МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

имени М.В. ЛОМОНОСОВА

На правах рукописи

Пензар Дмитрий Дмитриевич

Вычислительное предсказание эффектов мутаций в регуляторных районах генов

1.5.8 – Математическая биология, биоинформатика

Диссертация на соискание ученой степени

кандидата биологических наук

Научный руководитель:

д.б.н.

Иван Владимирович Кулаковский

Москва – 2025

Список сокращений и условных обозначений

ДНК – дезоксирибонуклеиновая кислота

РНК – рибонуклеиновая кислота

SNP – однонуклеотидный полиморфизм

3’НТО – 3’ нетранслируемые области

5’НТО – 5’ нетранслируемые области

ПЦР – полимеразная цепная реакция

QTL – Quantitative Trait Loci, локус количественных признаков

eQTL – expression Quantitative Trait Loci, локусы, отвечающие за изменение в экспрессии

sQTL – splicing Quantitative Trait Loci, локусы, отвечающие за изменение в частоте включения экзона

GWAS – Genome-Wide Association Studies, полногеномные ассоциативные исследования

CRISPR – Clustered Regularly Interspaced Short Palindromic Repeats, антивирусная защита бактерий, использующаяся в множестве технологий направленного редактирования генома с тем же названием

MPRA (МПРЭ) – massively parallel reporter assay, массовый параллельный эксперимент с репортерами

ТФ – транскрипционный фактор

п.о – пар оснований

н.т – нуклеотидов

кб п.о – 1000 ("килобаза") пар оснований

мб п.о – миллион ("мегабаза") пар оснований

ChIP-Seq – Chromatin Immunoprecipitation Sequencing, высокопроизводительное секвенирование результатов иммунопреципитации хроматина

АС – аллель-специфичное событие

АСС – событие аллель-спецфичного связывания транскрипционного фактора

АСД – событие аллель-специфичной доступности хроматина

АСЭ – событие аллель-специфичной экспрессии

АСМ – событие аллель-специфичной (хроматиновой) метки

ПВМ – Позиционно-весовая матрица

SELEX – Systematic Evolution of Ligands by EXponential Enrichment, Систематическая эволюция лигандов экспоненциальным обогащением

HT-SELEX – high-throughput SELEX, высокопроизводительный SELEX

обучающая выборка – выборка, на которой происходит обучение модели машинного обучения

тестовая выборка – выборка, на которой происходит тестирование

LSTM – Long short-term memory, Долгая краткосрочная память, разновидность архитектуры рекурентных нейронных сетей

1. Введение

Реализация генотипа организма в онтогенезе является одной из центральных тем генетики как науки. Понимание причин фенотипических различий между индивидуумами в популяции является классическим фундаментальным вопросом, а для человеческой популяции – еще и важнейшим шагом на пути к персонализированной медицине и повсеместному внедрению методов генной терапии в клиническую практику [1–3]. Преобладающая часть различий как между отдельными индивидуумами, так и между конкретным индивидуумом и аннотированным "референсным" геномом, приходится на однонуклеотидные геномные варианты [4]. Эффекты мутаций, затрагивающих последовательность белков, научились предсказывать с приемлемой точностью и даже инкрементальные темпы дальнейшего улучшения предсказаний позволяют рассчитывать на скорое достижимое клинически-достаточного уровня достоверности [5,6].

Интерпретация возможного влияния некодирующих замен затруднена сложностью, неоднозначностью и многоуровневостью регуляторного кода [7], ведь эффект мутации может действовать на различных уровнях – от транскрипции и сплайсинга до стабильности и активности мРНК [1,8].

Представленная работа посвящена биоинформатическим методам предсказания влияния вариантов в некодирующих участках генома на транскрипцию генов. Рассматриваются как классические подходы, основанные на позиционно-весовых матрицах, так и новые методы на основе классического машинного обучения и искусственных нейронных сетей в парадигме т.н. "глубокого обучения". В работе обсуждаются ограничения существующих подходов на основе полногеномных моделей, обученных предсказывать эпигенетические геномные разметки по нуклеотидной последовательности [9–13], демонстрируется завышенное качество этих подходов, вызванное характерной для геномных задач утечкой данных. Проводится оценка реально достижимой точности подобных моделей в задаче предсказания аллель-специфичного связывания факторов транскрипции. Также в работе представлен новый метод машинного обучения на основе современных полносверточных нейронных сетей для работы с данными массовых параллельных репортерных экспериментов (МПРЭ) [14–20]. Демонстрируется применимость разработанного метода для моделирования регуляторных последовательностей различных организмов, а итоговое качество предсказаний превосходит современные альтернативные подходы, включая полногеномные модели. Наконец, демонстрируется возможность модификации полученного решения для генерации некодирующих нуклеотидных последовательностей с заданными свойствами.

1.1. Актуальность темы исследования

Уже достигнутая доступность и продолжение снижения стоимости высокопроизводительного секвенирования постепенно переводят прочтение индивидуального генома из области продвинутого исследовательского инструментария в рутинную лабораторную практику [21,22]. Использование полногеномной информации об индивидуальных вариантах для ранней диагностики заболеваний и подбора персонализированной терапии перестает ограничиваться стоимостью лабораторной работы, и "бутылочным горлышком" становится эффективность и применимость вычислительных методов для аннотации индивидуального генома, в частности, полнотой баз данных, необходимых для аннотации и интерпретации функциональных последствий конкретных геномных вариантов [21,23–25]. В то время как для аннотации замен в белок-кодирующих районах уже существуют общепринятые и хорошо себя зарекомендовавшие подходы [5,6], инструменты для анализа нуклеотидных замен в некодирующих областях генов, на которые приходится порядка 90% клинически значимых мутаций [26–31], требуют активного развития и новых решений. Сегодня перспективным направлением считается использование методов искусственного интеллекта, в частности, ансамблей деревьев решений и моделей глубокого обучения для вычислительного представления предсказания активности регуляторных областей генов, использующих различные омиксные данные, полученных как в полногеномных и полнотранскритомных исследованиях в живых клетках, так и в результате массовых параллельных репортерных экспериментов [1].

Для предсказания активности регуляторных районов генов и эффектов однонуклеотидных замен в них сегодня перспективными принято считать «полногеномные» вычислительные модели, обученные, например, на данных об экспрессии генов и эпигенетических профилях генома, таких, как доступность хроматина для фрагментации нуклеазами, локализация различных модификаций гистонов или участков связывания факторов транскрипции [1,32]. Однако, уже понятно что полногеномных данных оказывается недостаточно: даже достаточно совершенные полногеномные модели не справляются с оценкой вклада малых изменений, таких как однонуклеотидные варианты, в регуляцию экспрессии генов [9–13]. Новое решение пришло с развитием МПРЭ, которые позволяют одновременно измерять активность тысяч и миллионов различных последовательностей вне контекста генома и напрямую оценивать эффект однонуклеотидных замен [14–19]. Для обработки и обобщения таких данных особенно хорошо подходят модели машинного обучения, получившие бурное развитие именно в последние годы. В то же время все еще не существует общепринятых стандартов и рекомендация по получению наиболее оптимальных моделей данного типа для МПРЭ.

Суммируя вышесказанное, безусловно актуальной является разработка и применение новых вычислительных методов и моделей на основе геномных данных и данных параллельных репортерных экспериментов для функциональной аннотации однонуклеотидных вариантов в регуляторных районах генов.

1.2. Научная новизна исследования

В диссертационной работе впервые продемонстрировано, что оценка качества предсказаний для моделей машинного обучения, обученных на омиксных данных, завышена в задаче предсказания эффектов однонуклеотидных замен в регулярных регионах по данным МПРЭ в связи с утечкой информации [33].

В работе впервые в большом масштабе успешно применены методы классического машинного обучения для предсказания аллель-специфичного связывания факторов транскрипции [34].

Разработан новый вычислительный метод на основе глубокого обучения специально оптимизированный для результатов высокопроизводительных МПРЭ [35]. Продемонстрирована возможность адаптации нейросети для рационального дизайна промоторных последовательностей генов с заданным уровнем активности при помощи впервые примененного для данной задачи подхода на основе диффузионных процессов [35].

1.3. Степень научной разработанности темы

С прочтением генома человека и развитием высокопроизводительных омиксных методов предпринималось множество попыток разработать вычислительные биоинформатические инструменты для приоритизации вариантов. Одним из первых способов можно считать картирование локусов количественных признаков (Quantitative Trait Loci, QTL), изучающее связь между частотами аллелей и фенотипом – например, молекулярным, таким как экспрессия гена (expression Quantitative Trait Loci, eQTL) или частотой включения экзона (splicing Quantitative Trait Loci, sQTL) [23,36]. К родственному методу следует отнести определение потенциально значимых геномных вариантов в ходе полногеномных ассоциативных исследований (Genome-Wide Association Studies, GWAS) за счет статистического анализа разниц частот аллелей на основании полногеномного набора вариантов у особей, обладающих и не обладающих каким-либо категориальным признаком [24]. Первая общая проблема данных методов - это неспособность напрямую определять каузальные варианты среди множества кандидатов в области неравновесия по сцеплению: методы на основе статистических ассоциаций указывают на локус, содержащий целый список вариантов, статистически ассоциированных с изучаемым признаком, и любой

вариант в локусе может быть причинным, "каузальным" [18,37]. Вторая проблема: сильная зависимость чувствительности детекции от объема выборки. Третья проблема, связанная со второй: трудность в определении ассоциаций для вариантов, редко встречающихся в популяции, и невозможность оценки эффекта индивидуальных соматических мутаций.

По современным представлениям, до 90% геномных вариантов, ассоциированных с наследственными болезнями и развитием злокачественных опухолей, расположено в некодирующих районах генома [26–31]. В свою очередь среди некодирующих вариантов наибольшая доля приходится на мутации в регуляторных областях, контролирующих транскрипцию – промоторах и энхансерах. В среднем, относительно референсной геномной сборки, индивидуальный геном содержит порядка 500 тысяч вариантов, расположенных в регуляторных регионах [38].

Наиболее надежным способом выявления причинных вариантов, в том числе регуляторных вариантов, влияющих на экспрессию генов, является прямая экспериментальная верификация их эффектов традиционными методами молекулярной биологии [39–41] или одновременное тестирование множества вариантов в массовых параллельных репортерных экспериментах [18,42,43] и скринингах при помощи высокопроизводительных методов, основанных на технологии CRISPR[44–46]. Однако представляется невозможным даже при помощи самых высокопроизводительных подходов перебрать все пространство возможных вариантов и их взаимодействий в контексте различных типов клеток эукариотического организма.

Решением становится использование вычислительных моделей и предсказательных алгоритмов [1,2]. С точки зрения механизма влияния вариантов в энхансерах и промоторах на экспрессию гена наиболее простой молекулярный механизм состоит в изменении аффинности участка связывания фактора транскрипции (ТФ), активатора или репрессора, в зависимости от аллеля. Таким образом, наиболее простые и широко применяемые методы основаны на использовании наборов позиционно-весовых матриц (ПВМ), описывающих характерные ДНК-паттерны в регуляторных регионах, с которыми происходит связывание факторов транскрипции [47,48]. Несмотря на простоту сравнения оценок ПВМ между аллелями, данный подход все еще остается де-факто стандартом для аннотации и приоритизации регуляторных однонуклеотидных вариантов, локализованных в промоторах и энхансерах [48,49]. Из подходов, основанных на классическом машинном обучении, популярность получил подход gkmSVM/deltaSVM(gapped-kmer/delta Support Vector Machine) [50], основанный на методе опорных векторов и показавший хорошее качество предсказаний на различных задачах, в том числе, занявший первое место в нескольких открытых соревнованиях по предсказанию влияния однонуклеотидных вариантов на регуляцию экспрессии генов [50,51]. В этом методе впервые была предложена следующая схема косвенного предсказания эффектов регуляторных мутаций: 1)

модель обучается отличать открытые участки хроматина или участки связывания транскрипционных факторов от случайных геномных последовательностей; 2) разница между предсказаниями полученной модели для доступности хроматина в зависимости от аллеля используется как оценка эффекта варианта с точки зрения его влияния на экспрессию.

Следующим шагом стало использование искусственных нейронных сетей для моделирования функционально значимых регуляторных участков генома [52]. Подход, предложенный в gkmSVM/deltaSVM, был адаптирован для искусственных нейронных сетей и одновременно расширен – модели стали обучать по нуклеотидной последовательности предсказывать тысячи эпигенетических разметок генома, полученных по результатам омиксных экспериментов [32,53–57]. Одновременно с увеличением числа сигналов предсказываемых моделью, начали предприниматься попытки увеличить размер контекста последовательности ДНК (“геномного окна”), который может использовать нейронная сеть для предсказания эпигенетического сигнала в данной позиции [32,53,55,56].

Нейросетевые модели демонстрируют хорошую согласованность между предсказаниями и данными насыщающего мутагенеза в промоторах, и успешно определяют некоторую часть причинных eQTL [32,53–57]. Однако, накапливаются многочисленные свидетельства в пользу того, что полногеномные нейросетевые модели плохо учитывают дальние взаимодействия и индивидуальные различия в геномах и плохо предсказывают паттерны экспрессии генов, специфичные для конкретных типов клеток [9–13]. Обучение на имеющихся персональных геномных последовательностях людей не исправляет ситуацию, лишь незначительно улучшая качество предсказаний моделей в пределах популяций, откуда происходят персональные геномы и не превосходя качество информированных об этих вариантах линейных моделей, сохраняя при этом их недостатки [58–60].

Предпринимаются попытки улучшить качество нейросетевых моделей за счет предобучения на геномных последовательностях различных организмов [61,62]. Однако этот подход не приводит к улучшению качества предсказания, иногда приводя к противоположному эффекту [63,64]. Дообучение на результатах секвенирования транскриптомов отдельных клеток и мультимодальных данных из отдельных клеток также значимо не улучшает качество моделей на упомянутых ранее задачах [65–67].

В связи с этим высказывается мнение, что полногеномных данных в принципе недостаточно для расшифровки регуляторного кода и необходимо прибегнуть к обучению моделей на результатах МПРЭ [9–13]. Использующиеся для этого нейросетевые архитектуры до сих пор представляли собой простейшие сверточные сети или неадаптированные архитектуры на основе трансформеров [68–72]. При этом, к сожалению, современные достижения в области дизайна архитектур нейронных сетей и их обучения практически не используются [73–75].

Помимо задачи оценки эффекта вариантов, широкое распространение получает применение нейронных сетей для задач генерации новых объектов в самых различных областях [76–81], включая задачи генетики и молекулярной биологии [70,82–96]. В частности, модели на основе диффузионных процессов [78,97] являются наиболее перспективным направлением развития данной области, однако вопрос их применения для получения последовательностей с заданными свойствами, в частности, с использованием для обучения данных МПРЭ, исследован достаточно слабо, несмотря на его практическую важность для задач синтетической биологии и генной терапии. Потенциально, продвинутые генеративные модели могли бы ускорить прогресс в расшифровке регуляторного кода и построения лучших предсказательных моделей за счет появления возможности проведения крупномасштабных вычислительных экспериментов и применения подобных моделей в активном обучении.

1.4. Цель и задачи исследования

Цель работы: создание новых вычислительных методов для предсказания эффектов однонуклеотидных замен в регуляторных районах генома человека на основе данных современных высокопроизводительных омиксных методов.

Задачи работы

1. Оценить эффективность обучения и тестирования вычислительных моделей для предсказания регуляторных эффектов однонуклеотидных вариантов на основе данных параллельных репортерных экспериментов с мутагенезом насыщающей ПЦР.
2. Разработать вычислительный метод для предсказания участков аллель-специфичного связывания факторов транскрипции, определенных на основе результатов экспериментов по иммунопреципитации хроматина с последующим глубоким секвенированием.
3. Разработать нейросетевой подход для предсказания активности промоторов и изменений их активности в зависимости от однонуклеотидных вариантов по данным массовых параллельных репортерных экспериментов. Адаптировать построенную нейросетевую модель для генерации промоторных последовательностей с заданным уровнем активности.

1.5. Объект и предмет исследования

Объектом исследования являются регуляторные регионы геномов эукариот, контролирующие транскрипцию генов.

Предметом исследования являются нуклеотидные последовательности регуляторных районов, замены в них, и биологическая активность районов, систематически измеренная с помощью современных высокопроизводительных методов молекулярной биологии.

Работа опирается на результатах применения массовых параллельных репортерных экспериментов, выполненных в клетках дрожжей и клеточных линиях человека. Такие крупномасштабные данные позволяют использовать новые архитектуры нейронных сетей для моделирования структуры и активности регуляторных районов и оценки влияния мутаций на экспрессию генов.

1.6. Методология и теоретические основы исследования

Теоретические основы исследования опираются на классические работы в области вычислительного анализа регуляторных последовательностей генома и систематический анализ

современных литературных источников по теме, что детально отражено в обзоре литературы. Методология исследования построена по современным принципам, изложенным в ключевых обзорах по проблемам использования методов машинного обучения для задач геномики. В целом, в исследовании использовались различные методы анализа данных, биоинформатики и вычислительной биологии, отвечающие принятым мировым стандартам. В работе уделено особое внимание проблеме переобучения и утечки данных, используются методы кросс-валидации и независимые тестовые выборки. Точность моделей проверена на результатах независимых экспериментов.

1.7. Теоретическая и практическая значимость работы

В работе изучается проблема утечки информации при обучении геномных моделей для предсказания эффектов регуляторных однонуклеотидных вариантов, а представленный нейросетевой метод опережает наилучшие из существующих в области решений в широком спектре задач регуляторной геномики. Удалось выявить ключевые элементы нейросетевой архитектуры, критически важные для успешного применения модели, и продемонстрировать биологическую осмысленность выучиваемого моделью сигнала. Наконец, в работе была разработана методика дизайна регуляторных последовательностей с заданной активностью, что имеет ценность для решения задач синтетической биологии, включая оптимизацию регуляторных районов генов для генной терапии. Таким образом, полученные в работе результаты имеют высокий уровень теоретической и научно-практической значимости.

Теоретическая значимость исследования обусловлена следующим:

1) продемонстрированы сложности прямого использования данных насыщающего мутагенеза для обучения моделей, предсказывающих эффект мутации в регуляторных районах генома;
2) создан новый нейросетевой метод на основе глубокого обучения для предсказания активности регуляторных районов генома, превосходящей имеющиеся аналоги, и предложены методы его адаптации к новым задачам;
3) предложен новый метод генерации регуляторных последовательностей с заданными свойствами.

Практическая значимость работы заключается в следующем:

1) Разработанные в работе методы и веса обученных моделей размещены в открытом доступе и могут быть использованы сторонними исследователями (https://github.com/autosome-ru/LegNet, https://github.com/autosome-ru/human_legnet), в том числе, для фунциональной аннотации некодирующих однонуклеотидных вариантов;

2) Разработанный пакет для подбора типов моделей для работы с короткими нуклеотидными последовательностями также предоставлен в открытый доступ (https://github.com/de-Boer-Lab/random-promoter-dream-challenge-2022) и может быть использован для дальнейших разработок и улучшения качества решения в задачах предсказания активности регуляторных регионов;

3) Предложенный метод генерации последовательностей с заданной экспрессией может быть использован для рационального дизайна генноинженерных конструкций в задачах генной терапии.

1.8. Положения, выносимые на защиту

- Показано, что данные параллельных репортерных экспериментов на основе мутагенеза насыщающей ПЦР в значительной степени отражают локальные зависимости в геномных сигналах. Это приводит к неоправданному завышению качества модели в случае использования простых традиционных разбиений доступной для обучения и тестирования моделей выборки данных. В некоторых случаях, например в соответствующей задаче соревнования CAGI5 (Critical Assessment of Genome Interpretation 5, 2018 год), это приводит к невозможности использовать традиционные подходы для оценки реальной точности моделей.
- На основе случайного леса с использованием геномных признаков разработана модель, достигающая приемлемого качества в задаче предсказания аллель-специфичного связывания факторов транскрипции в отдельных хорошо изученных типах клеток.
- Разработана новая сверточная нейронная сеть LegNet для предсказания активности регуляторных последовательностей и их влияния на экспрессию репортерных генов. Модель показала наилучшее качество среди всех моделей в независимом исследовании на промоторах дрожжей. Предложенный подход хорошо переносится на другие типы данных, в том числе, хорошо показывает себя на результатах МПРЭ, полученных в клетках человека, и превосходит по точности предсказаний имеющиеся альтернативы.
- Архитектура LegNet при помощи подхода “холодная диффузия” успешно адаптирована для генерации регуляторных последовательностей с заданной экспрессией.

1.9. Личный вклад автора

В работе (D. D. Penzar et al. 2019) лично автором проведен детальный анализ данных МПРЭ на основе насыщающего мутагенеза отдельных промоторов и обучение вычислительных моделей. В работе (Abramov et al. 2021) под руководством автора диссертации было проведено обучение и тестирование классических моделей на основе деревьев решений для предсказания аллель-специфичного связывания факторов транскрипции. В работе (D. Penzar et al. 2023)

непосредственно автором выполнен дизайн архитектуры нейронной сети, подбор методики ее обучения и всестороннее тестирование модели, а также проведено абляционное исследование и исследование пользы ансамблирования различных моделей на итоговое качество. В работе (Rafi et al. 2024) автором выполнен дизайн архитектуры наилучшего решения и разработана архитектура пакета для комбинации архитектур и подбора оптимальной модели. В работе (Agarwal et al. 2025) автором выполнена адаптация архитектуры нейронной сети к новым данным, подбор методики ее обучения, проведено абляционное исследование, изучена зависимость качества предсказания сети в зависимости от размера обучающего набора и протестирована способность нейросети предсказывать события аллель-специфичного связывания.

1.10. Структура и объем диссертации

Диссертационная работа состоит из титульного листа, оглавления, списка сокращений и условных обозначений, введения, обзора литературы, материалов и методов, результатов, заключения, выводов, списка литературы, списка публикация по теме диссертации и приложений. Работа изложена на 166 страницах, иллюстрирована 55 рисунками, 6 таблицами и 1 приложением. Список литературы состоит из 334 источников.

1.11. Апробация результатов исследования

Результаты работы были представлены на 6 международных конференциях:

1. 8 ноября 2022, RSG-DREAM-2022 (RECOMB/ISCB Conference on Regulatory & Systems Genomics with DREAM Challenges), Лас-Вегас, США, онлайн, приглашенный устный доклад, “NogiNet: repurposing EfficientNetV2 for accurate promoter sequence-to-expression modeling”;
2. 23-24 ноября 2022, Life of Genomes, Казань, Россия, стендовый доклад, “Использование современных сверточных архитектур нейронных сетей для предсказания экспрессии гена по последовательности промотора”;
3. 30-31 мая 2023, AIPPA-2023 (Artificial Intelligence- Possibilities for Practical Applications 2023), Алматы, Казахстан, приглашенный устный доклад, “Использование современных сверточных архитектур нейронных сетей для предсказания экспрессии гена по последовательности промотора”;
4. 3-6 августа 2023, MCCMB-2023 (Moscow Conference on Computational Molecular Biology 2023), устный доклад, “LegNet: a novel approach to modeling regulatory sequences with deep convolutional networks”;
5. 5-10 августа 2024, BGRS-2024 (устный доклад, “Machine learning for rational design and reliable prediction of activity of gene regulatory regions”;

6. 22-25 октября 2024, APBJC-2024 (Asia & Pacific Bioinformatics Joint Conference 2024), онлайн, устный доклад, Окинава, Япония, “LegNet allows for state-of-the-art prediction of activity and rational design of eukaryotic regulatory regions”.

1.12. Публикации по теме исследования

По результатам исследования опубликовано 6 печатных работ, в том числе 6 статей в рецензируемых научных журналах, индексируемых в WoS и Scopus. На странице 147 представлен список публикаций по теме диссертации.

2. Обзор литературы

2.1. Однонуклеотидные варианты в геноме человека и их связь с патологиями

Одной из центральных проблем современной генетики является предсказание фенотипических особенностей организма на основе его индивидуального генотипа [1], и ее решение является необходимым требованием для развития персонализированной медицины и генной терапии.

Большая часть различий в генотипах индивидуумов в человеческой популяции приходится на однонуклеотидные варианты [4]. По расположению в геноме важные варианты делятся на кодирующие – локализованные в участках, кодирующих белки и влияющие на их структуру и функции напрямую, и на некодирующие – находящиеся в других участках генома. В частности, среди последних, особый интерес представляют варианты, влияющие на экспрессию генов на разных уровнях регуляции (регуляция транскрипции, сплайсинга и прочее) (**рис 1. A**)

Влияние кодирующих мутаций хорошо изучено и предсказывается с удовлетворительным качеством современными методами [5,6]. Их недостатки и дальнейшее улучшение выходит за рамки данной работы.

Однако от 80 до 90% геномных вариантов, ассоциированных с наследственными болезнями и опухолями, находится в некодирующих районах генома [26–31]. В свою очередь среди некодирующих вариантов, наибольшая доля клинически значимых приходится на мутации в регуляторных областях, контролирующих транскрипцию – промоторы и энхансеры (**рис 1. B**). В среднем, индивидуальный геном содержит порядка 500 тысяч отличающихся от референсных вариантов, расположенных в регуляторных регионах [38]. При этом даже однонуклеотидные варианты могут приводить к возникновению новых энхансеров, как это происходило, по-видимому, в эволюции нервной системы человека [99].

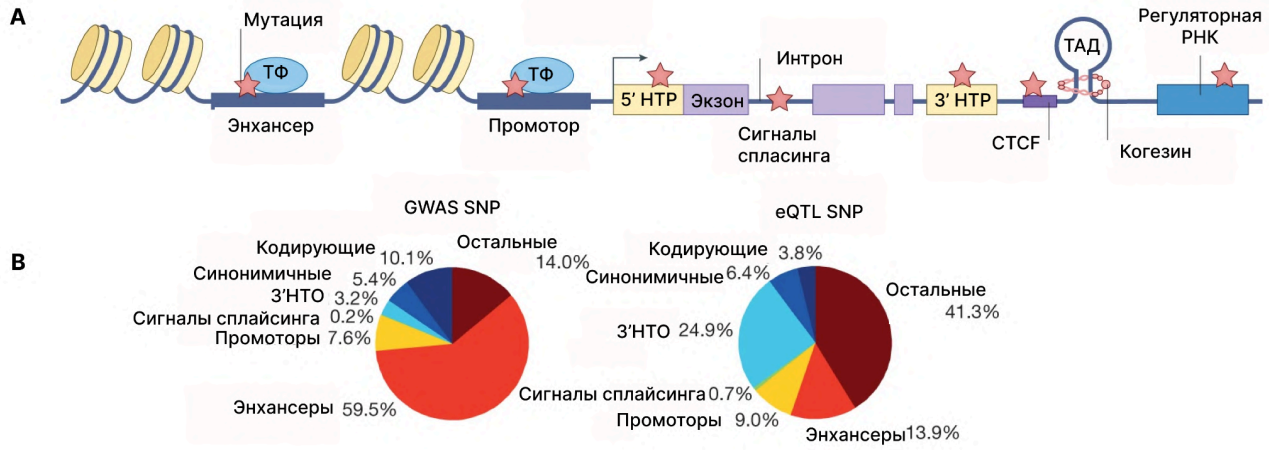


Рисунок 1. А. Примеры того, как вариант в некодирующих участках может влиять на экспрессию гена. Мутации в энхансерах и промоторах могут нарушать связывания с фактором транскрипции, оказывать влияние на трансляцию и стабильность мРНК за счет мутаций в 5'НТР и 3'НТР, вносить изменения в доменную организацию хроматина через нарушение связывание архитектурного белка CTCF или вносить изменения в последовательности регуляторных РНК. Адаптировано из [100] **B.** Среди аннотированных мутаций, которые играют значимую роль в развитии аутоиммунных заболеваний, преобладающую часть составляют мутации в некодирующих областях. Справа, доля мутаций в промоторах и энхансерах, также преобладает (GWAS) или не уступает (eQTL) долям остальных регионов. Адаптировано из [28].

2.2. Примеры клинически значимых регуляторных вариантов

Регуляторные мутации, ассоциированные с пониженным или повышенным риском развития заболевания, могут располагаться на самых разных расстояниях от старта транскрипции.

Показано, что регуляторные варианты, локализованные в консервативной части промотора гена бета-глобина, а именно в -26..-31 позициях мотива связывания TFIID (TATA-бокс), уменьшают экспрессию гена, приводя к развитию бета-талассемии [101].

Вариант rs10993994 в промоторе гена бета-микросеминопротеина (MSMB), расположенный на расстоянии 90 п.о от старт-кодона, нарушает связывание транскрипционного фактора CREB, приводя к повышенному риску развития рака простаты [102].

Вариант rs17079281 в промоторе гена дискоидина (DCBLD1), расположенный на расстоянии 160 п.о от старт-кодона, разрушает сайт связывания ТФ YY1, что приводит к подавлению экспрессии онкогена, снижая риск развития рака легких [103] (**рис. 2. А**).

Первые 15кб первого интрона гена рецептора интерлейкина 2 (IL2RA) содержат 5 энхансеров, регуляторные варианты в которых связаны с разрушением сайтов связывания различных транскрипционных факторов и ассоциированы с повышенным риском развития аутоиммунных заболеваний [104].

Важно понимать, что более 90% ассоциированных с заболеванием вариантов в человеческом геноме локализованы в участках далеких от гена, на который они оказывают влияние [105]. При этом на эти варианты приходится 60-75% объясненной дисперсии в экспрессии человеческих генов [23].

Замена rs12740374 в регуляторном регионе, расположенном на расстоянии 40кб от гена сортилина (SORT1), создает сайт связывания ТФ C/EBP и приводит к оверэкспрессии данного гена в печени, что, в свою очередь уменьшает уровень липопротеинов низкой плотности и увеличивает риск инфаркта миокарда [23].

Регуляторные варианты в энхансере гена фактора обмена гуанидина (RasGRP1), находящемся на расстоянии 60кб от стартового кодона, уменьшают экспрессию этого гена за счет нарушения связывания ТФ RUNX1 и CBFB, что приводит к увеличенному риску развития аутоиммунных заболеваний [106].

Регуляторный вариант rs356168 в энхансере гена альфа-синуклеина (SNCA), расположенном на расстоянии 80кб от стартового кодона, увеличивает экспрессию гена за счет создания сайта связывания ТФ EMX2, снижая риск развития болезни Паркинсона [107] (**рис. 2. В**).

Аналогично, мутации в энхансере гена IRX3, расположенного на расстоянии 500 кб п.о от регулируемого гена, приводят к повышенному риску развития ожирения [23]. Мутации же в энхансере, находящемся на расстоянии 1.5мб п. о регулируемого им гена SOX9 может приводить к развитию синдром Робена [108].

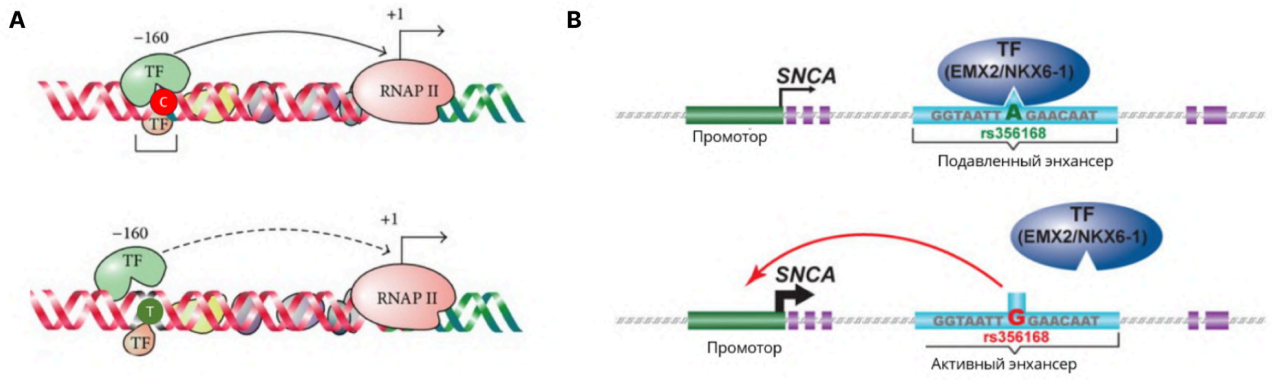


Рисунок 2. А. Регуляторный однонуклеотидный вариант в промоторе гена DCBLD1 влияет на связывание транскрипционного фактора YY1, приводя к подавлению экспрессии онкогена и снижая риск рака легких. Адаптировано из [103]. TF – транскрипционный фактор. **В.** Защитный вариант аллеля энхансера альфа-синуклеина приводит к пониженной вероятности развития болезни Паркинсона. Мутация в нём нарушает связывание ТФ EMX2, увеличивая риск развития болезни. Адаптировано из [107]

2.3. Детекция потенциально каузальных индивидуальных вариантов

В соответствии с нейтральной теорией эволюции Кимуры [109] большая часть нуклеотидных мутаций не влияет на фенотип значимо, в связи с чем встает вопрос о выявлении среди них тех, которые действительно имеют какой-то эффект на организм. Для выявления потенциально важных вариантов используются такие наблюдательные исследования как картирование локусов количественных признаков (QTL) [23,36] и полногеномные ассоциативные исследования GWAS [24,26].

2.3.1. QTL

Первым методом для определения потенциально каузальных вариантов является QTL. QTL изучает связь между генотипом и количественным фенотипом, включая молекулярные фенотипы, измеримые в омиксных экспериментах, но не подразумевает автоматически разрешения с точностью до отдельных вариантов, и в классических подходах фокусируется на маркерных локусах. Однако же сегодня наиболее распространенным видом данного анализа являются экспрессионные QTL (eQTL), в которых устанавливается связь между экспрессией мРНК, измеренной с помощью высокопроизводительных омиксных методов, и индивидуальными вариантами.

В случае работы с человеческими геномами необходимость тестирования большого числа гипотез (на каждый 5тыс нуклеотидов индивидуального генома приходится 1 отличие от референсного [38]) одновременно приводит к низкой мощности данной методики.

В среднем слабое влияние далеких регуляторных вариантов приводит к тому, что большая часть находимых вариантов расположена вблизи генов. Кроме того, часть исследований для увеличения мощности заранее исключает из рассмотрения удаленные от изучаемых генов и локусов варианты, что еще больше уменьшает число детектируемых дистальных QTL. В то же время за примерно 60-75% вариабельности в экспрессии отвечают именно далекие варианты. Детектируемые варианты могут не быть причинными [23] – на практике, для каждого молекулярного признака детектируются десятки и сотни вариантов, лишь несколько из которых являются каузальными [18]. В случае eQTL сложно детектировать варианты с очень большим эффектом на экспрессию, так они чаще отбраковываются под действием естественного отбора [18,25].

2.3.2. GWAS

Полногеномные ассоциативные исследования направлены на изучение связи между генотипом и фенотипом за счет сравнения аллельных частот вариантов между группами людей,

обладающих определенным фенотипом (например, наличие заболевания) и не обладающими им, или же группами людей, обладающих разной степенью проявления признака в случае количественных характеристик [24]. В случае использования микрочипов для генотипирования дополнительную трудность представляет то, что результатом являтся не список всех аллелей, которые могут отличаться от референсных, а лишь части из них, остальные же примерно восстанавливаются при помощи процедуры импутации. Качество импутации сильно зависит от изученности популяции – для европейских популяций она выше, чем для многих других, включая крайне разнородную российскую [110].

Из-за неравновесного сцепления генов [111] результатом GWAS являются не каузальные варианты, а широкий список вариантов, каждый из которых статистически значимо связан с исследуемым признаком. Эти варианты группируются в регионе, где расположен каузальный вариант. В исследовании [37] размер таких групп в среднем составлял 158 вариантов. Чаще всего каузальным признается наиболее значимый из этих вариантов, однако в реальности следует считать такие варианты в первую очередь маркерами. В частности, наличие нескольких каузальных вариантов в одном регионе может смещать наиболее значимую связь с собственно каузальных вариантов на их соседей. К этому же эффекту могут приводить ограничения процедуры импутации. Неучтенные в анализе ковариаты (возраст, пол, родственные связи и прочее) также могут приводить к появлению значимых вариантов, не являющимися биологически обоснованными [24].

Так как в ходе подобных исследований одновременно тестируется большое количество статистических гипотез, исследователи прибегают к поправке на множественное тестирование. Несмотря на ее необходимость для снижения количества ложноположительных результатов, ее применение снижает мощность процедуры и устанавливает нижнюю границу для эффекта и частоты казуального варианта, который может быть найден. Таким образом, детекция каузальных, но редких вариантов или вариантов с ограниченным эффектом при помощи GWAS затруднена [24].

Оба разобранных подхода не позволяют, среди прочего, предсказывать эффекты редких вариантов и новых мутаций, не наблюдающихся в изучаемой популяции, что может быть необходимо, например, для задач геномного редактирования или аннотации новых мутаций, возникающих в гипермутирующих раковых клетках.

2.4. Приоритизация индивидуальных вариантов с помощью функциональной аннотации

Размер списка потенциальных вариантов, ассоциированных с данным признаком, полученный из GWAS и QTL может быть уменьшен при помощи различных статистических

процедур [112,113]. В 10-20% случаев эти процедуры позволяют выбрать единственный вариант, а иначе – получить меньший по размеру достоверный интервал, в котором заданной апостериорной вероятностью находится каузальный вариант [20,113]. Достоверные интервалы легче поддаются биологической интерпретации, но только информации о генотипах не хватает для точного определения каузального варианта [20,114].

Наиболее надежным способом приоритизации причинных вариантов среди кандидатных является прямая экспериментальная проверка при помощи точечных экспериментов [39–41], массовых параллельных экспериментов с репортерами [18,42,43] и массовых скринингов при помощи CRISPR [44–46]. Стоит, однако, отметить что даже результаты экспериментальной проверки могут быть противоречивыми. Например, эффект варианта rs34500389 в промоторе бета-гемоглобина оценен как вариант с отсутствующим эффектом в массовом параллельном эксперименте [115], в то время как в последующей точечной проверке в работе [116] данный вывод ставится под сомнение.

До или параллельно экспериментальной проверки для приоритизации причинных вариантов можно использовать спектр различных биоинформатических методов. Основная идея данных методов заключается в выборе вариантов, тем или иным образом согласующихся с экспериментальными данными функциональной геномики и вычислительными моделями [24].
Можно проверить то, попадет ли вариант в область активного хроматина, район старта транскрипции, известную регуляторную или консервативную области генома, установленный экспериментально сайт связывания транскрипционного фактора или сайт сплайсинга [117,118], ассоциирован ли вариант с каким-либо аллель-специфичным событием, контактирует ли согласно структурной организации хроматина энхансер, где расположен вариант, с промотором гена, на который оказывается воздействие [24].

Взамен экспериментальной информации о геномных разметках можно использовать предсказательные модели различной сложности. Для вариантов, влияющих на регуляцию транскрипции, традиционным подходом является использование мотивов связывания транскрипционных факторов (ТФ) и оценка на их основе изменения силы связывания ТФ с участком, где расположен вариант [48]. Такой фокус на ТФ связан с их исключительной важностью в регуляторных процессах и тем, что, согласно современному консенсусу, сайты связывания ТФ являются атомарной единицей регуляторной грамматики [71].

Помимо традиционных методов для функциональной аннотации вариантов в последнее время широкое применяются методы машинного обучения – построение моделей, на основе нуклеотидной последовательности предсказывающих активность хроматина, наличие гистоновых меток и другие эпигенетические треки [118]. Потенциально такой подход позволяет учитывать более тонкие элементы регуляторной грамматики, такие как расстояние между мотивами ТФ их

ориентацию. В теории, методы машинного обучения способны работать значимо лучше классических за счет возможности учитывать огромные массивы данных в своих предсказаниях. В связи с этим особый интерес представляет развитие данных методов и преодоление их текущих недостатков, например, разобранных в разделе “Проблемы современных предсказательных моделей”.

Далее мы более подробно остановимся на упомянутых вариантах приоритизации.

2.5. Функциональная аннотация вариантов на основе полногеномных омиксных данных

На данный момент на основе полногеномного секвенирования создан целый спектр методов, которые позволяют прямо или косвенно оценивать активность регуляторных участков генома и, в частности, изучать связывание с ними ТФ (**рис. 3**) [119,120].

Остановимся на самых распространённых из них – методах оценки доступности хроматина, детекции связывания факторов транскрипции, наличия тех или иных модификаций гистонов ("гистоновых меток") и прямого измерения транскрипционной активности генов.

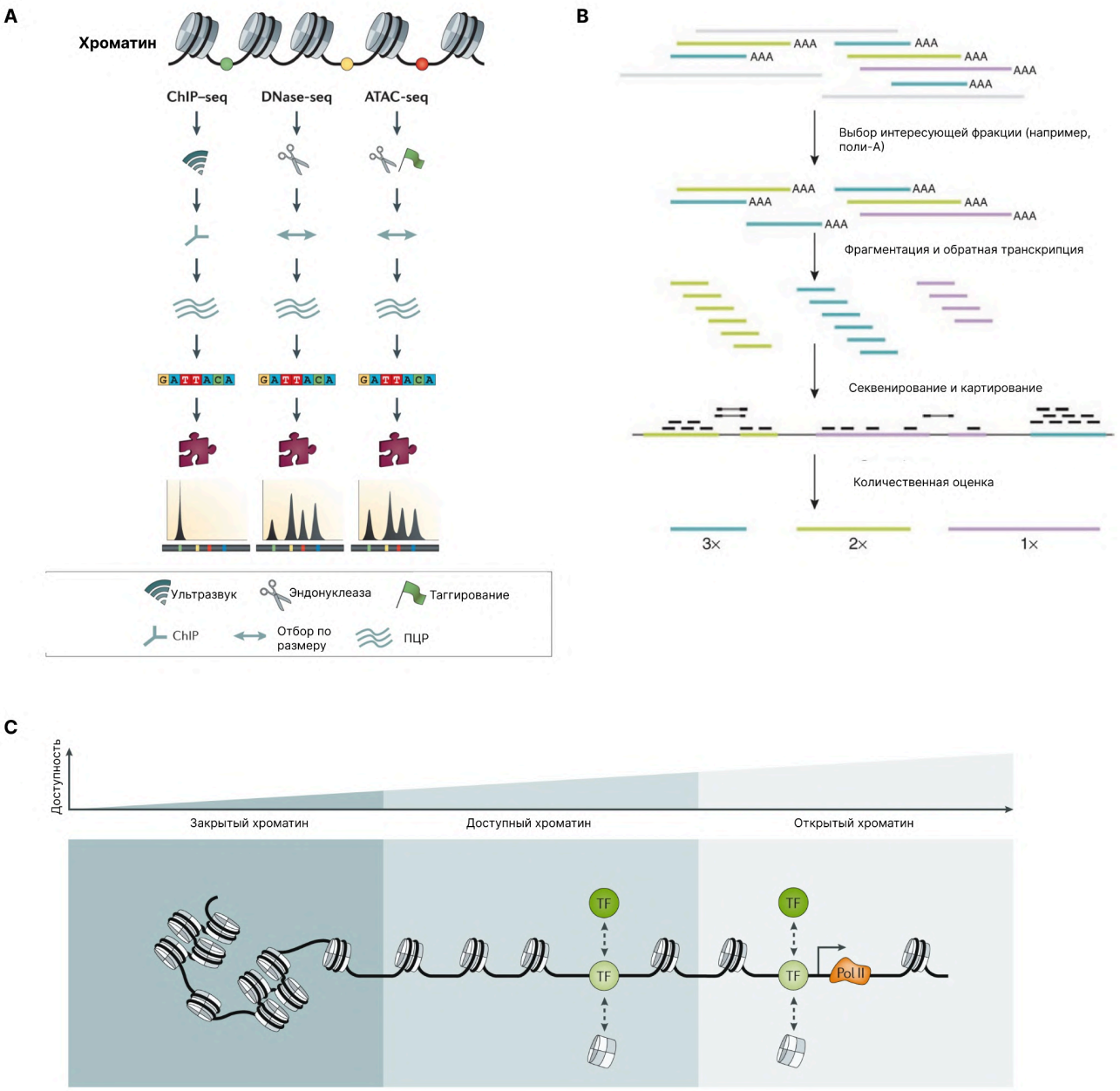


Рисунок 3.. Методы получения информации об активности участков генома. **A.** ChIP-Seq – получение информации о связывании ТФ и гистоновых меток. DNAse-seq – метод определения доступности хроматина с использованием неспецифичной ДНКазы 1. ATAC-seq – метод определения доступности хроматина для транспозиций на основе активности гиперактивной мутантной формы транспозазы Tn5. ChIP – коиммунопреципитация. Адаптировано из [120] **B.** RNA-Seq – измерение экспрессии участков генома за счет секвенирования РНК. РНК подвергается обратной транскрипции, а затем секвенируется. С помощью дополнительной фильтрации можно выбирать, какие именно РНК (например, какой длины) или какие участки РНК (например, 5' или 3' концы), секвенировать. Адаптировано из [119]. **C.** В отличие от закрытого хроматина, доступный хроматин позволяет транскрипционным фактором связаться с нужным сайтом, инициировать ремоделирование хроматина и установить открытый хроматин. Рисунок адаптирован из работы [121]. TF (transcription factor) - фактор транскрипции, PolII - РНК-полимераза II.

2.5.1. Доступность хроматина

Доступность хроматина — степень, с которой макромолекулы в ядре способны физически контактировать с комплексом хроматина и ДНК. Она определяется топологической организацией нуклеосом и других хроматин-связывающих факторов, затрудняющих доступ к хроматину [121].

Принято разделять два состояния хроматина: гетерохроматин – закрытый хроматин – в котором плотность нуклеосом высока, и эухроматин – открытый хроматин – где их плотность сравнительно низка. В открытом хроматине обычно располагаются активные в данной клетке участки транскрипции и регуляторные области (энхансеры, инсуляторы, промоторы). По тому, насколько часто гетерохроматин пребывает в закрытом состоянии, его делят на конститутивный, находящийся в конденсированной форме почти всегда (центромеры, теломеры, области тандемных повторов), и факультативный, который может быть конденсирован только в данном клеточном типе или на одной из гомологичных хромосом [105,122]. При этом дихотомическое разделение на гетерохроматин и эухроматин в целом искусственно. Существует широкий спектр доступности хроматина, начиная с полностью недоступного (**закрытого**) до **доступного** и полностью **открытого** хроматина [121] (**рис. 1. C**), что следует учитывать при биоинформатическом анализе и, в частности, при построении моделей машинного обучения [123].

Действие многих транскрипционных факторов проявляется в изменении доступности хроматина в области регуляторной области, с которой они связались [124,125]. Верно и обратное – за исключением особых ТФ, которые связываются в областях гетерохроматина[126], большая часть транскрипционных факторов, включая наиболее изученные в рамках крупномасштабного проекта ENCODE, связываются преимущественно с открытым хроматином – на него приходится порядка 90% регионов, с которыми взаимодействуют эти факторы транскрипции – [121,122] несмотря на то, что в клетках человека для каждого клеточного типа области доступного хроматина составляют лишь небольшой процент от всего генома [121].

Наиболее распространенными методами детекции доступности хроматина являются DNase-Seq и ATAC-Seq [127]. Оба метода основаны на фрагментации генома по открытым участкам хроматина при помощи нуклеазы и транспозазы соответственно и секвенировании прилегающих к разрывам или к местам транспозиций участков [121]. Оба метода имеют систематические ошибки, связанные с предпочтительными регионами, с которыми взаимодействуют используемые в них ферменты [105].

2.5.2. Участки связывания факторов транскрипции

Во многих случаях активность регуляторной последовательности преимущественно определяется связыванием с ТФ и силой этого связывания – аффинностью [128]. Таким образом,

мутация может изменить активность регуляторной последовательности через ухудшение связывания ТФ или через образование нового сайта связывания [29,41,103,107,129–132] (рис U).

Эксперименты по определению доступности хроматина позволяют получить лишь информацию об общей открытости регионов, а не составу ТФ которые обеспечивают активность данного региона.

Прямую информацию о связывании с участками генома конкретных транскрипционных факторов позволяет получить метод ChIP-Seq (**Ch**romatin **I**mmunoprecipitation followed by deep **S**equencing) [133] и его модификации [105].

2.5.3. Гистоновые метки

Различные модификации гистоновых белков – гистоновые метки – также регулируют доступность хроматина. Хроматин разной степени доступности в разных участках генома характеризуется различными модификациями. Например, триметилирование Lys4 гистона H3 (H3K4me3) характерно для промоторов активно экспрессирующихся генов. Для полностью подавленных регионов (к примеру, содержащих ретротранспозоны и повторы) характерна метка H3K9me3, которая привлекает комплексы, метилирующие ДНК, тем самым полностью подавляя транскрипцию данного участка) [134].

Информацию о профиле данных меток получают при помощи ChIP-Seq экспериментов.

2.5.4. Измерение транскрипционной активности генов

При помощи метода секвенирования RNA-Seq можно измерять уровень экспрессии генов [135] (рис W2). Метод CAGE (Cap analysis of gene expression) за счет отбора кэпированных РНК в дополнение к количественной оценке [136] экспрессии позволяет точно установить локализацию участков инициации транскрипции. В частности, таким образом была собрана база промоторных последовательностей и их активности в человеке и других млекопитающих в ходе проекта FANTOM5 [137]. Эти данные позволили определить клеточно-специфичные сайты промоторов с разрешением, близким к однонуклеотидному [137] и широко используются для создания моделей человеческих промоторов [138].

2.5.5. Аллель-специфичная регуляция экспрессии генов

Омиксные данные позволяют определять наличие гетерозиготных аллелей в геноме человека и, за счет разницы в числе прочтений для вариантов аллеля, находить аллель-специфичные события (АС), такие как аллель-специфичное связывания (АСС) ТФ – из экспериментов ChIP-Seq, аллель-специфичную доступность (АСД) хроматина – из данных DNase-Seq и ATAC-Seq (**рис. 4 А**), и аллель-специфичную экспрессию (АСЭ) – из RNA-Seq [34,139].

Эти данные затем могут быть использованы как для обучения моделей машинного обучения, так и для их валидации. Прямое обучение на данных по АС затруднено, поскольку в существующих база данных содержится недостаточно АС-вариантов и их достоверность ограничена в силу технических причин – например, овердисперсии при оценке статистической значимости аллельного дисбаланса. Тем не менее, данные по локализации АС являются полезными для валидации методов машинного обучения [34]. Также возможно проведение экспериментов, направленных на более точное определение аллель-специфичных событий за счет фазирования генома [69] (**рис. 4 В)**. В числе прочего подобный метод позволяет сопоставлять между собой разные АС, позволяя изучать связь между ними.

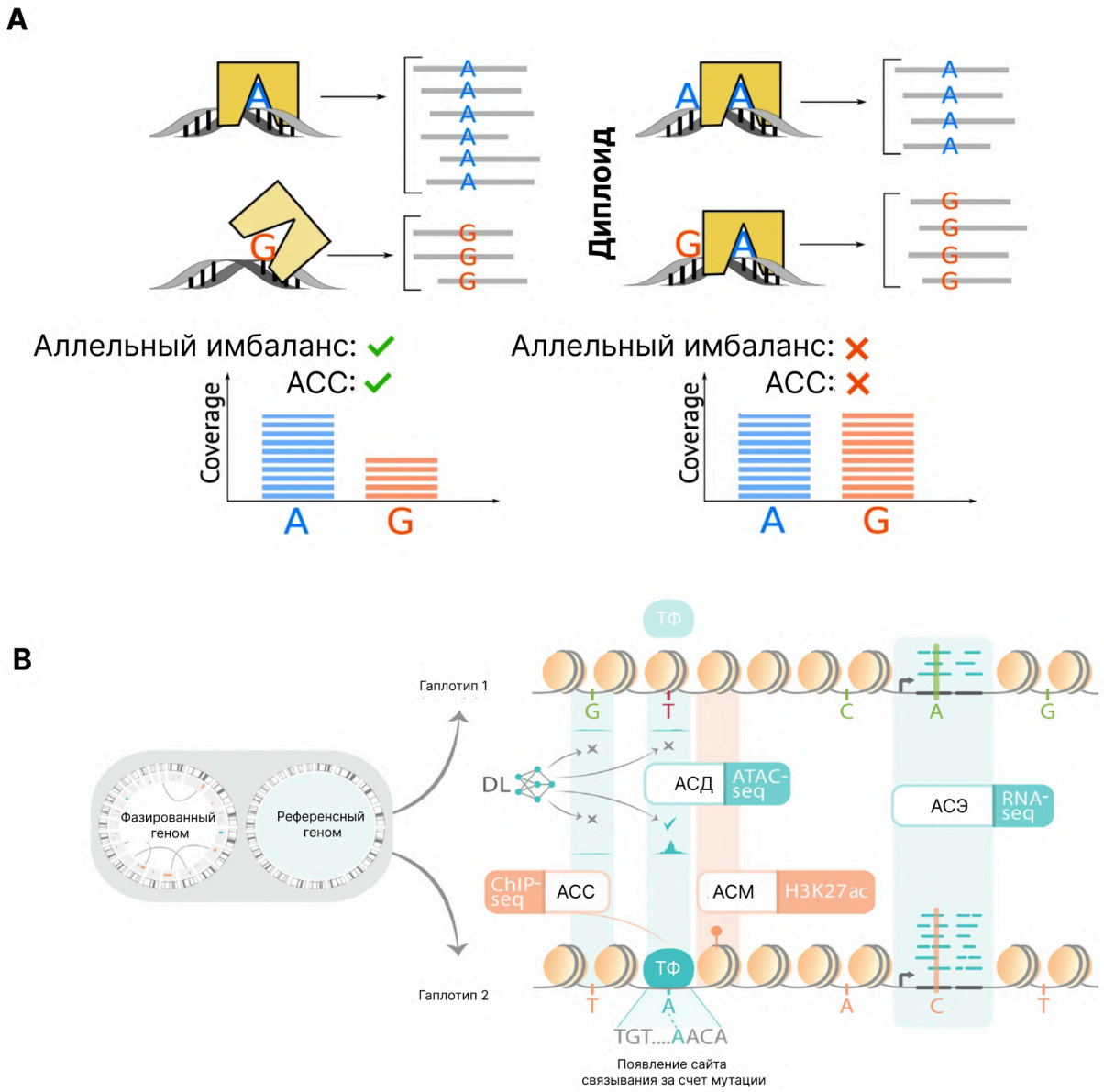


Рисунок 4. А. Схема получения информации об аллель-специфичном связывании из данных ChIP-Seq. За счет того, что в клетке каждая аутосома имеет гомологичную, которая отличается в некоторых локусах. В этих локусах можно может наблюдаться разница в числе ридов с разными аллелями. Адаптировано из [34] **В.** Улучшенная схема идентификации событий аллель-специфичной регуляции генов за счет использования фазированных геномов и непротиворечивого разделения наблюдений на основе различных омиксных протоколов по гаплотипам. Позволяет сопоставлять между собой разные АС [69].

2.5.6. Омиксные эксперименты для профилирования единичных клеток

Доступность хроматина может сильно меняться в ходе развития органов и процессов клеточной дифференцировки. В единичных клетках или однородных группах клеток одних и тех же тканей или органов разные геномные регионы имеют различную доступность. Обычные (bulk) DNase-Seq и ATAC-Seq же выдают усредненную картину, которая может плохо отражать реальное положение дел (**рис. 5**). Для учета особенностей одиночных клеток широкое распространение

приобретают single-cell методы, которые позволяют получать отдельную аннотацию доступных регионов для каждой клетки или клеток одного подтипа. Наиболее популярным на сегодняшний день среди таких методов является scATAC.

Аналогичные проблемы существуют и для других сигналов, включая RNA-Seq. Потому и в этом случае разработаны методы, позволяющие оценивать экспрессию генов на уровне отдельных клеток (исторически для single RNA-Seq были разработаны раньше) [140].

Кроме того, интенсивно разрабатываются методы, позволяющие получать мультимодальные измерения для каждой клетки, например, одновременно оценивать и доступность хроматина, и уровень экспрессии [141,142].

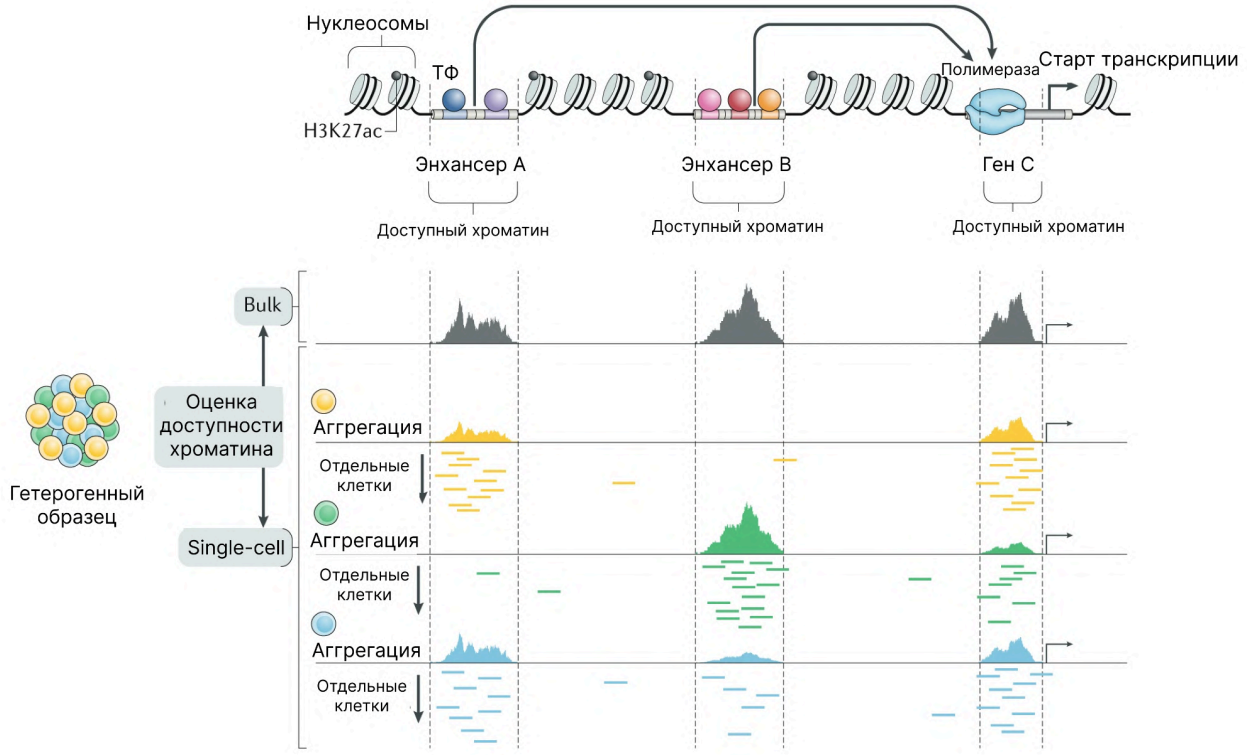


Рисунок 5. Классические методы оценки доступности хроматина усредняют сигнал по большому количеству различных групп клеток, в результате чего получаемый средний сигнал может плохо отображать реальную доступность регионов в различных типах клеток в пределах одной изучаемой ткани. Изображение адаптировано из [105].

