Дизайн хода остова белков

Дизайн белков, НТУ Сириус

Головин А.В. [1] [2] Дианкин И.Д. [1] [2]

[1]НТУ Сириус, ЦИИиИТ

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

Сириус, 2022

Содержание

Дизайн остова или формы белка

- **Локальный дизайн:** Вставки и делеции для достижения необходимой формы локального окружения
- **Глобальный дизайн формы:** Подгон последовательности под фолд и предсказание структуры из последовательности,

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

10.1101/2021.11.20.469408

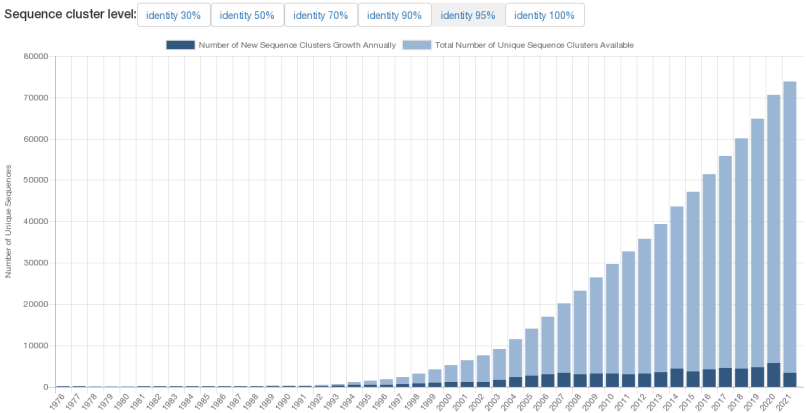
Сравнительное моделирование

- 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.

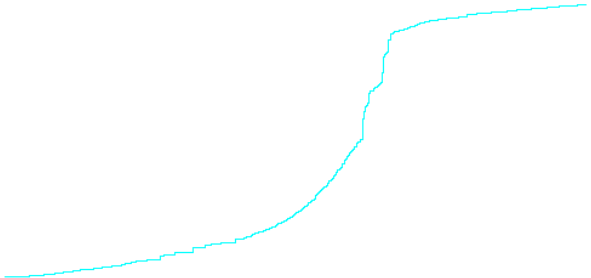
Известные структуры и последовательности

- Сейчас известно порядка 10^4 структур уникальных белков.
- UniProt это 565,928 белков. TrEMBL 225,013,025.
- Для 50% последовательностей можно предсказать способ укладки.

PDB

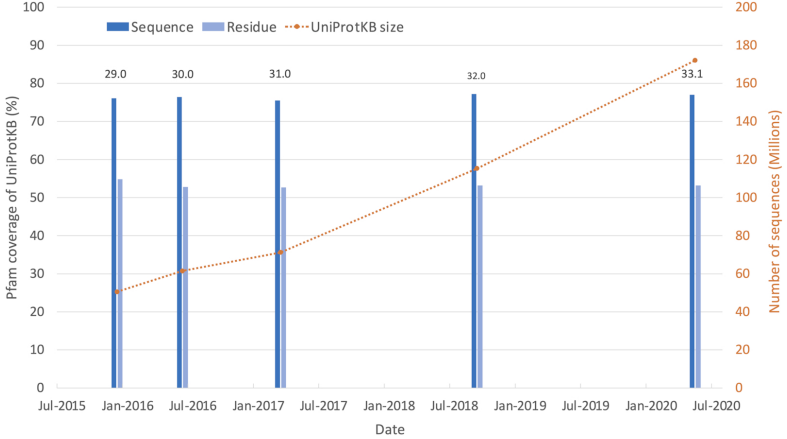

Sequence cluster level:
identity 30%
identity 50%
identity 70%
identity 90%
identity 95%
identity 100%
Number of New Sequence Clusters Growth Annually
Total Number of Unique Sequence Clusters Available
Number of Unique Sequences
80000
70000
60000
50000
40000
30000
20000
10000
0
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

