Короткий обзор ML в моделировании

Дизайн белков, НТУ Сириус

Головин А.В. [1] [2]

[1]НТУ Сириус, ЦИИиИТ

[2]МГУ им М.В. Ломоносова, Факультет Биоинженерии и Биоинформатики

Сириус, 2022

» ML в хемоинформатике

- Улучшение анализа HTS данных, в основном регрессии
- Улучшение предсказания афинности, токсичности, фармокинетике для заданных соединений. Регресии и не только.
- Генерация новых соединений под указанную задачу

» Дескрипторы

- 0D Формула: молекулярный вес, количество атомов и связей
- 1D Химические графы: фрагменты, функциональные группы
- 2D Топология структуры: индексы Weiner, Balaban, Randic, BCUTS
- 3D Геометрия молекулы: WHIM, autocorrelation, 3D-MORSE, GETAWAY
- 4D Химическая информация: Volsurf, GRID, Raptor

» 1D

- Одномерные дескрипторы - это скаляры: количество атомов, количество связей, молекулярный вес, суммы атомных свойств или количество фрагментов
- Просты в вычислении, страдают от проблем вырожденности, когда различные соединения сопоставляются с идентичными значениями дескриптора
- Одномерные дескрипторы обычно используются вместе с многомерными дескрипторами или выражаются как вектор из нескольких одномерных дескрипторов.

» 2D

- Двумерные химические дескрипторы являются наиболее частым типом дескрипторов
- Включают топологические индексы, молекулярные профили и двухмерные дескрипторы автокорреляции
- Важной особенностью 2D-дескрипторов является инвариантность графа, когда на значения дескрипторов не влияет перенумерация узлов (вершин) графа.
- Система $Mold^2$ быстро генерирует до 200 типов 2D-дескрипторов для больших составных наборов данных.
- Коммерческие программные пакеты включают систему DRAGON, до 5000 дескрипторов.

» 3D

- 3D дескрипторы извлекают химические особенности из трехмерных геометрий и наиболее чувствительны структурным изменениям
- Могут включать дескрипторы автокорреляции, данные об заместителях , дескрипторы поверхности, объема и квантово-химические дескрипторы
- Трехмерные химические дескрипторы полезны для идентификации «каркасов» - отдельных химических каркасов со сходной связывающей активностью
- Ключевым ограничением является вычислительная сложность генерации конформеров и выравнивания структур

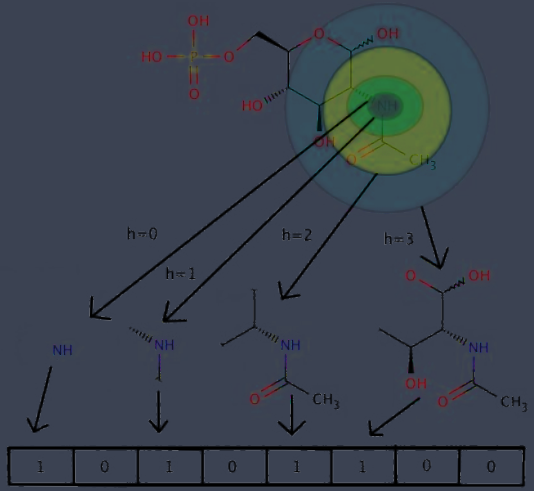
» 3D

- Предсказанные конформации могут не соответствовать соответствующим биоактивным конформациям.
- Химические дескрипторы 4D являются расширением дескрипторов 3D, которые одновременно рассматривают несколько структурных конформаций
- Эш и Фурчес применили MD киназы ERK2 для вычисления трехмерных дескрипторов по сеткой на основе траектории 20 нс и показали, что такие четырехмерные химические дескрипторы могут эффективно отличать наиболее активные ингибиторы ERK2 от неактивных с более высокой степенью обогащения.

» Fingerprints

- FP это многомерные векторы, элементами которых являются значения химических дескрипторов
- MACCS представляют собой двумерные двоичные FP (0 и 1), каждый из которых 166 бит указывает на наличие или отсутствие определенных ключей подструктуры
- Daylight FP и ECFP позволяют извлекать патерны до определенной длины или диаметра из графа структуры, и могут динамически индексировать представления с использованием хэш-функций, что часто обеспечивают более высокую специфичность.

» ECFP



10.1021/ci100050t

» Fingerprints

* Последние разработки - это continuous kernel и встроенные нейронные FP. Это внутренние представления, полученные с помощью SVM и нейронных сетей.
* Duvenaud et al. распространил концепцию свертки на молекулы, представленные в виде двумерных молекулярных графов.

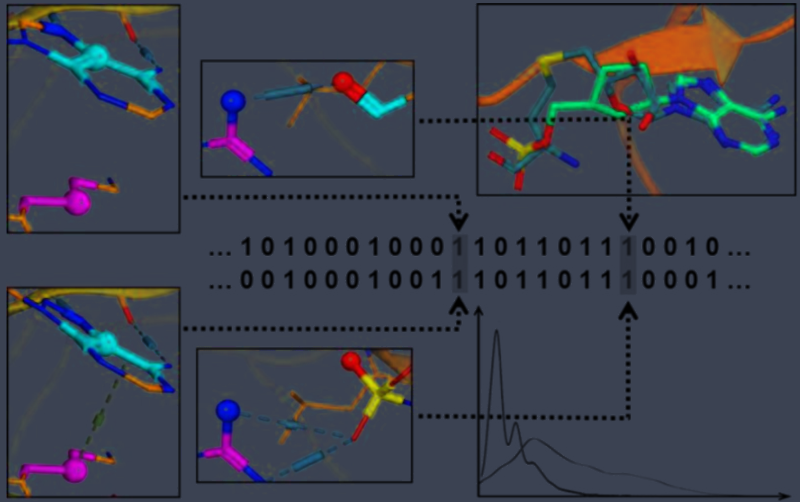
» 3D fingerprints

- 3D FP включают химические характеристики, основанные на фармакофорных паттернах, свойствах поверхности, молекулярных объемах или взаимодействия молекул
- MIF, реализованное в GRID. FP на основе MIF помещает лиганд в сетку с фиксированным интервалом и вычисляет электронный, стерический и гидрофобный вклад независимо в каждой точке сетки.

