Машинное обучение для предсказания эффектов мутаций в геноме

Пензар Дмитрий Дмитриевич

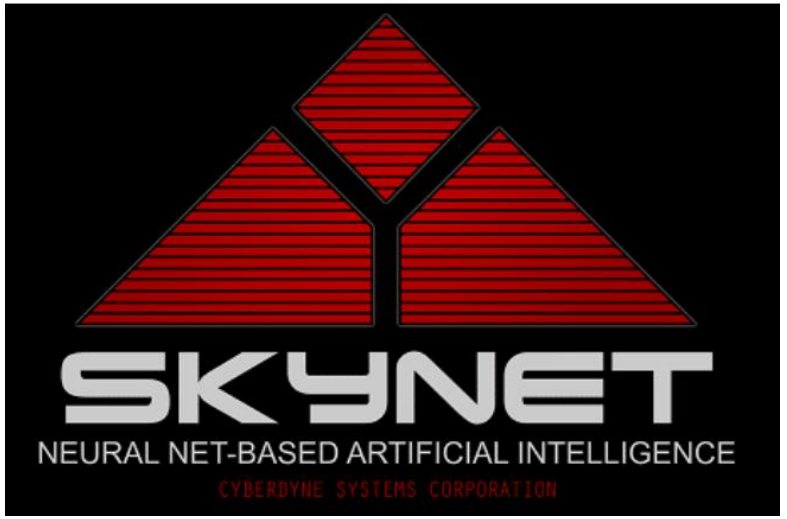
Преподаватель ФББ МГУ, сотрудник ИОГЕН РАН

Зачем?

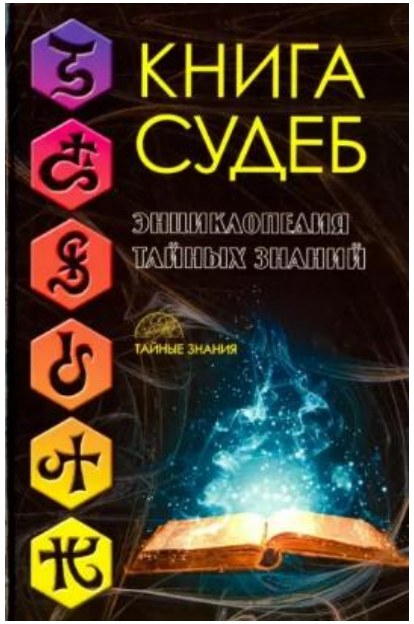
1. Большое число заболеваний обусловлено генетическими причинами
2. Часто мутация может увеличивать риск возникновения того или иного заболевания
3. Даже если мы не говорим про заболевание, хорошо бы уметь давать рекомендации по образу жизни в зависимости от генетики человека

Схема в идеальном мире

Геном человека



Сильный искуственный интеллект

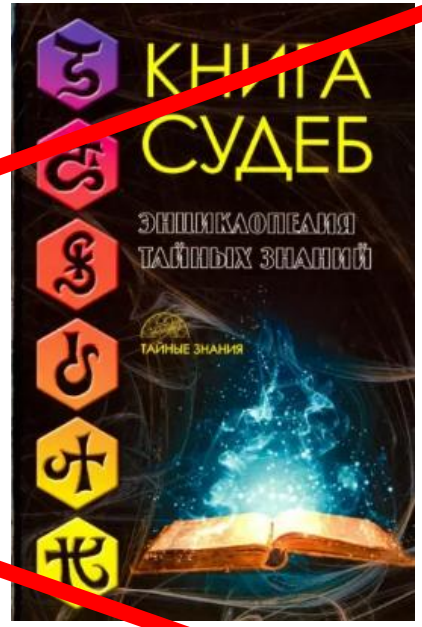


Прогноз на все случаи жизни

Схема в идеальном мире

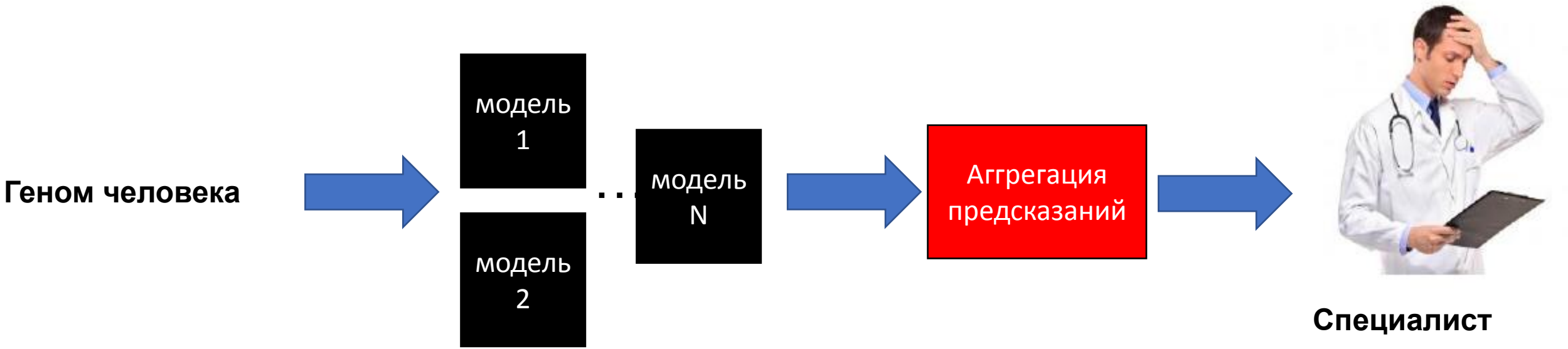
Геном человека

Сильный искуственный интеллект



Прогноз на все случаи жизни

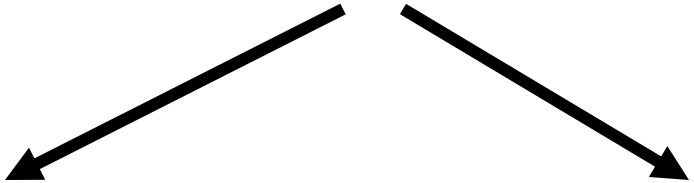
Схема в идеальном мире



Откуда эти модели получить

1. Молекулярная биология
2. Биоинформатический анализ
3. Машинное обучение на биологических данных

Мутации



Кодирующие мутации

1. *Нейтральные мутации*
2. Изменение структуры белка
3. Нарушение каталитической активности
4. Нарушение связывания с другими белками

…

Регуляторные мутации

1. *Нейтральные мутации*
2. Изменение доступности хроматина
3. Сплайсинг
4. Эффекты на уровне регуляции трансляции

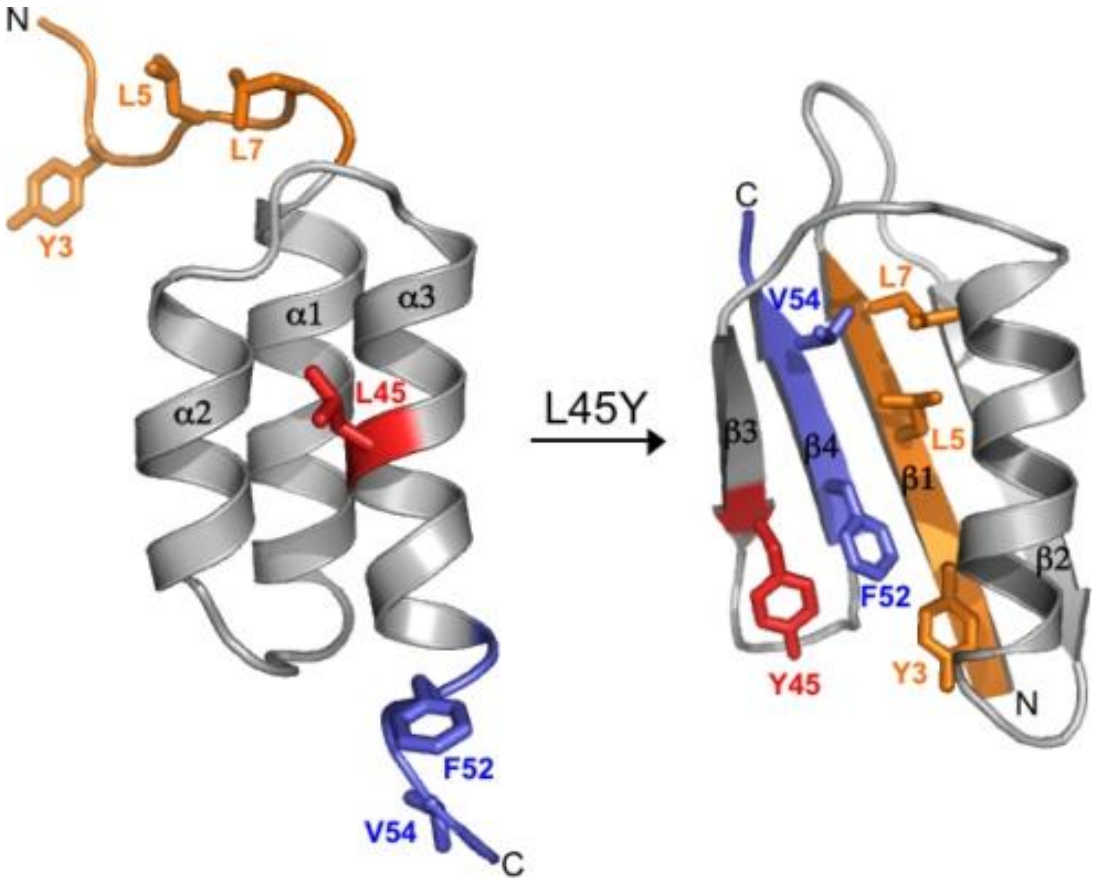
…

Общие проблемы

1. Малое количество данных
2. Большой шум в данных
3. Плохая переносимость
4. Условия эксперимента != реальные условия