UNIPROT



PFAM

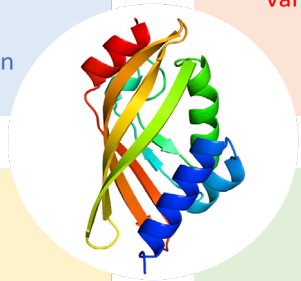
Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models



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
- Неперекрывющиеся 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

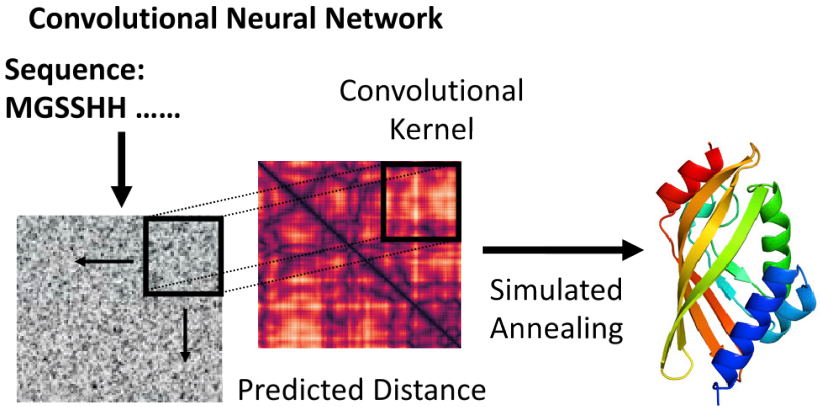
Представление структуры

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

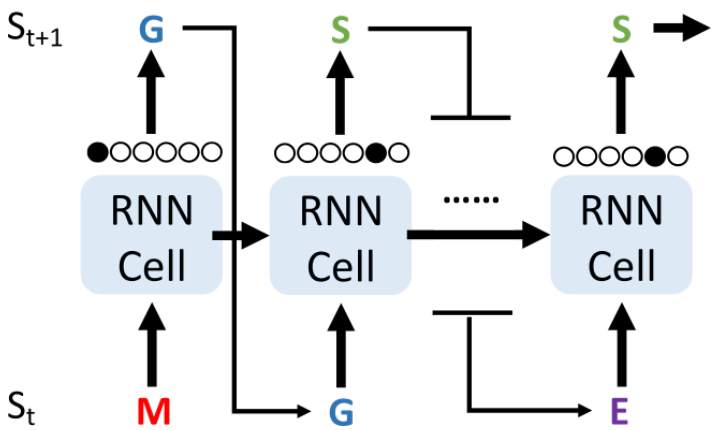
Оценочные функции и силовые поля

- 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

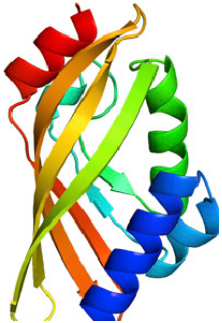


Recurrent NN

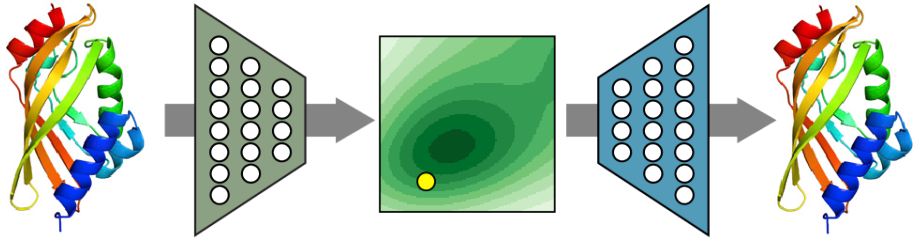


Sequence:

