

Лекция 4. Дизайн стабильности белков

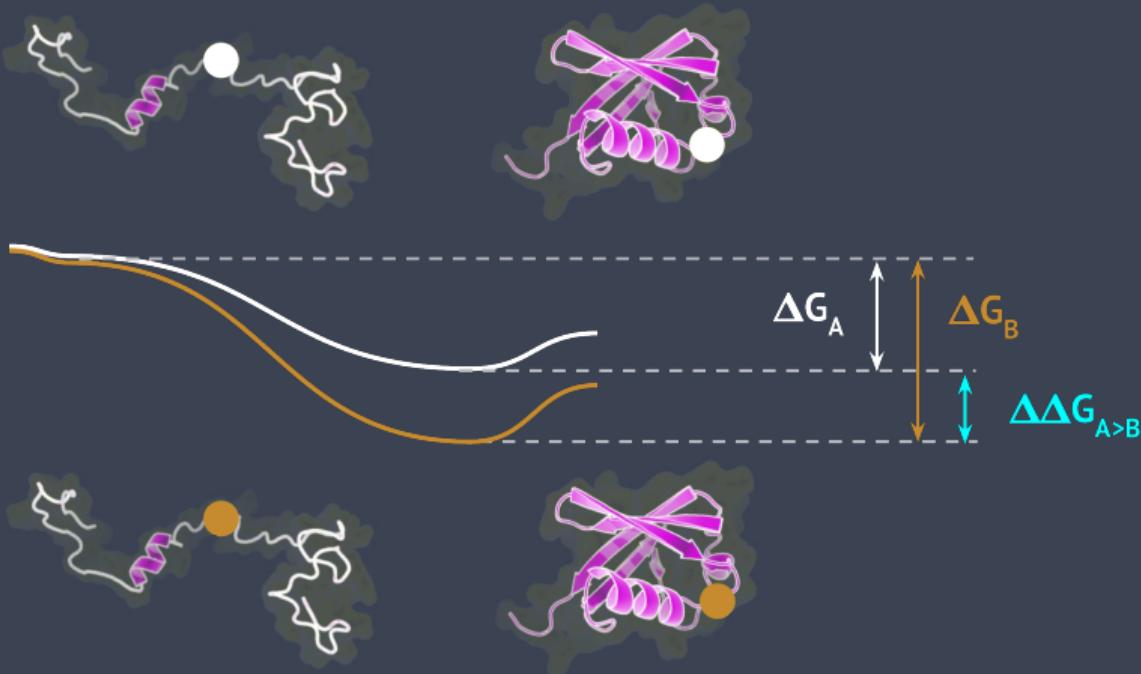
Курс: Методы машинного обучения в дизайне белков

Головин А.В.¹

¹МГУ им М.В. Ломоносова, Факультет Биоинженерии и Биоинформатики

Москва, 2023

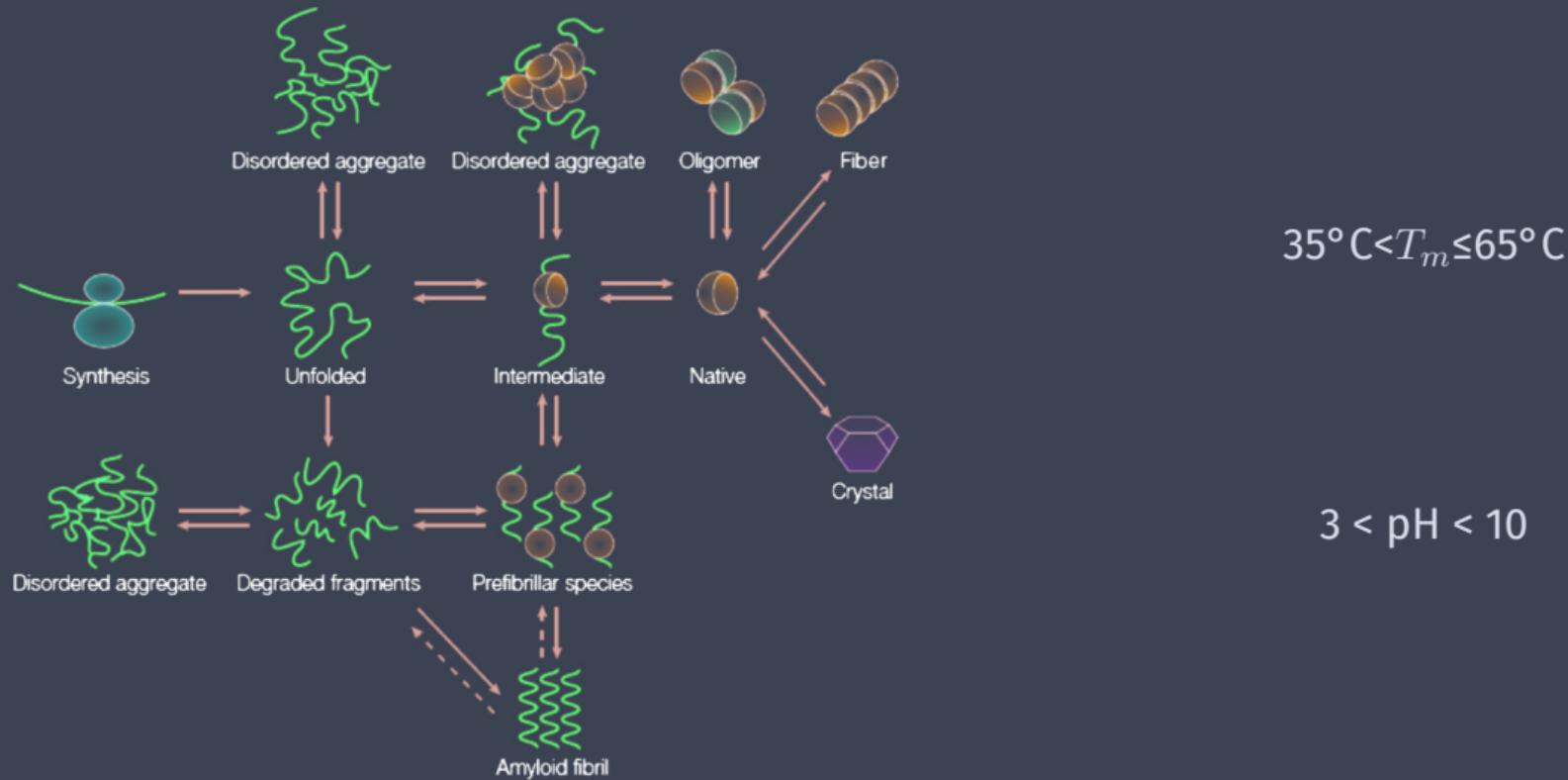
» Фолдинг и стабильность



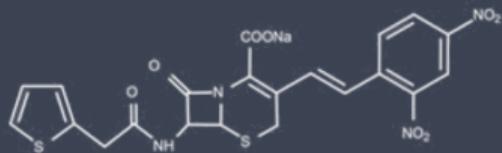
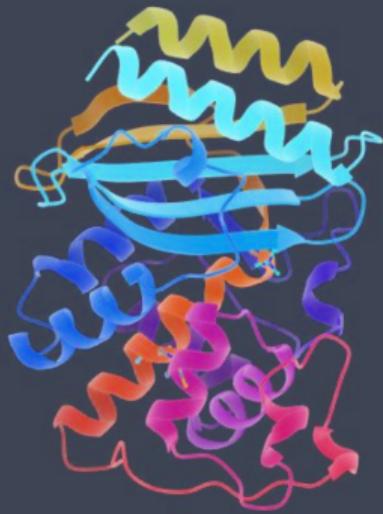
Задача дизайна
стабильности: указать
на замену $A>B$, которая
приведет к
положительному
 $\Delta G_{A>B}$



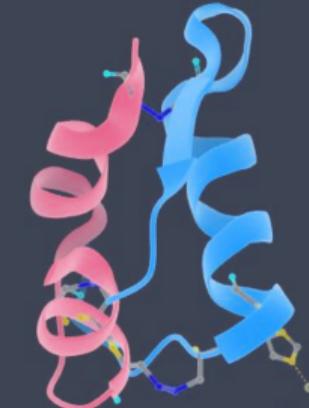
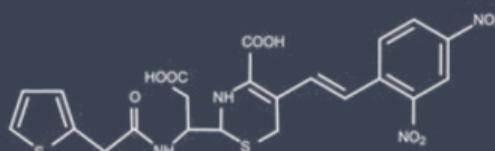
» Денатурация и агрегация