» 3D fingerprints

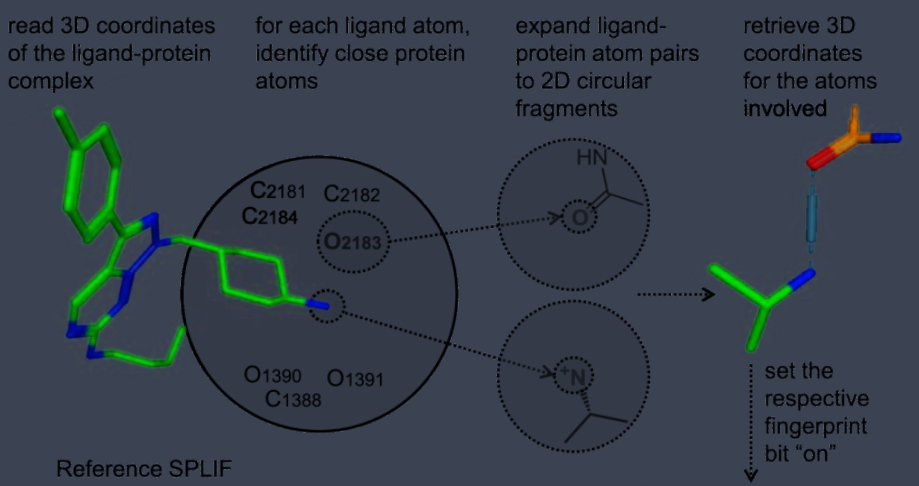
- FP на основе MIF можно затем использовать в сравнительном анализе молекулярного поля (CoMFA) путем установления взаимосвязей между точками трехмерной сетки и активностями соединения.
- Зависимость от относительной ориентации молекул внутри сетки является основным ограничением.
- Баскин и Жохова недавно представили подход непрерывного молекулярного поля (CMF), который заменяет сетку непрерывной функцией

» SPLIF



10.1021/ci500319f

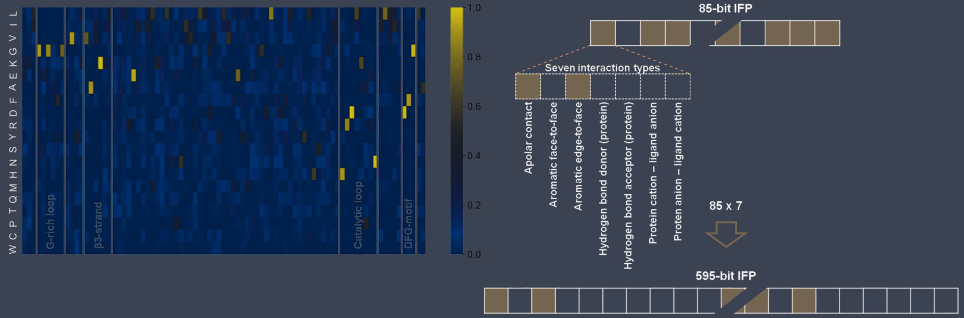
» SPLIF


read 3D coordinates of the ligand-protein complex
for each ligand atom, identify close protein atoms
expand ligand-protein atom pairs to 2D circular fragments
retrieve 3D coordinates for the atoms involved
C2181
C2182
C2184
O2183
O1390
O1391
C1388
HN
O
N
set the respective fingerprint bit “on”
Reference SPLIF

» Ингибиторы киназ

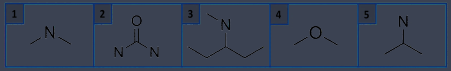
* Киназы ключевой класс ферментов в передаче сигналов
* Известно много ингибиторов киназ
* Сайт связывания очень похож, это связывание ATP

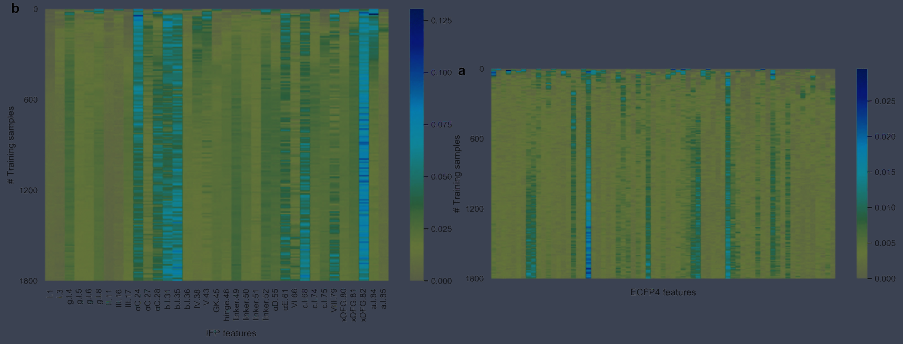
» Interaction fingerprints



Включение мишень-специфичной информации через IPF улучшило предсказание связывания примерно на 10% по сравнению с использованием традиционных FP.
doi.org/10.1186/s13321-020-00434-7

