

# Лекция 2. Введение в дизайн белков

Курс: Методы машинного обучения в дизайне белков

Головин А.В.<sup>1</sup>

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

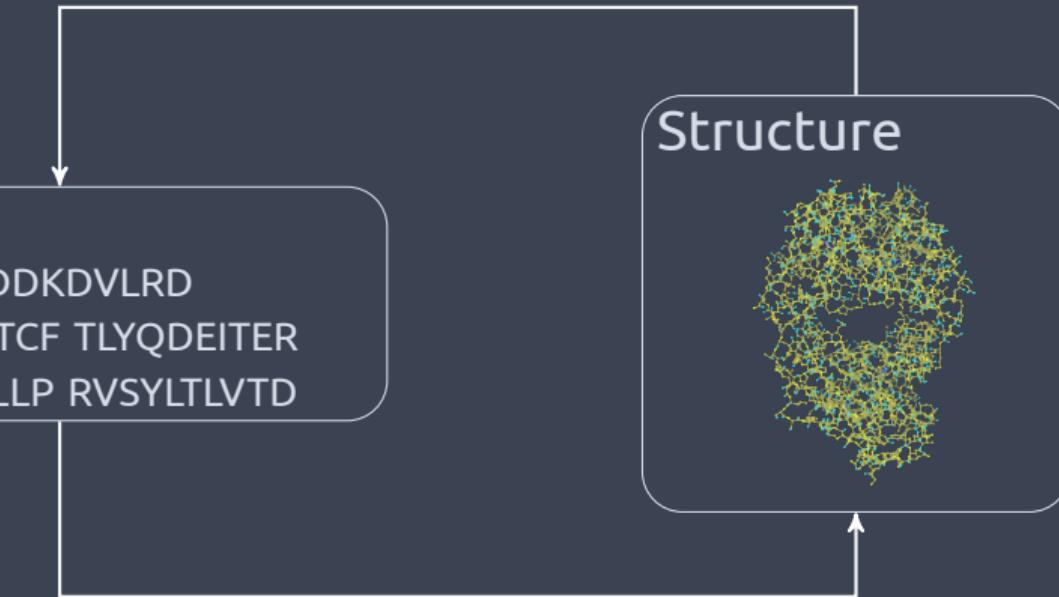
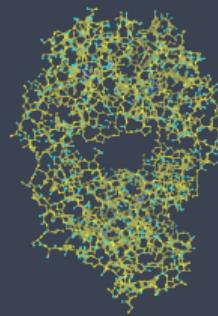
Москва, 2023

# » Main task of protein design

Sequence

MTDDKDVLRD  
VWFGRIPTCF TLYQDEITER  
EAEPYYLLLP RVSYLTIVTD

Structure



## » Основные проблемы:

- \* Монте-Карло: 100 а.к.  $3N$  степеней свободы, получаем  $10^{48}$  конформаций.
- \* **Парадокс Левинталя:** "Промежуток времени, за который полипептид приходит к своему скрученному состоянию, на много порядков меньше, чем если бы полипептид просто перебирал все возможные конфигурации".
- \* Для решения разумно использовать накопленные знания для моделирования.



## » Последовательность-структура

### Причины парадокса Левинталя:

- \* Теоретические модели, не соответствуют тому, что природа старается оптимизировать;
- \* В ходе эволюции были отобраны только те белки, которые легко сворачиваются;
- \* белки могут сворачиваться разными путями, не обязательно следуя глобально оптимальному пути.
- \* Считается, что структура определяется последовательностью, но иногда нужны другие факторы.
- \* Структура более консервативна чем последовательность



## » Сравнительное моделирование

- \* Зачем искать конформации если можно представить, что при подобии последовательностей подобны и структуры.
- \* Надо оценить насколько вероятно, что отличие в последовательности может привести изменению способа укладки цепи.
- \* Надо отфильтровать ошибки полученные при определении структуры.



## » Известные структуры и последовательности

- \* Сейчас известно порядка  $1.6 \times 10^5$  структур уникальных белков.
- \* UniProt это 562 000 белоков.
- \* Для 50% последовательностей можно предсказать способ укладки.



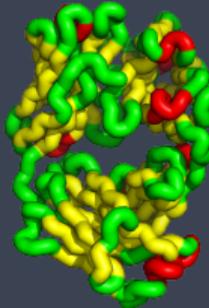
# » Степень идентичности и сравнительное моделирование

0

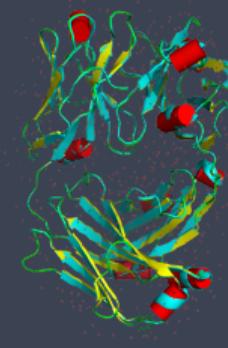
30

60

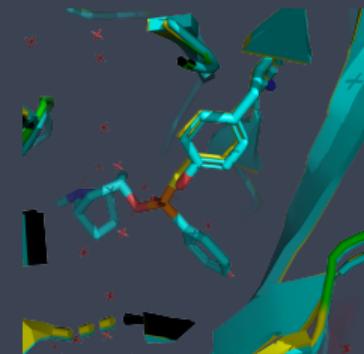
100



Фолд, мотивы



MR, Мутагенез



Докинг

Sali,A. & Kuriyan,J. Trends Biochem. Sci. 22, M20–M24 (1999)



## » Как это реализовать?

- \* Надо найти белок заготовку с известной структурой.
- \* Построить первичное выравнивание.
- \* Улучшить выравнивание.
- \* Построить ход основной цепи.
- \* Моделирование петель
- \* Достроить/моделировать положение боковых радикалов
- \* Проверка модели



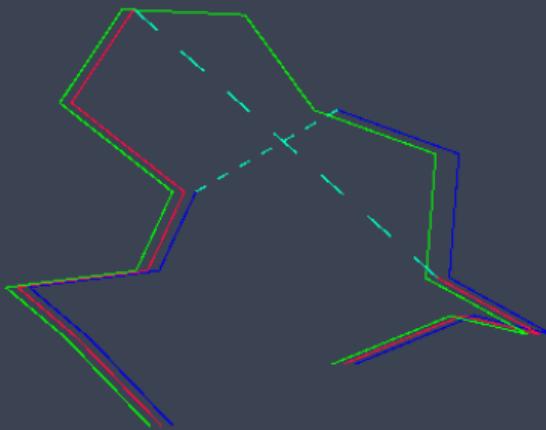
## » Поиск белка заготовки

- \* Поиск по PDB с помощью:
  - \* Blast
  - \* Psi-Blast
  - \* Методов распознавания упаковки
- \* Используя биологическую информацию.
- \* Функциональное аннотирование в базах данных.
- \* Используя информацию об активных сайтах, или мотивы.



# » Улучшение выравнивания

1	2	3	4	5	6	7	8	9	10	11	12	13	14
PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL	CYS
PHE	ASN	VAL	CYS	ARG	THR	PRO	---	---	---	GLU	ALA	ILE	CYS
PHE	ASN	VAL	CYS	ARG	---	---	---	THR	PRO	GLU	ALA	ILE	CYS



Из книги "Professional Gambling" от Gert Vriend



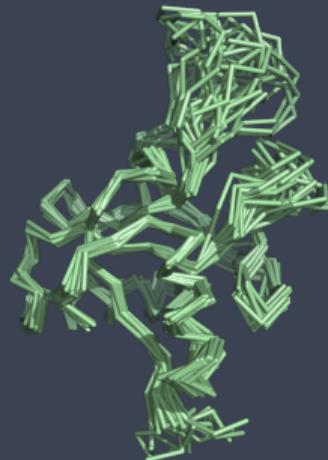
## » Качество белка заготовки

- \* Выбор качественного белка заготовки очень важен.
- \* Лучший вариант не обязательно обладает лучшей степенью идентичности.
  - \* Белок 1: ID 93%, 3.5 ангстрема разрешение. Хуже.
  - \* Белок 2: ID 90%, 1.5 ангстрема разрешение. Лучше!