» Области применения



↓
 β -Lactamase

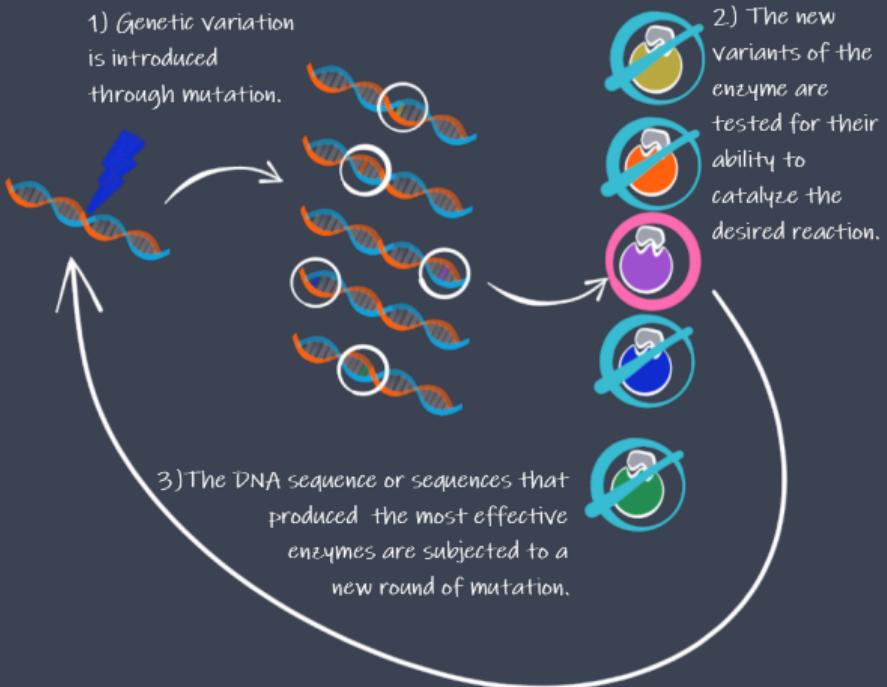


Получение коммерчески-применимых
ферментов

Стабилизация терапевтических
пептидов и белков



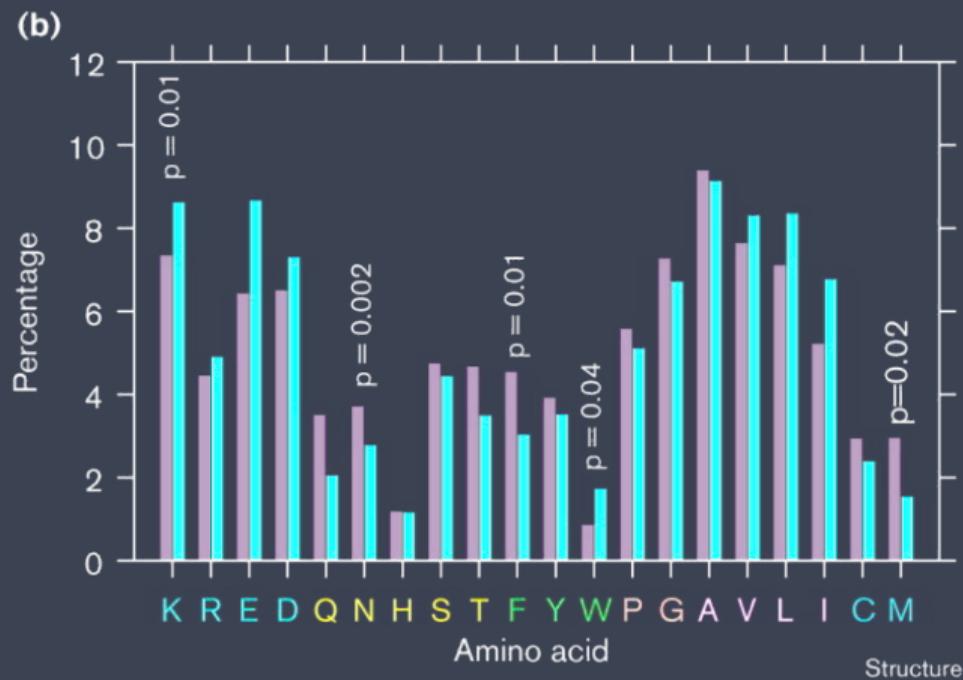
» Направленная эволюция



» Особенности термостабильных белков экстремофилов

В последовательностях:

- * Ile, Val, Leu, Trp ↑
- * Gly ↓, Pro ↑
- * Ser, Gln, Cys, Met ↓
- * Больше полярных АК

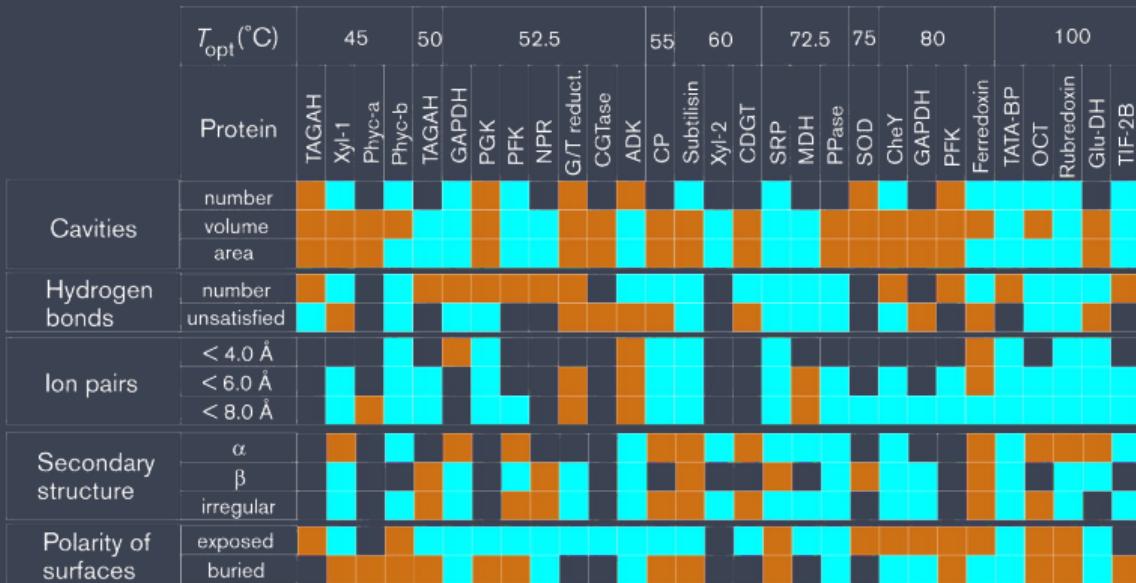


Красные это термофилы

» Особенности термостабильных белков экстремофилов

В структурах:

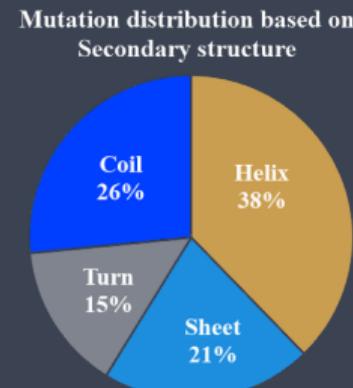
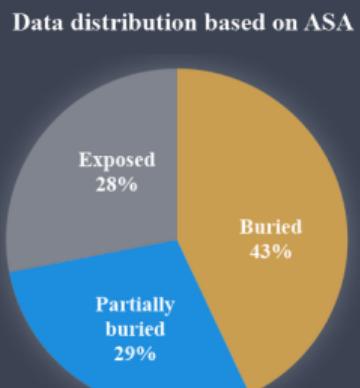
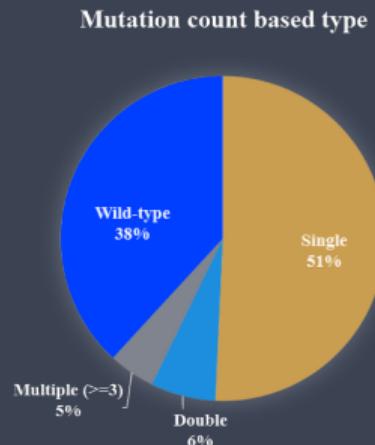
- * Короткие петли
- * Меньше полостей
- * Больше ионных пар на поверхности белка
- * Плотно упакованные концы
- * Более протяженные участки вторичной структуры