Кодирующие мутации

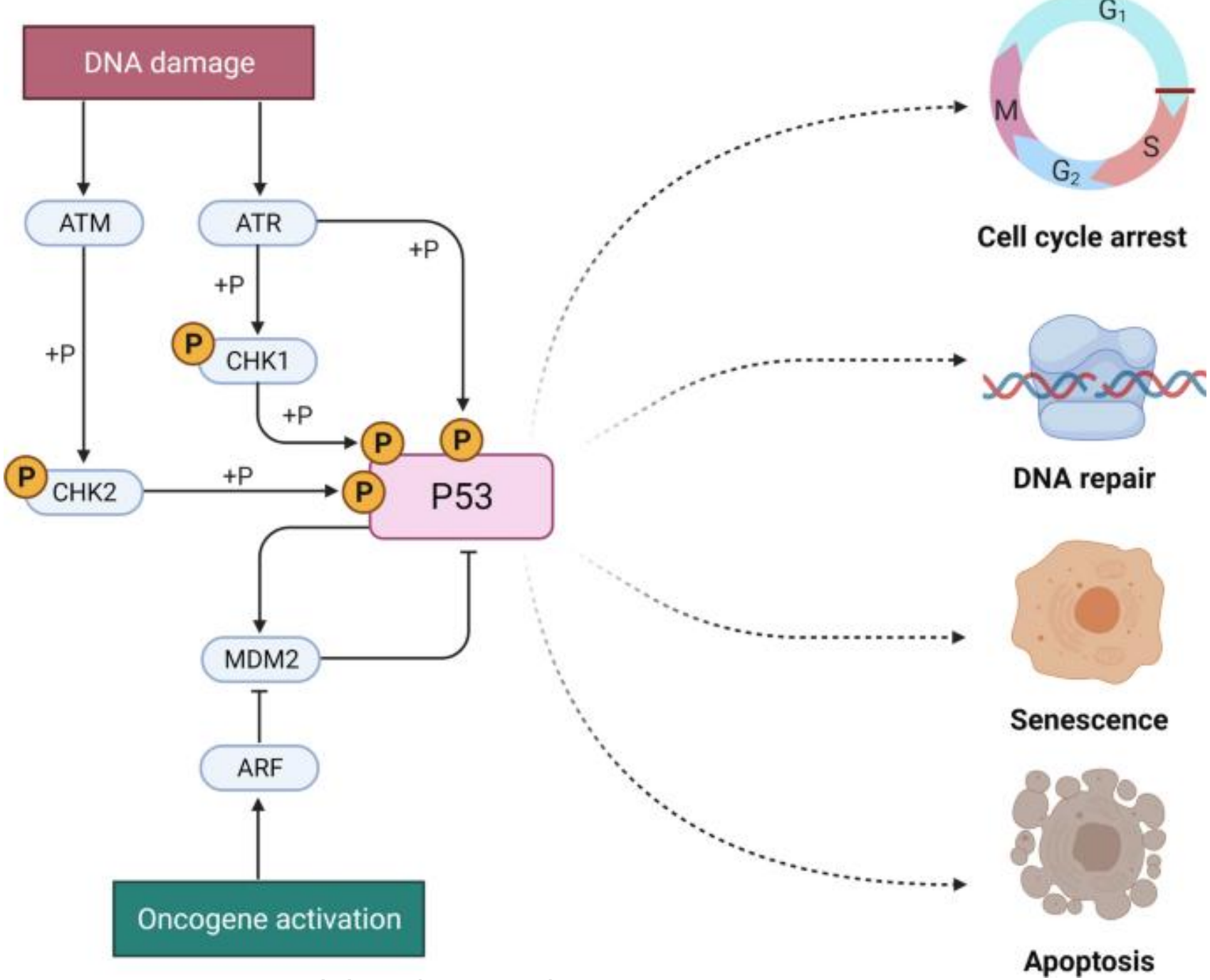
Может быть достаточно одной мутации в последовательности, чтобы поменять структуру белка до неузнаваемости. Но чаще всего эффект (хотя бы визуально) меньше



10.1073/pnas.0906408106

Серповидноклеточная анемия

P53



doi.org/10.1186/s12935-021-02396-8

AlphaFold2

Научились предсказывать структуру белка по последовательности значительно лучше предыдущих методов

Качество методов предсказания структуры, больше — лучше

Research

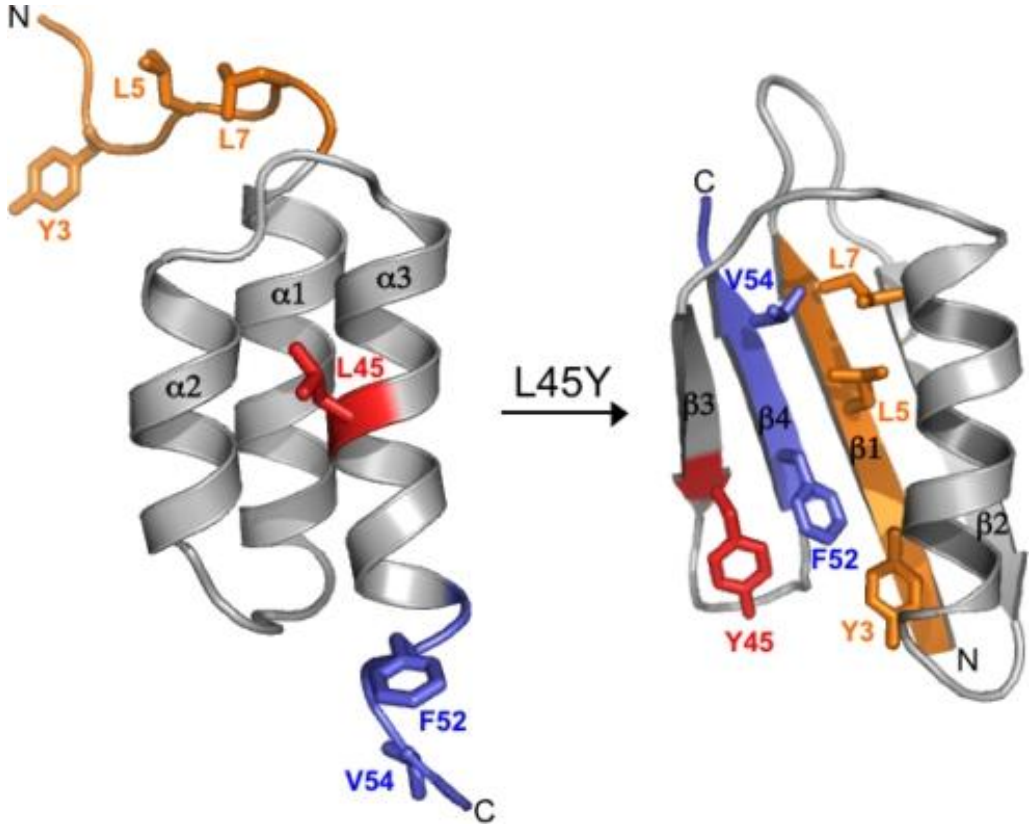
AlphaFold: a solution to a 50-year-old grand challenge in biology

November 30, 2020

1. 10.1038/s41586-021-03819-2
2. https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

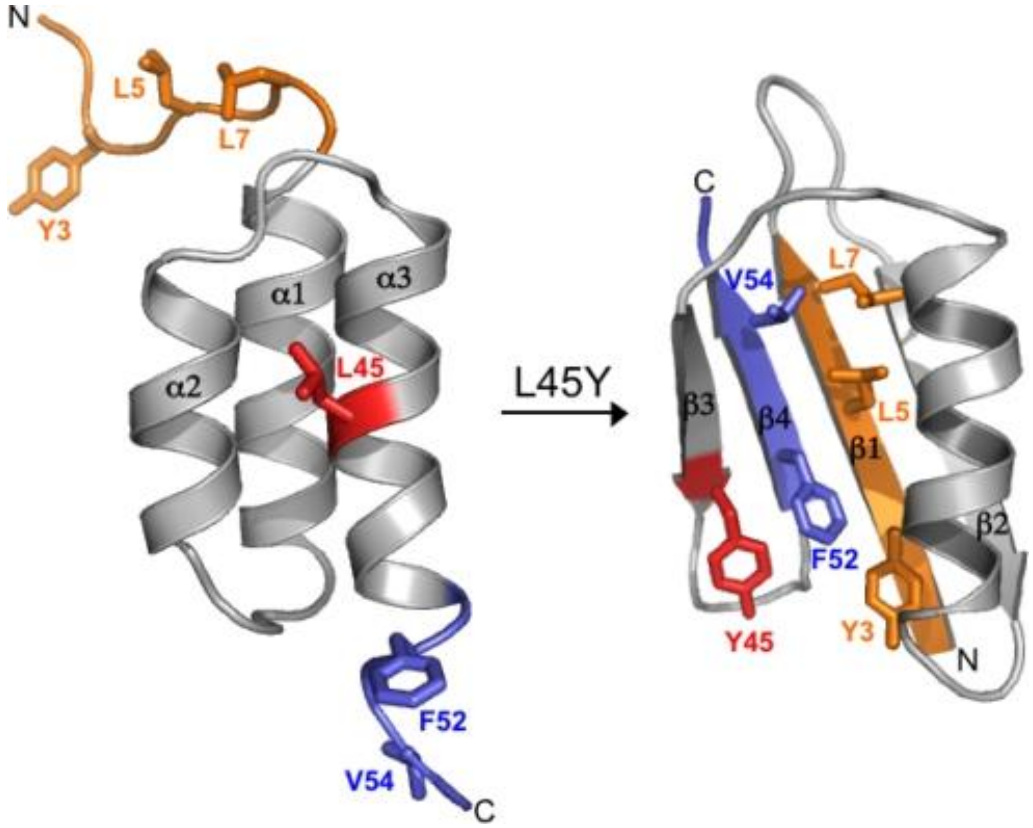
AlphaFold2

Первоначальные сообщения – AlphaFold2 решает любую задачу.
В том числе предсказывает корректно и приведенный выше пример



AlphaFold2

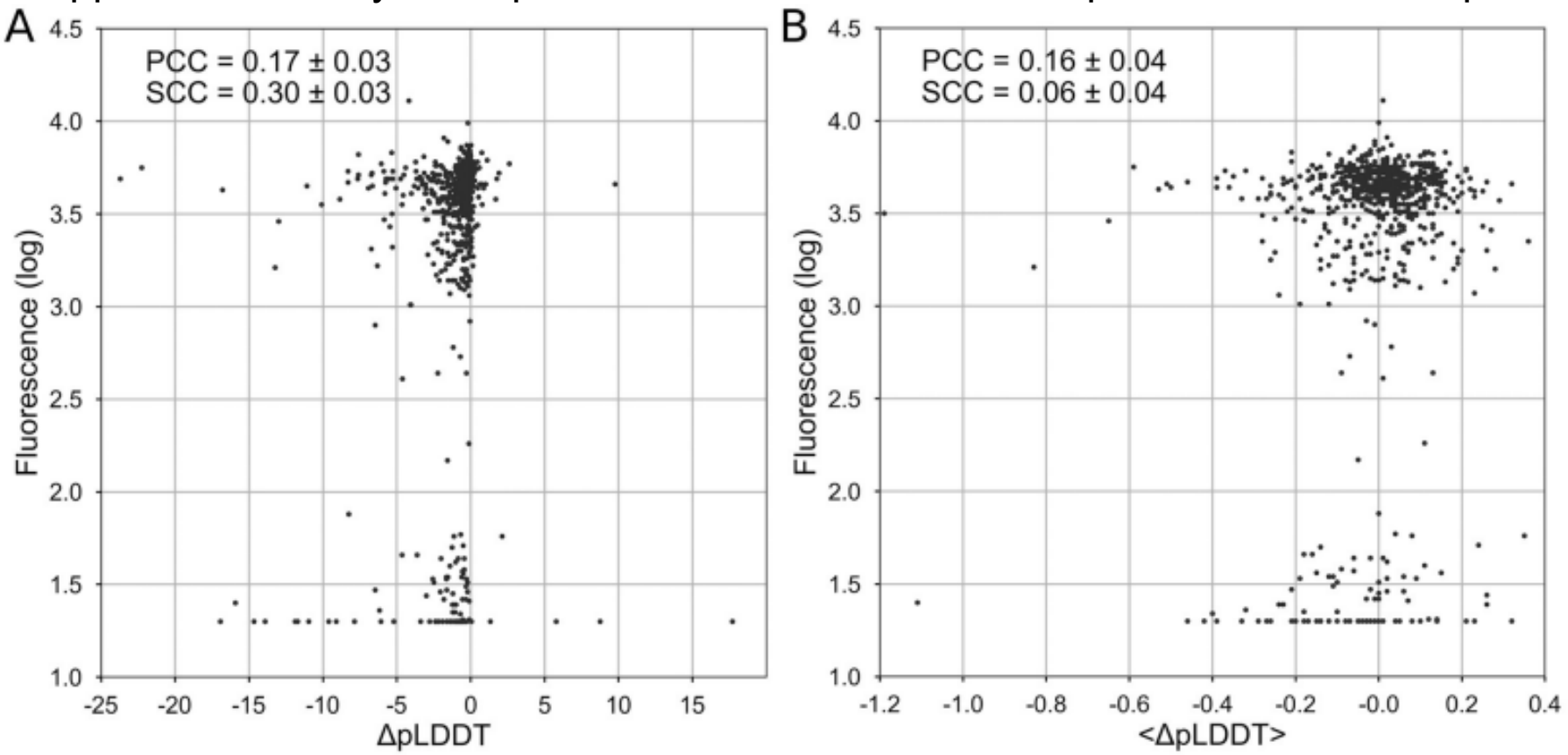
Первоначальные сообщения – AlphaFold2 решает любую задачу.
В том числе предсказывает корректно и приведенный выше пример