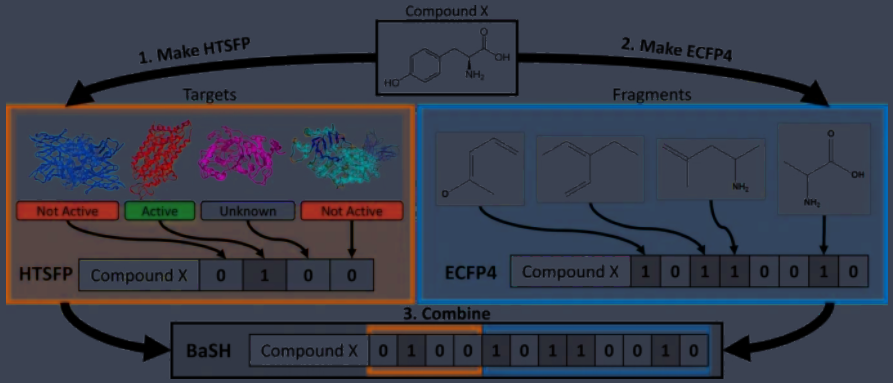
» Features





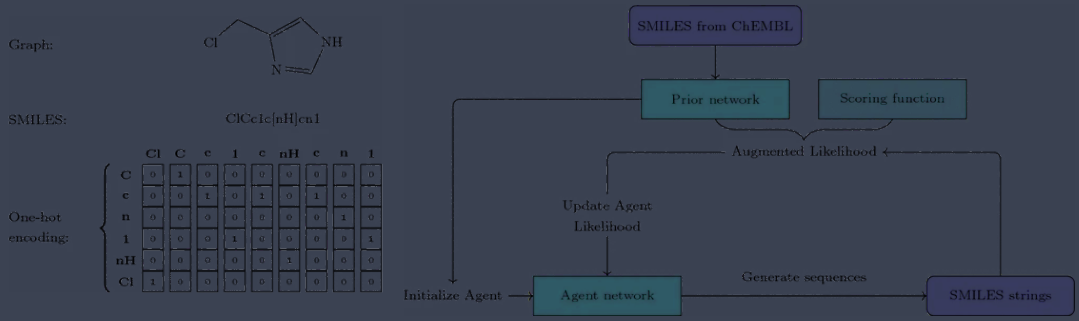
doi.org/10.1186/s13321-020-00434-7

» Включение bioassay



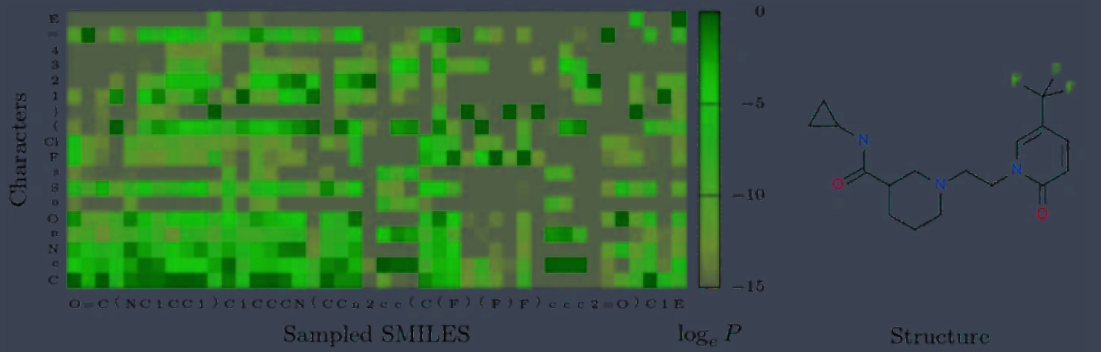
Комбинация структурных и данных об активности позволяет улучшить производительность и показывает эффективный переход между scafolds
10.1186/s13321-019-0376-1

» Агент



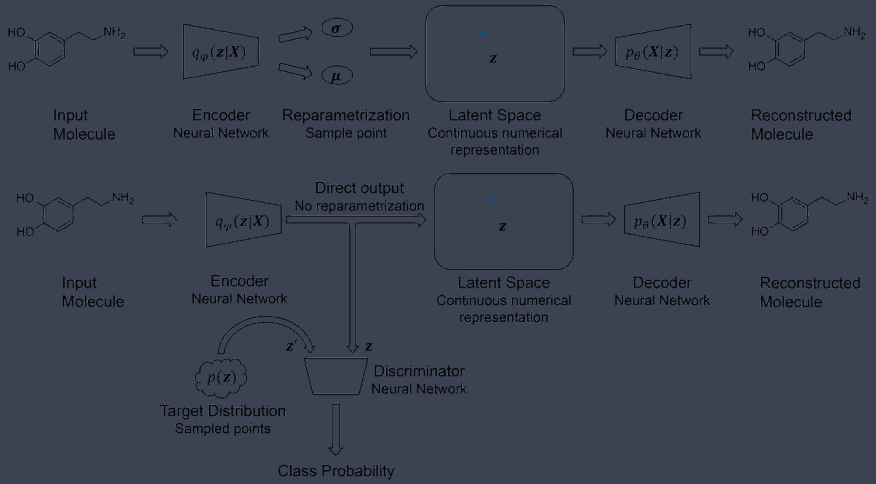
doi.org/10.1186/s13321-017-0235-x

» Генерация по подобию



doi.org/10.1186/s13321-017-0235-x

» Generative Autoencoder

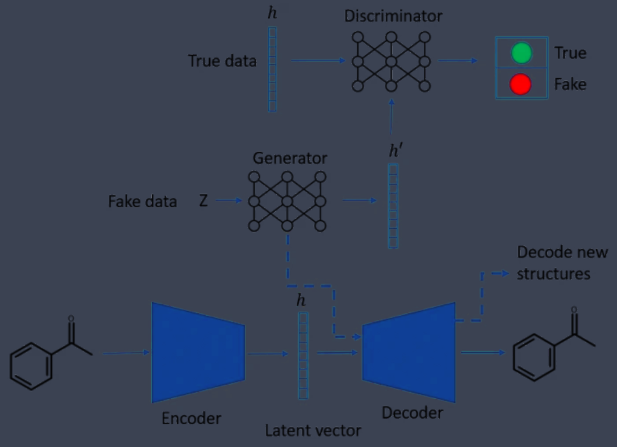


разработан для использования с RNN

» AE и GAN

* Предварительно обученный автокодер использовался для сопоставления молекулярной структуры с латентным вектором
* GAN был обучен с использованием латентных векторов в качестве входных и выходных данных
* После обучения GAN выбранные скрытые векторы были отображены обратно в структуры

» AE и GAN


h
Discriminator
True data
True
Fake
h′
Generator
Fake data
Z
Decode new
structures
h
Encoder
Latent vector
Decoder

» DeepChem

* Имеет множество модулей для Featurization молекул
* Упрощенное использование TensorFlow, Pytorch и тд
* Коллекция рецептов

» OpenChem

Model

`Smiles2Label`
- Prediction of properties from sequential input

`Graph2Label`
- Prediction of property from molecular graphs

`GraphRNNModel`
- Generation of molecular graphs

`GenerativeRNN`
- Generation of SMILES strings

Module

`Embedding`
- Token embeddings
- Positional embeddings

`Encoder`
- Recurrent encoder (RNN, GRU, LSTM)
- Convolutional encoder
- Graph convolutional encoder

`MLP`
- Multi-layer perceptron with custom activation functions

Task
- Classification
- Regression
- Multi-task
- Generation

Data layer
- SMILES string
- Protein sequences

Graphs
- Molecular graphs

» Направления для работы

- Методы основанные на подобии рассматривают подобные вещества и белки, равномерное распределение отсутсвует.
- Описание features сделать количественным.
- Методы основаны на datasets. Нужна адаптация под успешные предсказания.

» Направления для работы

* Объеденение баз данных. Комбинирование макисмально доступного количества данных для пары белок-ингибитор.
* Правильно включение структурно-функциональных данных для лигандов и белков.

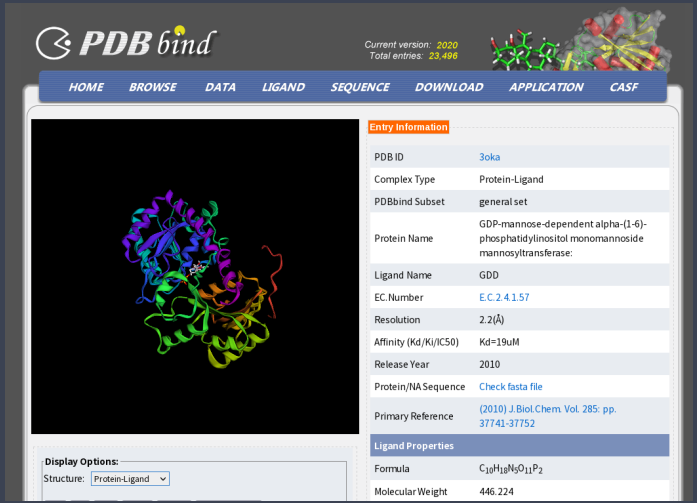
» Основные направления применения ML

- Перевзвешивание (Rescoring)
- Подготовка библиотеки из химического разнообразия
- Поиск сайтов связывания, выявление правил взаимодействия
- Генерация новых соединений с высокой аффинностью

