СУПЕРКОМПЬЮТЕРНЫЙ ДОКИНГ НА ОСНОВЕ ГЛОБАЛЬНОЙ ОПТИМИЗАЦИИ С ПОМОЩЬЮ ТЕНЗОРНЫХ ПОЕЗДОВ

Алексей СУЛИМОВ, Данил КУТОВ, Дмитрий ЖЕЛТКОВ, Евгений ТЫРТЫШНИКОВ, Владимир СУЛИМОВ

DIMONTA

Москва Россия

Научно-Исследовательский Вычислительный Центр МГУ имени М.В. Ломоносова

Пять составляющих быстрой и эффективной разработки лекарств

1. Актуальная био-мишень – врачи, биологи, биохимики, молекулярные биологи, фармацевты…

2. Наличие 3D-структуры биомишени (декартовы координаты каждого атома)

3. Высокая точность вычисления свободной энергии связывания белок-лиганд

   Error ΔG_{bind} < 1 kcal/mol

4. Возможность и простота синтеза ингибиторов

5. Наличие надежных тестовых систем для заданной био-мишени: определение ингибирующей активности лигандов

Можно ли с помощью докинга аккуратно вычислить ΔG_{bind} ?

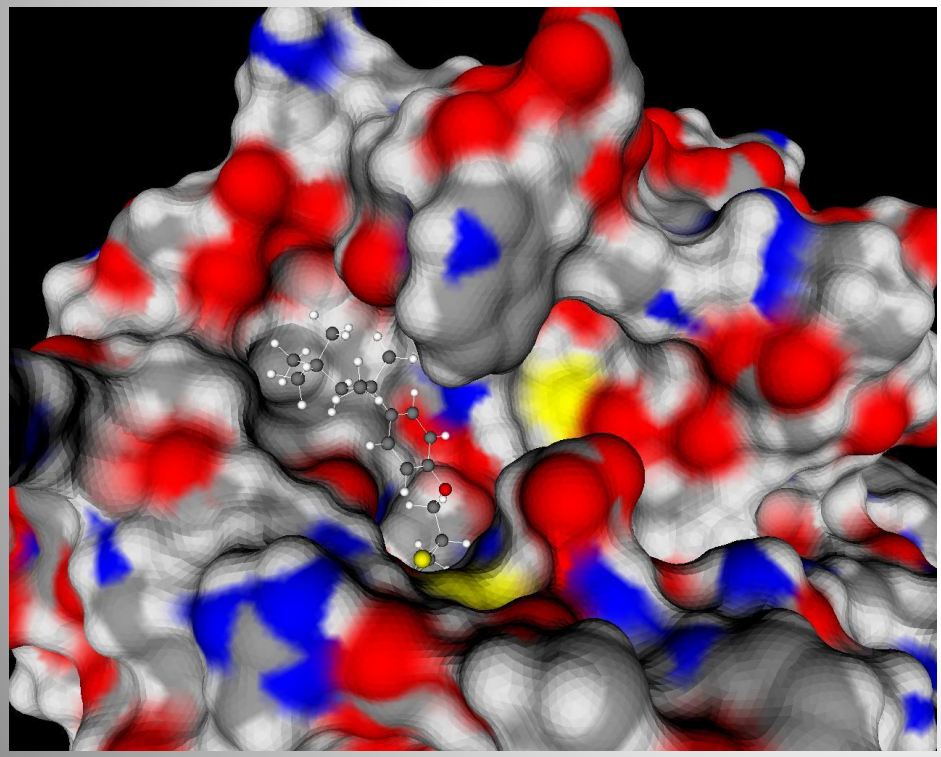
- Найти глобальный минимум (**minimum minimorum**) энергии комплекса белок-лиганд
- Найти спектр низкоэнергетических локальных минимумов – дают основной вклад в ΔG_{bind}
- Заданное силовое поле без подгоночных параметров
- Детальное исследование энергетической поверхности белок-лиганд
- Использование суперкомпьютерных мощностей
- Точность докинга:

Точность позиционирования

Точность вычисления энергии связывания

Парадигма докинга:

Положение лиганда в активном центре белка-мишени соответствует глобальному минимуму (**minimum minimorum**) энергии системы белок-лиганд



Докинг – нахождение **глобального минимума** целевой энергетической функции

Точность докинга зависит от:

- Силового поля – межмолекулярные и внутри-молекулярные взаимодействия
- Учет растворителя, модель
- Модели белка-мишени и лигандов
- Метод вычисления свободной энергии связывания
- **Алгоритм докинга**
- Доступные вычислительные мощности для докинга 1 лиганда