Нюанс: пример есть в обучающих данных – модель его видела

AlphaFold2: не-а

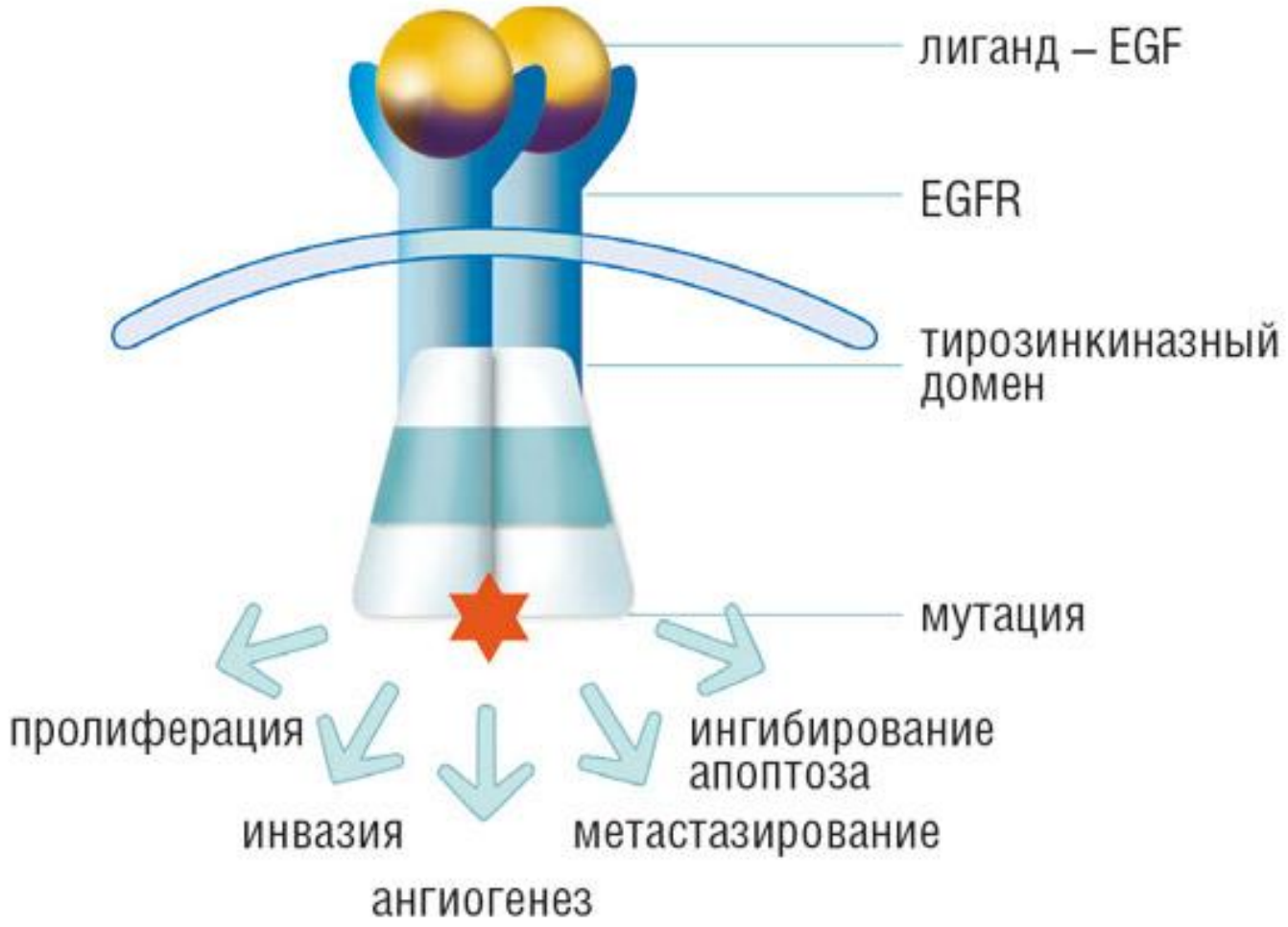
Авторы показали на нескольких датасетах (на картинке – флуоресценция мутантов белка GFP), что корреляция между экспериментальными данными и предсказанием от AlphaFold2 крайне слабая



doi.org/10.1371/journal.pone.0282689

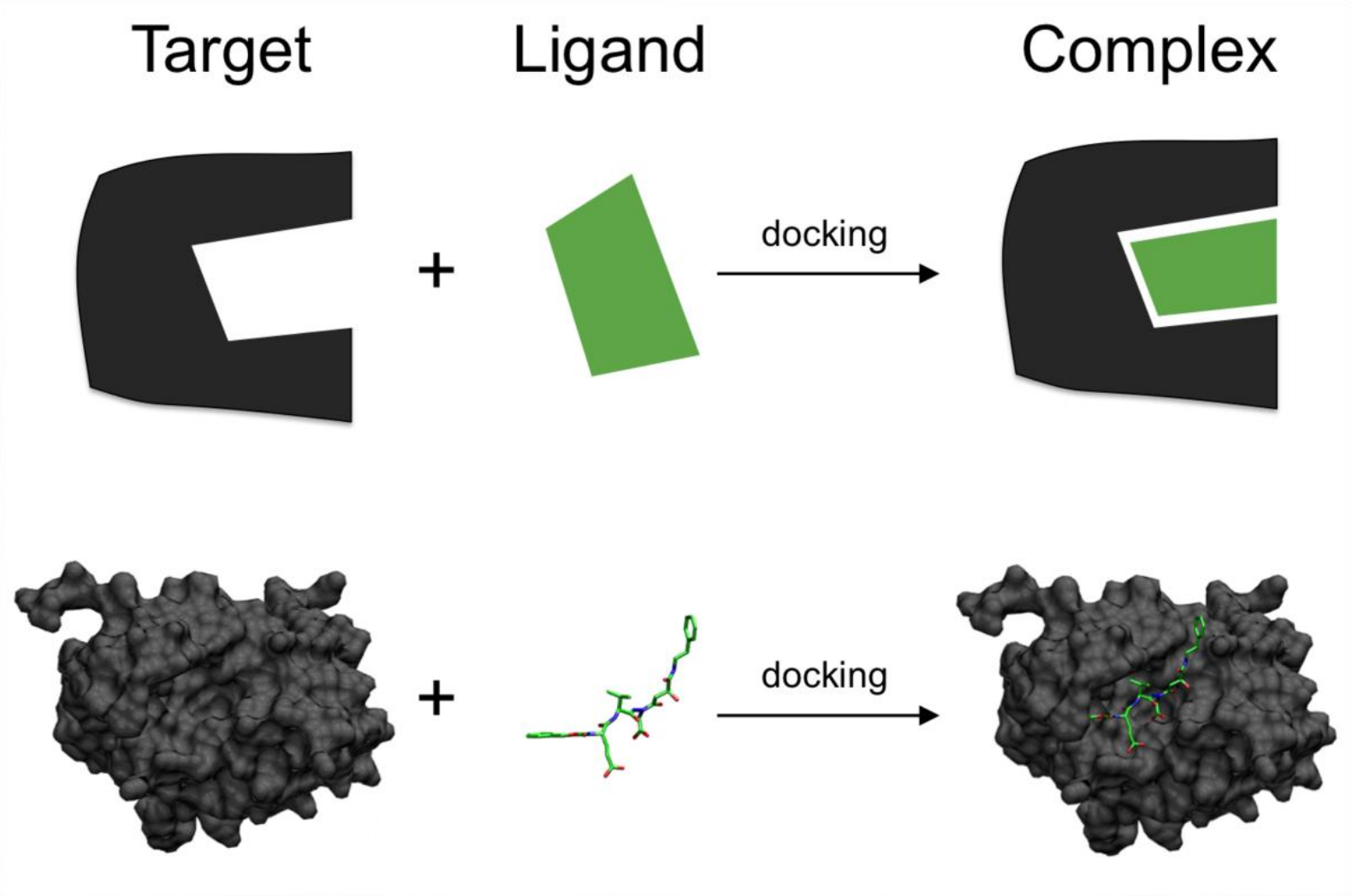
Мутации, влияющие на связывание

Мутации в рецепторе эпидермального фактора роста – частая причина развития опухолей



Докинг

Можем пытаться предсказывать такие эффекты.
А также подбирать лекарства



Cracking nuts with a sledgehammer: when modern graph neural networks do worse than classical greedy algorithms

Maria Chiara Angelini, Federico Ricci-Tersenghi

The recent work ``Combinatorial Optimization with Physics-Inspired Graph Neural Networks'' [Nat Mach Intell 4 (2022) 367] introduces a physics-inspired unsupervised Graph Neural Network (GNN) to solve combinatorial optimization problems on sparse graphs. To test the performances of these GNNs, the authors of the work show numerical results for two fundamental problems: maximum cut and maximum independent set (MIS). They conclude that "the graph neural network optimizer performs on par or outperforms existing solvers, with the ability to scale beyond the state of the art to problems with millions of variables."

In this comment, we show that a simple greedy algorithm, running in almost linear time, can find solutions for the MIS problem of much better quality than the GNN. The greedy algorithm is faster by a factor of 10^4 with respect to the GNN for problems with a million variables. We do not see any good reason for solving the MIS with these GNN, as well as for using a sledgehammer to crack nuts.

In general, many claims of superiority of neural networks in solving combinatorial problems are at risk of being not solid enough, since we lack standard benchmarks based on really hard problems. We propose one of such hard benchmarks, and we hope to see future neural network optimizers tested on these problems before any claim of superiority is made.