2.6. Оценка качества методов предсказания эффектов регуляторных вариантов

Практический интерес представляет получение модели, способной предсказывать влияние однонуклеотидных вариантов (SNV) в геноме человека на молекулярные процессы в клетках.

Данные о влиянии таких вариантов могут быть получены из GWAS и QTL исследований, данных об аллель-специфичных событиях, экспериментальных данных, напрямую измеряющих

эффекты вариантов, и баз данных, например Clinvar [101], агрегирующих информацию из разных источников**.**

Однако обучение модели на таких данных затруднено в силу их сравнительно небольшого количества, высокой коррелированности информации между различными позициями одной последовательности в случае данных насыщающего мутагенеза, и большого уровня шума в случае информации о событиях аллель-специфичного связывания и популяционных исследований. По этой причине эти данные чаще используют для валидации и/или дообучения уже имеющихся моделей [1–3].

Обычно на данных тестируется возможность модели решать все или поднабор из следующих задач:

1) классификация однонуклеотидных вариантов на значимые и незначимые;

2) предсказание знака эффекта – приводит ли вариант к повышению или к понижению количественного признака;

3) предсказание величины эффекта (абсолютной или относительной).

При этом учитывается, что точность модели в таких задачах ограничена не только качеством самой модели, но и уровнем шума в самих данных (irreducible noise).

Помимо упомянутых задач можно оценивать насколько уверенность предсказания модели коррелирует с экспериментальной [51], однако данная задача встречается редко.

В случае задач 1 и 2 в качестве метрик качества, как правило, используется ROC-AUC (area under receiver operating characteristic curve, площадь под рабочая характеристикой приёмника) и PR-AUC (area under precision-recall curve, площадь под кривой точность-полнота). Обе метрики по-разному работают в условиях разного баланса классов, и потому обычно репортируются вместе.

В случае задачи 3 в качестве оценки качества как правило используются корреляции Пирсона и Спирмена. Аналогично, так как обе корреляции характеризуются разным поведением, часто учитывают обе.

2.7. Оценка влияния однонуклеотидных вариантов на связывания факторов транскрипции при помощи позиционно-весовых матриц

Одним из простейших способов предсказания влияния регуляторных вариантов на связывание транскрипционных факторов с ДНК является использование позиционно-весовых матриц (ПВМ) [47,48].

Позиционно-весовая матрица описывает характерный ДНК-паттерн - набор последовательностей, с которыми предпочтительно связывается конкретный фактор транскрипции.

ПВМ представляет собой матрицу Lx4, которая содержит вероятности нахождения в каждой позиции каждого из четырех нуклеотидов. Модель, таким образом, предполагает независимость вкладов отдельных нуклеотидов в специфичность ДНК-белкового узнавания.

Оценка силы связывания для последовательности размера L в простейшем случае считается как логарифм правдоподобия этой последовательности при условии нашей модели. Для последовательностей большего размера оценка получается следующим образом:

1) оценка считается для каждой подпоследовательности размера L;
2) берется максимум или другая функция [143], агрегирующая данные оценки в одно число.

Также различные программы для оценки значимости силы связывания при помощи PWM позволяют оценить p-value предсказанного значения [48]. В случае, когда необходимо измерить эффект регуляторного варианта в последовательности, достаточно посчитать оценки только для подпоследовательностей, затронутых им.

ПВМ может быть построена на основе экспериментальных данных по определению сайтов связывания данного транскрипционного фактора. Было разработано большое количество экспериментальных протоколов [144]. Среди них наиболее популярны методы: иммунопреципитация хроматина с последующим высокопроизводительным секвенированием [145], систематическая эволюция лигандов экспоненциальным обогащением (SELEX и его развитие с использованием высокопроизводительного секвенирования HT-SELEX) [146], или данные белок-связывающих микрочипов [147].

Несмотря на кажущуюся простоту ПВМ модели, на практике оказывается достаточно сложно существенно превзойти качество предсказания ею связывания и эффекта одиночных замен на коротких последовательностях [49]. Более того, предсказываемая ею оценка силы связывания хорошо коррелирует с энергией связывания данного ТФ с фрагментом ДНК [148].

Однако, при использовании ПВМ для полногеномного анализа возникает проблема большого числа ложноположительных предсказаний связывания [149]. В случае, если мы заранее не знаем фактор, связывающийся с участком где расположен вариант, приходится перебирать более 1500 возможных ТФ [144] и соответствующих им ПВМ, что, опять же, увеличивает риск ложноположительного срабатывания, в результате чего возникает проблема агрегации результатов полученных для разных факторов.

Помимо этого, ПВМ не содержит информацию о сложном характере влияния расстояния до старта инициации транскрипции, в то время как расположение мотива связывания может не просто влиять на силу воздействия транскрипционного фактора, но и полностью менять его роль с активаторной на ингибиторную [150]. Слабость сайтов связывания ТФ может компенсироваться их оптимальным положением и ориентацией, обеспечивая тканесфецифичную экспрессию [151].

Чтобы учесть взаимодействие транскрипционных факторов, можно использовать предсказание ПВМ в качестве признаков, передаваемых на вход другой модели, обучаемой предсказывать связывание ТФ или эффекты регуляторных вариантов [51,152,153].

По своей сути ПВМ являются марковскими моделями нулевого порядка. Можно использовать модели большего порядка, а также различные модификации марковских моделей. Вопреки тому, что они могут исправлять часть недостатков ПВМ, в большинстве случаев ПВМ оказывается наилучшим представлением мотива [154]. Более того, для задач предсказания эффектов регуляторных мутаций, оцененных при помощи протокола SNP-SELEX [155] показано, что правильно подобранная ПВМ работает на уровне или превосходит более сложные методы машинного обучения [49].

2.8. Утечка данных при работе с геномными данными

Перед тем, как перейти непосредственно к применению методов машинного обучения, необходимо обсудить важную, особенно применительно к биологической области, проблему, которая может возникнуть при обучении модели – утечке данных (data leakage).

При так называемом индуктивном машинном обучении – когда мы обучаем модель на одних данных, чтобы потом работать с другими – нам наиболее важна **генерализация модели** – то, как она будет вести себя новых, не виденных ею ранее данных. В общем случае, так как мы не можем собрать все новые данные (иначе бы смысл машинного обучения терялся), мы ограничиваемся **оценкой** генерализации модели как качества модели на отложенной **тестовой выборке**, которая не использовалась при обучении и подборе гиперпараметров модели.

Утечка информации – ситуация, при которой из-за неправильной процедуры сборки, предобработки первоначальных данных, их разбиения на обучение и тест, выбора конкретной процедуры обучения или выбора модели машинного обучения, информация о тестовых данных явно или неявно учитывается получаемой моделью. В результате этого оценка качества модели завышается, и неверной оказывается оценка способности модели к генерализации [156].

Проблема эта актуальна и для задач биоинформатики, молекулярной биологии, генетики и медицины [157]. Например, при работе с белковыми структурами важно учитывать гомологию и степень сходства последовательностей [158–160], при разработке лекарств – сходство структур белков, положения и строения их активных центров и лигандов [161,162]. Это верно и для биомедицинских приложений машинного обучения, а именно, необходимо контролировать чтобы в обучение и тест шли данные от разных пациентов и отсутствовали дубликаты, а методы (например, оборудование для измерения сигнала) получения данных разных классов были схожи [163–166].

Возникает данная проблема и при работе с нуклеотидными последовательностями. Например, GC-состав последовательности может коррелировать с ее способностью специфически связываться с факторами транскрипции и с ее регуляторной активностью в целом, и это необходимо учитывать либо при формировании выборки из последовательностей отрицательного класса [167], либо при сравнении модели со случайным предсказателем. В работе [168] по предсказанию экспрессии показывается, что без использования разбиения обучающей и тестовой выборок по хромосомам, оценка качества модели оказывается сильно завышенной из-за локальной схожести эпигенетических сигналов и сравнительно малым количеством клеточно специфичных районов.

Еще более наглядным примером служит работа [169], показывающая возможность обучения модели-предсказателя энхансер-промоторных взаимодействий, показывающей F1-score порядка 0.80. В последовавшем за ней опровержении [170] было показано, что достигаемое качество по большей части обусловлено некорректным разбиением, использовавшимся авторами изначальной работы – пары в работе отбирались в обучение и валидацию случайным образом. Это приводило к тому, что пары, имеющие общий промотор, а, следовательно, общие промоторные признаки и схожие “оконные” признаки (эпигенетические сигналы в интервале между энхансером и промотором), попадали как в обучающую, так и в тестовую выборки. В случае разбиения пар на основании их положения в геноме (близкие пары оказывались вместе либо в тестовой выборке, либо в обучающей), предложенного в работе [170], оценка качества модели стремительно падала, немногим превосходя качество случайного предсказания. В случае же разбиения на основании промотора (все пары, относящиеся к одному промотору, вместе попадают или в тест, или в обучение), также рассмотренного в работе [170], утечки данных на эпигенетических признаках не происходило.

Стоит отметить, что утечка данных при работе с геномными данными в биологии может обусловлена не только неправильным разбиением, но и другими причинами, включая предобработку данных и выбор важных признаков до разбиение на обучение и тест [171]. Кроме того, из-за обилия гомологичных участков и функциональных повторов, задача разбиения генома способом, полностью предотвращающим утечку данных считается некоторыми исследователями неразрешимой [172].

2.9. Классическое машинное обучение при работе с регуляторными последовательностями

2.9.1. gkm-SVM и delta-SVM

Одним из хороших примеров применения классического машинного обучения для эффектов мутаций в регуляторных регионах является модель gkm-SVM [173–177] (**рис. 6**), основанная на методе опорных векторов.

Ранние методы использовали в качестве признаков последовательности наличие или встречаемость в ней подпоследовательностей длины k (k-меров). Такой подход имел ряд сильных ограничений, т.к не учитывал схожести разных k-меров, при малых значениях k приводил к малоинформативным признакам, а при больших – к чрезвычайно разреженному представлению последовательностей [175].

В gkmSVM вместо обычных k-меров используются k-меры с пропусками (gapped kmers), которые позволяют при подсчете k-меров учитывать не только точное вхождение k-мера, но и все k-меры похожие на него, что увеличивает качество модели и уменьшает склонность к переобучению. Метод широко использовался для предсказания различных биохимических профилей последовательности, например силы связывания белка с РНК, участков ChIP-Seq или DNase-Seq [173,174,176]. Для оценки эффекта мутации далее достаточно сравнить предсказания модели на последовательности с мутацией и без неё [50].

Так как предсказание при помощи SVM работает достаточно долго, авторы предлагают метод deltaSVM, заключающийся в предварительном преподсчете предсказания модели для всех возможных k-меров. Каждое из полученных предсказаний gkmSVM считается далее весом k-мера.

Далее предсказание эффекта замены (в частности, однонуклеотидного варианта) осуществляется следующим образом:

1) исходная последовательность до и после замены разбивается на k-меры;
2) каждому получившемуся k-меру присваивается вес;
3) считается разница между суммой весов k-меров из исходной последовательности и суммой весов k-меров.

Хотя это позволяет выполнять предсказание за линейное время, обучение исходной модели имеет в лучшем случае квадратичную от числа объектов сложность, что делает его неприменимым для больших объемов данных.

Аналогично ПВМ, предсказания моделей gkmSVM, обученных на разных экспериментальных данных, также можно объединять при помощи другой модели машинного обучения [51].

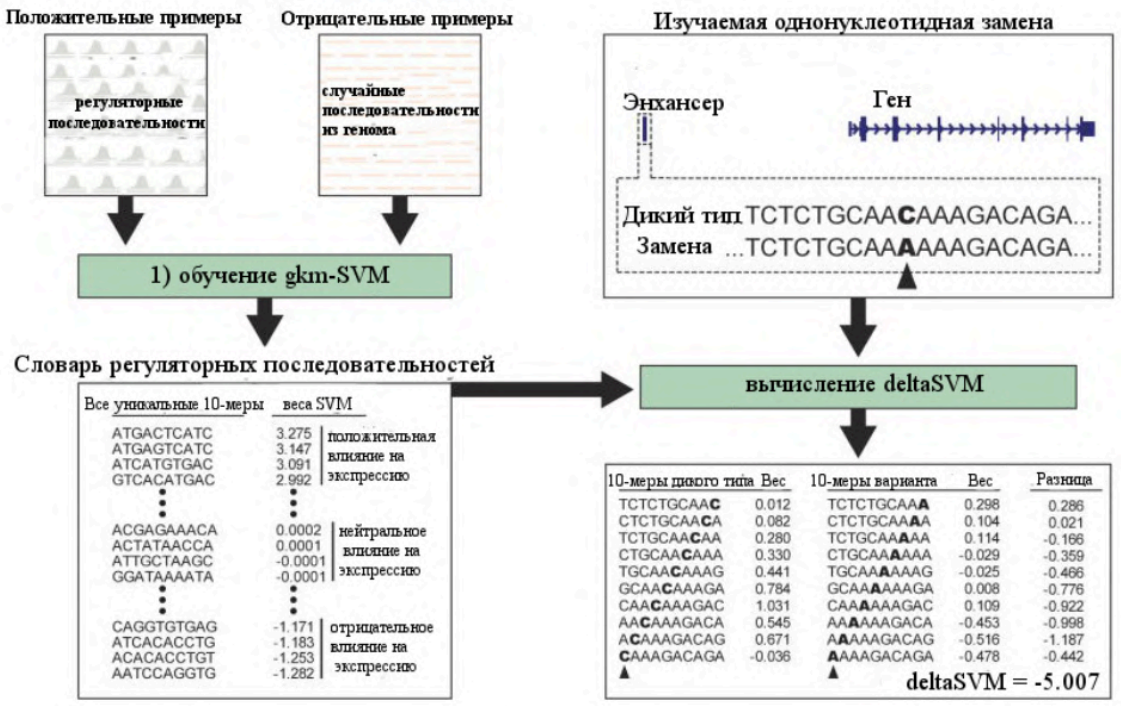


Рисунок 6. Схема deltaSVM. Первый шаг – обучение gkm-SVM. Далее при помощи gkm-SVM оценивается вес каждого уникального 10-мера. Далее вычитая веса кмеров из последовательности без мутации и из последовательности с мутацией, можно получить скор deltaSVM. Адаптировано из работы [50]

2.9.2. Решающие деревья и модели на их основе

Решающие деревья (Decision trees) воспроизводят логические схемы, позволяющие принять решение о классификации объекта с помощью ответов на иерархически организованную систему вопросов (**рис. 7, А**). Вопрос, задаваемый на последующем иерархическом уровне, зависит от ответа, полученного на предыдущем уровне [178].

В силу чувствительности к шуму и склонности к переобучению в чистом виде деревья решений большого распространения в биологии не получили. Большее распространение получили методы, основанные на деревьях решений и лишенные по крайней мере части их недостатков и основанные на ансамблях — объединение набора моделей, используемых для решения одной и той же задачи [179].

Первым из широко использующихся методов данного типа является **случайный лес** (**рис. 7, В**). Он представляет собой набор из большого числа решающих деревьев, каждое из которых обучалось на подвыборках исходного обучающего набора. Предсказание этого набора вычисляется как среднее предсказаний каждого из решающих деревьев.

Случайный лес 1) устойчив к шуму; 2) способен работать в условиях значительного преобладания числа признаков над числом объектов; 3) в первом приближении не требует сложной процедуры подбора гиперпараметров, что, в частности, позволяет использовать его когда выделение отдельной выборки для их выбора затруднено. В силу этого он до сих пор широко

используется в биологии [51,55,153,180–183], в том числе он может быть использован для работы нуклеотидными последовательностями. В частности, он использовался для предсказания эффектов 5’UTR на трансляцию по данным массовых параллельных репортерных экспериментов [184] и предсказания влияния мутаций в экзонах на события сплайсинга [185].

В случае регуляторной геномики наибольшее распространение получило использование случайного леса для агрегации предсказаний других моделей, таких как ПВМ [153] или нейронные сети [51,55].

Вторым методом, основанным на деревьях решений, является **бустинг** (**рис. 7, C**)**.** Идея подхода состоит в построении ансамбля моделей, каждая последующая модель в которой исправляет предсказания предыдущих. Наиболее популярным стала разновидность этого метода – **градиентный бустинг**. В нем на каждом шаге построения ансамбля ищется такая модель, которая аппроксимировала бы антиградиент (отрицательный градиент) функции ошибки текущего ансамбля предыдущих моделей [186]. Существует несколько реализаций данного подхода, наиболее известные – xgboost [187], lightgbm [188], catboost [189]. Метод показывает лучшие результаты при работе с данными, не имеющими внутренней структуры (табличные данные) [190,191].

Метод широко используется при работе с биологическими данными [183,192,193], в частности, он использовался для предсказания событий сплайсинга [194] и внутриклеточной локализации мРНК [195] по данным на основе МПРЭ. В регуляторной геномике градиентный бустинг также в основном используются для агрегации предсказаний нейронных сетей [51,196,197].

Использование ансамблей на основе деревьев решений для предсказания активности регуляторных последовательностей или эффектов мутаций в них по последовательности напрямую распространения не получило.

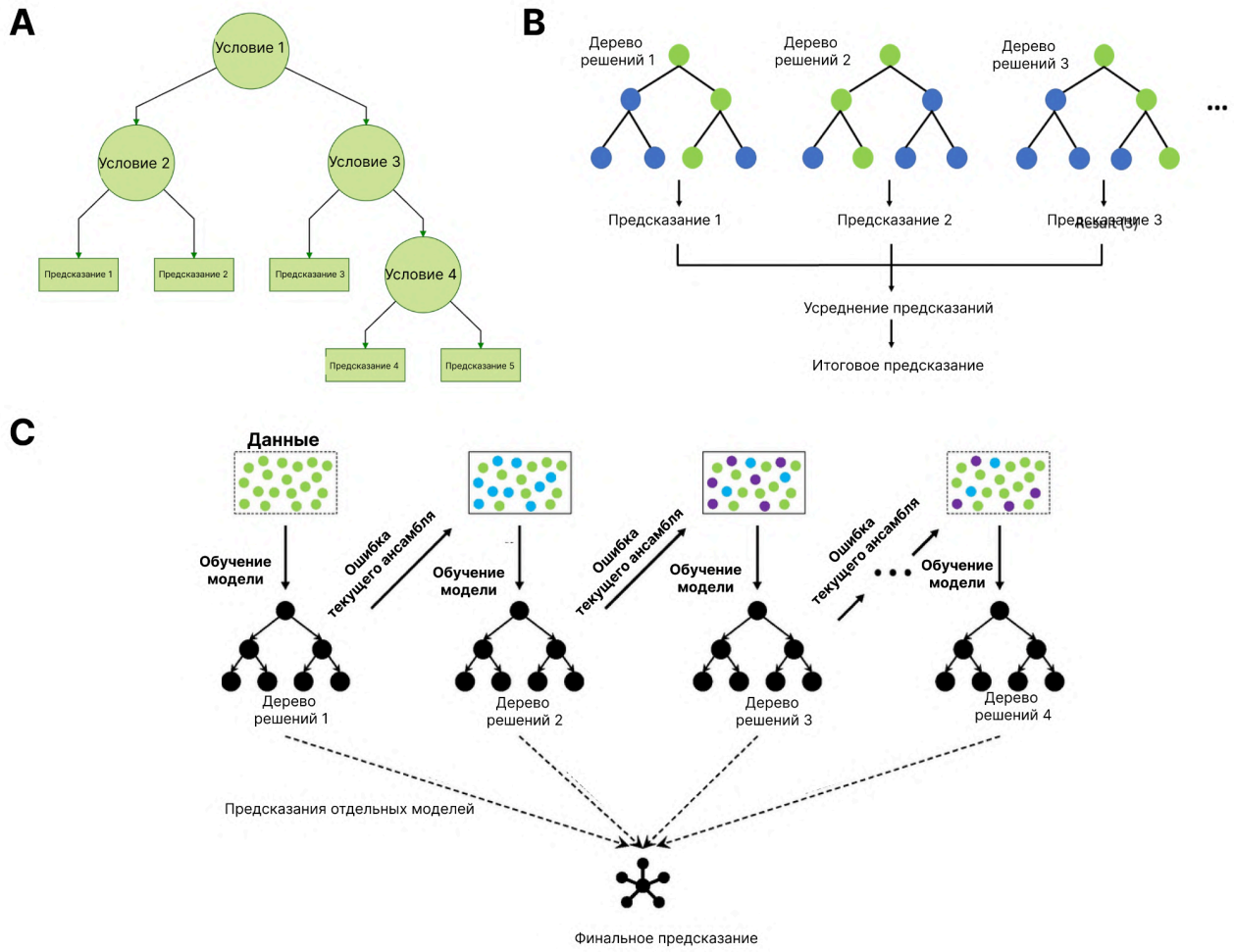


Рисунок 7. **A.** Дерево решений. Адаптировано с сайта machinelearningtheory.org **B.** Схема работы случайного леса. В первом приближении случайный Множество деревьев решений, обученных на подвыборках исходного обучающего набора. Для получения предсказания ансамбля голоса отдельных деревьев усредняются. Адаптировано из [198] **C.** Схема работы бустинга. Каждое дерево решений в ансамбле корректирует предсказания предыдущих, таким образом финальное предсказание представляет собой сумму предсказаний отдельных деревьев. Адаптировано из [199]

2.10. Основы методов глубокого машинного обучения, используемых при работе с регуляторными последовательностями

Следующим шагом в развитии методов для работы с регуляторными последовательностями стали искусственные нейронные сети.

2.10.1. Общая схема обучения и предсказания

Как уже упоминалось ранее, прямые данные о влиянии регуляторных вариантов малочисленны, получены в разных исследованиях и сильно зашумлены. По этой причине приобрело популярность развитие метода, предложенного авторами gkmSVM [171–175]. А именно – учить модели предсказывать эпигенетические треки по последовательности, а далее

использовать разницу в предсказаниях между последовательностями с референсным и альтернативным вариантом для оценки его эффекта. Такие модели в зарубежной литературе называются **seq2activity модели**, в данной работе мы будем называть их **полногеномными моделями**.

Таким образом, вначале модель учится на полногеномных омиксных данных (DNase-Seq, ATAC-Seq, ChIP-Seq, хроматиновые метки, RNA-Seq, CAGE, и т.д.) предсказывать по последовательности соответствующий сигнал в этой области. Затем обученную модель можно использовать несколькими способами.

В случае обучения модели на задачу предсказания одного сигнала, как это было в случае gkmSVM, как оценку влияния варианта можно напрямую использовать разницу в предсказании для референсной и альтернативной последовательностей [200].

В случае нейросетей часто прибегают к обучению моделей сразу на несколько задач (**мультитаргетные модели**): предсказание разных экспериментов из разных клеточных линий и т.д [53,55] (**рис. 8**). Аналогично можно использовать предсказания сразу большого числа моделей, обученных на отдельные задачи [51].

В этом случае задача усложняется, так как и для референсной, и для альтернативной последовательности мы получаем сразу несколько чисел (векторов), которые необходимо объединить в одно предсказание.

Для этого используются два семейства подходов – **zero-shot предсказание** или **дообучение** [3]**.**

В случае zero-shot предсказания мы используем методы агрегации предсказаний модели для референсной и альтернативной последовательностей, которые сами не являются обучаемыми (**рис. 9**) . В качестве таких методов можно:

1) использовать среднее (опционально – абсолютной) разности предсказаний в референсной и альтернативной последовательностях [55];

2) использовать разницу или среднее разниц для наиболее подходящего эпигенетического сигнала/сигналов, исходя из априорных соображений – например, использовать предсказания сигналов, измеренных в соответствующем клеточном типе [55];

3) использовать максимальную абсолютную разницу предсказаний в референсной и альтернативных последовательностях [201];

4) трактовать предсказания модели как представление референсной и альтернативной последовательностей и использовать в качестве оценки вреда варианта расстояние между последовательностями [57].

Помимо прочего, многие модели обучаются предсказывать не одно значение сигнала для поданной на вход последовательности, а по значению на каждую непересекающуюся

подпоследовательность – бин. В этом случае помимо агрегации сигналов с разных эпигенетических треков необходимо также решить эту задачу для бинов [32,55].

Методы на основе дообучения делятся на два подкласса – методы **пробинга** (probing) и собственно дообучения (**finetuning**).

В случае методов пробинга предсказания модели для референсной и альтернативной последовательностей используются в качестве входных признаков для другого метода машинного обучения – **метамодели** [53,55,202] (рис ММ. В). В качестве методов машинного обучения для пробинга используют линейную регрессию [55,202] (**линейный пробинг**), случайные леса [55], градиентный бустинг [53,197] или многослойный персептрон [203].

Для линейной регрессии можно использовать L1-регуляризацию (Lasso-регрессия), что позволяет работать с обычно сильно коррелированными предсказаниями мультитаргетных моделей и выбирать из предсказаний для большого числа различных клеточных линий наиболее релевантные. Аналогично для такой постановки хорошо подходит случайный лес.

Помимо прочего, встречаются иерархические подходы, в которых для обучения метамодели используются предсказания сети в нескольких (возможно, пересекающихся) окнах[197].

В случае собственно дообучения нейросети, обученной на задаче предсказания эпигенетических сигналов, ее веса частично или полностью размораживаются и она обучается на новую задачу [202] (**рис. 9 С**). При этом при обучении модели могут использоваться меньшая скорость обучения, дополнительный штраф на слишком сильное отклонение весов от значений соответствующих весов изначальной сети [61,204,205].

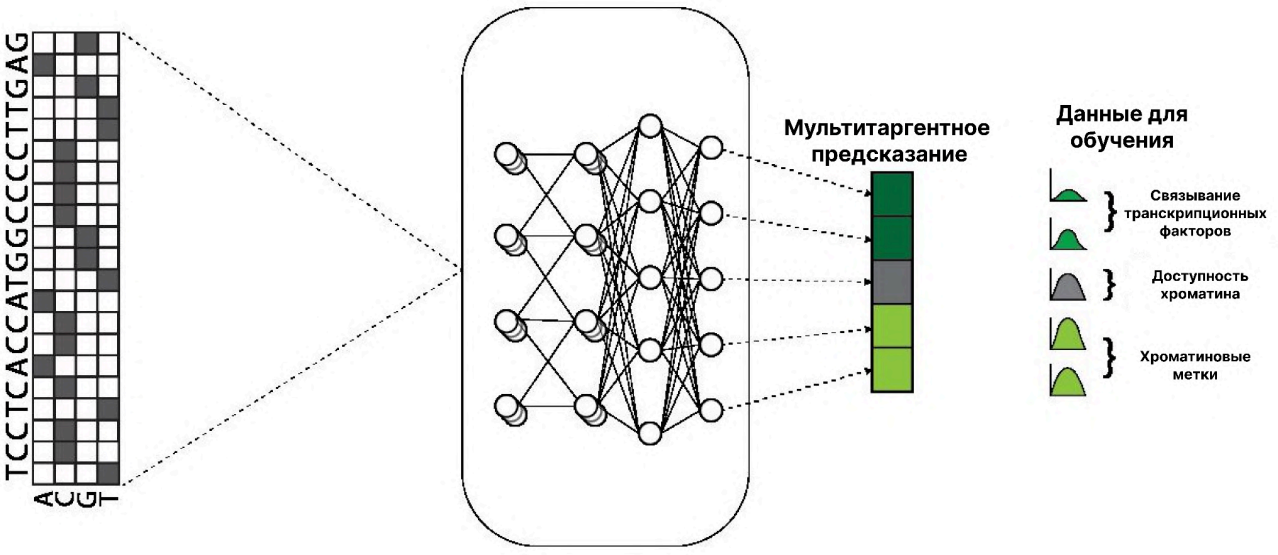


Рисунок 8. Схема обучение полногеномной нейросетевой модели. Модель учится по последовательности предсказывать сигналы множество различных экспериментов, например данных ChIP-Seq и доступности хроматина. Адаптировано из [3]

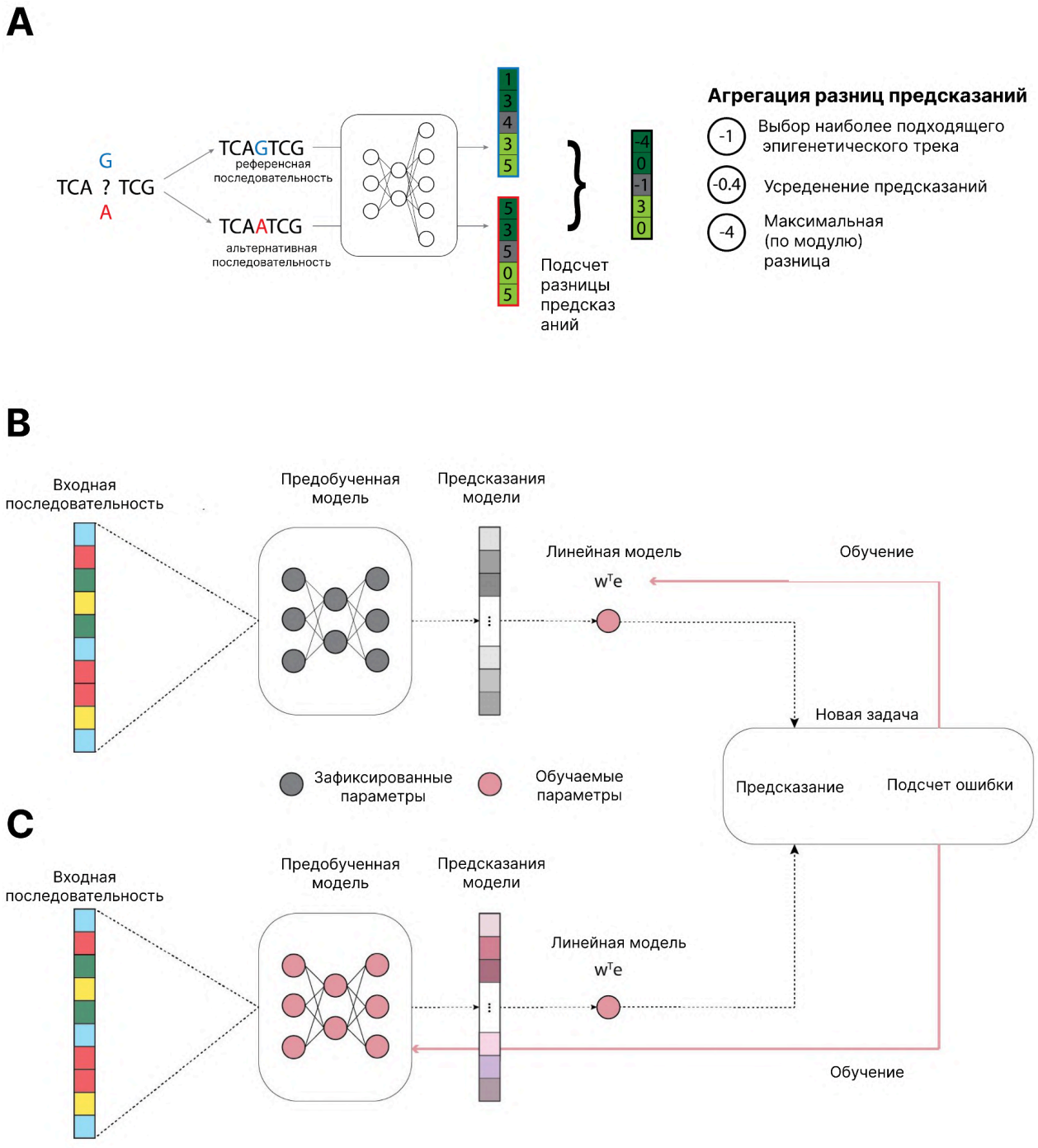


Рисунок 9. Различные варианты предсказания эффектов регуляторных мутаций при помощи полногеномной модели. **A. Zero-shot** предсказание. Для получения оценки эффекта варианта разницы предсказанных эпигенетических треков для референсной и альтернативной последовательностей агрегируется простой функцией, например, средним или максимумом. Также из всех треков может выбираться наиболее соответствующий задаче исходя из биологических знаний. **B.** Пробинг – на предсказаниях модели для референсного и альтернативного вариантов учится предсказывать эффект мутации дополнительная модель машинного обучение, часто – линейная модель, в этом случае пробинг называют линейным. **C.** Модель целиком дообучается на новой задаче, с частично или полностью размороженными весами всех слоёв. Адаптировано из [3]

2.10.2. Сверточные нейронные сети

Сверточные сети подходят для данных с выраженной внутренней локальной структурой, в том числе для нуклеотидных последовательностей [206]. Для представления нуклеотидов чаще всего используются так называемое one-hot кодирование — каждому нуклеотиду сопоставляется вектор длиной 4, в котором только одна позиция (в зависимости от нуклеотида) равна 1 [206] (**рис. 10а**). В простейшем случае (**рис. 10**) затем это представление подается в серию из сверточных слоев и слоев пулинга (feature extractor), а выход из этой серии подается на вход одному или нескольким полносвязным слоям (head, голова сети), которые дают итоговое предсказание.

Можно заметить аналогию между операцией свертки на первом слое и матрицей ПВМ. Можно предположить, что при некоторых условиях нейронная сеть способна выучить мотивы сайтов связывания транскрипционных факторов на уровне весовых матриц в первом сверточном слое. Это является хорошим популярным обоснованием того, почему сверточные сети должны работать для последовательностей ДНК. Существуют даже методы, пытающиеся при помощи нейронных сетей находить матрицы ПВМ на основе экспериментальных данных [207]. Однако в общем случае показано, что на первых слоях сверточных нейронных сетей не обязательно выучиваются мотивы связывания ТФ – например, веса сверток могут соответствовать отдельным частям мотивов, причем использование больших размеров ядер сверток (гипотетически стимулирующих нейронную сеть выучивать мотивы целиком) не влияет на это поведение [208].

Несмотря на отсутствие однозначного соответствия между свертками и мотивами, достаточно глубокие сверточные нейронные сети способны выучивать и мотивы связывания транскрипционных факторов [209], и более сложные взаимодействия между ними [54,210].

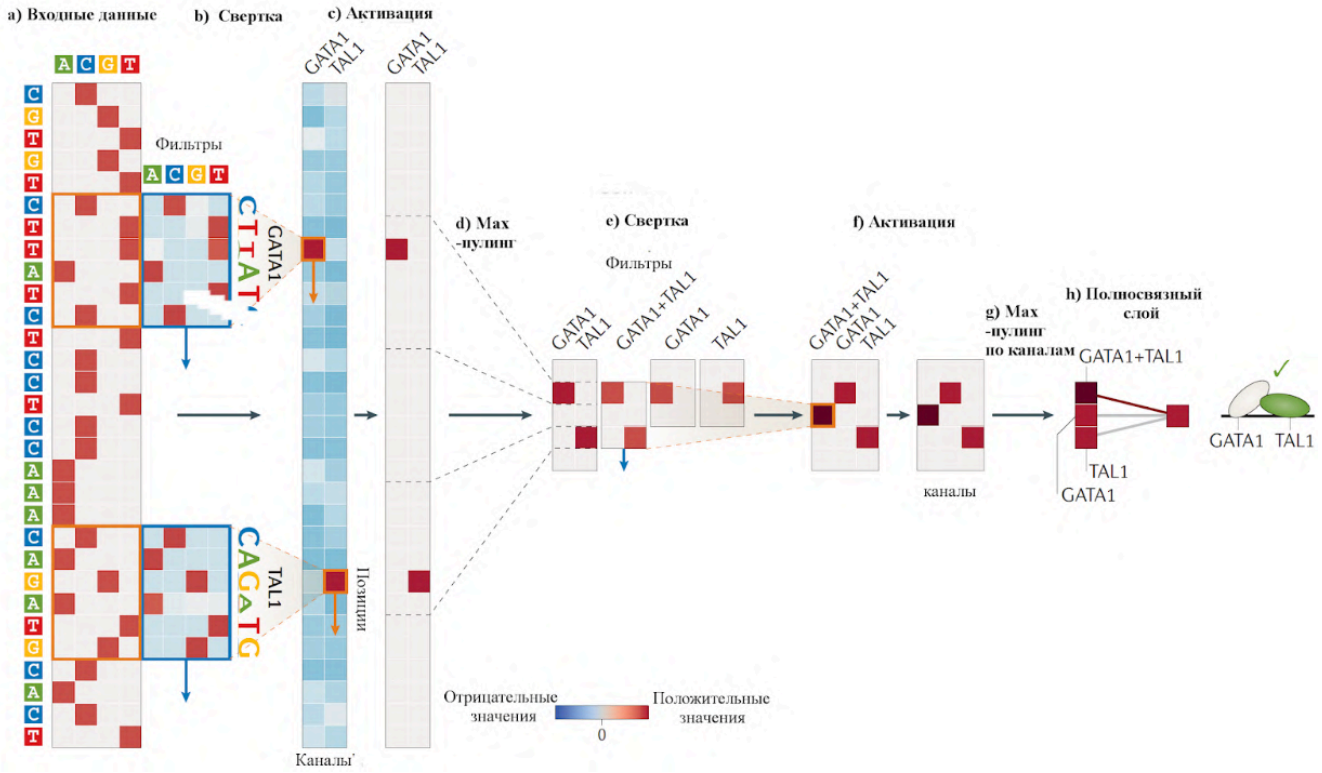


Рисунок 10. Структура простой сверточной сети для работы с нуклеотидными последовательностями. **a)** Входная последовательность кодируется при помощи one-hot кодирования. **b)** Первый сверточный слой сканирует последовательность, используя фильтры, которые могут в первом приближении восприниматься ПВМ сайтов связывания транскрипционных факторов. **c)** С помощью активационной функции (в данном случае - ReLU (возвращает само значение x, если x > 0, иначе 0)), негативные значения зануляются. d) Из каждой области с помощью max-пулинга берем максимальные значения. **e)** Второй слоя свертки ищет вхождения индивидуальных транскрипционных факторов и их групп. **g)** max-пулинг по каждому каналу. **h)** полносвязный слой, чтобы сделать финальное предсказание. Адаптировано из [206]

2.10.3. Остаточные соединения и батч-нормализация

При обучении глубоких нейронных сетей могут возникать проблемы с распространением сигнала, приводящие к затуханию или взрыву градиентов, в результате чего нейронная сеть перестает обучаться. Для того чтобы предотвратить такое поведение и улучшить сходимость обучения нейронной сети, существует несколько подходов, самыми известными из которых являются остаточные соединения (residual connections) [211] и батч-нормализация [212] (**рис. 11**).

Первый подход – остаточные соединения – состоит в добавлении к выходу блока нейронной сети его входа. Это облегчает нейросети передачу информации между слоями без изменений и улучшает характеристики градиента при обратном распространении ошибки.

Второй подход – батч-нормализация. В данном подходе для каждого признака A_i объекта в батче производится следующая трансформация:

1) признак нормализуется по формуле:

$$\hat{A}_i = \frac{(A_i - \mu_i)}{\sigma_i}$$

где μ_i и σ_i – среднее и стандартное отклонение признака, оцененные по батчу (для стандартного отклонения вводится поправка, чтобы оно не могло равняться 0)

2) полученный нормализованный признак трансформируется по следующей формуле

$$B_i = \hat{A}_i \cdot \gamma_i + \beta_i$$

где γ_i и β_i – обучаемые параметры слоя.

Во время предсказания вместо среднего и дисперсии признака, подсчитанных по батчу, используется их скользящее среднее, подсчитанное за время обучения нейронной сети.

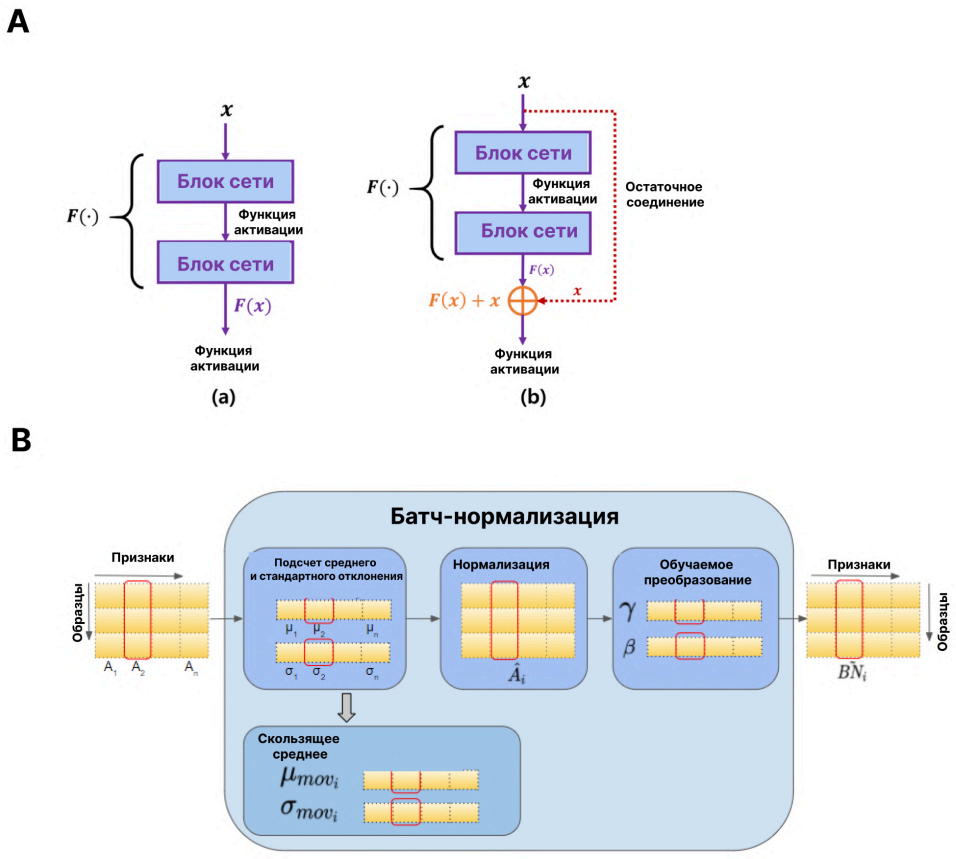


Рисунок 11. Методы улучшения сходимости обучения нейронных сетей. **A.** Остаточные соединения. Вход блока складывается с его выходом, за счет чего нейросети становится легче сохранять уже выученный сигнал. Адаптировано из [213]. **B.** Батч-нормализация. Каждый входной признак блока нормализуются за счет вычитания среднего и деления на стандартное отклонение, подсчитанные по батчу. Затем, чтобы дать возможность сети отменить данную трансформацию или видоизменить ее результат, каждый признак домножается на обучаемый параметр γ и к нему добавляется обучаемый параметр β. Во время тестирования вместо средних и стандартного отклонения, подсчитанных по батчу, используются их скользящие средние за время обучения нейронной сети. Адаптировано из [214].

2.10.4. Рецептивное поле и размер контекста сети

При работе с нейронными сетями возникает характеристика, которую нужно учитывать – на каком расстоянии в исходной последовательности может агрегировать сигналы нейронная сеть.
В случае сверточных сетей размер окна, который она может учесть, называется **рецептивным полем**. Для архитектур, которые пришли из обработки языков, вместо этого термина чаще используют родственный термин – **размер контекста**. В рамках данной работы два данных термина будут использоваться как синонимы.

В случае нейронных сетей для регуляторных регионов размер контекста определяет, к примеру, то, сможет ли нейросеть (или ее часть) учитывать взаимодействие разных транскрипционных факторов, связывающихся с данной последовательностью на определенном расстоянии друг от друга (**рис 12. A)**. Например, нейронная сеть с рецептивным полем в 100кб п.о даже теоретически сможет предсказать эффект только части вариантов, упомянутых в разделе “Примеры клинически значимых вариантов”

2.10.5. Рецептивное поле сверточной сети

В простейшем случае, когда в сети есть только последовательные сверточные блоки (**рис. 12 B**), размер рецептивного поля можно оценить по формуле:

$r = 1 + \sum_{i=1}^{n} (k_i - 1)$, где k_i - размер ядра *i*-го слоя.

Т.е размер рецептивного поля будет расти линейно по мере увеличения глубины сети, что плохо применимо для работы с большими последовательностями.

Существует несколько способов увеличения рецептивного поля. Первый из них – дилатированные свертки [56,215] вставляют между позициями фильтра “пробелы” длины d (**рис. 12 C**), и в этом случае формула для подсчета рецептивного поля выглядит следующим образом:

$r = 1 + \sum_{i=1}^{n} d_i(k_i - 1)$, где k_i - размер ядра i-го слоя, а d_i - размер “пробела”, dilation factor.

При этом dilation factor каждого последующего слоя, как правило, увеличивают в некое число раз, что приводит к экспоненциальному росту рецептивного поля [56], однако сильно замедляет работу сети.

Двумя другими вариантами решения проблемы увеличения рецептивного поля являются свертки с шагом (strided convolutions) и слои пулинга. В первом случае свертка делается с некоторым шагом s_i, что уменьшает размеры выхода свертки (**рис. 12 D**). Во втором случае входной вектор разбивается на (обычно непересекающиеся) части длины s_i, в каждой из которых считается максимальное, минимальное или среднее значения, которые и служат выходом слоя

(**рис. 12 Е**). В обоих случаях размер входного вектора уменьшается примерно в s_i раз. В обоих случаях же случаях каждый из таких слоёв сам по себе не увеличивает рецептивное поле нейронной сети, но увеличивает вклад всех последующих сверток в рецептивное поле, которое считается по формуле [216]:

$$r = 1 + \sum_{i=1}^{n} [(k_i - 1) \prod_{j=1}^{i-1} s_i]$$

Опять же, при таком подходе добавление каждого нового слоя свертки с шагом или слоя пулинга увеличивает размер итогового рецептивного поля экспоненциально. К минусам этого подхода можно отнести то, что за счет агрегации информации сеть с такими слоями может быть склонна хуже учитывать вклад одиночных нуклеотидных замен [217].

Если в сети признаки, выученные сверточной частью, далее целиком передаются в полносвязные слои, то при достаточном количестве обучающих данных она в принципе может выучить все взаимодействия, т.е ее рецептивное поле фактически равно длине поданной на вход последовательности. Однако это фиксирует длину, с которой может работать нейронная сеть, и значимо увеличивает число параметров в ней – полносвязные слои могут содержать параметров больше, чем все сверточные до этого.

По этой причине обычно признаки, выученные сверточной частью, сжимают при помощи усреднения значений признаков по каналам (global average pooling). В этом случае, если при предсказании сверточная сеть не смогла увидеть некоторое взаимодействие до слоя усреднения, оно с большой вероятностью не будет обнаружено и далее.

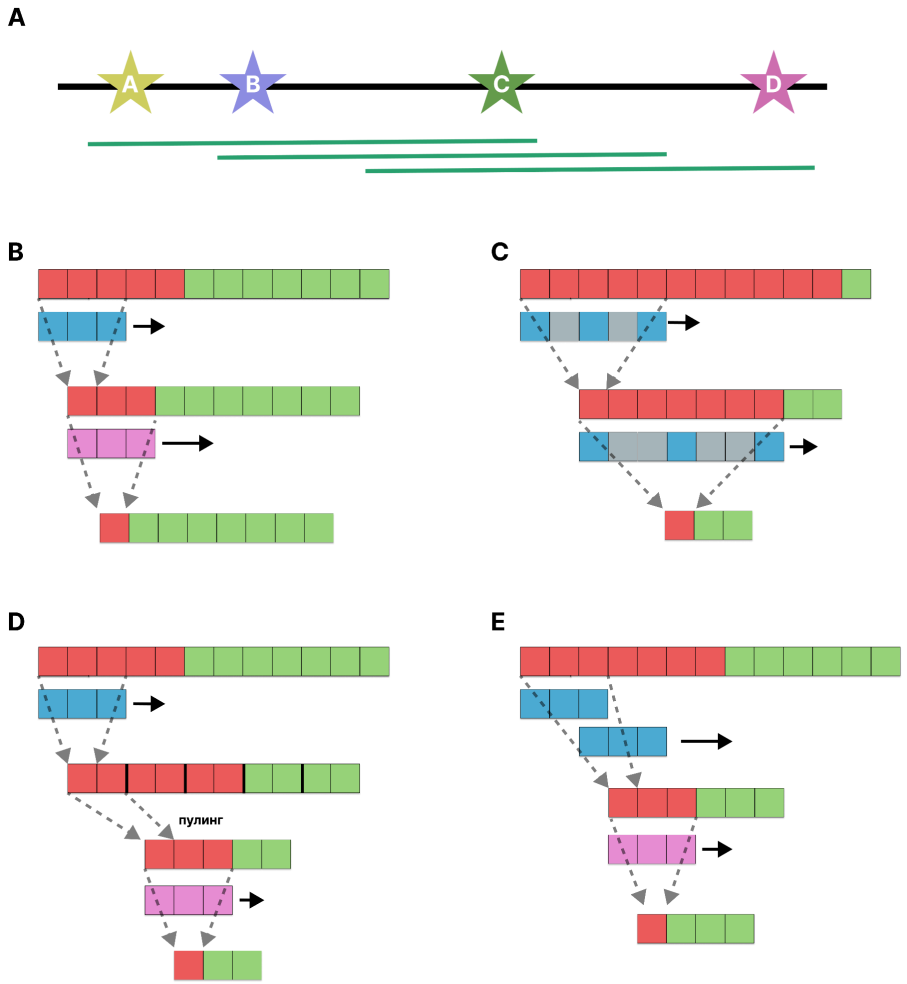


Рисунок 12. A. Рецептивное поле (зеленая линия) определяет то, какую долю последовательности может увидеть нейронная сеть и какие взаимодействия выучить. В данном случае модель может увидеть взаимодействие факторов A,B и C и факторов C и D, но не факторов A и D. **B.** Рецептивное поле двух последовательных сверток – каждое значение выходного слоя обращает внимание на 5 значений из входных данных. **C.** Рецептивное поле двух дилатированных сверток ($d_1 = 1$, $d_2 = 2$) **D.** Между двумя сверточными слоями применяется пулинг размера 2. **E.** Используются strided convolution, в которой фильтр применяется не в каждой позиции, а с определенным шагом.

2.10.6. Рекуррентные слои

Рекуррентные слои основаны на идее последовательного чтения последовательности и обработки каждой ее части при помощи одних и тех же блоков. Поведение этих блоков зависит от скрытого состояния, которое модифицируется по мере прочтения последовательности (**рис. 13 A**). Существуют примеры удачного применения гибридных архитектур из сверток и специальной версии рекуррентных блоков – LSTM (Long short-term memory) [218] к нуклеотидным последовательностям [219,220], в частности, в регуляторной геномике [221,222]. Для биологических задач применяются и другие модификации рекуррентных блоков [223].

2.10.7. Слои внимания

Слои внимания представляют другую возможность увеличить рецептивное поле нейросети, избегая чрезмерного увеличения числа параметров в модели. Общая идея заключается в том, что вместо того, чтобы выучивать веса, с которыми мы учитываем те или иные позиции во входной последовательности при вычислении новых признаков для данной, мы будем вычислять эти веса на лету, в зависимости от текущих значений признаков для данной позиции [224]. Несмотря на успешное применение слоев внимания во многих архитектурах для работы с регуляторными последовательностями [32,55,225], на данный момент остается не до конца ясным, действительно ли они вносят принципиальный вклад в качество полученных решений или же могут быть заменены правильно подобранными архитектурами полностью сверточных сетей.

2.10.8. Замена слоев внимания

Слои на основе механизма внимания требуют квадратичных от размера входной последовательности затрат на подсчет ответа, что делает сложным их использование для достаточно больших контекстов, которые могут быть важны для геномных задач. По этой причине предпринимаются попытки применения альтернатив, по-прежнему позволяющих учесть информацию со всей последовательности, но обладающих меньшей вычислительной сложностью. Среди подобных подходов можно отметить HyenaDNA [226], использующую параметризованные длинные свертки, Mamba-DNA (Caduceus) [227], основанную на модели пространства состояний. Стоит также упомянуть и изящное решение на основе использования рекуррентной памяти, позволяющей применять слои внимания к отдельным участкам последовательности, GENA-RMT [228].

2.10.9. Реальный размер рецептивного поля

Следует учитывать, что реальное рецептивное поле нейронной сети может отличаться от рассчитываемого теоретически, как для сверточных сетей [229], так и для рекуррентных сетей [230] и архитектур на основе механизма внимания [231]. Нейронная сеть может не учитывать достаточно далеких взаимодействий, если информации о них в обучающей выборке слишком мало и/или она шумная. Также ей может мешать затухание сигнала и другие явления, возникающие на практике при обучении нейронных сетей.

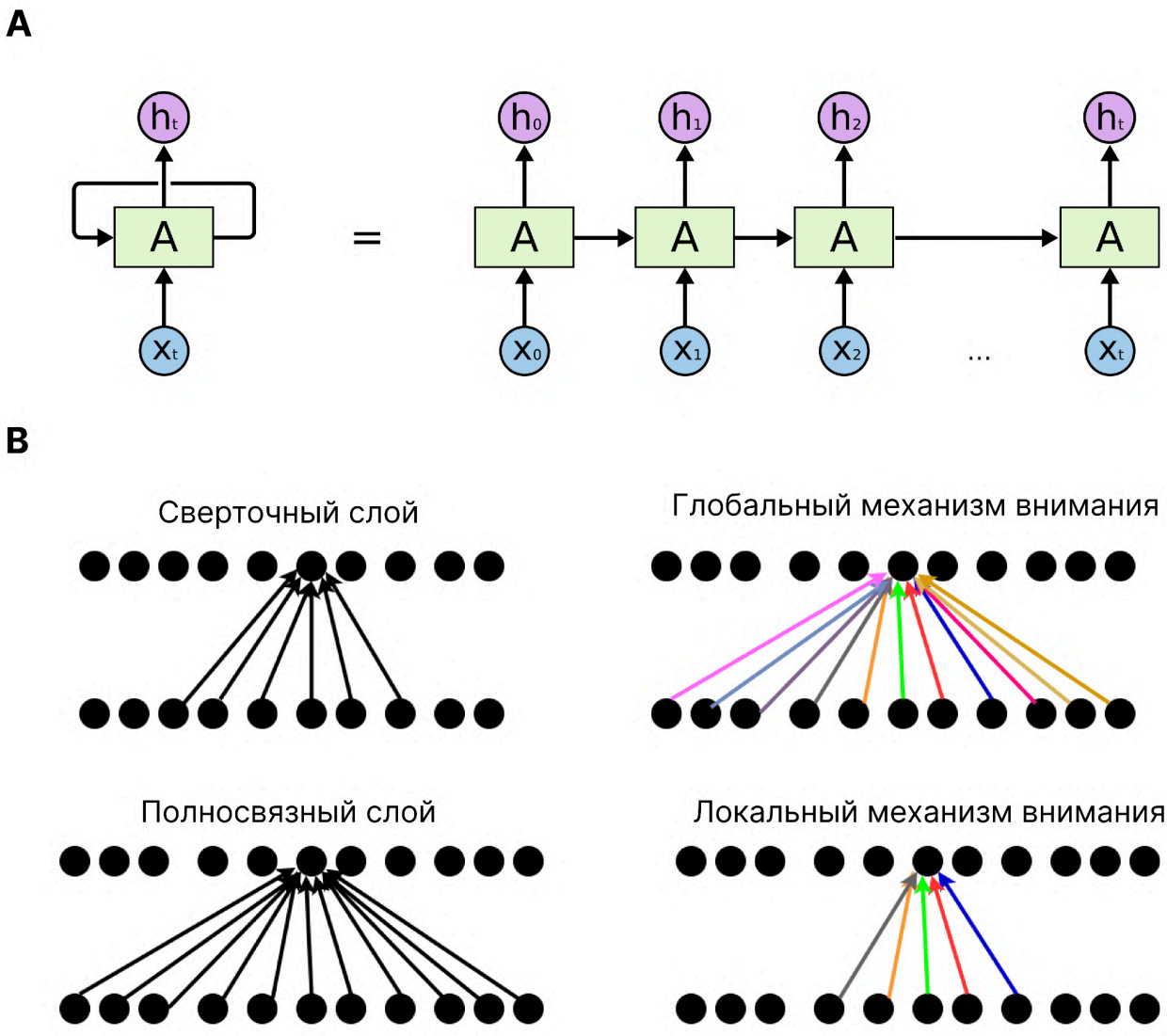


Рисунок 13. А. Общая схема рекуррентного блока. h_t – скрытое состояние в момент времени t, x_t – наблюдение в момент времение t. Адаптировано из [232]. **В.** Различия между сверткой, полностью связным слоем, глобальным и локальным слоями внимания. В случае слоев внимания веса, с которыми учитываются признаки с предыдущего слоя вычисляются на ходу, что для больших последовательностей позволяет уменьшить число параметров в нейронной сети.

2.10.10. Аугментация учебной выборки

Модели машинного обучения, как правило, начинают обгонять классические методы при достаточно большом размере выборки, на малом же они склонны переобучаться.

Одним из способов для увеличения размера выборки и борьбы с переобучением является аугментация – добавление в обучающую выборку модифицированных по определенным правилам исходных данных. Данный подход хорошо работает в случае задач компьютерного зрения и работы с естественным языком [233,234].