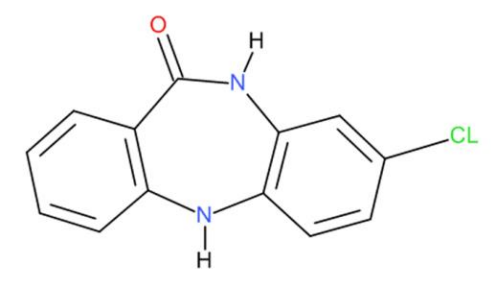
Точность позиционирования

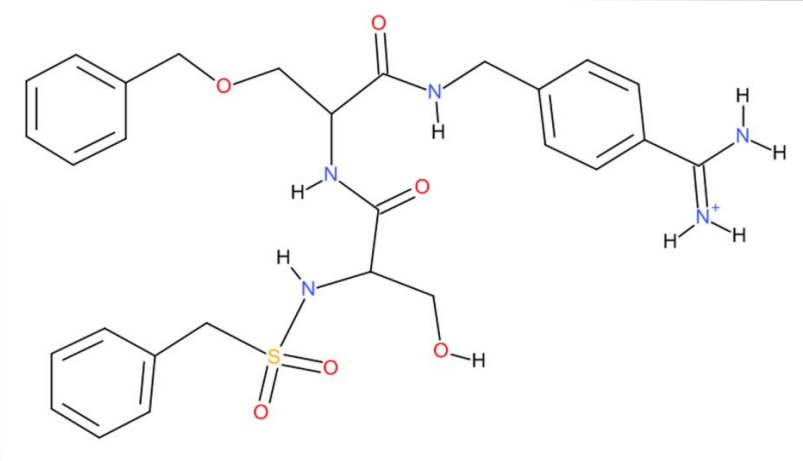
Точность вычисления энергии связывания

Тестовый набор комплексов белок-лиганд

- Набор **30 комплексов** белок-лиганд из Protein Data Bank, RMSD ≤ 2 Å, лиганды различных размеров и гибкости

- Поза локально оптимизированного нативного лиганда была близка к исходной позе нативного лиганда: RMSD < 1.5