https://doi.org/10.48550/arXiv.2206.13211

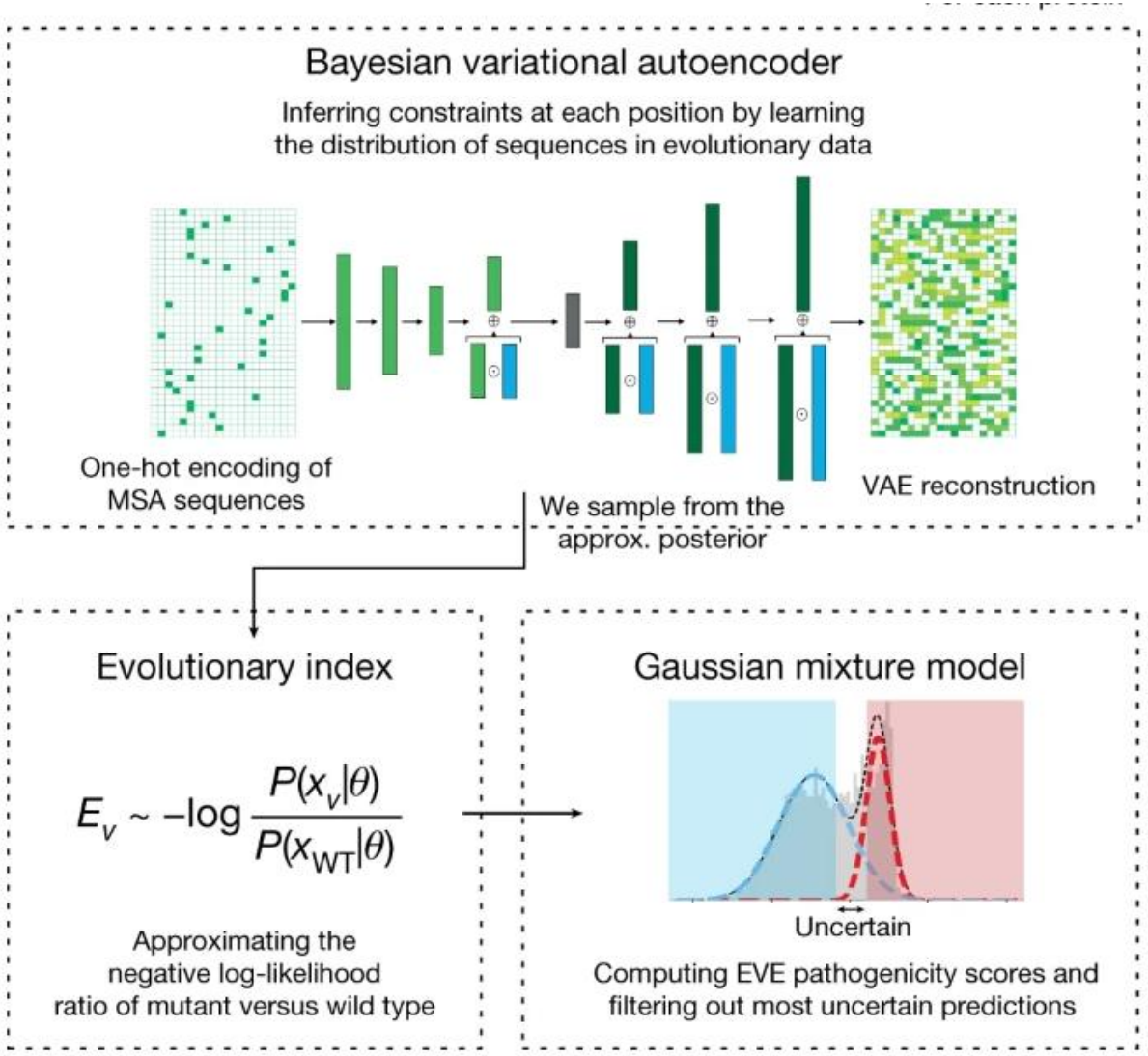
PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences

Martin Buttenschoen, Garrett M. Morris, Charlotte M. Deane

The last few years have seen the development of numerous deep learning-based protein-ligand docking methods. They offer huge promise in terms of speed and accuracy. However, despite claims of state-of-the-art performance in terms of crystallographic root-mean-square deviation (RMSD), upon closer inspection, it has become apparent that they often produce physically implausible molecular structures. It is therefore not sufficient to evaluate these methods solely by RMSD to a native binding mode. It is vital, particularly for deep learning-based methods, that they are also evaluated on steric and energetic criteria. We present PoseBusters, a Python package that performs a series of standard quality checks using the well-established cheminformatics toolkit RDKit. Only methods that both pass these checks and predict native-like binding modes should be classed as having ``state-of-the-art'' performance. We use PoseBusters to compare five deep learning-based docking methods (DeepDock, DiffDock, EquiBind, TankBind, and Uni-Mol) and two well-established standard docking methods (AutoDock Vina and CCDC Gold) with and without an additional post-prediction energy minimisation step using a molecular mechanics force field. We show that both in terms of physical plausibility and the ability to generalise to examples that are distinct from the training data, no deep learning-based method yet outperforms classical docking tools. In addition, we find that molecular mechanics force fields contain docking-relevant physics missing from deep-learning methods. PoseBusters allows practitioners to assess docking and molecular generation methods and may inspire new inductive biases still required to improve deep learning-based methods, which will help drive the development of more accurate and more realistic predictions.

Unsupervised – обучение на неразмеченных данных

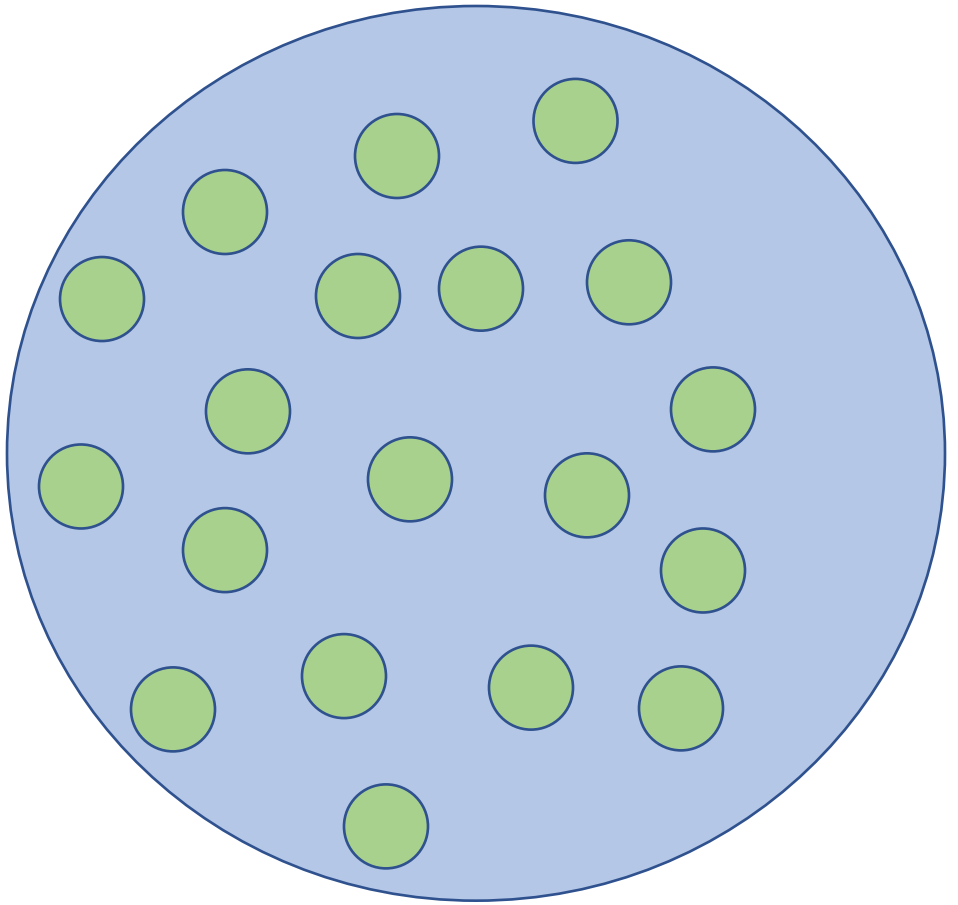
1. Выучим модель, которая переводит **природные** последовательности в некое непрерывное пространство. Мы их наблюдаем – значит, звери с ними живут достаточно успешно
2. Если наша последовательность с мутацией оказывается близко к природным – она не вредная. Если далеко — вредная



10.1038/s41586-021-04043-8

Unsupervised – обучение на неразмеченных данных

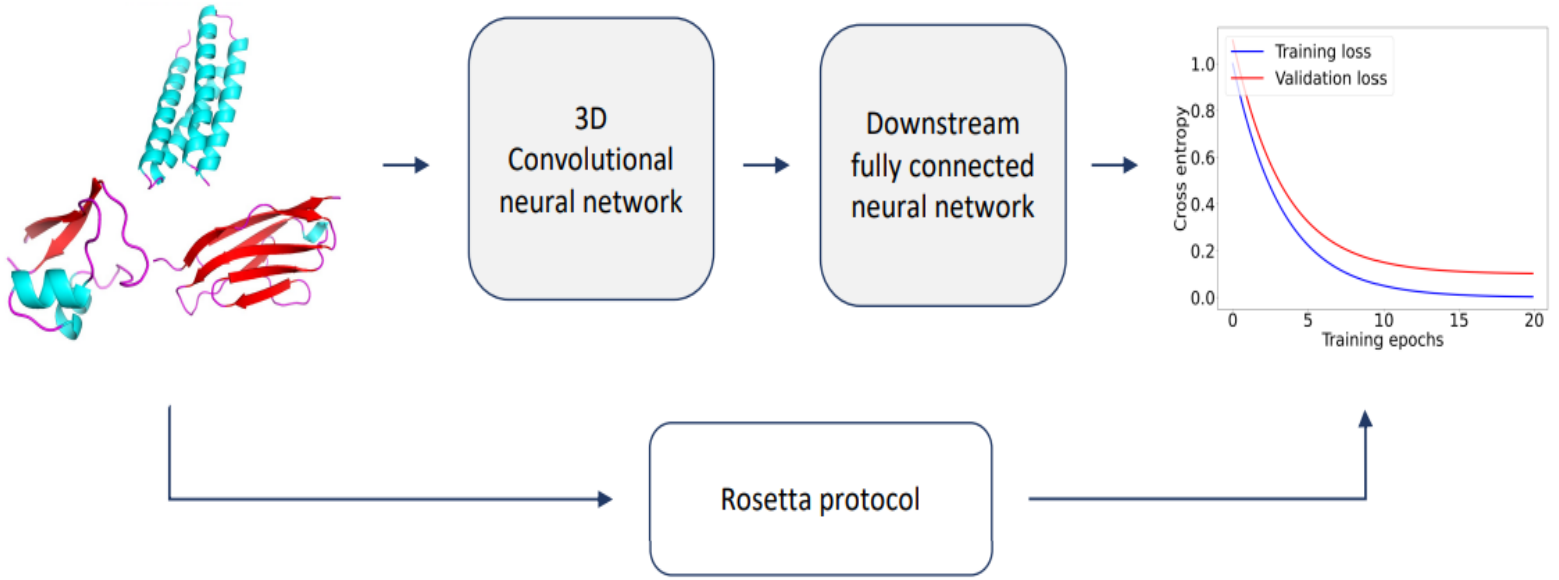
1. Выучим модель, которая переводит **природные** последовательности в некое непрерывное пространство. Мы их наблюдаем – значит, звери с ними живут достаточно успешно
2. Если наша последовательность с мутацией оказывается близко к природным – она не вредная. Если далеко — вредная



10.1038/s41586-021-04043-8

Unsupervised(Self-supervised) + Supervised

1. Сначала учим хорошее представление на неразмеченных данных
2. Используем его на размеченных синтетических данных – эффекты мутаций предсказаны при помощи пакета для молекулярного моделирования Rosetta.
3. Проверяем на реальных данных – предсказывает!
4. Что интересно, один из немногих примеров данной архитектуры нейросетей, хорошо работающих для белков



10.1101/2022.07.14.500157

Появление “массовых” данных под задачу

Mega-scale experimental analysis of protein folding stability in biology and protein design

Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J. Weinstein, Niall M. Mangan, Sergey Ovchinnikov, Gabriel J. Rocklin

doi: https://doi.org/10.1101/2022.12.06.519132

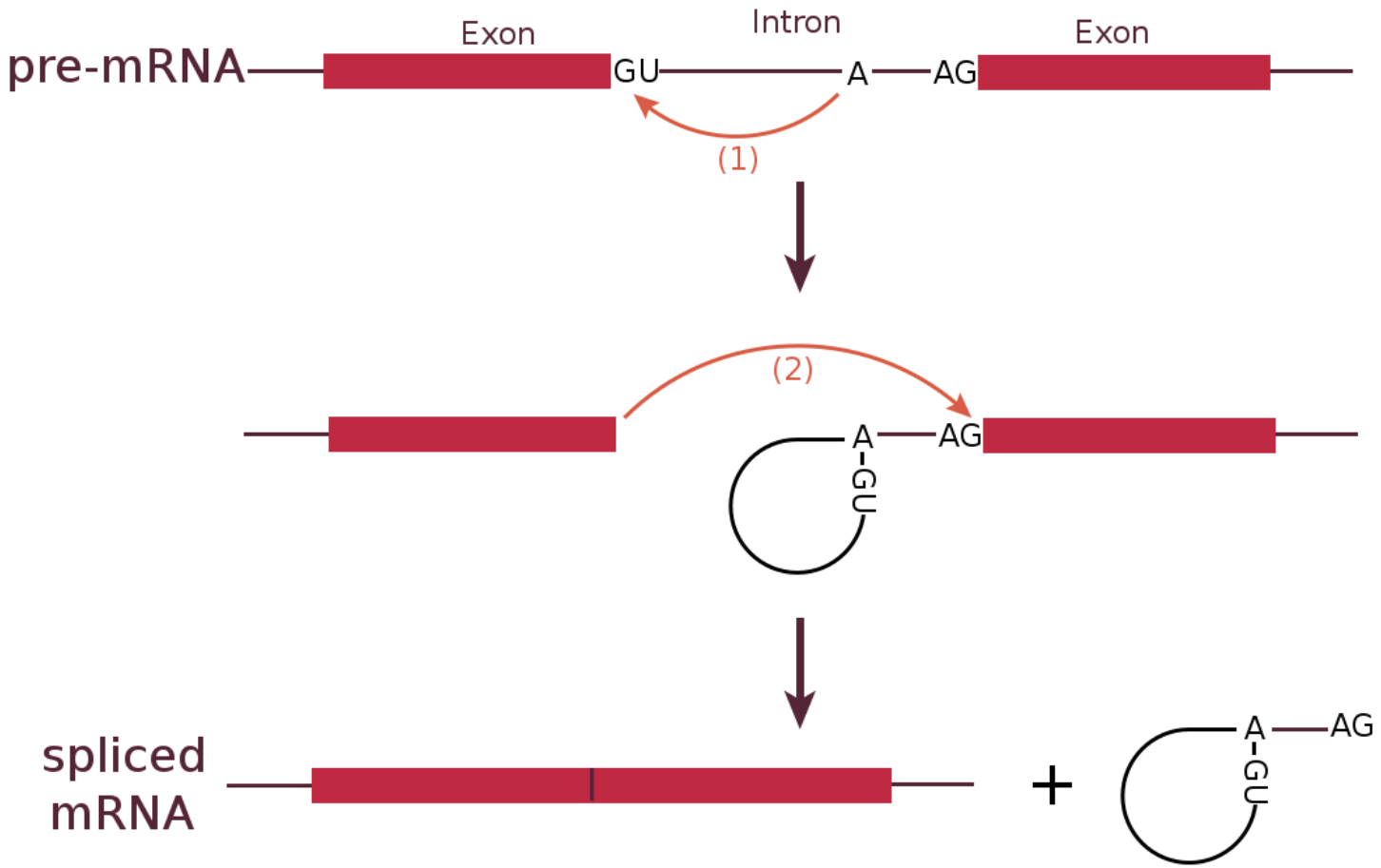
New mega dataset combined with deep neural network makes a progress in predicting impact of mutation on protein stability

Marina A Pak, Nikita V Dovidchenko, Satyarth Mishra Sharma, Dmitry N Ivankov

doi: https://doi.org/10.1101/2022.12.31.522396

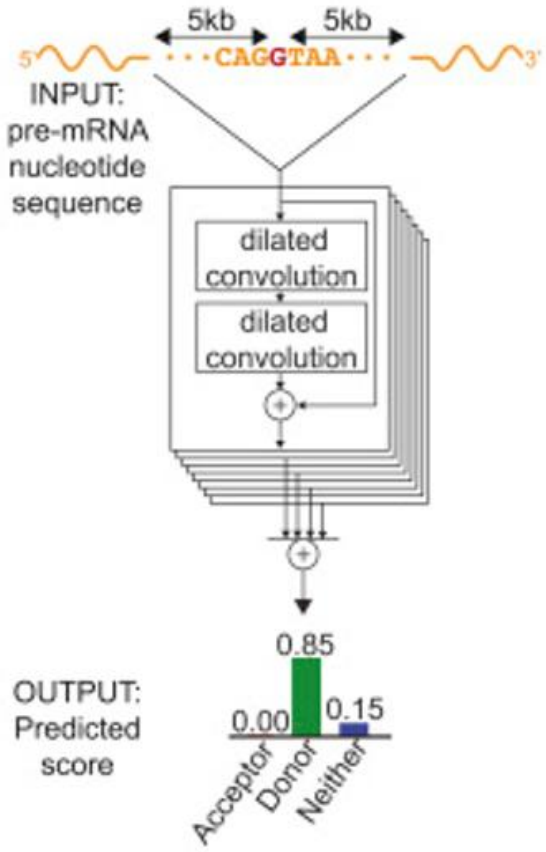
Регуляторные мутации

Сплайсинг


pre-mRNA
Exon
Intron
Exon
GU
A
AG
(1)
(2)
A
AG
A-GU
spliced
mRNA
+
A
AG
A-GU

Регуляторные мутации. Сплайсинг

1. Данных о точечных эффектах замен на сплайсинг мало
2. Научимся предсказывать для конкретной позиции вероятность быть сайтом сплайсинга
3. Будем использовать разницу оценок для последовательности с мутацией и без как оценку эффекта мутации



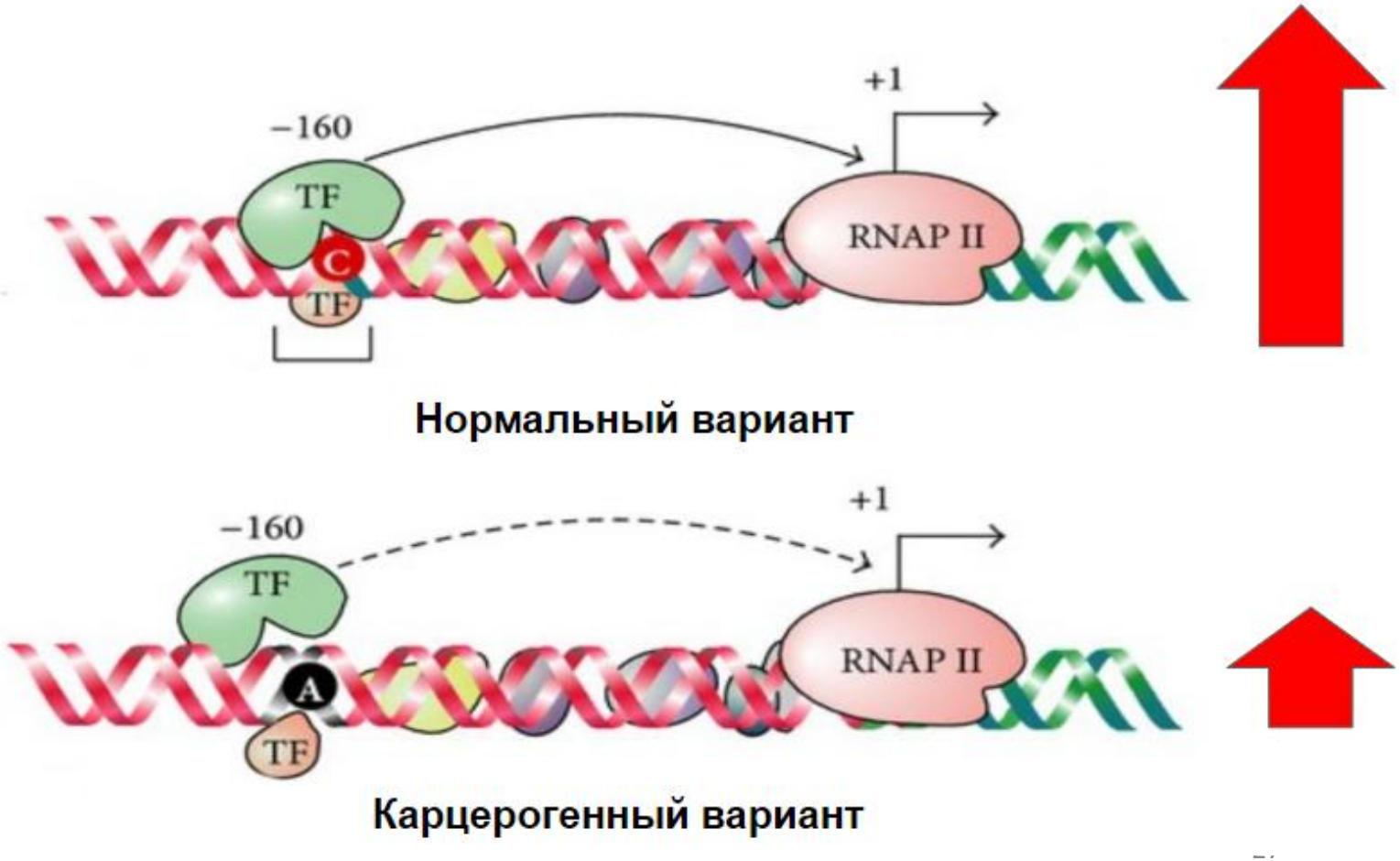
Model(ATGCACCACACAC) – Model(ATGCACGACACAC)
=
Эффект мутации

10.1016/j.cell.2018.12.015

Регуляторные мутации. Связывание транскрипционного фактора

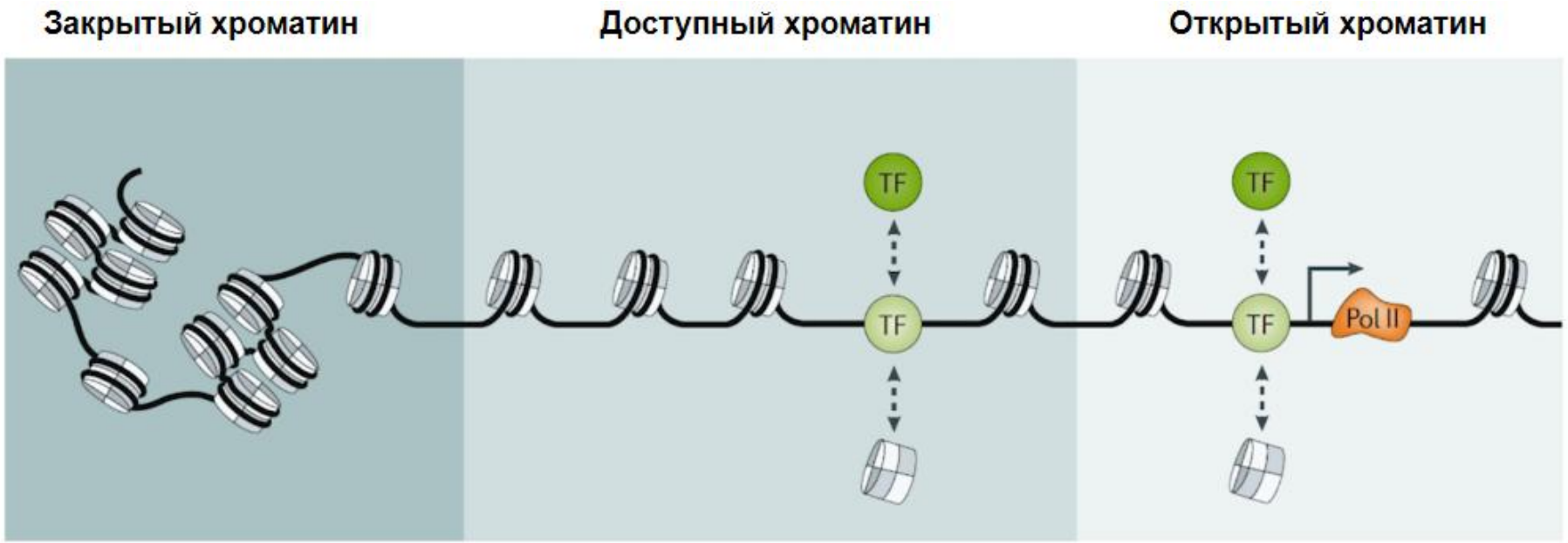
Многие важные регуляторные мутации расположены в сайтах связывания транскрипционных факторов