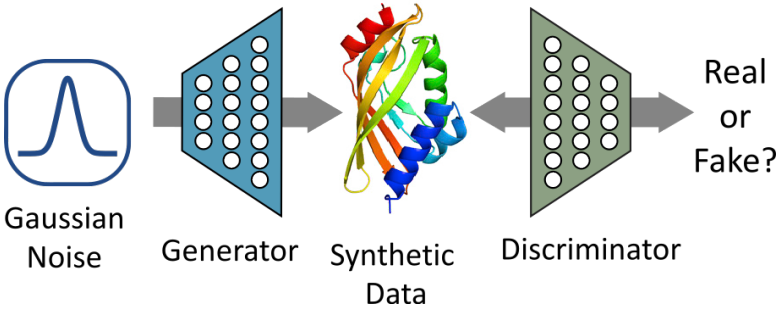
MGSSHH .....



Variational Auto Encoder



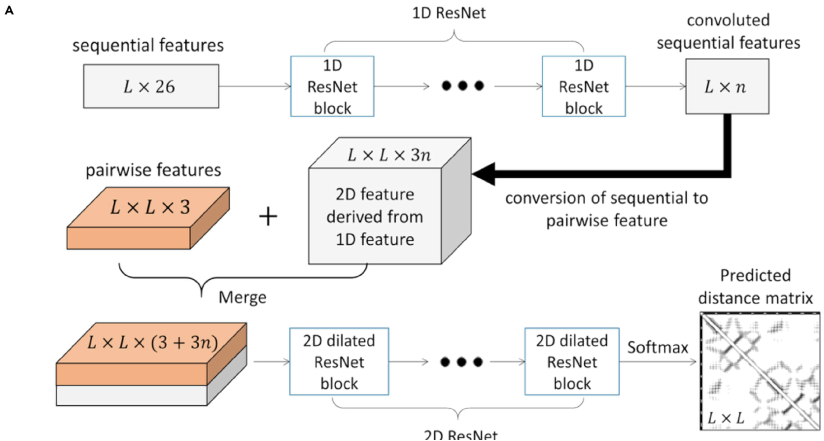
Generative Adversarial Network



Основная идея

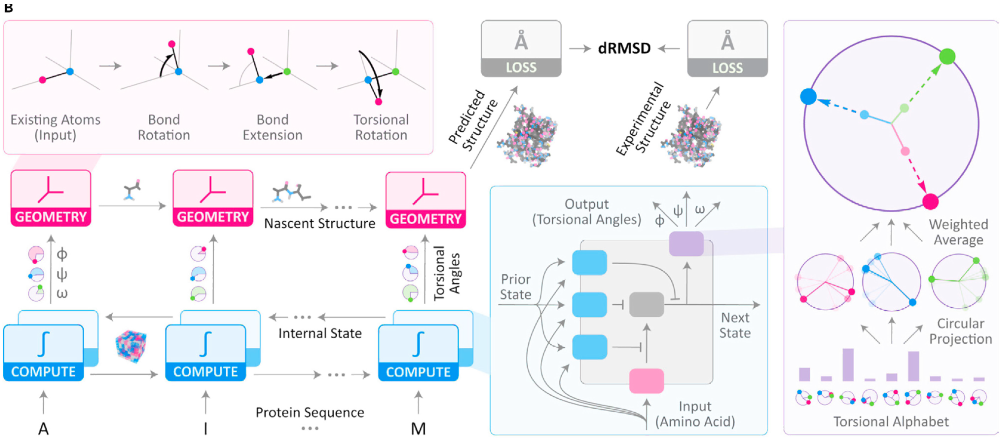
- Классические методы опираются на силовые поля и сложные протоколы
- Новая идея: контактирующие остатки эволюционируют вместе
- Нужна информация об гомологах, большие MSA
- RaptorX и AlphaFold