» Перевзвешивание

- Докинг достаточно производительный метод, можно получить много данных о положении лигандов в белке
- Классические оценочные функции это парные взаимодействия
- Результат геометрии связывания лиганда можно оценить с помощью оценочной функции из ML

» Данные, PDBbind



19,443 Записей

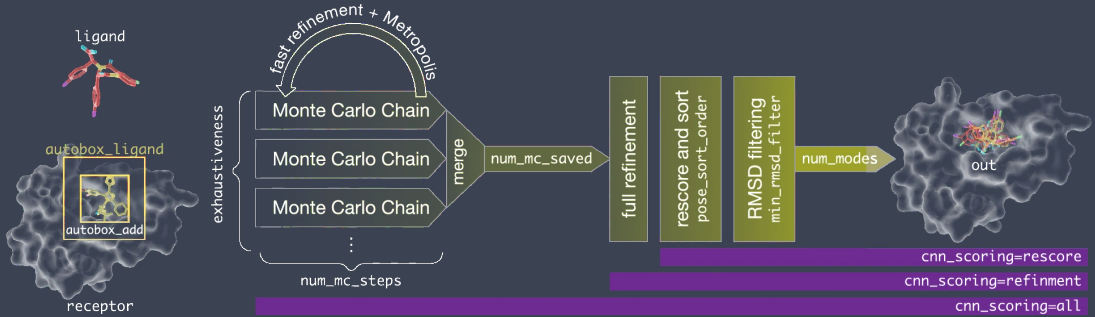
» Представление

- CNN исходно оптимизированы на изображения
- Сайт связывания представляется как 3D сетка
- Атомам присваиваются типы, 30 у лигнадов, 16 у белков
- Атомы представляются как распределение плотности

$$A(d,r) = \begin{cases} e^{-\frac{2d^2}{r^2}} & 0 \le d < r \\ \frac{4}{e^2 r^2} d^2 - \frac{12}{e^2 r} d + \frac{9}{e^2} & r \le d < 1.5r \\ 0 & d \ge 1.5r \end{cases}$$

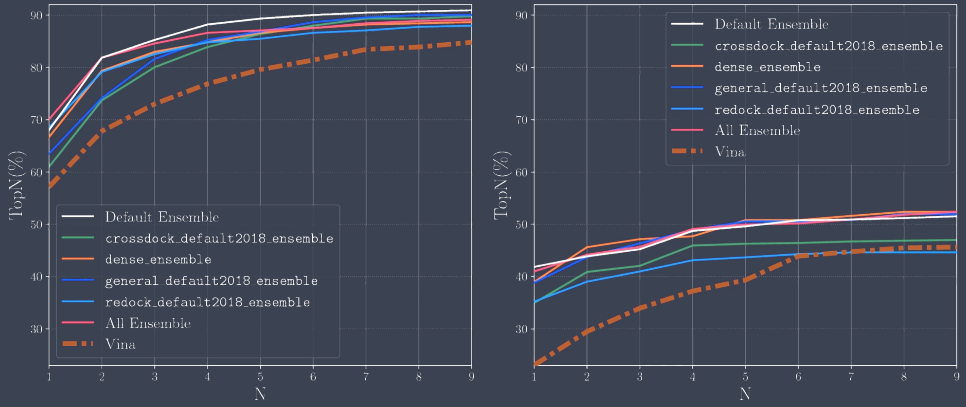
10.1021/acs.jcim.6b00740

» GNINA



10.1186/s13321-021-00522-2

» GNINA



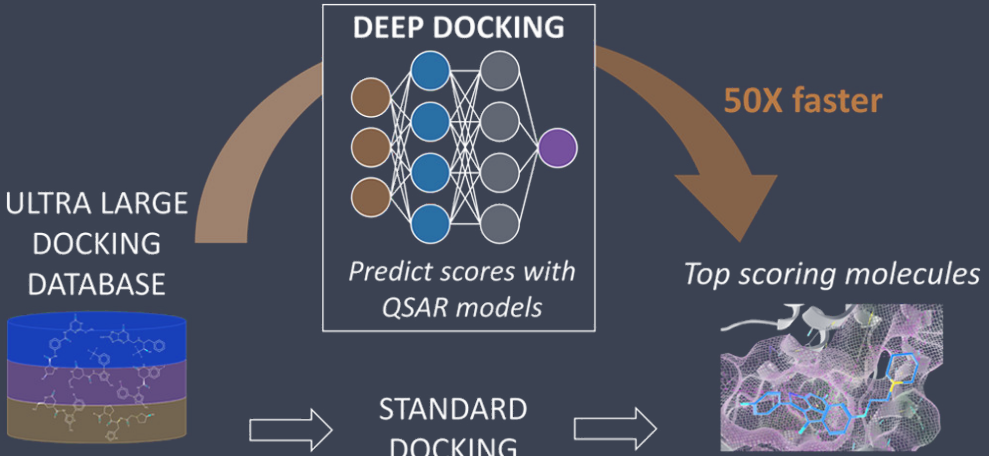
(a) Redocking Ensembles (b) Cross-docking Ensembles

TopN is the percentage of targets ranked above or at N with a RMSD less than 2 Å , 10.1186/s13321-021-00522-2

» Профилирование библиотек

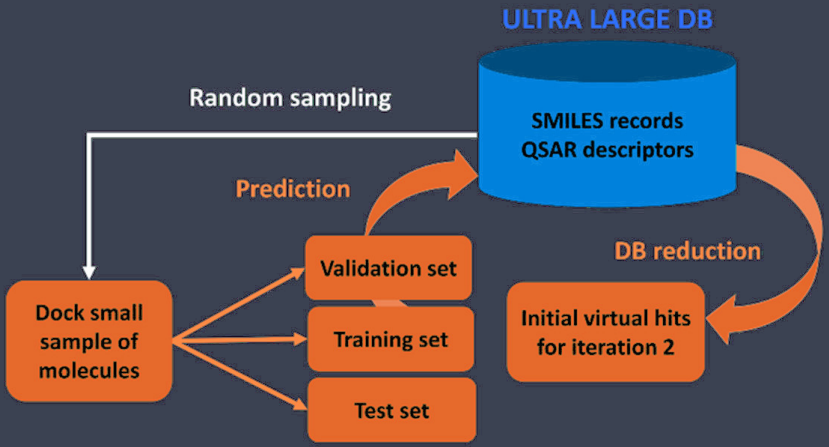
* Химическое разнообразие сравнимо с 10^{23}
* Даже первичная генерация молекул требует гигантских ресурсов
* Профилирование разнообразия под конкретную задачу