Можно решать вместе с более общей проблемой



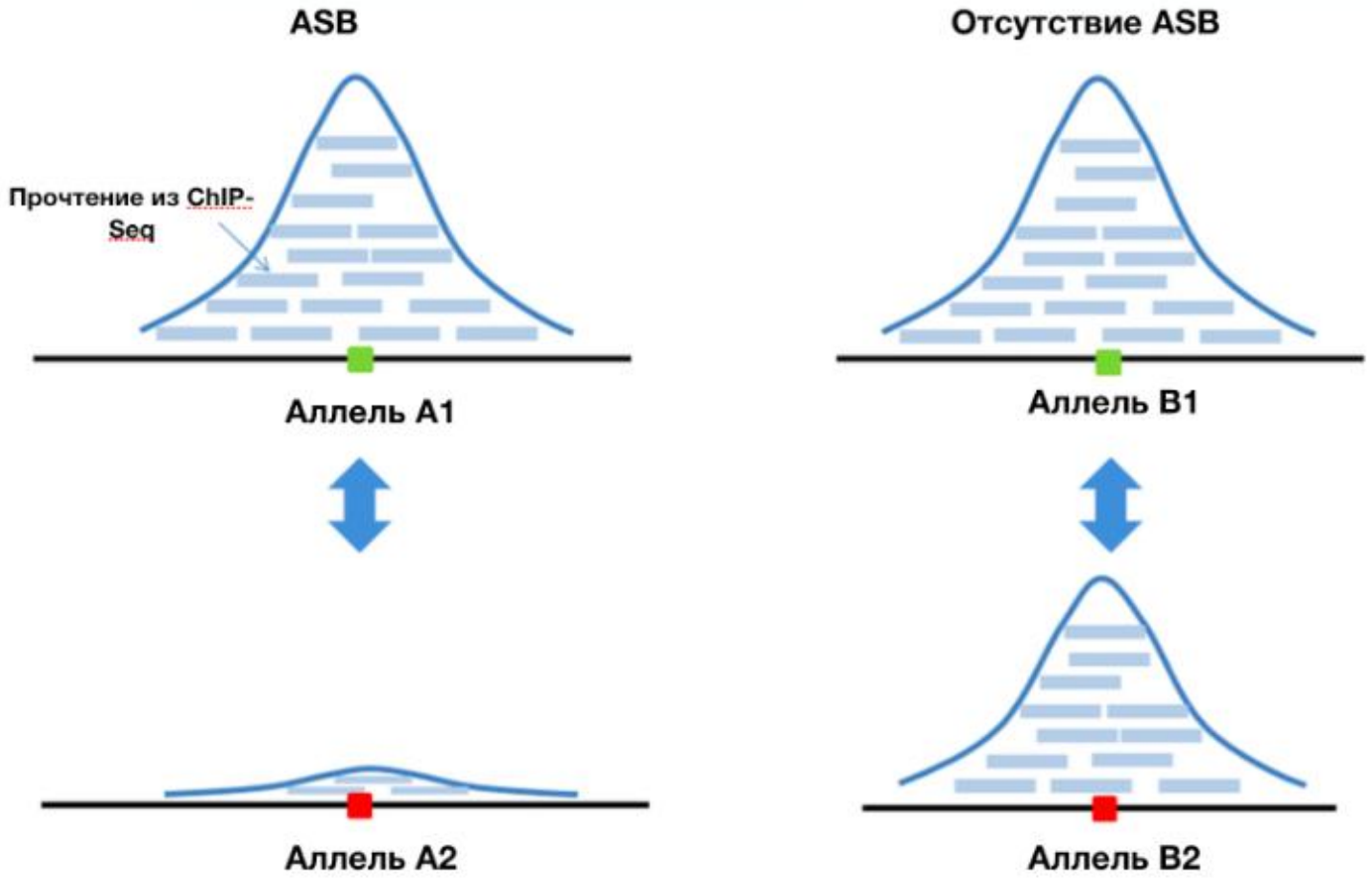
Регуляторные мутации. Открытость хроматина

Самый частый эффект мутации в регуляторных регионах – изменения открытости хроматина. Из-за изменения связывания конкретно ТФ или нет – не всегда важно



Прямых данных мало и они «прямые»

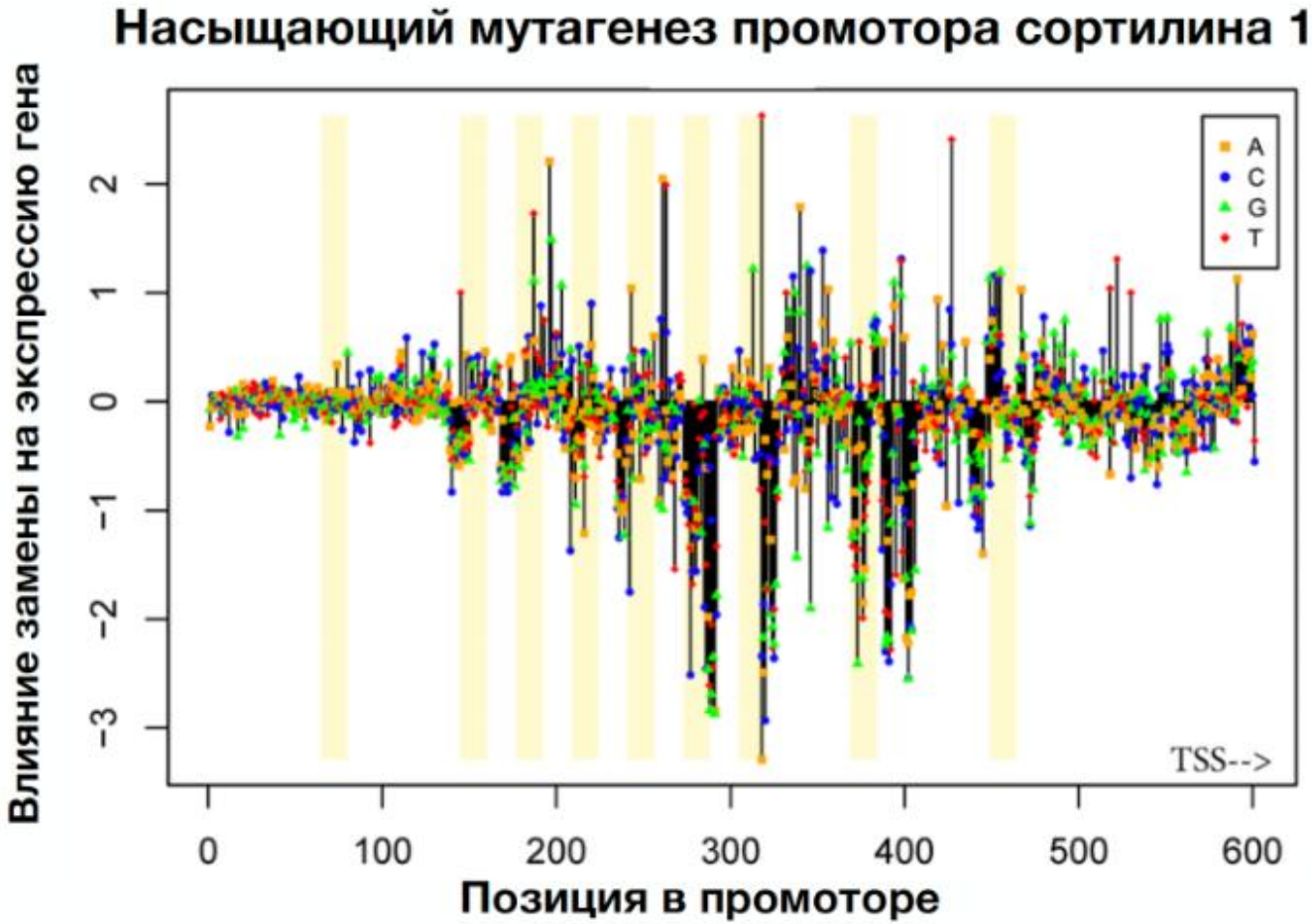
1. Есть данные об аллель-специфичном связывании
2. Большая часть – очень шумная и получена косвенным методом
3. Можно использовать только для валидации



10.1038/s41467-021-23007-0

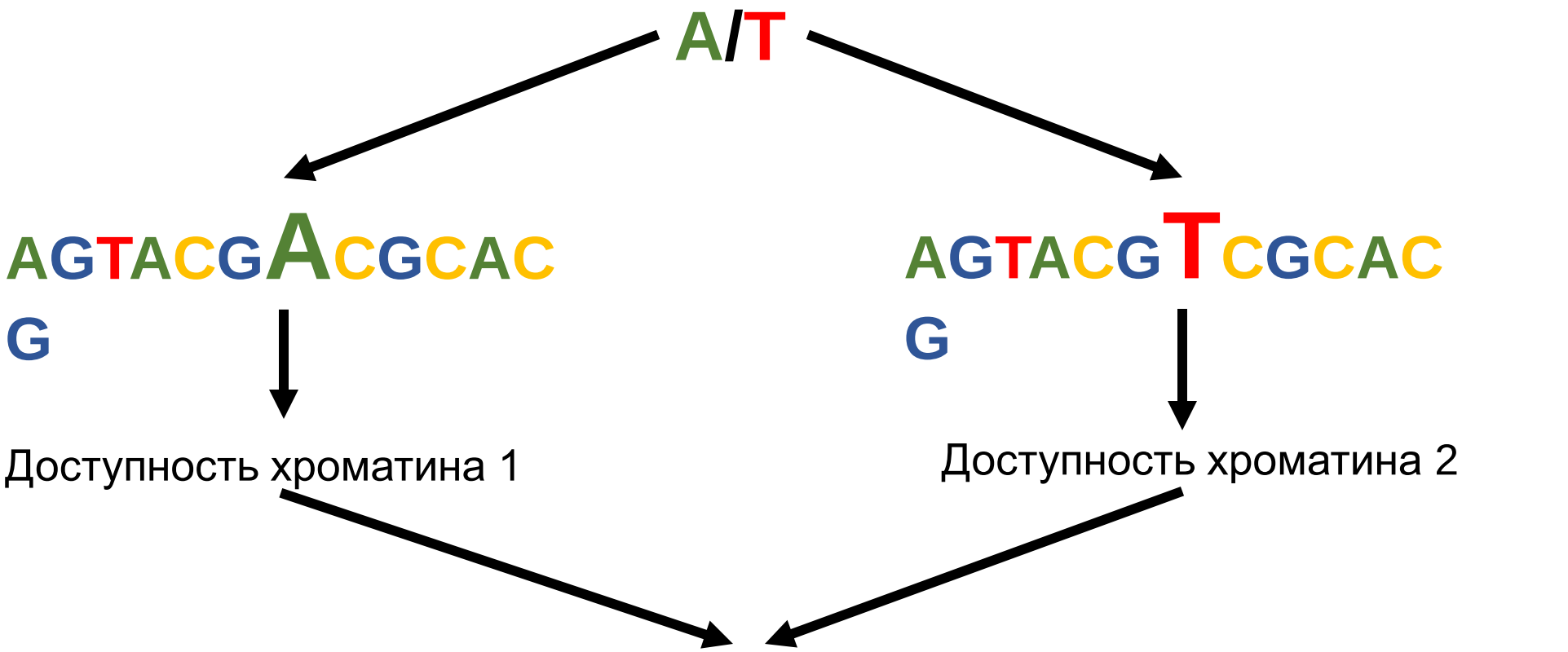
Прямых данных мало и они «прямые»

