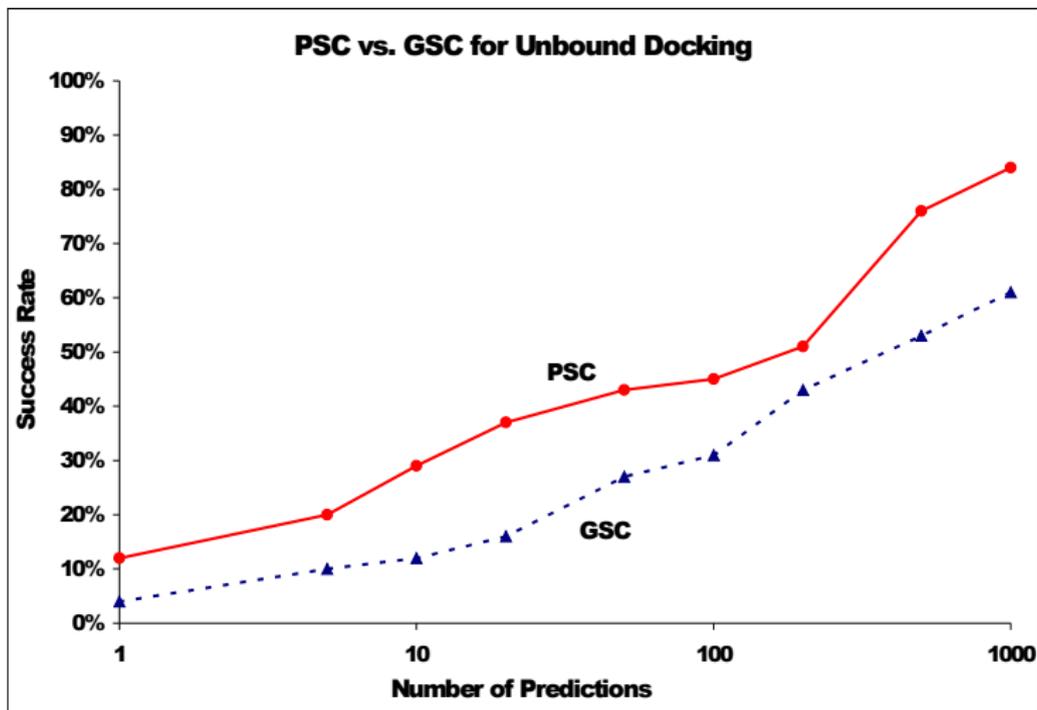
## Решеточная комплементарность поверхности