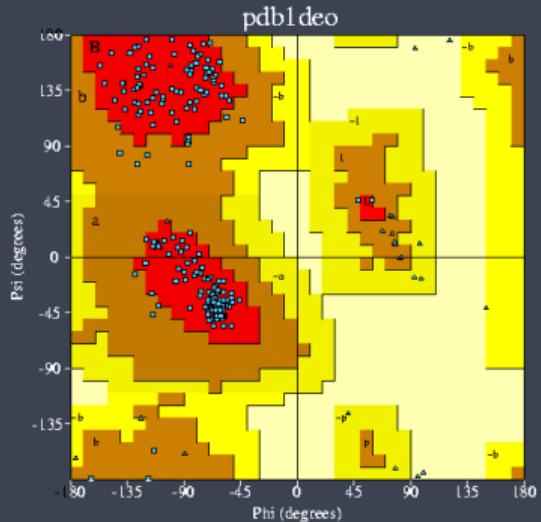


## » Если структура белка заготовки получена ЯМР

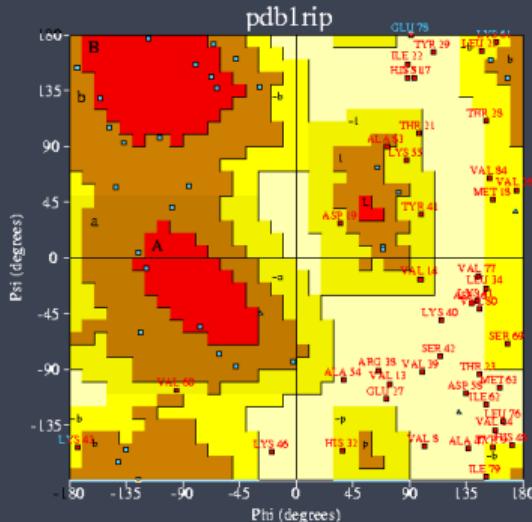
- \* Определимся какие области определены лучше.
- \* Соотнесём с выравниванием.
- \* Если низкая гомология выпадает на “подвижные” области, то структура подходит.



## » Качество заготовки, Рамачандран



PCA, хорошая структура.



ЯМР, сомнительные данные.



## » Построение остова

- \* Генерируем координаты остова моделируемого белка для остатков из выравненных областей.
- \* Не обязательно использовать координаты, могут подойти дистанционные ограничения.
- \* Большинство исследователей предпочитают Modeller. Modeller использует дистанционные ограничения.



## » Моделирование петель

- \* Эмпирическое моделирование:
  - \* Поиск подходящего фрагмента по PDB
  - \* Использовать базы данных (LIP, etc..)
- \* Молекулярная механика.
- \* Монте-Карло.
- \* Rosetta:
  - \* Поиск фрагментов близких по последовательности.
  - \* Комбинирование результатов поиска с помощью Монте-Карло.

Комбинации выше перечисленных.



## » Моделирование боковых радикалов

- \* Если идентичность последовательностей высока то можно ожидать высокую консервативность третичных контактов.
- \* Если анализ показывает, что важные контакты консервативны то:

**Лучше оставить конформацию боковых радикалов из заготовки чем моделировать.**



## » Моделирование боковых радикалов

- \* Конформация боковых радикалов зависит от конформации основной цепи.
- \* Существуют базы данных ротамеров.
- \* Некоторые исследователи считают, что SCWRL метод самый удачный.

Это эмпирический метод на основе теории графов.

<http://dunbrack.fccc.edu/SCWRL3.php>



## » Точность моделирования боковых радикалов

- \* Высокая точность моделирования достигается для боковых радикалов внутри глобулы.
  - \* Причина: в экспериментах остатки на поверхности более подвижны.
  - \* Вычислительное проще упаковать гидрофобные остатки, чем учесть полярные контакты и водородные связи с водой или с участием воды.



## » Улучшение модели

- \* Методы минимизации энергии.
- \* Моделирование молекулярной динамики (оптимизация гидрофобики)
- \* Моделирование Монте-Карло.
- \* Любой известный подход для оптимизации структуры.



## » Ошибки

- \* Обычно ошибки не исправляются на последующих этапах моделирования.
  - \* Хорошее выравнивание не исправит плохой выбор белка заготовки.
  - \* Хорошее моделирование петель не исправит плохое выравнивание.
- \* При обнаружении ошибки необходимо повторять некоторые этапы.



## » Проверка

- \* Большинство программ для моделирования по гомологии выдают правильные значения для связей и валентных углов.
- \* Карта Рамачандрана в большинстве случаев для модели выглядит также, как для белка заготовки
- \* Проверка на ориентацию или положение заряженных остатков может быть полезна.
- \* Использование любых экспериментальных данных:
  - \* Остатки активного центра.
  - \* Места модификаций.
  - \* Места контактов.

ProQ сервер оптимизирован на поиск правильной модели а не нативной структуры.



## » Ресурсы для гомологичного моделирования

- \* Modeller
- \* SwissModel
- \* Eva-CM
- \* Nest И т.д.



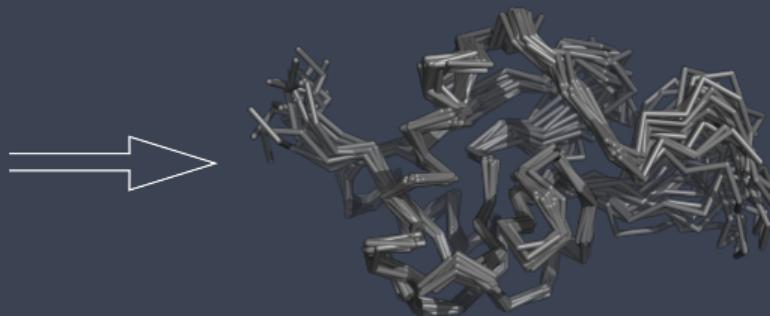
## » Предсказание структуры белка *Ab initio*

- \* Теоретически можно использовать молекулярную динамику.
- \* Моделирование отжига, как в МД так и в Монте-Карло.
- \* На основе фрагментов, Rosetta



## » *Ab initio, Rosetta*

- \* Метод использует информацию о предсказании вторичной структуры
- \* Сравниваем фрагменты от 3 до 9 остатков с библиотекой известных структур. Строим эти фрагменты.
- \* Соединяем эти фрагменты и используем Монте-Карло для оптимизации третичной структуры.



## » Ab initio, Rosetta

- \* Для определения хорошей конформации использую специальные потенциалы, которые делают модель похожей на нативную
- \* Что можно использовать:
  - \* Потенциалы для третичных контактов
  - \* Гидрофобные потенциалы
  - \* Потенциал для уменьшения радиуса вращения молекулы
  - \* Водородные связи и т.д.

Можно добавить знание об дисульфидных мостиках, местах связывания катионов металлов и т.д.

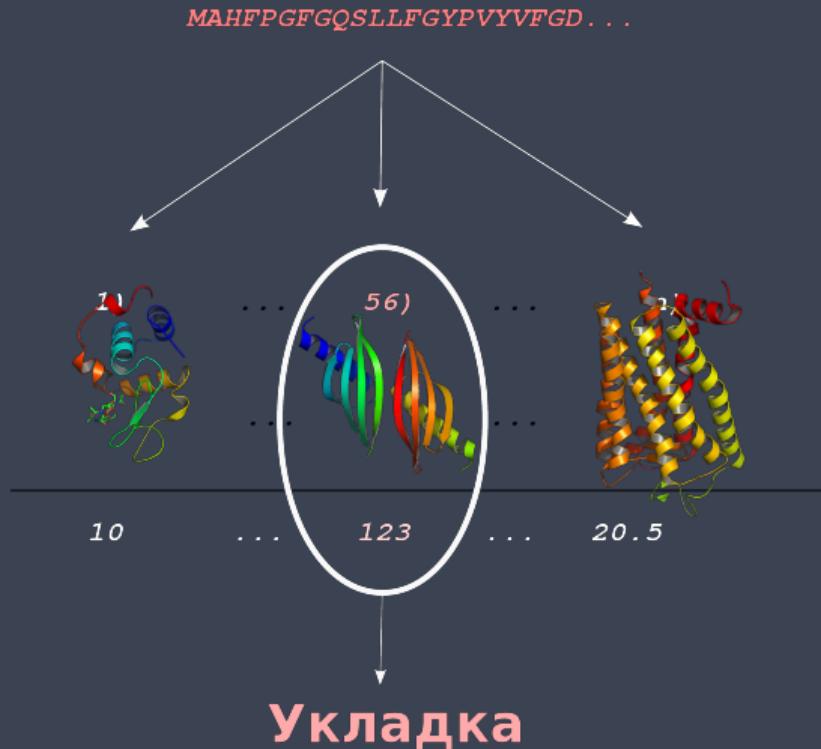


## » Threading — протягивание нити

- \* Сравниваем последовательность со всеми известными способами укладки.
- \* Используем потенциалы для определения тенденций в известных способах укладки.
  - \* Каждую аминокислоту из модели помещаем в позиции белков разных укладок
  - \* Определяем как хорошо эта аминокислота подходит белку заготовке на основе парных взаимодействий
  - \* Но основе суммарного результата определяем белок заготовку.