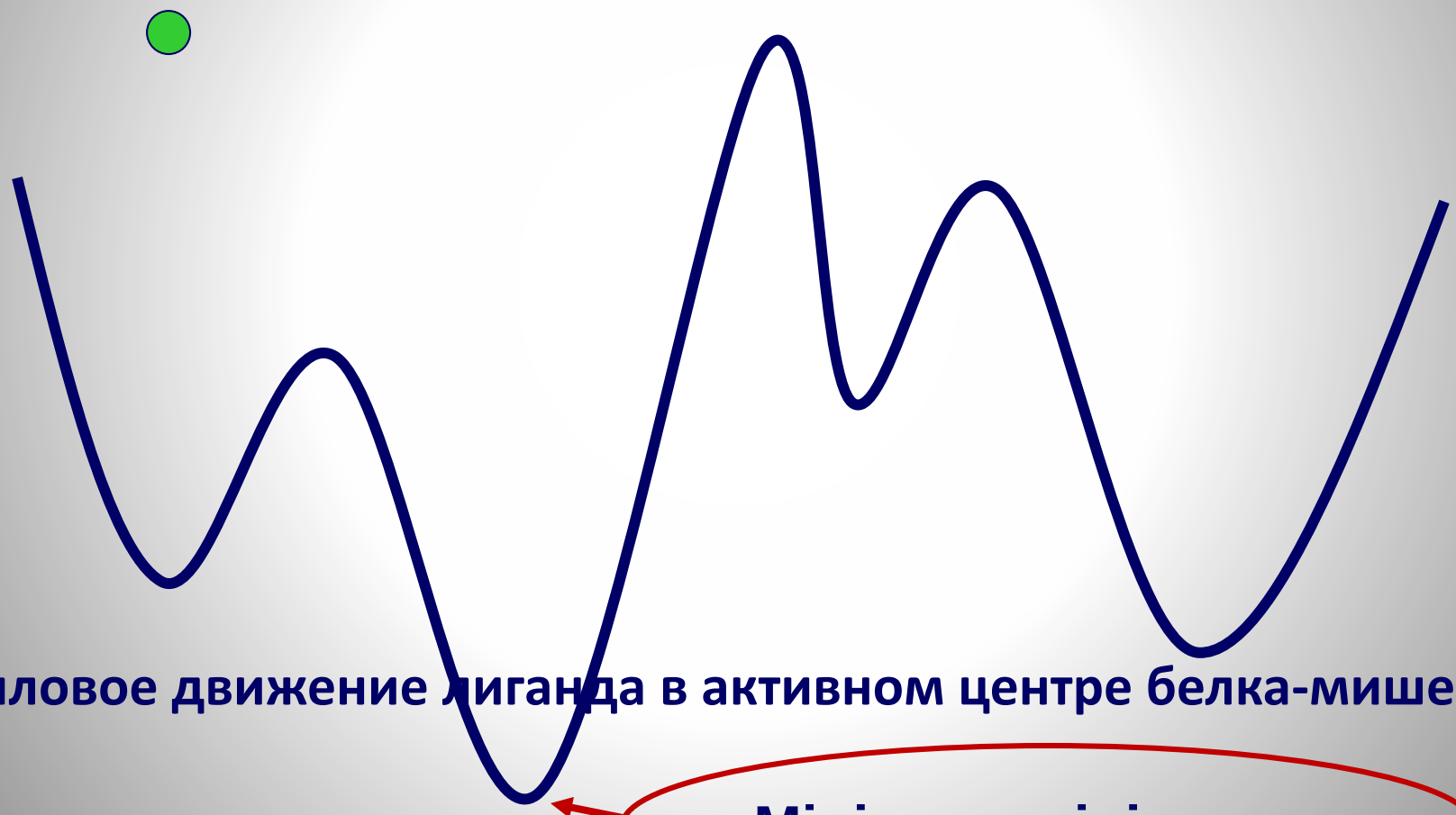


74 atoms, 19 torsions

FLM minima ⟷ SOL-T minima

Все минимумы с низкой энергией, найденные с помощью исчерпывающего поиска с применением суперкомпьютера

Минимумы с низкой энергией, найденные с помощью нового алгоритма тензорных поездов



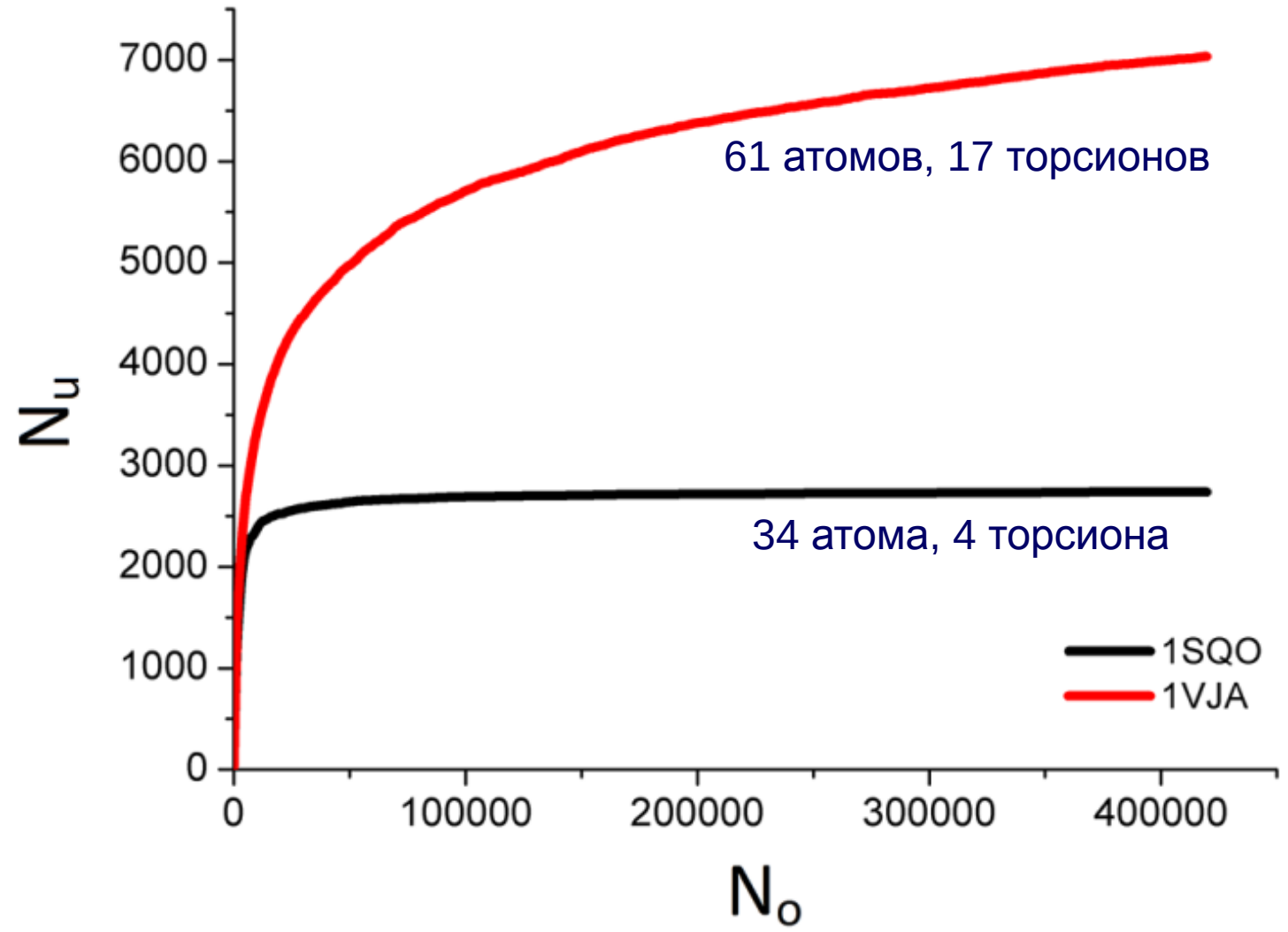
Тепловое движение лиганда в активном центре белка-мишени