1. Есть данные об эффекте мутаций в конкретных регуляторных последовательностях
2. До недавнего времени – данных мало
3. Большая часть – очень шумная и получена косвенным методом
4. Можно использовать только для валидации



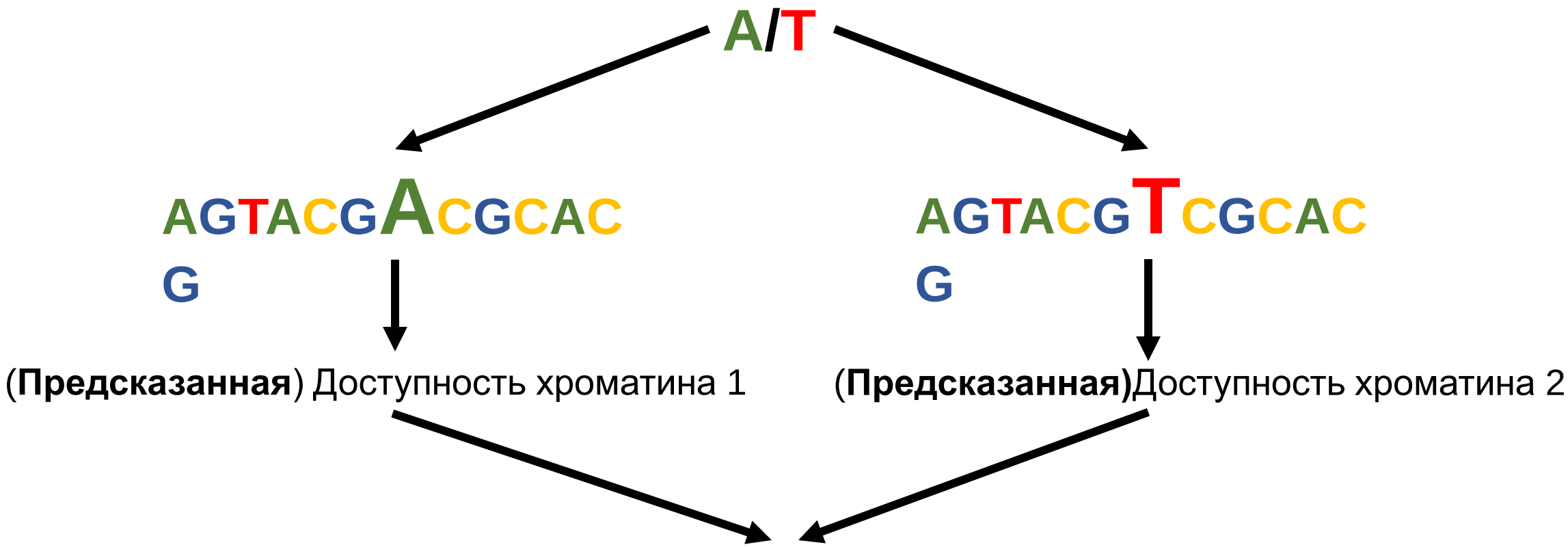
10.1002/humu.23797

Как мы предсказывали эффект, зная доступность?



Влияние замены = Доступность хроматина 2 – Доступность хроматина 1

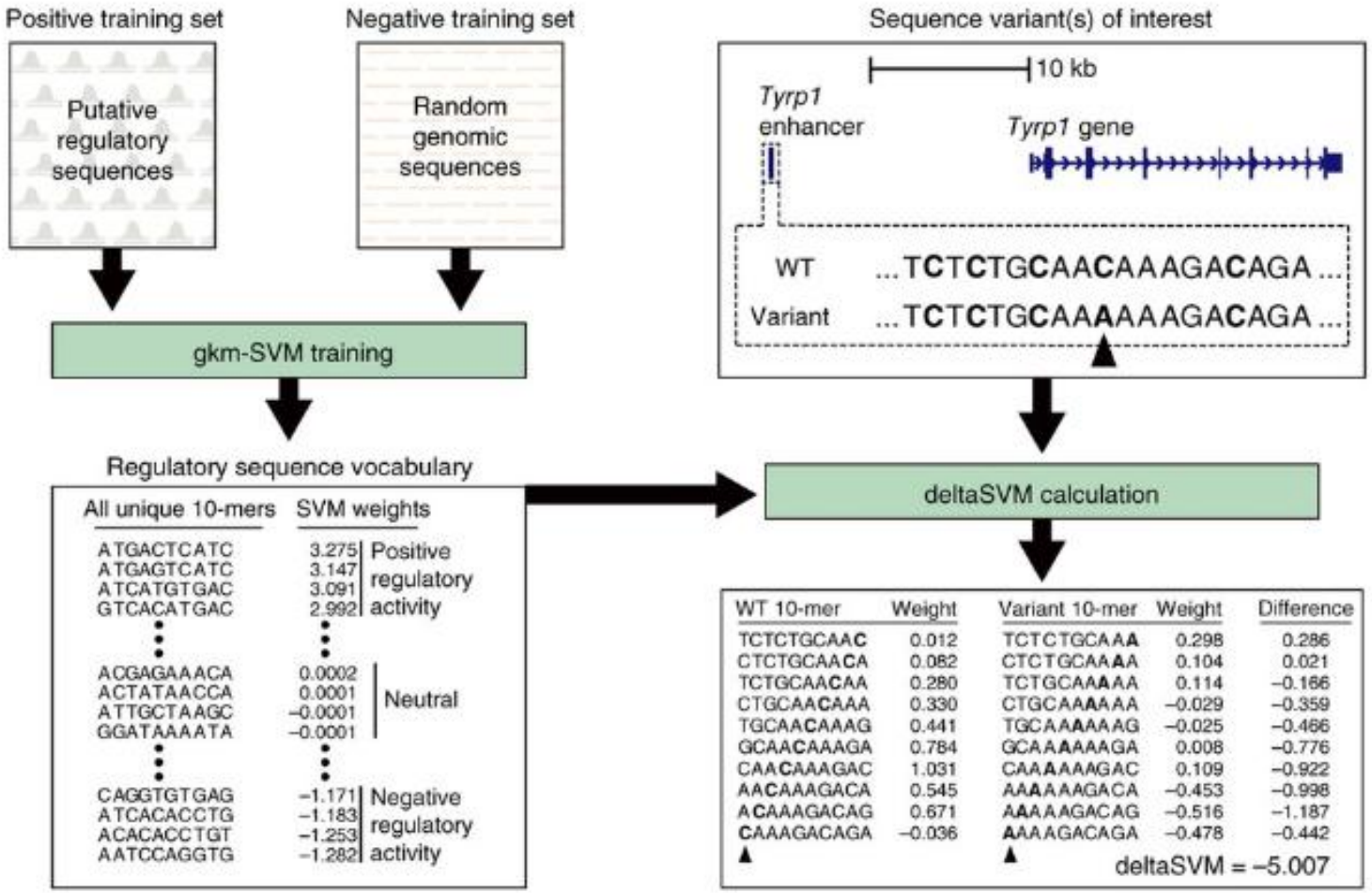
Общая схема **косвенного*** предсказания



** Для увеличения импакта статьи пишем "**zero-shot learning**", т.к мы учимся на одних данных, и предсказываем другие, не обучаясь на них специально*

Gkm/deltaSVM

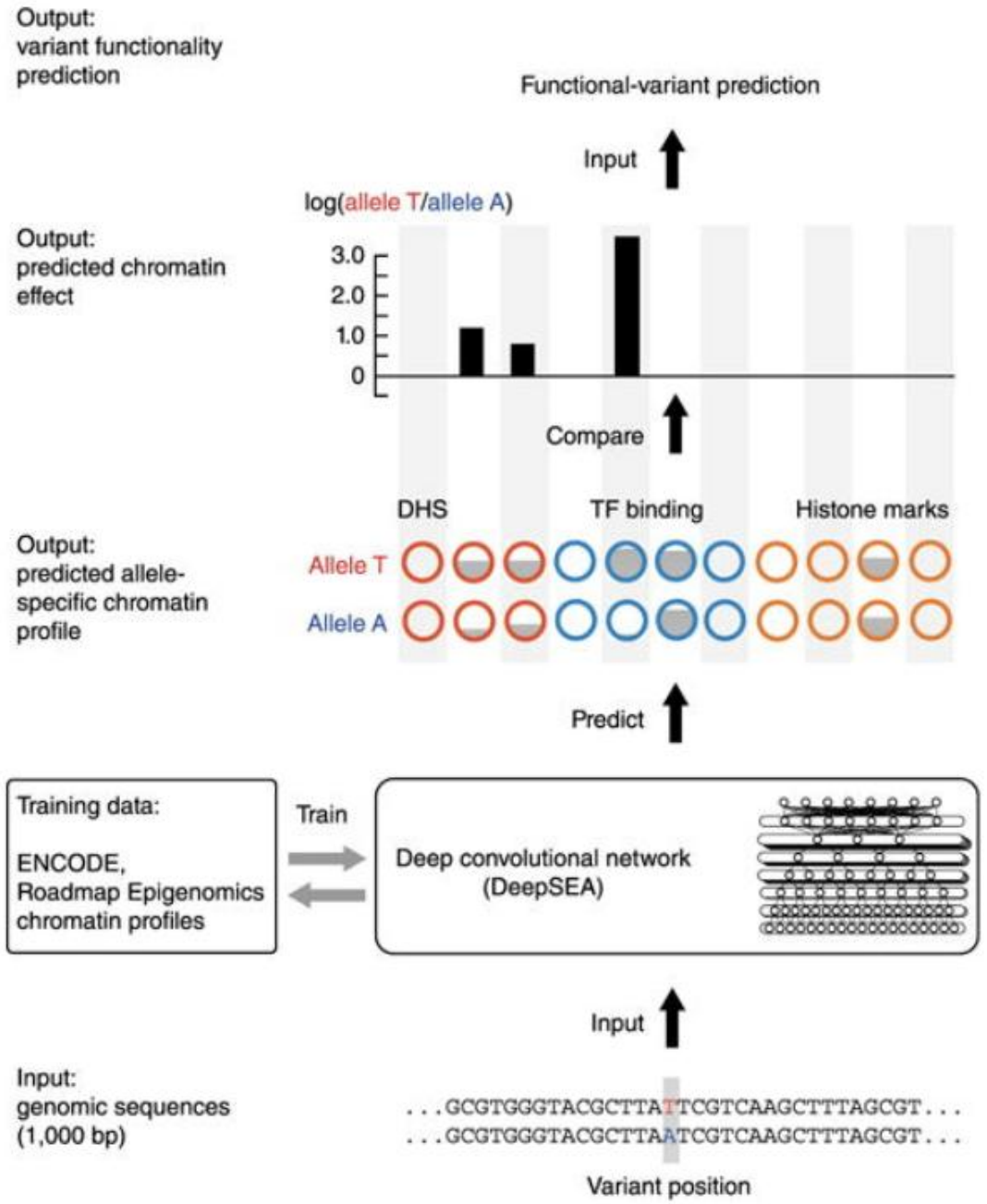
1. Основана на методе опорных векторов и его последующем «огрублении»
2. Предсказываем доступность хроматина бинарно – доступен или нет
3. Из SVM можно вытащить число, которое в нулевом приближении соответствует уверенности модели в предсказании класса
4. Используем это число в предыдущей схеме