# » Threading — протягивание нити



## » Threading — недостатки

- \* Взаимодействия в белке не всегда описываются парными контактами.
- \* Потенциалы часто основываются на профилях последовательностей.

Есть гибридные методы Rosetta/Threading: I-Tasser



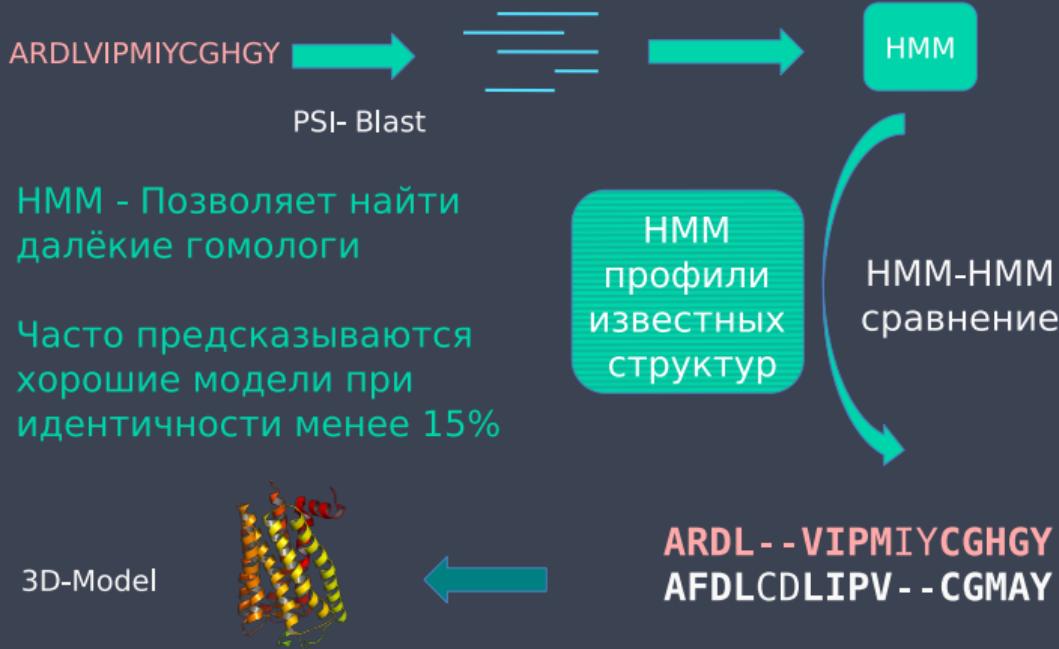
## » Распознавание укладки, Phyre2



## » Phyre2



## » Phyre2

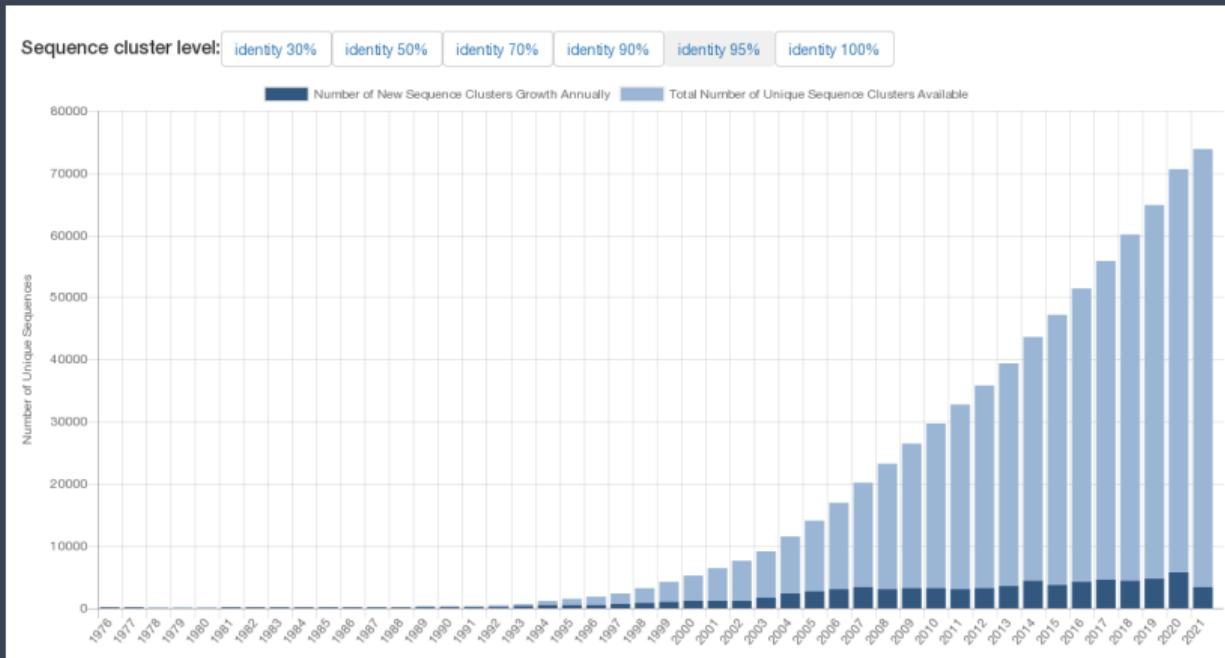


## » Мета серверы

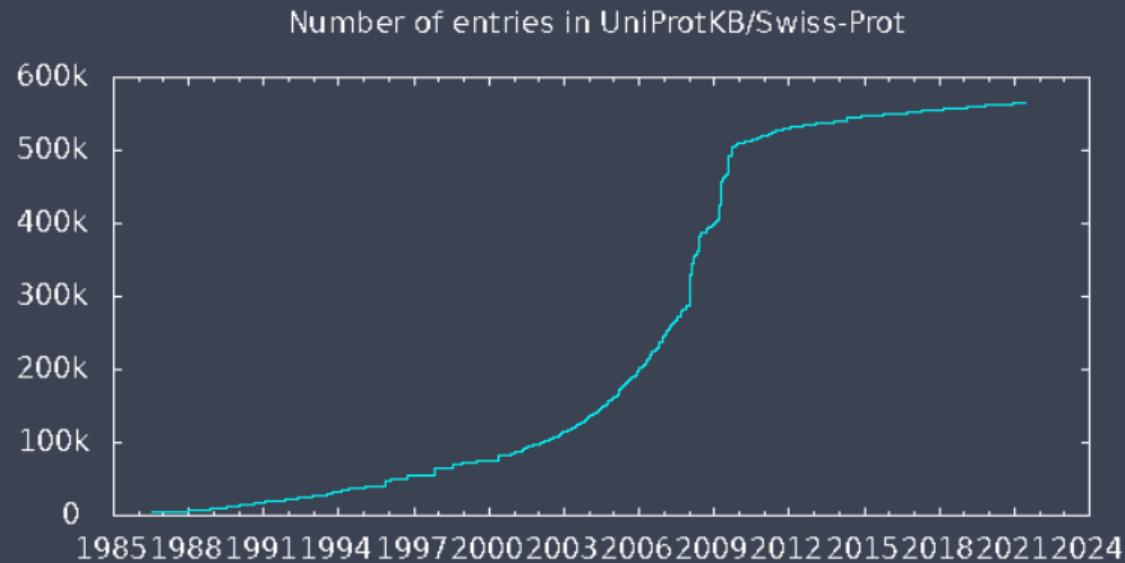
- \* Сравнение разных методов.
  - \* Большинство методов предсказывают правильную укладку в первых 10-20 результатах.
- 
- \* Удаление структур с высоким значением параметров модели, но с единственной укладкой.
  - \* Суперпозиция результатов, взвешивание.
  - \* Часто выдают только позиции атомов остова.