Идея подхода состоит в том, что некоторые преобразования (например, поворот изображения) преобразуют изображение в похожее, которое вполне могло бы встретиться в реальности (**рис. 14 А)**. За счет добавления таких данных мы неявно сообщаем модели эту информацию. Таким образом, в большинстве случаев аугментация работает за счет увеличения информации, которая содержится в обучающей выборке. При этом для изображений известны аугментации, применение которых приводит к данным, в реальной жизни не встречающимся, например, MixUp, CutOut и CutMix [235,236] (**рис. 14 В**). Предполагается, что такие аугментации работают за счет добавления неявных членов к оптимизируемой функции [235].

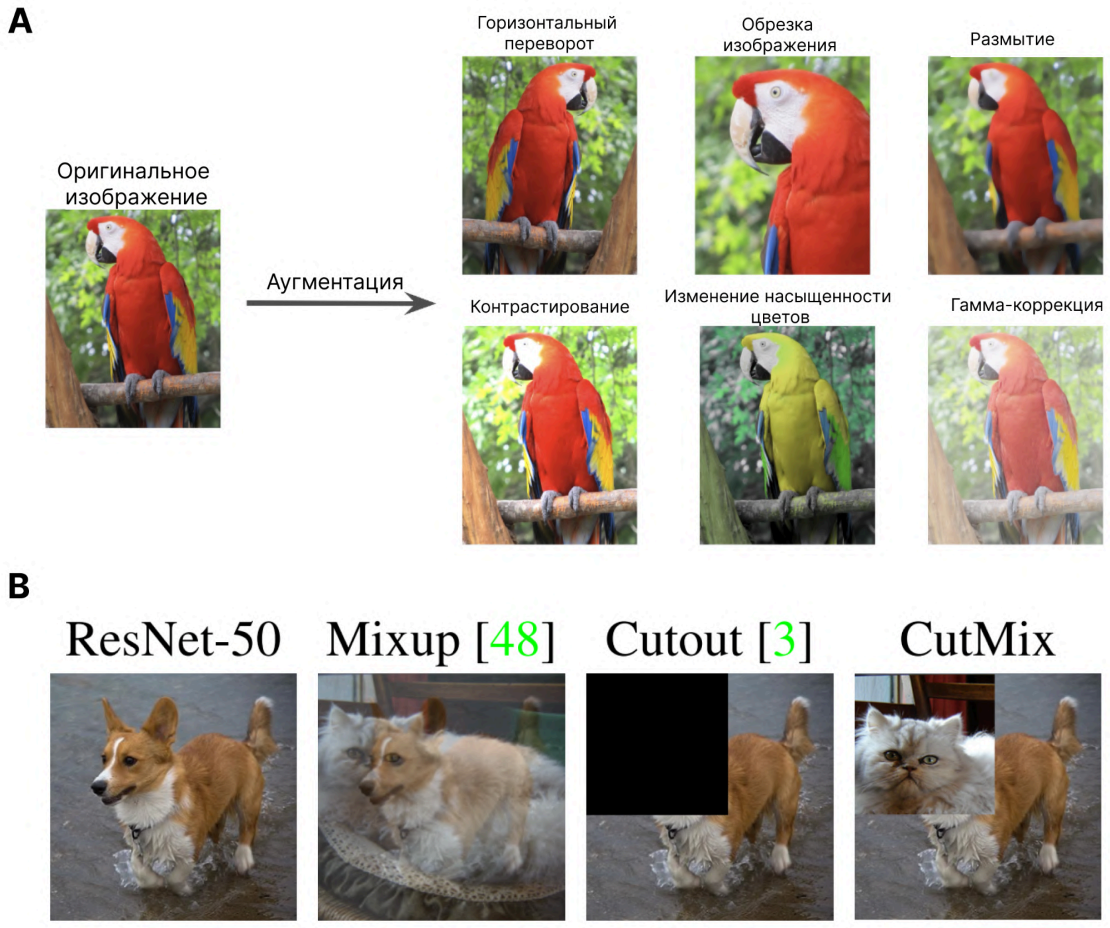


Рисунок 14. А. Классические аугментации, применяемые для изображений, которые преобразуют их в изображения, похожие на встречающиеся в реальности. Адаптировано из [237] **В.** “Нетипичные” аугментации, которые создают из изображений объекты, не встречающиеся в реальности. Адаптировано из [236].

В случае работы с нуклеотидными последовательностями аугментация также широко используется [55] и предлагаются новые методы аугментации, специфичные для нуклеотидных последовательностей [238,239]. Наиболее часто используется преобразование последовательности в обратно комплементарную ей [55,240], что является валидным, так как многие биохимические разметки генома либо инвариантны к ориентации последовательности (например, факторы

транскрипции, как правило, взаимодействуют с двухцепочечной ДНК), либо эксперимент, использовавшийся для получения данных, не позволяет однозначно установить ориентацию (**рис. 15 A**).

Также часто используется аугментация за счет сдвигов – когда известны фланкирующие участки по отношению к последовательности, для которой необходимо сделать предсказание – и можно сдвинуть последовательность в одну или другую сторону на несколько нуклеотидов, что также позволяет включить дополнительные объекты в обучающую выборку, так как предполагается маловероятным то, что существенный сигнал содержится в краевых участках анализируемой последовательности [55] (**рис. 15 B**).

Одним из недавно опубликованных подходов для аугментации является EvoAug, использующий аугментации, “вдохновленные” мутационными процессами, которые встречаются в реальности [238] (**рис. 15 C**). Этот подход является достаточно спорным, так как, к примеру, аугментация на основе точечных мутаций сообщает модели информацию о том, что одиночные замены не важны, что потенциально может приводить к ухудшению качества предсказания моделью таких замен в реальности. Авторы предлагают решать данную проблему при помощи последующего дообучения на исходных данных, однако не демонстрируют валидности данного подхода, в частности, не тестируют обученные таким образом нейронные сети на данных об однонуклеотидных заменах. Более биологически обоснованный подход на основе эволюции применяется в работе [239], в котором последовательность заменяется на гомологичную (**рис. 15 D**).

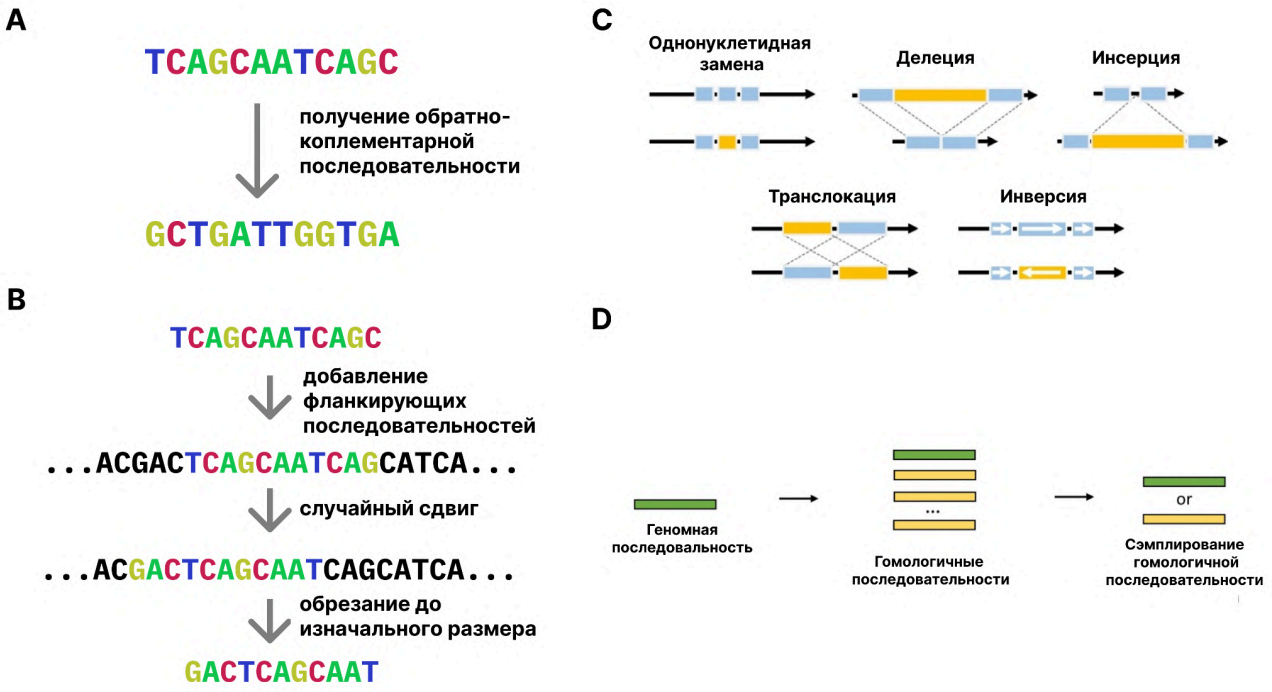


Рисунок 15. A. Аугментация путем замены последовательности на обратную комплементарную. **B.** Аугментация за счет сдвига: к целевой последовательности добавляются окружающие ее участки, немного сдвигаем окно, в котором собираемся делать предсказание и обрезаем последовательность до исходного размера. **C.** Аугментации, вдохновленные эволюционными процессами. **D.** Аугментация за счет замены исходной последовательности на гомологичную последовательность из межвидового выравнивания.

2.10.11. Аугментация во время предсказания

Помимо использования аугментации во время обучения модели, можно использовать аугментацию и во время предсказания для получения нескольких предсказаний для одного объекта. Далее эти предсказания можно усреднять или же использовать распределение полученных предсказаний для оценки уверенности модели [241]. В случае нуклеотидных последовательностей чаще всего используется усреднение предсказаний для прямой и обратной комплементарной цепей, что часто улучшает корреляцию предсказания с реальной величиной [55,72]. Аналогично можно использовать и усреднение предсказаний для небольших сдвигов последовательности [55].

2.10.12. Эквивариантные слои

В области применения машинного обучения в биологии часто возникают попытки как-то модифицировать базовые блоки нейронных сети для того, чтобы учесть в них знания из предметной области. Как правило, эти попытки не приводят к значимому увеличению качества или же оказывается, что того же качества можно было добиться без них. К примеру, AlphaFold2 [242]

использует большое количество блоков, модифицированных под задачу предсказания структуры белка, в то время как в развитии метода – AlphaFold3 эти блоки оказываются заменены на стандартные, а в обучение модели вносится большее количество аугментаций [6].

Аналогично, в области регуляторной геномики также предпринимаются такие попытки, самая заметная из которых – попытка учесть в архитектуре сети то, что часть сигналов инвариантны к ориентации цепи, при помощи **эквивариантных слоев** [227,243,244]. Учет инвариантности может осуществляться разными способами, самым простым из которых является использование эквивариантных сверток, которые применяют к прямой цепи веса без модификации, а к обратной – обратно-комплементарные веса (**рис. 16**), а затем либо конкатенируют, либо агрегирует ответы для обоих цепей.

Тем не менее, как показано в работе [240], использование обратно-комплементарной аугментации во время обучения сети и получения ее предсказания работает не хуже или лучше подходов на основе эквивариантности.

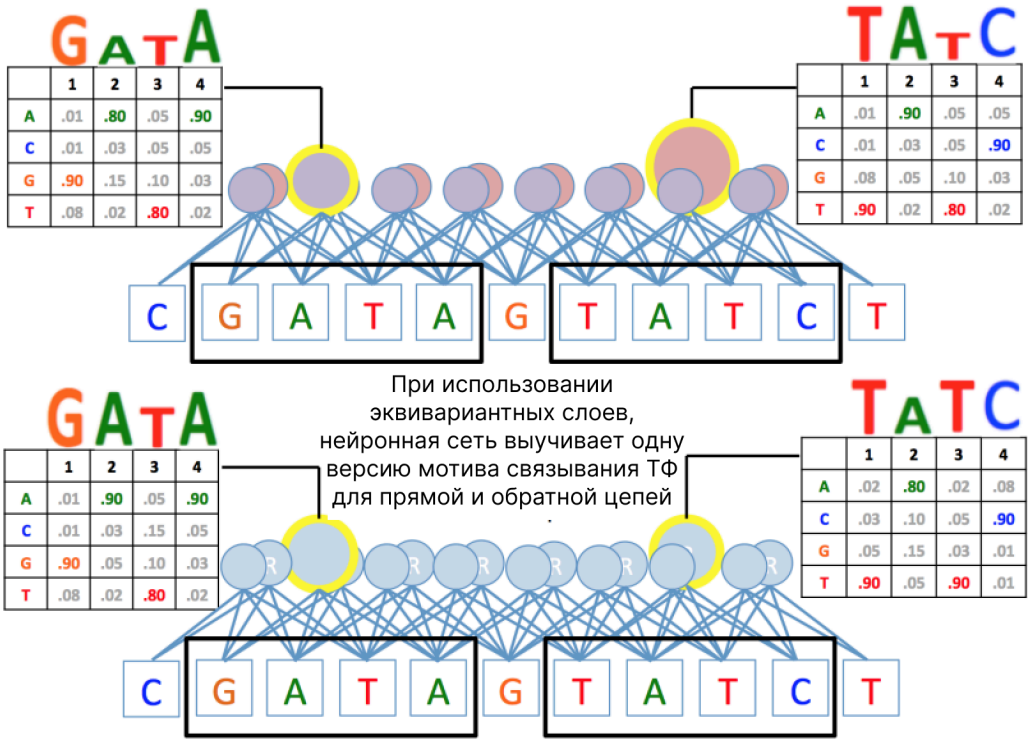


Рисунок 16. Пример работы слоя свертки эквивариантной к направлению цепи ДНК, Изображение адаптировано из [243]

2.11. Полногеномные нейросетевые модели

2.11.1. DeepSEA и ее модификации

Одной из первых нейронных сетей, которую стали широко использовать в геномных исследованиях, стала нейросеть DeepSEA [53].

Данная нейросеть была обучена предсказывать по нуклеотидной последовательности длины 1000 п.н. результаты 919 экспериментов, проведенных в рамках проекта ENCODE [245] по определению ДНКазной доступности участков связывания транскрипционных факторов и других геномных признаков, измеренных для разных типов клеток в разных условиях.

Для предсказания эффектов регуляторных замен в последовательности получали предсказания нейросети для последовательности с заменой и без нее, а затем на них и разнице между предсказаниями (абсолютной и логарифмической) тренировали модель градиентного бустинга, показавшую хорошее качество для данной задачи.

В дальнейшем авторы обучили нейронную сеть Beluga на большем окне (2000 пар оснований, и использовали ее в составе фреймворка ExPecto для предсказания эффектов регуляторных мутаций при помощи градиентного бустинга[197].

Последней на данный момент моделью, использующей данный подход, является Sei, обученная на данных 21 тысячи экспериментов о доступности хромтина и связывания факторов транскрипции, и использующей окно в 4000 пар оснований. Авторы демонстрируют, что предсказания сети можно использовать как эмбеддинги последовательности в пространстве и похожие по свойствам регуляторные последовательности группируются друг с другом. В частности, показывается, что расстояние между исходной и мутированной последовательностями является хорошей оценкой значимости мутации [57].

2.11.2. Basset

Глубокая сверточная сеть с достаточно простой архитектурой из нескольких сверточных слоев, выход которой подается напрямую в последовательность из нескольких полносвязных слоев [54] . Как и в случае с предыдущим семейством нейросети, используется преимущественно для предсказания эпигенетических треков. До сих пор часто применяется в регуляторной геномике [69,246,247] несмотря на моральное устаревание и неоптимизированность архитектуры.

2.11.3. Сравнение задач классификации и регрессии для предсказания эпигенетического профиля

При работе с геномными данными о доступности хроматина долгое время был широк принят подход, сводящий задачу предсказания активности участка к задаче классификации “активных” и “неактивных” районов. В качестве положительного класса выделялись активные регуляторные участки на основании некоего произвольно выбранного порога, а затем дополнительно формировалась негативная выборка неактивных районов генома.

Хотя этот подход позволяет создавать модели, более устойчивые к шуму [175], при наличии достаточно количества данных можно подойти к задаче предсказания доступности хроматина и с точки зрения регрессии – например, предсказывать число прочтений, которые легли на данную позицию в геноме или на данный отрезок (бин) [56].

Несмотря на то, что первый подход все еще используется [57], подробный анализ в [123] показывает, что модели, обученные в регрессионной постановке почти всегда показывают качество лучше, чем их классификационные аналоги (**рис. 17**).

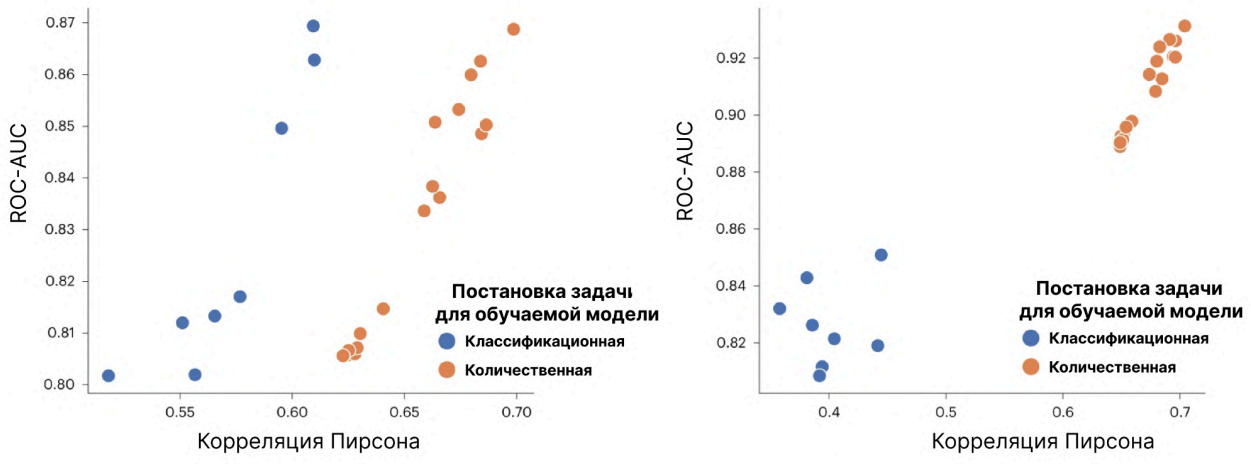


Рисунок 17. Сравнение регрессионных и классификационных моделей для решения задач предсказания полногеномных сигналов. Показывается, что в любой постановке регрессионные модели показывают лучшее качество.Адаптировано из [123].

2.11.4. Basenji и Basenji2

Одними из первых успешных регрессионных моделей можно считать полностью сверточные модели Basenji [56].

Данная модель предсказывает число прочтений в каждом из 4000 полногеномных экспериментов для каждого из бинов (непересекающихся подпоследовательностей) размера 128 нуклеотидов переданной на вход последовательности. За счет использования дилатированных

сверток, она принимает на вход последовательности существенно большего размера, нежели аналоги – 131 тысяч пар оснований.

При этом стоит отметить, что для предсказания конкретного бина используется информация только из окна размером 20 тысяч п.о вокруг него, т.е. архитектура не до конца эффективно использует переданную ей последовательность.

Вторая версия модели использует уже 5000 эпигенетических экспериментов для генома человека и добавляет результаты 1600 эпигенетических экспериментов, проведенных для мыши, что, как показывают авторы, дополнительно улучшает качество итогового предсказания на человеке [248].

2.11.5. Enformer

Долгое время state-of-the-art геномной моделью для работы с регуляторными участками ДНК считалась архитектура Enformer. Авторы берут за основу постановку из Basenji, но в исходную архитектуру добавляются слои внимания, которые теоретически увеличивают рецептивное поле модели и позволяют ей, в отличие от предшественника, извлекать всю информацию из переданной на вход последовательности (**рис. 18 А**) [55].

Помимо прочего, авторы показывают, что предсказание эффекта мутации можно делать не только путем последующего дообучения модели на нужную задачу, но и при помощи усреднения сигнала по наиболее похожим на целевую клеточным линиями или при помощи усреднения по всем экспериментам. Из-за того, что число параметров Enformer больше, чем у Basenji2, возникает вопрос, действительно ли дополнительное качество берется от использования слоёв внимания и бОльшего рецептивного поля или же от просто большего размера модели. В пользу последнего говорят результаты исследования [10], где показывается, что Enformer превосходит Basenji2 на всех подаваемых размерах последовательностей.

Также стоит отметить, что для дообучения модели на задачу предсказания экспрессии и эффектов однонуклеотидных замен авторы используют в качестве дообучаемой модели Lasso-регрессию и случайный лес, что существенно уменьшает вероятность переобучения итоговой модели и демонстрирует хорошие результаты на тестовых наборах.

2.11.6. Borzoi

Дальнейшее развитие модели Enformer – Borzoi [32] (**рис. 18 В**) – работает с большим входным окном (524 тысяч пар оснований) и предсказывает результаты экспериментов в более высоком разрешении, используя бины размера 32 п.н вместо 128. Кроме того, Borzoi учится предсказывать результаты экспериментов RNA-Seq, что увеличивает выборку и позволяет ему

лучше предсказывать эффекты eQTL, нежели Enformer. Более того, модель частично выучивает влияние мутаций на полиаденилирование и сплайсинг.

Интерес, помимо прочего, представляет подход, который используется авторами для того чтобы не увеличивать слишком критично размер модели и требования по памяти при обучении и предсказании – бОльшая часть модели все еще работает с разрешением 128 пар оснований, но в конце разрешение увеличивается до 32 пар оснований при помощи U-Net подобной части [249].

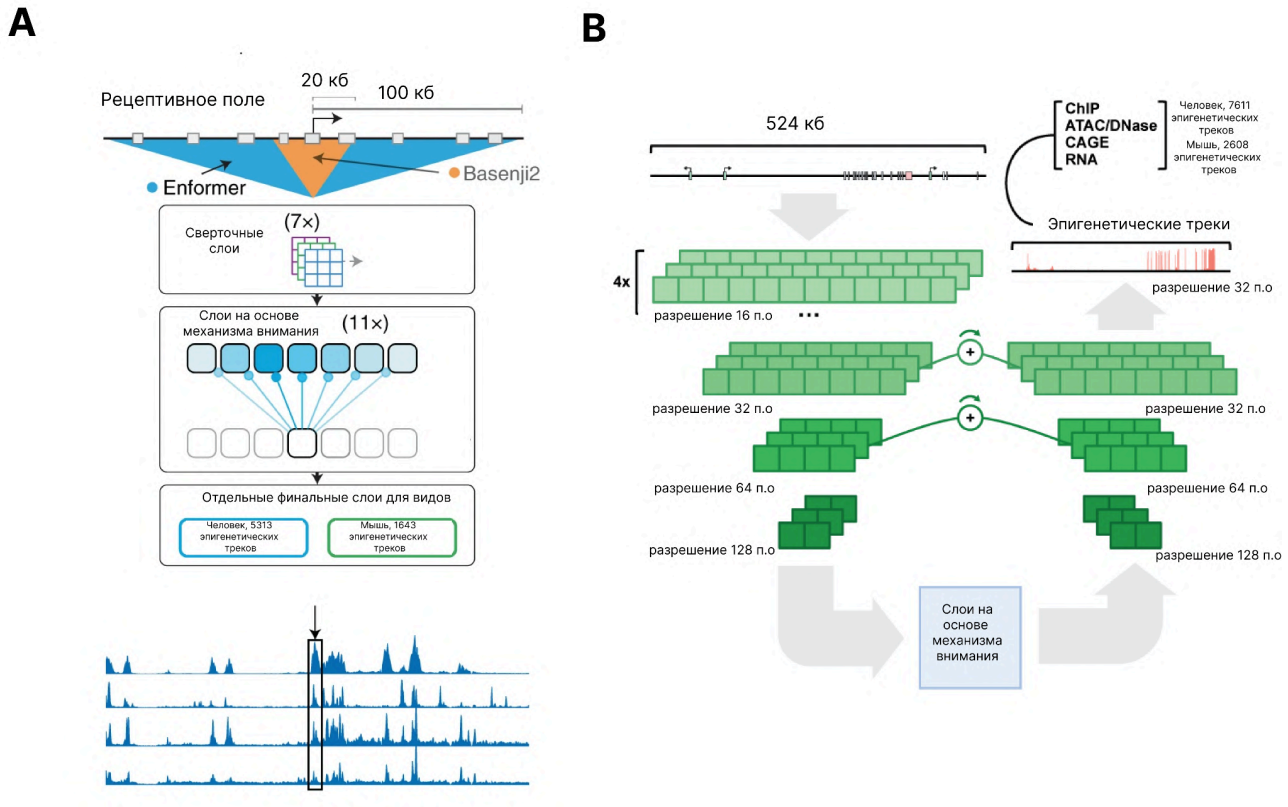


Рисунок 18. **A.** Общая схема работы Enformer. Адаптировано из [55]. **B.** Архитектура Borzoi. Адаптировано из [32].

2.11.7. BPNet

Клеточные типы очень сильно отличаются по активным в них транскрипционным факторам и правилам регуляторного синтаксиса [250]. При этом наличие общих закономерностей в решаемых задачах является необходимым условием того, чтобы мультитаргетное обучение работало лучше, чем обучение отдельных моделей на каждую задачу. Потому возникает вопрос о его пользе на данном уровне понимания регуляторной грамматики и имеющихся экспериментальных данных. Более того, некоторые авторы предполагают, что обучение моделей сразу на большом количестве экспериментов для разных клеточных линий закладывает в эти модели ложные закономерности, которые не могут быть потом убраны дообучением или другими методами, что ухудшает предсказание таких моделей (особенно в случае сигналов, специфичных для конкретных типов клеток) и затрудняет интерпретацию их предсказаний[250].

В связи с этим актуальна разработка архитектуры моделей и подходы для их обучения на небольшом числе экспериментов из одного источника. Среди таких работ можно отметить полностью сверточную архитектуру BPNet [200] и ее продолжение для работы с доступностью хроматина, ChromBPNet [251], которая используют отдельную подмодель для выучивания систематической ошибки экспериментов, вызванной предпочтениями используемых для фрагментации ДНК ферментов (bias model), и отдельную для выучивания биологически обусловленной части сигнала (accessibility model).

На данный момент в литературе нет достоверных данных о том, действительно ли обучение только на одной клеточной линии дает какие-то плюсы в сравнении с одновременным обучением на большом массиве информации из множества экспериментов, но мало свидетельств и обратного.

2.12. Проблемы современных предсказательных моделей

Недавние исследования свойств имеющихся нейросетевых решений для работы с регуляторными регионами демонстрируют ряд проблем, свойственных для полногеномным моделям.

Так, Enformer при проверке на персонализированных геномах показывает качество предсказания значительно хуже, чем методы, основанные на линейной регрессии. Для существенной части генов предсказанные им значения значимо отрицательно коррелируют с реальными изменениями экспрессии (**рис. 19 A**). Авторы демонстрируют, что эффект сохраняется и если заменить Enformer простой сверточной нейросетью [9].

В статье [13] показывается, что Enformer (как и простая сверточная архитектура на основе Basset) плохо предсказывает участки откытого хроматина, специфичные для определенных типов клеток (**рис. 19 B**). Авторы предпринимают попытку объяснить это мультитаргетной постановкой задачи для модели и демонстрируют, что для простой архитектуры обучение индивидуальных моделей улучшает качество. Однако используемая авторами для демонстрации данной закономерности архитектуры слишком проста и не оптимизирована, а также результаты оказываются не консистентными между разными наборами данных. Предположение же о том, что модели может не хватать параметров для выучивания специфики сразу всех клеточных линий в обучении, не подтверждается тем, что модели с большим числом весов в сравнении с Enformer не показывают устойчивого роста качества в сравнении с ним [32,62].

В работе [10] демонстрируется, что большая часть качества предсказания Enformer объясняется длиной последовательности ~ 32кб п.о, и его предсказания важности однонуклеотидных замен в энхансерах слабо коррелирует с реальными данными (**рис. 19 C-D**). На основании этого и проведения похожих тестов для Basenji2 делается вывод о том, что текущие модели практически не обращают внимание на далекие взаимодействия и не способны

предсказывать влияния нарушений в них. В других работах показывается, что модели с контекстом, не превыщающим 5кб п.о не уступают по качеству моделям с значительно большим рецептивным полем [13,252].

Добавление в обучение модели информации о хроматиновых контактах (Hi-C), чтобы помочь модели выучивать дальние взаимодействия, не улучшает значительно качество используемой в работе[225] модели.

Характерным примером того, что текущие модели неустойчиво выучивают регуляторную грамматику, является статья [11]. В данной работе вновь демонстрируется невозможность современных моделей предсказывать эффекты индивидуальных вариантов. Что более интересно, показывается, что корреляции предсказаний с реальными значениями эффектов для разных полногеномных моделей могут диаметрально отличаться (**рис. 19 E**). Более того, в работе [12] авторы демонстрируют, что это верно не только для двух моделей с разными архитектурами, но и для двух моделей с одинаковой архитектурой, но разными начальными инициализациями весов.

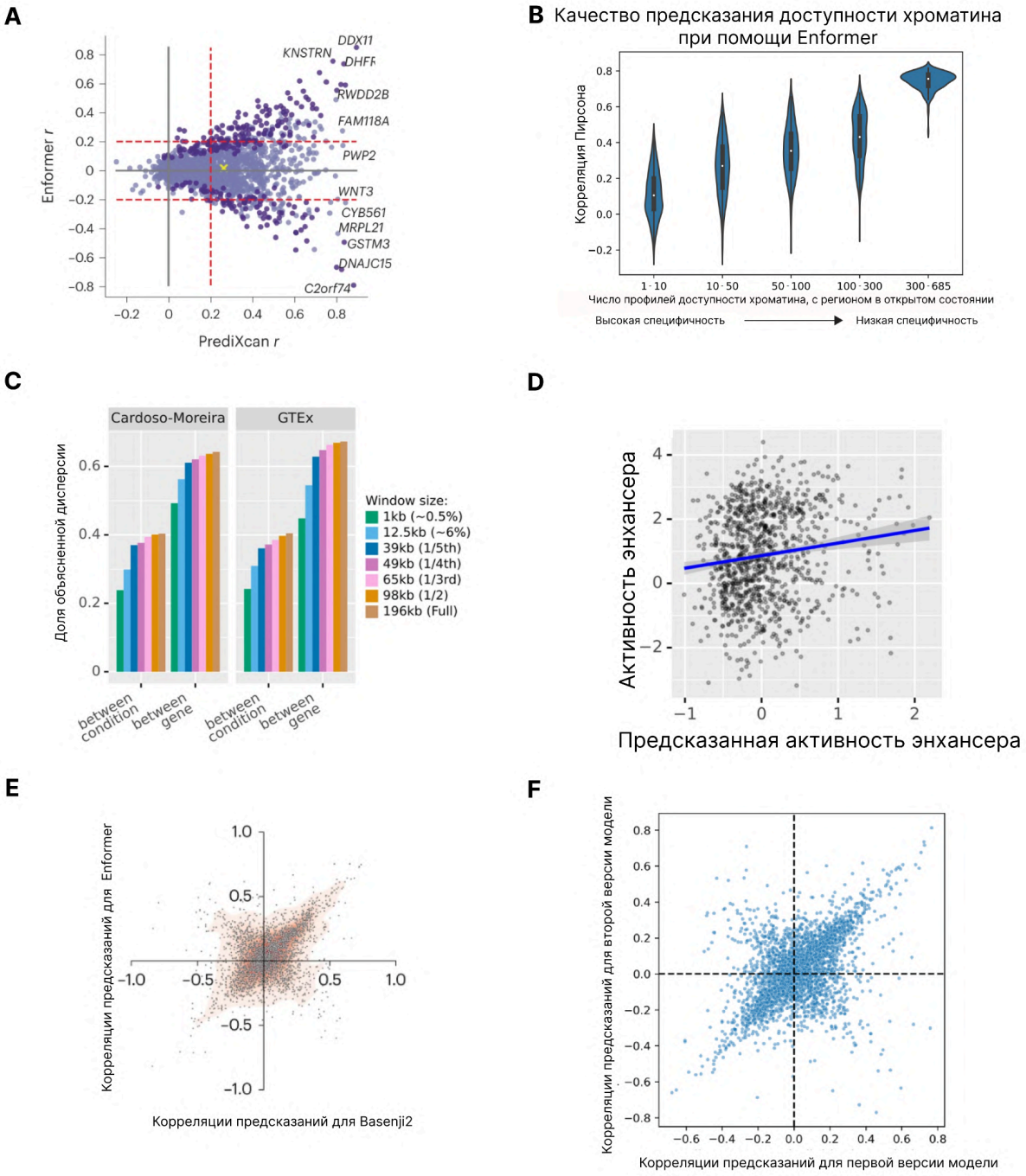


Рисунок 19. A. Для части генов предсказания Enformer отрицательно коррелируют с экспериментальными значениями. Адаптировано из [9]. **B.** Enformer плохо предсказывает пики, специфичные для клеточных линий. Адаптировано из [13]. **C.** При увеличении размера последовательности, которая подается на вход Enformer, уже на размере окна в 32тыс. п. о. наступает насыщение в доле объясненной дисперсии. Адаптировано из [10]. **D.** Enformer плохо предсказывает влияние замен в энхансерах. Адаптировано из [10]. **E.** Корреляции предсказаний двух моделей – Enformer и Basenji2, обученных на схожих наборах данных, могут иметь противоположный знак. Адаптировано из [11]. По каждой оси – корреляция предсказанных моделью эффектов мутаций и реальных эффектов для одного гена для данной модели. **F.** Подобное же наблюдается для двух моделей с архитектурой Basenji2, обученных на разных разбиениях одних и тех же данных. По каждой оси – корреляция предсказанных моделью эффектов мутаций и реальных эффектов для одного гена для данной модели. Адаптировано из [12].

Еще одним примером того, что современные модели для работы с ДНК недостаточно полно и точно выучивают регуляторную грамматику, является сравнение модели Puffin, обученной на полногеномных данных о регуляции транскрипции [138], и результатов двух позднее вышедших работ [150,210].

Последние две работы утверждают, что существует целый набор транскрипционных факторов, которые могут работать и как активаторы, и как ингибиторы транскрипции, в зависимости от их расположения по отношению к старту транскрипции. Puffin достаточно хорошо угадывает позиционно-специфичную активность в случае фактора NRF1, в случае факторов NFY и SP угадывает лишь общую форму, не угадывая размер эффектов, а в случае фактора YY1 предсказывает совсем другую форму зависимости (рис 20).

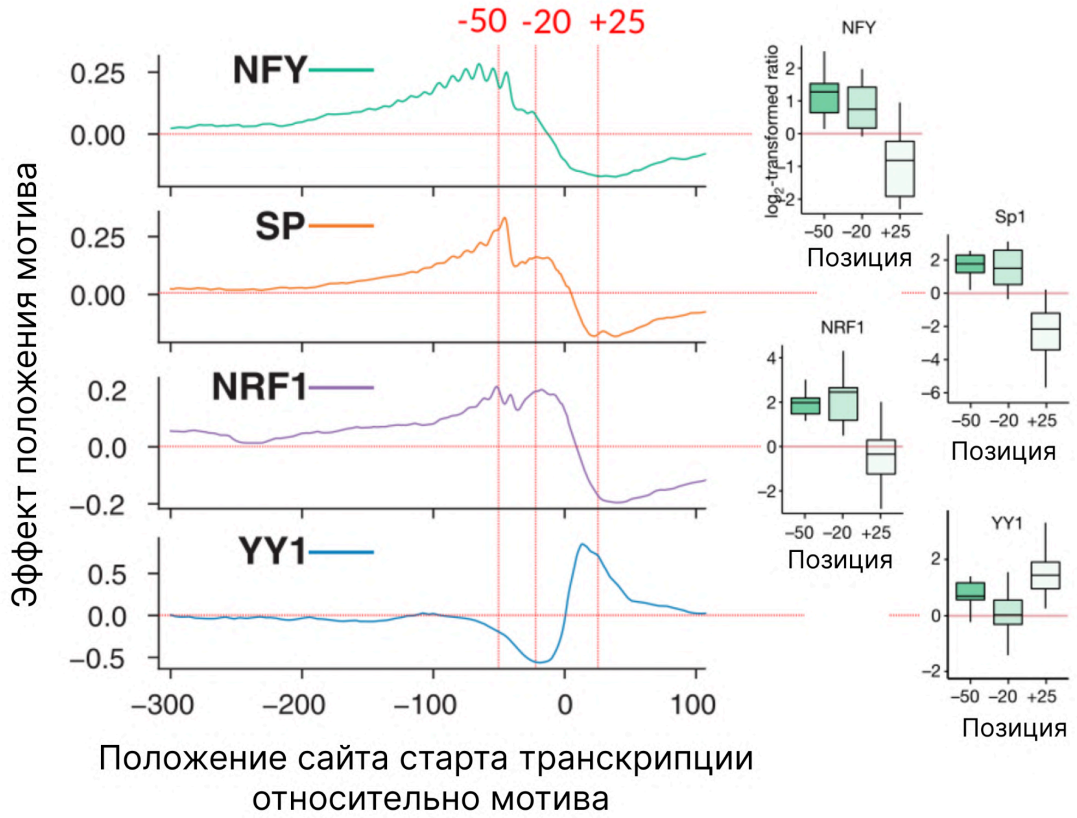


Рисунок 20. А. Согласно данным из статьи [150] фактор YY1 на небольшом удалении от сайта инициации транскрипции работает как активатор. Для факторов Sp1, NFY, NRF1 активируют транскрипцию если связываются до сайта инициации транскрипции и играют роль репрессоров, если связываются после. Изображение адаптировано из работы [150] **В.** Puffin [138] предсказывает для всех факторов YY1 и SP1 поведение, отличное от наблюдаемого экспериментально, что указывает на то, что он недостаточно точно выучил регуляторную грамматику. Изображение адаптировано из [138].

2.13. Использование специальных функций ошибки

Предсказание отличий в экспрессии между генами в одном типе клеток является более простой задачей, чем предсказание специфичной для различных типов клеток экспрессии одного гена. Еще более сложной задачей является предсказание изменений в экспрессии гена, вызванных индивидуальными вариантами [1,2].

Предполагается, что данное явление как минимум частично можно объяснить тем, что разница в уровнях экспрессии между генами в одной ткани в среднем куда больше, нежели разница в экспрессии одного гена между клеточными типами, которая в свою очередь в среднем больше изменений, вызванных индивидуальными заменами. При обучении модель, таким образом, наибольшую роль отводит первой задаче, уделяя двум другим меньше внимания [1,2].

Предпринимаются попытки учесть данную проблему введением функции ошибки, обращающей большее внимание на клеточную специфичность [67,253] или индивидуальную вариабельность [59] экспрессии. По-видимому, такой подход позволяет уменьшить выраженность проблемы, но не решает ее полностью.

2.14. Использование персонифицированных геномов

Еще одной попыткой преодолеть проблему недостаточного качества моделей глубокого обучения при предсказании эффектов индивидуальных вариантов является использование в обучении информации из персонифицированных геномов[1].

В работах [59,60] показывают, что дообучение на данных персонифицированных геномов с соответствующей измеренной экспрессией архитектуры Enformer позволяет улучшить корреляцию предсказаний с реальными значениями в случае крови и нервной ткани. Однако:

1) качество модели не превышает таковое у линейных моделей, используемых для этих задач;
2) качество модели не улучшается на регуляторных элементах, данных о заменах, в которых не было в обучении, что свидетельствует о том, что модель плохо выучила общую регуляторную грамматику;
3) качество модели на не включенных в обучение регуляторных элементах не зависит от размера выборки регуляторных элементов, включенных в обучение;
4) при обучении модели на одной популяции, ее качество резко падает при тестировании на другой, что соответствует поведению линейных моделей и связано с различием частот вариантов в популяциях и их сцеплением.

Таким образом, применение таких моделей для задач медицинской диагностики сопряжено с теми же проблемами, что и использование линейных моделей [58].

Авторы работы [59] использовали вклады каждого варианта, полученные с помощью in-silico насыщающего мутагенеза на основе дообученной нейронной модели в качестве весов линейной модели. Таким образом была получена линейная аппроксимация дообученной нейросетевой модели. При этом качество такой аппроксимации оказалось сопоставимым с качеством дообученной модели, что указывает на то, что дообучение не помогает модели выучить какие-либо нелинейные взаимодействия между вариантами.

При анализе дообученных нейросетевых моделей оказывается, что они ранжируют вклады причинных вариантов выше, чем линейная модель [60]. Отсюда можно сделать предположение, что они могут использоваться для выявления причинных вариантов. Однако в силу малого объема выборок, использовавшихся в работе [60], данное предположение стоит воспринимать с осторожностью.

Помимо прочего, в работе [60] демонстрируется сублинейный характер роста качества модели в зависимости от числа добавленных индивидов даже на включенных в обучение регуляторных элементах. Авторы работы утверждают, что рост качества, по-видимому, должен достигать некой верхней планки достаточно быстро, что ставит под сомнение возможность того, что добавление сильно большего числа индивидуальных геномов позволит улучшить предсказание – уже на порядка 500 геномах модель выходит на плато. Однако демонстрируемая авторами зависимость (**рис. 21**) может объясняться и отрицательной степенной зависимостью качества модели от числа добавленных геномов, что в машинном обучении наблюдается для множества моделей и задач и получило название scaling law (закона масштабирования) [254].

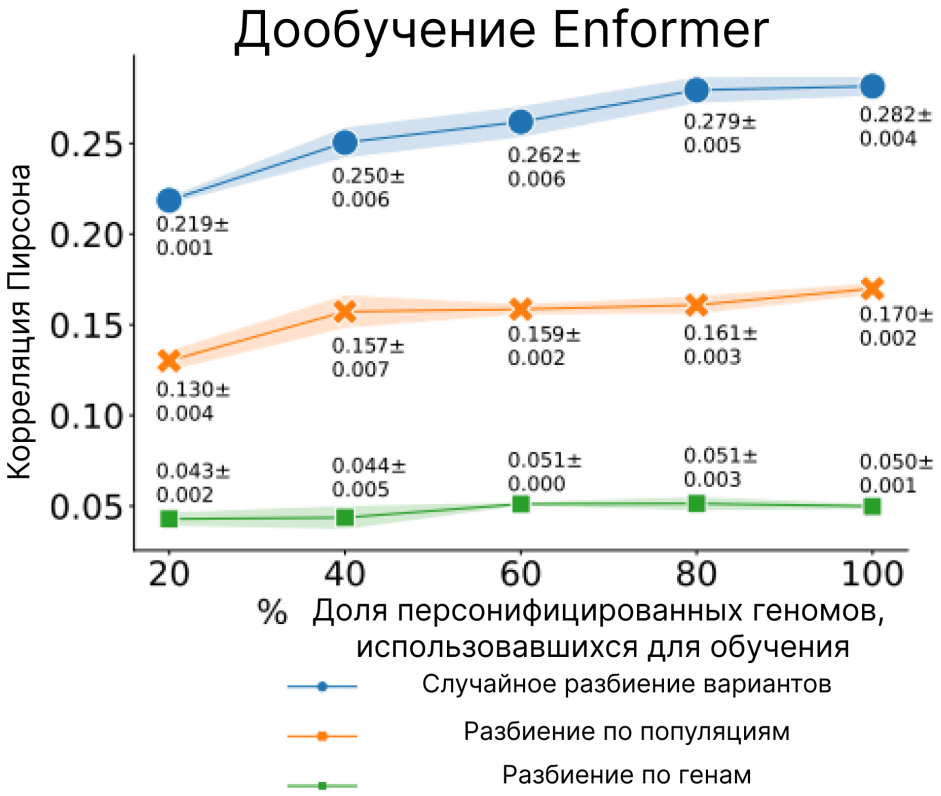


Рисунок 21. Зависимость корреляции предсказанной дообученными моделями экспрессии с реальной от размера использовавшегося для дообучения набора данных. Авторами работы[60] демонстрируется: 1) в случае случайного разбиения данных о мутациях между обучающей и тестовой выборками, без учета генов и популяции , откуда пришла мутация, дообучение модели улучшает качество предсказания(синяя кривая) 2) аналогично происходит если разбиение происходит без учета гена, но одна из популяций целиком отправляется в тестовую выборку, но достигаемое качество в целом ниже (оранжевая кривая); 3) если же разбиение происходит с учетом гена – все мутации, относящиеся к данному гену, отправляются только либо в тестовую выборку, либо обучающую – то добавление новых геномов не улучшает качества предсказания модели и оно остается околослучайным.

2.15. Использование данных секвенирования единичных клеток

Как обсуждалось в разделе 2.12, проблема с данными, на которых училась большая часть сегодняшних полногеномных моделей, состоит в том, что они фактически представляют собой сигнал, усредненный по большому числу различных клеток из гетерогенных популяций, что может мешать модели выучить общую “регуляторную грамматику” с достаточной точностью. Потому одним из возможных направлений улучшения качества предсказаний полногеномных моделей является использование данных экспериментов по профилированию единичных клеток (single cell) [65–67]. Особый интерес представляет использование т.н. “мультимодальных” экспериментов, измеряющих для одной клетки одновременно несколько сигналов, например, scATAC-seq и scRNA-seq, что потенциально позволяет точнее установить взаимосвязь между активностью регионов и экспрессию генов [66,255,256].

Действительно, в то время как полногеномные модели нашли лишь ограниченное применение для задач генерации последовательностей с требуемой клеточной специфичностью [257], и существуют данные о том, что Enformer плохо подходит для данной задачи [247], есть примеры успешного дизайна регуляторных последовательностей с заданной клеточной специфичностью при помощи моделей, обученных на данных single-cell [93].

Аналогично, в задачах, которые уже решались полногеномными моделями на приемлемом уровне, модели, дообученные на single-cell данных показывают улучшение качества предсказания [65–67].

В то же время на данный момент нет убедительных свидетельств того, что использование этого типа данных позволяет решить указанные выше проблемы полногеномных моделей.

2.16. Языковые модели для ДНК

Основные подходы к решению данной задачи основываются на обучение с учителем – модель учат по входной последовательности предсказывать целевое значение.

В качестве одной из альтернатив такому подходу предлагают предобучение моделей специальным подходом, называемым самообучением (self-supervised learning). В нем перед моделью на неразмеченных данных формулируется задача, решение которой может быть плохо применимо на практике, но заставляет модель выучить внутреннее представление данных. Например – предсказание пропущенных слов в предложении [258] или собирание пазла, полученного из картинки [259]. Затем выученное моделями представление может быть использовано для решения других задач (downstream tasks) путем дообучения этих моделей целиком или использования их внутренних представлений как признаков для других моделей.

Этот подход показывает лучшие результаты в областях обработки натуральных языков (natural language processing, NLP) [260] и компьютерного зрения [261,262]. В случае биологии он показывает хорошие результаты при работе с последовательностями белков – позволяет предсказывать эффекты вредных мутаций [263,264] или наличие сигнальных пептидов в белке [265]. Использовался он в качестве одного из компонентов в AlphaFold2 [242] – нейросетевой архитектуры для предсказания трехмерной структуры белка.

Модели, получаемые в ходе применения этого подхода к геномным последовательностям, получили общее название ДНК-моделей или языковых моделей для последовательности ДНК [61,204,226,227,266].

При обучении таких последовательностей используется референсный геном человека [266], геномы людей из проекта 1000 геномов [38,61], геномы млекопитающих, эукариот [61,204]. Данные геномы трактуются как тексты, которые бьются на куски — предложения, переводимые в последовательности токенов (нуклеотидов [226], k-меров [62], подслов (BPE) [61]), часть из

которых зануляется (маскируется), а часть меняется на случайные. Задачей модели является восстановить исходные предложения (masked language modelling, маскированное языковое моделирование) [258]. Также возможна постановка, при которой модели на вход дается последовательность токенов, и ее задачей становится предсказание следующего токена (next token prediction) [226,267] (**рис. 22 А**).

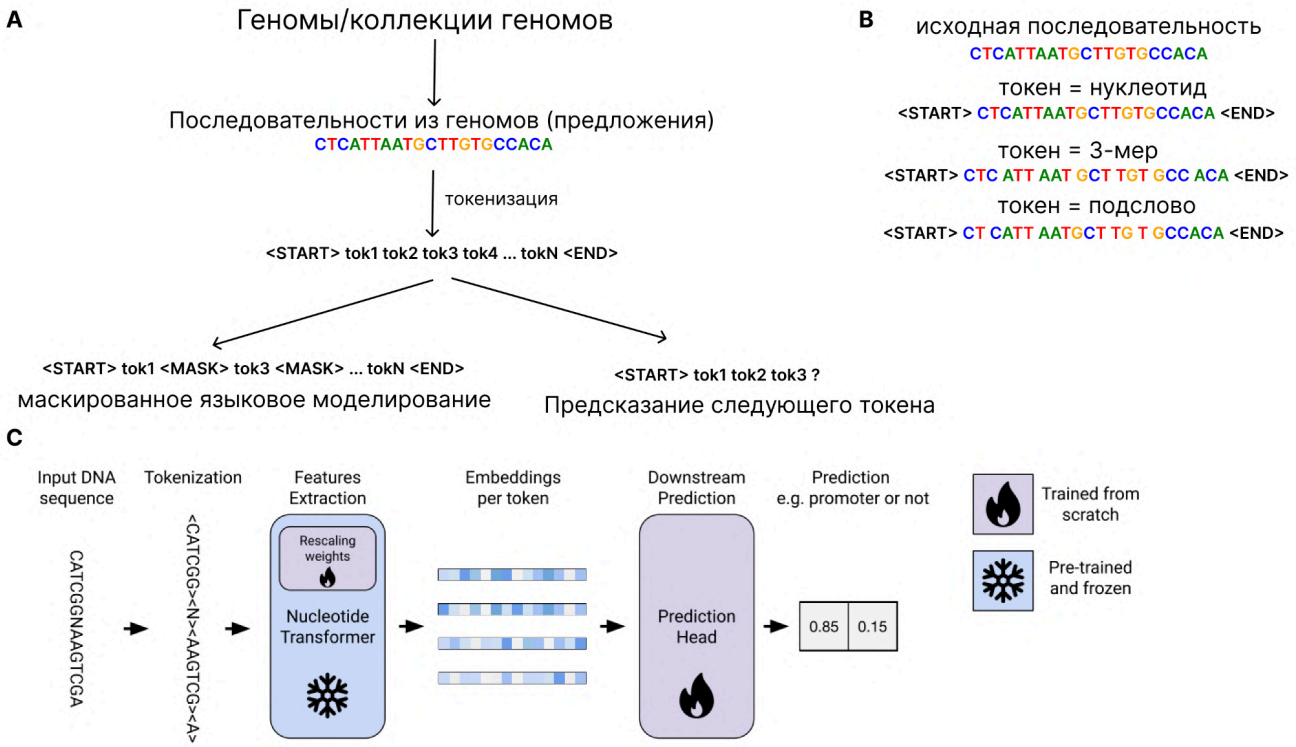


Рисунок 22. А. Схема обучения ДНК-языковых моделей. Геном или коллекция геномов бьются на отдельные последовательности. Каждая последовательность подвергается токенизации – разделении на отдельные части, которые могут представлять собой нуклеотиды, к-меры или подслова разного размера. **B.** Различные способы токенизации – разбиение на нуклеотиды, 3-меры или подслова на основе какого-то детерминированного алгоритма. **C.** Пример применения языковой модели к реальной задаче на примере Nucleotide Transformer. Эмбединги получаемые из предобученной модели используются для дообучения на новой задаче. При этом сама модель также может дообучаться либо используются методы, позволяющие дообучать только ее часть весов. Адаптировано из [62]

Предполагается, что в ходе решения подобной задачи модель может выучить внутреннее представление последовательностей генома, которое будет среди прочего содержать информацию о регуляторной грамматике.

Однако во многих работах показывается, что представления, выученные ДНК-моделями, часто работают хуже более традиционных supervised моделей [203,204] не помогают в части задач, и могут даже оказаться хуже, чем one-hot кодирование [64,204]. Дообучение ДНК-моделей также может уступать по качеству простым supervised моделям[268].

Бенчмарки, показывающие преимущество ДНК-моделей, страдают из-за слабого биологического обоснования решаемых проблем [269]. Часть из этих бенчмарков показывает результаты, противоположные ранее опубликованным с использованием тех же моделей и наборов данных. Например, в статье [227] утверждается, что ROC-AUC Enformer в задаче классификации однонуклеотидных вариантов находится в районе 0.56-0.67, в то время как в самой статье Enformer и независимом бенчмарке это качество оказывается в районе 0.74-0.75 [55,203].

Высказываются как сомнения в том, что текущие ДНК-модели учат какую-либо полезную информацию [2], так и сомнения в теоретической обоснованности такого подхода в целом – может ли сложная клеточно специфичная грамматика генома, развивавшаяся в течении более 2 млрд лет, быть выучена на основе сравнительно небольшого количества геномов, отсеквенированных на данный момент и не содержащих множество вариантов последовательностей в силу их исключительной вредности для организма и исключения из популяции естественным отбором [1,3].

2.17. Проблема недостаточного размера генома

На данный момент выдвигается идея, что полногеномных данных недостаточно для полноценного изучения грамматики регуляторных регионов [71,172]. Геном человека слишком короткий, чтобы даже теоретически содержать примеры всех возможных взаимодействий более 1500 человеческих транскрипционных факторов в большом числе независимых контекстов [71]. Более того, последовательности в геноме не являются истинно случайными, и схожесть между отдельными участками часто обуславливается гомологией, а не функциональными зависимостями, что приводит к существенно большему сходству различных геномных участков, по сравнению со случайными нуклеотидными последовательностями той же длины (**рис. 23 A**).

Допустим, задача состоит в предсказании эффекта промоторной последовательности на экспрессию гена. Последовательность промотора неоднородна – в ней есть участки как функционально значимые, так и не являющиеся таковыми. Однако в силу того, в ходе эволюции эти участки переиспользуются вместе, и в неважных не успевает накопиться достаточное количество мутаций – модель может спутать корреляцию (важный и неважный участки в силу сцепления расположены рядом) с причинно-следственной связью (оба участка важны для осуществления функции), как показано на **рис. 23 B**. В то же время при работе со случайными последовательностями такой опасности практически нет – оба участка встретятся независимо, в результате чего модель сможет правильно определить более важный (**рис. 23 C**).

Это приводит как к уже упомянутым проблемам с утечкой данных, так и и к проблемам с уменьшением эффективного объема используемых данных в силу вырожденности примеров в обучающей выборке.

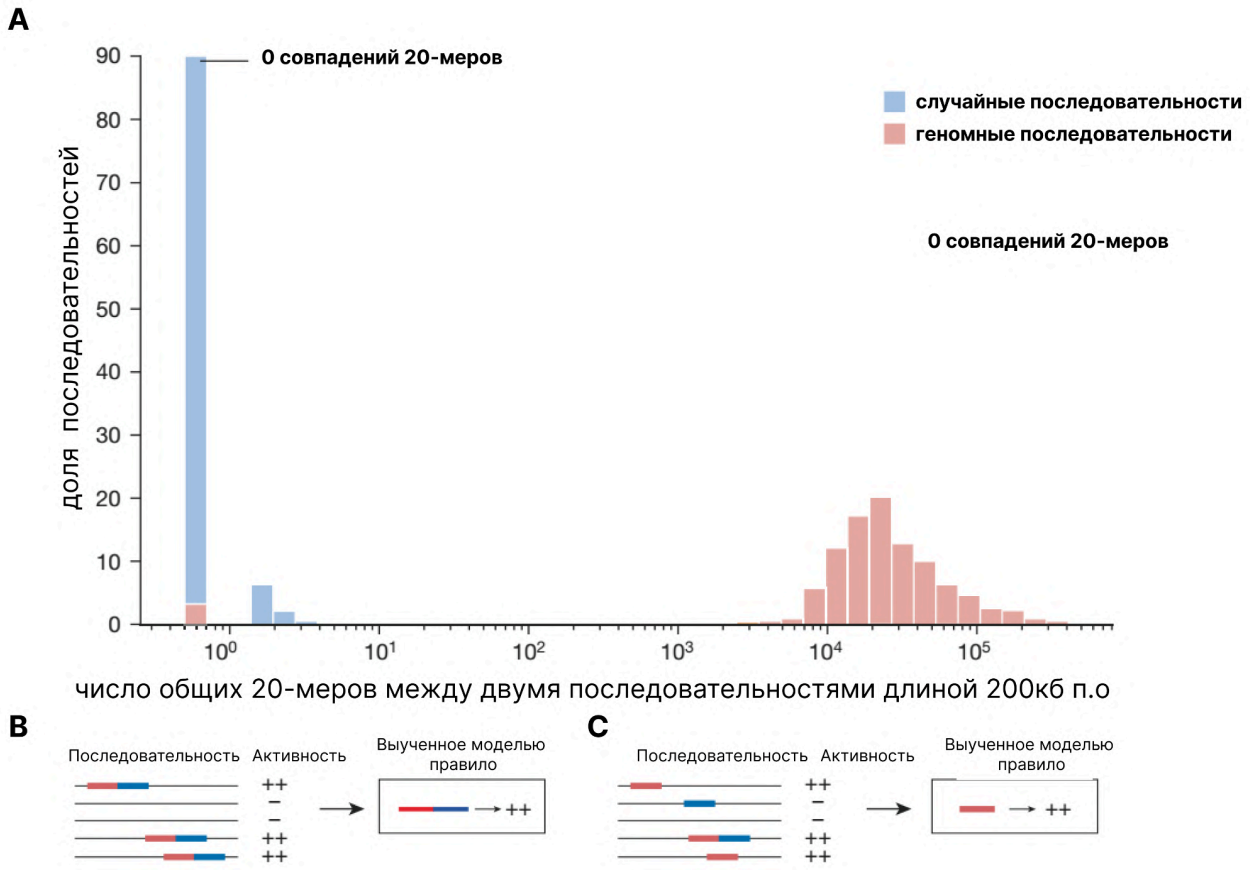


Рисунок 23. **A.** Распределение числа идентичных 20-меров в парах случайных и геномных последовательностей длины 200кб. В геномных последовательностях число совпадающих 20-меров значительно больше того, что мы бы ожидали по случайным причинам. **B.** В случае геномных последовательностей может ситуация, когда функционально важный участок (красный) сопровождается функционально неважным (синий), оставшимся практически без изменений из-за отсутствия действия на него негативного отбора. В этом случае модель, обученная на геноме может ошибочно предположить, что важно объединение этих двух участков, а не только красный. **C.** В случае со случайными последовательностями будут встречаться отдельно и важный, и неважный участки, в результате чего модель могла бы выучить правильную закономерность.