Архитектуры, растояния



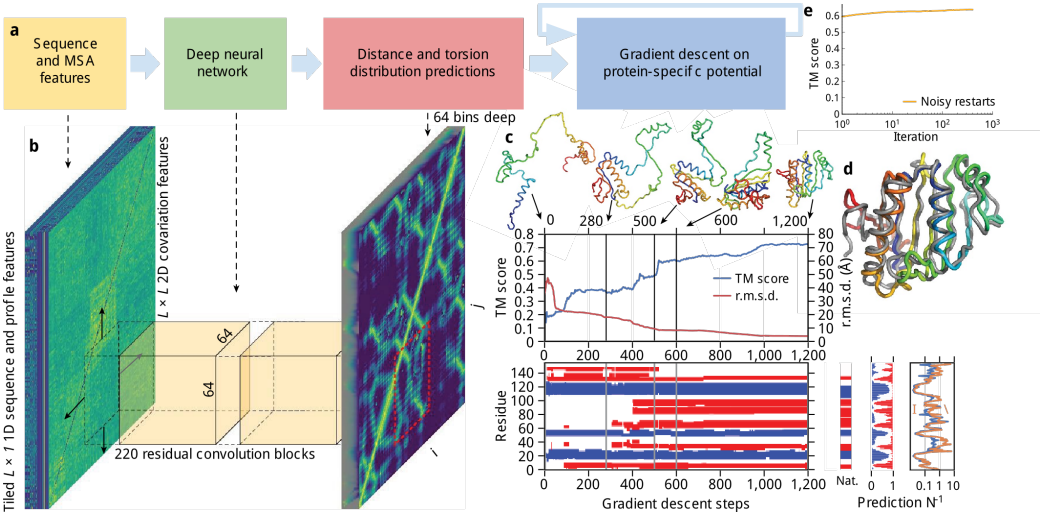
10.1002/prot.25810

Архитектуры, end2end



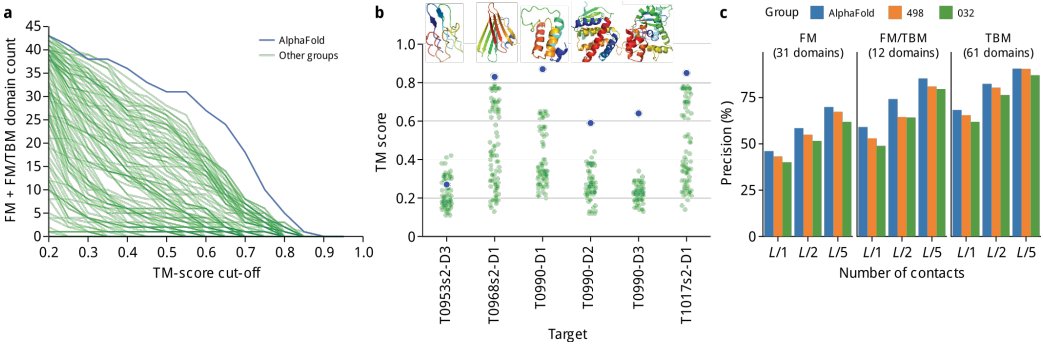
10.1016/j.cels.2019.03.006

Alphafold 1, идея



10.1038/s41586-019-1923-7

Alphafold 1, CASP

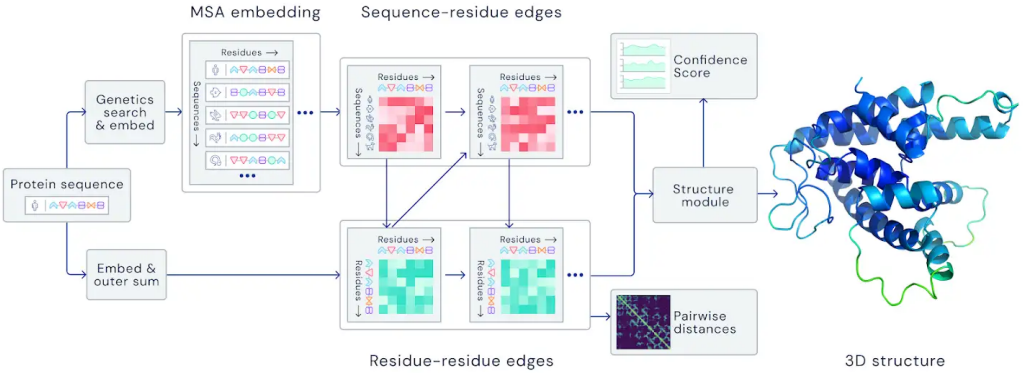


10.1038/s41586-019-1923-7

Alphafold 2, метод