## » PDB

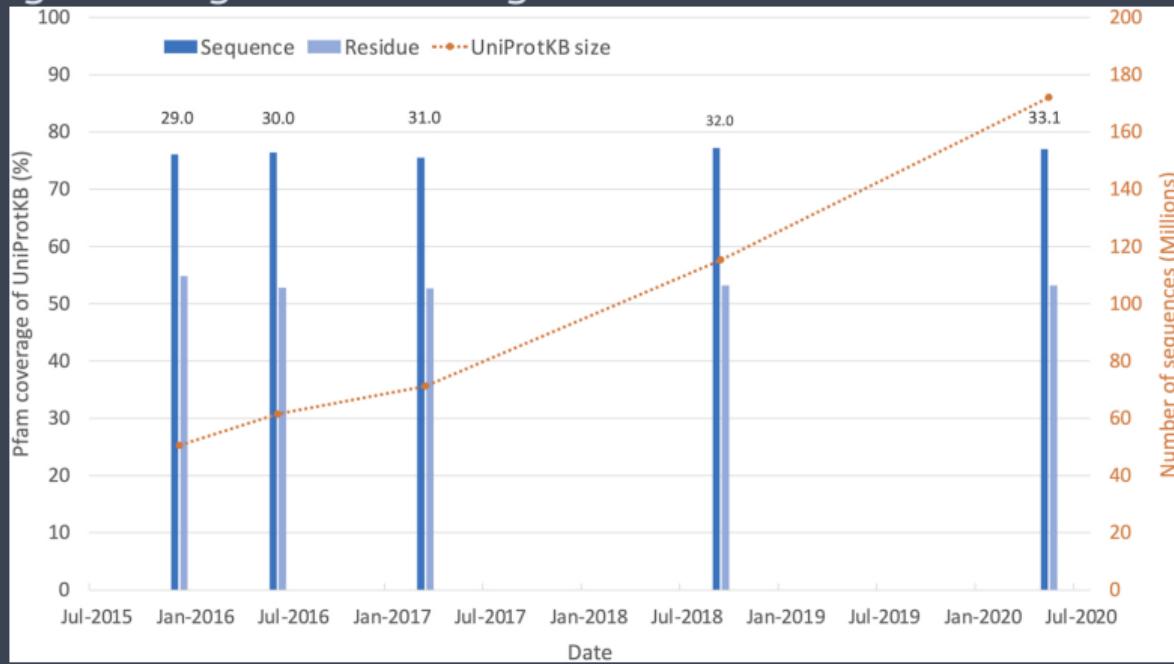


# » UNIPROT



# » PFAM

Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models



# » Разнообразие

## Sequence Feature

Amino Acid Sequence  
Chemical Descriptors  
Evolutional Information

## Learned Feature

Variational Auto-Encoders  
ProtVec  
UniRep

## Protein Graph

Voxelization

Torsion Angles

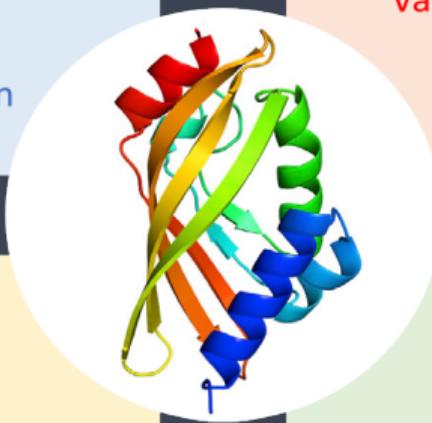
Full Structure

## Protein Surface

Secondary Structure

Inter-residue Distance

Structural Feature



10.1016/j.patter.2020.100142



## » Представление последовательности

- \* Естественное представление это аминокислота = целое число
- \* Можно добавить MSA, PSSM как реальное число
- \* Вторичная структура как 3 или 8 букв
- \* Данные об коэволюции



## » Экстракция представлений

- \* NLP алгоритмы: Word2Vec, Doc2Vec, BioVec, ProtVec
- \* Неперекрывающиеся трипептиды
- \* mLSTM (RNN), фиксированое описание для пептидов
- \* BERT и GPT3 хорошо сработали для предсказания вторичной структуры
- \* AE и VAE были удачно применены для связи последовательности со стабильностью



## » Представление структуры

- \* Прямое использование координат атомов затруднительно
- \* Voxels, 3D сетка окружения для CNN
- \* Торсионные углы, малые изменения сильно меняют структуру
- \* Попарные расстояния или карты контактов
- \* Графы для GNN, можно отделить ферменты от белков, предсказания интерфейсов
- \* Представление поверхности, MASIF



## » Оценочные функции и силовые поля

- \* MM Силовые поля достаточно хороши для стандартных взаимодействий
- \* ML используется для внедрения квантовых явлений при сохранении производительности
- \* Точность может достигать очень затратных QM методов.
- \* SchNet, ANI-1x, PhysNet

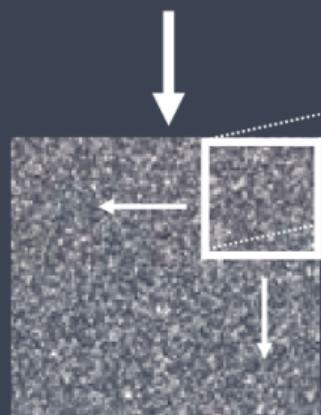


# » Convolutional NN

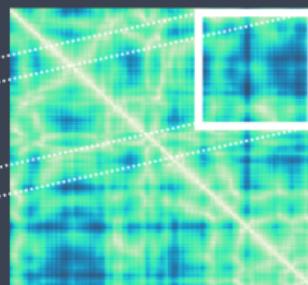
## Convolutional Neural Network

Sequence:

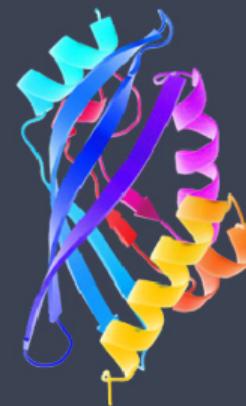
MGSSHH .....



Convolutional  
Kernel



Simulated  
Annealing



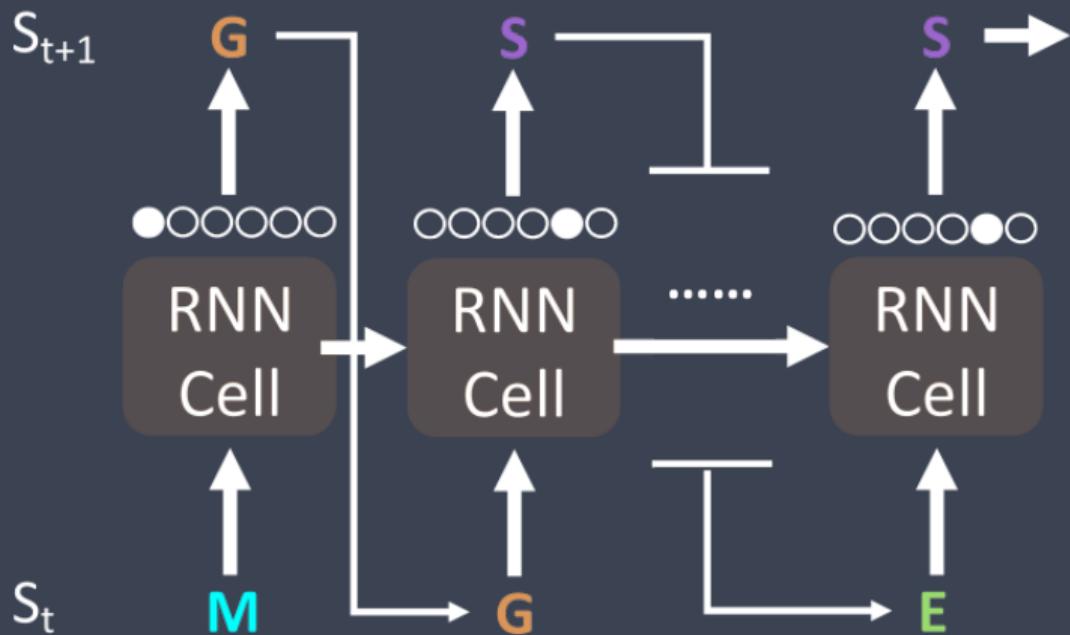
Pairwise Features

Matrix

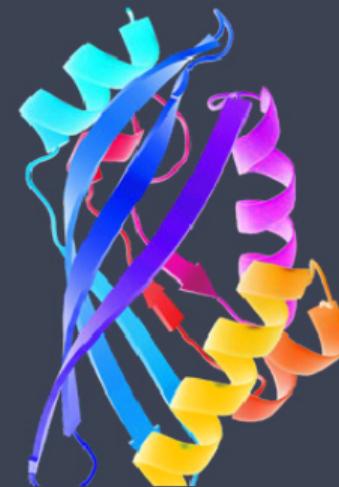
[10.1016/j.patter.2020.100142](https://doi.org/10.1016/j.patter.2020.100142)