» ProTermDB

ProThermDB

HOME BROWSE STATISTICS TUTORIAL UPLOAD RELATED RESOURCES DOWNLOADS CITE US CONTACT US



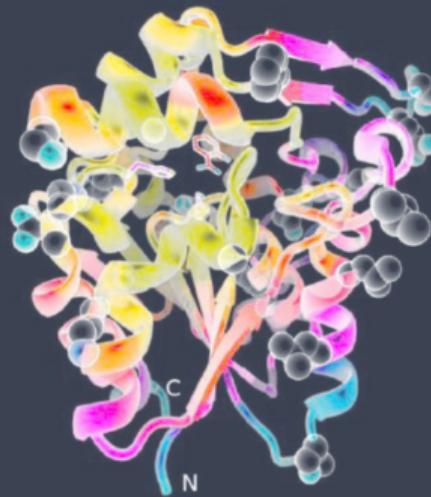
■ Single ■ Double ■ Multiple (≥ 3) ■ Wild-type ■ Buried ■ Partially buried ■ Exposed ■ Helix ■ Sheet ■ Turn ■ Coil



» Вычислительный дизайн стабильности

This figure displays a multiple sequence alignment of 20 different DNA or protein sequences. The sequences are arranged vertically on the left, and the alignment spans 100 positions across the top. Each position is represented by a colored box, where the color indicates the nucleotide or amino acid at that specific site. The color key is as follows: black (conservation), red (high conservation), green (moderate conservation), blue (low conservation), and yellow (very low conservation). The alignment shows significant conservation at certain positions, particularly around positions 50-60 and 95-100.

последовательность

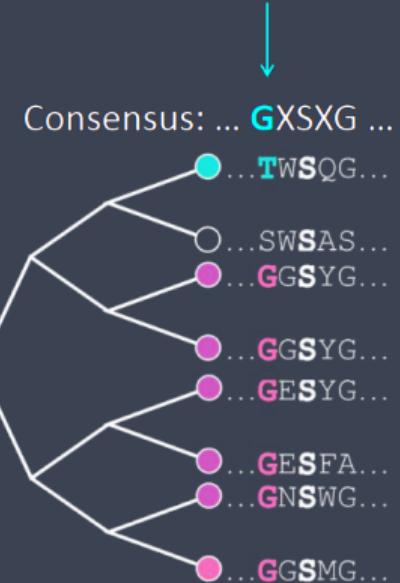


Структура



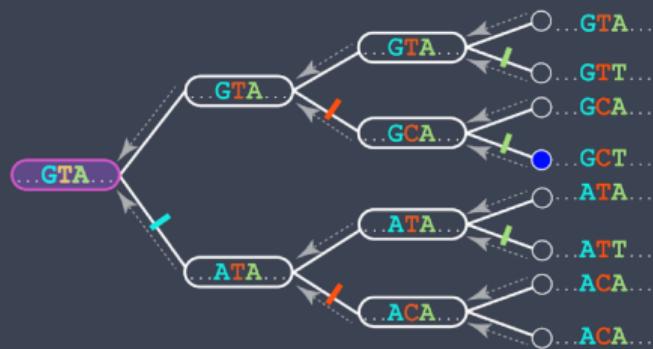
» Back-to-consensus

- * Предполагается, что консенсусные остатки в функционально разнообразных гомологах отвечают за стабильность, а отличия отражают случайные дестабилизирующие мутации, которые оказались нейтральными при отборе
- * Замена в позиции, отличающейся от гомологов, на консенсусную АК приведет к увеличению стабильности



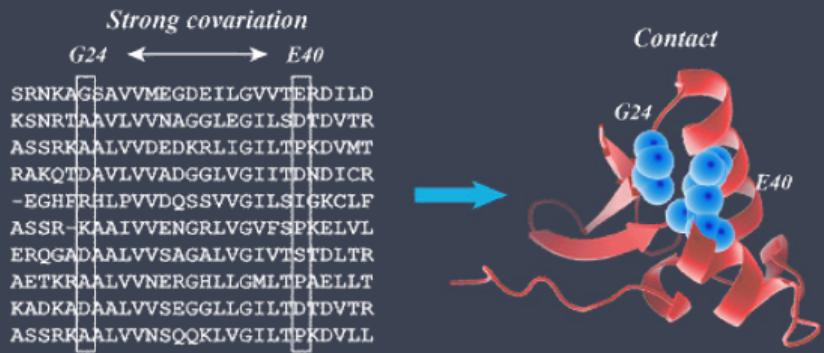
» Востановление предшественника

- * Жизнь произошла от термофильных организмов. Предковые гомологи всех белков - термостабильные.
- * Построение филогенетических деревьев и поиск предковой последовательности приведет к созданию термостабильного белка.

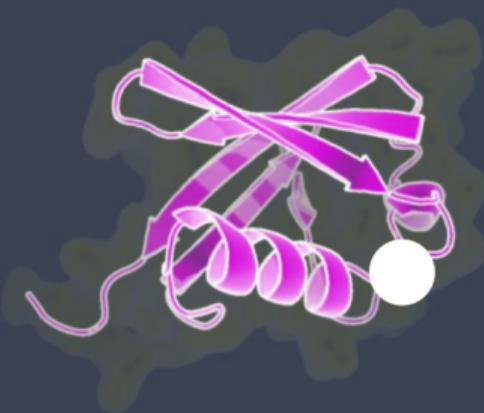


» Сочетание структурных данных и выравниваний

- * Рациональный анализ структур
- * Поиск ковариирующих позиций
- * Сохранение остатков вторичной структуры
- * Сохранение функционально важных позиций



» Оценка ddG единичных замен

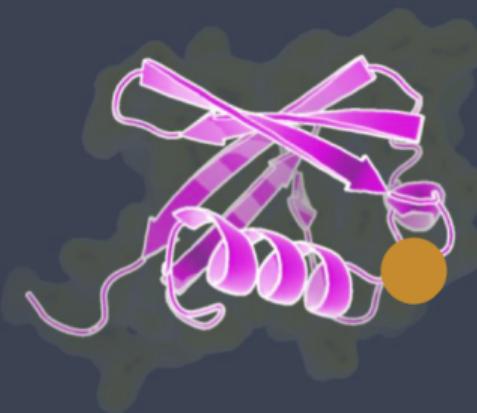


Physics-based

Оценка межатомных
взаимодействий

Descriptors-based

Descriptors-based
Воспроизведение
известных структур с
применением ML



Knowledge-based

Воспроизведение
статистических данных

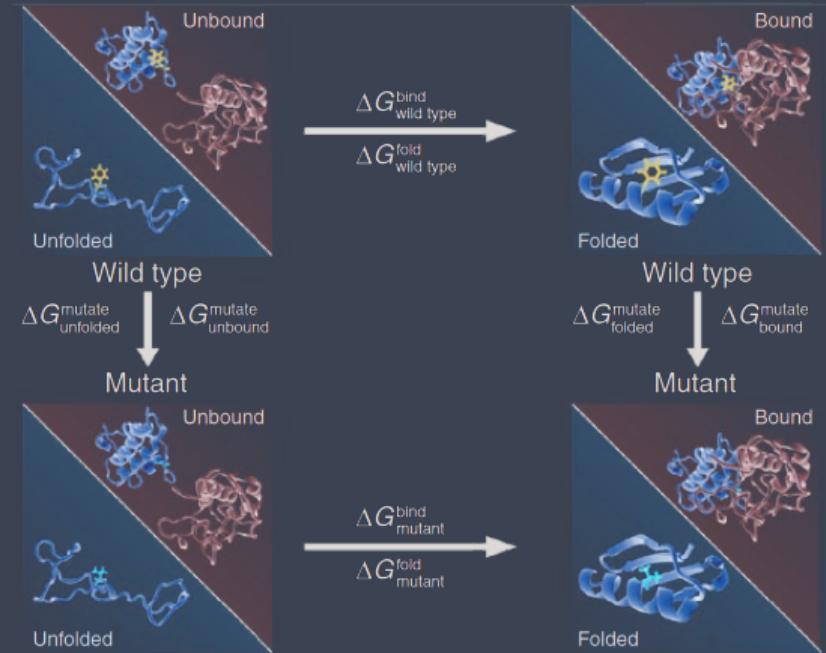


» Оценка ddG единичных замен: CC/PBSA

- * Physics-based

$$\Delta G_{CC/PBSA} = \Delta G_{EL} + \Delta G_{VdW} + \Delta G_S$$

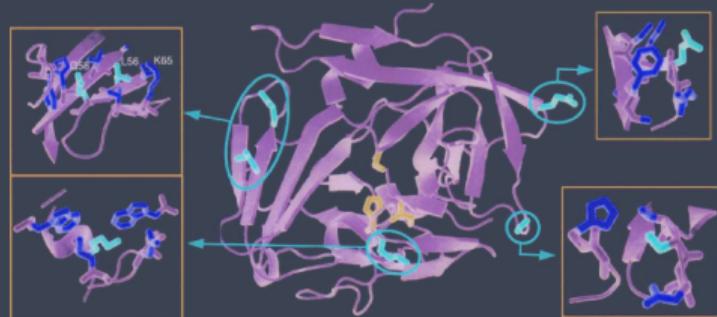
- * Вычислительно затратный



» Оценка ddG единичных замен: PopMusic

- * Physics-based

$$\Delta\Delta G_P = \sum_{i=1}^{13} a_i(A) \Delta\Delta W_i + a_{14}(A) \Delta V_+ + \\ + a_{15}(A) \Delta V_- + a_{16}(A)$$