2.18. Массовые параллельные эксперименты с репортерами

Одним из подходов для решения этих проблем является использование массовых параллельных экспериментов с репортерами (МПРЭ, massively parallel reporter essays, MPRA), которые позволяют оценивать влияние на экспрессию нативных геномных, специально сконструированных и случайных последовательностей [14–17,19] (**рис. 24 А**). МПРЭ могут быть разделены на категории несколькими разными способами.

Первое разделение основывается на том, происходит ли интеграция вектора, несущего регуляторный элемент, в геном, или же измерение активности производится эписомально (**рис, 24 В**). Эписомальное измерение проводить проще, однако оно сопряжено с рядом минусов:

1) число клеточных культур, которые легко могут быть трансфецированы ограничено[270];

2) измерение активности производится в сильно отличающемся от генома контексте, что приводит к более низкой корреляции измеренной таким образом активности с активностью последовательности в ее исходном геномном контексте и к диапазону наблюдаемой активности, отличающемуся от геномного [183].

Второе разделение основывается на том, как именно происходит измерение активности регуляторных последовательностей (**рис. 24 C**). Наиболее частыми способами является:

1) измерение активности при помощи **сортировки клеток на основе флуоресценции**. В этом случае в качестве репортерного гена, экспрессия которого зависит от активности целевой последовательности, выступает флуоресцирующий белок, например YFP [68,70]. Также в этом семействе подходов часто используется второй флуоресцирующий белок, экспрессия которого не зависит от целевой последовательности. Это позволяет выполнить нормировку на независящие от регуляторной последовательности факторы – размер клетки, куда попала плазмида, общую активность в ней транскрипции и т.д. [183]. Данный подход, в силу того что оценивает активность последовательности по флуоресценции белка, требует дополнительных изменений в дизайне эксперимента и плазмиды для разделение вкладов транскрипции, стабильности мРНК и трансляции [183];

2) измерение активности последовательности на основе измерения уровня транскрипта репортерного гена [17,18,51,72,115,210,247,270–272]. При этом последовательность транскрипта не содержит всю информацию о регуляторном окружении. По этой причине используются дополнительные баркоды для того, чтобы связать транскрипт с соответствующей регуляторной последовательностью. Далее число прочтений транскрипта нормируется на число прочтений последовательности репортера, содержащего регуляторную последовательности. Полученное значение используется как мера активности регуляторной последовательности [183].

Наконец третье разделение основано на природе последовательностей, активность которых измеряется в эксперименте (**рис. 24 D)**:

1) полностью случайные последовательности. Библиотеки таких последовательностей очень легко готовить и потому подобные подходы позволяют измерять активность миллионов случайных последовательностей [68,70]. Минусом такого подхода является тот факт, что он не позволяет целенаправленно исследовать области низких и высоких значений экспрессий, изучение которых представляет наибольший практический интерес;

2) геномные последовательности. Иногда используются просто нарезанные последовательности из генома [18], однако для получения более точной информации об активности регуляторных последовательностей чаще выбирают определенную фракцию участков генома – например, открытых в целевой клеточной линии или содержащих определенную метку.

Минусом этого подхода является невозможность наблюдать активность последовательностей, которые были бы удалены из генома естественным отбором;

3) последовательности на основе известных геномных вариантов, использование которых позволяет сравнить активность референсного и альтернативного вариантов. Это может быть использовано для приоритизации регуляторных вариантов [18,42,43];

4) последовательности полученные путем внесения однонуклеотидных замен в регуляторный элемент, влияющего на экспрессию клинически значимого гена. Такой тип экспериментов называется насыщающим мутагенезом [51,115] и может использоваться для валидации методов машинного обучения. Возникают сомнения в том, может ли он быть использован для обучения моделей, предсказывающих эффекты мутаций в регуляторных участках, отличных от использовавшегося в эксперименте [183];

5) синтетические последовательности, специально сконструированные для изучение регуляторной грамматики. Так может исследоваться влияние расстояния между мотивами на взаимодействие двух ТФ или влияние числа мотивов связывания одного и того же ТФ на активность регуляторной последовательности [71,273]. Так же могут использоваться сконструированные последовательности с предсказанными свойствами, например, клеточной специфичностью [247].

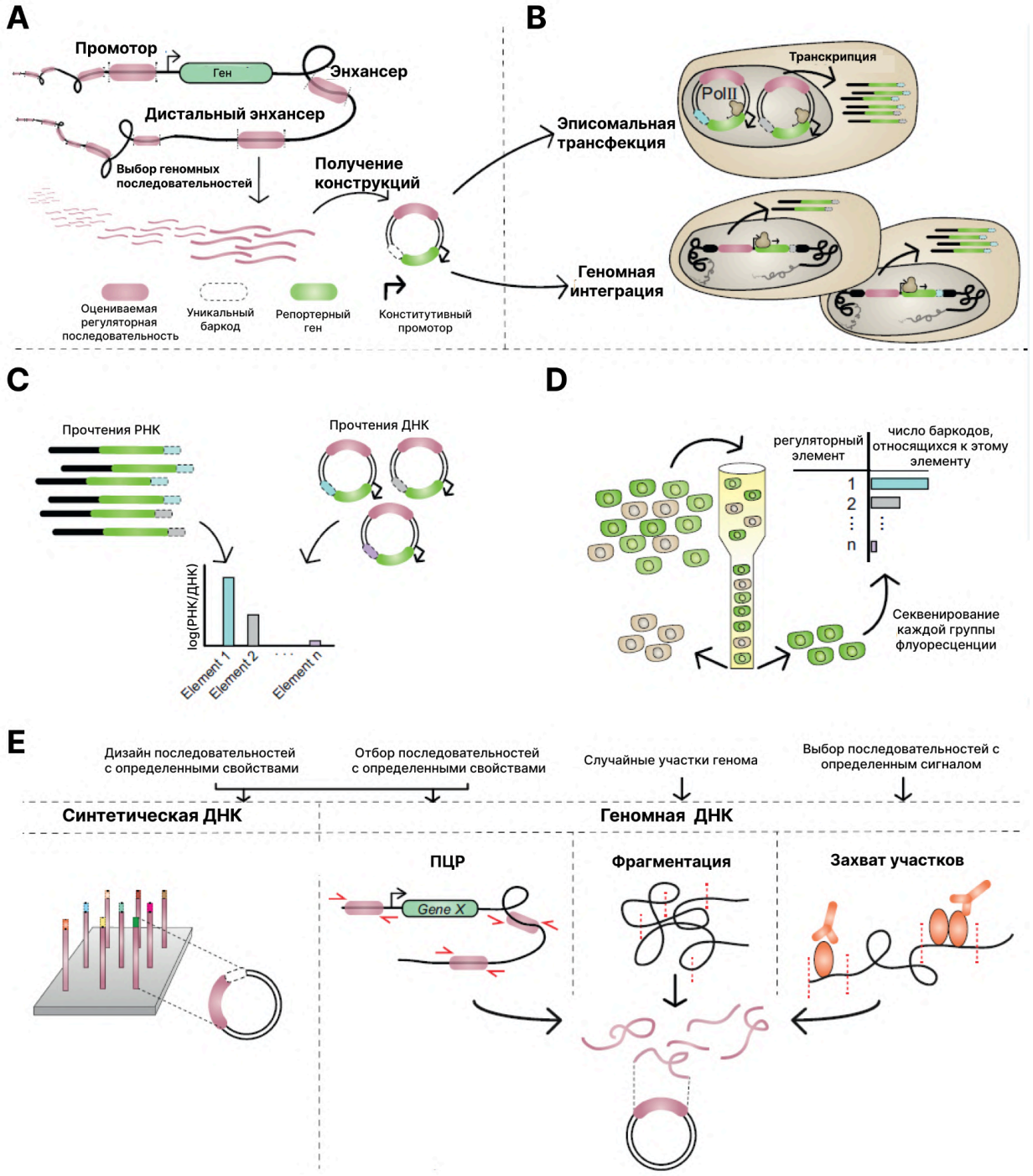


Рисунок 24. Схема массового параллельного репортерного эксперимента. Адаптировано из [274] **A.** Исследуемые последовательности помещаются в генетический вектор. **B.** Вектор далее либо трансфецируется в клеточную линию, либо же используется вектор, способный к интеграции в геном изучаемого организма. **C-D.** Измерение активности последовательности может осуществляться либо на основе измерения количества прочтений транскрипта репотерного гена (C), либо при помощи сортировки на основе флуоресценции (D). **E.** В эксперимент может измеряться активность разных последовательностей – 1) синтетических, специально сконструированных или случайных; 2) выделенных с заранее заданных участков генома в частности с использованием насыщающего мутагенеза; 3) случайные последовательности, полученные фрагментацией генома; 4) последовательности открытых участков хроматина или участков генома, содержащих определенный сигнал.

В недавних работах показано, что модели, обученные на данных массовых параллельных экспериментов, значительно лучше выучивают позиционно-специфичные регуляторные сигналы нежели модели, обучаемые на полногеномных данных [71,210].

К общим минусам массовых параллельных экспериментов стоит отнести:

1) сравнительно малый размер последовательностей которые могут быть измерены, что не позволяет модели учитывать дальние взаимодействия;
2) исследуемые последовательности находятся вне своего геномного контекста и часто дополнительно помещаются в искусственную последовательность, что может приводить к изменению их активности и опять же не позволяет учитывать дальние взаимодействия;
3) в большинстве методов используется стандартный минимальный промотор, а не тот с которым в норме взаимодействует регуляторная последовательность.

В связи с этим вероятнее всего будущее области лежит в комбинировании полногеномных данных и данных MPRA. Это косвенно подтверждается тем, что использование предсказаний модели, обученной на геномных данных в качестве входа для модели линейной регрессии, параметры которой подбираются на части данных репортерного эксперимента, значительно улучшает корреляцию на отложенной выборке между наблюдаемыми и предсказанными значениями [72].

2.19. Нейросетевые архитектуры, применяемые при работе с MPRA

На данный момент при работе с MPRA чаще всего используются простые архитектуры, такие как небольшие сверточные нейросети [10,70,72,275], Basset [69,276], или неадаптированные к задаче трансформерные архитектуры [70].

Также были попытки использовать модели, основанные на биохимических представлениях о том, как должна регулироваться экспрессия гена. Например, с использованием оценок силы связывания ТФ с регуляторной последовательностью при помощи ПВМ, однако данные модели значимо уступают даже простым нейросетевым моделям [68,70]

Можно показать (**рис. 25**) что предсказания моделей, обученных на данных MPRA, хорошо коррелируют с полногеномными данными, т.е. они по крайней мере частично выучивают грамматику общую и для генома, и для специфики эксперимента, в котором определяют их активность.

Представляет собой большой интерес развитие данного направления и разработка улучшенных архитектур, основанных на современных сверточных моделях [73–75].

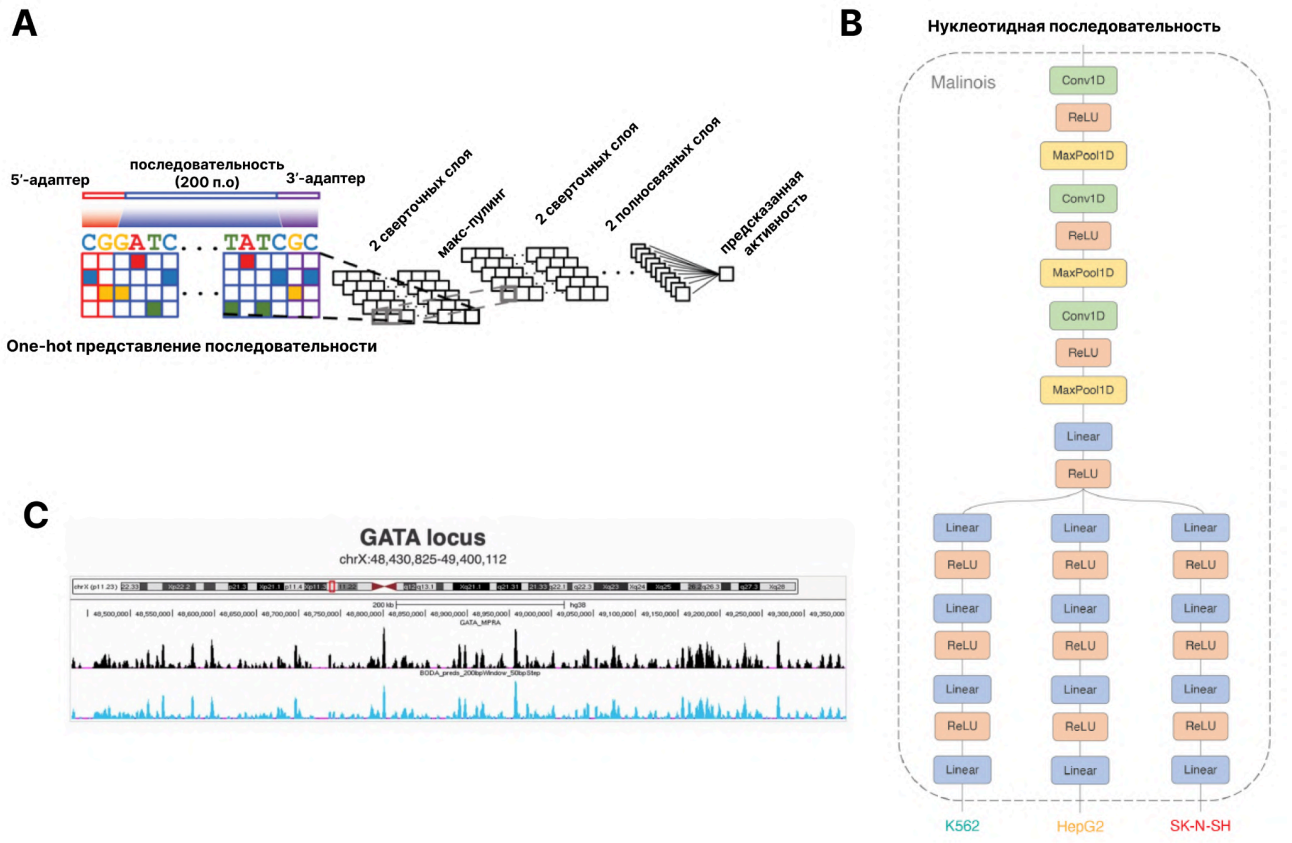


Рисунок 25. A. Архитектура сверточной нейронной сети MPRAnn из работы [72]. **B.** Архитектура нейросети Malinois из работы [276]. Используется архитектура Basset, но на каждую клеточную линию добавлены отдельные головы из линейных слоёв. **C.** Сравнение предсказания нейросети Malinois и геномных данных. Нейросеть предсказывает сигнал (RNA-Seq) за пределами промотора, который видела в ходе эксперимента [276].

2.20. Интерпретация предсказаний модели

При работе с биологическими данными важным является не только получение модели, предсказания которой хорошо коррелируют с реальностью, но и возможность объяснить предсказания данной модели с точки зрения текущих представлений о молекулярных механизмах и вовлеченных в процесс белках. В случае нуклеотидных последовательностей это может как помочь лучше понимать выученную моделью регуляторную грамматику, так и находить ложные корреляции, выученные ею. В то же время получить информацию о важности тех или иных элементов в нуклеотидной последовательности напрямую на основе весов модели не представляется возможным, так как уже отмечалось, что даже веса сверток первого слоя модели могут не выучивать биологически легко интерпретируемые признаки [208].

По этой причине создаются специальные методы, позволяющие понять, какие паттерны в последовательности были выделены моделью как важные.

2.20.1. Насыщающий мутагенез in silico

Насыщающий мутагенез in silico (in silico saturation mutagenesis, ISM, **рис. 26 А**) основан на подражании экспериментальному пути определения важности позиций в регуляторной последовательности – насыщающему мутагенезу, откуда и происходит название, и результаты которого могут совпадать с in silico аналогом для хорошо обученной модели [55,277].

В данном подходе для каждой позиции последовательности мы сравниваем предсказания модели между референсной последовательностью и последовательностью с одной из трех возможных замен в данной позиции [278]. Разница между этими предсказаниями считается оценкой важности данной замены, а из оценок важностей каждой замены можно оценить и важность позиции в целом (например, беря максимум или среднее из полученных значений). Кроме того, на основании важности мутаций в каждой позиции можно определять мотивы факторов транскрипции, которые и влияют на предсказание модели [55,72].

2.20.2. LIME

Идея LIME (Local Interpretable Model-Agnostic Explanations) [279] состоит в аппроксимации предсказаний модели в некоторой окрестности интересующей точки при помощи более простой модели (суррогатной модели), которую легко интерпретировать, например линейной модели с L1-регуляризацией или деревом решений.

Это достигается за счет следующих шагов:

1. Для интересующего объекта набирают дополнительные объекты из локальной окрестности вокруг исходного;
2. Для этих объектов получают предсказания от модели, которую требуется интерпретировать;
3. Дополнительные объекты взвешиваются на основе удаленности от исходного;
4. На предсказаниях основной модели учится суррогатная модель с учетом весов объектов.
5. При помощи этой модели делается предсказание для исходного объекта.

Данный подход страдает от неустойчивости – даже для похожих объектов он может давать очень разные объяснения и зависит от выбора изначальных параметров. К тому же в случае последовательностей не совсем понятна его польза в сравнении с in silico насыщающим мутагенезом.

2.20.3. MAVE-NN и SQUID

MAVE-NN [280] – метод получения интерпретируемых моделей на основе MAVE (Multiplex assays of variant effect) – общего названия экспериментов для оценки влияния нуклеотидных замен

в нуклеотидных последовательностях и аминокислотных замен в белках. Он позволяет учитывать как аддитивное влияние однонуклеотидных замен на итоговую величину, так и вклад эпистатических взаимодействий между разными однонуклеотидными заменами в различных позициях последовательности. В частности, метод позволяет моделировать насыщающий эффект мутаций на предсказываемую величину.

Метод SQUID (Surrogate Quantitative Interpretability for Deepnets) [281] состоит в применении MAVE-NN к данным не реальных экспериментов, а к полученным искусственно, например, при помощи in silico насыщающего мутагенеза. Таким образом он представляет собой генерализацию LIME-подхода, но, согласно утверждениям авторов, лучше адаптирован для нуклеотидных последовательностей, в том числе лучше обнаруживает сайты связывания известных факторов транскрипции.

При этом стоит отметить, что в исходной статье приводится достаточно мало значимых доводов в пользу этих утверждений. Часть доводов вызывает сомнение – например, авторы утверждают, что SQUID работает лучше, чем ISM для задачи предсказания эффектов однонуклеотидных замен (корреляция Пирсона ~0.45 vs 0.48). Авторы объясняют это тем, что суррогатная модель, используемая их подходом, “сглаживает” предсказания исходной модели, делая их более самосогласованными, что и улучшает итоговое качество. Такой эффект действительно может наблюдаться в машинном обучении при дистилляции модели – замены более сложной модели более простой [282]. Заметим, однако, что раздельно по промоторным последовательностям, исследовавшимся в работе, отличия между ISM и SQUID малы и не носят систематический характер (**рис. 26. В**).

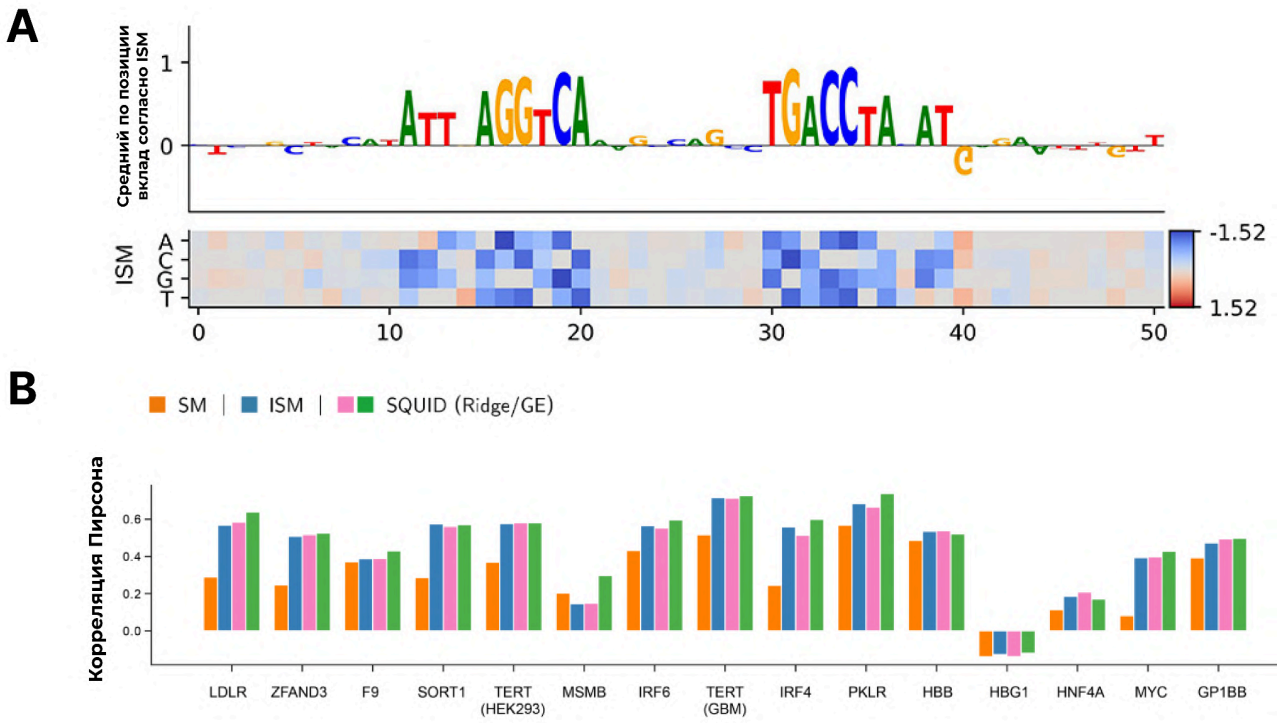


Рисунок 26. A – иллюстрация результата работы метода насыщающего мутагенеза in silico. В нижней части картинки показаны отдельные вклады замен нуклеотида на альтернативные. В верхней – величина соответствующего нуклеотида пропорциональна усредненному значение эффектов замен в данной позиции. Можно предположить, что модель принимает свое решение на основе наличия двух мотивов связывания ТФ. Адаптировано из [283] **B** – Сравнение качества различных методов интерпретации предсказаний нейронных сетей на задаче предсказания эффектов однонуклеотидных замен. Заметно, что насыщающий мутагенез in silico работает лучше или на уровне лучшего метода для каждой из исследовавшихся регуляторных последовательностей [281].

2.20.4. Карты значимости

Подход основан на вычислении градиента входа нейронной сети по ее выходу для определения того, какие участки входного объекта большего всего повлияли на итоговое решение модели. Метод впервые был применен к изображениям [284], но даже на них получаемые им значимости неустойчивы и неточно выделяют области, которые учитывает при предсказании модель.

В случае нуклеотидных последовательностей возникает дополнительная проблема за счет категориальной природы данных. Эта проблема может быть частично решена при помощи коррекции на эту природу, предложенной в работе [285].

2.20.5. DeepLIFT

Метод DeepLIFT [286] предлагает следующий способ подсчета вклада каждого признака:

1. Для каждого выхода каждого слоя нейронной сети подсчитывается "референсное" значение за счет пропускания через сеть некоего базового объекта. Для изображений это может быть, к примеру, полностью черное или белое изображение.
2. Далее через нейронную сеть пропускается интерпретируемый объект и метод сравнивает разницу между активациями нейронов каждого слоя с "референсной" активацией.
3. За счет агрегации этой разницы вдоль нейронной сети вычисляется "вклад" каждого входного признака объекта.

В случае с нуклеотидными последовательностями проблема данного подхода возникает уже на этапе выбора базового объекта — должна ли это быть случайная последовательность, полученная перемешиванием исходной, или что-то другое — метод неустойчив к выбору базового объекта.

В работе [287] предлагается модификация данного метода – DeepSHAP – за счет подсчета вкладов на основе сразу нескольких базовых объектов, что увеличивает устойчивость получаемых оценок значимости нуклеотидов.

В работе [277] обучали на данных пространственных single-cell экспериментов для неокортекса человека, мозга мыши и курицы обучили отдельные модели, предсказывающие на основе нуклеотидной последовательности регуляторного региона клеточные типы, в которых он будет активен. После этого для клеточно-специфичных регионов этих видов были оценены вклады отдельных нуклеотидов в предсказания моделей. Как показали авторы, на основе корреляции этих вкладов между разными моделями, можно судить о степени гомологии соответствующих клеточных типов.

2.20.6. Метод интегрированных градиентов

Метод [288] является развитием карт важности, а также заимствует идею базового объекта у DeepLift:

1. Вместо того, чтобы рассматривать только интерпретируемый объект, метод рассматривает путь от базового объекта до изучаемого в пространстве признаков;
2. На этом пути равномерно выбираются точки;
3. Для каждой точки этого пути вычисляется градиент нейронной сети по входу;
4. Далее за счет суммирования произведений этих градиентов на разницу между интерпретируемым объектом и точкой пути, для которой градиенты были подсчитаны, получают итоговый вклад каждого признака.

Метод обладает большей устойчивостью, нежели DeepLIFT, и дает менее шумные важности признаков, чем карты важности.

2.20.7. Выбор метода, наилучшим образом подходящего для нуклеотидных последовательностей

Несмотря на обилие разработанных методов для интерпретации предсказания моделей, в том числе и ориентированных специально на интерпретацию предсказаний моделей для работы с нуклеотидными последовательностями, на данный момент не существует работ, надежно и достоверно демонстрирующих их преимущество перед наиболее прямым способом на основе насыщающего мутагенеза in silico.

2.20.8. TF-MoDISco

Все разобранные ранее методы в случае нуклеотидной последовательности выдают значимость каждой позиции. В работе [209] представлен алгоритм, который на основе оценок важности в каждой позиции набора последовательностей выделяет наиболее часто встречающиеся паттерны в этих последовательностях и опционально позволяет проверить, совпадают ли эти последовательности с мотивами связывания известных транскрипционных факторов.

2.20.9. Интерпретируемые модели

Стоит отметить попытки разработки нейронных сетей для работы с нуклеотидными последовательностями, которые интерпретируемы по построению [138,207,289]. Однако данные модели либо на реальных задачах не показывают заявленной интерпретируемости, либо серьезно уступают по качеству моделям без подобных ограничений.

2.21. Генерация последовательностей

Одним из важных направлений работы с геномными регуляторными последовательностями является генерация последовательностей с заданными свойствами. Решение этой задачи позволит:

1) изучать регуляторную грамматику, проводя эксперименты in silico при помощи генеративной модели;
2) генерировать последовательности с заданной клеточной специфичностью и свойствами для дизайна биотехнологических конструкций и генной терапии;
3) использовать генеративную модель как часть схемы метода активного обучения, в ходе которого в нескольких итерациях сгенерированные последовательности измеряются экспериментально и используются для дообучения модели машинного обучения [290].

Существует несколько подходов к генерации последовательностей различной степени эффективности.

2.21.1. Дизайн последовательностей на основе правил

Дизайн последовательностей на основе правил не подразумевает использования методов машинного обучения, а, вместо этого, предлагает использовать знания о сайтах транскрипционных факторов и других особенностях, которые характерны для последовательностей, проявляющих заданные свойства. К сожалению, в ряде работ было показано, что полученные таким образом последовательности часто не обладают желаемыми свойствами, не воспроизводя поведения известных энхансеров, на основе знания о которых происходил их дизайн [291,292].

2.21.2. Генерация последовательностей на основе предсказаний оракула

Во второй группе подходов модели машинного обучения используются, но только на этапе оценки свойств сгенерированных последовательностей, в качестве предсказателей этих свойств – оракулов. Генерация же осуществляется другими частями вычислительного конвейера.

Самым простым подходом генерации последовательностей-кандидатов является одноэтапная генерация случайных последовательностей заданной длины при помощи генератора случайных чисел и последующая оценка их при помощи заранее обученной предсказательной модели. Этот подход несмотря на простоту позволяет, например, создавать энхансеры дрозофилы с требуемой активностью [272].

Однако в случае сложности требуемых свойств (например, избирательная экспрессия только в одном подтипе нервной ткани) нет гарантии даже при генерации большого числа случайных последовательностей (например, миллиарда как в работе [272]) получить последовательность с заданными свойствами.

Поэтому на практике чаще используется многоэтапная генерация, в основе которой, как правило, лежат идеи генетического алгоритма.

2.21.3. Генетический алгоритм

Генетический алгоритм – эвристический метод оптимизации заданной метрики на основе имитации процессов естественной эволюции (таких, как мутация, кроссинговер и отбор).

В генетическом алгоритме, как правило, присутствуют следующие этапы:

1. **Инициализация**: на этом этапе генерируются изначальные последовательности-кандидаты. В случае генерации нуклеотидных последовательностей это могут быть либо случайные последовательности, либо последовательности, случайно взятые из генома, либо геномные последовательности с известными регуляторными свойствами. Свойства текущих последовательностей оцениваются при помощи модели-оракула. Далее на основе этих свойств

считается оценка того, насколько каждая из последовательностей решает поставленную задачу, и эта оценка (или ее производные) используется как мера приспособленности последовательности.

2. **Скрещивание:** выбираются пары особей с вероятностью пропорциональной их приспособленности и между особями пары происходит скрещивание. В случае нуклеотидных последовательностей обычно обменивают участки последовательностей.

3. **Мутационный процесс:** выбираются особи с вероятностью пропорциональной их приспособленности и в их описание вводятся мутации. В случае нуклеотидных последовательностей обычно в отобранных последовательностях обычно осуществляют однонуклеотидные замены.

4. **Отбор:** из всех последовательностей отбираются последовательности с наилучшей приспособленностью. Часть последовательностей может отбираться случайным образом на основе приспособленности с целью уменьшить вероятность быстрого схождения алгоритма в локальный минимум.

5. **Терминация**: алгоритм повторяет шаги 2-4 до тех пор, пока не будет получена последовательность с необходимыми свойствами или не будет достигнут предел по числу итераций.

На стохастичность алгоритма можно дополнительно влиять за счет параметра “температуры”, который определяет, насколько сильно приспособленность последовательности влияет на ее шансы быть выбранной для скрещивания или мутационного процесса и быть отобранной на следующие шаги алгоритма.

Подход широко используется при дизайне нуклеотидных последовательностей с заданными свойствами – дизайне промоторов дрожжей, клеточно специфичных промоторов и энхансеров человека и регуляторных последовательностей растений [70,85,91,93,276,293]. В случае дизайна нуклеотидных последовательностей часто используют вариант генетического алгоритма без стадии скрещивания, называя данную разновидность алгоритма **направленной эволюцией** [70]. Также на этапе мутационного процесса может производиться вставка в последовательность мотивов связывания известных транскрипционных факторов [93]. Помимо этого, при определении, в каких позициях последовательности стоит произвести мутацию, могут использоваться значимости позиций, полученные из предсказательной модели разобранными ранее методами интерпретации предсказаний [91].

К минусам генетического алгоритма стоит отнести его склонность к попаданию в локальный минимум и сложность подбора оптимальных гиперпараметров [294]. Также, скорее всего, данный алгоритм будет плохо работать в случае дизайна последовательностей очень большого размера в силу своей переборной природы. Однако на данный момент нет исследований, показывающих, что методы на основе генетических алгоритмов как-либо уступают далее разбираемым методам,

целиком основанным на машинном обучении, в задаче генерации нуклеотидных последовательностей с заданными свойствами. Более того, в других областях биологии, таких как генерация веществ с заданными свойствами, показано, что генетические алгоритмы не уступают другим подходам [295].

2.21.4. Методы на основе градиентов (максимизация активации)

Одним из интересных подходов, который работает для нуклеотидных последовательностей, в то же время приводя к нереалистичным результатам в случае работы с изображениями [296], является генерация последовательности за счет пропускания градиентов нейронной сети до поданной на вход последовательности и изменения этой последовательности в соответствии с градиентами [297,298] с целью максимизации/минимизации предсказаний нейросети.

Дискретная природа последовательности не позволяет распространять градиент напрямую, из-за чего авторы прибегают к репараметризации с использованием распределения Гумбеля (Gumbel-Softmax). В результате оптимизируются вероятности нуклеотидов в заданных позициях, т.е., по сути, позиционно-весовая матрица. После того, как процедура оптимизации окончена, можно получить последовательности путем генерации из данной позиционно-весовой матрицы. Кроме того, метод может стартовать не со случайной позиционной-весовой матрицы, а, например, с требованием не сильно отклоняться от наперед заданной последовательности. Метод успешно используется в биологии и позволяет получать регуляторные последовательности человека с высокой клеточно специфичной активностью [276].

2.21.5. Генеративные модели

Следующим семейством методов, которые можно использовать для генерации последовательностей с заданными свойствами, являются генеративные модели на основе нейронных сетей.

Наиболее перспективными архитектурами генеративных моделей являются:

1) Генеративно-состязательные сети (GAN);
2) Модели на основе механизмов внимания;
3) Диффузионные модели.

2.21.6. Генеративно-состязательные сети

Идея генеративно-состязательных моделей (GAN) состоит в совместном обучении двух нейросетей – генератора и дискриминатора [299]. От первой нейросети – генератора – требуется создавать новые синтетические данные на основе поданного в него случайного шума,

неотличимые от настоящих. Задача же дискриминатора –отличать реальные данных от искусственных, сгенерированных генератором.

В математическом виде это записывается следующим образом:

$$\min_{\theta_g} \max_{\theta_d} [E_{x \sim p_{data}} \log D_{\theta_d}(x) + E_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))]$$

где D– – дискриминатор, G – генератор, θ – параметры нейросетей, p_{data}– распределение реальных данных, p_z – распределение случайного шума, подаваемого на вход генератору.

Обучение этих двух сетей происходит поэтапно. На первом этапе выполняется шаг минимизации по параметрам генератора, таким образом от формулы выше остается только:

$$\min_{\theta_g} E_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))]$$

1. В генератор подается случайный шум, на основе которого он генерирует синтетические данные;
2. В дискриминатор подаются сгенерированные синтетические данные, для которых он предсказывает, являются ли они настоящими;
3. Чем менее реальными считает дискриминатор сгенерированные данные, тем большая ошибка распространяется в генератор, веса которого обновляются с целью уменьшения этой ошибки. Веса дискриминатора при этом остаются неизменными.

На втором этапе выполняется максимизация по параметрам дискриминатора. Теперь в дискриминатор подаются и реальные, и синтетические данные от генератора и он штрафуется, если не может отличить одни от других. Его веса обновляются, при этом веса генератора не изменяются. Процесс повторяется до схождения ошибок генератора и дискриминатора к фиксированным значениям. В идеальном случае дискриминатор становится неспособен отличить реальные данные от сгенерированных (генератор побеждает), на практике, однако, такого результата удается достичь не всегда.

Этот метод позволяет получать безусловные GAN. Один из способов их применения для получения нуклеотидных последовательностей с заданными свойствами заключается в генерации последовательностей “похожих” на нативные или относящиеся к определенной группе (например, промоторов). Дальнейшая их оценка осуществляется дополнительно обученной предсказательной сетью. Роль же генеративной модели заключается только в генерации последовательностей, как для дизайна бактериальных промоторов в работе [86], или же для получения клеточно специфичных энхансеров человека в работе [290].

На данный момент нет свидетельств того, что такой способ использования GAN работает лучше разобранного ранее, когда последовательности генерируются просто случайным выбором или отбираются из генома.

Другой способ генерации при помощи GAN заключается в использовании предсказательной модели для оптимизации вектора, который подается на вход генератору – фактически, мы выполняем поиск в латентном пространстве генератора последовательностей с необходимыми нам свойствами [88,90].

В теории и этот подход уступает подходу, позволяющему сразу генерировать последовательности с заданными свойствами при помощи условных GAN. Для получения условной разновидности GAN во время обучения в генератор и дискриминатор подаются метки классов, к которым должны принадлежать генерируемые объекты или другие требуемые свойства [300]. Такой метод обеспечивает лучшую стабильность обучения генеративных состязательных сетей в случае, когда последовательности, принадлежащие к разным классам, очень сильно отличаются, и позволяет генерировать последовательности с необходимыми свойствами за один шаг.

2.21.7. Использования языковых моделей

Для генерации нуклеотидных последовательностей в теории можно использовать языковые модели, обученные на задаче предсказания следующего токена (Рис. Р. А.) [301,302]. В частности, в работе [95] показано, что можно использовать данных подход для генерации регуляторных последовательностей дрожжей и человека с заданной экспрессией.

2.21.8. Диффузионные модели

Изначально диффузионные модели были применены к изображениям и основываются на идеях прямых и обратных диффузионных процессов [78,303]. В ходе прямого диффузионного процесса исходное изображение постепенно зашумляется за счет добавления небольших порций гауссовского шума на каждом шаге до тех пор, пока итоговое изображение не станет неотличимо от гауссовского шума (суммарное число шагов T подбирается исследователем) (**рис 27. А**).

Задачей обучаемой диффузионной модели является восстановление исходного изображения на основе переданного зашумленного изображения и номера шага прямого диффузионного процесса, в ходе которого это изображение было получено.

Обученную диффузионную модель затем можно использовать следующим образом

1. Генерируем изначальный гауссовский шум;

Для i от T до 1:

2. Подаем его на вход диффузионной модели, говоря, что этот шум – результат i шага прямого диффузионного процесса;

3. Полученное сгенерированное изображение при помощи прямого диффузионного процесса зашумляем до шага i-1;

На выходе получаем итоговое сгенерированное изображение.

Авторы работы cold diffusion [97] показывают, что в процессе диффузии важны не тип вносимого шума, а постепенное добавление шума и процесс обучения модели при котором для каждого небольшого изменения в распределении модель запоминает, как его можно исправить так, чтобы вернуть объект в предыдущее состояние (**рис 27. В**).

Были предложены также подходы, основанные на использовании специальных распределений для моделирования диффузионного процесса в случае категориальных переменных [304,305]. В работе [257] напротив используется стандартный диффузионный процесс на латентном представлении последовательностей по аналогии с работой для генерации последовательностей в высоком разрешении [78]. Все упомянутые работы помимо прочего предлагают способы генерации нуклеотидных последовательностей с заданными свойствами на основе условных моделей.

Также для генерации последовательностей с заданными свойствами при помощи диффузии существует два подхода, специфичных для диффузионных моделей.

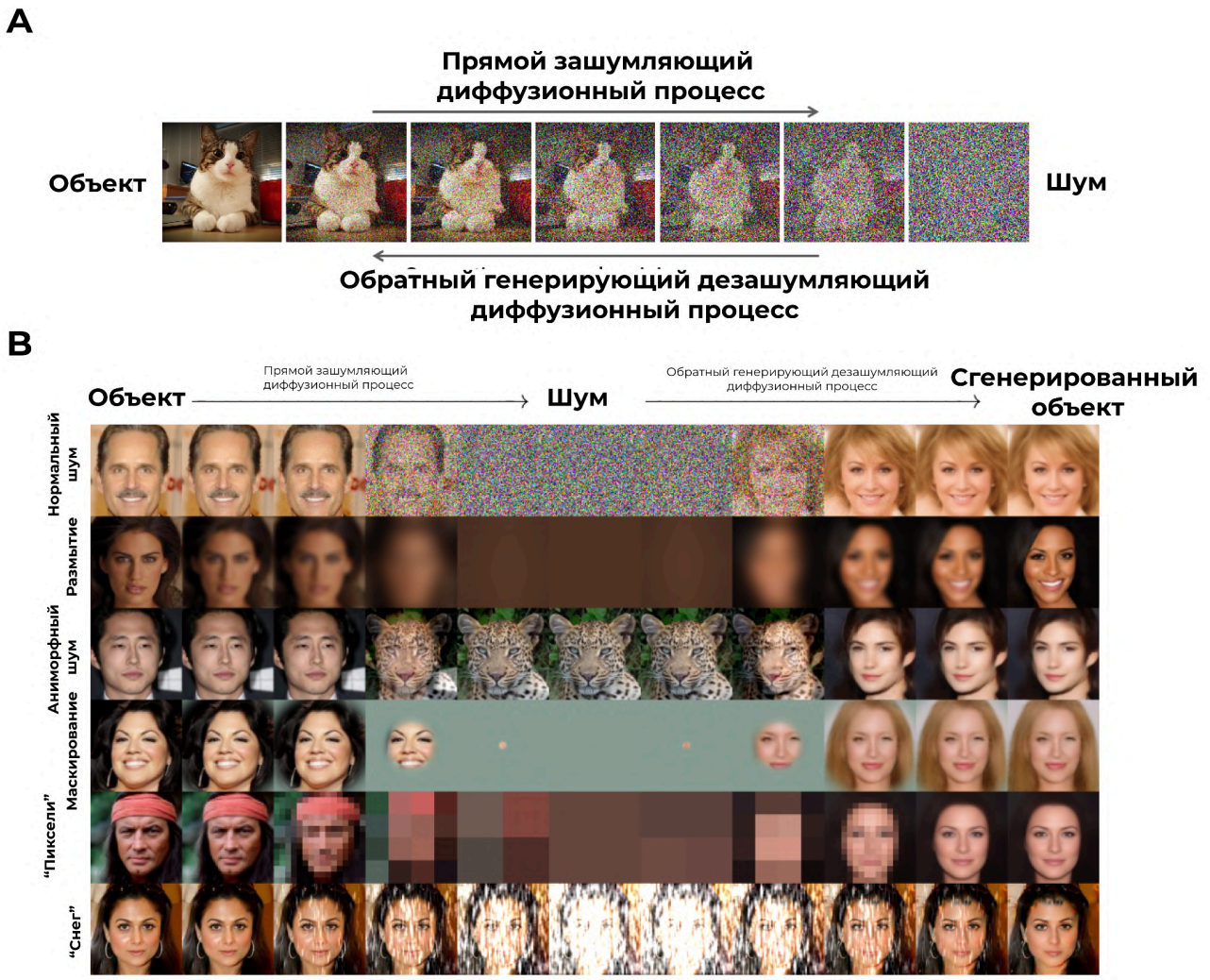


Рисунок 27. А – иллюстрация работы прямого и обратного диффузионных процессов. Адаптировано из [306]**. В** В качестве зашумляющего процесса не обязательно использовать гаусовский, другие, даже очень экзотичные также позволяет генерировать объекты.Адаптировано из [97].

Первый подход – диффузия, направляемая классификатором [307] – использует предсказательную модель для того, чтобы на основе градиентов ее предсказаний по входу модифицировать генерируемый объект. Варьируя вклад данных градиентов в процесс генерации можно балансировать требования к генерируемому объекту и его похожесть на реальные данные.
Второй подход – диффузия, свободная от классификатора – состоит в том, что в условную диффузионную модель время от времени не подается условие, за счет чего одновременно учится и условная диффузионная модель, и безусловная. Далее на каждом шаге используются обе модели, что позволяет, опять же, балансировать требования к генерируемому объекту и его похожесть на реальные данные [308].

В то время как в других задачах генерации биологических сущностей, например, белковых структур или низкомолекулярных веществ, эти технологии были достаточно подробно

исследованы [309,310], для нуклеотидных последовательностей вопрос применимости данных подходов остается малоисследованным[311].

Также представляет интерес развитие подхода для генерации нуклеотидных последовательностей при помощи генерализации диффузионных моделей – flow matching models[311].

2.22. Перспективы

На данный момент предложено множество подходов для решения задачи предсказания активности регуляторной последовательности и эффектов мутаций в ней. Однако ни одна из предложенных моделей не показывает приемлемое качество на независимых клинически значимых примерах, и потому применение этих моделей применение на практике существенно ограниченно.

Одним из возможных объяснений этого является недостаточность генома и геномных экспериментов для выучивания регуляторной грамматики и необходимость использования синтетических данных, получаемых при помощи массовых параллельных экспериментов с репортерами. В то время как количество подобных данных стремительно растёт, на текущий момент отсутствуют модели машинного обучения, специально адаптированные для работы с этими данными или их интеграции с полногеномными.

Также в последних работах отмечается возможность использования моделей, обученных на данных МПРЭ, для генерации последовательностей с заданными свойствами.

В связи с этим представляется необходимой разработка оптимизированной модели для работы с новым типом данных о регуляторных последовательностях, позволяющей эффективно предсказывать их активность, а также тестирование возможности использования такой модели в задаче генерации регуляторных последовательностей с заданными свойствами.

3. Материалы и методы

3.1. Предсказания эффектов регуляторных мутаций по данным насыщающего мутагенеза[1]

3.1.1. Результаты МПРЭ с насыщающим мутагенезом промоторов и энхансеров человека

В 2018 году консорциумом CAGI (Critical Assessment of Genome Interpretation) проводилось соревнование по использованию вычислительных методов для аннотации генома. Одной из дисциплин соревнования было предсказание эффектов однонуклеотидных замен в регуляторных участках человеческого генома. В рамках этой дисциплины участникам были предоставлены данные, полученные в ходе массового параллельного эксперимента с репортерами, и содержат информацию о результатах насыщающего мутагенеза 5 энхансеров (генов IRF4, IRF6, MYC, SORT1, ZFAND3) и 9 промоторов (генов F9, GP1BB, HBB, HBG, HNF4A, LDLR, MSMB, PKLR, TERT). Каждая из регуляторных последовательностей тестировалась в клеточной линии, избранной как модель ткани, для которой наблюдались нарушения экспрессии соответствующего гена. Промотор гена теломеразы TERT был протестирован в двух клеточных линиях.

Сводная информация об энхансерах и промоторах приведена в **таблице 1** и **таблице 2**

[1] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: **Penzar D.**, Zinkevich A.O., Vorontsov I.E., Sitnik V.V., Favorov A.V., Makeev V.J., Kulakovskiy I.V. What Do Neighbors Tell About You: The Local Context of Cis-Regulatory Modules Complicates Prediction of Regulatory Variants // Front. Genet.– 2019.– Vol. 10.– P. 1078. doi: 10.3389/fgene.2019.01078. JIF (для WoS) = 2.8, (0.70/0.40)

Таблица 1. Энхансеры, использовавшиеся в соревновании.

Энхансер	**Ассоциированное заболевание**	**Клеточная линия**
IRF4	Расстройство пигментации	SK-MEL-28
IRF6	Расщепление нёба	HaCaT
MYC	Различные виды онкологических заболеваний	HEK293T
SORT1	Избыток липопротеинов низкой плотности, инфаркт миокарда	HepG2
ZFAND3	Диабет второго типа	MIN6

Таблица 2. Промоторы, использовавшиеся в соревновании

Промотор	**Ассоциированное заболевание**	**Клеточная линия**
F9	Гемофилия B	HepG2
GP1BB	Врожденная тромбоцитопатия (синдром Бернарда-Сулье)	HEL 92.1.7
HBB	Талассемия	HEL 92.1.7
HBG	Персистенция фетального гемоглобина	HEL 92.1.7
HNF4A (P2)	Сахарный диабет взрослого типа у молодых	HEK293T
LDLR	Семейная гиперхолистеролемия	HepG2
MSMB	Рак простаты	HEK293T
PKLR	Недостаток пируват-киназы	K562
TERT	Различные типы рака	HEK293T, GBM

Таблица 3. Соответствие между реальными и нерелевантными координатами. Нумерация нуклеотидов с 1.

Регуляторная последовательность	**Реальное начало**	**Реальный конец**	**Нерелевантное начало**	**Нерелевантный конец**
F9	chrX:138612622	chrX:138612924	chr3:138612622	chr3:138612924
GP1BB	chr22:19710789	chr22:19711173	chr3:19710789	chr3:19711173
HBB	chr11:5248252	chr3::5248438	chr3:5248252	chr3:5248438
HBG1	chr11:5271035	chr11:5271308	chr3:5271035	chr3:5271308
HNF4A	chr20:42984160	chr20:42984444	chr3:42984160	chr3:42984444
IRF4	chr6:396143	chr6:396593	chr3:396143	chr3:396593
IRF6	chr1:209989135	chr1:209989734	chr3:11966705	chr3:11967305
LDLR	chr19:11199907	chr19:11200224	chr3:11199907	chr3:11200224
MSMB	chr10:51548988	chr10:51549578	chr3:51548988	chr3:51549578
MYC	chr8:128413074	chr8:128413673	chr3:128413074	chr3:128413673
PKLR	chr1:155271187	chr1:155271655	chr3:155271186-	chr3:155271655
SORT1	chr1:109817274	chr1:109817873	chr3:109817274	chr3:109817873
TERT	chr5:1295105	chr5:1295362	chr3:1295104	chr3:1295362
ZFAND3	chr6:37775276	chr6:37775853	chr3:37775275	chr3:37775853

3.1.2. Дополнительные данные о результатах МПРЭ с насыщающим мутагенезом

Для того чтобы оценить генерализуемость решений, полученных при обучении на данных конкурса CAGI5, на другие данные, мы использовали информацию из массовых параллельных экспериментов с репортерами для энхансеров генов ALDOB и ECR11 человека [312]. Конструкции для этих энхансеров повторяют конструкции для энхансеров из CAGI 2018, однако в работе [312] они трансфецировались в клетки печени мыши путем инъекции в хвостовую вену. Данные, полученные в данном эксперименте, были обработаны авторами аналогично данным CAGI 2018 и использовались как независимый набор для верификации.

3.1.3. Признаки на основе DeepSEA

В соответствии с процедурой, предложенной авторами нейросети DeepSEA [53], в качестве признаков для модели использовались 919 предсказаний DeepSEA для референсного аллеля, 919 предсказаний для альтернативного аллеля, 919 разниц в предсказаниях между референсным и альтернативным аллелем, а также e-value этих разниц и логарифм отношения предсказаний для референсного и альтернативного аллелей.

3.2. Предсказания событий аллель-специфичного связывания[2]

Для того чтобы определить, насколько возможно обучение модели с использованием признаков DeepSEA на данных аллель-специфичного связывания из опубликованной нами базы данных ADASTRA [34], мы обучали модели случайного леса с использованием признаков DeepSEA и признаков, основанных на мотивах связывания транскрипционных факторов [313].

Нами использовались данные по 10 ТФ и 10 типам клеток с наибольшим количеством, обнаруженных в нашем исследовании аллель-специфичных событий. Гиперпараметры случайного леса выставлялись по умолчанию, кроме числа деревьев, которые было выбрано равным 500.

[2] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: Abramov S., Boytsov A., Bykova D., **Penzar D.**, Yevshin I., Kolmykov S.K., Fridman M.V., Favorov A.V., Vorontsov I.E., Baulin E., Kolpakov F., Makeev V.J., Kulakovskiy I.V. Landscape of allele-specific transcription factor binding in the human genome // Nature Communications – 2021.– Vol. 12, № 1.– P. 2751. doi: 10.1038/s41467-021-23007-0. JIF (для WoS) = 14.7 (1.20/0.20)

Модели на вход подавалось 4 группы признаков (**таблица 4**), включая признаки из нейронной сети DeepSEA и признаки, полученные при помощи базы HOCOMOCO [314] и программы PERFECTOS-APE[48].

Таблица 4. Признаки, использованные для предсказания аллель-специфичных событий связывания

Тип признака	Обозначение признака	Описание признака
Аннотация вариантов при помощи мотивов из базы HOCOMOCO[314]	motif_log_pref, motif_log_palt	отрицательный десятичный логарифм p-value вхождения мотива [48] для референсного и альтернативного аллелей
	motif_fc	отношение логарифмов p-value вхождение мотивов для референсного и альтернативного аллелей
	motif_pos	позиция мотива относительно варианта
Аллель-специфичная доступность хроматина из DNase-Seq экспериментов из работы[315]	numphets_dnase	число образцов, гетерозиготных по данному аллелю
	reads1_dnase, reads2_dnase	покрытие референсного и альтернативных аллелей
	totalReads_dnase, pctRef_dnase	общее покрытие однонуклеотидного варианта и доля референсного аллеля
	qvalue_dnase	q-value аллель-специфичной доступности
	is_imbalanced_dnase_bool	проходит ли однонуклеотидный вариант порог на значимость аллель-специфичной доступности
Признаки, полученные из нейронной сети DeepSEA[53]	919 признаков для референсного аллеля	919 выходов нейронной сети DeepSEA для референсного аллеля
	919 признаков для альтернативного аллеля	919 выходов нейронной сети DeepSEA для альтернативного аллеля
	919 разниц признаков	919 разниц выходов нейронной сети DeepSEA для референсного и альтернативного аллелей
	919 e-value разниц	e-value разниц выходов нейронной сети
	919 отношений разниц	919 отношений выходов нейронной сети DeepSEA для

		референсного и альтернативного аллелей

3.3. Предсказание активности синтетических промоторов в дрожжах[3]

В 2022 году проводилось соревнование по предсказанию регуляторной активности 80 п.о. участков, измеренной при помощи массового параллельного репортерного эксперимента, методами машинного обучения.

В ходе конкурса было запрещено использовать предобученные модели и ансамбли моделей глубокого машинного обучения.

3.3.1. Данные соревнования DREAM-2022

Данные соревнования содержали данные об активности (в смысле влияния на экспрессию репортерного гена) порядка 6.7млн 80-нт промоторов в дрожжах *Saccharomyces cerevisiae*, культивировавшихся в виноградном соке сорта Шардоне. В эксперименте GPRA клетки дрожжей были трансфецированы плазмидой, содержащей два репортерных гена:

1) репортерный ген YFP (Yellow Fluorescent Protein, желтый флуоресцирующий белок), контролируемый 80-нт случайными последовательностями ДНК, вставленными в промоторный контекст

2) экспрессируемый конститутивно репортеный ген RFP (Red Fluorescent Protein, красный флуоресцирующий белок).

Клетки дрожжей при помощи клеточного сортировщика были разделены на 18 групп (бинов, 0-17) по экспрессии на основе логарифма отношения светимости белков YFP и RFP (**рис. 28**). После секвенирования каждой из этих групп для каждой последовательности промотора была получена информация о том, в каких группах она встречалась. Влиянием данной последовательности на экспрессию считалось взвешенное среднее номеров групп, где последовательность встречалась [68].

[3] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: **Penzar D.**, Nogina D., Noskova E., Zinkevich A., Meshcheryakov G., Lando A., Rafi A.M., de Boer C., Kulakovskiy I.V. LegNet: a best-in-class deep learning model for short DNA regulatory regions // Bioinformatics.– 2023.– Vol. 39, № 8. doi: 10.1093/bioinformatics/btad457. JIF (для WoS) = 4.4 (0.95/0.45). Rafi A.M., Nogina D., **Penzar D.**, Lee D., Lee D., Kim N., Kim S., Kim D., Shin Y., Kwak I.-Y., Meshcheryakov G., Lando A., Zinkevich A., Kim B.-C., Lee J., Kang T., Vaishnav E.D., Yadollahpour P., Random Promoter DREAM Challenge Consortium, Kim S., Albrecht J., Regev A., Gong W., Kulakovskiy I.V., Meyer P., de Boer C.G. A community effort to optimize sequence-based deep learning models of gene regulation. // Nature Biotechnology – 2024. doi: 10.1038/s41587-024-02414-w. JIF (для WoS) = 33.1 (1.5/0.30)

Тестовые данные в конкурсе были представлены несколькими различными категориями последовательностей:

1) случайные последовательности;

2) нативные последовательности были получены путем выбора случайных 80 п.о участков из нативных промоторов дрожжей;

3) последовательности с высокой и низкой экспрессией были получены генетическим алгоритмом [316] с использованием в качестве функции оценки модель, обученную ранее на случайных последовательностях, протестированных в качестве промоторов в клетках дрожжей из работы [68];

4) “сложные для анализа последовательности” – последовательности, для которых наиболее сильно отличались предсказания биохимической и сверточных моделей из [68] – были также получены при помощи генетического алгоритма;

5) пары последовательностей, отличающихся друг от друга одной заменой, были взяты из работы [68], а также получены внесением случайных замен в предыдущие группы последовательностей;

6) Пары случайных последовательностей, содержащие мотив связывания одного из факторов транскрипции (Reb1, Hsf1), но в консенсус мотива внесены 1-3 однонуклеотидные замены;

7) Пары случайных последовательностей, отличающиеся положением мотива связывания одного из охарактеризованных факторов транскрипции (Skn7, Mga1, Ume6, Mot3, Azf1).

Данные о составе количественном составе тестового набора данных приведены в **Таблице 5**.