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

# Парная комплементарность поверхности


 $\mathbf{R}_{PSC}$ 

 $\mathbf{L}_{PSC}$

## PSC vs. GSC и Success Rate

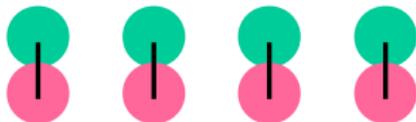


# Почему так?

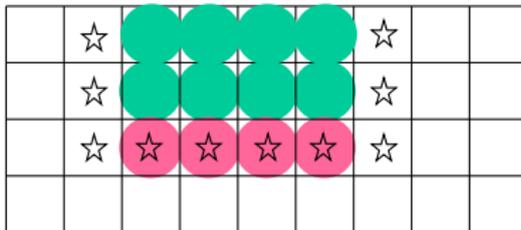
**A**



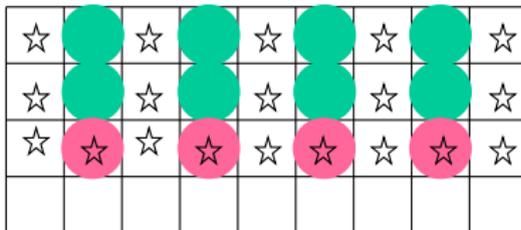
**B**



**C**



**D**



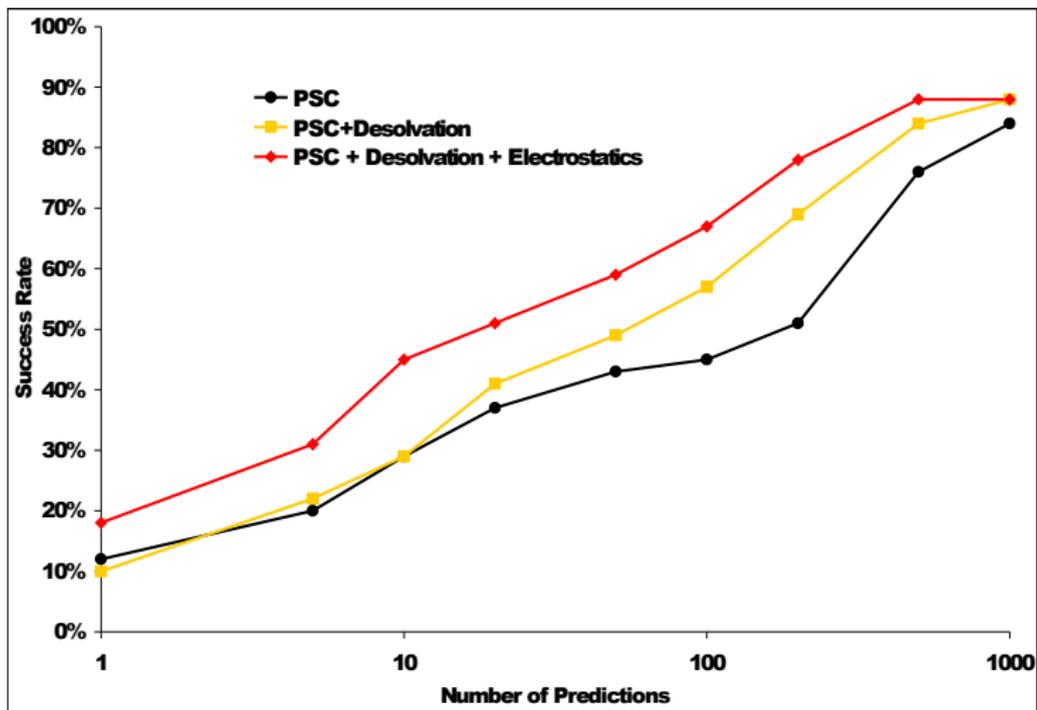
# Функция оценки энергии связывания

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

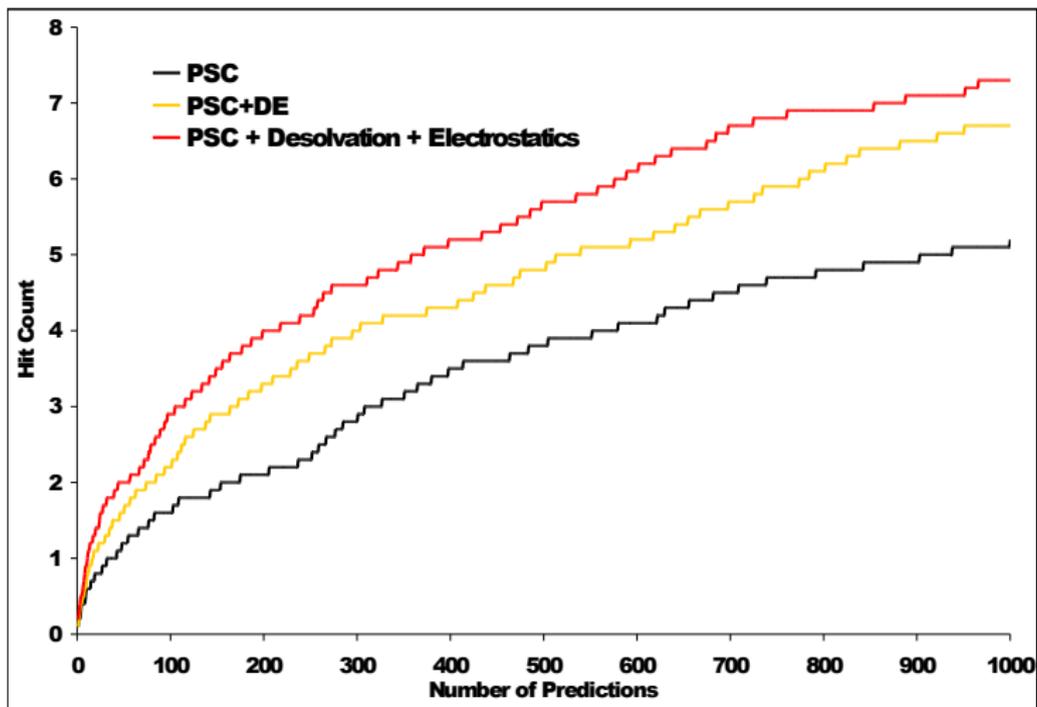
- $\Delta E_{VdW}$  - это комплементарность поверхности
- $\Delta G_{desol}$  - это гидрофобика
- $\Delta E_{el.}$  - это электростатика
- $\Delta G_{const}$  - это изменение вращательной и прочих энтропий.