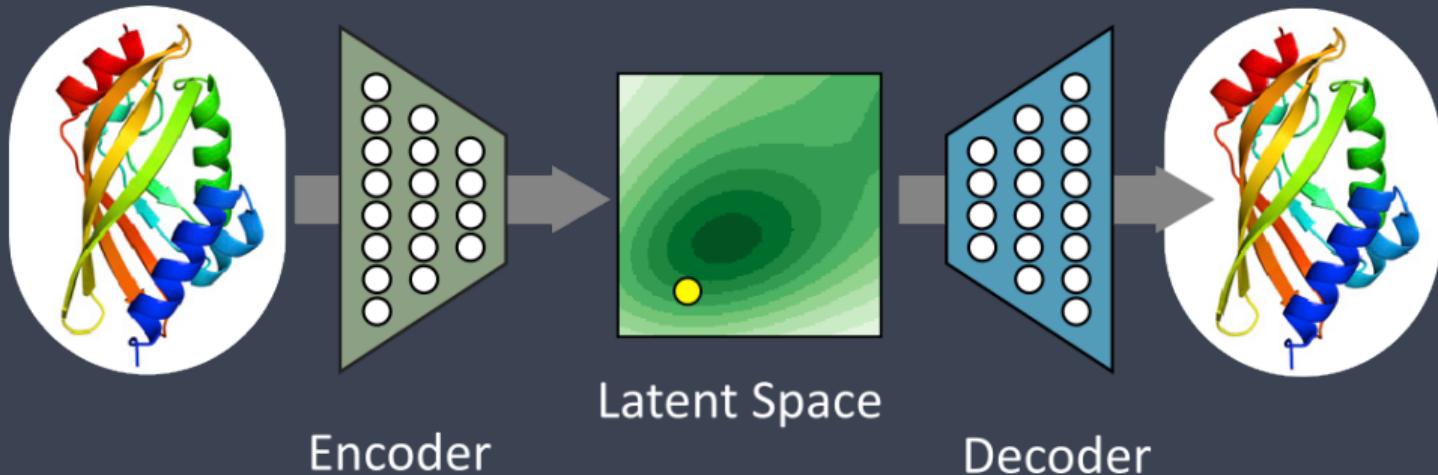
## » Recurrent NN



Sequence:  
MGSSHH .....



# » Variational Auto Encoder



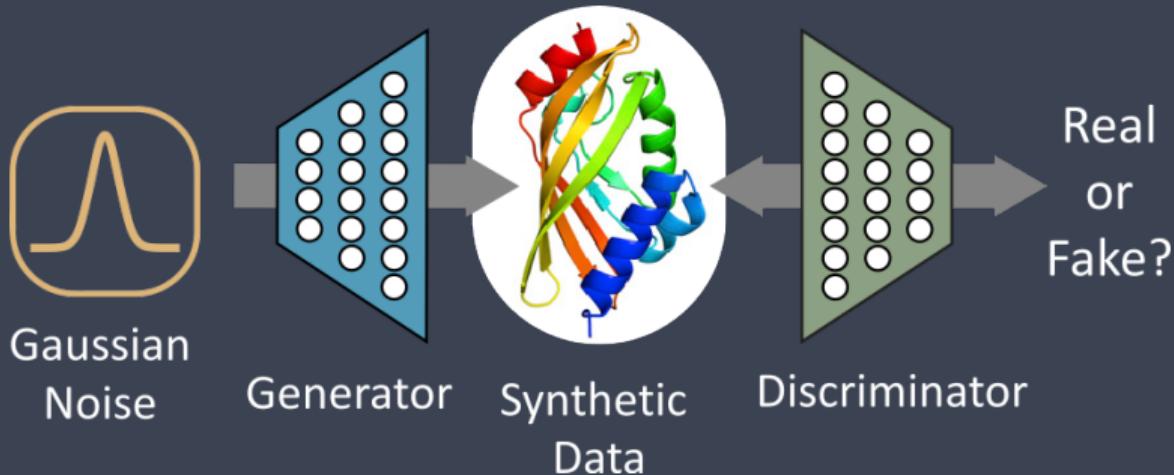
Encoder

Latent Space

Decoder



# » Generative Adversarial Network



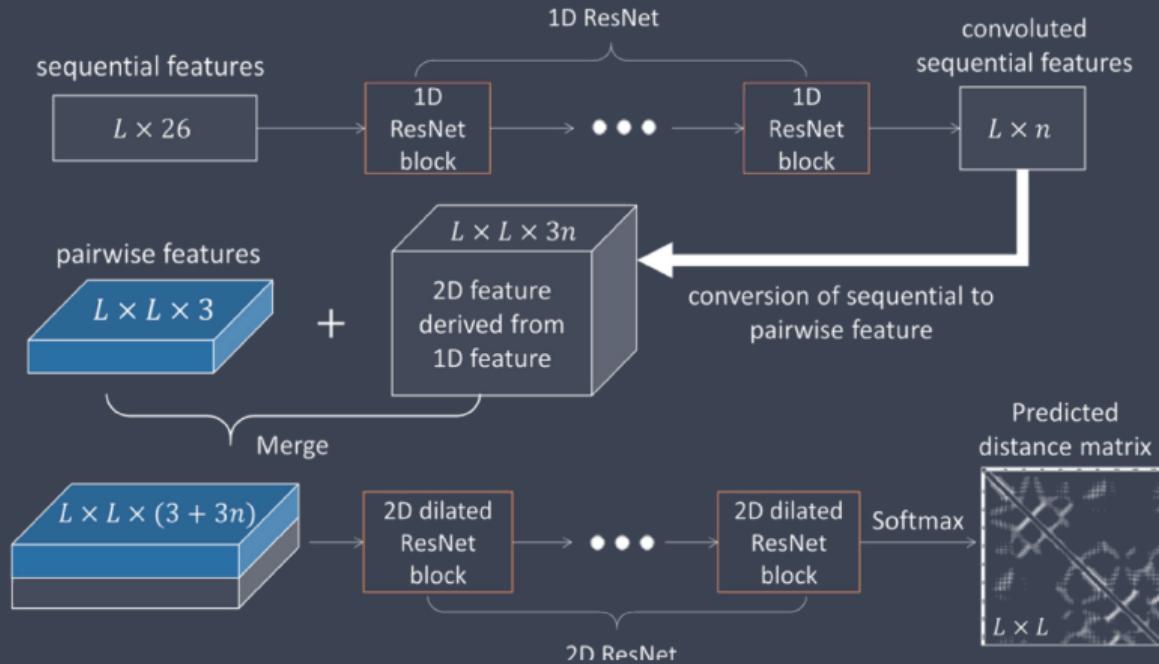
## » Основная идея

- \* Классические методы опираются на силовые поля и сложные протоколы
- \* Новая идея: контактирующие остатки эволюционируют вместе
- \* Нужна информация об гомологах, большие MSA
- \* RaptorX и AlphaFold