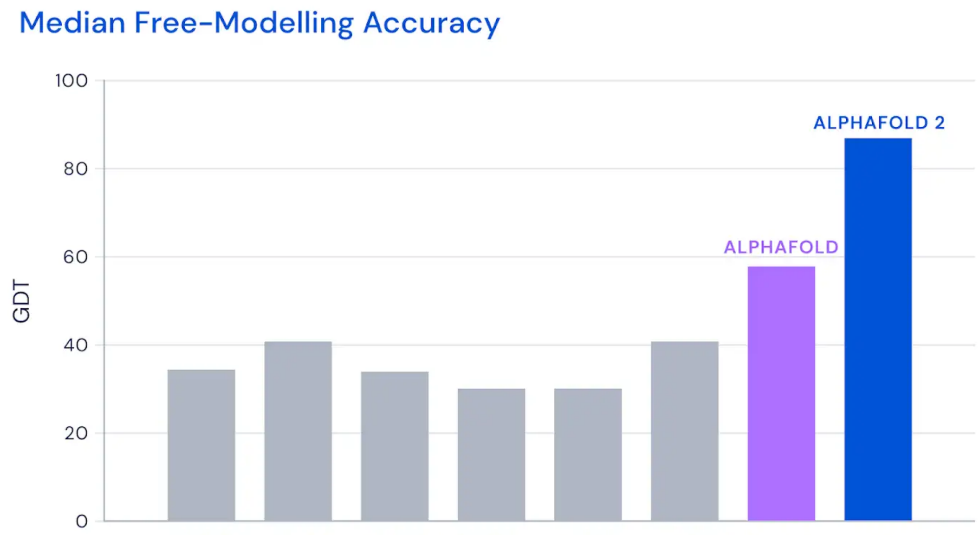
- Deep-learning architectures overly favor sequence-local interactions
- Solution: Developed a novel, attention-based deep learning architecture to achieve self-consistent structure prediction
- Shallow MSA
- Deep learning algorithm to attend arbitrarily over the full MSA, instead of using pairwise co-evolution features