- * WWW available

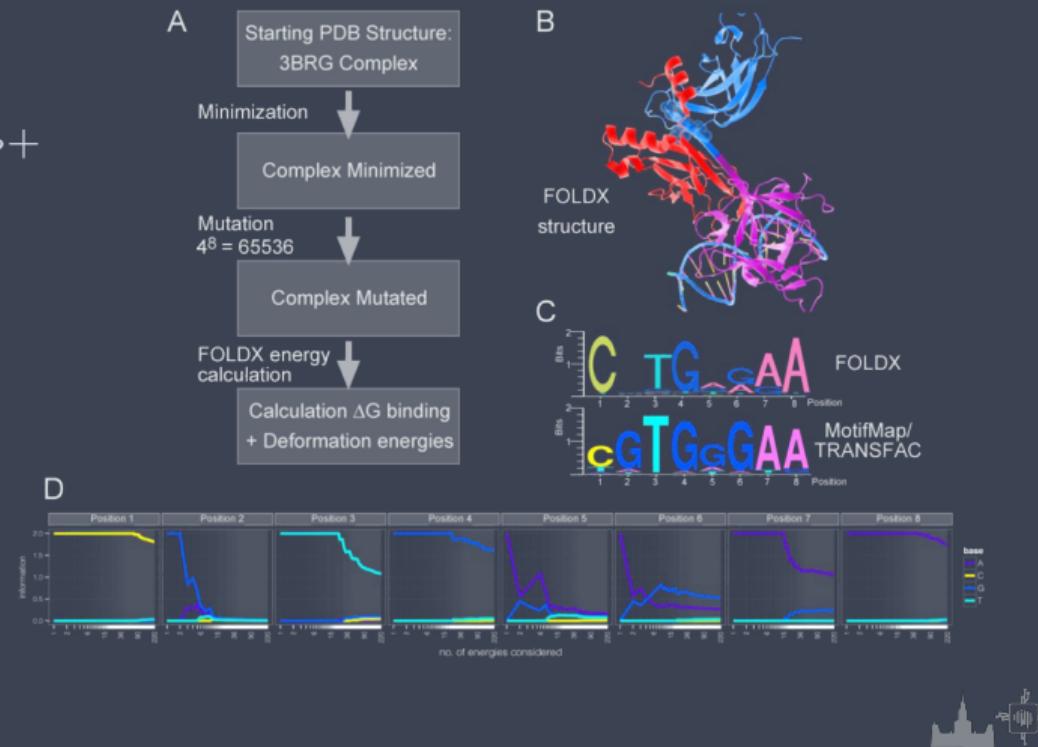


» Оценка ddG единичных замен: FoldX

- * Physics-based + Knowledge-based

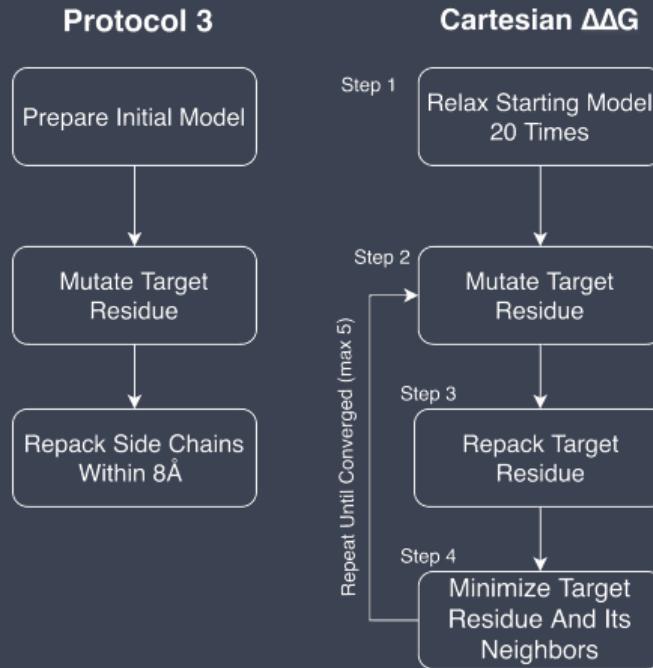
$$\Delta G = \Delta G_{vdw} + \Delta G_{solH} + \Delta G_{solP} + \\ + \Delta G_{hbond} + \Delta G_{wb} + \Delta G_{el} + \\ + \Delta S_{mc} + \Delta S_{sc}$$

- * Есть в виде веб-сервера
- * Умеет работать с ДНК
- * Умеет работать с димерами
- * Умеет учитывать стабильность при смене pH или ионной силы раствора



» Оценка ddG единичных замен: Rosetta

- * Knowledge-based в сочетании с Монте-Карло динамикой для минимизации структур
 - * До 2016 года для подсчета ddG применялись кастомные протоколы
 - * После - cartesian_ddG
 - * Вычислительно затратен



» Оценка ddG единичных замен: I-Mutant

- * Predictor-based
- * Support vector machine regression
- * Есть в виде веб-сервера
- * Умеет работать с SNP
- * Умеет работать на последовательности без структуры

I-Mutant Suite
Predictor of effects of single point protein mutation

Prediction of protein stability changes upon single point mutation from:

Protein Structure (if available)
 Protein Sequence

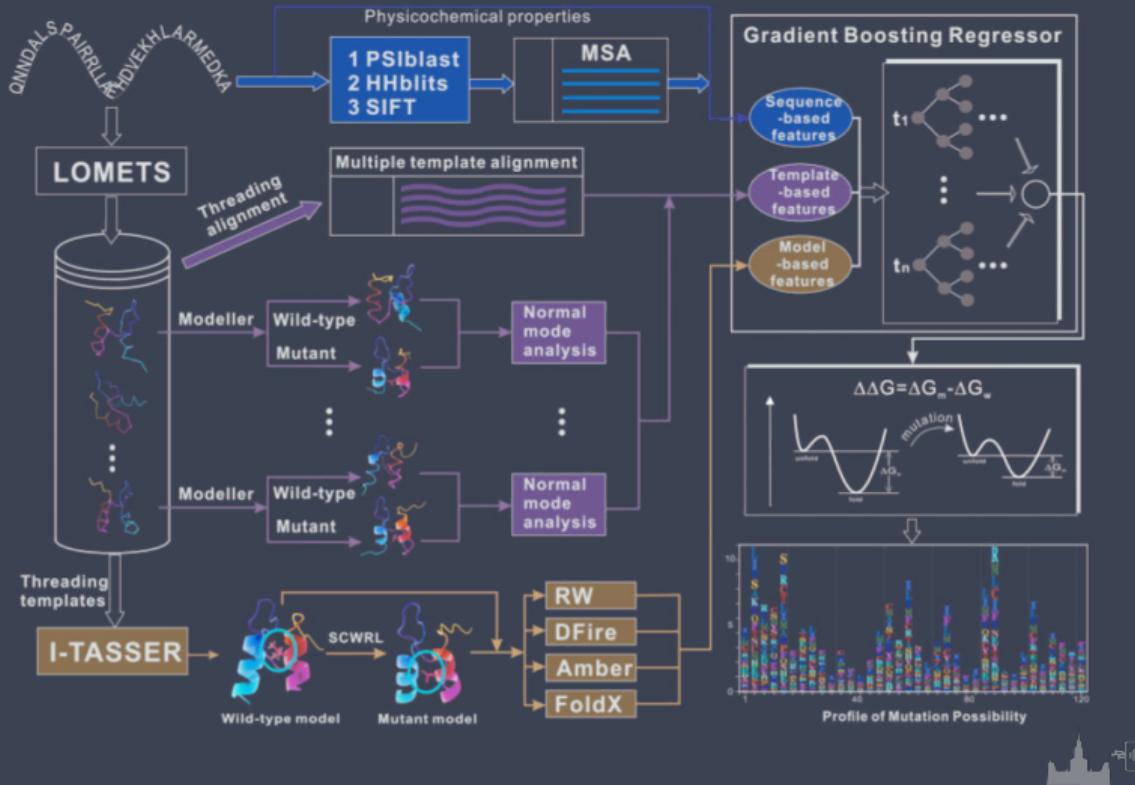
Prediction of Disease associated single point mutation from

Protein Sequence



» Оценка $\Delta\Delta G$ единичных замен: STRUM

- * Predictor-based
- * Gradient boosting regression
- * Был в виде веб-сервера - сейчас сервер временно мёртв



» Оценка ddG единичных замен: Критика

FoldX as Protein Engineering Tool: Better Than Random Based Approaches?