# » Архитектуры, расстояния

A

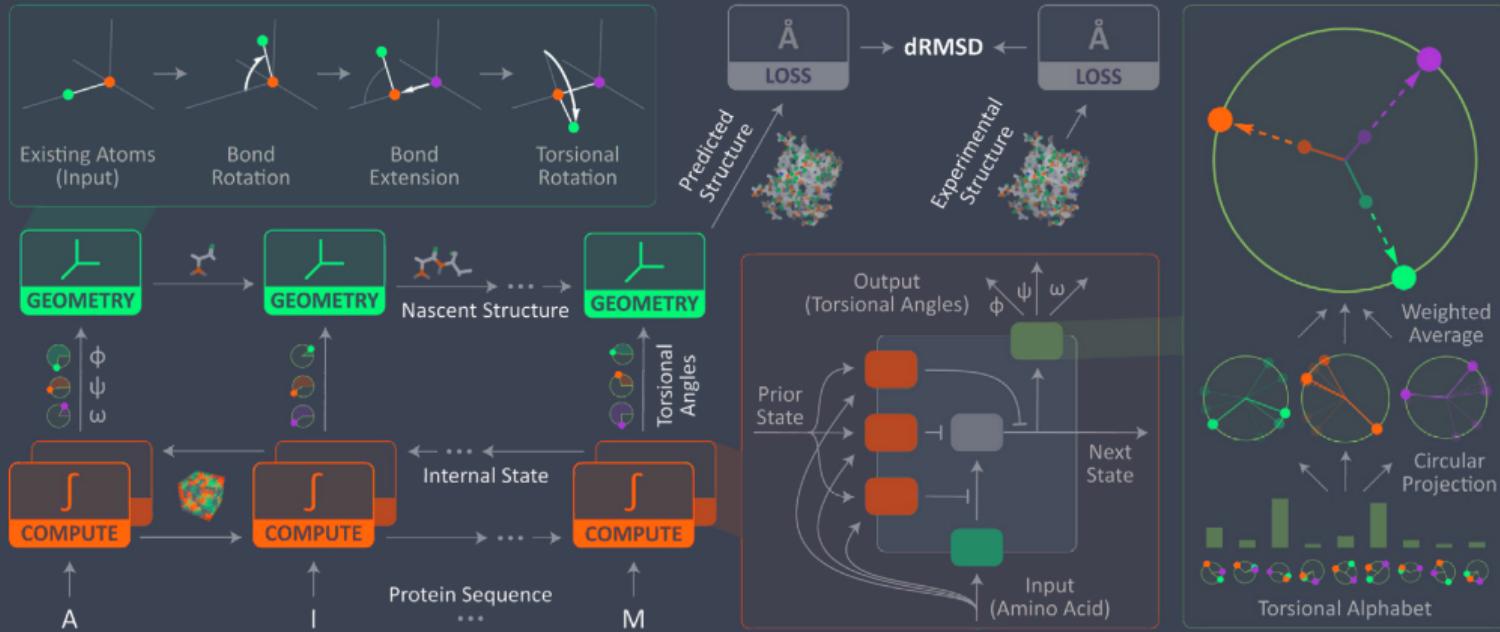


10.1002/prot.25810



# » Архитектуры, end2end

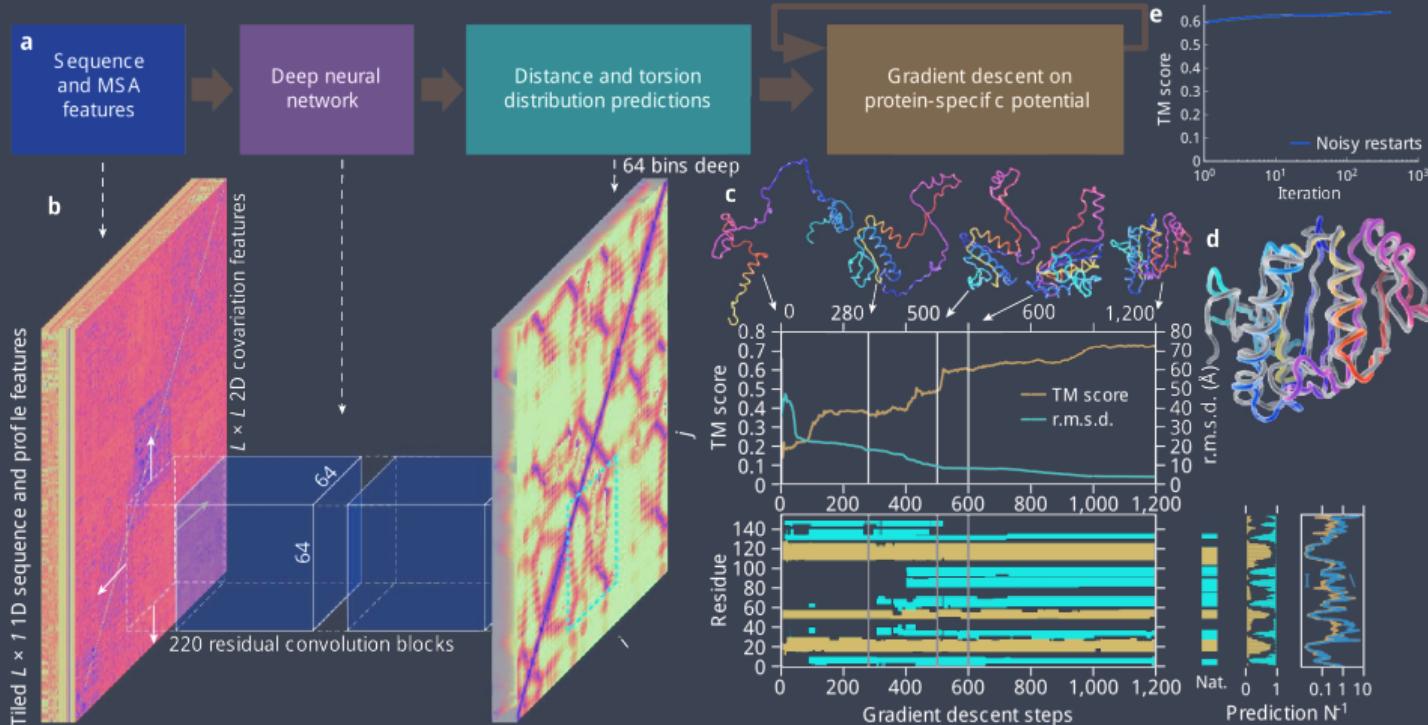
B



10.1016/j.cels.2019.03.006



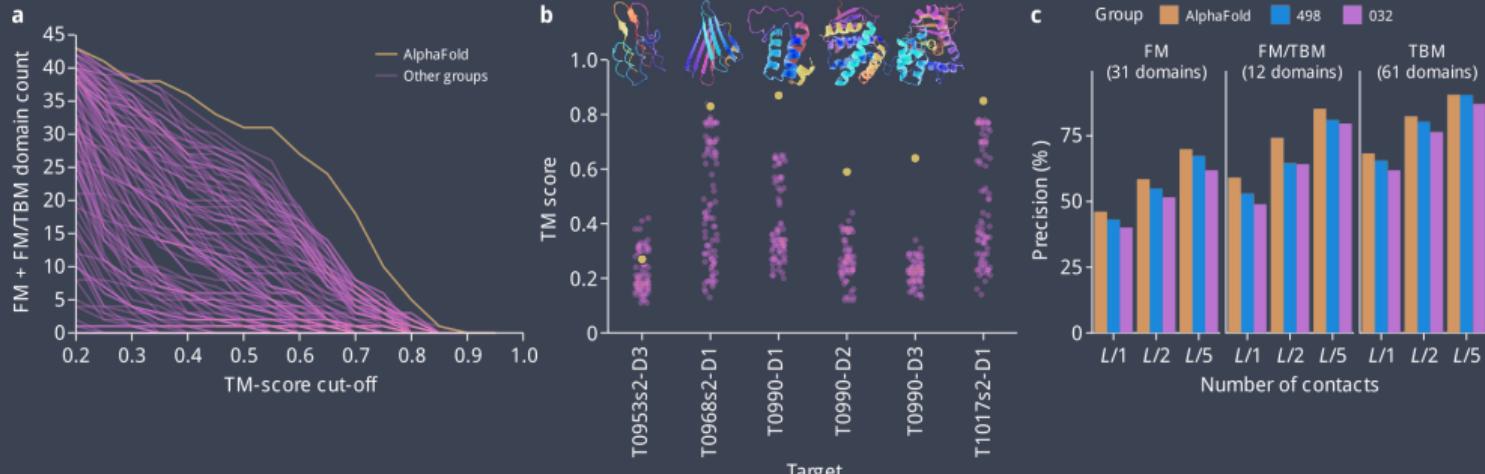
# » AlphaFold 1, идея



10.1038/s41586-019-1923-7



# » AlphaFold 1, CASP



10.1038/s41586-019-1923-7

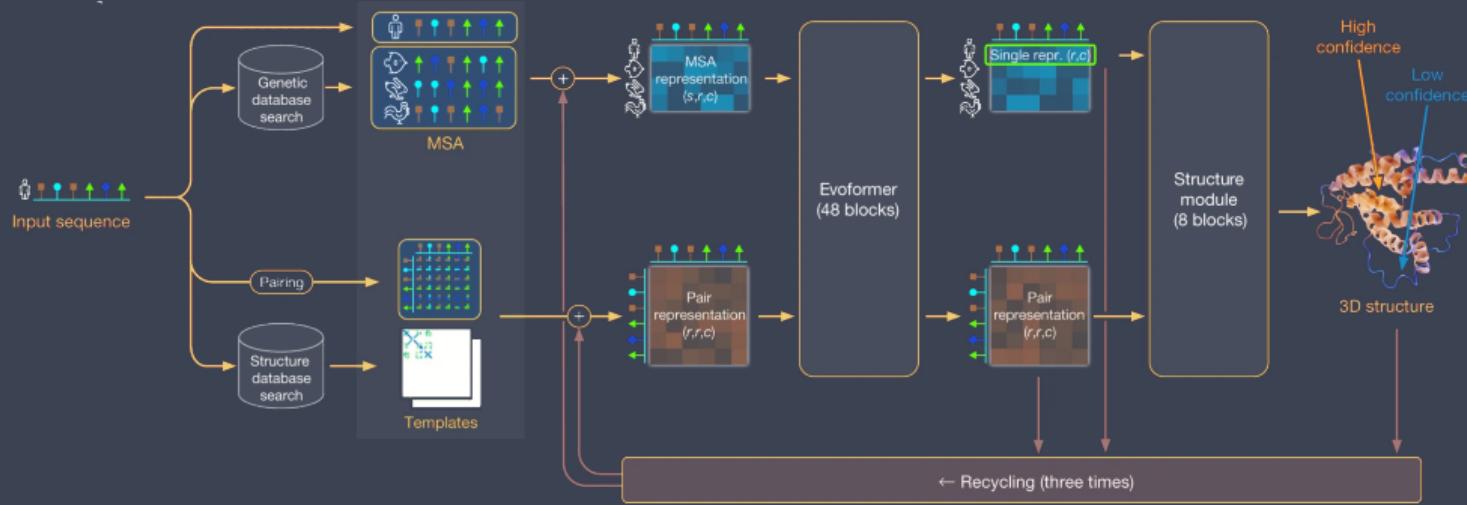


## » AlphaFold 2, метод

- \* Проблема: Архитектуры глубокого обучения излишне отдают предпочтение локальным взаимодействиям
- \* Решение: Разработана новая архитектура глубокого обучения, основанная на внимании, для достижения самосогласованной структуры.
- \* Маленькие MSA
- \* Алгоритм глубокого обучения для полного присутствия на протяжении всего MSA, вместо использования функций парной коэволюции.



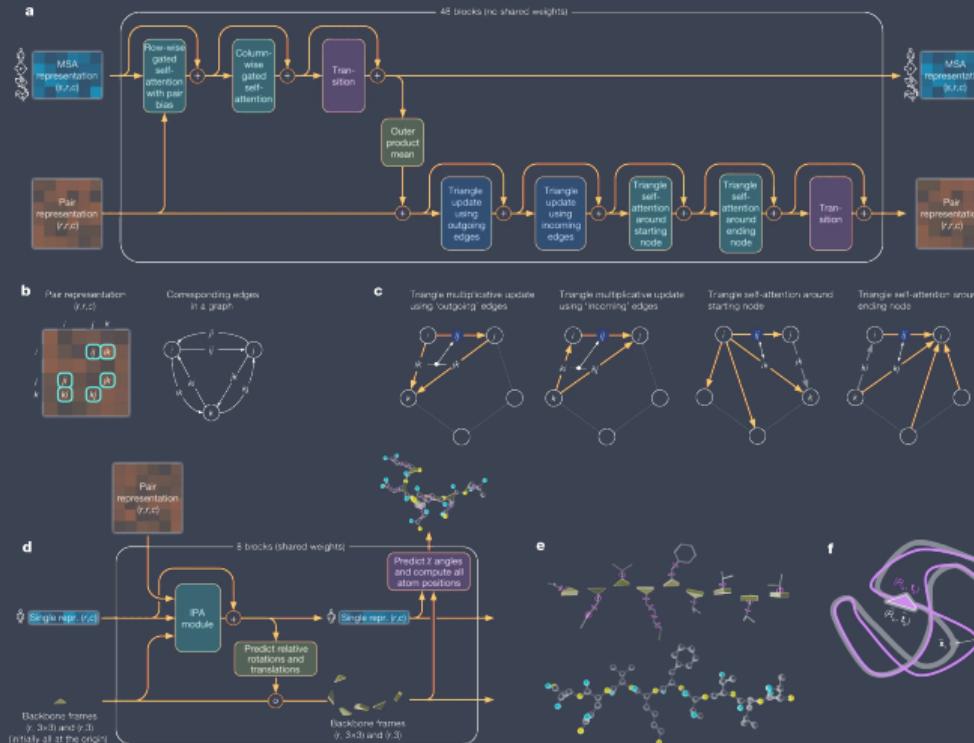
# » AlphaFold 2, метод



10.1038/s41586-021-03819-2



# » AlphaFold 2, метод



10.1038/s41586-021-03819-2