» Deep Docking, шаг 1

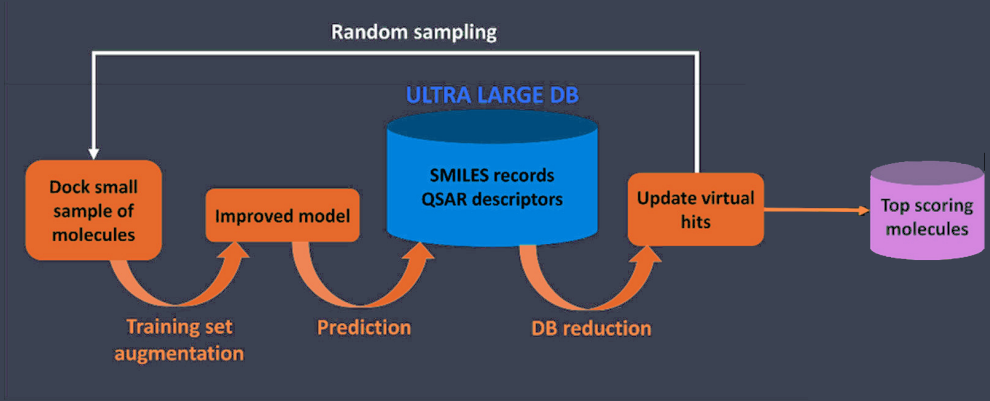


10.1021/acscentsci.0c00229

» Deep Docking, шаг 1



» Deep Docking, шаг 2



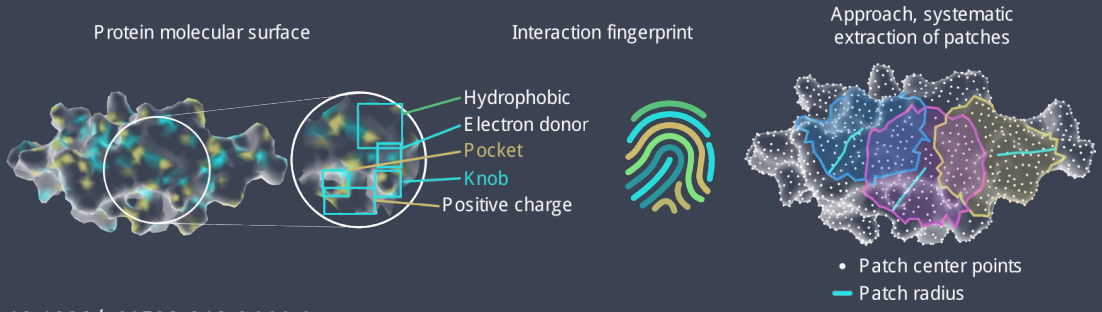
» Данные, представление, архитектруа

- Отбор из базы данных, $10^3 - 10^6$
- Используются молекулярные QSAR дексрипторы, точнее Morgan fingerprints
- ”feed-forward”DNN сети в библиотеке Keras

» Выявление сайтов связывания

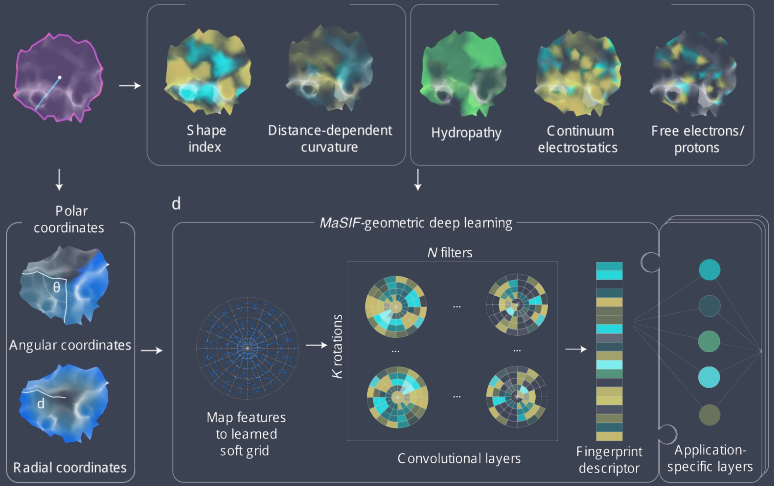
- Поверхность белка определяет взаимодействие
- В поверхности важна не только ”геодезия”но и типы атомов
- Составление ”fingerprints”
- Обучение, выявление потенциальных сайтов и поиск подходящих отпечатков