Oliver Buß✉, Jens Rudat✉, Katrin Ochsenreither✉

A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation 

Jianwen Fang✉

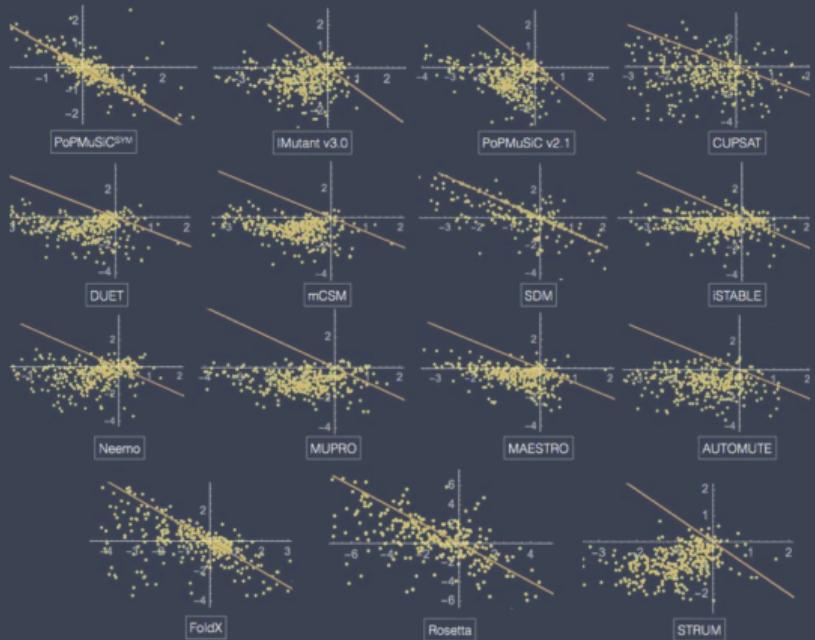
Quantification of biases in predictions of protein stability changes upon mutations

Fabrizio Pucci✉, Katrien V Bernaerts, Jean Marc Kwasigroch, Marianne Rooman



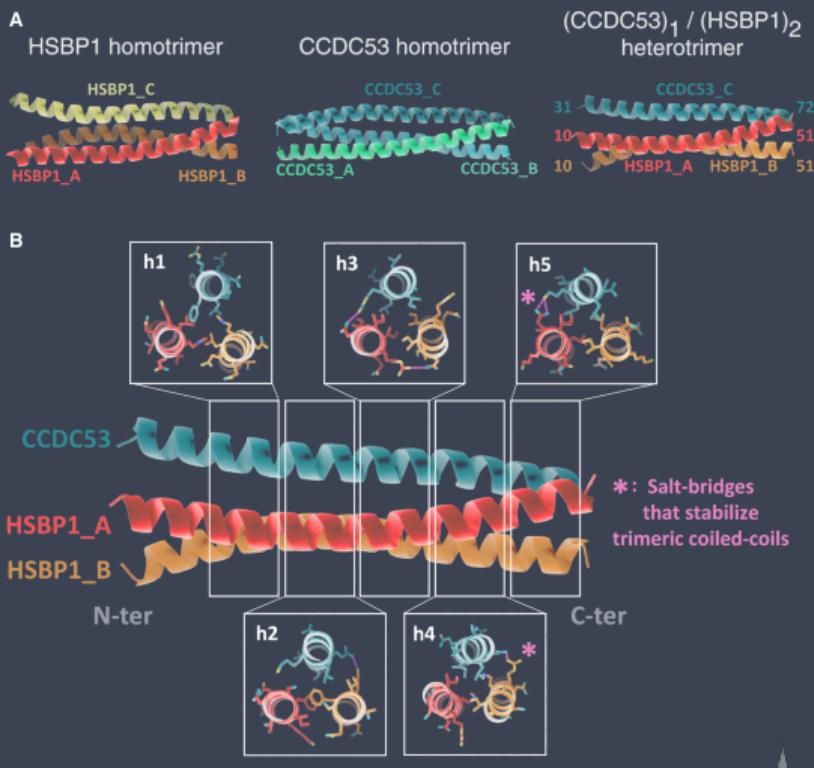
» Оценка ddG единичных замен: Критика

- * Критерий асимметричности ddG
 $\Delta\Delta G_{WT \rightarrow mut} = -\Delta\Delta G_{mut \rightarrow WT}$

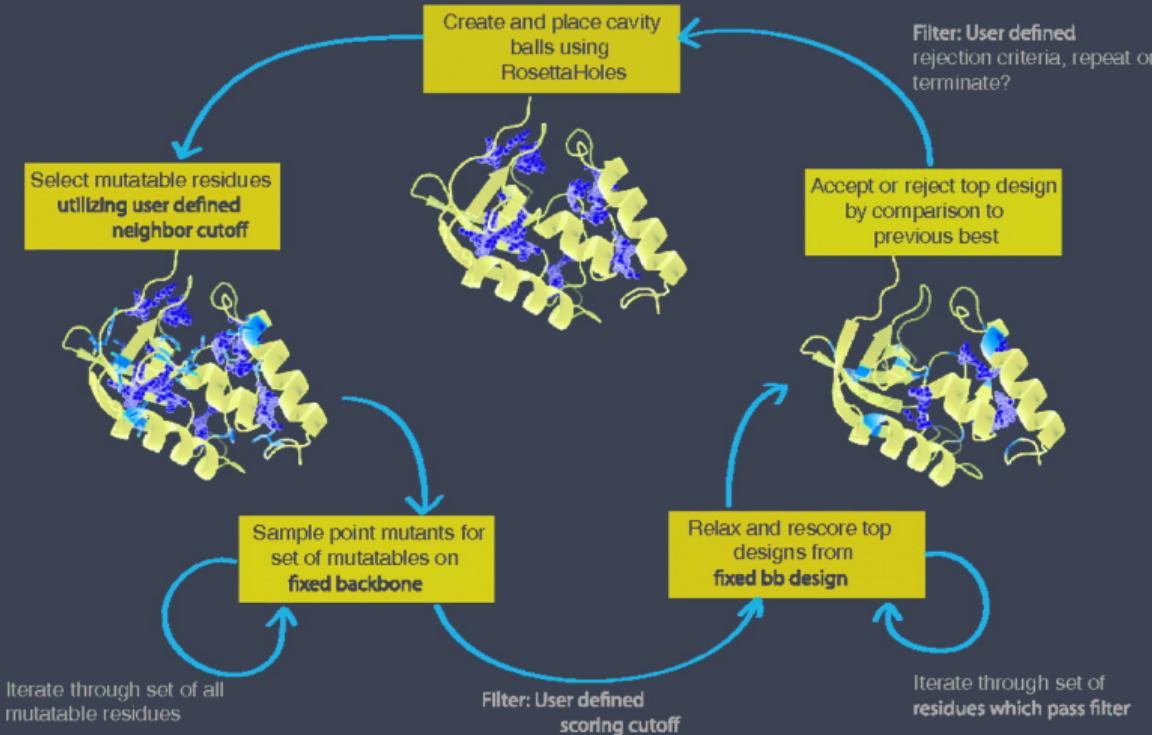


» Дизайн гидрофобных ядер

- * Наибольшие успехи - с помощью алгоритмов Rosetta
 - * Получены и стабилизированы неприродные фолды