Minimum minimorum

Программа докинга FLM – Find Local Minima

- FLM не использует предварительно рассчитанную сетку энергий атомов лиганда
- Жесткий протеин *(в данной работе)*
- Локальная оптимизация энергии по всем атомам лиганда, начиная из случайных поз лиганда
- Силовое поле MMFF94 в вакууме
- Параллельные многопроцессорные вычисления: **8191 ядер** несколько часов на суперкомпьютере Lomonosov – может считать сколь угодно
- 1024 различных низкоэнергетических минимума

Число обновлений локальных минимумов (*Nu*) в зависимости от числа сделанных оптимизаций (*No*) для комплексов 1SQO (black line) и 1VJA (red line)

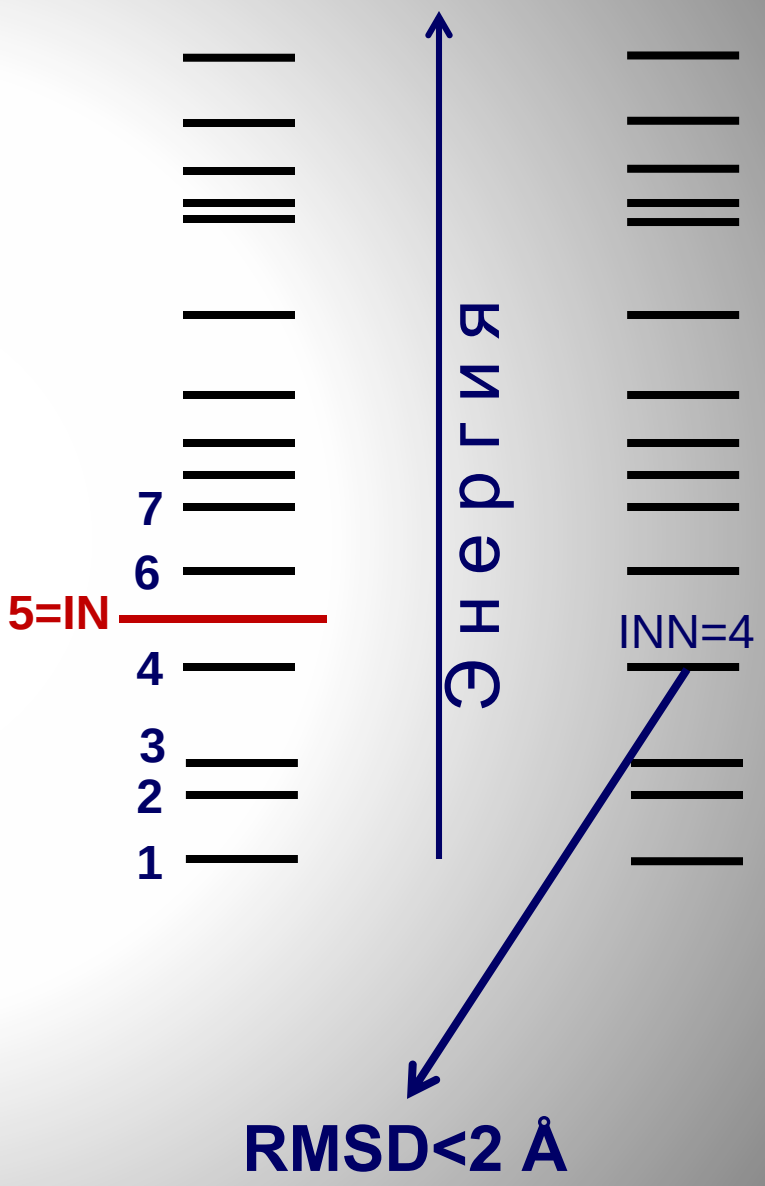


Индексы минимумов энергии комплекса белок-лиганд

- Каждый минимум имеет целочисленный индекс, соответствующий его положению в списке минимумов, отсортированных по возрастанию их энергии. Минимум с самой низкой энергией имеет индекс 1