» MASIF, общая идея



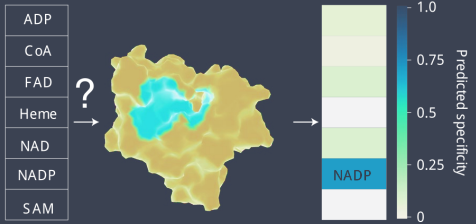
10.1038/s41592-019-0666-6

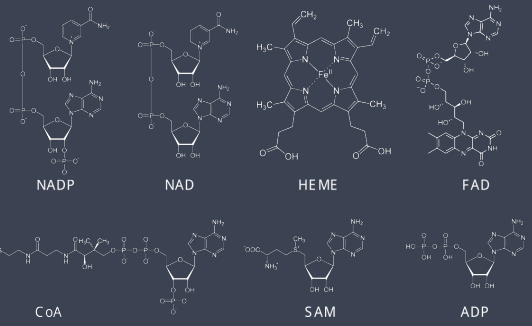
» MASIF, реализация



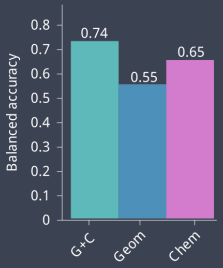
Данные:PDBbind, SAbDab

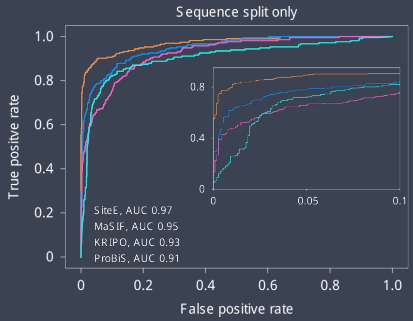
» Результат

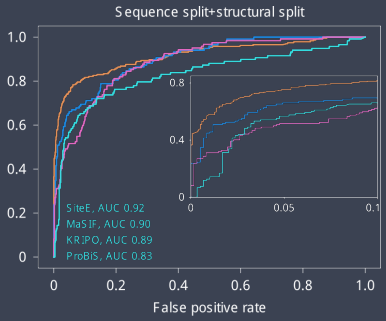

ADP
CoA
FAD
Heme
NAD
NADP
SAM
?
NADP
1.0
0.75
0.5
0.25
0
Predicted specificity


NADP
NAD
HEME
FAD
CoA
SAM
ADP

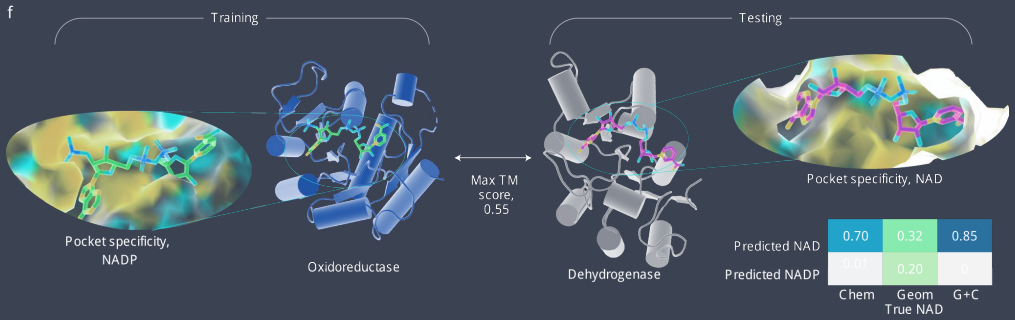
» Аккуратность







» Пример

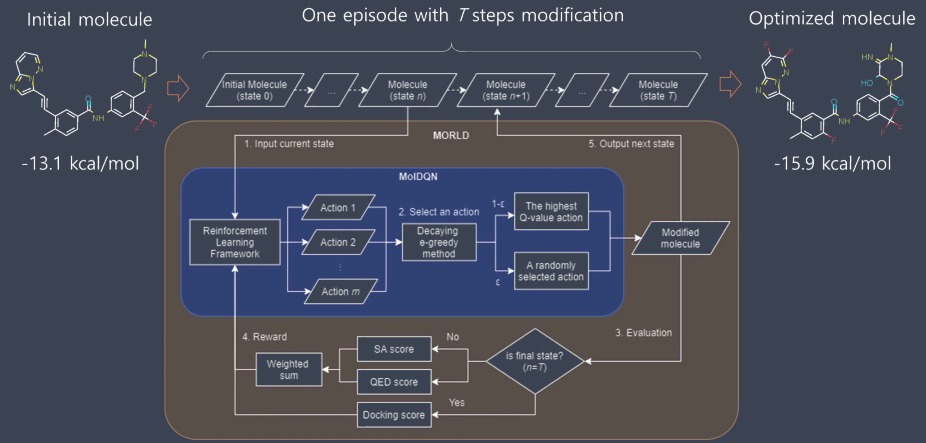


https://github.com/lpdi-epfl/masif

» Оптимизация соединений

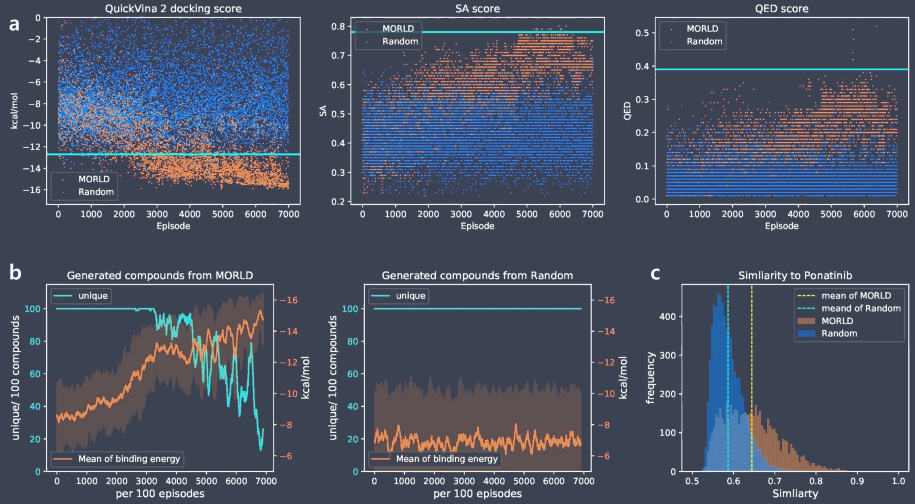
- Наличие соединения в сайте связывания
- Обучение происходит в процессе работы
- Вознаграждение и штрафы приходят из докинга

» MORD

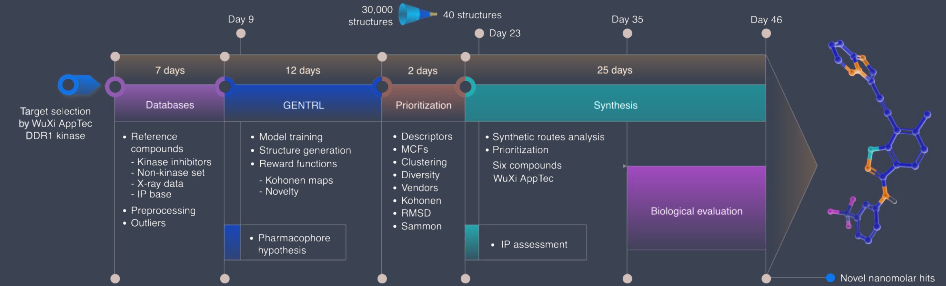


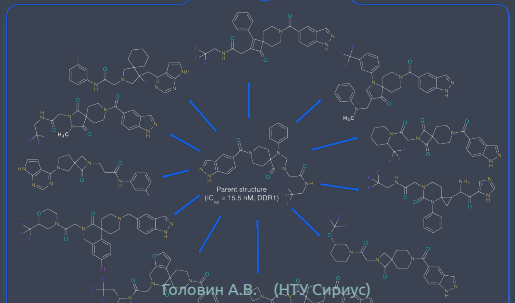
10.1038/s41598-020-78537-2

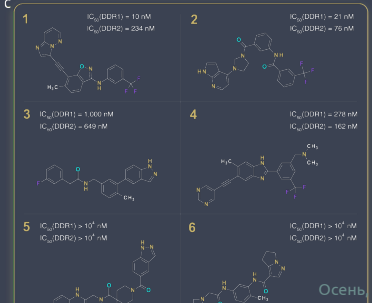
» MORLD, результаты



» GENTRL


30,000 structures
40 structures
Day 9
Day 23
Day 35
Day 46
7 days
12 days
2 days
25 days
Target selection by WuXi AppTec DDR1 kinase
Databases
GENTRL
Prioritization
Synthesis
• Reference compounds
- Kinase inhibitors
- Non-kinase set
- X-ray data
- IP base
• Preprocessing
• Outliers
• Model training
• Structure generation
• Reward functions
- Kohonen maps
- Novelty
• Pharmacophore hypothesis
• Descriptors
• MCFs
• Clustering
• Diversity
• Vendors
• Kohonen
• RMSD
• Sammon
• Synthetic routes analysis
• Prioritization
Six compounds WuXi AppTec
• IP assessment
Biological evaluation
Novel nanomolar hits


Parent structure
(IC50 = 15.5 nM, DDR1)


c
1
IC50(DDR1) = 10 nM
IC50(DDR2) = 234 nM
2
IC50(DDR1) = 21 nM
IC50(DDR2) = 75 nM
3
IC50(DDR1) = 1,000 nM
IC50(DDR2) = 649 nM
4
IC50(DDR1) = 278 nM
IC50(DDR2) = 162 nM
5
IC50(DDR1) > 10^4 nM
IC50(DDR2) > 10^4 nM
6
IC50(DDR1) > 10^4 nM
IC50(DDR2) > 10^4 nM