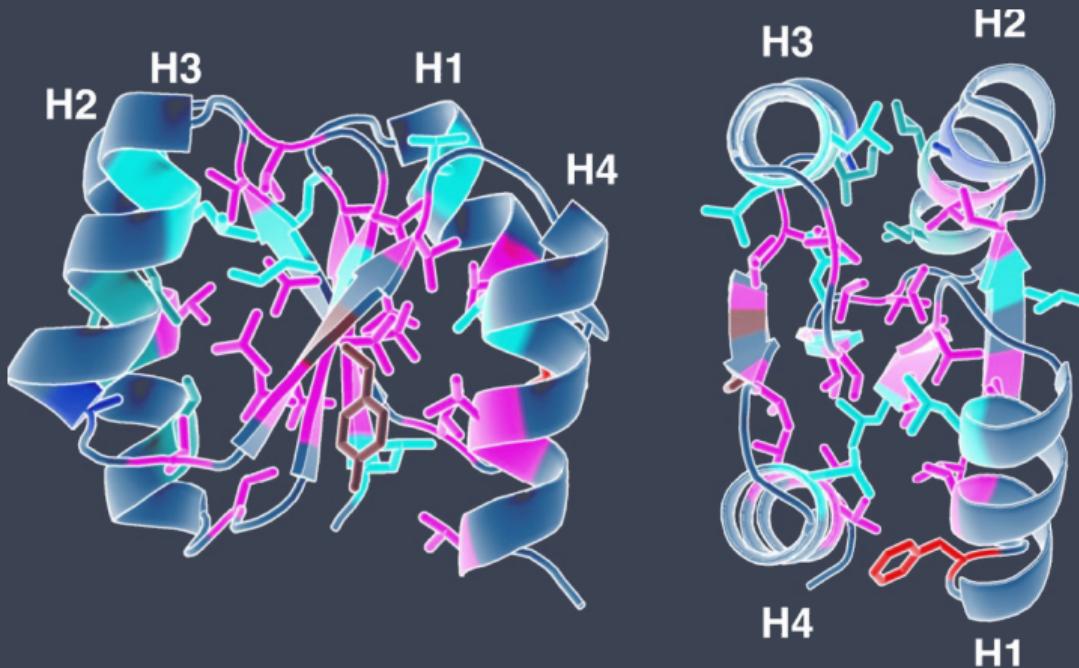
» Rosetta VIP



10.1073/pnas.1115172109



» Недоупакованное гидрофобное ядро

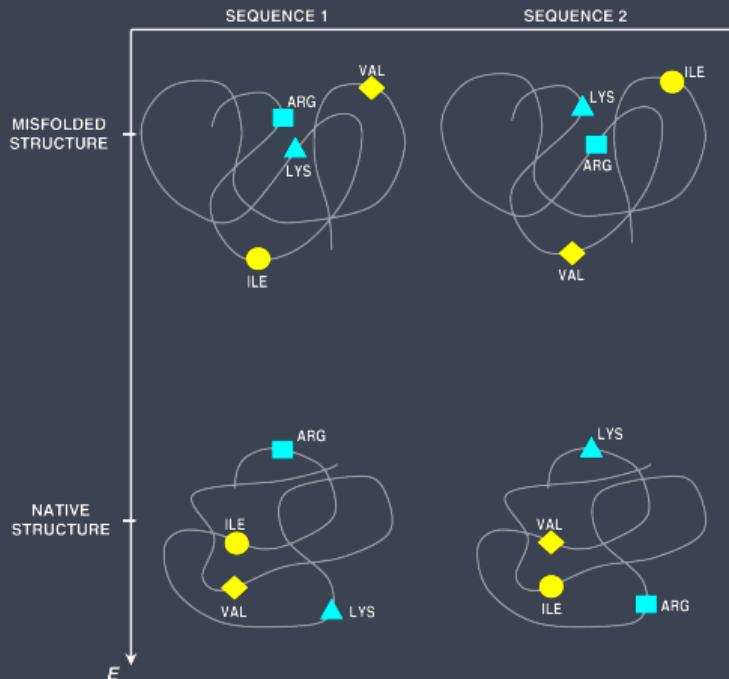


Robust folding of a de novo designed ideal protein even with most of the core mutated to valine. <https://doi.org/10.1073/pnas.2002120117>



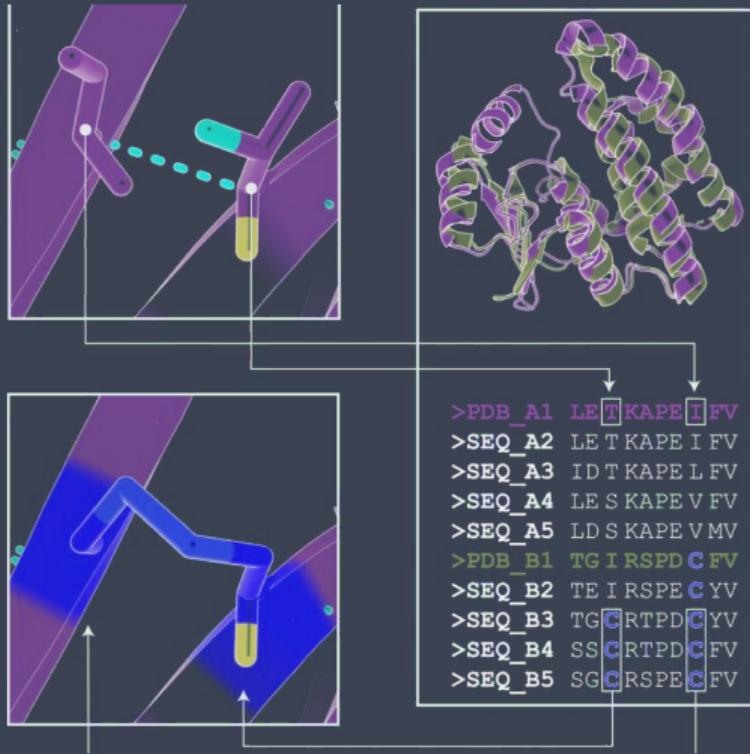
» Внесение заряженных АК

- * Формирование солевых мостиков
- * Негативный дизайн - дестабилизация несвернутых состояний



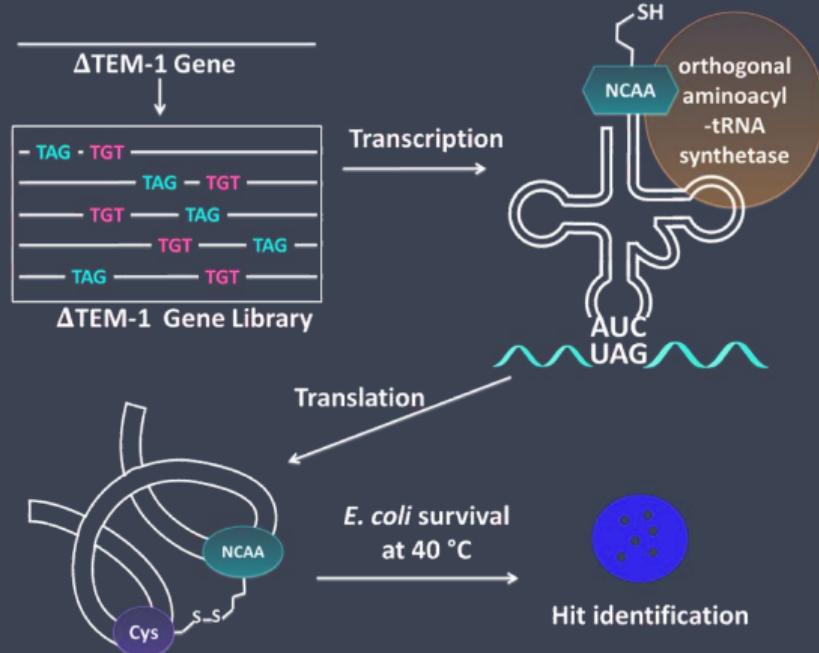
» Дизайн дисульфидных связей

- * MODIP, DbD - поиск на основе статистической скоринг-функции
- * SSbondPre - предиктор на основе нейросети