## » Последовательности и MSA

- \* Evoformer: создание массива  $N_{seq} \times N_{res}$ , который представляет собой обработанный MSA
- \* Evoformer: Массив  $N_{res} \times N_{res}$ , представляющий пары остатков (структура).
- \* Ключевыми нововведениями в блоке Evoformer являются новые механизмы обмена информацией внутри MSA и парных представлений, напрямую связывает пространственные и эволюционные связи.



## » Структурный модуль

- \* Явная трехмерная структура в виде вращения и трансляции каждого остатка белка (глобальные каркасы твердого тела).
- \* Ключевая инновация : нарушение структуры цепочки, чтобы обеспечить одновременное локальное уточнение всех частей структуры
- \* Учет непредставленных атомах боковой цепи, а также термин потери, который существенно увеличивает влияние на правильность ориентации остатков.
- \* Итеративное уточнение с использованием всей сети



## » Производительность AlphaFold2

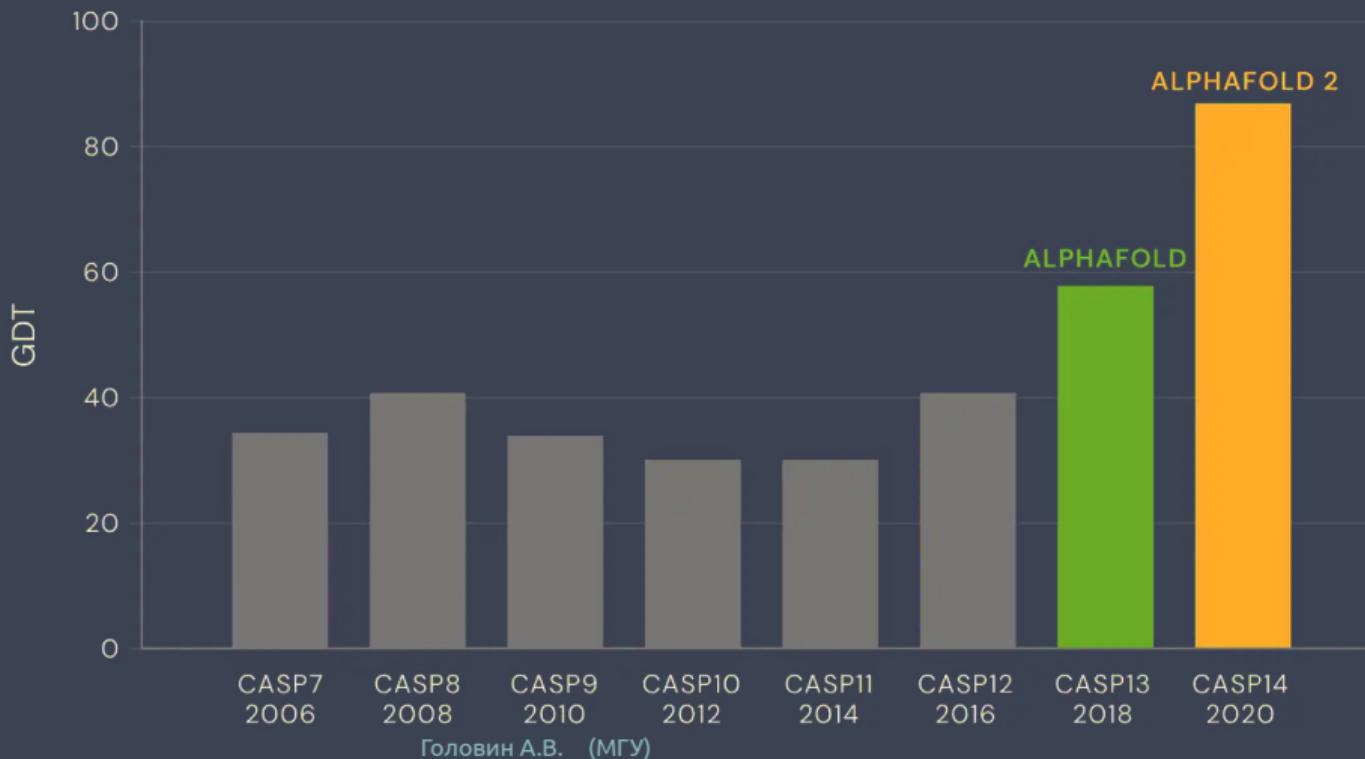
- \* Размер белка ограничены размером RAM GPU
- \* RTX 3090, 400 а.к., около 5 минут