» История

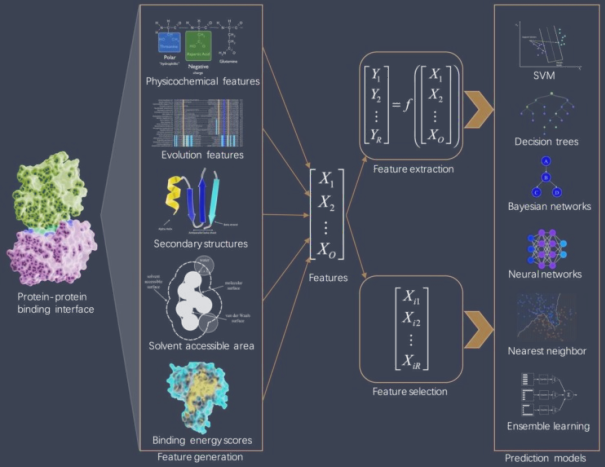
- Первые работы в 2001 году
- Суть задачи: для двух данных последовательностей предсказать взаимодействие
- Типы представлений: состав, доменный состав, мотивы, профили гидрофобности, геномные особености, филогенетические особености

» 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"

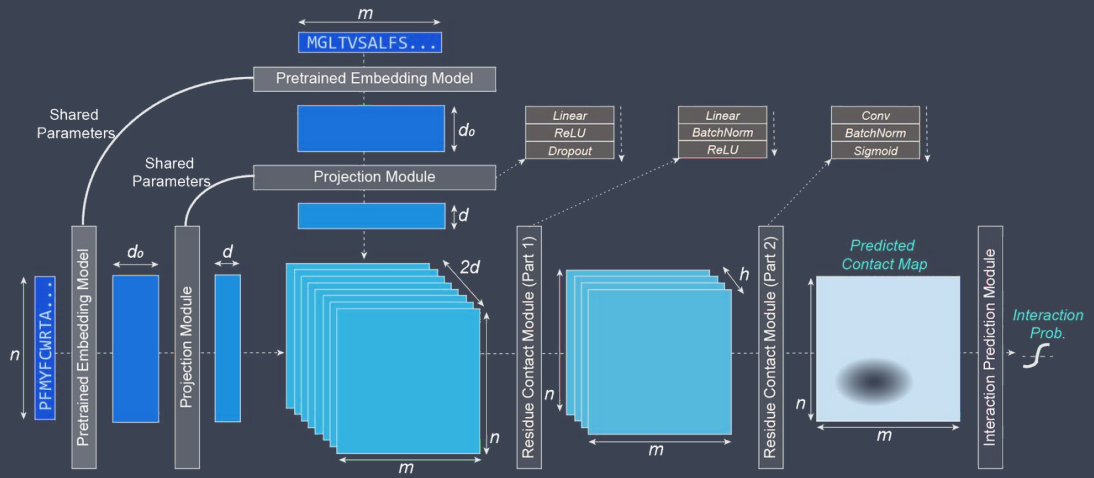


10.3390/molecules23102535

» Достижения

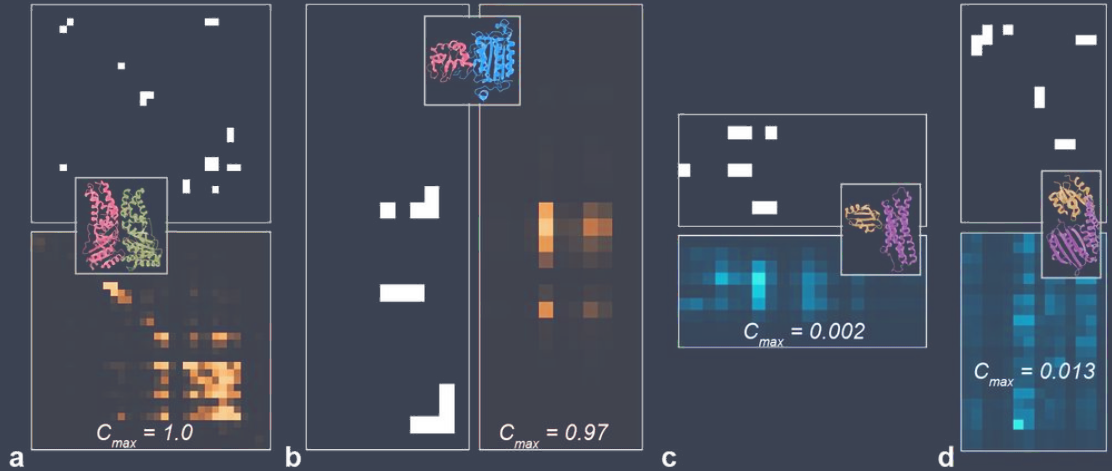
- Существенный прогресс, но есть еще ”вызовы”
- ddG из экспериментов не унифицированно
- Малое наполнение данными
- Часто случается переобучение
- Предикторы ’hot spots’ используют последовательность и структурную информацию для представления, но 3D не используется полностью
- *Перспективным считается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

» D-SCRIPT



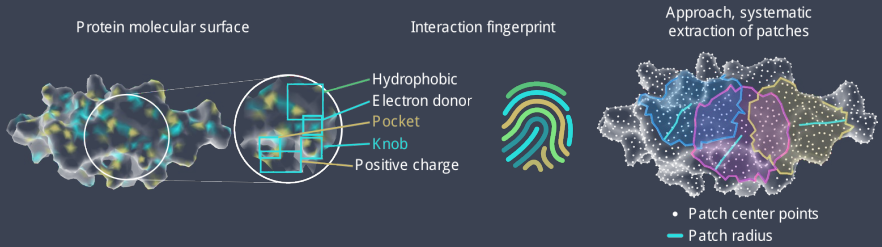
10.1016/j.cels.2021.08.010

» D-SCRIPT, результат



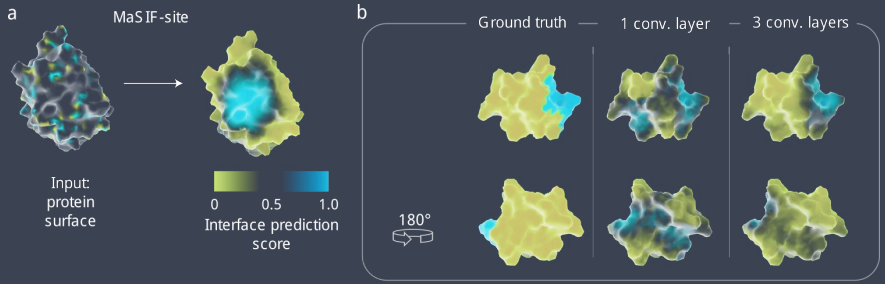
*When D-SCRIPT correctly predicts an interaction, its contact maps are significantly similar to the ground truth.