- **IN** – (**I**ndex of **N**ative) это индекс локально оптимизированного нативного лиганда

- **INN** – (**I**ndex of **N**ear **N**ative) это индекс минимума с наинизшей энергией, в котором поза лиганда отличается от нативной позы меньше, чем на RMSD < 2 Å



SOL-T – программа докинга на основе тензорных поездов

1. Жесткий белок

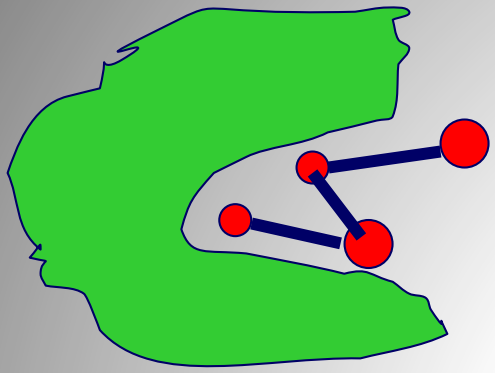
2. Гибкий лиганд: трансляции и вращения лиганда как целого и все внутренние торсионы

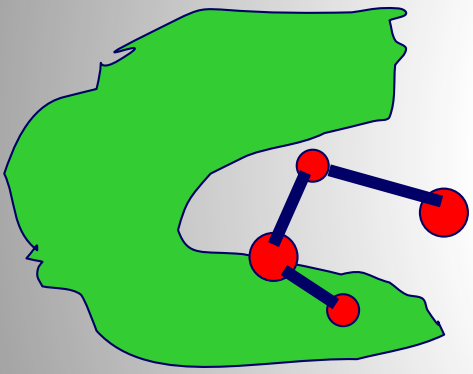
3. Силовое поле MMFF94 в вакууме

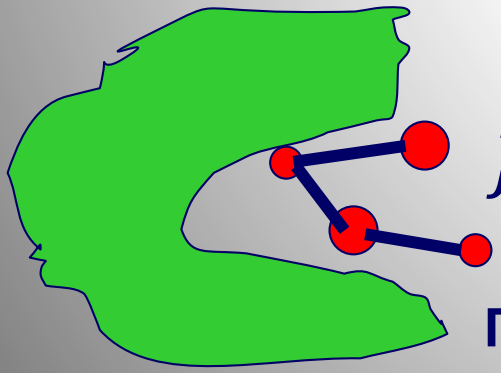
4. Новый алгоритм докинга на основе алгоритма глобальной оптимизации TT-CROSS – конечный алгоритм