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

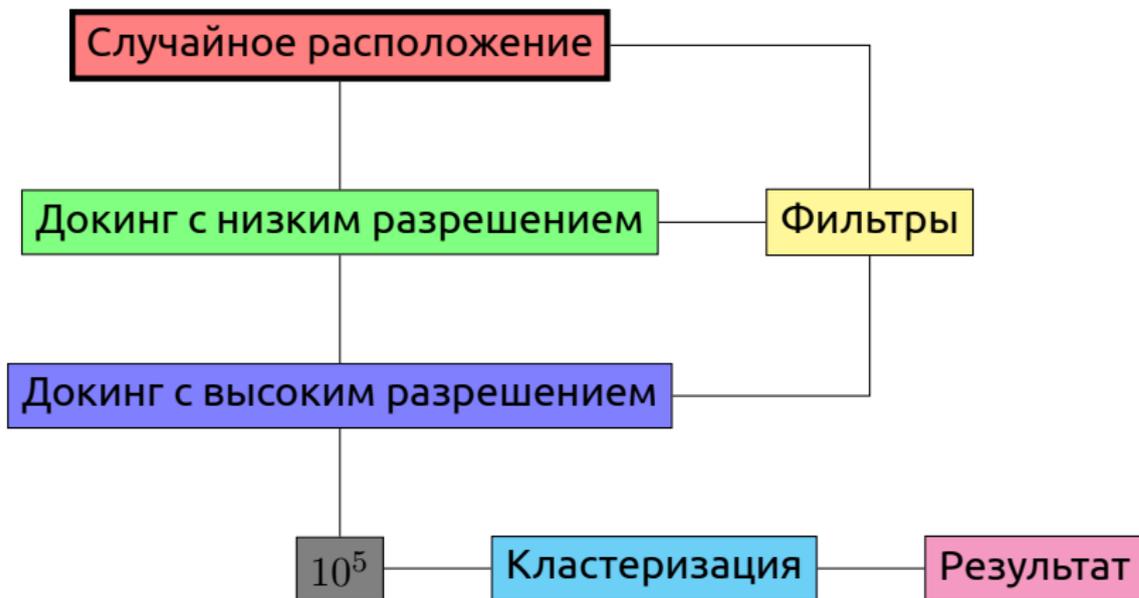
## Влияние на Success rate



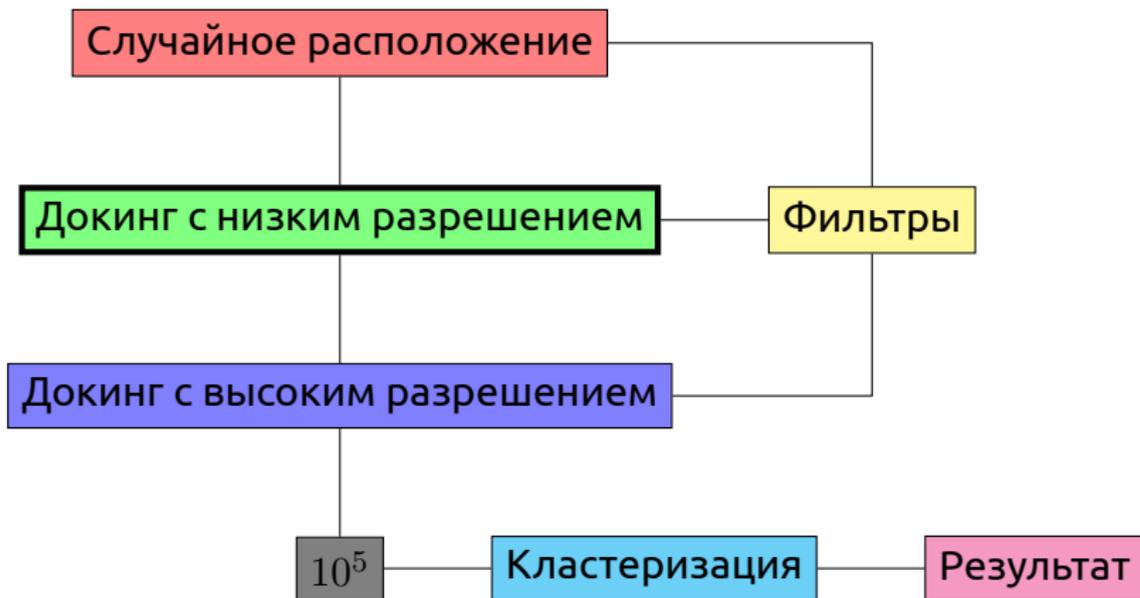
# Влияние на Hit Count



# Rosettadock Алгоритм



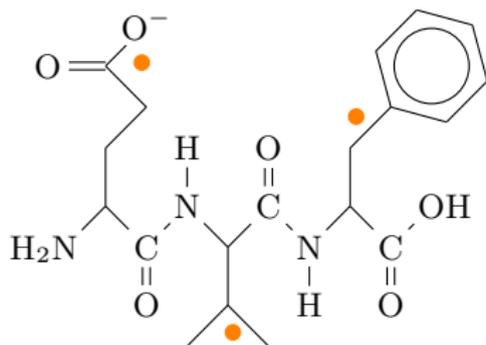
## Rosettadock Алгоритм



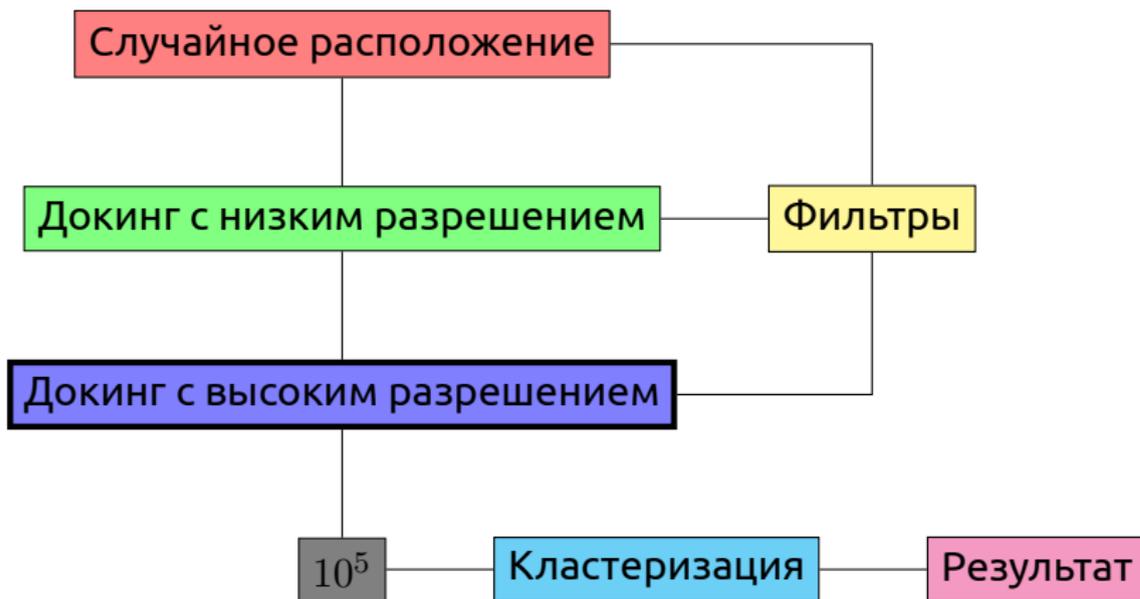
# Поиск с низким разрешением

## 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.



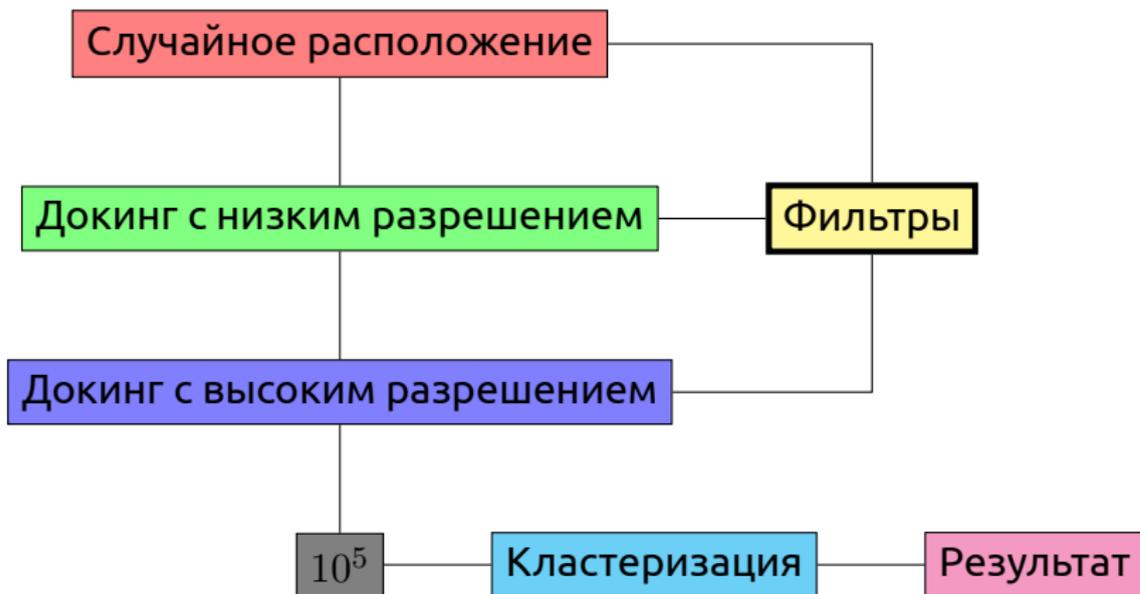
# Rosettadock Алгоритм



# Уточнение с высоким разрешением

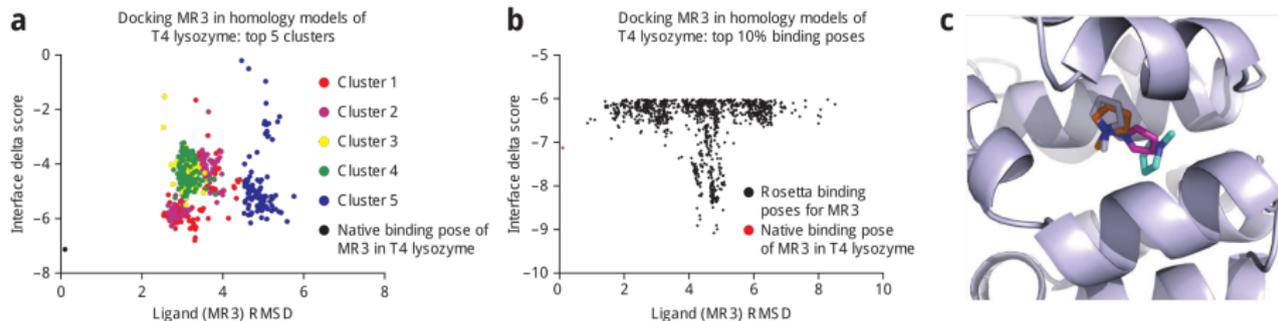
- Из библиотеки ротамеров добавляются полноатомные боковые цепи
- Используется полноценная оценка энергии (ММ)
- Монте-Карло + оптимизация геометрии
- Циклическое использование оптимизации положения как твердого тела и полноатомная оптимизация положения боковых радикалов

## Rosettadock Алгоритм



# Пример докинга лиганда

## 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).

# Beware of docking!

Yu-Chian Chen<sup>1,2,3</sup>

<sup>1</sup>Human Genetic Center, Department of Medical Research, China Medical University Hospital, Taichung, Taiwan

<sup>2</sup>Research Center for Chinese Medicine and Acupuncture, China Medical University, Taichung 40402, Taiwan