» MASIF

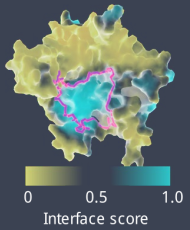


MaSIF-site: классификатор, на входе поверхность белка, на выходе прогнозируемая оценка для каждой вершины поверхности на вероятность участия в PPI

» MASIF, предсказание участков

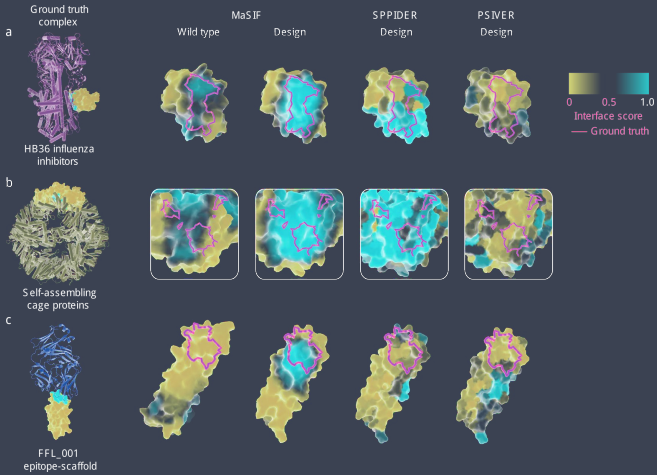


» MASIF, сравнение



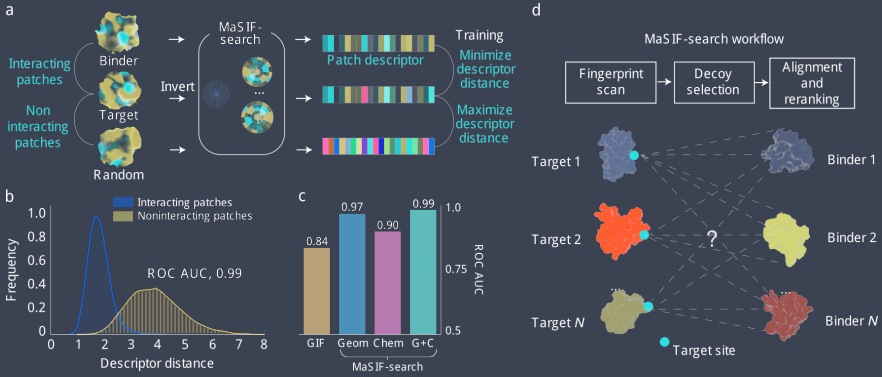
* PSIVER - Naïve Bayes classifier (NBC) and a kernel density estimation method (KDE)
* SPIDER - SAS метрика + MSA,

» 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 hydropathy.

» ”FATALITY”or not?

Table 1 | **Results for large-scale docking benchmark 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			t ime (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

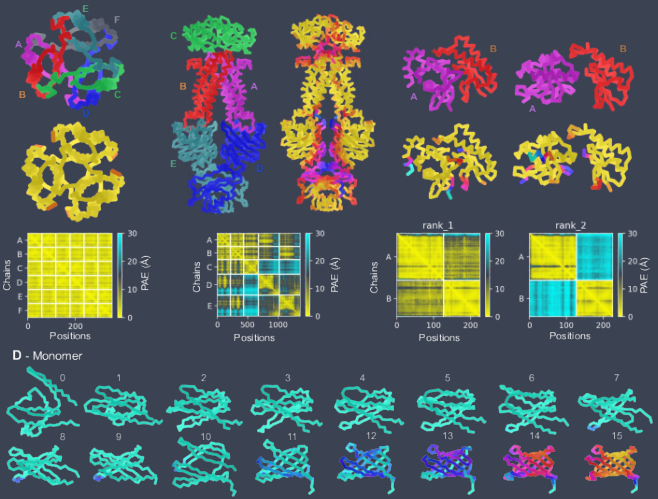
No. of solved complexes in the top, number of target–binder complexes within 5 Å iRMSD found in the top 100, top ten or top one (for holo cases) or top 1,000, top 100 and top ten (for apo cases). Time (min), CPU time in minutes for each program, which excludes precomputation time for MaSIF-search.

Table 2 | **Results for large-scale docking benchmark 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			t ime (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.

» ColabFold



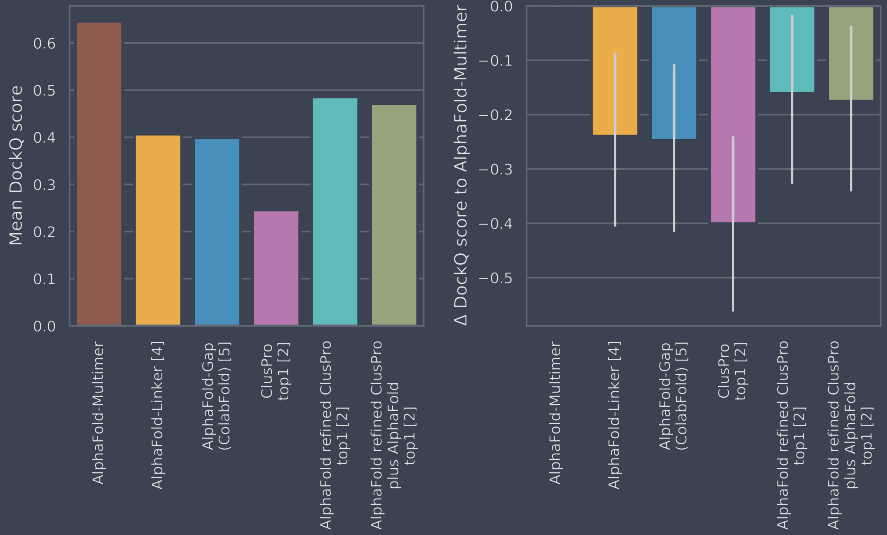
10.1101/2021.08.15.456425v1.full.pdf

» AlphaFold-Multimer

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

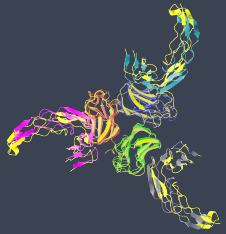
10.1101/2021.10.04.463034v1

» AlphaFold-Multimer

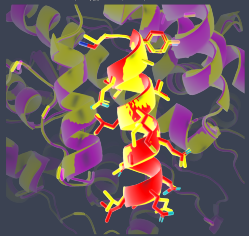


» AlphaFold-Multimer

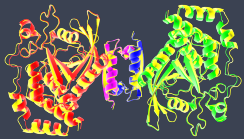
(a) A2B2C2 heteromer
TM-score = 98.0, N_{res} = 1,246, PDB ID = 6E3K



(b) A3B3 heteromer
TM-score = 89.3, N_{res} = 795, PDB ID = 7KID



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



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

» AlphaFold-Multimer