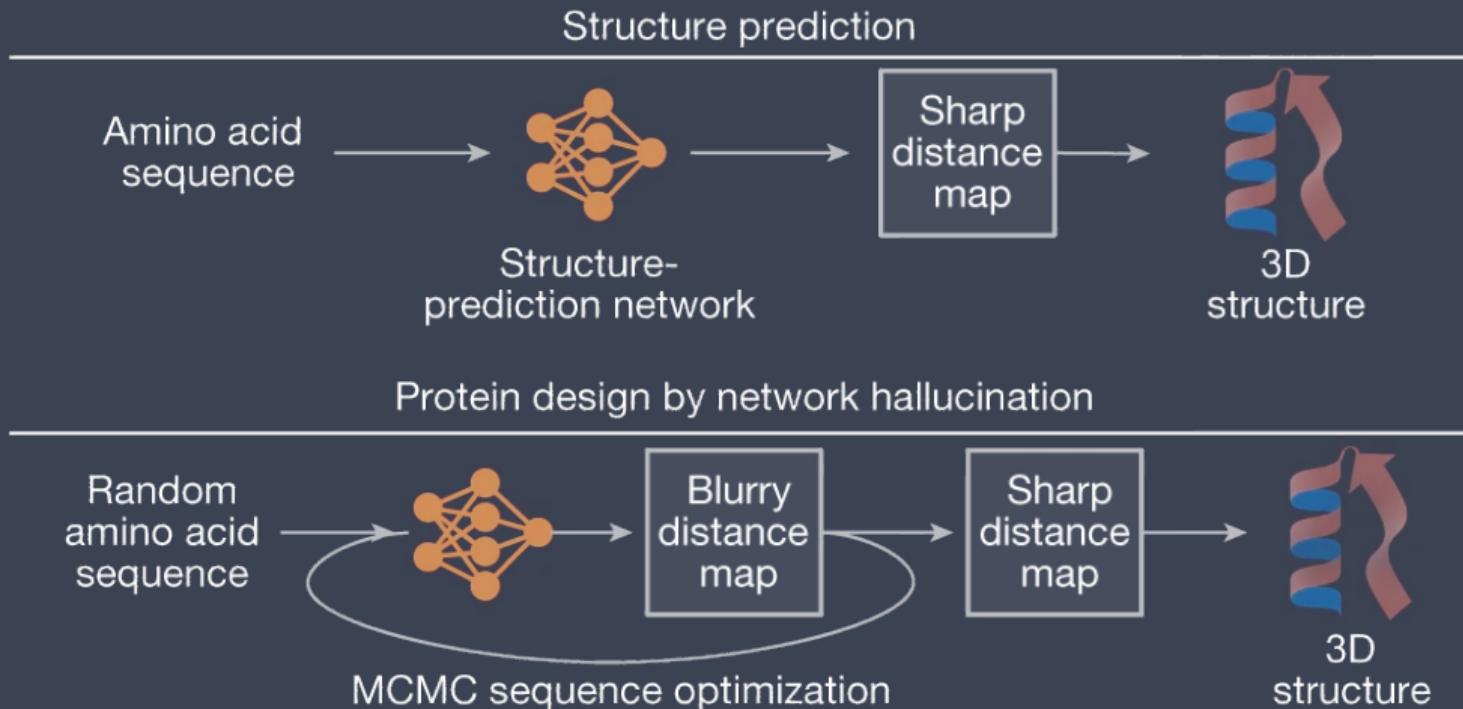


» Дизайн дисульфидных связей с нестандартными АК

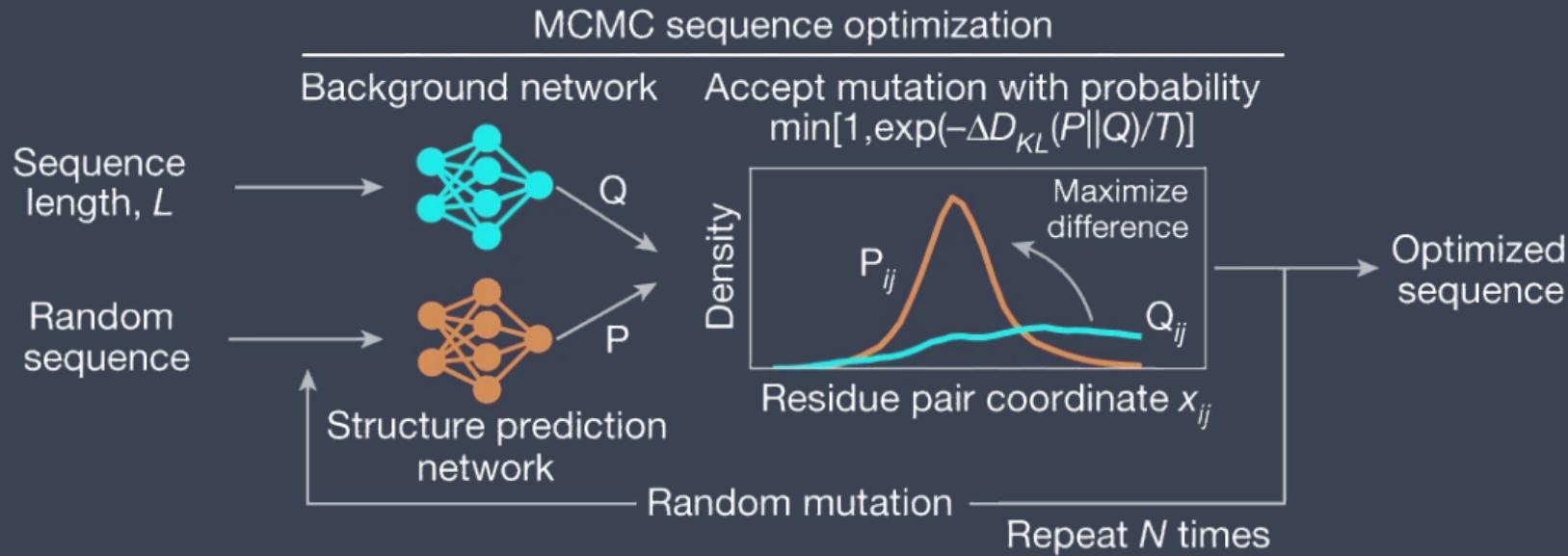
- * Дисульфидные мостики формируют тиол-содержащие АК
- * Используются методы биоинформатики для поиска возможных позиций для мутации на тиол-содержащие АК и методы биоинженерии для вставки таких АК в процессе трансляции



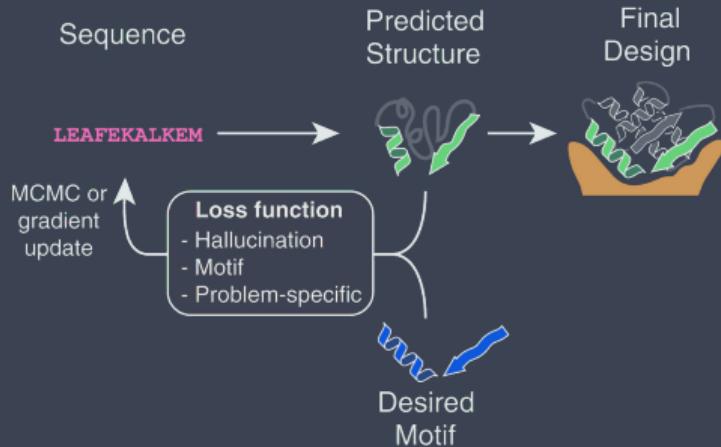
» "network hallucination"



» "network hallucination"



» Ограниченные галлюцинации



- * Сложная функцию потери, которая сочетает в себе часть из галлюцинаций с частью реконструкции мотива.
- * Подход с ограниченными галлюцинациями требует больших вычислительных ресурсов, поскольку для каждого шага градиентного спуска во время оптимизации последовательности требуется прямой и обратный проход через сеть.