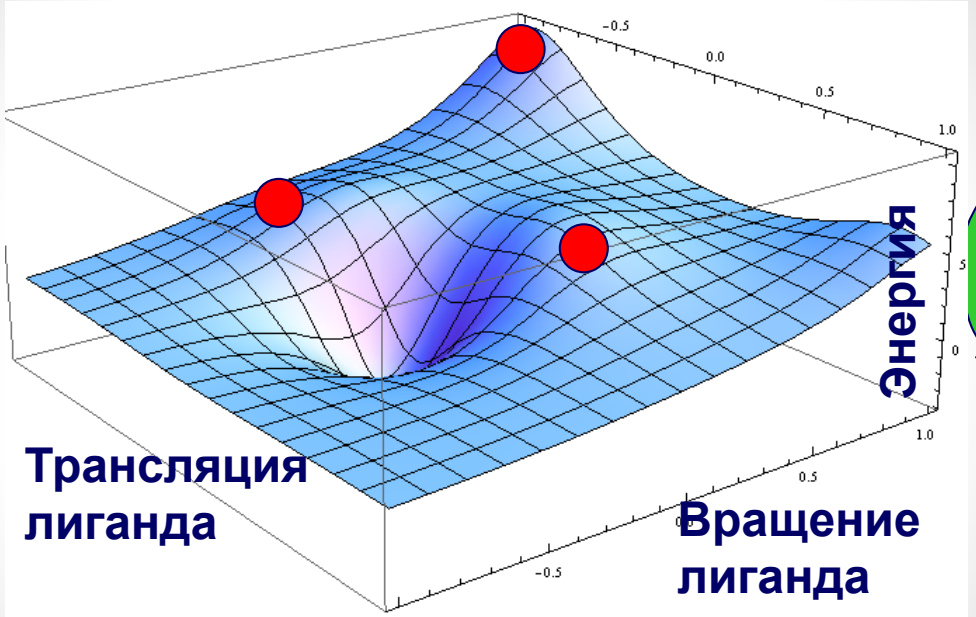
5. Локальная оптимизация

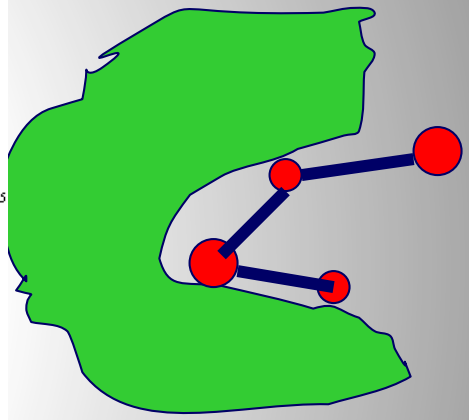
Перевод непрерывной функции многих переменных в многомерный тензор











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

Поиск максимального по модулю элемента тензора

Разложение в тензорный поезд

Тензор $A \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ может быть представлен в виде:

$$A(i_1, \dots, i_d) =$$

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

$r_1, \dots, r_{d-1}$ - **ТТ-ранги** тензора

$G_1, \dots, G_d$ - **ТТ-ядра (вагоны)** тензора

Преимущества ТТ-разложения

- Для хранения тензора требуется запомнить – только $\boldsymbol{O}(\boldsymbol{dnr^2})$ элементов вместо всех элементов тензора $\boldsymbol{n^d}$.

- Быстрый метод получения аппроксимации тензора, представленного многомерной функцией (TT-Cross) – вычисление $\boldsymbol{O}(\boldsymbol{dnr^2})$ элементов и выполнение $\boldsymbol{O}(\boldsymbol{dnr^3})$ арифметических операций.

- Быстрая и надежная тензорная арифметика (большинство операций стоят $\boldsymbol{O}(\boldsymbol{dnr^3})$ или даже меньше), при этом тензорные арифметические операции это последовательности стандартных матричных операций и методов.

$$r = \max_{i=\overline{1,d-1}} r_i$$

$$n = \max_{i=\overline{1,d}} n_i$$

Глобальная ТТ оптимизация: поиск максимального по модулю элемента тензора

1. Генерация подматриц неких матриц развертки, используя наборы элементов тензора
2. Интерполяция подматриц, используя крестовый метод интерполяции с рангом не выше r_{max}: набор интерполяционных точек для каждой подматрицы содержит элементы с большим значениями модуля.
3. Локальная оптимизация интерполяционных точек (поз лиганда), добавление проекций оптимизированных точек (поз лиганда) к тензору и к наборам интерполяционных точек.
4. Обновление каждого набора интерполяционных точек матрицы развертки объединением набора точек для предыдущей матрицы развертки с набором для последующей матрицы развертки.
5. Добавление лучших точек к набору точек для каждой матрицы развертки и переход к п.1, используя полученный набор точек (поз лиганда) как элементы тензора.

Глобальная оптимизация с помощью ТТ

- Поиск максимального по модулю элемента тензора

- Сложность - $O(dnr_{max}^2)$ вычислений функции энергии, $O(dnr_{max}^3)$ арифметических операций, $O(dr_{max})$ локальных оптимизаций

- Высокий параллелизм вычислений, сложность параллельности - $O(r_{max})$ вычислений функции энергии, $O(r_{max})$ арифметических операций, $O(1)$ локальных оптимизаций .

Варианты программы SOL-T

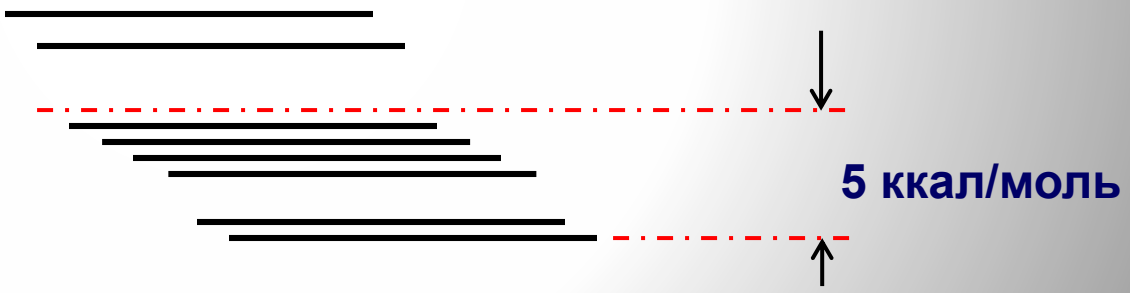
- SOL-T16, r_{max}=16
- SOL-T32, r_{max}=32
- SOL-T64, r_{max}=64

r_{max} – это максимальный ранг для аппроксимации матрицы свертки

Критерии сравнения минимумов, найденных с помощью FLM и SOL-T

- Сравнение энергий найденных минимумов

- Сравнение числа локальных минимумов в заданном интервале энергий над глобальным