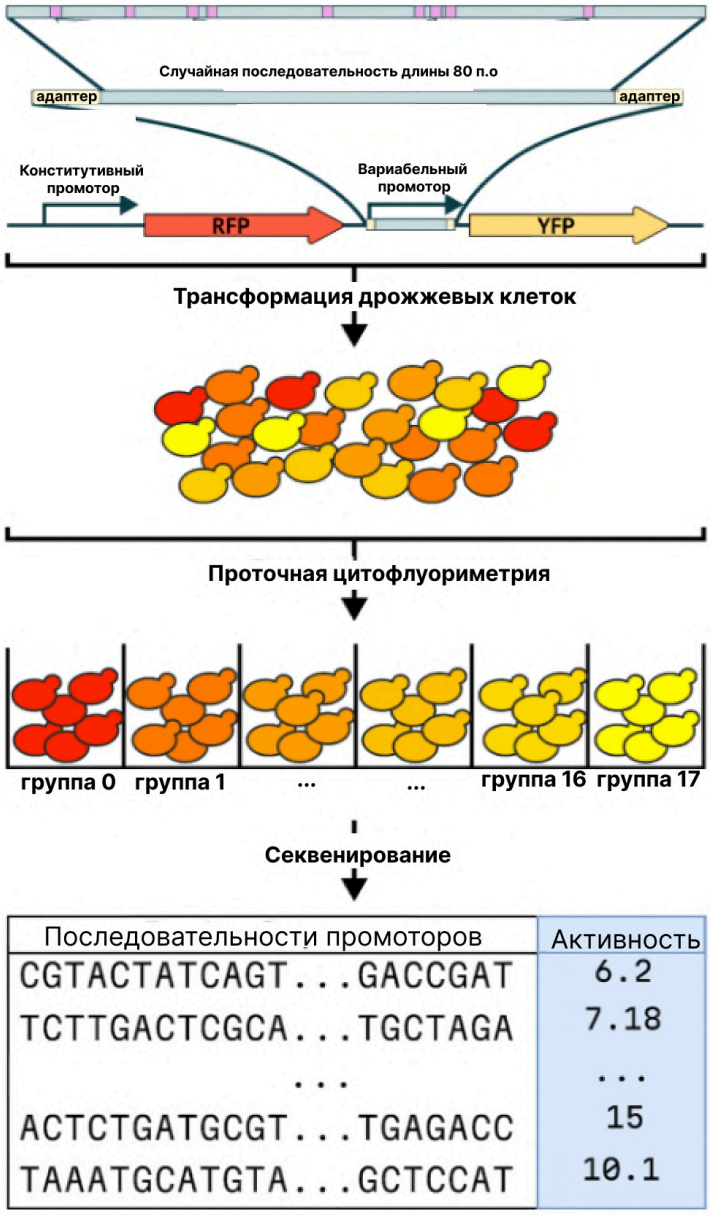


Рисунок 28. Схема гигантского массового параллельного репортерного эксперимента (GPRA). Для вставки используется вектор, содержащий два флуорисцирующих белка – RFP, под контролем конститутивного промотора и YFP, который контролировался участком, куда осуществлялась вставка. Вектор трансформировался в клетки дрожжей, которые экспрессировали красный и желтый белок в соотношении, определявшимся отношением регуляторных активностей конститутивного промотора и вставленной последовательности. При помощи клеточного сортера клетки делились по соотношению светимости YFP и RF на 18 групп. Последовательности плазмиды из клеток каждой группы секвенировались. Каждой отсеквенированной таким образом регуляторной последовательности присваивалась активность равная номеру группы. В случае, если последовательность обнаруживалась в нескольких группах, ей присваивалось значение активности, равное взвешенному среднему арифметическому. Адаптировано из [68].

Таблица 5. Количественный состав тестового набора данных в конкурсе DREAM-2022

Группа последовательностей	**Название группы в конкурсе**	**Количество**	**Вес последовательностей в метрике конкурса**
Все	All	71,103	1
С высокой экспрессией	High	968	0.3
С низкой экспрессией	Low	997	0.3
Нативные	Native	997	0.3
Случайные	Random	6349	0.3
Сложные	Challenging	1,953	0.5
SNVs	SNVs	44,340 пар	1.25
Замены в мотивах	Motif pertubation	3,287 пар	0.3
Разное положение мотива в последовательности	Motif tiling	2,624 пар	0.4

3.3.2. Ранее опубликованные данные МПРА

Для того чтобы оценить генерализуемость полученного в ходе конкурса решения на другие данные об активности регуляторных последовательностей дрожжей, использовались данные из исследования [70], которые были получены согласно уже описанному протоколу, для дрожжей *S. cerevisiae*, культивируемых в двух средах: YPD (стандартная полная среда, содержащая дрожжевой экстракт, пептон и декстрозу, 30 млн последовательностей) и SD-Ura (среда без урацила, 20 млн последовательностей).

Тестовые данные были получены в той же работе в независимых экспериментах и включали только измерения, полученные для нативных (т.е. присутствующих в геноме дрожжей) последовательностей промоторов (3928 для стандартной полной среды, 3977 для среды без урацила), измеренные с высокой точностью (активность каждой последовательности измерялась в среднем в 100 клетках дрожжей) [70].

3.3.3. Альтернативные модели

Наше решение соревновалось с решениями других участников. Среди них стоит отметить решения 2го и 3го места.

Решение команды BHI, занявшей второе место, представляло собой гибридную архитектуру из сверточных блоков и LSTM. Первый блок их нейронной сети состоял из двух параллельных сверточных блоков с разными размерами ядер, выходы которых конкатенировались и подавались на вход двунаправленному LSTM-блоку, выход которого подавался на вход еще двум параллельным конволюциями с разными размерами ядер. Выход этого слоя переводился в вектор и подавался в финальный блок из трёх последовательных полносвязных слоев, на выходе которого предсказывалась экспрессия целевой последовательности.

Решение команды Unlock_DNA, занявшей третье место, основывалось на архитектуре BERT [258] с дополнительными модификациями, включая добавленные в середину каждого блока BERT сверточные блоки.

Обе команды использовали обратно-комплементарную аугментацию как при обучении модели, так и на этапе предсказания.

3.3.4. Процедура обучения модели

Модель обучалась при помощи расписания обучения OneCycleLearningRate [317] с модификациями FastAI [318] (**рис. 29 А**).

Параметры OneCycleLearningRate были выбраны с использованием 1/10 части тренировочных данных. Для выбора максимальной скорости обучения для OneCycleLearningRate (0.005) мы использовали тест LR-range [319] (**рис. 29 В**).

В качестве оптимизатора изначально использовался AdamW (weight_decay=0.01) [320], в постконкурсном анализе он был заменен на недавно опубликованный оптимизатор Lion (weight_decay=0.1) [321].

Каждая эпоха обучения состояла из 1000 батчей размером 1024 тензора отвечающих входным последовательностям. Модель обучалась 80 эпох. Для окончательной модели использовались гиперпараметры, подобранные на валидации на 1/10 части тренировочной выборки. Финальная

модель обучалась на всём наборе тренировочных данных. Мы использовали ту же инициализацию весов, что и в EfficientNetV2.

3.3.5. Параметры диффузионной модели

Обучение диффузионной модели длилось в течении 200 эпох по 1000 батчей с использованием оптимизатора AdamW (weight_decay=0.01), скоростью обучения равной 0.001, числом последовательностей в батче равным 1024 и соотношением 4 к 1 между обучающим и валидационным наборами.

В данном исследовании для генерации последовательностей дрожжей с заданной экспрессией, мы использовали T=100 и shift=30. Для новых задач эти параметры необходимо подбирать отдельно.

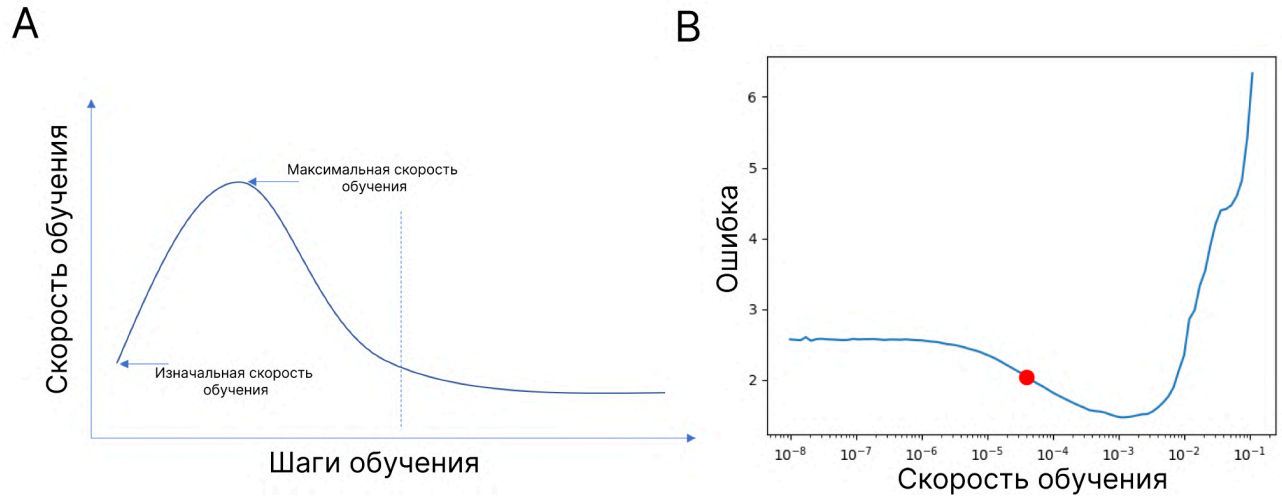


Рисунок 29. **A.** Расписание обучения OneCycleLearningRate. Обучение стартует с низкой скорости обучения, доводит ее до некой максимальной скорости, и затем начинает постепенно понижать скорость обучения, доводя до скорости еще более низкой, чем изначальная. Адаптировано из [322]. **B.** Learning rate тест состоит в обучении модели в течении одной эпохи с постепенным увеличением скорости обучения после каждой итерации. Максимальная скорость обучения, при которой ошибка сети еще не начинает расти, показывает район, в котором стоит искать максимальную скорость обучения для расписания OneCycleLearningRate. Адаптировано из [323].

3.4. Данные об активности регуляторных элементов в клетках человека[4]

3.4.1. Независимые библиотеки участков

Членами консорциума ENCODE были проведены эксперименты при помощи технологии lentiMPRA по измерению влияния некодирующих последовательностей длины 230 п.н. на транскрипцию репортерного гена в клеточных линиях HepG2 (клеточная линия гепатоцеллюлярной карциномы человека), K562 (клеточной линией иммортализованного миелогенного лейкоза) и WTC11 (индуцированные плюрипотентные стволовые клетки).

В числе исследуемых последовательностей были выбраны (**рис. 30 B)**:

1) потенциальные энхансеры (регионы открытого хроматина в соответствующем типе клеток);
2) промоторы, центрированные на начало старта транскрипции;
3) последовательности энхансеров, перемешанные с сохранением динуклеотидного состава;
4) последовательности с известной активностью в данных клеточных линиях по данным предыдущих работ [324–326].

Помимо этого, в клеточной линии HepG2 была измерена активность предполагаемых “нейтральных” синтетических элементов, которые не должны функционировать как промоторы или энхансеры. В случае клеточной линии K562 дополнительно измерялась активность 15 тысяч некодирующих участков, выбранных из областей гетерохроматина. В случае клеточной линии WTC11 из-за низкой эффективности трансдукции была измерена активность существенно меньшего числа регуляторных элементов. Кроме этого, для данных, полученных для этой клеточной линии, наблюдалось меньшее соотношение сигнал-шум [72].

Активность каждого элемента была рассчитана как логарифм по основанию 2 от нормализованного количества молекул РНК со всех баркодов, соответствующих элементу, делённого на нормализованное количество молекул ДНК со всех штрих-кодов, соответствующих элементу. Активность всех регуляторных элементов измерялась в прямой и обратно-комплементарной ориентациях (**рис. 30 A**).

[4] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: Agarwal V., Inoue F., Schubach M., **Penzar D.**, Martin B.K., Dash P.M., Keukeleire P., Zhang Z., Sohota A., Zhao J., Georgakopoulos-Soares I., Noble W.S., Yardımcı G.G., Kulakovskiy I.V., Kircher M., Shendure J., Ahituv N. Massively parallel characterization of transcriptional regulatory elements // Nature.– Springer Science and Business Media LLC, 2025.– P. 1–10. doi: 10.1038/s41586-024-08430-9. JIF (для WoS) = 50.5, (2.75/0.25)

3.4.2. Объединенная коллекция протестированных последовательностей

Объединенная библиотека (**рис. 30 C**) состояла из:

1) примерно 19,000 потенциальных энхансеров, активных в каждой из трех клеточных линий, отобранных равномерно из предыдущих независимых библиотек, чтобы охватить широкий диапазон активности;

2) подмножества промоторов, демонстрирующих высокую дисперсию экспрессии, а также широкий диапазон средней экспрессии среди различных типов клеток из предыдущих наборов;

3) последовательностей энхансеров, перемешанных с сохранением динуклеотидного состава;

4) набора положительных и отрицательных контролей, использующих синтетические элементы, разработанные для проявления активности в клетках HepG2;

5) природные элементы, для которых имеются доказательства специфической активности в клетках K562.

Активность элементов была измерена только для прямой ориентации относительно репортерного гена.

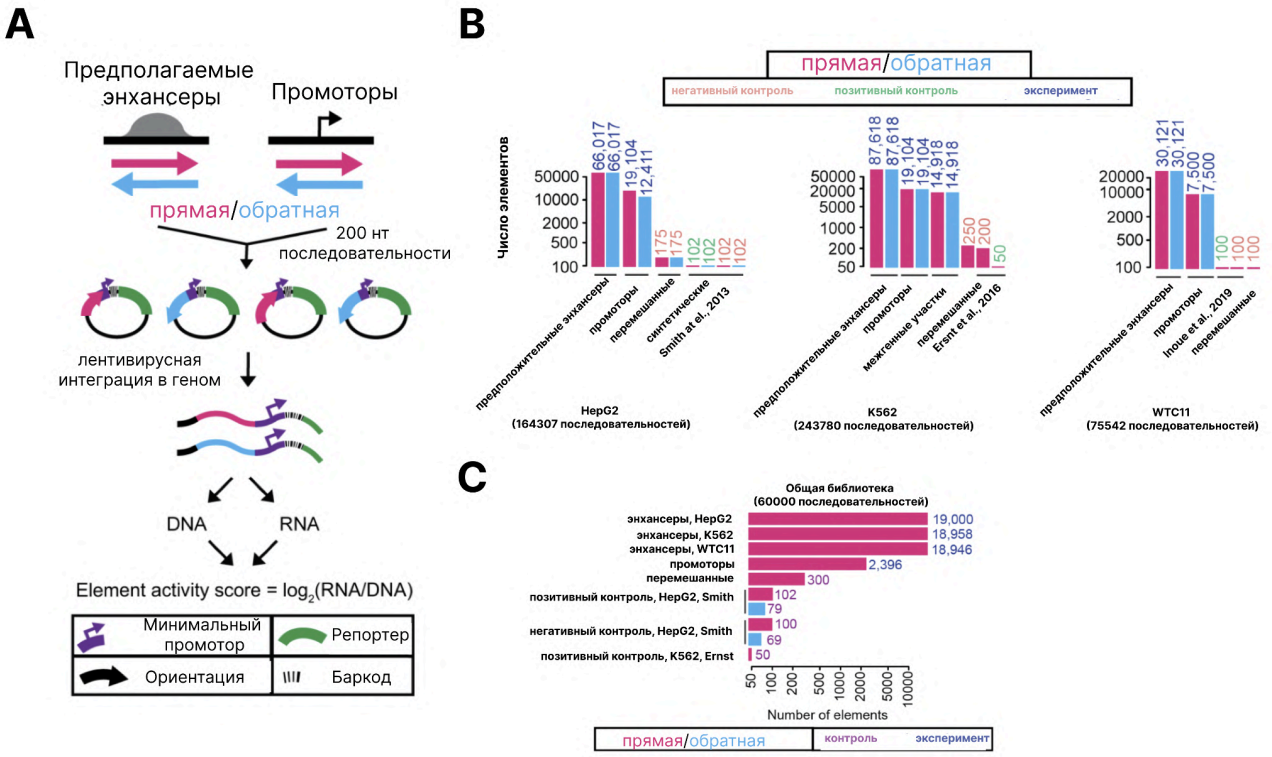


Рисунок 30. **A.** Схема lentiMPRA. Тысячи регуляторных элементов, включая предполагаемые энхансеры (регионы открытого хроматина в соответствующем типе клеток) и промоторы (регионы, центрированные на сайтах инициации транскрипции белок-кодирующих генов), встраиваются в репортерные плазмиды в обоих ориентациях вместе с баркодами. Библиотеки трансфецированы в клетки HepG2, K562 и WTC11 с использованием лентивируса, и интегрированная ДНК и транскрибированные РНК баркоды секвенируются для количественной оценки активности CRE. **B.** Состав библиотек HepG2, K562 и WTC11. В каждую библиотеку включены тысячи предполагаемых энхансеров и промоторов, негативные контроли (последовательности с перемешанными динуклеотидами или элементы, не продуцирующие сигнала по результатам предыдущих исследований), и позитивные контроли (элементы с подтвержденной активностью по результатам предыдущих исследований). Полосы окрашены в зависимости от тестируемой ориентации, с сопровождающими числами, указывающими количество тестируемых элементов в каждой категории. Цвета чисел соответствуют типу элемента. **C.** Состав совместной библиотеки, протестированной в клетках HepG2, K562 и WTC11. Было протестировано похожее соотношение предполагаемых энхансеров (подписаны как энхансеры), отобранных из каждого типа клеток, а также меньшее количество промоторов, негативных контролей (последовательности с перемешанными динуклеотидами или элементы, не имеющие сигнала по результатам предыдущих исследований) и позитивных контролей (элементы с подтвержденной активностью по результатам предыдущих исследований), включенных в каждую библиотеку. В данном случае элементы тестировались только в одной ориентации.

3.4.3. Оценка качества моделей

В данном исследовании для оценки качества моделей использовалась вложенная кроссвалидация. Последовательности каждого набора данных были разбиты на 10 случайных

фолдов одинакового размера. Если активность последовательности была измерена в прямом и обратном положении, то оба измерения попадали в один и тот же фолд.

Для каждого возможного разбиения 10 полученных фолдов на 8 обучающих, 1 валидационный и один тестовый, производилось обучение модели и получение предсказаний для тестового фолда. Предсказания для одного и того же тестового фолда усреднялись.

Для последовательностей, которые присутствовали в данных в прямом и в обратном положении, предсказания для них также усреднялись.

Итоговое качество модели вычислялось как корреляция Пирсона между предсказанными значениями и реальными.

3.4.4. Альтернативные модели

Наше решение сравнивалось с четырьмя моделями, предложенными авторами оригинальных данных.

3.4.4.1. Биохимическая модель

Первая модель, называемая авторами **биохимической**, использовала разметку участков связывания ТФ по данным ChIP-seq, локализацию модификации гистонов (гистонового ChIP-seq) и доступность хроматина (DNase-seq и ATAC-seq), доступные для клеток HepG2, K562 и WTC11 для сборки генома человека hg38 из проекта ENCODE[245]. Для клеточной линии WTC11 были также включены данные, полученные для клеточной линии H1-ESC (часто используемая клеточная линия стволовых клеток человека). В результате для HepG2 использовались данные 1506 эпигенетических разметок, для K562 – 1206 треков и 277 треков для WTC11. Для каждой последовательности из библиотеки были извлечены средние сигналы каждого трека из соответствующего участка генома человека. Сигналы были логарифмически трансформированы (с добавлением псевдосчета 0.1). Наконец, для каждого типа клеток реплики, соответствующие одному эксперименту, были усреднены. В результате было получено 655 признаков для HepG2, 447 признаков для K562 и 122 признака для WTC11.

Сама модель представляла собой LASSO регрессию с параметром регуляризации, подбираемым при помощи вложенной кросс-валидации.

3.4.4.2. SeiMPRA и EnformerMPRA

В моделях SeiMPRA и EnformerMPRA использовались предсказания моделей Sei [57] и Enformer [55] соответственно. В случае независимых библиотек, предсказание выполнялось для каждой последовательности в обеих ориентациях (с добавленными адаптерами, чтобы симулировать условия MPRA эксперимента). Полученные для двух ориентаций предсказания

усреднялись, в результате для каждой последовательности было получено 21907 признаков (SeiMPRA) и 5313 признаков (EnfomerMPRA). Для общей библиотеки, использовались предсказания для последовательностей только в прямой ориентации.

Так как Sei и Enformer требуют входные последовательности длины 4000 п.о и 196608 п.о, то все последовательности были дополнены с обеих сторон N (неизвестный нуклеотид) таким образом, чтобы исходная последовательность оставалась в центре.

Полученные признаки, как и в случае биохимической модели, подавались на вход LASSO регрессии с параметром регуляризации, подбираемом при помощи вложенной кросс-валидации.

3.4.4.3. MPRAnn

Сверточная нейронная сеть, состоящая из 2 сверточных слоёв, слоя макс-пулинга, еще 2 сверточных слоя, финальный выход которых преобразуется в одномерный вектор и пропускается через два линейных слоя (**рис. 48 В**). Для оптимизации модели использовался режим обучения со скоростью обучения 0.001 и ранней остановкой обучения модели в случае неуменьшения валидационной ошибки в течении 10 эпох. В целом обучение длилось 100 эпох, использовался оптимизатор Adam. Как и в случае предыдущих моделей, при обучении оптимизировалась среднеквадратичная ошибка.

3.4.5. Процедура обучения

Модель MPRALegNet обучалась при помощи процедуры, описанной для изначальной модели LegNet со следующими изменениями:

1) Использовался оптимизатор AdamW с weight_decay=0.1;
2) Было добавлено обрезание градиента (clip_val=1) для избежания взрыва градиента из-за использования расписания обучения OneCycleLearningRate;
3) Во время одной эпохи через модель пропускалось не 1000 батчей, а число батчей, необходимое чтобы модель за одну эпоху видела все объекты датасета;
4) Число эпох было подобрано на первом разбиении вложенной кросс валидации (1й фолд – тестовый, 2й – валидация, 3-10 – обучение) и равнялось 25 эпохам для датасетов HepG2 и K562 и 20 эпохам для датасета WTC11.
5) Оптимизировалась среднеквадратичная ошибка, так как решалась регресионная задача.

Аналогичные параметры использовались при обучении Legformer.

4. Результаты

4.1. Утечка данных при обучении моделей по данным параллельных репортерных экспериментов с мутагенезом насыщающей ПЦР[5]

Модели команд, разделивших первое место в конкурсе [51] использовали признаки на основе нейронной сети DeepSEA [53], предсказывающей для поданной на вход 1000 п.о. нуклеотидной последовательности доступность хроматина и связывания ТФ с ней.

В ходе участия в конкурсе нами было замечено, что разбиение, предложенное авторами (**рис. 31. А**), может приводить к утечке данных, в связи с тем, что близкие позиции в геноме часто обладают похожей функциональной важностью. Если признаки позиций, где произошла мутация, предоставляемые модели, позволяют ей устанавливать близость этих позиций в геноме, она может научиться предсказывать эффект мутации в позиции из тестового набора на основе агрегации эффектов рядом расположенных позиций из обучающего набора, не используя знаний о регуляторной грамматике.

Для того чтобы примерно оценить вклад утечки данных от участков репортера, участвующих в обучении, к соседним валидационным участкам, мы решили сравнить поведение моделей, обученных на биологически релевантных признаках, и моделей, обученных на признаках, не несущих биологической информации, но по остальным свойствам напоминающие биологически релевантные.

Для этого был создан набор “нерелевантных” признаков DeepSEA, полученных для геномных регионов, не имеющих отношения к исходным, но имеющих тот же размер. Так как использовавшиеся в соревновании репортеры не включали ни одного регуляторного района от генов, расположенных на третьей хромосоме, мы использовали в качестве “ложных” регионов регионы с третьей хромосомы, имеющие те же координаты начала и конца, что и использованные в экспериментах регуляторные элементы. По построению полученные таким образом признаки не должны нести никакой биологически релевантной информации. Любое качество на валидационной выборке, отличное от случайного, получаемое с помощью этих признаков, можно объяснить только утечкой данных.

---

[5] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: **Penzar D.**, Zinkevich A.O., Vorontsov I.E., Sitnik V.V., Favorov A.V., Makeev V.J., Kulakovskiy I.V. What Do Neighbors Tell About You: The Local Context of Cis-Regulatory Modules Complicates Prediction of Regulatory Variants // Front. Genet.– 2019.– Vol. 10.– P. 1078. doi: 10.3389/fgene.2019.01078. JIF (для WoS) = 2.8, (0.70/0.40)

Мы обучили случайный лес [327] при помощи пакета scikit-learn [328] с числом деревьев 500 (для остальных гиперпараметров использовались значения по-умолчанию) отдельно на настоящих признаках, полученных при помощи нейросети DeepSEA, и на нерелевантных признаках с другой хромосомы [33]. В результате мы оказалось, что модель, обученная на полностью нерелевантных признаках, не только показывает качество лучше случайного, но выступает наравне или на уровне с многими решениями конкурса [51] (**рис. 31**)

Это можно объяснить тем, что модель автоматически выучивает эвристику, согласно которой хорошим предсказанием эффекта мутации в оцениваемой позиции является средний эффект в ближайших позициях из обучения. В то же время можно заметить, что модель на основе реальных признаков все же значимо превосходит модель, обученную на нерелевантных. В связи с этим возникает вопрос: применима ли полученная модель на практике. В то же время можно заметить, что модель на основе реальных признаков все же значимо превосходит модель, обученную на нерелевантных. В связи с этим возникает вопрос: применима ли полученная модель на практике.

Для того чтобы ответить на данный вопрос, мы предсказали при помощи полученной модели эффект мутаций на опубликованных ранее данных насыщающего мутагенеза. Проводилась оценка двух версий модели – обученной только на тренировочной выборке из конкурса и обученной на всех данных с использованием тех же гиперпараметров. В связи с использованием тех же гиперпараметров модели и слабой склонностью случайных лесов к переобучению в таких условиях, можно надеяться, что такая модель за счет большого количества данных сможет лучше выучить регуляторную грамматику, а значит, и предсказать эффекты однонуклеотидных мутаций. Однако, согласно результатам оценки обеих моделей оказывается, что:

1) модель, обученная на тренировочной выборке, показывает крайне слабое, хоть и отличное от случайного, качество на публичных данных;

2) модель, обученная на полной выборке, показывает качество равное случайному предсказанию.

Таким образом, практическая применимость моделей, обученных или дообученных при помощи данных о насыщающем мутагенезе, оказывается под большим вопросом.

Таблица 6. AUROC и AUPRC на независимом наборе данных.

Модель	AUROC	AUPRC
RF(DeepSEA), CAGI 2018, тренировочные данные	0.6	0.2
RF(DeepSEA), CAGI 2018, полные данные	0.5	0.15

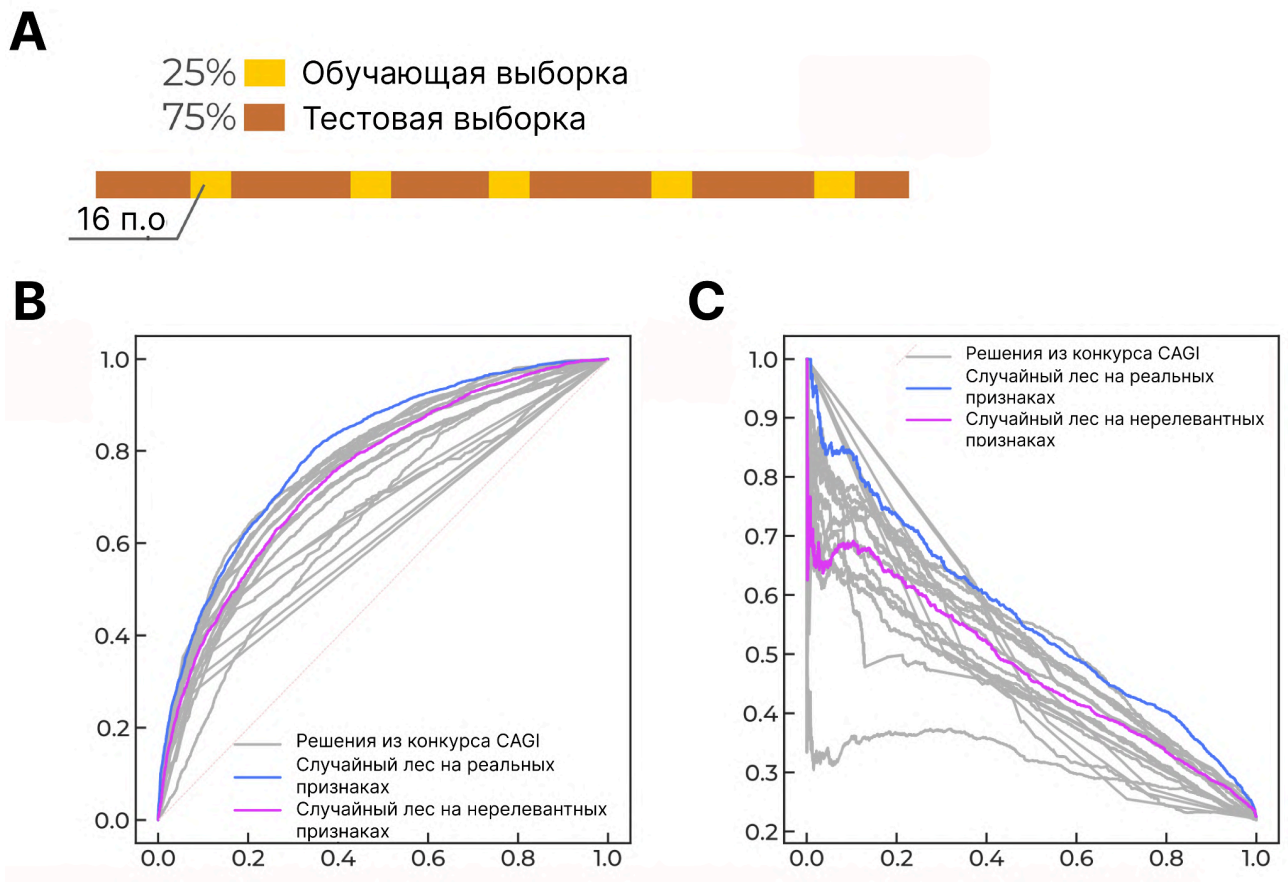


Рисунок. 31. A. Схема разбиения данных CAGI5, предложенное авторами конкурса. Для каждого репортера, обучающая выборка SNV (в сумме составляющая 25%) состоит из множества блоков длины 16 распределенных по координатам репортера. **B-C.** Синим показана модель, обученная на настоящих признаках DeepSEA, розовым — на нерелевантных, серым цветом показаны решения участников. В случае использования вместо реальных признаков, полученных со случайной последовательности генома такой же длины, качество решения по сравнению с лучшим (синим) падает, но все еще остается далеко от уровня случайного. На графике B изображены ROC-кривые, на графике C – PR-кривые.

4.2. Предсказания событий аллель-специфичного связывания[6]

Для того чтобы определить, насколько возможно обучение модели с использованием признаков DeepSEA на данных аллель-специфичного связывания [34], мы обучили модели случайного леса с использованием признаков DeepSEA и признаков, основанных на мотивах связывания транскрипционных факторов [313].

Задача предсказания аллель-специфичного связывания может быть формализована двумя путями:

[6] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: Abramov S., Boytsov A., Bykova D., **Penzar D.**, Yevshin I., Kolmykov S.K., Fridman M.V., Favorov A.V., Vorontsov I.E., Baulin E., Kolpakov F., Makeev V.J., Kulakovskiy I.V. Landscape of allele-specific transcription factor binding in the human genome // Nature Communications – 2021.– Vol. 12, № 1.– P. 2751. doi: 10.1038/s41467-021-23007-0. JIF (для WoS) = 14.7 (1.20/0.20)

1) общая классификация – предсказать, приводит ли данный однонуклеотидный вариант к событию аллель-специфичного связывания хотя бы для одного ТФ или хотя бы в одном типе клеток;

2) ТФ-специфичное или клеточно-специфичное предсказание – предсказать, приводит ли данный однонуклеотидный вариант к событию аллель-специфичного связывания для данного конкретного транскрипционного фактора или в данной конкретной клеточной линии.

Модели для обеих подзадач были обучены и проверены с использованием метода кросс-хромосомной валидации: итеративно для каждой из 22 аутосом одна хромосома выбиралась для валидации, а остальные 21 использовались для обучения. На каждой итерации оценивалось качество модели на отложенной хромосоме, и вычислялись средние значения ROC-AUC и PR-AUC.

Для первой подзадачи среднее качество моделей на уровне ТФ и типов клеток составило 0.74 и 0.73 ROC-AUC, и 0.44 и 0.56 PR-AUC соответственно (**рис. 32 A**). Для второй подзадачи для каждого ТФ и каждого типа клеток была обучена отдельная модель (**рис. 32 C-D**). Качество моделей различалось для разных ТФ и типов клеток, с наивысшим ROC-AUC в 0.72 и 0.81 для CTCF (среди ТФ) и HepG2 (среди типов клеток), и наивысшим PR-AUC 0.35 и 0.64 для CTCF и A549.

Анализ вклада признаков (**рис. 32 E-F**) показал, что все модели использовали признаки на основе нейронной сети DeepSEA. При этом среди всех признаков преимущественно использовались признаки, относящиеся к необходимым клеточным линиям или факторам транскрипции. Помимо этого, модели также использовали информацию из экспериментальных данных DNase-Seq, и информация об аллельном дисбалансе в среднем оказывалась для моделей важнее, чем информация о покрытии, что согласуется с более ранними исследованиями [315,329]. В случае предсказания сайтов аллель-специфичного связывания отдельных ТФ, признаки на основе мотивов также оказались полезными для моделей.

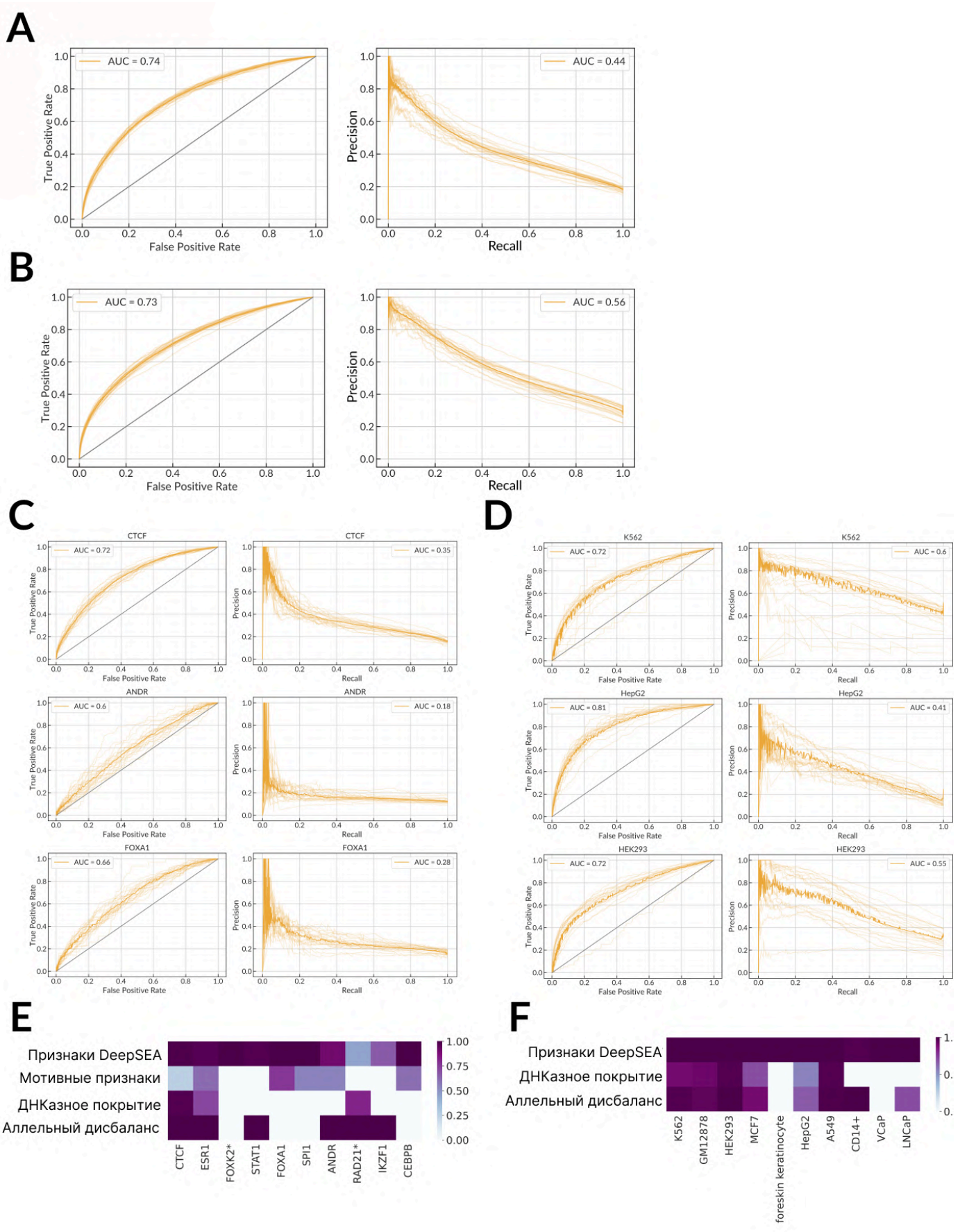


Рисунок 32. A-B. ROC-кривые и PR-кривые для задачи предсказания события аллель-специфичного связывания на общем наборе данных из 10 ТФ/10 типах клетках с наибольшим числом аллель-специфичных событий. **C-D**. ROC-кривые и PR-кривые для отдельных классификаторов для 3 транскрипционных факторов и типов клеток с

наибольшим количеством ASB. Значения площади под кривой показаны в легендах. Прозрачные кривые обозначают результаты индивидуальных исключений хромосом, яркие сплошные линии показывают усредненные кривые. **E-F**. Теплокарты, суммирующие относительную ранговую важность различных признаков (слева - модели для отдельных клеточных типов, справа – модели для отдельных ТФ), которые сгруппированы в четыре категории (см. **Таблицу 4**): (1) Выходы из финального слоя DeepSEA, (2) Признаки, связанные с мотивами, (3) Общее покрытие по данным DNase-Seq, (4) Признаки, связанные с дисбалансом в DNase-Seq. Цветовая шкала тепловой карты обозначает наилучшие относительные ранги признаков из каждой группы. В ранжировании были учтены только значимые признаки (т.е, которые показали результаты лучше, чем случайно перемешанные признаки), при этом значимость определялась с использованием меры важности Джини (вычисляемой как общее уменьшение нечистоты узлов, взвешенное по доле выборок, достигающих этого узла, усредненное по всем деревьям в случайном лесе[327]). Звездочкой (*) отмечены факторы транскрипции, для которых отсутствует известный мотив в базе данных HOCOMOCO[314], и, следовательно, отсутствуют признаки, связанные с мотивами.

4.3. Архитектура LegNet и ее применение к данным DREAM-2022[7]

4.3.1. Представление входных данных

Последовательности из обучающей выборки были дополнены с 5'-конца нуклеотидами из плазмиды, использовавшейся в эксперименте, до общей длины 150 пар оснований и закодированы в матрицы 4x150 с использованием one-hot кодирования (**рис. 33**).

Авторами конкурса для каждой последовательности предоставлялось взвешенное среднее групп, в которых она встречалась. При этом точной информации о том, сколько раз в эксперименте в каждой из групп встречалась конкретная последовательность, авторами конкурса не предоставлялось. По этой причине мы рассматривали последовательности с целыми значениями оценок как "**синглтоны**", т.е. они, вероятно, наблюдались только один раз, в то время как нецелые оценки были получены путем усреднения двух или более наблюдений. Такие последовательности важно было каким-то особенным образом пометить для модели, так как данные об их измеренной активности являются более шумными, поскольку основаны на единичных наблюдениях (**рис. 34**).

Чтобы предоставить эту информацию модели, мы ввели бинарный канал **is_singleton** (1 для синглтонов, 0 для других обучающих последовательностей). Поскольку ориентация регуляторных элементов относительно сайтов начала транскрипции может влиять на их активность, то есть, эффект от присутствия регуляторного элемента в прямой и обратных ориентациях может отличаться, нельзя было напрямую прибегать к аугментации добавлением в датасет обратно-комплементарных последовательностей.

Поэтому, помимо дополнения данных путем добавления для каждой последовательности ее обратно-комплементарного аналога, мы добавили в представление последовательности дополнительный канал **is_reverse**, указывающий, подана последовательность в прямой или обратно комплементарной форме, заполняя канал 0 или 1 соответственно.

Таким образом, результирующий вход нейронной сети имеет размерность 6x150 (**рис. 33**).

Для предсказания активности последовательностей из тестовой выборки использовалась аугментация во время тестирования, которая заключалась в усреднении предсказаний, сделанных для прямой (is_reverse=0) и обратно комплементарной (is_reverse=1) последовательности.

---

[7] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: **Penzar D.**, Nogina D., Noskova E., Zinkevich A., Meshcheryakov G., Lando A., Rafi A.M., de Boer C., Kulakovskiy I.V. LegNet: a best-in-class deep learning model for short DNA regulatory regions // Bioinformatics.– 2023.– Vol. 39, № 8. doi: 10.1093/bioinformatics/btad457. JIF (для WoS) = 4.4 (0.95/0.45). Rafi A.M., Nogina D., **Penzar D.**, Lee D., Lee D., Kim N., Kim S., Kim D., Shin Y., Kwak I.-Y., Meshcheryakov G., Lando A., Zinkevich A., Kim B.-C., Lee J., Kang T., Vaishnav E.D., Yadollahpour P., Random Promoter DREAM Challenge Consortium, Kim S., Albrecht J., Regev A., Gong W., Kulakovskiy I.V., Meyer P., de Boer C.G. A community effort to optimize sequence-based deep learning models of gene regulation. // Nature Biotechnology – 2024. doi: 10.1038/s41587-024-02414-w. JIF (для WoS) = 33.1 (1.5/0.30)

Аугментация по время тестирования за счет усреднения предсказаний для значений канала is_singleton 0 и 1 приводила к ухудшению качества, потому окончательные прогнозы для оценки делались с указанием is_singleton=0, то есть, считая, что все тестовые последовательности имеют достоверно измеренную активность (что соответствует постановке организаторов конкурса).

Использование эквивариантных сверток [240] не улучшило качества используемой модели. Добавление в обучающие данные аугментаций по типу MixUp [235] и CutMix [236] приводило у ухудшению качества предсказаний модели.

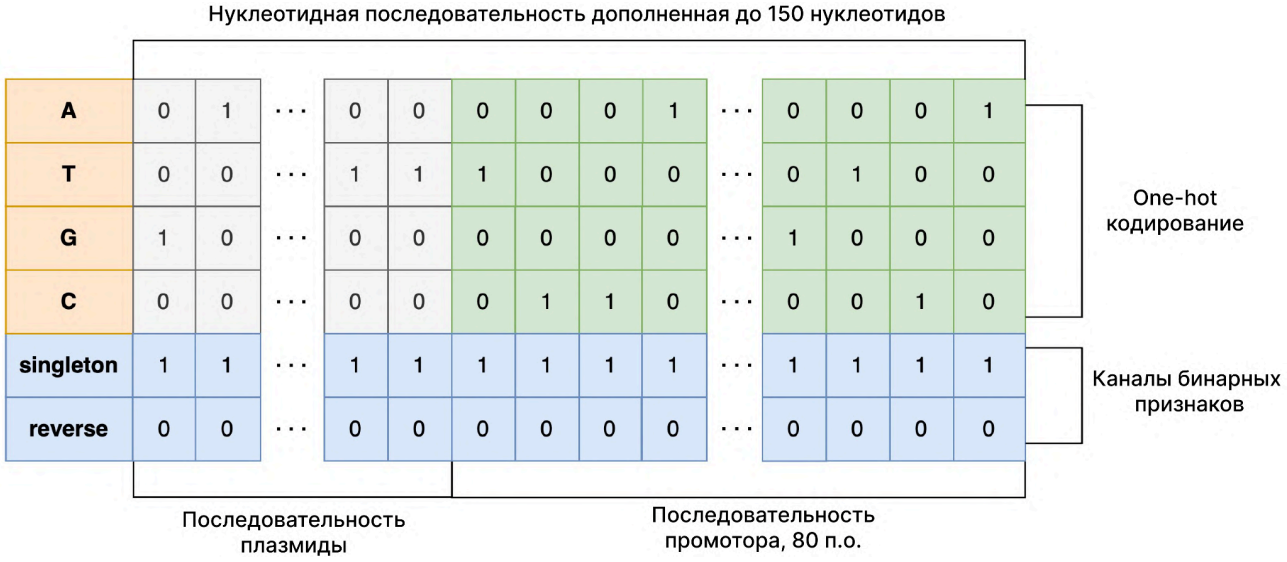


Рисунок 33. Представление последовательности, используемое нейросетью LegNet для решения задачи конкурса DREAM-2022. Последовательность предоставляется как 6-канальное одномерное изображение, в котором 4 канала содержат one-hot кодирование нуклеотидов на соответствующих позициях, а оставшиеся два являются константными и кодируют информацию о том, 1) измерена ли активность последовательности в эксперименте лишь единоразово или многократно, 2) подана она в прямой или обратно-комплементарной ориентации.

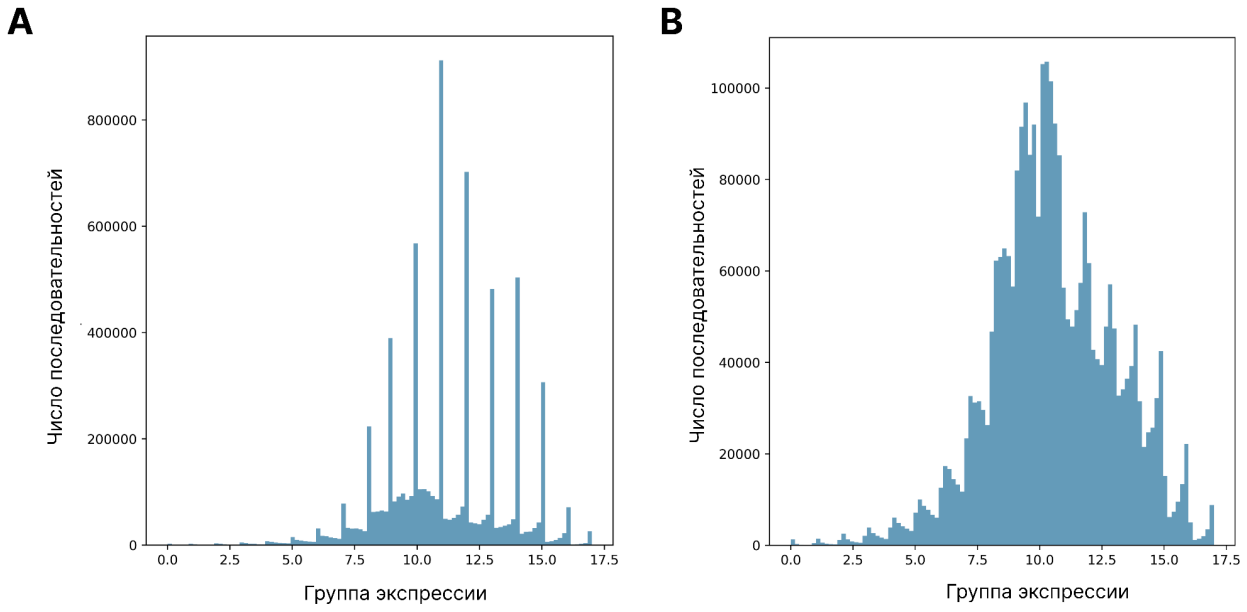


Рисунок 34. А. Распределение всех значений активности последовательностей в тренировочной выборке. Заметно, что целые значения превалируют над остальными. С большой вероятностью последовательности, которым соответствует целое значение активности наблюдались в эксперименте один раз. **В.** Распределение значений активности последовательностей в тренировочной выборке с отфильтрованными целыми значениями.

4.3.2. Модификация задачи регрессии с учетом особенностей данных

Конкурсная задача предсказания непрерывной величины значения активности представляется как задача регрессии. Однако, этот подход не позволяет использовать особенности природы предоставленных данных.

Мы изменили формулировку задачи для машинного обучения из регрессионной в классификационную (т.н. мягкая классификация), и для этого преобразовали целевые оценки экспрессии в вероятности классов, используя приведенную ниже схему. В работе [68] показано, что наблюдаемое значение экспрессии является реализацией нормально распределенной величины. По этой причине каждому значению экспрессии мы ставим в соответствие вероятностное распределение:

$$\rho \sim N(\mu = e + 0.5,\ sd = 0.5),$$

где e - значение измеренной экспрессии.

Классом в данной постановке называется целочисленное значение экспрессионной группы. Таким образом, при условии нумерации 18 групп с нуля, для каждого группы i от 1 до 16 вероятность того, что последовательность принадлежит данному классу, находится в промежутке $[i,\ i + 1)$. В случае нулевого и 17-го групп вероятность находится в промежутке $(-\inf,\ 1)$ и $[17,\ +\inf)$ соответственно.

Сдвиг среднего в нормальном распределении на 0.5 является поправкой на непрерывность и позволяет уменьшить разницу между исходной экспрессией для последовательности и математическим ожиданием дискретизированного нормального распределения.

Стандартное отклонение нормального распределения подбиралось при помощи кросс-валидации.

В качестве функции ошибки модели мы выбрали дивергенцию Кульбака-Лейблера между распределением вероятностей, полученных из тренировочных данных, и предсказанным вектором 18-ти значений от 0 до 1. Для получения итогового предсказанного значения e_{pred} на этапе предсказания на тестовых данных полученные значения на промежутке $[0,\ 1]$ домножаются на величину соответствующей им группы:

$$e_{pred} = \sum_{i=0}^{17} i \cdot p_i$$

Слой модели, отвечающий за эту операцию и объединенный с предшествующей операцией softmax, называется soft-argmax [330].

4.3.3. Архитектура нейронной сети

Для решения задачи предсказания экспрессии по последовательности за основу была взята архитектура EfficientNetV2 [74], которая на начало 2022 года была признана наилучшей в задачах обработки изображений. Архитектура была адаптирована для работы с нуклеотидными последовательностями - блоки для работы с двухмерными изображениями были заменены на блоки для работы с одномерными сигналами, было изменено количество блоков и порядок их следования. Помимо этого, residual connections исходной архитектуры были заменены конкатенацией, что, как было показано для DenseNet [331], улучшает сходимость нейросети и помогает работать с шумными данными (**рис. 35 A**).

Первый блок LegNet (блок Stem, **рис. 35 B**) представляет собой стандартную свертку с размером фильтра 7. За ним следуют батч-нормализация и SiLU активация.

Выход первого блока передается в последовательность из шести блоков свертки видоизменённой архитектуры EfficientNetV2 (**рис. 35 C**). Основные модификации состоят в том, что стандартные residual-связи между блоками были заменены residual-конкатенацией по каналам.

Все свертки использовались с паддингом для сохранения размера выходного тензора равным размеру входного. Перед передачей в выходной блок нейронной сети тензор пропускается через блок той же структуры, что и Stem-блок для приведения к нужному числу каналов.

В EfficientNet-подобного блоке используется модифицированная версия обычной свертки, позволяющая уменьшить число параметров модели – групповая свертка, в которой для вычисления выходного канала используется только несколько входных каналов. Число используемых входных каналов называется размером группы и является гиперпараметром. В нашем случае мы использовали размер группы, равный 2.

Блок Squeeze and Excitation (SE-блок, **рис. 35 E**), используемый в составе EfficientNet-подобного блока, является модификацией части оригинальной архитектуры EfficientNetV2. Основная идея SE блока заключается в том, чтобы динамически перераспределять весовые коэффициенты между различными каналами признаков, основываясь на их важности. Это позволяет модели лучше фокусироваться на наиболее значимых признаках данных. На первом этапе данный блок сжимает информацию о каждом канале признаков до одного значения при помощи поканального усреднения. Полученные признаки таким образом пропускаются через среднюю часть блока, оканчивающуюся сигмоидальной функцией активации, в результате чего получаются веса, соответствующие важности каждого канала. Затем входной тензор домножается на эти веса.

Количество параметров в билинейном блоке внутри блока SE было уменьшено с помощью низкорангового представления параметризованного тензора через реализацию канонического полиадического разложения библиотеки TensorLy.

Последний блок (**рис. 35 D**) состоит из одного сверточного слоя с размером фильтра 1, за которым следует поканальное усреднение (Global Average Pooling) и активация soft-max. Мы использовали 256 каналов для первого блока и [128, 128, 64, 64, 64, 64] каналов для шести EfficientNet-подобных блоков соответственно.

4.3.4. Результаты конкурса DREAM-2022

В качестве основных метрик авторами конкурса использовались взвешенные квадрат корреляции Пирсона и корреляция Спирмена

$$PearsonScore = \sum_{i}^{subsets} w_i \times PearsonR^2_i \Big/ \sum_{i}^{subsets} w_i$$

$$SpearmanScore = \sum_{i}^{subsets} w_i \times SpearmanR_i \Big/ \sum_{i}^{subsets} w_i$$

Для того чтобы определить устойчивость полученных оценок, авторы также проводили процедуру бутстрепа, считая метрики 10000 раз по случайно выбранным 10% тестовых данных и подсчитывая для каждой команды ранг, который она заняла.

В результате решение на основе LegNet заняло первое место во всех номинациях, устойчиво превосходя решения всех других участников (**рис. 36**). Что особенно важно, наша модель превзошла остальные решения в задаче предсказания эффектов однонуклеотидных мутаций (**рис. 36 C-D**).

Также стоит отметить, что модели по-разному справлялись с разными подкатегориями данных. С наибольшим качеством предсказывалась экспрессия случайных последовательностей, в то время как для нативных последовательностей качества моделей отличались уже более существенно. Также все модели, включая LegNet, столкнулись с проблемой в предсказании последовательностей с очень высокой и низкой экспрессией (**рис. 37**). Это может быть объяснено и малым числом таким последовательностей в обучающем датасете, и тем, что экспериментальный шум для крайних групп экспрессии был выше.

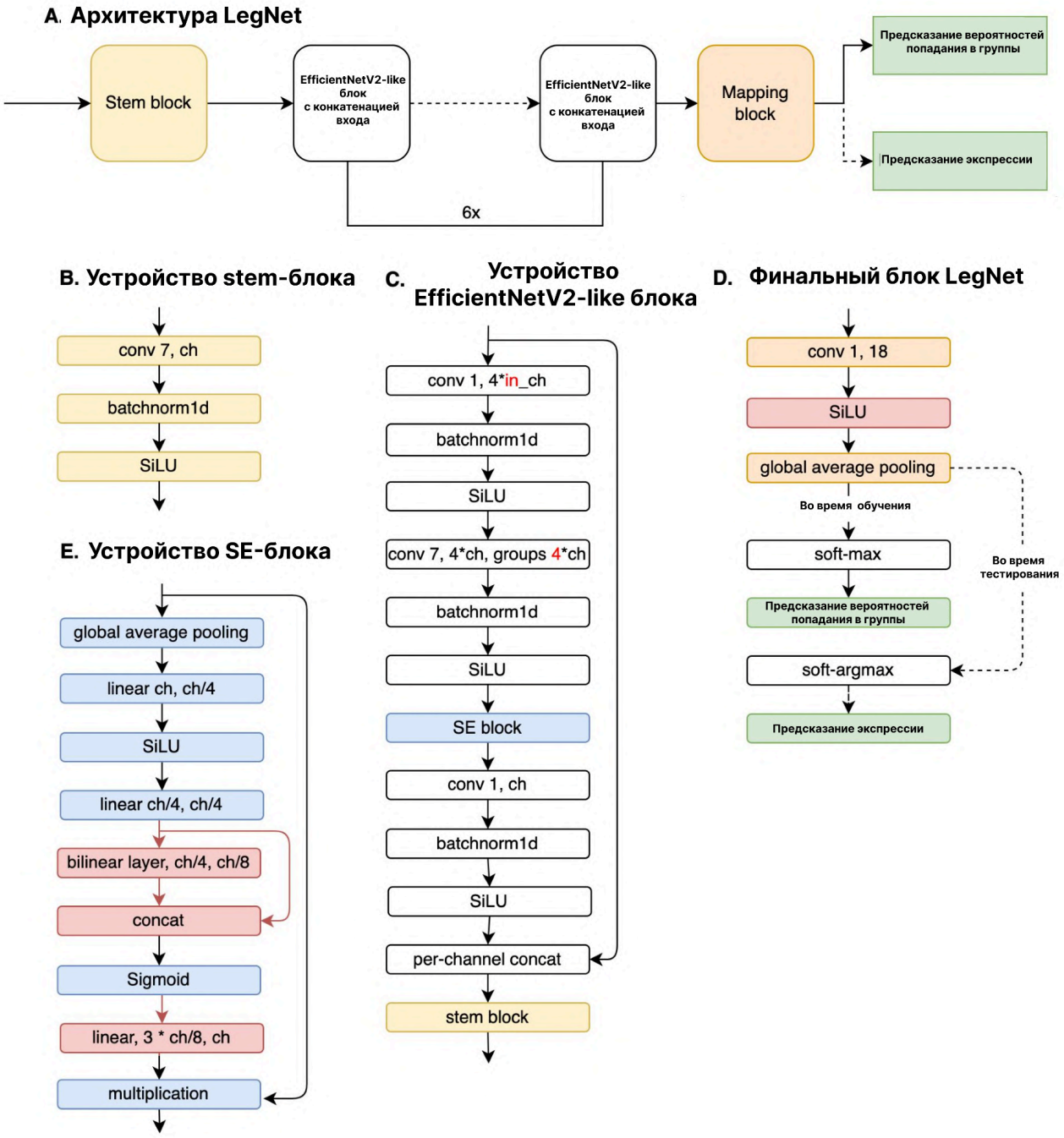


Рисунок 35. Схема архитектуры LegNet. **A** - обзор архитектуры. **B** - структура блока Stem. **C** - сверточный блок адаптированный из EfficientNetV2. **D** - последний слой нейросети. Пунктирные линии соответствуют сценариям при валидации и тесте. **E** - структура SE-блока. Красные блоки соответствуют блокам, которые были удалены в ходе постконкурсного анализа решения. Global average pooling – поканальное усреднение. mulitplication – поточечное произведение.

К интересным результатам также стоит отнести то, что модели лучше справлялись с предсказанием эффектов мутаций в мотивах – как очень простой задачей, с которой хорошо справляются и позиционно-весовые матрицы, чем с задачей предсказания эффекта изменения положения мотива относительной старта транскрипции, что согласуется с наблюдением, что моделям в среднем сложнее выучить позиционно-зависимые эффекты.

Последнее, что можно отметить это то, что последовательности, специально задизайненные как “сложные” – те, на которых модели из предыдущей работы авторов конкурса давали разные ответы, не оказались проблемными ни для одной из моделей конкурса. Это указывает на то, что, вероятнее всего, отличия ответов моделей можно объяснить их переобученностью, и, как следствие, их высокой несогласованностью, а не реальной сложностью предсказания экспрессии для отобранных последовательностей.