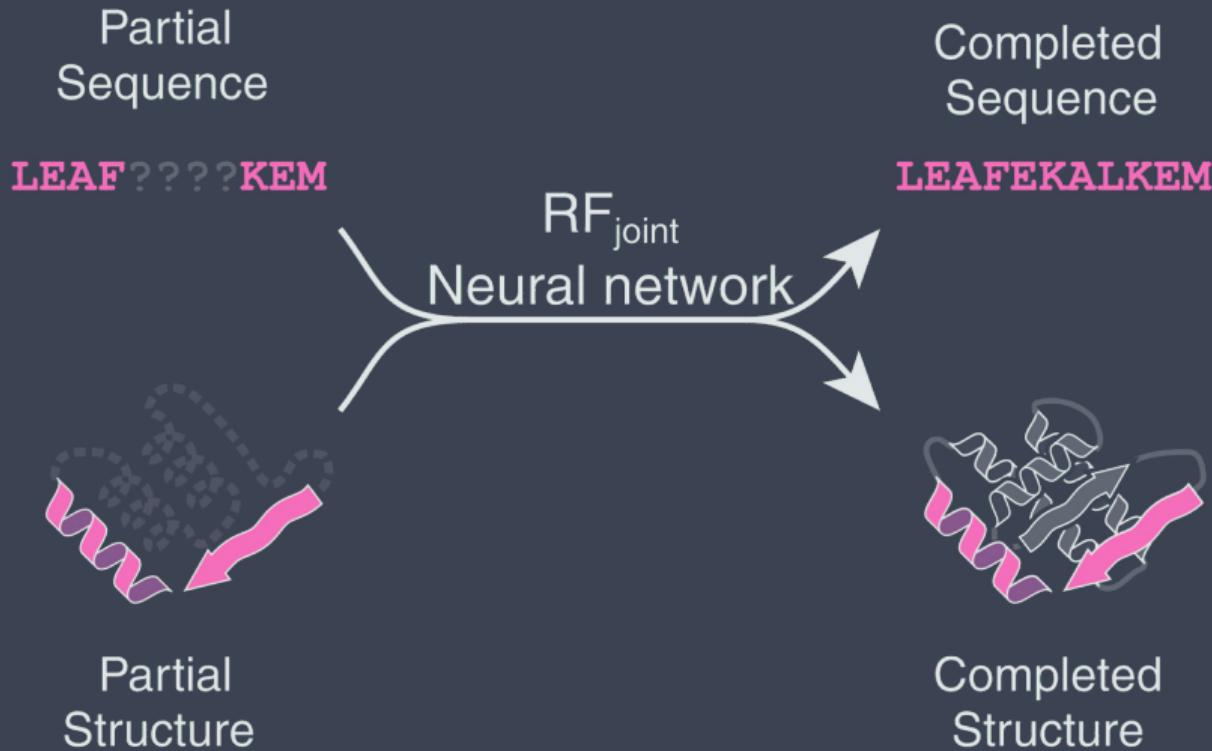
» Inpainting



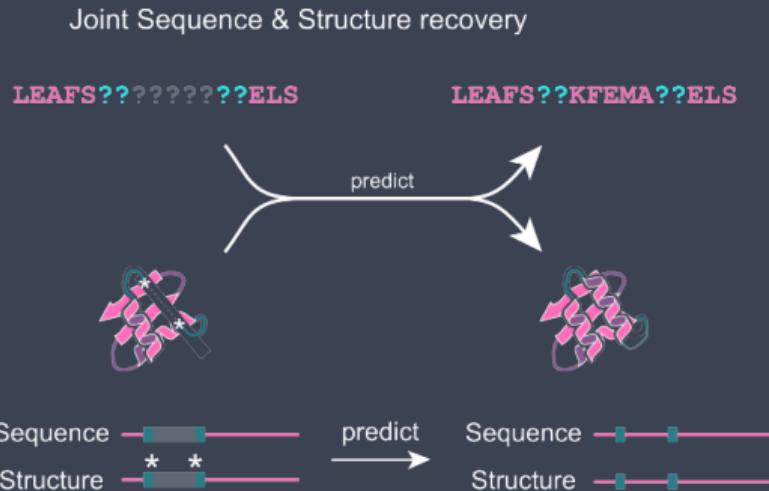
- * Широкий спектр проблем проектирования структуры белков можно аналогичным образом сформулировать как проблемы восстановления недостающей информации



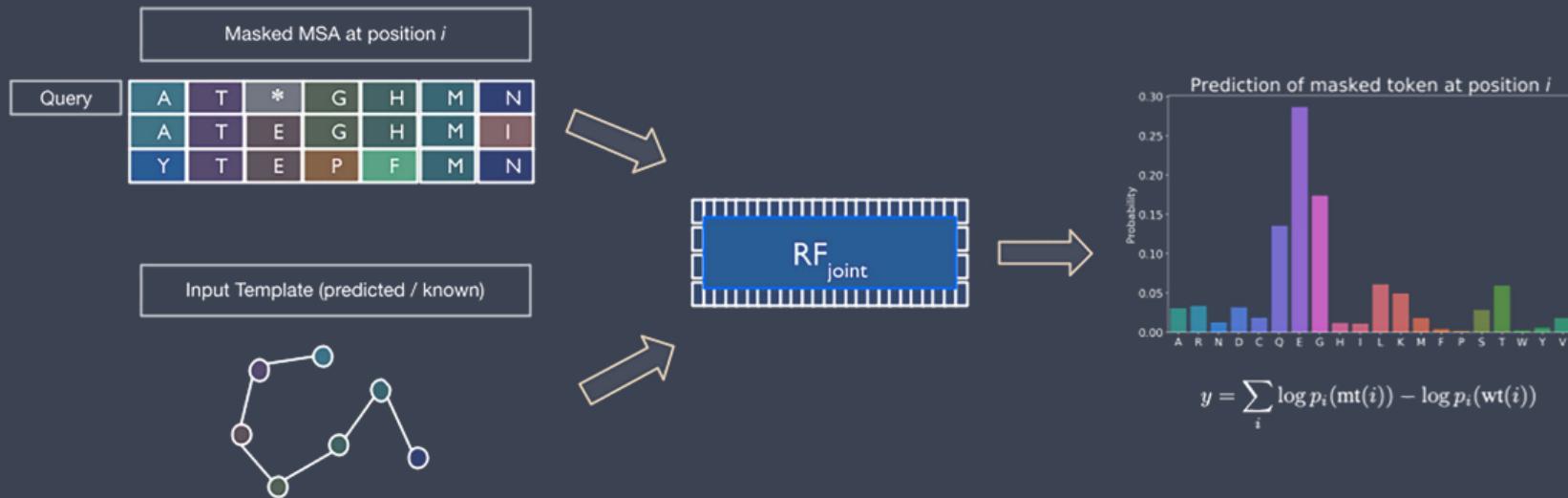
» Inpainting



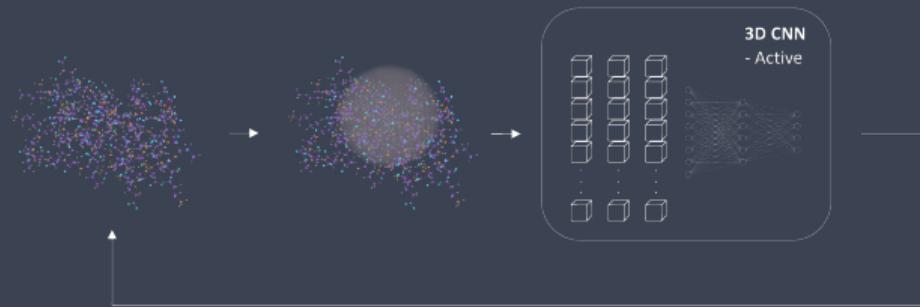
» RFjoint



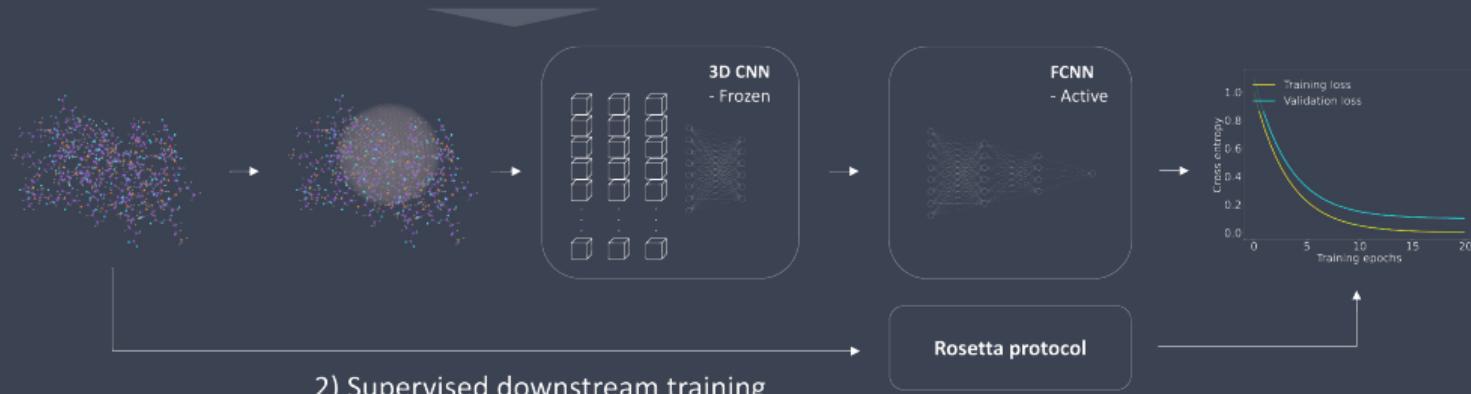
» Inpainting



» Rapid protein stability prediction



1) Self-supervised pre-training



2) Supervised downstream training



» Заключение

- * В процессе дизайна стабильности белка приходится иметь дело со сложным процессом фолдинга белка, предсказать который пока невозможно - приходится аппроксимировать
- * Выборки для обучения методов предсказания ограничены
- * Несмотря на локальные успехи - мало системных универсальных методов