- Сравнение поз лиганда в минимумах - RMSD

Сравнение энергий FLM vs SOL-T

PDB ID	Разность энергий глобальных минимумов, ккал/моль				Число минимумов в интервале 5 ккал/моль над глобальным минимумом			
	FLM	SOL-T16	SOL-T32	SOL-T64	FLM	SOL-T16	SOL-T32	SOL-T64
1B9V	0	2.3	1.3	1.2	212	7	18	40
1BR5	0	3.7	2.7	2.7	11	2	3	3
1C5Y	0	0	0	0	2	2	2	2
1DWC	0	36.9	26.1	5.7	11	0	0	0
1EFY	0	3.5	0	0	30	8	14	28
1F5L	0	0	0	1.5	1	1	1	1
1HPV	0	7.1	6.5	2.9	4	0	0	1
1I7Z	0	0	0	0	1	1	1	1
1J01	0	0	10.1	8.3	4	1	0	0
1K1J	0	0	0.06	0	9	3	1	1

Сравнение энергий FLM vs SOL-T

- Глобальный минимум найден программой FLM
- Минимумы с самой низкой энергией, найденные программой SOL-T, совпадают с глобальным минимумом или лежат выше него
- Для 50% комплексов глобальный минимум, найденный программой FLM, находится одним из вариантов SOL-T16, SOL-T32, SOL-T64
- Для 9 комплексов из 30 глобальный минимум, найденный FLM находится всеми тремя вариантами SOL-T: SOL-T16, SOL-T32 и SOL-T64

Сравнение энергий FLM vs SOL-T

- FLM находит максимальное количество минимумов (по сравнению с SOL-T) в интервале 5 ккал/моль над глобальным минимумом
- Эффективность докинга для SOL-T16, SOL-T32 и SOL-T64 одинакова
- SOL-T16 существенно быстрее, чем SOL-T32, 64
- Парадигма докинга выполняется не всегда
- Комплексы с простой поверхностью и комплексы со сложной энергетической поверхностью

Значения индексов IN/INN для минимумов, найденных FLM и SOL-T для 30 комплексов

PDB ID	Значения индексов IN/INN			
	FLM	SOL-T16	SOL-T32	SOL-T64
1B9V	inf/inf	216/308	467/242	872/872
1BR5	inf/121	76/25	155/24	258/149
1C5Y	1/1	1/1	1/1	1/1
1DWC	inf/629	2/inf	56/7	70/18
1EFY	192/88	33/27	84/inf	140/127
1F5L	1/1	1/1	1/1	1/1
1HPV	97/1	4/1	9/1	10/1
1I7Z	1/1	1/1	1/1	1/1
1J01	1/1	1/1	1/inf	1/inf
1K1J	5/1	3/1	2/1	2/1

Сложность энергетической поверхности белок-лиганд

30 тестовых комплексов белок-лиганд

разбиваются на 2 группы:

1. **Простая поверхность** белок-лиганд (лиганд - до 8 торсионов) – 9 комплексов из 30:
   - FLM, SOL-T16, SOL-T32, SOL-T64 – все находят *один глобальный минимум*
   - Выполняется парадигма докинга
2. **Сложная поверхность** белок-лиганд:
   - *Глобальные минимумы разные для* FLM и для всех вариантов SOL-T16, SOL-T32, SOL-T64
   - Парадигма докинга не выполняется