Alphafold 2, метод

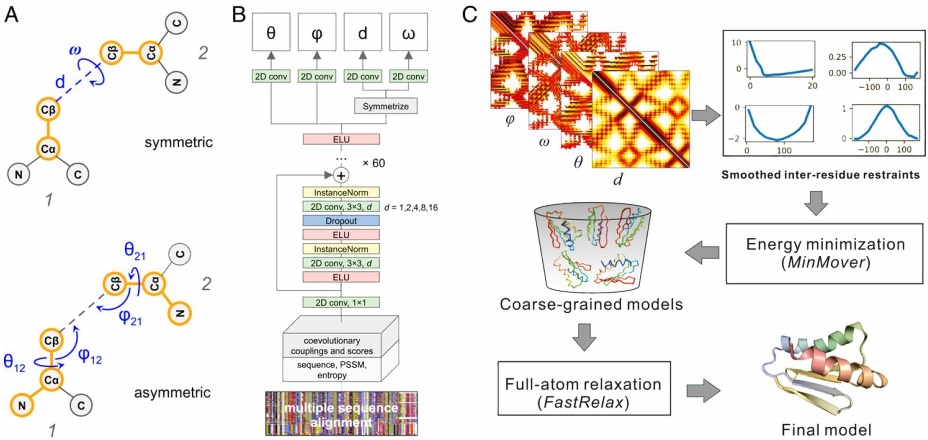


10.1038/s41586-021-03819-2

Alphafold 2, результат

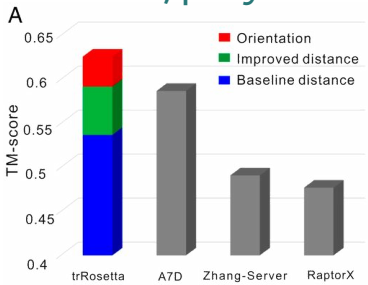


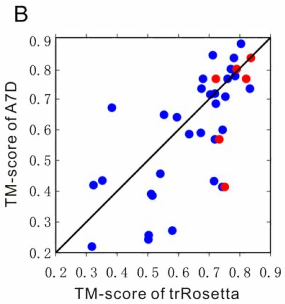
trRosetta, метод

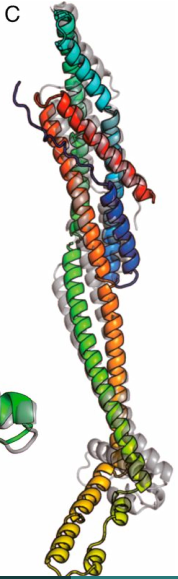


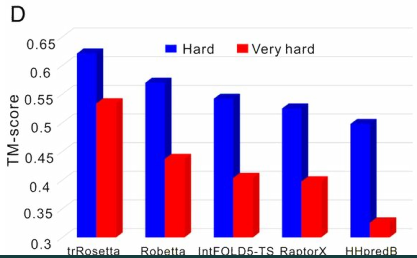
10.1073/pnas.1914677117

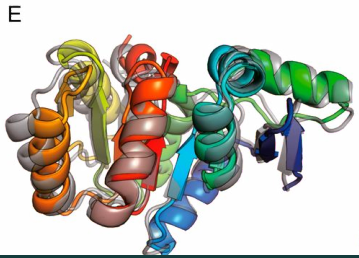
trRosetta, результат



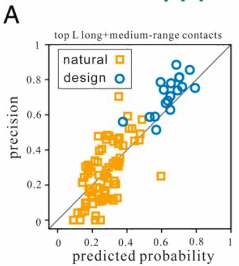


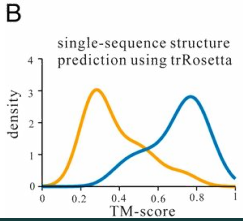


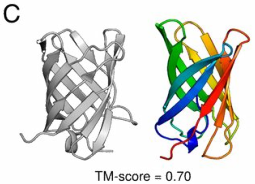


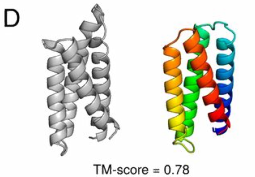


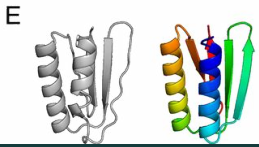
trRosetta, дизайн

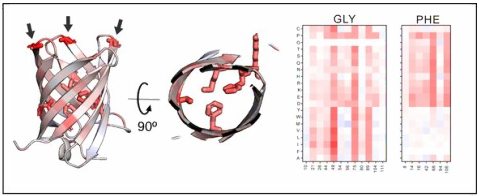


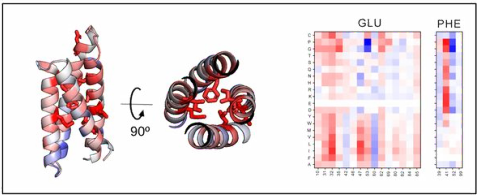


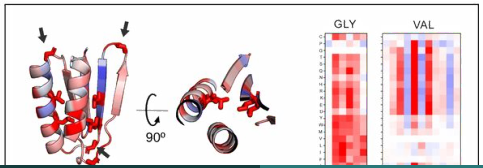












Заключение

- 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.