не нужно, 200 млн. моделей: <https://alphafold.ebi.ac.uk/>



# » AlphaFold 2, результат

## Median Free-Modelling Accuracy

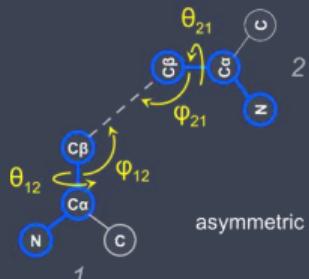
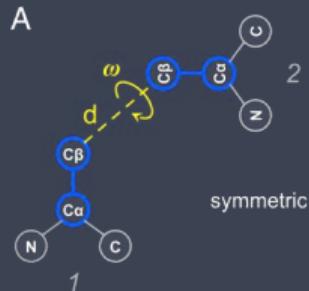


## » Оценка сообщества

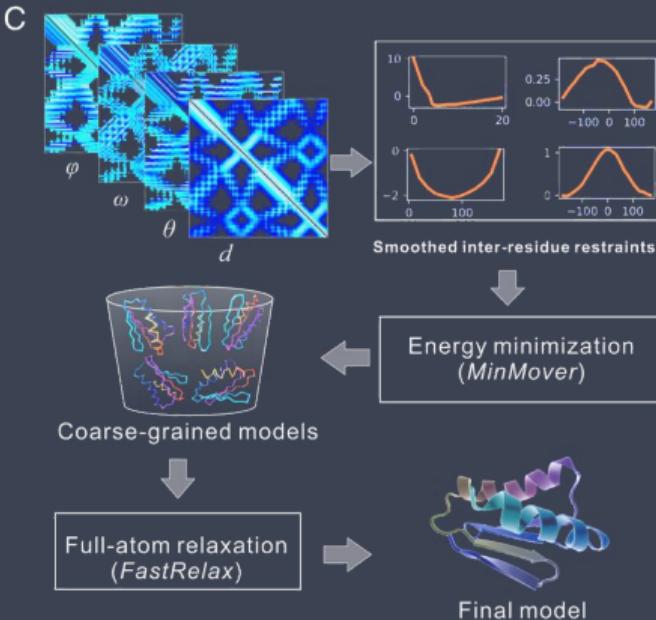
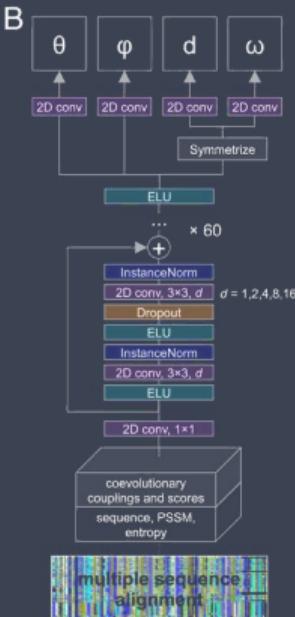
- \* прогнозирование вариантов, обнаружение карманов и построение моделей на основе экспериментальных данных. Структуры, предсказанные с помощью AF2, в среднем так же хороши, как и те, которые получены на основе экспериментальных структур.
- \* AF2 возвращает полное предсказание по белку, часто содержат сегменты белка, размещение которых является неопределенным. Эта неопределенность может привести к неправильным оценкам или выявлению структурного сходства, карманов, эффектов вариантов или плохого построения модели.
- \* AF2 не очень подходит для прогнозирования структур мутированных белков.
- \* AF2 превосходит стандартные подходы докинге белковых молекул, не требуя даже создания стартовых белковых структур.



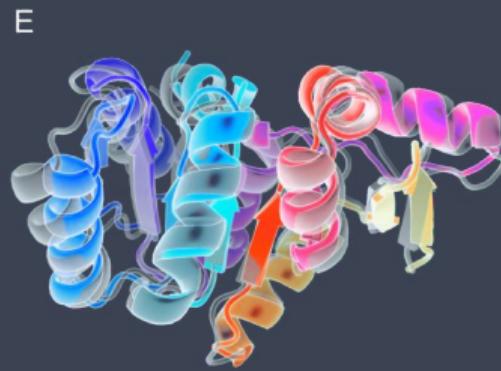
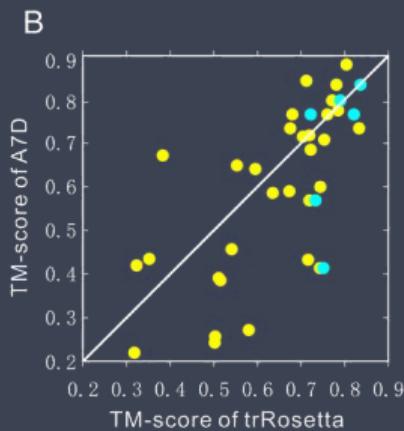
# » trRosetta, метод



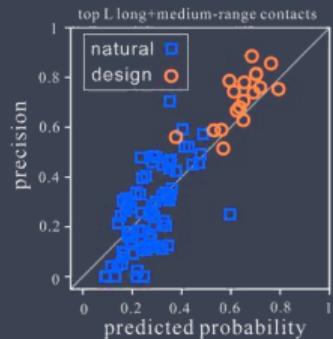
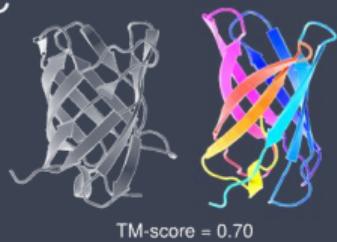
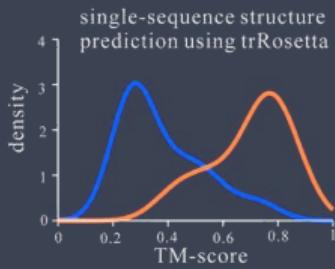
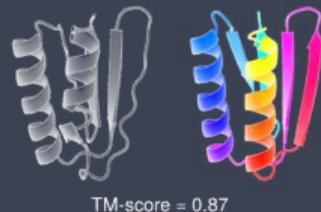
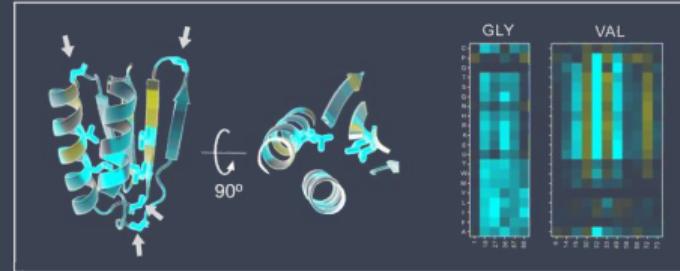
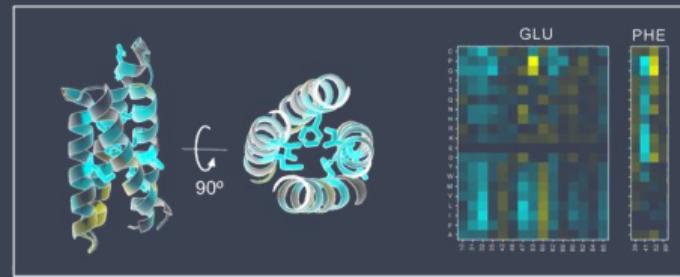
[10.1073/pnas.1914677117](https://doi.org/10.1073/pnas.1914677117)



# » trRosetta, результат



# » trRosetta, дизайн

**A****C****B****E****D**

## » Заключение

- \* Суть современного моделирования белков - эмпирическая
- \* Чем больше известной информации используется при моделировании тем точнее модель.
- \* Каждый метод имеет недостатки.
- \* Критический анализ модели позволяет выявить ошибки и улучшить модель.