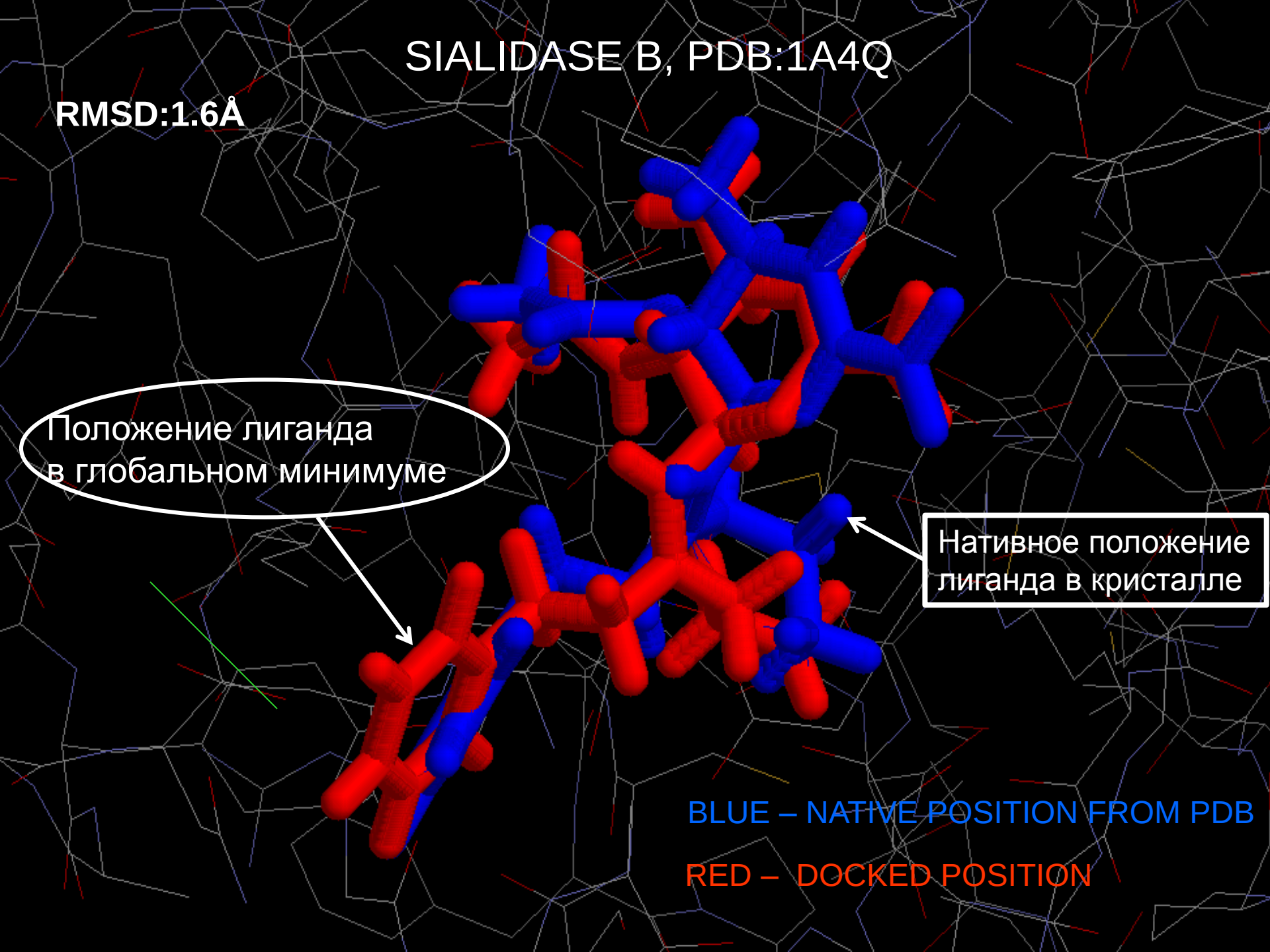
Парадигма докинга

- **Выполняется не всегда**

- Для 30% комплексов (для 9 из 30) выполняется **парадигма докинга** для комплексов с простой энергетической поверхностью в MMFF94:
  - *Локально оптимизированный нативный лиганд имеет наинизшую энергию*
  - *Найденный глобальный минимум, соответствует положению лиганда вблизи от нативного положения*

SIALIDASE B, PDB:1A4Q

RMSD:1.6Å



Времена счета FLM vs SOL-T

PDB ID	Время счета, core·hours				
	FLM	SOL-T16	SOL-T32	SOL-T64	Отношение времен SOL-T 16:32:64
1C5Y	23 059	18	205	149	1:12:9
1F5L	22 824	41	373	366	1:9:9
1I7Z	30 384	138	386	1232	1:3:9
1LQD	26 356	204	584	1898	1:3:9
1PPC	30 713	225	687	2468	1:3:11
1SQO	23 318	42	310	314	1:7:7
2PAX	15 831	23	60	244	1:3:11
3KIV	30 386	11	32	103	1:3:10
3PAX	30 713	33	81	238	1:2:7
4FSW	22 918	23	178	134	1:8:6

Выводы

1. Разработан и проверен новый алгоритм **бессеточного** докинга на основе тензорных поездов
2. Показана высокая эффективность ТТ-докинга
3. Новый алгоритм открывает возможности докинга:
   - с учетом **растворителя** в континуальных моделях
   - с учетом **подвижности атомов белка**
4. Для части тестовых комплексов показана выполнимость парадигмы докинга для целевой функции – энергии комплекса белок-лиганд в поле MMFF94 в вакууме

Спасибо за внимание

- ***…Surely every medicine is an innovation; and he that will not apply new remedies, must expect new evils…***
- ***…Каждый медицинский метод есть инновация; а кто не хочет применять новые средства, должен ждать новых бед…***

Sir Francis Bacon (1561-1626)

OF INNOVATIONS