doi.org/10.1038/ng.3331

DeepSEA

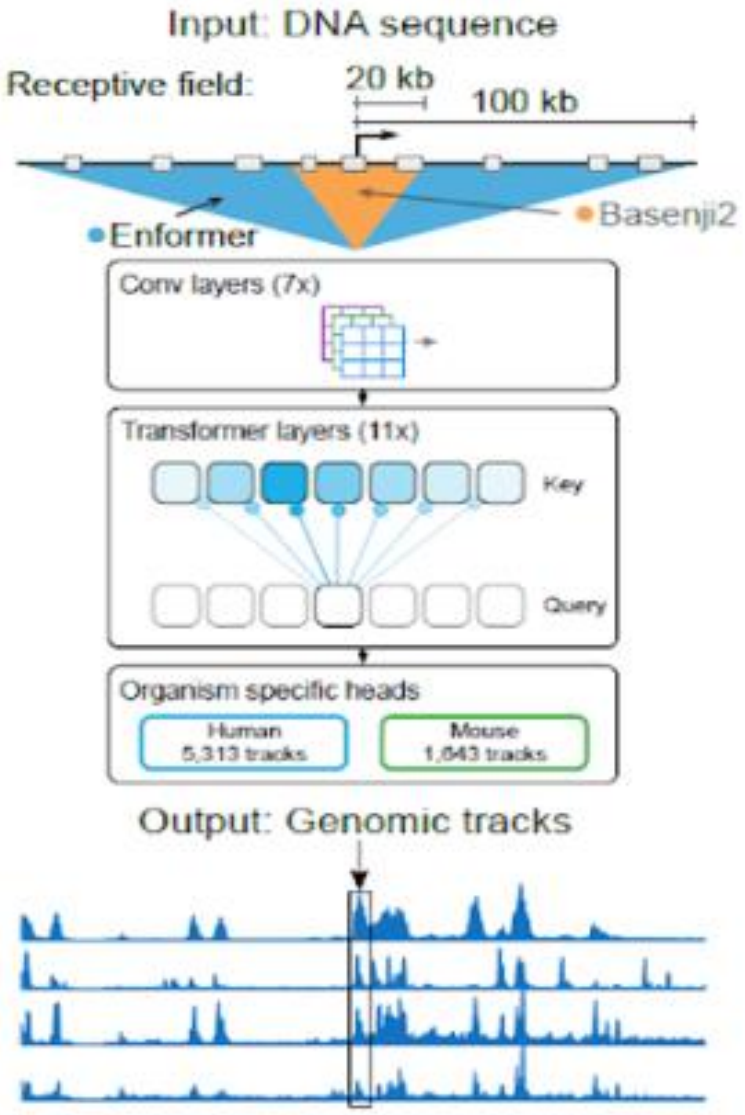
1. Учим нейронную сеть предсказывать результаты сразу 919 экспериментов (тоже – есть или нет связывание, классификация)
2. Для этого используем последовательность длины 1000н
3. Можем посчитать эти предсказания для последовательности без и с заменой
4. Далее обучаем на этих предсказаниях, используя маленький датасет, дополнительную модель
5. Потеряли zero-shot, приходится **дообучаться** на шумных данных



10.1038/nmeth.3547

Enformer

1. Учим нейронную сеть предсказывать **численные** результаты сразу **7000+** экспериментов (регрессия)
2. Используем последовательность сильно бóльшего размера (100кб)
3. Можем посчитать эти предсказания для последовательности без и с заменой
4. Далее обучаем на этих предсказаниях, используя маленький датасет, дополнительную модель
5. Но можно просто усреднить предсказания по клеточным линиям, близким целевой или активным ТФ – тоже получается хорошее предсказание – **zero-shot** вернулся



Задача решена?

1. DeepMind заявил, что да
2. Нет, не решена

Нюансы

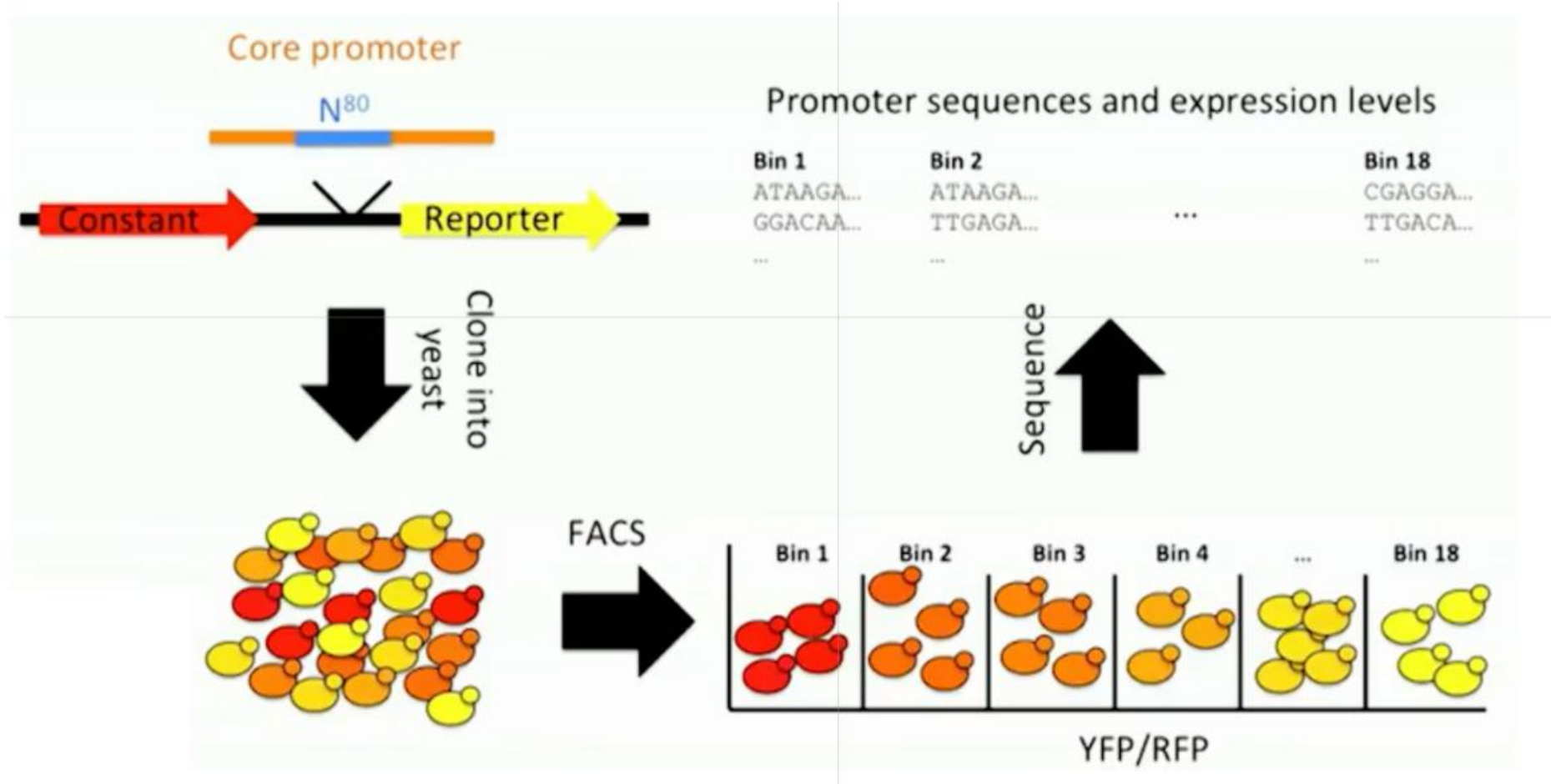
1. При корректном анализе оказывается, что Enformer никак не использует бóльшую глубину окна – не в состоянии улавливать эффекты, проявляющиеся на больших расстояниях (10.1101/2022.09.15.508087)
2. Главное заявление статьи – научились предсказывать эффекты мутаций на экспрессию – ложно. Аккуратное измерение этого не подтверждает (10.1101/2023.03.16.532969v1)
3. Интересно, что силу эффекта получается предсказать лучше, чем направление
4. Enformer обучался на очень большом числе экспериментов – нужны очень аккуратно выбирать данные для оценки его качества, иначе легко взять данные, которые модель уже видела (**data leakage**)
5. Мы показали (https://twitter.com/dmitrypenzar/status/1640667989371416578 , в процессе редактирования статьи), что при обучении на достаточно чистых прямых данных, можно победить Enformer. При условии, что эти данные он частично видел...

Прямые данные

Появляются методы, позволяющие напрямую оценивать эффект последовательности на экспрессию, и делать это массово и с сравнительно низким шумом

Massively parallel reporter essays (дрожжи)

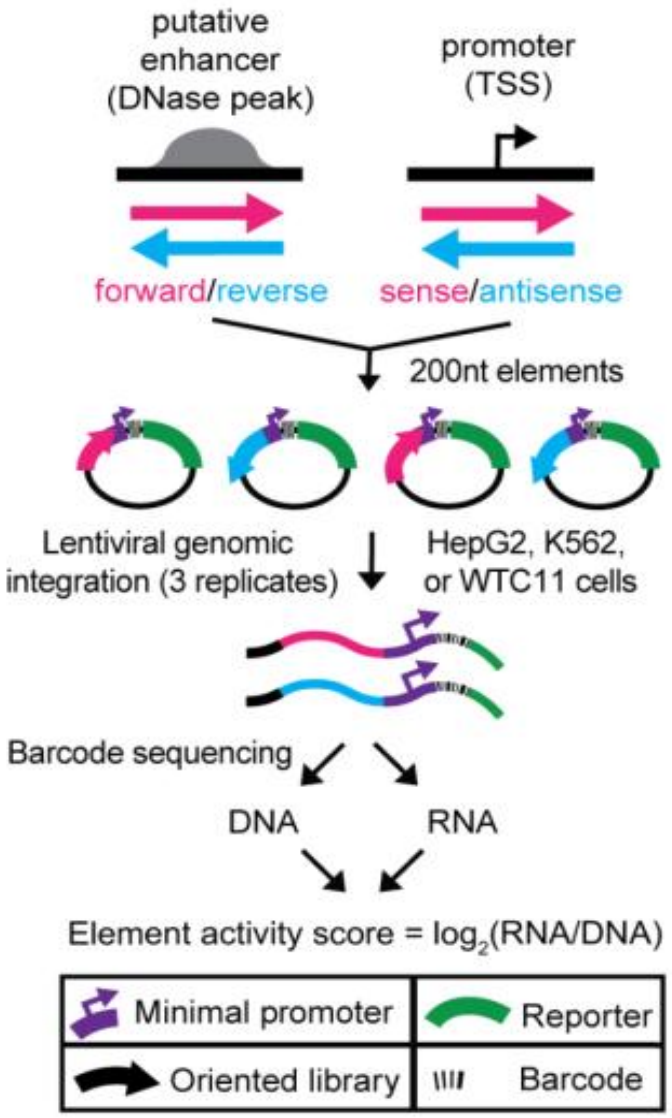
1. Вставляют случайные последовательности длины 80н перед геном экспрессируемого белка
2. Кроме того, в конструкции есть константно экспрессируемый белок
3. Если последовательность сильно увеличивает экспрессию – клетка будет светиться желтым. Если сильно понижает – красным. И т.д
4. Далее остается отсортировать клетки по светимости и отсеквенировать



10.1038/s41587-019-0315-8

Massively parallel reporter essays (человек)

1. Делают и эксперименты, дающие в том числе непрерывные измерения
2. Ограничения пока те же, что в прошлом эксперименте – размер последовательности небольшой, и конструкции не позволяют оценивать дальние взаимодействия



doi.org/10.1101/2023.03.05.531189