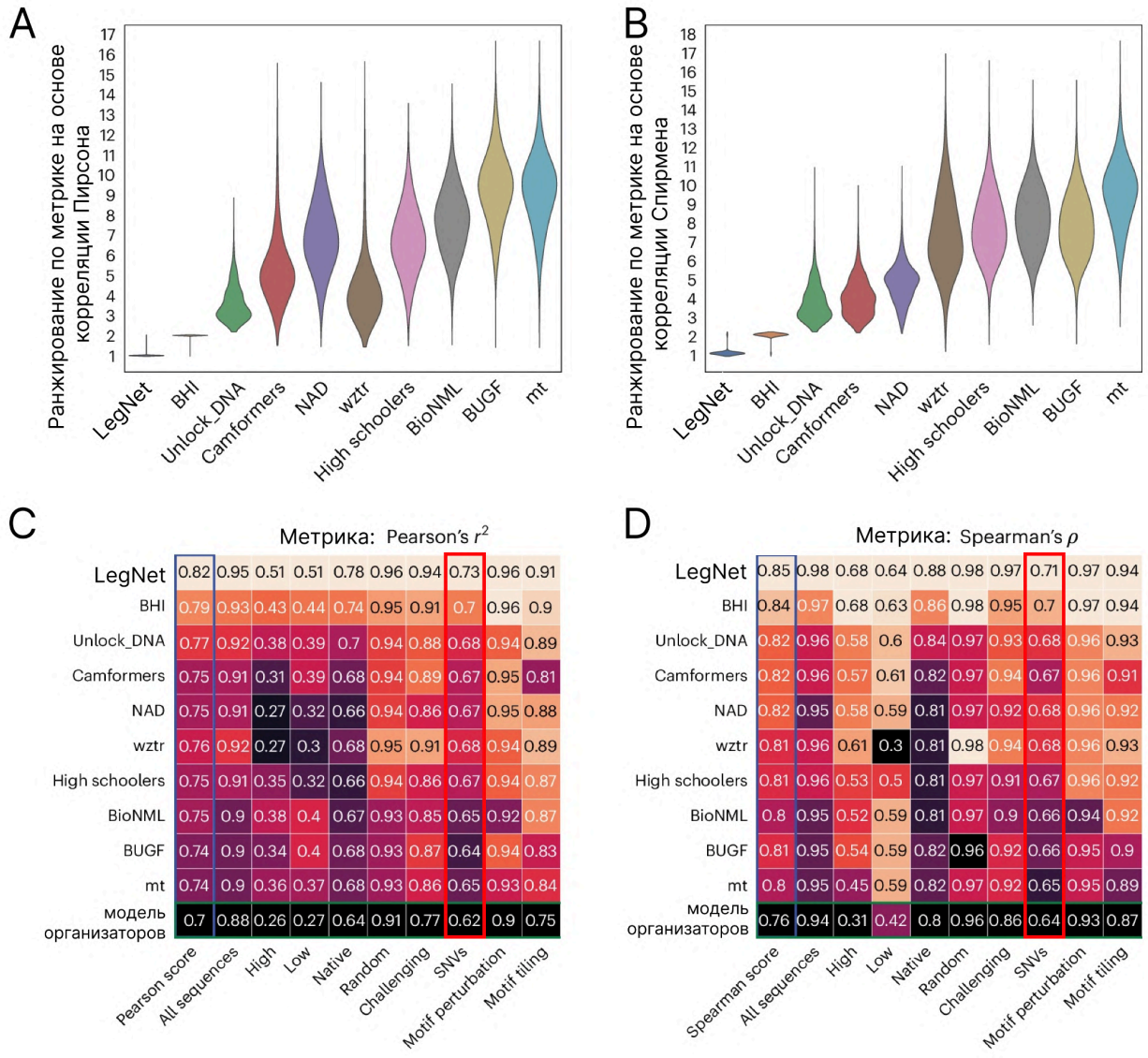


Рисунок 36. A-B. Модель LegNet (autosome.org) значимо превосходит все остальные решения в обоих конкурсных метриках. C-D. Во всех подкатегориях задач наша модель также превосходит остальные решения, что особенно важно, в задаче предсказания эффектов однонуклеотидных мутаций (SNV, красная рамочка).

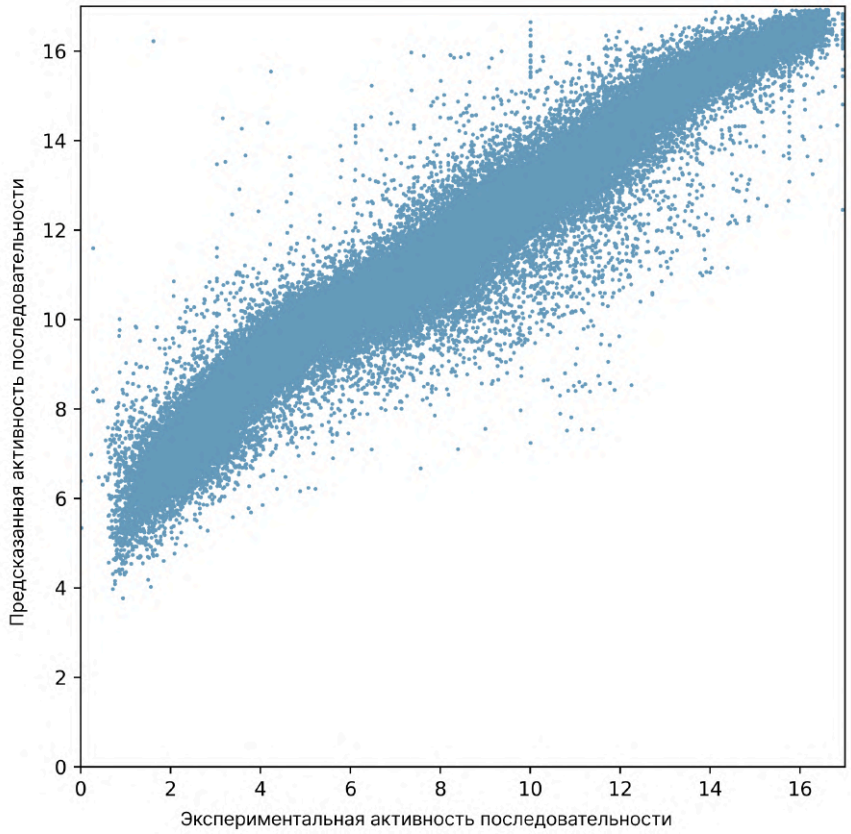


Рисунок 37. Сравнение предсказанных и реальных значений активности регуляторных элементов на тестовом датасете DREAM-2022. Модель плохо предсказывает значение активности ниже 4.

4.3.5. Пост-конкурсная оптимизация модели LegNet

Для того чтобы идентифицировать ключевые элементы представленного нами решения, мы последовательно проверили роль:

1) удаления и добавления отдельных слоев нейронной сети;
2) используемого алгоритма оптимизации и расписания обучения;
3) переформулировки задачи из регрессионной в задачу мягкой классификации;
4) использования канала синглетонов;
5) роль обратно-комплементарной аугментации.

Для каждого теста мы учили 5 моделей с использованием разных изначальных состояний генератора случайных чисел.

В частности, мы проверили эффект следующих архитектурных решений на качество модели:

1) SiLU-активации до слоя поканального усреднения;
2) различных модификаций SE-блока;
3) замены слоя конкатенации (concat) на слой поточечного сложения (add), который используется в оригинальной архитектуре EfficientNetV2 [74];
4) размера внутреннего представления, используемого в SE-блоке;

5) числа основных EfficientNetV2-like блоков модели.

Нами было показано, что в изначальной архитектуре:

1) можно заменить усложненный SE-блок на классический без ухудшения качества;

2) можно заменить группированные свертки на depthwise (крайний случай группированных сверток, в которой размер группы равен 1) без ухудшения качества, но уменьшением числа параметров модели;

3) можно убрать активацию перед поканальным усреднением, что улучшает качество и стабилизирует обучение.

Помимо этого, нами было показано, что замена оптимизатора AdamW на Lion также помогает улучшить качество предсказания модели.

Интересно, что схема обучения в целом важнее большей части гиперпараметров в архитектуре LegNet, за исключением наличия SE-блока.

4.3.6. Ансамблирование моделей

Для того чтобы оценить верхнюю планку качества, которого можно достичь на задаче конкурса, мы решили прибегнуть к ансамблированию разных версий нашей модели, полученных использованием разных изначальных состояний генератора случайных чисел.

Для этих целей мы обучили 100 различных моделей и далее 500 раз случайным образом выбирали 50 моделей и измеряли качество ансамбля, получаемого их последовательным добавлением.

В ходе данного эксперимента мы показали (**рис. 38**), что ансамбли моделей в среднем превосходят по качеству единичную модель.

Помимо этого, стоит отметить, что при ансамблировании неоптимизированной версии нашей модели, качество среднее качество ансамбля из 50 моделей оказалось ниже.

Таким образом, мы можем утверждать, что на задаче предсказания активности регуляторных последовательностей дрожжей из конкурса DREAM-2022 можно потенциально получить немного лучшее решение, чем имеющееся на данный момент. В то же время мы не можем утверждать, что качество данного решения будет ограничиваться качеством ансамбля из 50 текущих версий моделей.

Стоит также отметить, что в силу высокой скорости вычисления предсказаний, представленной нами модели, на практике не представляет проблемы использовать предсказание не индивидуальной модели, а ансамбля из нескольких.

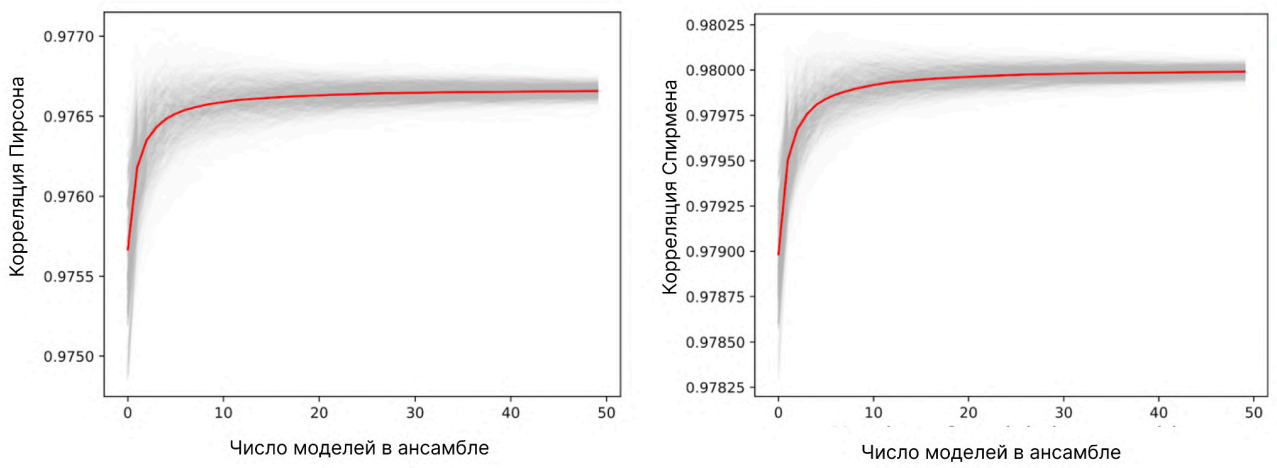


Рисунок 38. Ансамблирование в среднем улучшает корреляцию предсказаний с реальными значениями. Для простоты интерпретации приведены корреляции Пирсона и Спирмена без дополнительных модификаций. Левая панель: коэффициент корреляции Пирсона, правая панель: коэффициент корреляции Спирмена. Серые траектории: отдельные выборки фиксированных размер (ось X) от сотни обученных моделей; красная кривая: среднее значение. При интерпретации результатов стоит обратить внимание на нижний предел оси Y.

4.3.7. Предсказание активности дрожжевых промоторов по опубликованным ранее данным

Мы провели тренировку LegNet на опубликованных данных экспрессии репортерного гена в дрожжевых клетках для промоторов с известной последовательностью в среде YPD (комплексная среда: дрожжевой экстракт, пептон и глюкоза) и среде SD-Ura (синтетическая среда, не содержащая урацил).

В обоих случаях LegNet продемонстрировал высокое и стабильное качество предсказаний со значениями метрик, значительно превосходящими результаты полученные моделью на основе механизма внимания, опубликованной вместе с исходными данными[70] (**рис. 39**).

Нижняя граница предсказания, возникающая на уровнях экспрессии 4 (YPD) и 2,5 (SD-Ura), является известной проблемой, заключающейся в специфике тренировочных данных. Она возникала в предыдущих работах с данными из подобных экспериментов [68,70]. Вероятно, это явление вызвано тем, что клеточный сортер имеет ограниченное соотношение сигнал/шум в этом диапазоне.

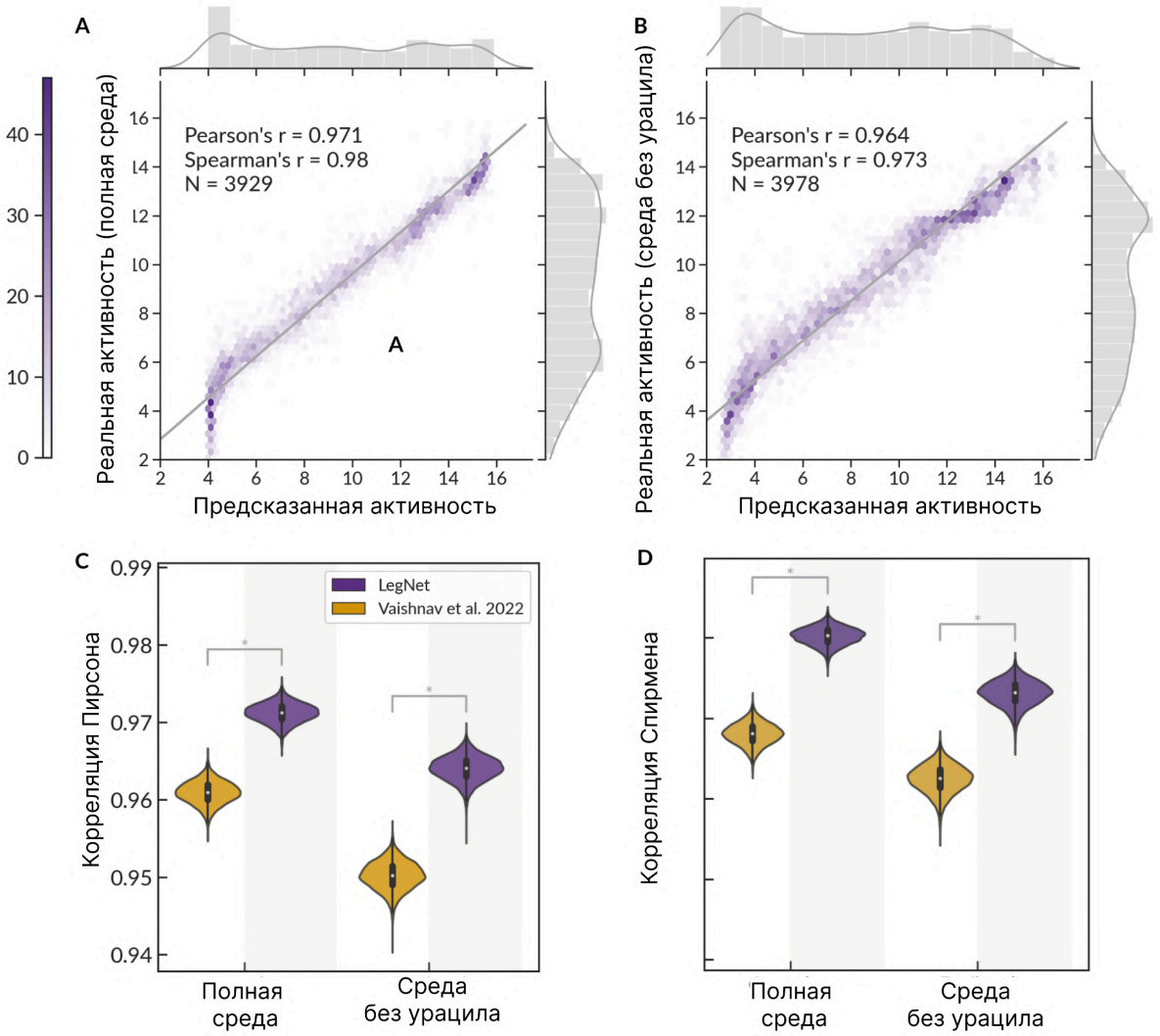


Рисунок 39. Предсказание экспрессии моделью LegNet по последовательности промотора на открытых данных. **A-B.** Предсказание экспрессии нативных промоторов для дрожжей, выращенных в полной среде. **A**. и среде без урацила **B**. **C-D**. Сравнение метрик модели LegNet и трансформерной модели [70] на нативных промоторах. Графики показывают распределение корреляции Пирсона (**C**) и корреляции Спирмена (**D**) между предсказанными и истинными значениями. Распределения метрик получены процедурой бутстрепа с n=10,000. *p <0.001 из теста зависимых корреляций Сильвера [332].

Мы также сравнили LegNet с более ранними подходами глубокого обучения, протестированными в [70] (**рис. 40**). Заметен разрыв между метриками LegNet (~0,96-0,98 корреляции Пирсона и Спирмена) и традиционными моделями глубокого обучения, такими как DeepSEA и DanQ (корреляции около ~0,92-0,94).

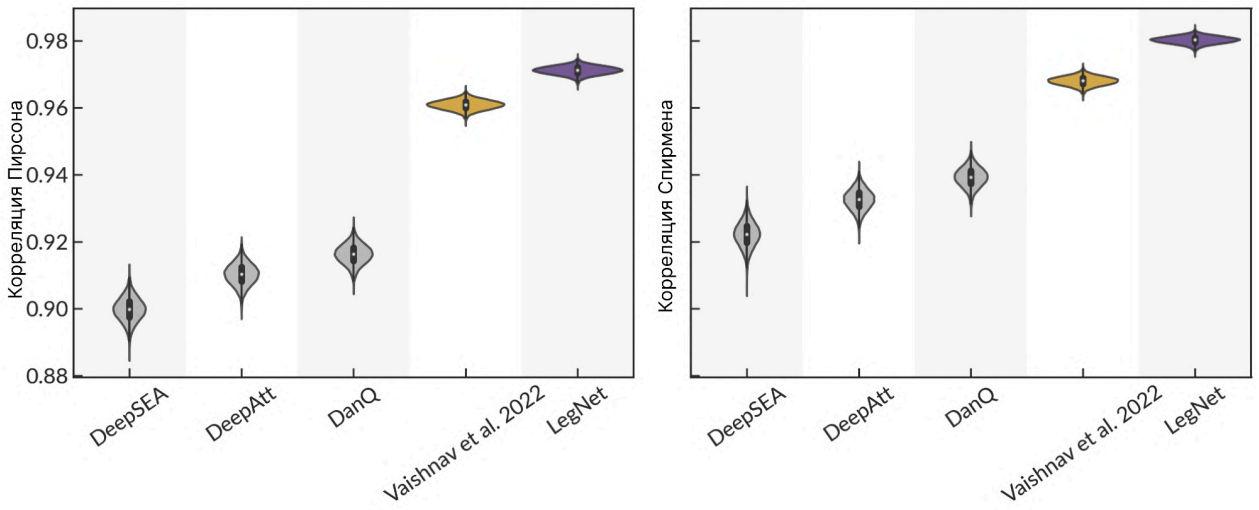


Рисунок 40. Предсказание экспрессии моделью LegNet по последовательности промотора на открытых данных в сравнении с DeepSEA, DeepAtt, DanQ и трансформенной моделью из [31]. Графики показывают распределение корреляции Пирсона (слева) и корреляции Спирмена (справа) между предсказанными и истинными значениями. Распределения метрик получены процедурой бутстрепа с n=10,000.

4.3.8. Оценка влияния замен в последовательности промотора

Как упоминалось ранее, в конкурсе DREAM-2022 LegNet превзошёл решения других участников во всех категориях промоторов, в том числе в оценке экспрессии промоторов с однонуклеотидными заменами. Мы протестировали LegNet в предсказании влияния множественных нуклеотидных замен на открытых данных, полученных в результате массового параллельного эксперимента (**рис. 41**). Схема проведения эксперимента была той же что и ранее, однако его целью была регистрация изменения экспрессии при случайном генетическом дрейфе [31]. Для 1000 уникальных случайных промоторных последовательностей в течение трех поколений произвольно вводились однонуклеотидные мутации и для каждого поколения производилось измерение уровня экспрессии.

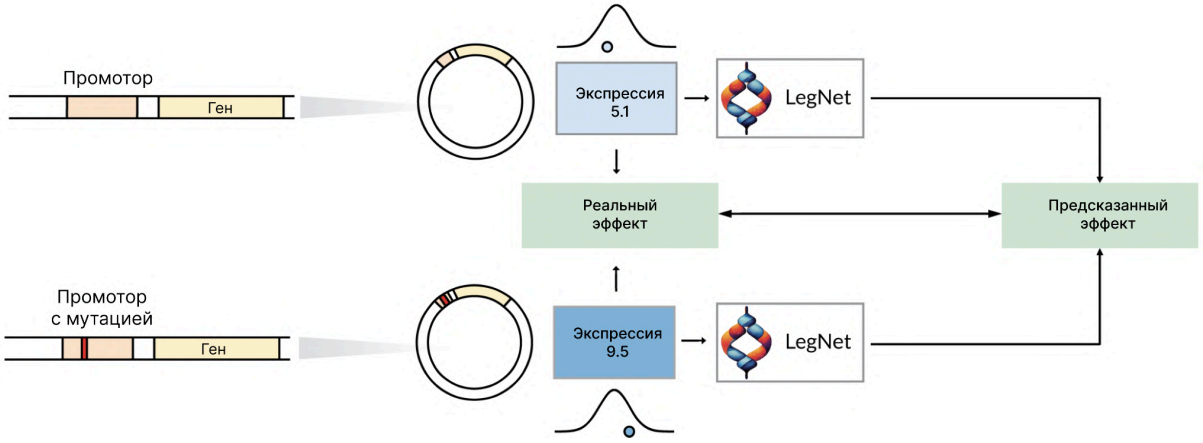


Рисунок 41. Схема оценки влияния замен при помощи LegNet. Исходные и мутированные последовательности промоторов по отдельности передаются в обученную нейронную сеть. Эффект мутации оценивается как разница между соответствующими предсказаниями и сравнивается с экспериментальными данными.

Мы оценили способность LegNet количественно оценивать разницу между экспрессиями исходной и мутировавшей промоторной последовательности в зависимости от числа нуклеотидных замен (1, 2 или 3) и сравнили результаты с моделью на основе механизма внимания из [31] (**рис. 42**). Во всех сценариях LegNet продемонстрировал значительное улучшение метрик предсказания по сравнению с моделью, предложенной авторами.

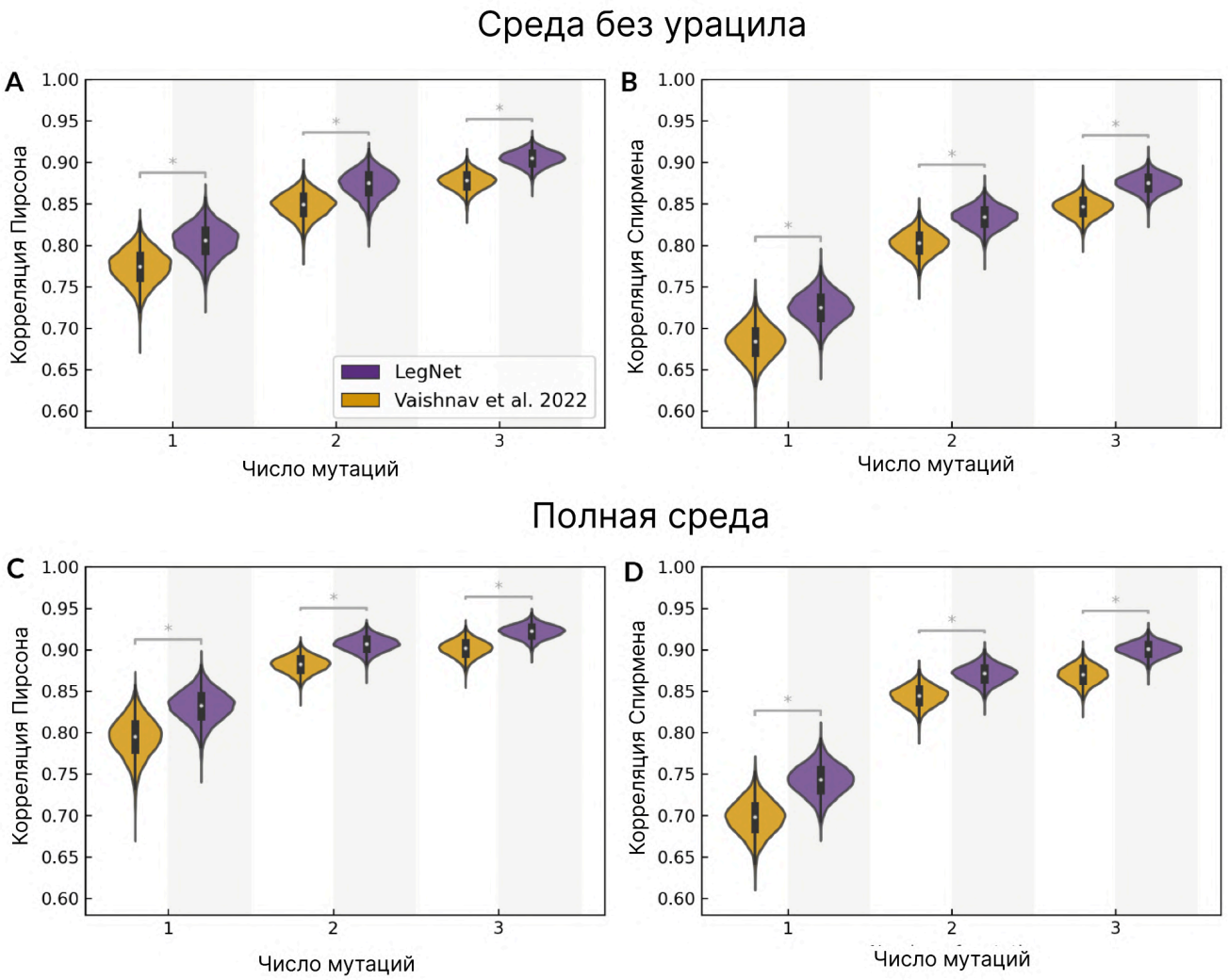


Рисунок 42. Предсказание влияния замен моделью LegNet по последовательности промотора на открытых данных для дрожжей, выращенных в среде SD-Ura - Defined (**A-B**) и YPD - Complex (**C-B**). Графики показывают распределение корреляции Пирсона (**A, C**) и корреляции Спирмена (**B, D**) между предсказанными и истинными значениями. Распределения метрик получены процедурой бутстрепа с n=10,000. *p < 0.001 из теста зависимых корреляций Сильвера [32].

4.3.9. Оптимизация решений конкурса

По результатам конкурса совместно с организаторами нами был разработан пакет PrixFixe для применения решений, предложенных участниками конкурса, к другим задачам регуляторной геномики (https://github.com/de-Boer-Lab/random-promoter-dream-challenge-2022).

В частности, нами было проведено исследование того, могут ли различные компоненты решений топ-3 команд привести к улучшению итогового решения конкурса.

Оказалось, что никакие компоненты, предложенные другими командами, не улучшают качество нейронной сети оптимизированной архитектуры LegNet на задаче предсказания активности дрожжевых промоторов.

В то же время предложенный нами режим обучения и формулировка оптимизируемой задачи как задачи мягкой классификации существенно улучшает качество решений 2го и 3го места (**рис. 43 A**). Это подтверждает вывод о важности этих двух компонентов, сделанный нами в независимой постконкурсной оптимизации. Стоит также отметить, что предложенная нами модель является наименьшей по числу используемых параметров (**рис. 43 B**).

Это лишний раз подчеркивает важность оптимизации решений моделей машинного обучения и учета биологической составляющей решаемой проблемы.

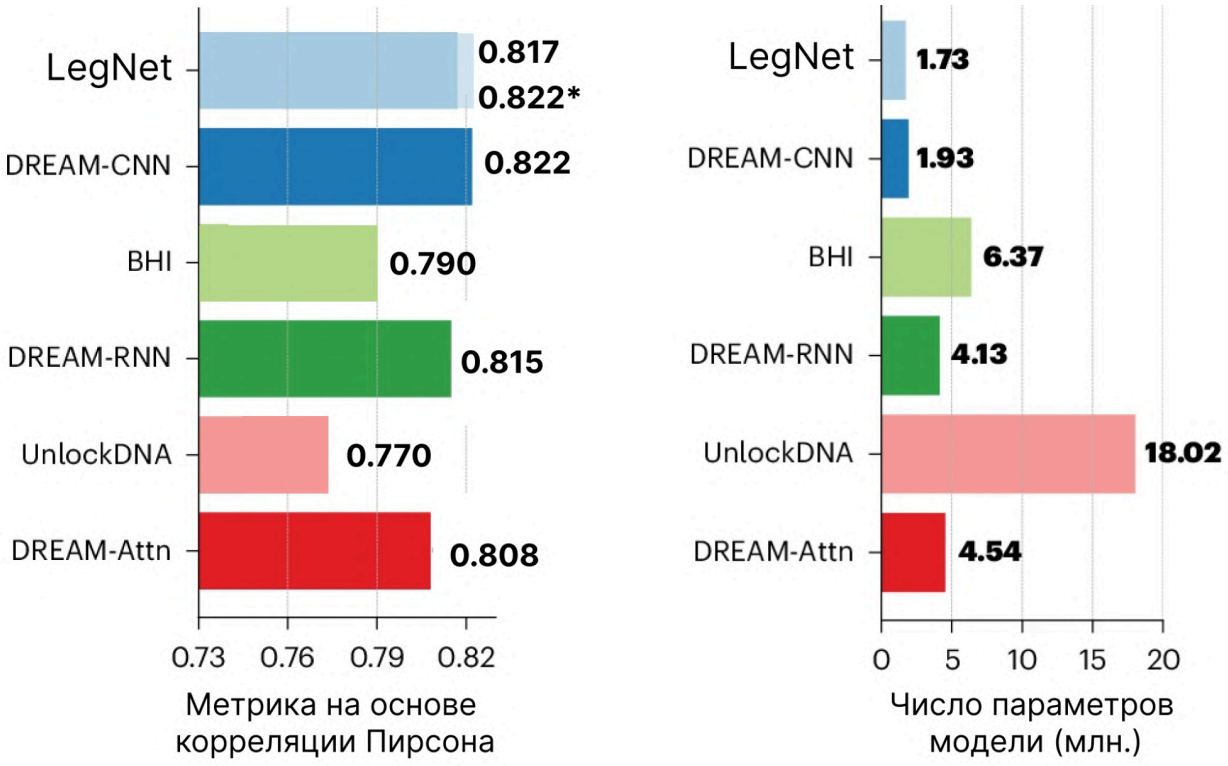


Рисунок 43. Постконкурсная оптимизация моделей участников конкурса DREAM-2022 при помощи пакета PrixFixe. Модель участников и результат их оптимизации при помощи пакета расположены парами (LegNet – DREAM-CNN, BHI – DREAM-RNN, UnlockDNA – DREAM-Attn) **A.** Качество моделей на задаче конкурса. Указано два качества – для исходной модели и оптимизированной нами после конкурса (качество оптимизированной модели указано *). **B.** Число параметров моделей. Для LegNet указано число параметров уже оптимизированной модели.

4.4. Генерация промоторных последовательностей с заданной активностью[8]

4.4.1. Холодная диффузия

Для того чтобы адаптировать нашу модель к задаче генерации регуляторных последовательностей в качестве основы мы решили использовать подход, основанный на методе **холодной диффузии** [97].

В качестве источника шума мы использовали внесение одиночной замены в случайное место последовательности (**рис. 45 А**). При этом не запрещались возвратные мутации. В силу того, что внесение k подобных мутаций в последовательность элементарно представляется в виде внесения этих мутаций итеративно, предложенный зашумляющий процесс отвечает требованиям холодного диффузионного процесса.

4.4.2. Подбор числа шагов диффузии

Мы провели дополнительный вычислительный эксперимент для определения числа шагов данного процесса, достаточного для того, чтобы предсказанная активность зашумленных последовательностей не отличалась от предсказанной активности последовательностей, сгенерированных случайно в предположении равной вероятности каждого нуклеотида (рис XY.).

Для этого вносились замены в стартовые последовательности трех видов:

1) последовательности из экспериментальных групп 5-6;
2) последовательности из экспериментальных групп (14-15);
3) последовательности из групп 5-15;

Данное разбиение использовалось так как модели, обученные в конкурсе, не могут надежно предсказывать попадание последовательностей в группы активности с номерам 1-4 и 16-17. Для каждой группы мы выбирали только 10000 последовательностей для ускорения подсчета.

Как видим из **рис. 44**, требуемый результат достигается при внесении порядка 300 мутаций.

[8] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: **Penzar D.**, Nogina D., Noskova E., Zinkevich A., Meshcheryakov G., Lando A., Rafi A.M., de Boer C., Kulakovskiy I.V. LegNet: a best-in-class deep learning model for short DNA regulatory regions // Bioinformatics.– 2023.– Vol. 39, № 8. doi: 10.1093/bioinformatics/btad457. JIF (для WoS) = 4.4 (0.95/0.45)

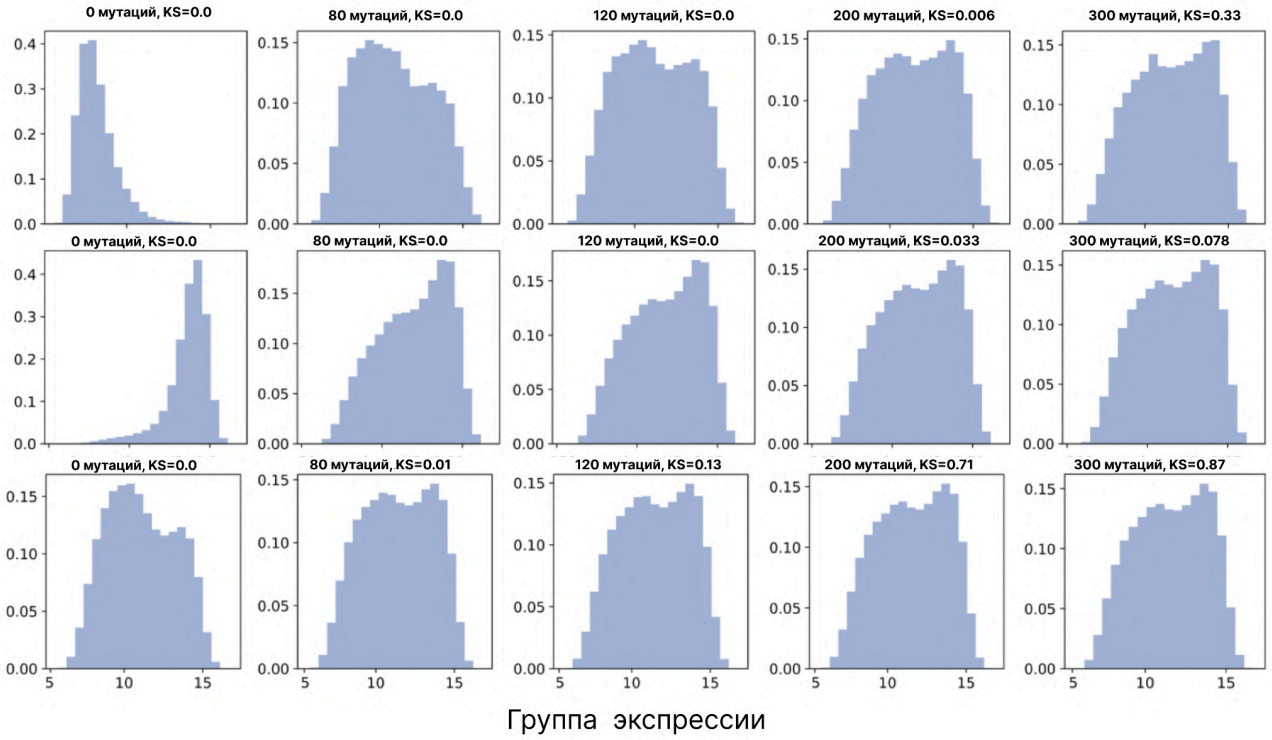


Рисунок 44. Оценка количества мутаций, достаточных для изменения распределения предсказанной экспрессии набора стартовых последовательностей в набор случайных последовательностей. Первый столбец (1): Распределения активности регуляторных последовательности, полученные путем случайной выборки 10 000 промоторов из определенных ячеек данных GPRA. Следующие столбцы (2-5): распределение выражений после введения заданного количества мутации в каждой последовательности. Верхний ряд: последовательности, выбранные из ячеек экспрессии номер. 5-6; средний ряд: последовательности, отобранные из 14-15 ячеек экспрессии; нижний ряд: последовательности, отобранные из 5-15 ячеек экспрессии; KS: тест Колмогорова-Смирнова сравнение эмпирического распределения с предсказанным выражением для случайного набора последовательности.

4.4.3. Архитектура диффузионной модели

Задачей генеративной модели во время обучения является восстановление исходной последовательности на основе поданной на вход зашумленной последовательности, изначальной экспрессии и числа шагов зашумляющего процесса.

Для получения архитектуры генеративной модели из архитектуры предсказательной изменению был подвергнут Mapping block изначальной архитектуры, который был заменен на одну поточечную свертку, которая редуцирует число каналов до 4 (**рис. 45 B**). Предполагается, что каждый из каналов содержит ненормированные вероятности соответствующего нуклеотида.

На вход же архитектура принимает 6-ти канальное одномерное изображение, 4 первых канала которого кодируют нуклеотиды, 5й канал (канал экспрессии) используется для передачи информации об экспрессии исходной последовательности, а 6й канал (канал мутаций) сообщает модели число шагов зашумляющего процесса, примененного к исходной последовательности (**рис. 45 C**).

4.4.4. Обучение диффузионной модели

В ходе обучения модели ей передавались последовательности, зашумленные при помощи диффузионного процесса (число шагов выбиралось случайно от 0 до 300). В качестве функции ошибки использовалась перекрестная энтропия между исходной последовательностью и предсказанными диффузионной моделью вероятностями нуклеотидов в каждой позиции.

4.4.5. Схема генерации последовательностей при помощи диффузионной модели

Генерация последовательностей с заданной активностью при помощи диффузионной модели осуществляется по следующей схеме (**Рис. 46**):

1. Генерируем случайную нуклеотидную последовательность.

Для i от T до 1:

2. Подаем ее на вход диффузионной модели, сообщаем модели, что подаем эту последовательность как результат i-го шага прямого диффузионного процесса, примененного к последовательности с целевой экспрессией E;

3. Полученную последовательность при помощи прямого диффузионного процесса зашумляем до шага i-1-shift;

На выходе получаем итоговую сгенерированную последовательность.

Положительное значение параметра дополнительного сдвига (shift) необходимо для того, чтобы заставить модель вносить большое число изменений за итерацию, дальше уходя от исходной случайной последовательности в сторону последовательностей с заданной экспрессией. Данная модификация была введена в ходе предварительного тестирования и приводит к лучшему качеству генерации. По всей видимости, ее необходимость связана с дискретной природой данных.

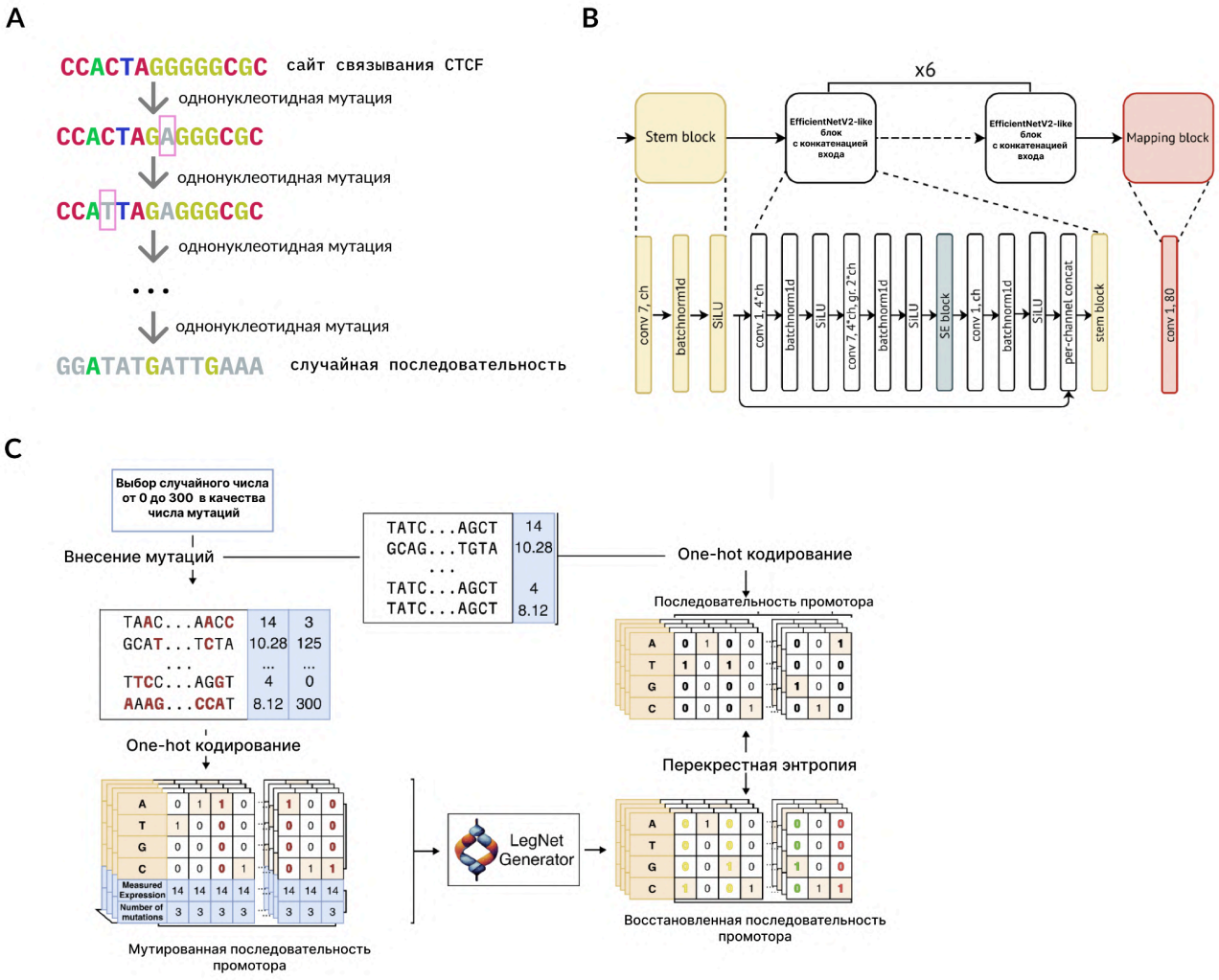


Рисунок 45. A. Зашумляющий процесс LegNet-диффузии. **B.** Архитектура генеративной модели. Принципиальное отличие от архитектуры предсказательной модели заключается в mapping block, выделенном красным. **C**. Схема обучения генеративной модели. В выходе модели зеленым покрашены правильно восстановленные позиции исходной последовательности; желтым – позиции, которые были неправильно определены как измененные; красным – позиции, которые остались нескорректированными. Для простоты иллюстрации предполагается, что выходом модели является one-hot закодированная последовательность, а не вероятности нуклеотидов.

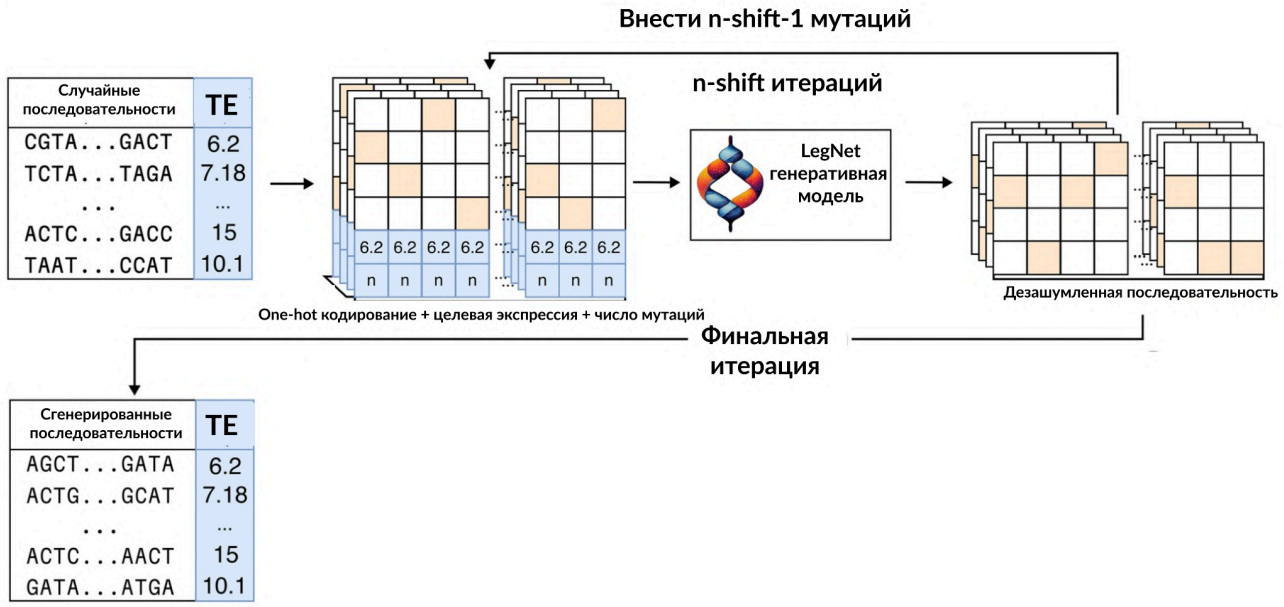


Рисунок 46. Схема генерации последовательностей с заданной экспрессией при помощи холодной диффузионной модели. TE – целевая экспрессия.

4.4.6. Оценка качества генерации регуляторных последовательностей дрожжей

Согласно описанной выше процедуре, вначале была обучена диффузионная модель с использованием данных конкурса DREAM-2022. После этого данная модель была применена для генерации последовательностей с заданной экспрессией.

Для вычислительной валидации качества генерации последовательностей мы использовали предсказательную модель, оценив при ее помощи влияние на экспрессию для сгенерированных последовательностей. Корреляции Пирсона и Спирмена между запрошенными значениями и предсказанными достигли значений 0.839 и 0.843, что свидетельствует о хорошем соответствии запрошенной и достигнутой активности (**рис. 47**).

Модель просили генерировать только последовательности из 5-17 экспрессионных бинов, т.к для меньших номеров бинов было слишком мало данных для обучения, и, кроме того, достоверность сгенерированных последовательностей нельзя было бы оценить при помощи предсказательной модели в силу её недостатков, упомянутых ранее.

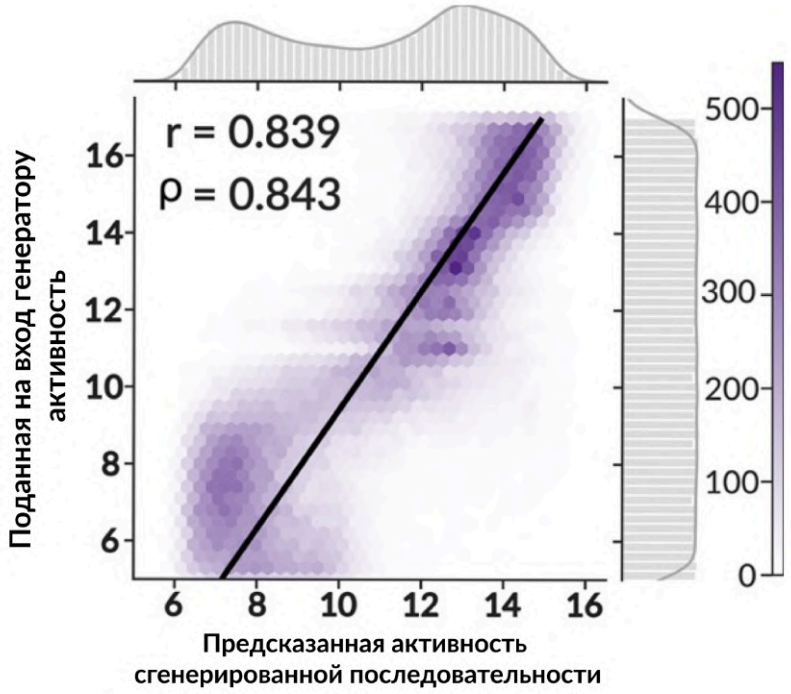


Рисунок 47. Корреляция между желаемой (целевой) и наблюдаемой (предсказанной LegNet) экспрессией для 110 592 сконструированных промоторов; цветовая шкала обозначает количество промоторов в контейнере; Коэффициенты корреляции Пирсона и Спирмена показаны в верхнем левом углу.

4.5. Предсказание активности регуляторных элементов человека[9]

4.5.1. Представление входных данных нейросетевой модели

Ко всем 200 п.о последовательностям, активность которых измерялась в МПРЭ, были добавлены последовательности адаптеров (5’ AGGACCGGATCAACT и 3′ – CATTGCGTGAACCGA), доводя итоговый размер последовательности, подаваемой в нейронную сеть, до 230 п.о.

Подаваемая на вход нейронной сети последовательность кодировалась в 4-мерный вектор с использованием one-hot кодирования. Была проведена аугментация данных путем добавления для каждой последовательности ее обратно-комплементарного аналога.

При оптимизации модели рассматривалась возможность добавления канала **is_reverse**, указывающего, подана ли последовательность в прямой или обратно комплементарной форме, заполняя канал 0 или 1 соответственно.

Также при обучении использовалась аугментация сдвигом последовательности, в ходе которой последовательность случайным образом сдвигалась на расстояние от 0 до 21 нуклеотида в сторону старта транскрипции.

Помимо этого сравнивалось качество моделей, обученных на усредненных активностях для прямой и обратной ориентаций (как предлагалось авторами данных) или обученных предсказывать активности для прямой и обратной ориентации без усреднения.

На этапе тестирования модель предсказывала 4 значения:

1) активность для прямой и обратной ориентации элемента относительно фиксированных фланкирующих (т.е. адаптерных) областей;

2) в качестве дополнительной аугментации — значения для полных обратных комплементарных последовательностей (т.е. получение обратной комплементарной последовательности элемента вместе с адаптерными областями) из шага (1).

Окончательный прогноз представлял собой среднее этих четырёх значений.

Во время тестирования не использовалась аугментация сдвигом, так она не приводила к улучшению предсказаний модели.

[9] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: Agarwal V., Inoue F., Schubach M., **Penzar D.**, Martin B.K., Dash P.M., Keukeleire P., Zhang Z., Sohota A., Zhao J., Georgakopoulos-Soares I., Noble W.S., Yardımcı G.G., Kulakovskiy I.V., Kircher M., Shendure J., Ahituv N. Massively parallel characterization of transcriptional regulatory elements // Nature.– Springer Science and Business Media LLC, 2025.– P. 1–10. doi: 10.1038/s41586-024-08430-9. JIF (для WoS) = 50.5, (2.75/0.25)

4.5.2. Адаптация архитектуры LegNet

Для работы с регуляторными участками человека в изначальную архитектуру нейронной сети LegNet (описанную в разделе 4.3.3) были внесены изменения (**рис 48. А)**:

1) после каждого EfficientNetV2-like блока был добавлен слой пулинга, который позволил увеличить рецептивное поле нейронной сети, что необходимо в связи с большим размером последовательностей;

2) размер ядра свертки и число блоков были подобраны таким образом, чтобы рецептивное поле модели соответствовало длине последовательности;

3) в связи с тем, что задача стала истинно регрессионной, был удален слой soft-argmax и добавлен дополнительный линейный слой после поканального усреднения.

Модификации (1) и (2), помимо прочего, позволили уменьшить число параметров модели, что было необходимо в связи с меньшими размерами обучающих наборов.

Для всех датасетов использовалось число сверток в четырех Efficient-like блоках 80, 96, 112 и 128 соответственно. Для датасетов HepG2 и K562 использовался размер ядра свертки, равный 11 для первого (stem) сверточного блока и 9 для остальных. В случае датасета WTC11 из-за его меньшего размера использовались ядра свертки размера 5 и 3 для первого и последующих слоев соответственно. Это привело к уменьшению рецептивного поля сети по сравнению с размером последовательности, но вместе с тем привело и к улучшению качества за счет уменьшения переобучения. Размеры датасета при этом, по-видимому, не позволяли выучить дальние взаимодействия. Финальная модифицированная архитектура для данной задачи получила название MPRALegNet.

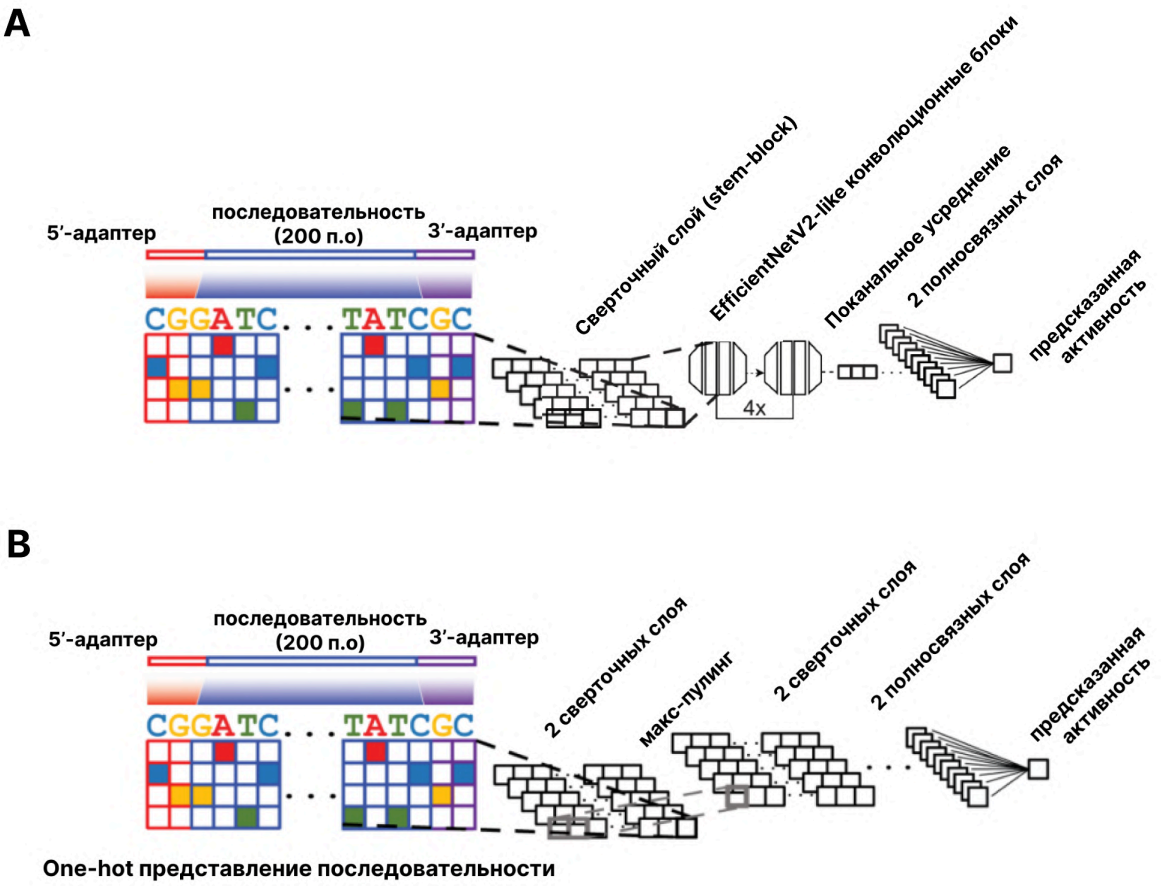


Рисунок 48. А. Архитектура MPRALegNet. **В.** Архитектура MPRAnn. Последовательность длины 230 п.о кодируется при помощи one-hot кодирования и подается на вход нейронной сети, состоящей из двух сверточных блоков, выход которых подается на вход слою макспулинга, а его выход, в свою очередь, пропускается еще через два сверточных блока. Их выход преобразуется в одномерный вектор и пропускается через два полносвязных слоя.

4.5.3. Подбор гиперпараметров модели

Подбор параметров осуществлялся при помощи вложенной кросс-валидации. Так как даже такая схема склонна к утечке данных, перебирался лишь ограниченный набор возможных значений гиперпараметров модели и схемы ее обучения.

К ним относились:

1) Использование аугментации сдвигом;

2) Аугментация за счет замены на обратно-комплементарную последовательность (с переворачиванием фланков);

3) Обучение модели на активностях, полученных усреднением активностей для прямой и обратной ориентацией или же обучение на не усредненных значениях;

4) Добавление канала с информацией о том, заменялась ли последовательность на обратно-комплементарную с переворачиванием фланков;

5) Необходимость аугментации во время тестирования.