<sup>3</sup>Department of Biomedical Informatics, Asia University, Taichung, 41354, Taiwan

**Docking is now routine in virtual screening or lead optimization for drug screening and design. The number of papers related to docking has dramatically increased over the past decade. However, there are many issues to consider when undertaking a docking study. Frequent problems or issues arise, such as the wrong binding site of the target protein, screening using an unsuitable small-molecule database, the choice of docking pose, high dock score but failed in molecular dynamics (MD) simulation, and lack of clarity over whether the compound is an inhibitor or agonist. These problems should be cause for caution and concern before performing docking. Some papers show comprehensive biochemistry experiments but only a simple docking figure. This review presents some evidence to show that the docking might be questionable, despite a high score. In some cases, the accuracy of docking can even change from 0% to 92.66%. Thus, please beware of docking!**

**Docking for structure-based drug design**

Since its beginning in the 1960s, docking has gone with the

is low (it even works well on a personal computer), docking-related papers have sharply increased over the past decade (Figure 1). However, can we or should we trust the results of these docking studies? In this paper I provide a critical survey of the field, pointing out the strengths and weaknesses of the current family of docking protocols.

Careful evaluation shows that accuracy is a major problem with docking studies, because if the docking is not approached with precision then these papers will be of little value [6–8]. Questionable docking results can be found, even in high-profile journals. There are frequent problems such as an inaccurate binding site of the target protein, screening using an unsuitable small-molecule database, the choice of docking pose, high dock score (binding affinity) but failed in MD simulation, lack of clarity over whether the compound is an inhibitor or agonist, or the docking results are inconsistent with bioassays. The worst case is often found in some very high profile journals, which show an excellent bioassay but only with a simple docking figure. These problems in the interpretation of docking should be cause for caution and