Финальная версия модели для последующих анализов включала все аугментации, кроме добавления канала с ориентацией последовательности (**рис. 49).**

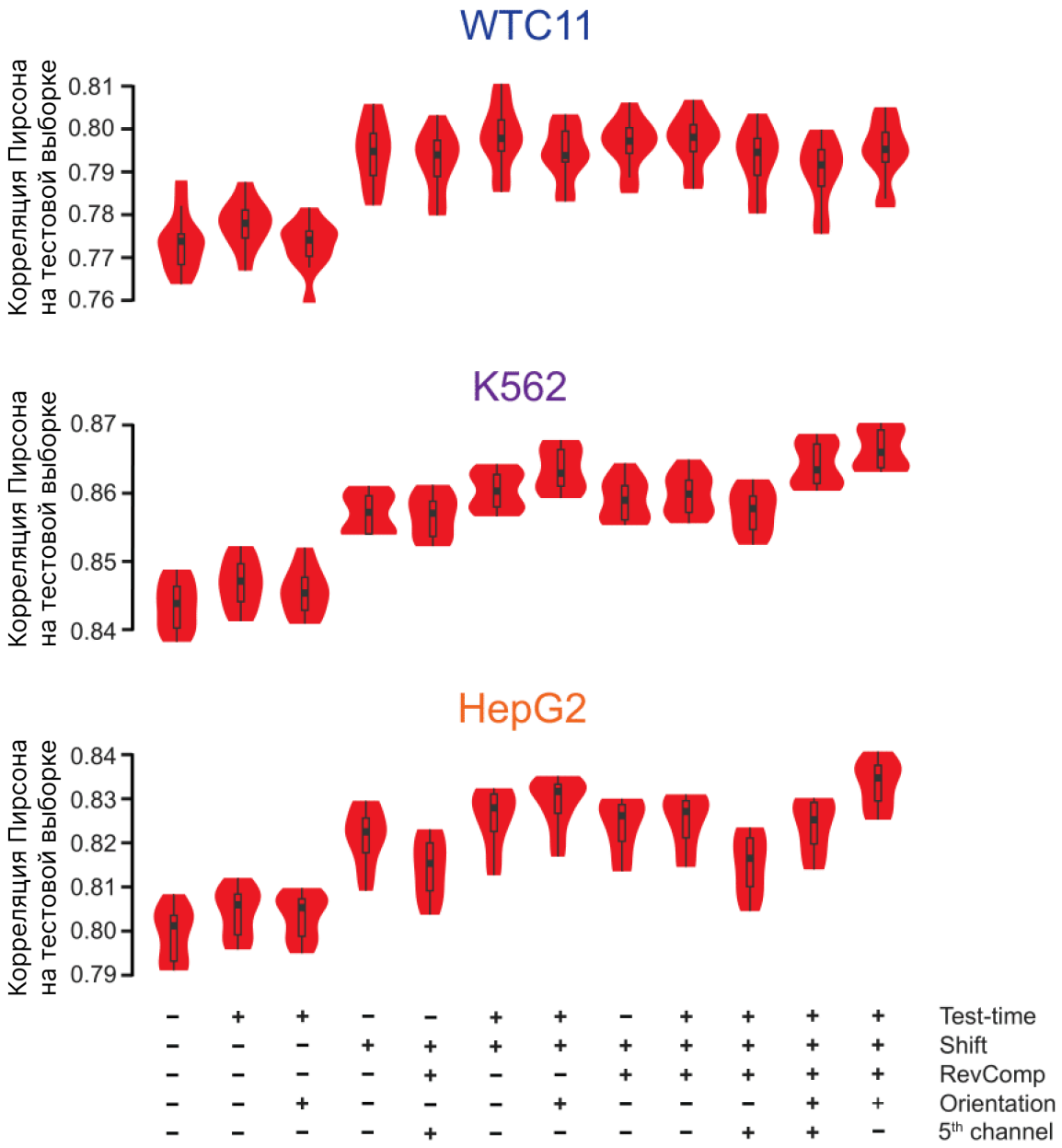


Рисунок 49. Скрипичные диаграммы, показывающие качество различных вариантов MPRALegNet на каждом из десяти фолдов кросс-валидации на тестовых данных, для разных типов аугментаций. «-» и «+» обозначают удаление или использование, соответственно, следующих аугментаций: 1) «test-time», при которой среднее предсказание вычисляется для различных аугментаций тестовой последовательности; 2) «shift», при которой последовательность случайным образом сдвигалась на 0 до +21 п.н. в сторону старта транскрипции; 3) «RevComp», при которой последовательность случайным образом заменялась на обратную комплементарную (включая адаптеры); 4) «Orientation», при которой модель обучалась на активности элемента для каждой ориентации отдельно, вместо усреднения по обеим ориентациям; и 5) «5th channel», при которой наряду с one-hot кодированием последовательности (т.е. первые 4 канала) использовался 5-й канал для указания, заменялась ли последовательность на обратно-комплементарную с переворачиванием фланков.

4.5.4. Независимые библиотеки

На независимых библиотеках все модели, предсказывающие сигнал на основе последовательности, смогли превзойти биохимическую модель. Что еще более важно, предложенная нами архитектура MPRALegNet показала лучшее качество в двух из трех типов клеток (**рис. 50 A,C**).

Стоит отметить, что качество EnformerMPRA и SeiMPRA может быть завышено по сравнению с биохимической моделью, так как они используют большее число признаков и информацию о дополнительных типах клеток и эпигенетических сигналах. Более того, так они обучались практически на всем геноме, у них была возможность наблюдать эпигенетические сигналы, ассоциированные с последовательностями из тестовых наборов.

Анализ, проведенный на подвыборках исходных датасетов показал, что качество модели MPRALegNet может быть улучшено при увеличении выборки (**рис 50. B**), причем качество модели зависит от данных по отрицательному степенному закону, что может соответствовать “закону шкалирования”, экспериментально показанному для многих моделей машинного обучения [254].

4.5.5. Анализ регуляторной грамматики, выученной моделью

Для того чтобы понять, какие регуляторные принципы удалось выучить нашей модели на независимых библиотеках, мы провели in silico мутагенез на каждой независимой библиотеке и затем использовали метод TF-MoDISco[209] для того, чтобы выделить паттерны, на которые обращает внимание нейронная сеть. Это позволило идентифицировать большой набор мотивов связывания транскрипционных факторов, относящихся к генам домашнего хозяйства, и, как предполагается, активирующих экспрессию во всех клеточных типах, включая факторы NRF1, USF1/2, TFEB/TFE3 и семейства факторов, ассоциированные с JUN/FOS, KLF(KLF/SPs), C/EBP и ETS. Был найден и фактор REST, известный репрессор транскрипции [333] (**Приложение 1**).

Три наиболее частых связывающих сайта транскрипционных факторов (TFBS), ассоциированных с активацией транскрипции во всех типах клеток, были мотивами, связанными с KLF, ETS и CTCF. В отличие от этого, наиболее специфичными для типа клеток были паттерны связывания HNF4A/G в клетках HepG2, димера GATA::TAL1 в клетках K562, и составной элемент, связываемый димером POU5F1::SOX2 в клетках WTC11 (**рис. 50 D)**.

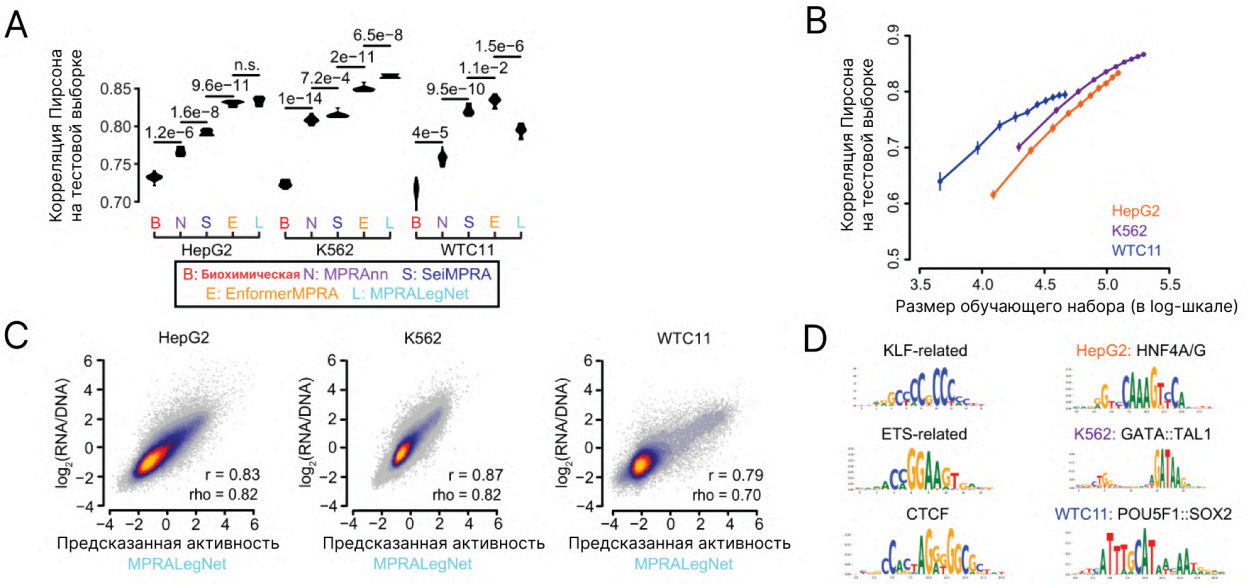


Рисунок 50. A. Скрипичные диаграммы, показывающие качество моделей, основанных на последовательностях, и биохимических моделей на десяти фолдах кросс-валидации, с улучшением по сравнению с предыдущей моделью, оценённым с помощью одностороннего парного t-теста. **B.** Влияние размера обучающего набора на качество модели. Данные для каждого типа клеток были уменьшены до каждого 10-го процентиля (т.е. от 10 до 100%). Для каждой точки отложены стандартное отклонение корреляций Пирсона по 90 моделям (10 фолдов тестовых данных x 9 обученных моделей, различающихся выбором валидационного набора). **C.** Диаграмма рассеивания, показывающая связь между предсказаниями MPRALegNet и наблюдаемыми показателями активности элементов для каждого типа клеток. **D.** Набор обогащённых мотивов, обнаруженных с помощью TF-MoDISco; слева показаны три основных мотива, обнаруженных в нескольких типах клеток, а справа — основной мотив, обнаруженный для каждого типа клеток.

Далее для каждой клеточной линии были выбраны топ-10 матриц ТФ, выделенных TF-MoDISCO [209]. Каждая регуляторная последовательность промотора и энхансера из независимой библиотеки была проверена при помощи FIMO [334] на наличие сайтов связывания данных ТФ. В результате для каждой последовательности был получен вектор, характеризующий число сайтов различных ТФ, в ней встретившихся.

Для того чтобы исследовать то, насколько хорошо модель выучила зависимость между числом сайтов данного ТФ и активностью регуляторной последовательности, нами были отобраны последовательности, имеющие сайты связывания только одного ТФ. Отобранные последовательности были разбиты на группы по числу сайтов данного ТФ, и группы из менее чем 10 последовательностей были отсеяны. Далее для каждой оставшейся группы была посчитана экспериментальная и предсказанная медианные активности. Оказалось, что для всех клеточных линий модель достаточно точно выучила зависимости активности регуляторного элемента от числа сайтов связывания (**рис. 51. A-C**).

При этом для большинства ТФ наблюдалась мультипликативная (в логарифмических координатах, соответственно, лог-аддитивная) зависимость активности регуляторной

последовательности от числа сайтов конкретного ТФ. Тем не менее для части, например, STAT1/4/5A/5B, наблюдалась субмультипликативная зависимость в силу насыщения эффекта, оказываемого добавлением нового сайта ТФ, на экспресию (**рис. 51. G**). Модель MPRALegNet смогла корректно воспроизвести данную динамику.

Для каждой возможной пары из ТФ, отобранных при помощи TF-MoDISco [209] на прошлом этапе, были взяты только последовательности, содержащие не более одного сайта связывания каждого из ТФ пары и не содержащие сайтов связывания других ТФ.

Далее каждая пара ТФ была охарактеризована при помощи коэффициента при члене, отражающем взаимодействие двух ТФ (ТФ1 * ТФ2) в линейной регрессии, предсказывающей активность регуляторного элемента как функцию присутствия мотива связывание ТФ1, ТФ2 и их взаимодействия:

$$log(RNA/DNA) \sim \text{ТФ1} + \text{ТФ2} + \text{ТФ1} * \text{ТФ2}$$

Соответственно, отрицательное значение коэффициента при взаимодействии говорит о субмультипликативном характере зависимости, а положительное – о супермультипликативном. Для оценки того, насколько модель MPRALegNet может учитывать взаимодействие факторов, на место экспериментального $log(RNA/DNA)$ было подставлено значение, предсказываемое моделью. Оказалось, что коэффициенты, получаемые для экспериментальных и предсказанных моделью значений активности, хорошо коррелируют друг с другом (**рис 51. D-E**), что говорит о том, что модель смогла выучить нелинейные взаимодействия между ТФ. На **рис. 51, H** приведен пример субмультипликативного взаимодействия ТФ HNF4A/G и ТФ NFYA/C, а на **рис 51, I** – сверхмультпликативного между факторами NRF1 и FOS::JUN. MPRALegNet отлично видит оба взамодействия.

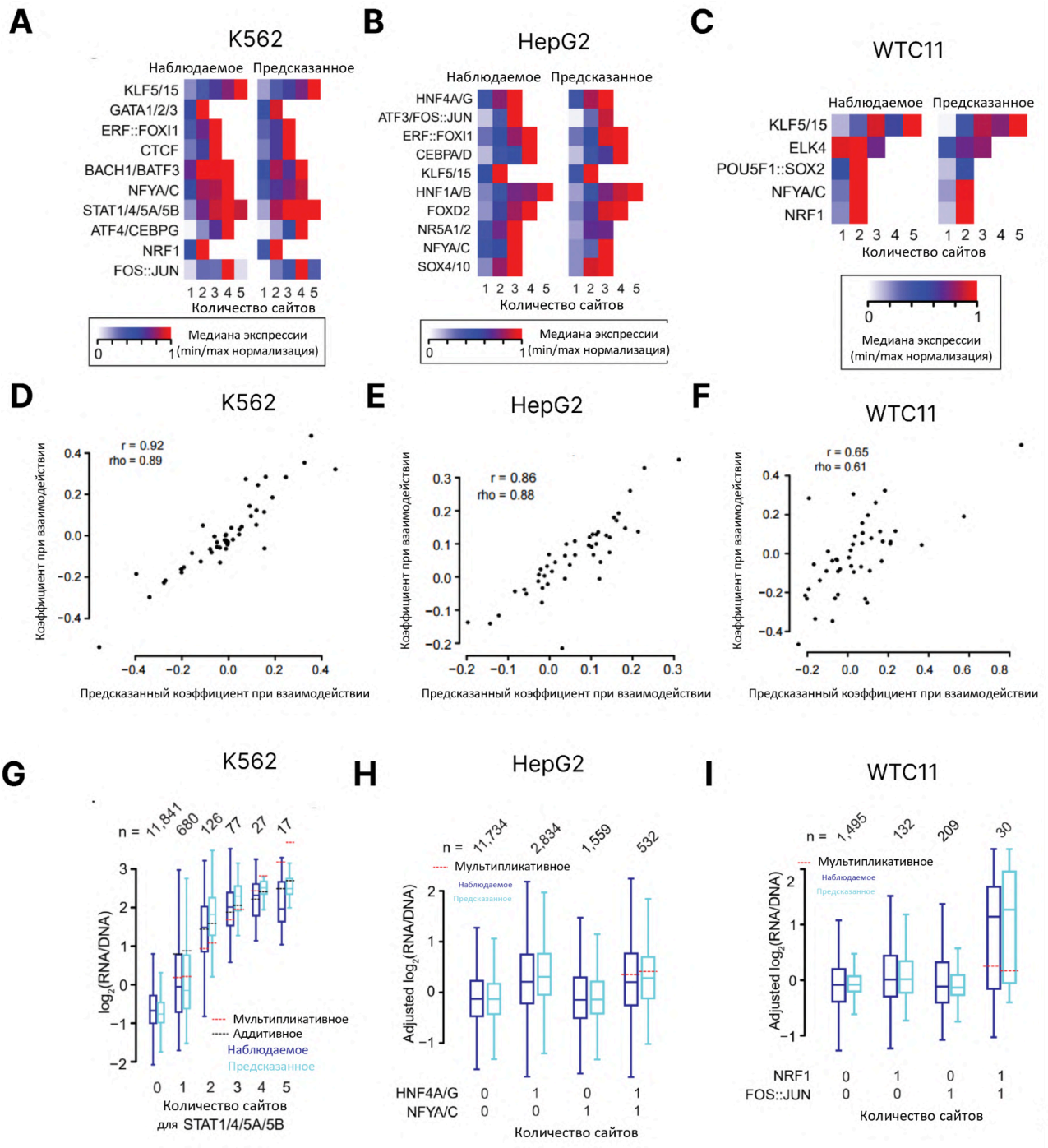


Рисунок 51. A-C. Медианы активностей (экспериментальных и предсказанных MPRALegNet) регуляторных последовательностей из трех клеточных линий, содержащих данное количество предсказанных сайтов связывания ТФ. **D-E**. Корреляция между коэффициентами при взаимодействии пар ТФ в регрессионных моделях, построенных на основе экспериментальных и предсказанных активностей. **G.** Зависимость экспериментальной и предсказанной активности регуляторной последовательности в клеточной линии K562 от количества предсказанных сайтов связывания ТФ STAT1/4/5A/5B. **H.** Субмультипликативный характер взаимодействия факторов HNF4A/G и NFYA/C. **I.** Сверхмультипликативный характер взаимодействия факторов NRF1 и FOS::JUN.

4.5.6. Предсказание аллель-специфичных событий

Для предсказания аллель-специфичных события при помощи MPRALegNet использовалась следующая процедура:

1) для заданного SNP (однонуклеотидного полиморфизма) получали 230 п.о. последовательности для референсного и альтернативного вариантов путем вырезания из генома окна в 230 п.о., центрированного на SNP;

2) для референсной и альтернативной последовательностей считались предсказания при помощи 90 моделей, обученных в ходе вложенной кросс-валидации на соответствующем датасете;

3) предсказания выполнялись для последовательностей в прямой и обратной ориентациях, таким образом получалось 180 предсказаний для референсной последовательности и 180 для альтернативной;

4) на основе этих предсказаний считалась средняя разница между предсказанными активностями для референсной и альтернативной последовательностей PredRef-PredAlt, а также дисперсии предсказаний для референсной и альтернативной последовательностей varRef и varAlt соответственно.

Эта процедура проводилась для 6 наборов аллель-специфичных событий в клетках K562/HepG2, обнаруженных в данных ChIP-Seq/ATAC-Seq/DNase-Seq с уровнем FDR 5% в соответствующих типах клеток в базах данных UDACHA (https://udacha.autosome.org, релиз IceKing) [139] и ADASTRA (https://adastra.autosome.org, релиз BillCipher) [34].

Аллель-специфичные rSNP не обязательно связаны с изменением экспрессии генов. Во многих случаях прогнозы модели для референсного и альтернативного аллеля ожидаются очень схожими, что ограничивает надежность оценки направления предсказанного эффекта. Чтобы учесть неопределенность при прогнозировании PredRef-PredAlt мы оценили стандартное отклонение, используя исходные оценки всех моделей ансамбля по формуле:

$$sd = \sqrt{varRef/N + VarAlt/N},$$

где N – это число моделей в ансамбле

В случае нейтральных вариантов распределение итоговых оценок должно следовать нормальному распределению со средним равным нулю и стандартным отклонением, оцененным по формуле выше, поэтому мы вычислили p-value для каждой PredRef-PredAlt на основе соответствующих значений z-статистики и использовали порог $p < 0.05$ для выбора предсказаний с высокой уверенностью.

Для всех шести тестируемых комбинаций источников ACC и типов клеток мы наблюдали значимые ассоциации между наблюдаемыми и предсказанными оценками как до, так и после

исключения случаев, в которых аллель-специфичное событие было незначимым или предсказания модели были слишком неопределенными (**рис. 52 А**).

Отсюда можно сделать вывод о том, что MPRALegNet успешно предсказывает аллель-специфичные события в соответствующих типах клеток.

4.5.7. Предсказание эффектов однонуклеотидных вариантов

Для того чтобы оценить качество предсказания моделью влияния однонуклеотидных замен в регуляторных участках человека (https://kircherlab.bihealth.org/satMutMPRA), были взяты эффекты всех однонуклеотидных вариантов в промоторах генов F9, LDLR, PKLR и энхансере гена SORT1 из работы по in vitro насыщающему мутагенезу в человеческих регуляторных элементах [115]. Данные регуляторные последовательности использовались так как измерение эффектов однонуклеотидных мутаций для них производилось в клеточных типах, на которых производилось обучение моделей.

Насыщающий мутагенез in silico для всех элементов был проведен с использованием MPRALegNet на основе координат GRCh38. Поскольку большинство элементов превышают длину 200 п.н., области были разделены на окна длиной 200 п.н. с шагом 150 п.н., что привело к перекрытию в 50 п.н. между соседними окнами. Предсказания усреднялись по перекрывающимся окнам. Последовательности, вырезанные из окна в 200 п.н. увеличивались до 230 п.н. путем добавления адаптеров из обучения с обеих сторон. Данные in silico и in vitro насыщающих мутагенезов сравнивались для всех вариантов, эффект которых был измерен в ходе эксперимента минимум 10 раз.

Сравнение предсказаний MPRALegNet для промотора PKLR с данными MPRA показало, что большинство значимых сайтов связывания транскрипционных факторов (например, GATA3, KLF9, SP5 и NFIB) детектируются, хотя предсказанные размеры эффектов относительно меньше для KLF4 и GATA2 (**рис. 52 В.**).

В целом, мы наблюдали корреляцию 0.49 для SORT1, 0.65 для PKLR, 0.66 для LDLR и 0.51 для F9 между модельными прогнозами и наблюдаемыми данными (**рис. 52 С.**), что подтверждает, что MPRALegNet, несмотря на обучение на активности cCRE, может частично моделировать регуляторные эффекты отдельных генетических вариантов. Эти результаты сопоставимы с результатами Enformer (0.63 для SORT1, 0.83 для PKLR, 0.62 для LDLR и 0.28 для F9).

Таким образом, MPRALegNet может быть использован для предсказания эффекта однонуклеотидных замен.

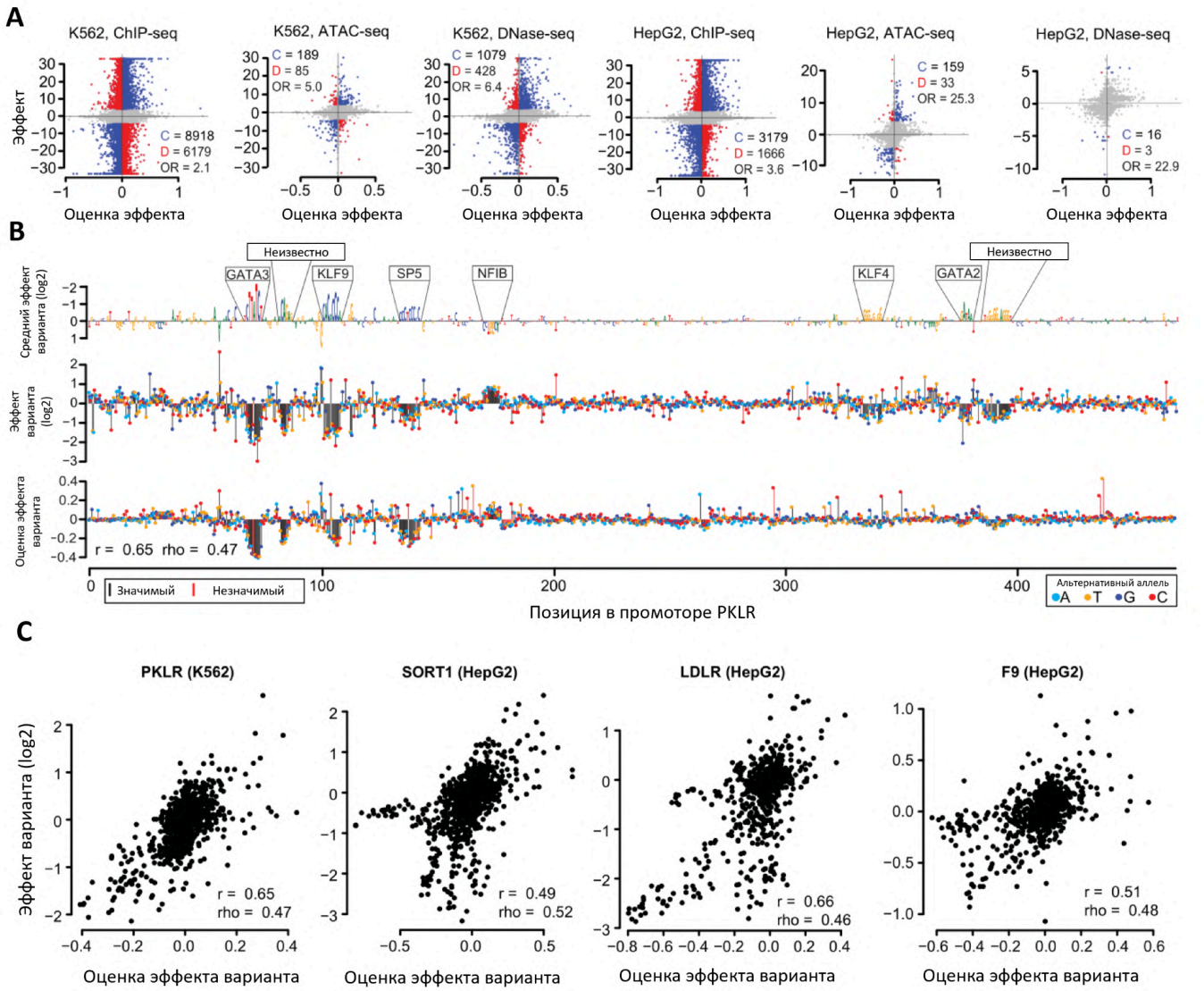


Рисунок 52. a. Диаграммы рассеяния предсказанных эффектов вариантов и наблюдаемых аллель-специфичных событий, обнаруженных в данных ChIP-seq, ATAC-seq и DNase-seq в клетках K562 и HepG2. Указано количество случаев, в которых предсказания и наблюдения совпадают (C; синий), не совпадают (D; красный), или не рассматриваются (серый), так как либо FDR для ACC > 0.05, либо предсказания модели слишком незначимо отличаются от 0 (p-значение > 0.05). Также указано соответствующее отношение шансов (OR), подсчитанное точным тестом Фишера. **b.** Данные насыщенного мутагенеза для промотора PKLR. Верхняя строка представляет собой референсную последовательность, масштабированную по среднему размеру эффекта среди всех альтернативных мутаций, с аннотацией значимых сайтов связывания транскрипционных факторов (TFBS), которые соответствуют известным мотивам. Измеренные размеры эффектов отдельных вариантов показаны во второй строке, а в нижней строке представлены прогнозы MPRALegNet с соответствующими значениями корреляции Пирсона (r) и Спирмена (rho). **c.** Диаграммы рассеяния, демонстрирующие корреляцию между предсказанными MPRALegNet эффектами генетических вариантов и наблюдаемыми эффектами вариантов, как было выявлено в эксперименте с насыщенным мутагенезом, тестирующем промотор PKLR, энхансер SORT1, промотор LDLR и промотор F9.

4.5.8. Общая библиотека

В случае общей библиотеки задачей модели было предсказать не абсолютную экспрессию в каждой клеточной линии, а “специфичность” – разницу между экспрессией в данной клеточной линии и средним. В данной постановке, опять же, биохимическая модель оказалось худшей. MPRALegNet также выступил лучше MPRAnn и практически наравне с SeiMPRA. Лучшей же моделью в данной постановке оказалась EnformerMPRA (**рис. 53**).

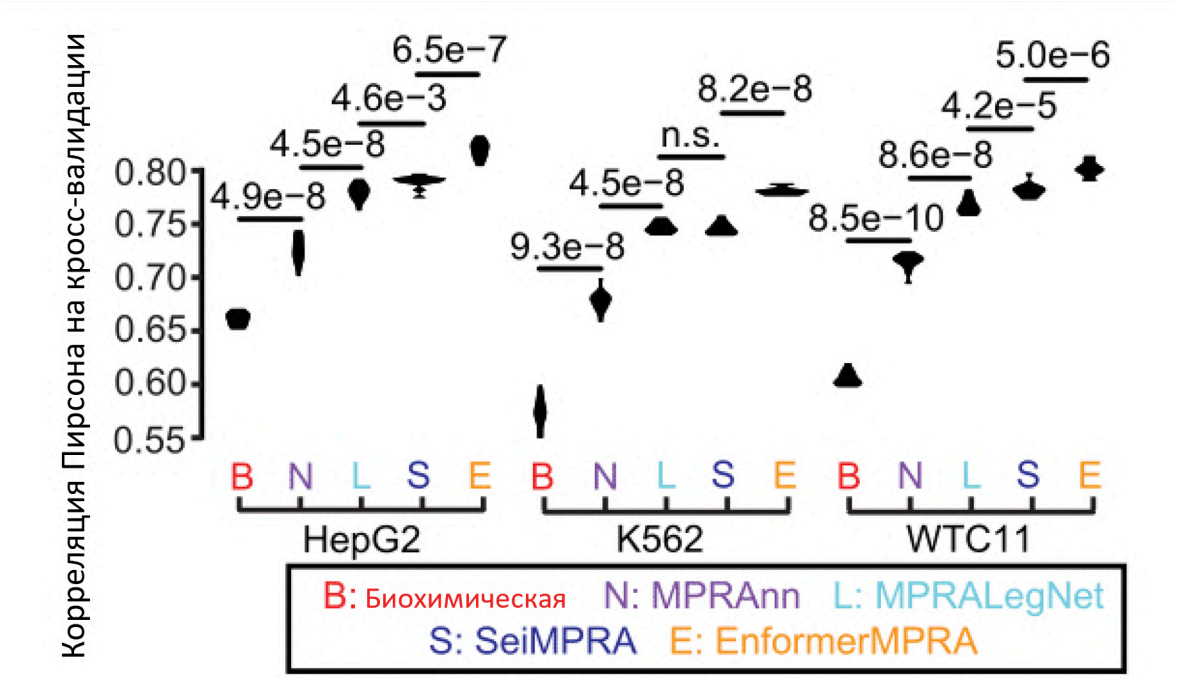


Рисунок 53. Качество обученных моделей MPRAnn, MPRALegNet, SeiMPRA и EnformerMPRA на каждом из десяти фолдов кросс-валидации для отложенных данных с оценкой улучшения относительно предыдущей модели с использованием одностороннего парного t-теста.

4.5.9. Использование признаков Enformer в LegNet

Далее мы решили проверить, позволит ли добавление предсказаний из Enformer в качестве признаков в модель LegNet улучшить качество предсказания.

Для того чтобы получить признаки из Enformer, мы использовали процедуру, похожую на описанную для EnformerMPRA, но усреднение признаков производилось только по предсказаниям для четырех центральных бинов, так все остальные соответствовали нуклеотидам из паддинга.

Архитектура MPRALegNet была модифицирована следующим образом:

1) Был добавлен отдельный линейный слой, сжимающий предсказания от Enformer в 16 признаков;

2) Эти 16 признаков конкатенируются к признакам, получаемым после слоя поканального усреднения.

Данная модификация модели получила название Legformer. В процедуру обучения дополнительных изменений не вносилось.

В случае независимых библиотек (**рис. 54**), данная модель достигла такого же качества на клеточной линии WTC11, что и MPRAEnformer. Что более удивительно – модель показала лучшее качество на двух других клеточных линиях, даже на K562, где MPRALegNet уже превосходила качество MPRAEnformer. В случае общей библиотеки Legformer также показала качество или сравнимое, или лучше MPRAEnformer (**рис. 55**).

Таким образом, признаки Enformer содержат информацию, выученную в ходе обучения предсказанию эпигенетических сигналов, дополняющую ту, которую можно выучить непосредственно из МПРЭ.

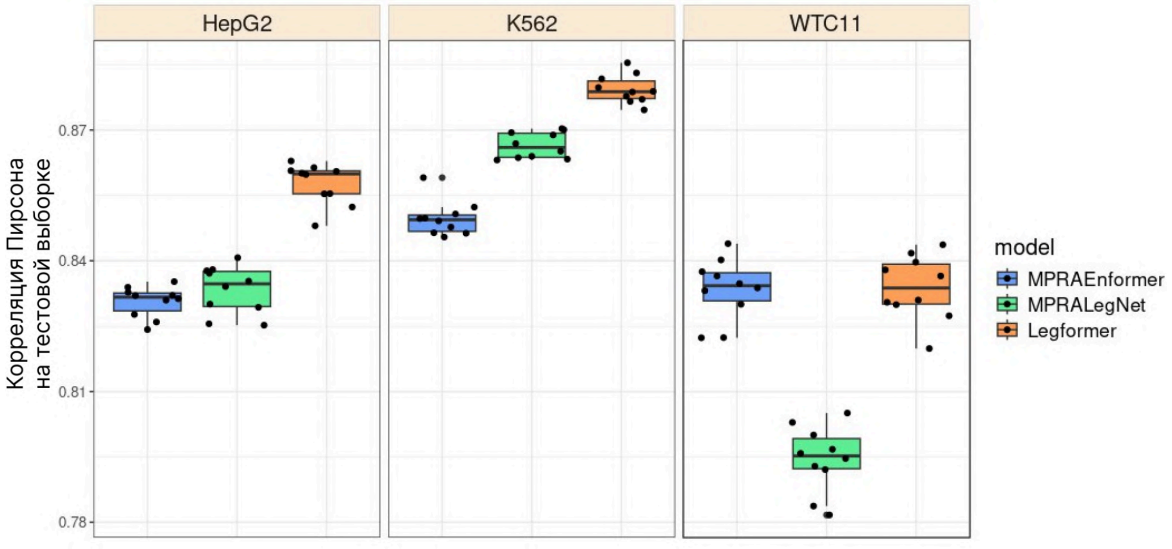


Рисунок 54. Сравнение качества (корреляции Пирсона) моделей MPRAEnformer, MPRALegNet и Legformer на десяти фолдах кросс-валидации на независимых библиотеках регуляторных участков

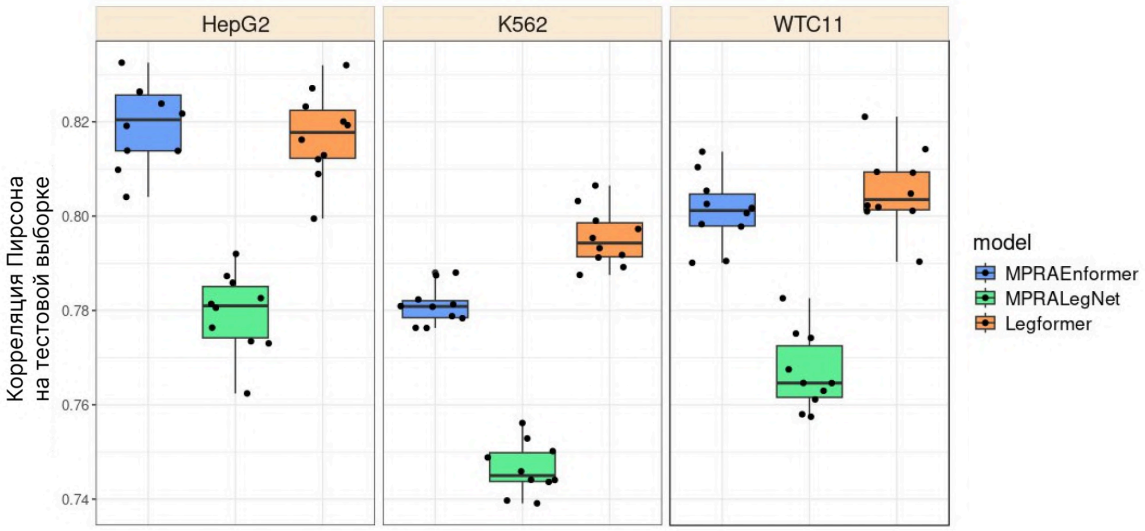


Рисунок 55. Сравнение качества (корреляции Пирсона) моделей MPRAEnformer, MPRALegNet и Legformer на десяти фолдах кросс-валидации на объединенной библиотеке регуляторных участков.

5. Заключение

В работе была разработана инновационная нейросетевая архитектура для предсказания активности регуляторных регионов и предложена методика ее обучения. Это позволило проанализировать данные массовых параллельных репортерных экспериментов, предсказать влияние однонуклеотидных замен на активность регуляторных регионов и участки аллель-специфичного связывания, а также интерпретировать полученные предсказания при помощи насыщающего мутагенеза *in silico*. Была показана возможность адаптации представленной архитектуры для генерации регуляторных регионов с заданной активностью.

В работе было показано, что данные насыщающего мутагенеза могут быть использованы для валидации моделей машинного обучения, но их использование для дообучения сопряжено с утечкой данных и, следовательно, с завышенной оценкой качества предсказаний. Было продемонстрировано, что переносимость предсказаний моделей на независимые экспериментальные данные ограничена и качество предсказаний может не превосходить таковое у случайной модели.

Было показано, что использование признаков, основанных на нейронных сетях, обученных предсказывать эпигенетические разметки генома, помогает моделям машинного обучения предсказывать события аллель-специфичного связывания.

В дальнейшем представляется перспективным создание мультимодальной модели, которая будет комбинировать данные из различных экспериментов, включая эпигенетические разметки, персонализированные геномы, данные экспериментов с единичными клетками, информации о трехмерной структуре хроматина и массовых экспериментов с репортерами. Мы предполагаем, что этот подход сможет преодолеть ограничения существующих моделей и позволит улучшить точность предсказания эффектов индивидуальных вариаций и упростит создание генноинженерных конструкций с заданной клеточной специфичностью.

6. Основные результаты и выводы

1. Обучение и тестирование вычислительных моделей для предсказания эффектов регуляторных вариантов на результатах массовых параллельных репортерных экспериментов с мутагенезом насыщающим ПЦР приводит утечке информации и значительному завышению оценки качества предсказаний. При тестировании на результатах независимых экспериментов такие модели демонстрируют значительное снижение точности предсказаний.
2. Достаточный объем учебной выборки для модели на основе случайного леса позволяет получать достоверные предсказания участков аллель-специфичного связывания в геноме для хорошо изученных типов клеток и факторов транскрипции. В качестве признаков необходимы как генерируемые полногеномными нейросетевыми моделями, так и оценки эффекта замен, полученные с помощью традиционных моделей мотивов связывания транскрипционных факторов.
3. Использование современных достижений в области дизайна и обучения моделей глубокого обучения позволило построить новую полносверточную нейросетевую архитектуру LegNet, хорошо подходящую для предсказания активности регуляторных регионов эукариот и эффектов однонуклеотидных вариантов по данным массовых параллельных экспериментов с репортерами. В этих задачах LegNet превосходит и традиционные биоинформатические подходы, и альтернативные нейросетевые решения. Адаптация LegNet на основе метода холодной диффузии позволяет создавать промоторные последовательности для достижения заданного уровня экспрессии целевого гена.

Научные статьи по теме диссертации, опубликованные в журналах SCOPUS, WOS, RSCI[10]

1. Agarwal V., Inoue F., Schubach M., **Penzar D**., Martin B.K., Dash P.M., Keukeleire P., Zhang Z., Sohota A., Zhao J., Georgakopoulos-Soares I., Noble W.S., Yardımcı G.G., Kulakovskiy I.V., Kircher M., Shendure J., Ahituv N. Massively parallel characterization of transcriptional regulatory elements // Nature.– Springer Science and Business Media LLC, 2025.– P. 1–10. doi: 10.1038/s41586-024-08430-9. JIF (для WoS) = **50.5**, (2.75/0.25)

2. Rafi A.M., Nogina D., **Penzar D.**, Lee D., Lee D., Kim N., Kim S., Kim D., Shin Y., Kwak I.-Y., Meshcheryakov G., Lando A., Zinkevich A., Kim B.-C., Lee J., Kang T., Vaishnav E.D., Yadollahpour P., Random Promoter DREAM Challenge Consortium, Kim S., Albrecht J., Regev A., Gong W., Kulakovskiy I.V., Meyer P., de Boer C.G. A community effort to optimize sequence-based deep learning models of gene regulation. // Nat. Biotechnol.– 2024. doi: 10.1038/s41587-024-02414-w. JIF (для WoS) = **33.1** (1.5/0.30)

3. **Penzar D.**, Nogina D., Noskova E., Zinkevich A., Meshcheryakov G., Lando A., Rafi A.M., de Boer C., Kulakovskiy I.V. LegNet: a best-in-class deep learning model for short DNA regulatory regions // Bioinformatics.– 2023.– Vol. 39, № 8. doi: 10.1093/bioinformatics/btad457. JIF (для WoS) = **4.4** (0.95/0.45)

4. Abramov S., Boytsov A., Bykova D., **Penzar D.**, Yevshin I., Kolmykov S.K., Fridman M.V., Favorov A.V., Vorontsov I.E., Baulin E., Kolpakov F., Makeev V.J., Kulakovskiy I.V. Landscape of allele-specific transcription factor binding in the human genome // Nat. Commun.– 2021.– Vol. 12, № 1.– P. 2751. doi: 10.1038/s41467-021-23007-0. JIF (для WoS) = **14.7** (1.20/0.20)

5. Ambrosini G., Vorontsov I., **Penzar D**., Groux R., Fornes O., Nikolaeva D.D., Ballester B., Grau J., Grosse I., Makeev V., Kulakovskiy I., Bucher P. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study // Genome Biol.– Springer Science and Business Media LLC, 2020.– Vol. 21, № 1.– P. 114. doi: 10.1186/s13059-020-01996-3. JIF (для WoS) = **10.1**, (1.12/0.15)

6. **Penzar D**., Zinkevich A.O., Vorontsov I.E., Sitnik V.V., Favorov A.V., Makeev V.J., Kulakovskiy I.V. What Do Neighbors Tell About You: The Local Context of Cis-Regulatory Modules Complicates Prediction of Regulatory Variants // Front. Genet.– 2019.– Vol. 10.– P. 1078. doi: 10.3389/fgene.2019.01078. JIF (для WoS) = **2.8,** (0.70/0.40)

[10] В скобках приведен объем публикации в условных печатных листах и вклад автора в условных печатных листах

Список литературы

1. Sasse A., Chikina M., Mostafavi S. Unlocking gene regulation with sequence-to-function models // Nat. Methods.– Springer Science and Business Media LLC, 2024.– Vol. 21, № 8.– P. 1374–1377.
2. Zeiltinger J., Roy S., Ay F., Mathelier A., Medina-Rivera A., Mahony S., Sinha S., Ernst J. Perspective on recent developments and challenges in regulatory and systems genomics // arXiv [q-bio.GN].– 2024.
3. Kathail P., Bajwa A., Ioannidis N.M. Leveraging genomic deep learning models for non-coding variant effect prediction // arXiv [q-bio.GN].– 2024.
4. Sachidanandam R., Weissman D., Schmidt S.C., Kakol J.M., Stein L.D., Marth G., Sherry S., Mullikin J.C., Mortimore B.J., Willey D.L., Hunt S.E., Cole C.G., Coggill P.C., Rice C.M., Ning Z., Rogers J., Bentley D.R., Kwok P.Y., Mardis E.R., Yeh R.T., Schultz B., Cook L., Davenport R., Dante M., Fulton L., Hillier L., Waterston R.H., McPherson J.D., Gilman B., Schaffner S., Van Etten W.J., Reich D., Higgins J., Daly M.J., Blumenstiel B., Baldwin J., Stange-Thomann N., Zody M.C., Linton L., Lander E.S., Altshuler D., International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms // Nature.– Springer Science and Business Media LLC, 2001.– Vol. 409, № 6822.– P. 928–933.
5. Cheng J., Novati G., Pan J., Bycroft C., Žemgulytė A., Applebaum T., Pritzel A., Wong L.H., Zielinski M., Sargeant T., Schneider R.G., Senior A.W., Jumper J., Hassabis D., Kohli P., Avsec Ž. Accurate proteome-wide missense variant effect prediction with AlphaMissense // Science.– 2023.– Vol. 381, № 6664.– P. eadg7492.
6. Abramson J., Adler J., Dunger J., Evans R., Green T., Pritzel A., Ronneberger O., Willmore L., Ballard A.J., Bambrick J., Bodenstein S.W., Evans D.A., Hung C.-C., O'Neill M., Reiman D., Tunyasuvunakool K., Wu Z., Žemgulytė A., Arvaniti E., Beattie C., Bertolli O., Bridgland A., Cherepanov A., Congreve M., Cowen-Rivers A.I., Cowie A., Figurnov M., Fuchs F.B., Gladman H., Jain R., Khan Y.A., Low C.M.R., Perlin K., Potapenko A., Savy P., Singh S., Stecula A., Thillaisundaram A., Tong C., Yakneen S., Zhong E.D., Zielinski M., Žídek A., Bapst V., Kohli P., Jaderberg M., Hassabis D., Jumper J.M. Accurate structure prediction of biomolecular interactions with AlphaFold 3 // Nature.– 2024.– Vol. 630, № 8016.– P. 493–500.
7. Trifonov E.N. Thirty years of multiple sequence codes // Genomics Proteomics Bioinformatics.– Oxford University Press (OUP), 2011.– Vol. 9, № 1-2.– P. 1–6.
8. Buccitelli C., Selbach M. mRNAs, proteins and the emerging principles of gene expression control // Nat. Rev. Genet.– Springer Science and Business Media LLC, 2020.– Vol. 21, № 10.– P. 630–644.
9. Sasse A., Ng B., Spiro A.E., Tasaki S., Bennett D.A., Gaiteri C., De Jager P.L., Chikina M., Mostafavi S. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings // Nat. Genet.– 2023.– Vol. 55, № 12.– P. 2060–2064.
10. Karollus A., Mauermeier T., Gagneur J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers // Genome Biol.– 2023.– Vol. 24, № 1.– P. 56.
11. Huang C., Shuai R.W., Baokar P., Chung R., Rastogi R., Kathail P., Ioannidis N.M. Personal transcriptome variation is poorly explained by current genomic deep learning models // Nat. Genet.– 2023.– Vol. 55, № 12.– P. 2056–2059.
12. Bajwa A., Rastogi R., Kathail P., Shuai R.W., Ioannidis N.M. Characterizing uncertainty in predictions of genomic sequence-to-activity models // bioRxiv.– 2023.
13. Kathail P., Shuai R.W., Chung R., Ye C.J., Loeb G.B., Ioannidis N.M. Current genomic deep learning models display decreased performance in cell type-specific accessible regions // Genome Biol.– 2024.– Vol. 25, № 1.– P. 202.
14. Patwardhan R.P., Lee C., Litvin O., Young D.L., Pe'er D., Shendure J. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis // Nat. Biotechnol.– 2009.– Vol. 27, № 12.– P. 1173–1175.
15. Hiatt J.B., Patwardhan R.P., Turner E.H., Lee C., Shendure J. Parallel, tag-directed assembly of locally derived short sequence reads // Nat. Methods.– 2010.– Vol. 7, № 2.– P. 119–122.
16. White M.A., Myers C.A., Corbo J.C., Cohen B.A. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks // Proc. Natl. Acad. Sci. U. S. A.– 2013.– Vol. 110, № 29.– P. 11952–11957.
17. Inoue F., Ahituv N. Decoding enhancers using massively parallel reporter assays // Genomics.– 2015.–

Vol. 106, № 3.– P. 159–164.
18. van Arensbergen J., Pagie L., FitzPatrick V.D., de Haas M., Baltissen M.P., Comoglio F., van der Weide R.H., Teunissen H., Võsa U., Franke L., de Wit E., Vermeulen M., Bussemaker H.J., van Steensel B. High-throughput identification of human SNPs affecting regulatory element activity // Nat. Genet.– Springer Science and Business Media LLC, 2019.– Vol. 51, № 7.– P. 1160–1169.
19. Romanov S.E., Laktionov P.P. Practical application of massively parallel reporter assay in biotechnology and medicine // Клиническая практика.– ECO-Vector LLC, 2023.– Vol. 13, № 4.– P. 74–87.
20. Siraj L., Castro R.I., Dewey H., Kales S., Nguyen T.T.L., Kanai M., Berenzy D., Mouri K., Wang Q.S., McCaw Z.R., Gosai S.J., Aguet F., Cui R., Vockley C.M., Lareau C.A., Okada Y., Gusev A., Jones T.R., Lander E.S., Sabeti P.C., Finucane H.K., Reilly S.K., Ulirsch J.C., Tewhey R. Functional dissection of complex and molecular trait variants at single nucleotide resolution // bioRxivorg.– 2024.
21. Bagger F.O., Borgwardt L., Jespersen A.S., Hansen A.R., Bertelsen B., Kodama M., Nielsen F.C. Whole genome sequencing in clinical practice // BMC Med. Genomics.– 2024.– Vol. 17, № 1.– P. 39.
22. Hawkes G., Beaumont R.N., Li Z., Mandla R., Li X., Albert C.M., Arnett D.K., Ashley-Koch A.E., Ashrani A.A., Barnes K.C., Boerwinkle E., Brody J.A., Carson A.P., Chami N., Chen Y.-D.I., Chung M.K., Curran J.E., Darbar D., Ellinor P.T., Fornage M., Gordeuk V.R., Guo X., He J., Hwu C.-M., Kalyani R.R., Kaplan R., Kardia S.L.R., Kooperberg C., Loos R.J.F., Lubitz S.A., Minster R.L., Naseri T., Viali S. 'itea, Mitchell B.D., Murabito J.M., Palmer N.D., Psaty B.M., Redline S., Shoemaker M.B., Silverman E.K., Telen M.J., Weiss S.T., Yanek L.R., Zhou H., NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Liu C.-T., North K.E., Justice A.E., Locke J.M., Owens N., Murray A., Patel K., Frayling T.M., Wright C.F., Wood A.R., Lin X., Manning A., Weedon M.N. Whole-genome sequencing in 333,100 individuals reveals rare non-coding single variant and aggregate associations with height // Nat. Commun.– Springer Science and Business Media LLC, 2024.– Vol. 15, № 1.– P. 8549.
23. Albert F.W., Kruglyak L. The role of regulatory variation in complex traits and disease // Nature Reviews Genetics.– Nature Publishing Group, 2015.– Vol. 16, № 4.– P. 197–212.
24. Uffelmann E., Huang Q.Q., Munung N.S., de Vries J., Okada Y., Martin A.R., Martin H.C., Lappalainen T., Posthuma D. Genome-wide association studies // Nat. Rev. Methods Primers.– Springer Science and Business Media LLC, 2021.– Vol. 1, № 1.– P. 1–21.
25. Mostafavi H., Spence J.P., Naqvi S., Pritchard J.K. Systematic differences in discovery of genetic effects on gene expression and complex traits // Nat. Genet.– Springer Science and Business Media LLC, 2023.– Vol. 55, № 11.– P. 1866–1875.
26. Hindorff L.A., Sethupathy P., Junkins H.A., Ramos E.M., Mehta J.P., Collins F.S., Manolio T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits // Proc. Natl. Acad. Sci. U. S. A.– 2009.– Vol. 106, № 23.– P. 9362–9367.
27. Edwards S.L., Beesley J., French J.D., Dunning A.M. Beyond GWASs: illuminating the dark road from association to function // Am. J. Hum. Genet.– 2013.– Vol. 93, № 5.– P. 779–797.
28. Farh K.K.-H., Marson A., Zhu J., Kleinewietfeld M., Housley W.J., Beik S., Shoresh N., Whitton H., Ryan R.J.H., Shishkin A.A., Hatan M., Carrasco-Alfonso M.J., Mayer D., Luckey C.J., Patsopoulos N.A., De Jager P.L., Kuchroo V.K., Epstein C.B., Daly M.J., Hafler D.A., Bernstein B.E. Genetic and epigenetic fine mapping of causal autoimmune disease variants // Nature.– 2015.– Vol. 518, № 7539.– P. 337–343.
29. Khurana E., Fu Y., Chakravarty D., Demichelis F., Rubin M.A., Gerstein M. Role of non-coding sequence variants in cancer // Nat. Rev. Genet.– 2016.– Vol. 17, № 2.– P. 93–108.
30. Rojano E., Seoane P., Ranea J.A.G., Perkins J.R. Regulatory variants: from detection to predicting impact // Brief. Bioinform.– 2019.– Vol. 20, № 5.– P. 1639–1654.
31. Walavalkar K., Notani D. Beyond the coding genome: non-coding mutations and cancer // Front. Biosci. .– 2020.– Vol. 25, № 10.– P. 1828–1838.
32. Linder J., Srivastava D., Yuan H., Agarwal V., Kelley D.R. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation // bioRxiv.– 2023.– P. 2023.08.30.555582.
33. Penzar D.D., Zinkevich A.O., Vorontsov I.E., Sitnik V.V., Favorov A.V., Makeev V.J., Kulakovskiy I.V. What Do Neighbors Tell About You: The Local Context of Cis-Regulatory Modules Complicates Prediction of Regulatory Variants // Front. Genet.– 2019.– Vol. 10.– P. 1078.
34. Abramov S., Boytsov A., Bykova D., Penzar D.D., Yevshin I., Kolmykov S.K., Fridman M.V., Favorov A.V., Vorontsov I.E., Baulin E., Kolpakov F., Makeev V.J., Kulakovskiy I.V. Landscape of allele-specific transcription factor binding in the human genome // Nat. Commun.– 2021.– Vol. 12, №

1.– P. 2751.
35. Penzar D., Nogina D., Noskova E., Zinkevich A., Meshcheryakov G., Lando A., Rafi A.M., de Boer C., Kulakovskiy I.V. LegNet: a best-in-class deep learning model for short DNA regulatory regions // Bioinformatics.– 2023.– Vol. 39, № 8.
36. Kearsey M.J. The principles of QTL analysis (a minimal mathematics approach) // J. Exp. Bot.– Oxford University Press (OUP), 1998.– Vol. 49, № 327.– P. 1619–1623.
37. Astle W.J., Elding H., Jiang T., Allen D., Ruklisa D., Mann A.L., Mead D., Bouman H., Riveros-Mckay F., Kostadima M.A., Lambourne J.J., Sivapalaratnam S., Downes K., Kundu K., Bomba L., Berentsen K., Bradley J.R., Daugherty L.C., Delaneau O., Freson K., Garner S.F., Grassi L., Guerrero J., Haimel M., Janssen-Megens E.M., Kaan A., Kamat M., Kim B., Mandoli A., Marchini J., Martens J.H.A., Meacham S., Megy K., O'Connell J., Petersen R., Sharifi N., Sheard S.M., Staley J.R., Tuna S., van der Ent M., Walter K., Wang S.-Y., Wheeler E., Wilder S.P., Iotchkova V., Moore C., Sambrook J., Stunnenberg H.G., Di Angelantonio E., Kaptoge S., Kuijpers T.W., Carrillo-de-Santa-Pau E., Juan D., Rico D., Valencia A., Chen L., Ge B., Vasquez L., Kwan T., Garrido-Martín D., Watt S., Yang Y., Guigo R., Beck S., Paul D.S., Pastinen T., Bujold D., Bourque G., Frontini M., Danesh J., Roberts D.J., Ouwehand W.H., Butterworth A.S., Soranzo N. The Allelic landscape of human blood cell trait variation and links to common complex disease // Cell.– 2016.– Vol. 167, № 5.– P. 1415–1429.e19.
38. 1000 Genomes Project Consortium, Auton A., Brooks L.D., Durbin R.M., Garrison E.P., Kang H.M., Korbel J.O., Marchini J.L., McCarthy S., McVean G.A., Abecasis G.R. A global reference for human genetic variation // Nature.– 2015.– Vol. 526, № 7571.– P. 68–74.
39. Musunuru K., Strong A., Frank-Kamenetsky M., Lee N.E., Ahfeldt T., Sachs K.V., Li X., Li H., Kuperwasser N., Ruda V.M., Pirruccello J.P., Muchmore B., Prokunina-Olsson L., Hall J.L., Schadt E.E., Morales C.R., Lund-Katz S., Phillips M.C., Wong J., Cantley W., Racie T., Ejebe K.G., Orho-Melander M., Melander O., Kotliansky V., Fitzgerald K., Krauss R.M., Cowan C.A., Kathiresan S., Rader D.J. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus // Nature.– Springer Science and Business Media LLC, 2010.– Vol. 466, № 7307.– P. 714–719.
40. Afanasyeva M.A., Putlyaeva L.V., Demin D.E., Kulakovskiy I.V., Vorontsov I.E., Fridman M.V., Makeev V.J., Kuprash D.V., Schwartz A.M. The single nucleotide variant rs12722489 determines differential estrogen receptor binding and enhancer properties of an IL2RA intronic region // PLoS One.– Public Library of Science (PLoS), 2017.– Vol. 12, № 2.– P. e0172681.
41. Uvarova A.N., Stasevich E.M., Ustiugova A.S., Mitkin N.A., Zheremyan E.A., Sheetikov S.A., Zornikova K.V., Bogolyubova A.V., Rubtsov M.A., Kulakovskiy I.V., Kuprash D.V., Korneev K.V., Schwartz A.M. rs71327024 Associated with COVID-19 Hospitalization Reduces CXCR6 Promoter Activity in Human CD4+ T Cells via Disruption of c-Myb Binding // Int. J. Mol. Sci.– 2023.– Vol. 24, № 18.
42. Choi J., Zhang T., Vu A., Ablain J., Makowski M.M., Colli L.M., Xu M., Hennessey R.C., Yin J., Rothschild H., Gräwe C., Kovacs M.A., Funderburk K.M., Brossard M., Taylor J., Pasaniuc B., Chari R., Chanock S.J., Hoggart C.J., Demenais F., Barrett J.H., Law M.H., Iles M.M., Yu K., Vermeulen M., Zon L.I., Brown K.M. Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma // Nat. Commun.– Springer Science and Business Media LLC, 2020.– Vol. 11, № 1.
43. Weiss C.V., Harshman L., Inoue F., Fraser H.B., Petrov D.A., Ahituv N., Gokhman D. The cis-regulatory effects of modern human-specific variants // Elife.– eLife Sciences Publications, Ltd, 2021.– Vol. 10.
44. Bock C., Datlinger P., Chardon F., Coelho M.A., Dong M.B., Lawson K.A., Lu T., Maroc L., Norman T.M., Song B., Stanley G., Chen S., Garnett M., Li W., Moffat J., Qi L.S., Shapiro R.S., Shendure J., Weissman J.S., Zhuang X. High-content CRISPR screening // Nat. Rev. Methods Primers.– Springer Science and Business Media LLC, 2022.– Vol. 2, № 1.– P. 1–23.
45. Morris J.A., Caragine C., Daniloski Z., Domingo J., Barry T., Lu L., Davis K., Ziosi M., Glinos D.A., Hao S., Mimitou E.P., Smibert P., Roeder K., Katsevich E., Lappalainen T., Sanjana N.E. Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens // Science.– 2023.– Vol. 380, № 6646.– P. eadh7699.
46. Ryu J., Barkal S., Yu T., Jankowiak M., Zhou Y., Francoeur M., Phan Q.V., Li Z., Tognon M., Brown L., Love M.I., Bhat V., Lettre G., Ascher D.B., Cassa C.A., Sherwood R.I., Pinello L. Joint genotypic and phenotypic outcome modeling improves base editing variant effect quantification // Nat. Genet.– Springer Science and Business Media LLC, 2024.– Vol. 56, № 5.– P. 925–937.
47. Stormo G.D., Schneider T.D., Gold L., Ehrenfeucht A. Use of the "Perceptron" algorithm to distinguish

translational initiation sites in E. coli // Nucleic Acids Res.– 1982.– Vol. 10, № 9.– P. 2997–3011.
48. Vorontsov I.E., Kulakovskiy I.V., Khimulya G. PERFECTOS-APE-Predicting Regulatory Functional Effect of SNPs by Approximate P-value Estimation // Bioinformatics.– 2015.
49. Boytsov A., Abramov S., Makeev V.J., Kulakovskiy I.V. Positional weight matrices have sufficient prediction power for analysis of noncoding variants // F1000Res.– 2022.– Vol. 11.– P. 33.
50. Lee D., Gorkin D.U., Baker M., Strober B.J., Asoni A.L., McCallion A.S., Beer M.A. A method to predict the impact of regulatory variants from DNA sequence // Nat. Genet.– 2015.– Vol. 47, № 8.– P. 955–961.
51. Shigaki D., Adato O., Adhikari A.N., Dong S., Hawkins-Hooker A., Inoue F., Juven-Gershon T., Kenlay H., Martin B., Patra A., Penzar D.D., Schubach M., Xiong C., Yan Z., Boyle A.P., Kreimer A., Kulakovskiy I.V., Reid J., Unger R., Yosef N., Shendure J., Ahituv N., Kircher M., Beer M.A. Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay // Hum. Mutat.– 2019.– Vol. 40, № 9.– P. 1280–1291.
52. Alipanahi B., Delong A., Weirauch M.T., Frey B.J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning // Nat. Biotechnol.– 2015.– Vol. 33, № 8.– P. 831–838.
53. Zhou J., Troyanskaya O.G. Predicting effects of noncoding variants with deep learning-based sequence model // Nat. Methods.– 2015.– Vol. 12, № 10.– P. 931–934.
54. Kelley D.R., Snoek J., Rinn J.L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks // Genome Res.– 2016.– Vol. 26, № 7.– P. 990–999.
55. Avsec Ž., Agarwal V., Visentin D., Ledsam J.R., Grabska-Barwinska A., Taylor K.R., Assael Y., Jumper J., Kohli P., Kelley D.R. Effective gene expression prediction from sequence by integrating long-range interactions // Nat. Methods.– 2021.– Vol. 18, № 10.– P. 1196–1203.
56. Kelley D.R., Reshef Y.A., Bileschi M., Belanger D., McLean C.Y., Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks // Genome Res.– 2018.– Vol. 28, № 5.– P. 739–750.
57. Chen K.M., Wong A.K., Troyanskaya O.G., Zhou J. A sequence-based global map of regulatory activity for deciphering human genetics // Nat. Genet.– 2022.– Vol. 54, № 7.– P. 940–949.
58. Martin A.R., Kanai M., Kamatani Y., Okada Y., Neale B.M., Daly M.J. Clinical use of current polygenic risk scores may exacerbate health disparities // Nat. Genet.– Springer Science and Business Media LLC, 2019.– Vol. 51, № 4.– P. 584–591.
59. Drusinsky S., Whalen S., Pollard K.S. Deep-learning prediction of gene expression from personal genomes // bioRxiv.– 2024.– P. 2024.07.27.605449.
60. Rastogi R., Reddy A.J., Chung R., Ioannidis N.M. Fine-tuning sequence-to-expression models on personal genome and transcriptome data // bioRxiv.– 2024.– P. 2024.09.23.614632.
61. Fishman V., Kuratov Y., Petrov M., Shmelev A., Shepelin D., Chekanov N., Kardymon O., Burtsev M. GENA-LM: A Family of Open-Source Foundational DNA Language Models for Long Sequences // bioRxiv.– 2023.– P. 2023.06.12.544594.
62. Dalla-Torre H., Gonzalez L., Mendoza-Revilla J., Carranza N.L., Grzywaczewski A.H., Oteri F., Dallago C., Trop E., de Almeida B.P., Sirelkhatim H., Richard G., Skwark M., Beguir K., Lopez M., Pierrot T. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics // bioRxiv.– 2023.– P. 2023.01.11.523679.
63. Advancing DNA Language Models: The Genomics Long-Range Benchmark [Electronic resource] // InstaDeep.– InstaDeep Ltd, 2024.– URL: https://www.instadeep.com/research/paper/advancing-dna-language-models-the-genomics-long-range-benchmark/ (accessed: 25.09.2024).
64. Tang Z., Koo P.K. Evaluating the representational power of pre-trained DNA language models for regulatory genomics // bioRxiv.– 2024.
65. Schwessinger R., Deasy J., Woodruff R.T., Young S., Branson K.M. Single-cell gene expression prediction from DNA sequence at large contexts // bioRxiv.– 2023.– P. 2023.07.26.550634.
66. Hingerl J.C., Martens L.D., Karollus A., Manz T., Buenrostro J.D., Theis F.J., Gagneur J. scooby: Modeling multi-modal genomic profiles from DNA sequence at single-cell resolution // bioRxivorg.– 2024.– P. 2024.09.19.613754.
67. Lal A., Karollus A., Gunsalus L., Garfield D., Nair S., Tseng A.M., Gordon M.G., Collier J.L., Diamant N., Biancalani T., Corrada Bravo H., Scalia G., Eraslan G. Decoding sequence determinants of gene expression in diverse cellular and disease states // bioRxiv.– 2024.– P. 2024.10.09.617507.
68. de Boer C.G., Vaishnav E.D., Sadeh R., Abeyta E.L., Friedman N., Regev A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters // Nat. Biotechnol.– 2020.– Vol. 38, № 1.– P.

56–65.
69. Atak Z.K., Taskiran I.I., Demeulemeester J., Flerin C., Mauduit D., Minnoye L., Hulselmans G., Christiaens V., Ghanem G.-E., Wouters J., Aerts S. Interpretation of allele-specific chromatin accessibility using cell state-aware deep learning // Genome Res.– 2021.– Vol. 31, № 6.– P. 1082–1096.
70. Vaishnav E.D., de Boer C.G., Molinet J., Yassour M., Fan L., Adiconis X., Thompson D.A., Levin J.Z., Cubillos F.A., Regev A. The evolution, evolvability and engineering of gene regulatory DNA // Nature.– 2022.– Vol. 603, № 7901.– P. 455–463.
71. Sahu B., Hartonen T., Pihlajamaa P., Wei B., Dave K., Zhu F., Kaasinen E., Lidschreiber K., Lidschreiber M., Daub C.O., Cramer P., Kivioja T., Taipale J. Sequence determinants of human gene regulatory elements // Nat. Genet.– Springer Science and Business Media LLC, 2022.– Vol. 54, № 3.– P. 283–294.
72. Agarwal V., Inoue F., Schubach M., Martin B.K., Dash P.M., Zhang Z., Sohota A., Noble W.S., Yardimci G.G., Kircher M., Shendure J., Ahituv N. Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types // bioRxiv.– 2023.
73. Hu J., Shen L., Albanie S., Sun G., Wu E. Squeeze-and-Excitation Networks // arXiv [cs.CV].– 2017.
74. Tan M., Le Q.V. EfficientNetV2: Smaller models and faster training // arXiv [cs.CV].– 2021.
75. Liu Z., Mao H., Wu C.-Y., Feichtenhofer C., Darrell T., Xie S. A ConvNet for the 2020s // arXiv [cs.CV].– 2022.
76. van den Oord A., Vinyals O., Kavukcuoglu K. Neural discrete representation learning // arXiv [cs.LG].– 2017.
77. Ramesh A., Dhariwal P., Nichol A., Chu C., Chen M. Hierarchical text-conditional image generation with CLIP latents // arXiv [cs.CV].– 2022.
78. Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. High-resolution image synthesis with latent diffusion models // 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).– IEEE, 2022.
79. Razzhigaev A., Shakhmatov A., Maltseva A., Arkhipkin V., Pavlov I., Ryabov I., Kuts A., Panchenko A., Kuznetsov A., Dimitrov D. Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion // arXiv [cs.CV].– 2023.
80. Božić M., Horvat M. A survey of deep learning audio generation methods // arXiv [cs.SD].– 2024.
81. Minaee S., Mikolov T., Nikzad N., Chenaghlu M., Socher R., Amatriain X., Gao J. Large Language Models: A survey // arXiv [cs.CL].– 2024.
82. Kadurin A., Nikolenko S., Khrabrov K., Aliper A., Zhavoronkov A. DruGAN: An advanced generative adversarial autoencoder model for de Novo generation of new molecules with desired molecular properties in silico // Mol. Pharm.– 2017.– Vol. 14, № 9.– P. 3098–3104.
83. Polykovskiy D., Zhebrak A., Vetrov D., Ivanenkov Y., Aladinskiy V., Mamoshina P., Bozdaganyan M., Aliper A., Zhavoronkov A., Kadurin A. Entangled conditional adversarial autoencoder for de Novo drug discovery // Mol. Pharm.– American Chemical Society (ACS), 2018.– Vol. 15, № 10.– P. 4398–4405.
84. Zhavoronkov A., Ivanenkov Y.A., Aliper A., Veselov M.S., Aladinskiy V.A., Aladinskaya A.V., Terentiev V.A., Polykovskiy D.A., Kuznetsov M.D., Asadulaev A., Volkov Y., Zholus A., Shayakhmetov R.R., Zhebrak A., Minaeva L.I., Zagribelnyy B.A., Lee L.H., Soll R., Madge D., Xing L., Guo T., Aspuru-Guzik A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors // Nat. Biotechnol.– Nature Publishing Group, 2019.– Vol. 37, № 9.– P. 1038–1040.
85. Sinai S., Wang R., Whatley A., Slocum S., Locane E., Kelsic E.D. AdaLead: A simple and robust adaptive greedy search algorithm for sequence design // arXiv [cs.LG].– 2020.
86. Wang Y., Wang H., Wei L., Li S., Liu L., Wang X. Synthetic promoter design in Escherichia coli based on a deep generative network // Nucleic Acids Res.– Oxford University Press (OUP), 2020.– Vol. 48, № 12.– P. 6403–6412.
87. Corso G., Stärk H., Jing B., Barzilay R., Jaakkola T. DiffDock: Diffusion steps, twists, and turns for molecular docking // arXiv [q-bio.BM].– 2022.
88. Zrimec J., Fu X., Muhammad A.S., Skrekas C., Jauniskis V., Speicher N.K., Börlin C.S., Verendel V., Chehreghani M.H., Dubhashi D., Siewers V., David F., Nielsen J., Zelezniak A. Controlling gene expression with deep generative design of regulatory DNA // Nat. Commun.– Springer Science and Business Media LLC, 2022.– Vol. 13, № 1.– P. 5099.
89. Ingraham J.B., Baranov M., Costello Z., Barber K.W., Wang W., Ismail A., Frappier V., Lord D.M., Ng-Thow-Hing C., Van Vlack E.R., Tie S., Xue V., Cowles S.C., Leung A., Rodrigues J.V., Morales-Perez C.L., Ayoub A.M., Green R., Puentes K., Oplinger F., Panwar N.V., Obermeyer F., Root

A.R., Beam A.L., Poelwijk F.J., Grigoryan G. Illuminating protein space with a programmable generative model // Nature.– Springer Science and Business Media LLC, 2023.– Vol. 623, № 7989.– P. 1070–1078.
90. Barazandeh S., Ozden F., Hincer A., Seker U.O.S., Cicek A.E. UTRGAN: Learning to generate 5' UTR sequences for optimized translation efficiency and gene expression // bioRxiv.– 2023.
91. Li T., Xu H., Teng S., Suo M., Bahitwa R., Xu M., Qian Y., Ramstein G.P., Song B., Buckler E.S., Wang H. Modeling 0.6 million genes for the rational design of functional cis-regulatory variants and de novo design of cis-regulatory sequences // Proc. Natl. Acad. Sci. U. S. A.– Proceedings of the National Academy of Sciences, 2024.– Vol. 121, № 26.– P. e2319811121.
92. Morrow A.K., Thornal A., Flynn E.D., Hoelzli E., Shan M., Garipler G., Kirchner R., Reddy A.J., Tabchouri S., Gupta A., Michel J.-B., Laserson U. ML-driven design of 3' UTRs for mRNA stability: biorxiv;2024.10.07.616676v1 // Synthetic Biology.– bioRxiv, 2024.
93. Taskiran I.I., Spanier K.I., Dickmänken H., Kempynck N., Pančíková A., Ekşi E.C., Hulselmans G., Ismail J.N., Theunis K., Vandepoel R., Christiaens V., Mauduit D., Aerts S. Cell-type-directed design of synthetic enhancers // Nature.– 2024.– Vol. 626, № 7997.– P. 212–220.
94. Frank C.J., Schiwietz D., Fuss L., Ovchinnikov S., Dietz H. Alphafold2 refinement improves designability of large de novo proteins // bioRxiv.– 2024.– P. 2024.11.21.624687.
95. Lal A., Garfield D., Biancalani T., Eraslan G. regLM: Designing realistic regulatory DNA with autoregressive language models // bioRxiv.– 2024.
96. Ivanenkov Y., Zagribelnyy B., Malyshev A., Evteev S., Terentiev V., Kamya P., Bezrukov D., Aliper A., Ren F., Zhavoronkov A. The hitchhiker's guide to deep learning driven generative chemistry // ACS Med. Chem. Lett.– 2023.– Vol. 14, № 7.– P. 901–915.
97. Bansal A., Borgnia E., Chu H.-M., Li J.S., Kazemi H., Huang F., Goldblum M., Geiping J., Goldstein T. Cold diffusion: Inverting arbitrary image transforms without noise // arXiv [cs.CV].– 2022.
98. Rafi A.M., Nogina D., Penzar D., Lee D., Lee D., Kim N., Kim S., Kim D., Shin Y., Kwak I.-Y., Meshcheryakov G., Lando A., Zinkevich A., Kim B.-C., Lee J., Kang T., Vaishnav E.D., Yadollahpour P., Random Promoter DREAM Challenge Consortium, Kim S., Albrecht J., Regev A., Gong W., Kulakovskiy I.V., Meyer P., de Boer C.G. A community effort to optimize sequence-based deep learning models of gene regulation // Nat. Biotechnol.– 2024.
99. Li S., Hannenhalli S., Ovcharenko I. De novo human brain enhancers created by single-nucleotide mutations // Sci. Adv.– 2023.– Vol. 9, № 7.– P. eadd2911.
100. Elliott K., Larsson E. Non-coding driver mutations in human cancer // Nat. Rev. Cancer.– 2021.– Vol. 21, № 8.– P. 500–509.
101. Landrum M.J., Lee J.M., Riley G.R., Jang W., Rubinstein W.S., Church D.M., Maglott D.R. ClinVar: public archive of relationships among sequence variation and human phenotype // Nucleic Acids Res.– Oxford University Press (OUP), 2014.– Vol. 42, № Database issue.– P. D980–D985.
102. Lou H., Yeager M., Li H., Bosquet J.G., Hayes R.B., Orr N., Yu K., Hutchinson A., Jacobs K.B., Kraft P., Wacholder S., Chatterjee N., Feigelson H.S., Thun M.J., Diver W.R., Albanes D., Virtamo J., Weinstein S., Ma J., Gaziano J.M., Stampfer M., Schumacher F.R., Giovannucci E., Cancel-Tassin G., Cussenot O., Valeri A., Andriole G.L., Crawford E.D., Anderson S.K., Tucker M., Hoover R.N., Fraumeni J.F. Jr, Thomas G., Hunter D.J., Dean M., Chanock S.J. Fine mapping and functional analysis of a common variant in MSMB on chromosome 10q11.2 associated with prostate cancer susceptibility // Proc. Natl. Acad. Sci. U. S. A.– Proceedings of the National Academy of Sciences, 2009.– Vol. 106, № 19.– P. 7933–7938.
103. Wang Y., Ma R., Liu B., Kong J., Lin H., Yu X., Wang R., Li L., Gao M., Zhou B., Mohan M., Yu H., Hou Z., Shen H., Qian B. SNP rs17079281 decreases lung cancer risk through creating an YY1-binding site to suppress DCBLD1 expression // Oncogene.– 2020.– Vol. 39, № 20.– P. 4092–4102.
104. Schwartz A.M., Demin D.E., Vorontsov I.E., Kasyanov A.S., Putlyaeva L.V., Tatosyan K.A., Kulakovskiy I.V., Kuprash D.V. Multiple single nucleotide polymorphisms in the first intron of the IL2RA gene affect transcription factor binding and enhancer activity // Gene.– Elsevier BV, 2017.– Vol. 602.– P. 50–56.
105. Minnoye L., Marinov G.K., Krausgruber T., Pan L., Marand A.P., Secchia S., Greenleaf W.J., Furlong E.E.M., Zhao K., Schmitz R.J., Bock C., Aerts S. Chromatin accessibility profiling methods // Nat. Rev. Methods Primers.– Springer Science and Business Media LLC, 2021.– Vol. 1, № 1.
106. Baars M.J.D., Douma T., Simeonov D.R., Myers D.R., Kulhanek K., Banerjee S., Zwakenberg S., Baltissen M.P., Amini M., de Roock S., van Wijk F., Vermeulen M., Marson A., Roose J.P., Vercoulen Y. Dysregulated RASGRP1 expression through RUNX1 mediated transcription promotes autoimmunity

// Eur. J. Immunol.– Wiley, 2021.– Vol. 51, № 2.– P. 471–482.
107. Soldner F., Stelzer Y., Shivalila C.S., Abraham B.J., Latourelle J.C., Barrasa M.I., Goldmann J., Myers R.H., Young R.A., Jaenisch R. Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression // Nature.– 2016.– Vol. 533, № 7601.– P. 95–99.
108. Long H.K., Osterwalder M., Welsh I.C., Hansen K., Davies J.O.J., Liu Y.E., Koska M., Adams A.T., Aho R., Arora N., Ikeda K., Williams R.M., Sauka-Spengler T., Porteus M.H., Mohun T., Dickel D.E., Swigut T., Hughes J.R., Higgs D.R., Visel A., Selleri L., Wysocka J. Loss of extreme long-range enhancers in human neural crest drives a craniofacial disorder // Cell Stem Cell.– Elsevier BV, 2020.– Vol. 27, № 5.– P. 765–783.e14.
109. Kimura M. Neutral theory of molecular evolution.– Cambridge University Press, 1985.
110. Cahoon J.L., Rui X., Tang E., Simons C., Langie J., Chen M., Lo Y.-C., Chiang C.W.K. Imputation accuracy across global human populations // Am. J. Hum. Genet.– Elsevier BV, 2024.– Vol. 111, № 5.– P. 979–989.
111. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future // Nat. Rev. Genet.– Springer Science and Business Media LLC, 2008.– Vol. 9, № 6.– P. 477–485.
112. Wang G., Sarkar A., Carbonetto P., Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping // J. R. Stat. Soc. Series B Stat. Methodol.– Oxford University Press (OUP), 2020.– Vol. 82, № 5.– P. 1273–1300.
113. Wellcome Trust Case Control Consortium, Maller J.B., McVean G., Byrnes J., Vukcevic D., Palin K., Su Z., Howson J.M.M., Auton A., Myers S., Morris A., Pirinen M., Brown M.A., Burton P.R., Caulfield M.J., Compston A., Farrall M., Hall A.S., Hattersley A.T., Hill A.V.S., Mathew C.G., Pembrey M., Satsangi J., Stratton M.R., Worthington J., Craddock N., Hurles M., Ouwehand W., Parkes M., Rahman N., Duncanson A., Todd J.A., Kwiatkowski D.P., Samani N.J., Gough S.C.L., McCarthy M.I., Deloukas P., Donnelly P. Bayesian refinement of association signals for 14 loci in 3 common diseases // Nat. Genet.– Nature Publishing Group, 2012.– Vol. 44, № 12.– P. 1294–1301.
114. Kanai M., Elzur R., Zhou W., Global Biobank Meta-analysis Initiative, Daly M.J., Finucane H.K. Meta-analysis fine-mapping is often miscalibrated at single-variant resolution // Cell Genom.– Elsevier BV, 2022.– Vol. 2, № 12.– P. 100210.
115. Kircher M., Xiong C., Martin B., Schubach M., Inoue F., Bell R.J.A., Costello J.F., Shendure J., Ahituv N. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution // Nat. Commun.– 2019.– Vol. 10, № 1.– P. 3583.
116. Ponomarenko M., Sharypova E., Drachkova I., Chadaeva I., Arkova O., Podkolodnaya O., Ponomarenko P., Kolchanov N., Savinkova L. Unannotated single nucleotide polymorphisms in the TATA box of erythropoiesis genes show in vitro positive involvements in cognitive and mental disorders // BMC Med. Genet.– Springer Science and Business Media LLC, 2020.– Vol. 21, № Suppl 1.– P. 165.
117. Shihab H.A., Rogers M.F., Gough J., Mort M., Cooper D.N., Day I.N.M., Gaunt T.R., Campbell C. An integrative approach to predicting the functional effects of non-coding and coding sequence variation // Bioinformatics.– 2015.– Vol. 31, № 10.– P. 1536–1543.
118. Schubach M., Maass T., Nazaretyan L., Röner S., Kircher M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions // Nucleic Acids Res.– 2024.– Vol. 52, № D1.– P. D1143–D1154.
119. Pepke S., Wold B., Mortazavi A. Computation for ChIP-seq and RNA-seq studies // Nat. Methods.– 2009.– Vol. 6, № 11 Suppl.– P. S22–S32.
120. Meyer C.A., Liu X.S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology // Nat. Rev. Genet.– 2014.– Vol. 15, № 11.– P. 709–721.
121. Klemm S.L., Shipony Z., Greenleaf W.J. Chromatin accessibility and the regulatory epigenome // Nat. Rev. Genet.– 2019.– Vol. 20, № 4.– P. 207–220.
122. Thurman R.E., Rynes E., Humbert R., Vierstra J., Maurano M.T., Haugen E., Sheffield N.C., Stergachis A.B., Wang H., Vernot B., Garg K., John S., Sandstrom R., Bates D., Boatman L., Canfield T.K., Diegel M., Dunn D., Ebersol A.K., Frum T., Giste E., Johnson A.K., Johnson E.M., Kutyavin T., Lajoie B., Lee B.-K., Lee K., London D., Lotakis D., Neph S., Neri F., Nguyen E.D., Qu H., Reynolds A.P., Roach V., Safi A., Sanchez M.E., Sanyal A., Shafer A., Simon J.M., Song L., Vong S., Weaver M., Yan Y., Zhang Z., Zhang Z., Lenhard B., Tewari M., Dorschner M.O., Hansen R.S., Navas P.A., Stamatoyannopoulos G., Iyer V.R., Lieb J.D., Sunyaev S.R., Akey J.M., Sabo P.J., Kaul R., Furey T.S., Dekker J., Crawford G.E., Stamatoyannopoulos J.A. The accessible chromatin landscape of the human genome // Nature.– 2012.– Vol. 489, № 7414.– P. 75–82.
123. Toneyan S., Tang Z., Koo P.K. Evaluating deep learning for predicting epigenomic profiles // Nat Mach

Intell.– 2022.– Vol. 4, № 12.– P. 1088–1100.
124. Laurette P., Strub T., Koludrovic D., Keime C., Le Gras S., Seberg H., Van Otterloo E., Imrichova H., Siddaway R., Aerts S., Cornell R.A., Mengus G., Davidson I. Transcription factor MITF and remodeller BRG1 define chromatin organisation at regulatory elements in melanoma cells // Elife.– 2015.– Vol. 4.
125. Ishii S., Kakizuka T., Park S.-J., Tagawa A., Sanbo C., Tanabe H., Ohkawa Y., Nakanishi M., Nakai K., Miyanari Y. Genome-wide ATAC-see screening identifies TFDP1 as a modulator of global chromatin accessibility // Nat. Genet.– Springer Science and Business Media LLC, 2024.– Vol. 56, № 3.– P. 473–482.
126. Razavi R., Fathi A., Yellan I., Brechalov A., Laverty K.U., Jolma A., Hernandez-Corchado A., Zheng H., Yang A.W.H., Albu M., Barazandeh M., Hu C., Vorontsov I.E., Patel Z.M., Codebook Consortium, Kulakovskiy I.V., Bucher P., Morris Q., Najafabadi H.S., Hughes T.R. Extensive binding of uncharacterized human transcription factors to genomic dark matter // bioRxivorg.– 2024.
127. Grandi F.C., Modi H., Kampman L., Corces M.R. Chromatin accessibility profiling by ATAC-seq // Nat. Protoc.– Springer Science and Business Media LLC, 2022.– Vol. 17, № 6.– P. 1518–1552.
128. Deplancke B., Alpern D., Gardeux V. The Genetics of Transcription Factor DNA Binding Variation // Cell.– 2016.– Vol. 166, № 3.– P. 538–554.
129. Killela P.J., Reitman Z.J., Jiao Y., Bettegowda C., Agrawal N., Diaz L.A. Jr, Friedman A.H., Friedman H., Gallia G.L., Giovanella B.C., Grollman A.P., He T.-C., He Y., Hruban R.H., Jallo G.I., Mandahl N., Meeker A.K., Mertens F., Netto G.J., Rasheed B.A., Riggins G.J., Rosenquist T.A., Schiffman M., Shih I.-M., Theodorescu D., Torbenson M.S., Velculescu V.E., Wang T.-L., Wentzensen N., Wood L.D., Zhang M., McLendon R.E., Bigner D.D., Kinzler K.W., Vogelstein B., Papadopoulos N., Yan H. TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal // Proc. Natl. Acad. Sci. U. S. A.– 2013.– Vol. 110, № 15.– P. 6021–6026.
130. Mirkov M.U., Verstockt B., Cleynen I. Genetics of inflammatory bowel disease: beyond NOD2 // Lancet Gastroenterol Hepatol.– 2017.– Vol. 2, № 3.– P. 224–234.
131. Fabbri C., Serretti A. Role of 108 schizophrenia-associated loci in modulating psychopathological dimensions in schizophrenia and bipolar disorder // Am. J. Med. Genet. B Neuropsychiatr. Genet.– 2017.– Vol. 174, № 7.– P. 757–764.
132. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes // Nature.– 2020.– Vol. 578, № 7793.– P. 82–93.
133. Furey T.S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions // Nat. Rev. Genet.– 2012.– Vol. 13, № 12.– P. 840–852.
134. Dong X., Greven M.C., Kundaje A., Djebali S., Brown J.B., Cheng C., Gingeras T.R., Gerstein M., Guigó R., Birney E., Weng Z. Modeling gene expression using chromatin features in various cellular contexts // Genome Biol.– 2012.– Vol. 13, № 9.– P. R53.
135. Kukurba K.R., Montgomery S.B. RNA Sequencing and Analysis // Cold Spring Harb. Protoc.– 2015.– Vol. 2015, № 11.– P. 951–969.
136. Shiraki T., Kondo S., Katayama S., Waki K., Kasukawa T., Kawaji H., Kodzius R., Watahiki A., Nakamura M., Arakawa T., Fukuda S., Sasaki D., Podhajska A., Harbers M., Kawai J., Carninci P., Hayashizaki Y. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage // Proc. Natl. Acad. Sci. U. S. A.– Proceedings of the National Academy of Sciences, 2003.– Vol. 100, № 26.– P. 15776–15781.
137. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest A.R.R., Kawaji H., Rehli M., Baillie J.K., de Hoon M.J.L., Haberle V., Lassmann T., Kulakovskiy I.V., Lizio M., Itoh M., Andersson R., Mungall C.J., Meehan T.F., Schmeier S., Bertin N., Jørgensen M., Dimont E., Arner E., Schmidl C., Schaefer U., Medvedeva Y.A., Plessy C., Vitezic M., Severin J., Semple C.A., Ishizu Y., Young R.S., Francescatto M., Alam I., Albanese D., Altschuler G.M., Arakawa T., Archer J.A.C., Arner P., Babina M., Rennie S., Balwierz P.J., Beckhouse A.G., Pradhan-Bhatt S., Blake J.A., Blumenthal A., Bodega B., Bonetti A., Briggs J., Brombacher F., Burroughs A.M., Califano A., Cannistraci C.V., Carbajo D., Chen Y., Chierici M., Ciani Y., Clevers H.C., Dalla E., Davis C.A., Detmar M., Diehl A.D., Dohi T., Drabløs F., Edge A.S.B., Edinger M., Ekwall K., Endoh M., Enomoto H., Fagiolini M., Fairbairn L., Fang H., Farach-Carson M.C., Faulkner G.J., Favorov A.V., Fisher M.E., Frith M.C., Fujita R., Fukuda S., Furlanello C., Furino M., Furusawa J.-I., Geijtenbeek T.B., Gibson A.P., Gingeras T., Goldowitz D., Gough J., Guhl S., Guler R., Gustincich S., Ha T.J., Hamaguchi M., Hara M., Harbers M., Harshbarger J., Hasegawa A., Hasegawa Y., Hashimoto T., Herlyn M., Hitchens K.J., Ho Sui S.J., Hofmann O.M., Hoof I., Hori F., Huminiecki L., Iida K., Ikawa T., Jankovic B.R., Jia H., Joshi A., Jurman G., Kaczkowski B., Kai C., Kaida K., Kaiho A., Kajiyama K., Kanamori-Katayama M., Kasianov A.S.,

Kasukawa T., Katayama S., Kato S., Kawaguchi S., Kawamoto H., Kawamura Y.I., Kawashima T., Kempfle J.S., Kenna T.J., Kere J., Khachigian L.M., Kitamura T., Klinken S.P., Knox A.J., Kojima M., Kojima S., Kondo N., Koseki H., Koyasu S., Krampitz S., Kubosaki A., Kwon A.T., Laros J.F.J., Lee W., Lennartsson A., Li K., Lilje B., Lipovich L., Mackay-Sim A., Manabe R.-I., Mar J.C., Marchand B., Mathelier A., Mejhert N., Meynert A., Mizuno Y., de Lima Morais D.A., Morikawa H., Morimoto M., Moro K., Motakis E., Motohashi H., Mummery C.L., Murata M., Nagao-Sato S., Nakachi Y., Nakahara F., Nakamura T., Nakamura Y., Nakazato K., van Nimwegen E., Ninomiya N., Nishiyori H., Noma S., Noma S., Noazaki T., Ogishima S., Ohkura N., Ohimiya H., Ohno H., Ohshima M., Okada-Hatakeyama M., Okazaki Y., Orlando V., Ovchinnikov D.A., Pain A., Passier R., Patrikakis M., Persson H., Piazza S., Prendergast J.G.D., Rackham O.J.L., Ramilowski J.A., Rashid M., Ravasi T., Rizzu P., Roncador M., Roy S., Rye M.B., Saijyo E., Sajantila A., Saka A., Sakaguchi S., Sakai M., Sato H., Savvi S., Saxena A., Schneider C., Schultes E.A., Schulze-Tanzil G.G., Schwegmann A., Sengstag T., Sheng G., Shimoji H., Shimoni Y., Shin J.W., Simon C., Sugiyama D., Sugiyama T., Suzuki M., Suzuki N., Swoboda R.K., 't Hoen P.A.C., Tagami M., Takahashi N., Takai J., Tanaka H., Tatsukawa H., Tatum Z., Thompson M., Toyodo H., Toyoda T., Valen E., van de Wetering M., van den Berg L.M., Verado R., Vijayan D., Vorontsov I.E., Wasserman W.W., Watanabe S., Wells C.A., Winteringham L.N., Wolvetang E., Wood E.J., Yamaguchi Y., Yamamoto M., Yoneda M., Yonekura Y., Yoshida S., Zabierowski S.E., Zhang P.G., Zhao X., Zucchelli S., Summers K.M., Suzuki H., Daub C.O., Kawai J., Heutink P., Hide W., Freeman T.C., Lenhard B., Bajic V.B., Taylor M.S., Makeev V.J., Sandelin A., Hume D.A., Carninci P., Hayashizaki Y. A promoter-level mammalian expression atlas // Nature.– 2014.– Vol. 507, № 7493.– P. 462–470.

138. Dudnyk K., Cai D., Shi C., Xu J., Zhou J. Sequence basis of transcription initiation in the human genome // Science.– 2024.– Vol. 384, № 6694.– P. eadj0116.
139. Buyan A., Meshcheryakov G., Safronov V., Abramov S., Boytsov A., Nozdrin V., Baulin E.F., Kolmykov S., Vierstra J., Kolpakov F., Makeev V.J., Kulakovskiy I.V. Statistical framework for calling allelic imbalance in high-throughput sequencing data // bioRxiv.– 2023.– P. 2023.11.07.565968.
140. Hayashi T., Ozaki H., Sasagawa Y., Umeda M., Danno H., Nikaido I. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs // Nat. Commun.– Nature Publishing Group, 2018.– Vol. 9, № 1.– P. 619.
141. Cao J., Cusanovich D.A., Ramani V., Aghamirzaie D., Pliner H.A., Hill A.J., Daza R.M., McFaline-Figueroa J.L., Packer J.S., Christiansen L., Steemers F.J., Adey A.C., Trapnell C., Shendure J. Joint profiling of chromatin accessibility and gene expression in thousands of single cells // Science.– American Association for the Advancement of Science (AAAS), 2018.– Vol. 361, № 6409.– P. 1380–1385.
142. Bravo González-Blas C., Matetovici I., Hillen H., Taskiran I.I., Vandepoel R., Christiaens V., Sansores-García L., Verboven E., Hulselmans G., Poovathingal S., Demeulemeester J., Psatha N., Mauduit D., Halder G., Aerts S. Single-cell spatial multi-omics and deep learning dissect enhancer-driven gene regulatory networks in liver zonation // Nat. Cell Biol.– Nature Publishing Group, 2024.– Vol. 26, № 1.– P. 153–167.
143. Orenstein Y., Shamir R. A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data // Nucleic Acids Res.– Oxford University Press (OUP), 2014.– Vol. 42, № 8.– P. e63.
144. Lambert S.A., Jolma A., Campitelli L.F., Das P.K., Yin Y., Albu M., Chen X., Taipale J., Hughes T.R., Weirauch M.T. The Human Transcription Factors // Cell.– 2018.– Vol. 175, № 2.– P. 598–599.
145. Johnson D.S., Mortazavi A., Myers R.M., Wold B. Genome-wide mapping of in vivo protein-DNA interactions // Science.– American Association for the Advancement of Science (AAAS), 2007.– Vol. 316, № 5830.– P. 1497–1502.
146. Hallikas O., Taipale J. High-throughput assay for determining specificity and affinity of protein-DNA binding interactions // Nat. Protoc.– Springer Science and Business Media LLC, 2006.– Vol. 1, № 1.– P. 215–222.
147. Berger M.F., Bulyk M.L. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors // Nat. Protoc.– Springer Science and Business Media LLC, 2009.– Vol. 4, № 3.– P. 393–411.
148. Berg O.G., von Hippel P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters // J. Mol. Biol.– 1987.– Vol. 193, № 4.– P. 723–750.
149. Wasserman W.W., Sandelin A. Applied bioinformatics for the identification of regulatory elements //

Nat. Rev. Genet.– 2004.– Vol. 5, № 4.– P. 276–287.
150. Duttke S.H., Guzman C., Chang M., Delos Santos N.P., McDonald B.R., Xie J., Carlin A.F., Heinz S., Benner C. Position-dependent function of human sequence-specific transcription factors // Nature.– 2024.
151. Farley E.K., Olson K.M., Zhang W., Rokhsar D.S., Levine M.S. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers // Proc. Natl. Acad. Sci. U. S. A.– Proceedings of the National Academy of Sciences, 2016.– Vol. 113, № 23.– P. 6508–6513.
152. Cornejo-Páramo P., Zhang X., Louis L., Yang Y.-H., Li Z., Humphreys D., Wong E.S. A Bag-Of-Motif Model Captures Cell States at Distal Regulatory Sequences // bioRxiv.– 2024.– P. 2024.01.03.574012.
153. Vorontsov I.E., Kozin I., Abramov S., Boytsov A., Jolma A., Albu M., Ambrosini G., Faltejskova K., Gralak A.J., Gryzunov N., Inukai S., Kolmykov S., Kravchenko P., Kribelbauer-Swietek J.F., Laverty K.U., Nozdrin V., Patel Z.M., Penzar D., Plescher M.-L., Pour S.E., Razavi R., Yang A.W.H., Yevshin I., Zinkevich A., Weirauch M.T., Bucher P., Deplancke B., Fornes O., Grau J., Grosse I., Kolpakov F.A., Makeev V.J., Hughes T.R., Kulakovskiy I.V. Cross-platform DNA motif discovery and benchmarking to explore binding specificities of poorly studied human transcription factors: biorxiv;2024.11.11.619379v2 // Bioinformatics.– bioRxiv, 2024.
154. Weirauch M.T., Cote A., Norel R., Annala M., Zhao Y., Riley T.R., Saez-Rodriguez J., Cokelaer T., Vedenko A., Talukder S., DREAM5 Consortium, Bussemaker H.J., Morris Q.D., Bulyk M.L., Stolovitzky G., Hughes T.R. Evaluation of methods for modeling transcription factor sequence specificity // Nat. Biotechnol.– 2013.– Vol. 31, № 2.– P. 126–134.
155. Yan J., Qiu Y., Ribeiro Dos Santos A.M., Yin Y., Li Y.E., Vinckier N., Nariai N., Benaglio P., Raman A., Li X., Fan S., Chiou J., Chen F., Frazer K.A., Gaulton K.J., Sander M., Taipale J., Ren B. Systematic analysis of binding of transcription factors to noncoding variants // Nature.– Springer Science and Business Media LLC, 2021.– Vol. 591, № 7848.– P. 147–151.
156. Apicella A., Isgrò F., Prevete R. Don’t Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning // arXiv [cs.LG].– 2024.
157. Kapoor S., Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science // Patterns (N. Y.).– Elsevier BV, 2023.– Vol. 4, № 9.– P. 100804.
158. Walsh I., Pollastri G., Tosatto S.C.E. Correct machine learning on protein sequences: a peer-reviewing perspective // Brief. Bioinform.– 2016.– Vol. 17, № 5.– P. 831–840.
159. Littmann M., Heinzinger M., Dallago C., Olenyi T., Rost B. Embeddings from deep learning transfer GO annotations beyond homology // Sci. Rep.– 2021.– Vol. 11, № 1.– P. 1160.
160. Ding F., Steinhardt J. Protein language models are biased by unequal sequence sampling across the tree of life // bioRxiv.– 2024.– P. 2024.03.07.584001.
161. Gomes J., Ramsundar B., Feinberg E.N., Pande V.S. Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity // arXiv [cs.LG].– 2017.
162. Corso G., Deng A., Fry B., Polizzi N., Barzilay R., Jaakkola T. Deep Confident Steps to New Pockets: Strategies for Docking Generalization // ArXiv.– 2024.
163. Gabriel R.A., Kuo T.-T., McAuley J., Hsu C.-N. Identifying and characterizing highly similar notes in big clinical note datasets // J. Biomed. Inform.– 2018.– Vol. 82.– P. 63–69.
164. Roberts M., Driggs D., Thorpe M., Gilbey J., Yeung M., Ursprung S., Aviles-Rivero A.I., Etmann C., McCague C., Beer L., Weir-McCall J.R., Teng Z., Gkrania-Klotsas E., Rudd J.H.F., Sala E., Schönlieb C.-B. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans // Nature Machine Intelligence.– Nature Publishing Group, 2021.– Vol. 3, № 3.– P. 199–217.
165. Maguolo G., Nanni L. A critic evaluation of methods for COVID-19 automatic detection from X-ray images // Inf. Fusion.– 2021.– Vol. 76.– P. 1–7.
166. Isensee F., Wald T., Ulrich C., Baumgartner M., Roy S., Maier-Hein K., Jaeger P.F. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation // arXiv [cs.CV].– 2024.
167. Khan A., Riudavets Puig R., Boddie P., Mathelier A. BiasAway: command-line and web server to generate nucleotide composition-matched DNA background sequences // Bioinformatics.– 2021.– Vol. 37, № 11.– P. 1607–1609.
168. Schreiber J., Singh R., Bilmes J., Noble W.S. A pitfall for machine learning methods aiming to predict across cell types: biorxiv;512434v2 // Bioinformatics.– bioRxiv, 2019.– P. 473.
169. Whalen S., Truty R.M., Pollard K.S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin // Nat. Genet.– 2016.– Vol. 48, № 5.– P. 488–496.
170. Xi W., Beer M.A. Local epigenomic state cannot discriminate interacting and non-interacting

enhancer-promoter pairs with high accuracy // PLoS Comput. Biol.– 2018.– Vol. 14, № 12.– P. e1006625.
171. Whalen S., Schreiber J., Noble W.S., Pollard K.S. Navigating the pitfalls of applying machine learning in genomics // Nat. Rev. Genet.– 2022.– Vol. 23, № 3.– P. 169–181.
172. de Boer C.G., Taipale J. Hold out the genome: a roadmap to solving the cis-regulatory code // Nature.– 2024.– Vol. 625, № 7993.– P. 41–50.
173. Ghandi M., Lee D., Mohammad-Noori M., Beer M.A. Enhanced regulatory sequence prediction using gapped k-mer features // PLoS Comput. Biol.– 2014.– Vol. 10, № 7.– P. e1003711.
174. Ghandi M., Mohammad-Noori M., Ghareghani N., Lee D., Garraway L., Beer M.A. gkmSVM: an R package for gapped-kmer SVM // Bioinformatics.– 2016.– Vol. 32, № 14.– P. 2205–2207.
175. Lee D., Karchin R., Beer M.A. Discriminative prediction of mammalian enhancers from DNA sequence // Genome Res.– 2011.– Vol. 21, № 12.– P. 2167–2180.
176. Lee D. LS-GKM: a new gkm-SVM for large-scale datasets // Bioinformatics.– 2016.– Vol. 32, № 14.– P. 2196–2198.
177. VandenBosch L.S., Luu K., Timms A.E., Challam S., Wu Y., Lee A.Y., Cherry T.J. Machine Learning Prediction of Non-Coding Variant Impact in Human Retinal cis-Regulatory Elements // Transl. Vis. Sci. Technol.– 2022.– Vol. 11, № 4.– P. 16.
178. Howard W.R. Pattern Recognition and Machine Learning20072Christopher M. Bishop. Pattern Recognition and Machine Learning. Heidelberg, Germany: Springer 2006. i‐xx, 740 pp., ISBN: 0‐387‐31073‐8 $74.95 Hardcover // Kybernetes.– 2007.– Vol. 36, № 2.– P. 275–275.
179. Geurts P., Irrthum A., Wehenkel L. Supervised learning with decision tree-based methods in computational and systems biology // Mol. Biosyst.– 2009.– Vol. 5, № 12.– P. 1593–1605.
180. Touw W.G., Bayjanov J.R., Overmars L., Backus L., Boekhorst J., Wels M., van Hijum S.A.F.T. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? // Brief. Bioinform.– 2013.– Vol. 14, № 3.– P. 315–326.
181. Fawagreh K., Gaber M.M., Elyan E. Random forests: from early developments to recent advancements // Systems Science & Control Engineering.– Taylor & Francis, 2014.– Vol. 2, № 1.– P. 602–609.
182. Boulesteix A.-L., Janitza S., Kruppa J., König I.R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics: Random forests in bioinformatics // WIREs Data Mining Knowl Discov.– 2012.– Vol. 2, № 6.– P. 493–507.
183. La Fleur A., Shi Y., Seelig G. Decoding biology with massively parallel reporter assays and machine learning // Genes Dev.– Cold Spring Harbor Laboratory, 2024.– Vol. 38, № 17-20.– P. 843–865.
184. Cao J., Novoa E.M., Zhang Z., Chen W.C.W., Liu D., Choi G.C.G., Wong A.S.L., Wehrspaun C., Kellis M., Lu T.K. High-throughput 5’ UTR engineering for enhanced protein production in non-viral gene therapies // Nat. Commun.– Springer Science and Business Media LLC, 2021.– Vol. 12, № 1.– P. 4138.
185. Soemedi R., Cygan K.J., Rhine C.L., Wang J., Bulacan C., Yang J., Bayrak-Toydemir P., McDonald J., Fairbrother W.G. Pathogenic variants that alter protein code often disrupt splicing // Nat. Genet.– Springer Science and Business Media LLC, 2017.– Vol. 49, № 6.– P. 848–855.
186. Friedman J.H. Greedy Function Approximation: A Gradient Boosting Machine // Ann. Stat.– Institute of Mathematical Statistics, 2001.– Vol. 29, № 5.– P. 1189–1232.
187. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.– ACM, 2016.– P. 785–794.
188. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y. LightGBM: a highly efficient gradient boosting decision tree // Proceedings of the 31st International Conference on Neural Information Processing Systems.– Red Hook, NY, USA: Curran Associates Inc., 2017.– P. 3149–3157.
189. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: unbiased boosting with categorical features // arXiv [cs.LG].– 2017.
190. McElfresh D., Khandagale S., Valverde J., C V.P., Feuer B., Hegde C., Ramakrishnan G., Goldblum M., White C. When do neural nets outperform boosted trees on tabular data? // arXiv [cs.LG].– 2023.
191. Jeffares A., Curth A., van der Schaar M. Deep learning through A telescoping lens: A simple model provides empirical insights on grokking, gradient boosting & beyond // arXiv [cs.LG].– 2024.
192. Babajide Mustapha I., Saeed F. Bioactive Molecule Prediction Using Extreme Gradient Boosting // Molecules.– 2016.– Vol. 21, № 8.
193. Li H., Peng J., Sidorov P., Leung Y., Leung K.-S., Wong M.-H., Lu G., Ballester P.J. Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data // Bioinformatics.– 2019.

194. Mikl M., Hamburg A., Pilpel Y., Segal E. Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries // Nat. Commun.– Springer Science and Business Media LLC, 2019.– Vol. 10, № 1.– P. 4572.
195. Mikl M., Eletto D., Nijim M., Lee M., Lafzi A., Mhamedi F., David O., Sain S.B., Handler K., Moor A.E. A massively parallel reporter assay reveals focused and broadly encoded RNA localization signals in neurons // Nucleic Acids Res.– Oxford University Press (OUP), 2022.– Vol. 50, № 18.– P. 10643–10664.
196. Zhou J., Troyanskaya O.G. Predicting effects of noncoding variants with deep learning-based sequence model // Nat. Methods.– 2015.– Vol. 12, № 10.– P. 931–934.
197. Zhou J., Theesfeld C.L., Yao K., Chen K.M., Wong A.K., Troyanskaya O.G. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk // Nat. Genet.– Nature Publishing Group, 2018.– Vol. 50, № 8.– P. 1171–1179.
198. Hao L., Kim J., Kwon S., Ha I.D. Deep learning-based survival analysis for high-dimensional survival data // Mathematics.– MDPI AG, 2021.– Vol. 9, № 11.– P. 1244.
199. Deng H., Zhou Y., Wang L., Zhang C. Ensemble learning for the early prediction of neonatal jaundice with genetic features // BMC Med. Inform. Decis. Mak.– Springer Science and Business Media LLC, 2021.– Vol. 21, № 1.– P. 338.
200. Avsec Ž., Weilert M., Shrikumar A., Krueger S., Alexandari A., Dalal K., Fropf R., McAnany C., Gagneur J., Kundaje A., Zeitlinger J. Base-resolution models of transcription-factor binding reveal soft motif syntax // Nat. Genet.– Nature Publishing Group, 2021.– Vol. 53, № 3.– P. 354–366.
201. Dey K.K., van de Geijn B., Kim S.S., Hormozdiari F., Kelley D.R., Price A.L. Evaluating the informativeness of deep learning annotations for human complex diseases // Nat. Commun.– Springer Science and Business Media LLC, 2020.– Vol. 11, № 1.– P. 4703.
202. Reddy A.J., Herschl M.H., Geng X., Kolli S., Lu A.X., Kumar A., Hsu P.D., Levine S., Ioannidis N.M. Strategies for effectively modelling promoter-driven gene expression using transfer learning // bioRxiv.– 2023.– P. 2023.02.24.529941.
203. Kao C.H., Trop E., Polen M., Schiff Y., de Almeida B.P., Gokaslan A., Pierrot T., Kuleshov V. ADVANCING DNA LANGUAGE MODELS: THE GENOMICS LONG-RANGE BENCHMARK.– 2024.
204. Dalla-Torre H., Gonzalez L., Mendoza-Revilla J., Lopez Carranza N., Grzywaczewski A.H., Oteri F., Dallago C., Trop E., de Almeida B.P., Sirelkhatim H., Richard G., Skwark M., Beguir K., Lopez M., Pierrot T. Nucleotide Transformer: building and evaluating robust foundation models for human genomics // Nat. Methods.– Springer Science and Business Media LLC, 2024.– P. 1–11.
205. Li X., Grandvalet Y., Davoine F. Explicit inductive bias for transfer learning with convolutional networks // arXiv [cs.LG].– 2018.
206. Eraslan G., Avsec Ž., Gagneur J., Theis F.J. Deep learning: new computational modelling techniques for genomics // Nat. Rev. Genet.– 2019.
207. Novakovsky G., Fornes O., Saraswat M., Mostafavi S., Wasserman W.W. ExplaiNN: interpretable and transparent neural networks for genomics // Genome Biol.– 2023.– Vol. 24, № 1.– P. 154.
208. Koo P.K., Eddy S.R. Representation learning of genomic sequence motifs with convolutional neural networks // PLoS Comput. Biol.– 2019.– Vol. 15, № 12.– P. e1007560.
209. Shrikumar A., Tian K., Avsec Ž., Shcherbina A., Banerjee A., Sharmin M., Nair S., Kundaje A. Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5 // arXiv [cs.LG].– 2018.
210. Barbadilla-Martínez L., Klaassen N., Franceschini-Santos V.H., Breda J., Hernandez-Quiles M., van Lieshout T., Urzua Traslaviña C.G., Yücel H., Boi M.C.L., Hermana-Garcia-Agullo C., Gregoricchio S., Zwart W., Voest E., Franke L., Vermeulen M., de Ridder J., van Steensel B. The regulatory grammar of human promoters uncovered by MPRA-trained deep learning // bioRxiv.– 2024.– P. 2024.07.09.602649.
211. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition // arXiv [cs.CV].– 2015.
212. Ioffe S., Szegedy C. Batch Normalization: Accelerating deep network training by reducing internal covariate shift // arXiv [cs.LG].– 2015.
213. Kim K.-S., Choi Y.-S. HyAdamC: A new Adam-based hybrid optimization algorithm for convolution neural networks // Sensors (Basel).– MDPI AG, 2021.– Vol. 21, № 12.– P. 4054.
214. Doshi K. Batch Norm Explained Visually — How it works, and why neural networks need it [Electronic resource] // Towards Data Science.– 2021.– URL: https://towardsdatascience.com/batch-norm-explained-visually-how-it-works-and-why-neural-networks-need-it-b18919692739 (accessed: 02.12.2024).

215. Yu F., Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions // arXiv [cs.CV].– 2015.
216. Adaloglou N. Understanding the receptive field of deep convolutional networks // AI Summer.– Sergios Karagiannakos, 2020.
217. Sharma S., Mehra R. Implications of Pooling Strategies in Convolutional Neural Networks: A Deep Insight // Found. Comput. Decision Sci.– 2019.– Vol. 44, № 3.– P. 303–330.
218. Hochreiter S., Schmidhuber J. Long short-term memory // Neural Comput.– 1997.– Vol. 9, № 8.– P. 1735–1780.
219. Stiehler F., Steinborn M., Scholz S., Dey D., Weber A.P.M., Denton A.K. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning // Bioinformatics.– Oxford University Press (OUP), 2021.– Vol. 36, № 22-23.– P. 5291–5298.
220. Gabriel L., Becker F., Hoff K.J., Stanke M. Tiberius: End-to-End Deep Learning with an HMM for Gene Prediction: biorxiv;2024.07.21.604459v1 // Bioinformatics.– bioRxiv, 2024.
221. Quang D., Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences // Nucleic Acids Res.– 2016.– Vol. 44, № 11.– P. e107.
222. Quang D., Xie X. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data // Methods.– Elsevier BV, 2019.– Vol. 166.– P. 40–47.
223. Schmidinger N., Schneckenreiter L., Seidl P., Schimunek J., Hoedt P.-J., Brandstetter J., Mayr A., Luukkonen S., Hochreiter S., Klambauer G. Bio-xLSTM: Generative modeling, representation and in-context learning of biological and chemical sequences // arXiv.– 2024.
224. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention Is All You Need // arXiv [cs.CL].– 2017.
225. Lin J., Luo R., Pinello L. EPInformer: a scalable deep learning framework for gene expression prediction by integrating promoter-enhancer sequences with multimodal epigenomic data // bioRxiv.– 2024.– P. 2024.08.01.606099.
226. Nguyen E., Poli M., Faizi M., Thomas A., Birch-Sykes C., Wornow M., Patel A., Rabideau C., Massaroli S., Bengio Y., Ermon S., Baccus S.A., Ré C. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution // ArXiv.– 2023.
227. Schiff Y., Kao C.-H., Gokaslan A., Dao T., Gu A., Kuleshov V. Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling // arXiv [q-bio.GN].– 2024.
228. Kuratov Y., Shmelev A., Fishman V., Kardymon O., Burtsev M. Recurrent memory augmentation of GENA-LM improves performance on long DNA sequence tasks.– 2024.
229. Luo W., Li Y., Urtasun R., Zemel R. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks // arXiv [cs.CV].– 2017.
230. Chen Y., Zhang X., Hu S., Han X., Liu Z., Sun M. Stuffed Mamba: State collapse and state capacity of RNN-based long-context modeling // arXiv [cs.CL].– 2024.
231. Kuratov Y., Bulatov A., Anokhin P., Rodkin I., Sorokin D., Sorokin A., Burtsev M. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack // arXiv [cs.CL].– 2024.
232. Understanding LSTM Networks [Electronic resource].– URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (accessed: 22.07.2024).
233. Shorten C., Khoshgoftaar T.M. A survey on Image Data Augmentation for Deep Learning // Journal of Big Data.– SpringerOpen, 2019.– Vol. 6, № 1.– P. 1–48.
234. Feng S.Y., Gangal V., Wei J., Chandar S., Vosoughi S., Mitamura T., Hovy E. A Survey of Data Augmentation Approaches for NLP // arXiv [cs.CL].– 2021.
235. Zhang H., Cisse M., Dauphin Y.N., Lopez-Paz D. mixup: Beyond Empirical Risk Minimization // arXiv [cs.LG].– 2017.
236. Yun S., Han D., Oh S.J., Chun S., Choe J., Yoo Y. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features // arXiv [cs.CV].– 2019.
237. Buslaev A., Iglovikov V.I., Khvedchenya E., Parinov A., Druzhinin M., Kalinin A.A. Albumentations: Fast and Flexible Image Augmentations // Information.– Multidisciplinary Digital Publishing Institute, 2020.– Vol. 11, № 2.– P. 125.
238. Lee N.K., Tang Z., Toneyan S., Koo P.K. EvoAug: improving generalization and interpretability of genomic deep neural networks with evolution-inspired data augmentations // Genome Biol.– 2023.– Vol. 24, № 1.– P. 105.
239. Duncan A.G., Mitchell J.A., Moses A.M. Improving the performance of supervised deep learning for regulatory genomics using phylogenetic augmentation // Bioinformatics.– 2024.– Vol. 40, № 4.
240. Zhou H., Shrikumar A., Kundaje A. Towards a Better Understanding of Reverse-Complement

Equivariance for Deep Learning Models in Genomics // Proceedings of the 16th Machine Learning in Computational Biology meeting / ed. Knowles D.A., Mostafavi S., Lee S.-I.– PMLR, 22--23 Nov 2022.– Vol. 165.– P. 1–33.
241. Wang G., Li W., Aertsen M., Deprest J., Ourselin S., Vercauteren T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks // Neurocomputing.– 2019.– Vol. 335.– P. 34–45.
242. Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., Bates R., Žídek A., Potapenko A., Bridgland A., Meyer C., Kohl S.A.A., Ballard A.J., Cowie A., Romera-Paredes B., Nikolov S., Jain R., Adler J., Back T., Petersen S., Reiman D., Clancy E., Zielinski M., Steinegger M., Pacholska M., Berghammer T., Bodenstein S., Silver D., Vinyals O., Senior A.W., Kavukcuoglu K., Kohli P., Hassabis D. Highly accurate protein structure prediction with AlphaFold // Nature.– 2021.– Vol. 596, № 7873.– P. 583–589.
243. Shrikumar A., Greenside P., Kundaje A. Reverse-complement parameter sharing improves deep learning models for genomics // bioRxiv.– 2017.– P. 103663.
244. Mallet V., Vert J.-P. Reverse-Complement Equivariant Networks for DNA Sequences // bioRxiv.– 2021.– P. 2021.06.03.446953.
245. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome // Nature.– 2012.– Vol. 489, № 7414.– P. 57–74.
246. Yuan H., Kelley D.R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks // Nat. Methods.– 2022.– Vol. 19, № 9.– P. 1088–1096.
247. Gosai S.J., Castro R.I., Fuentes N., Butts J.C., Mouri K., Alasoadura M., Kales S., Nguyen T.T.L., Noche R.R., Rao A.S., Joy M.T., Sabeti P.C., Reilly S.K., Tewhey R. Machine-guided design of cell-type-targeting cis-regulatory elements // Nature.– Nature Publishing Group, 2024.– P. 1–10.
248. Kelley D.R. Cross-species regulatory sequence activity prediction // PLoS Comput. Biol.– 2020.– Vol. 16, № 7.– P. e1008050.
249. Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation // arXiv [cs.CV].– 2015.
250. Zeitlinger J. Seven myths of how transcription factors read the cis-regulatory code // Curr. Opin. Syst. Biol.– Elsevier BV, 2020.– Vol. 23.– P. 22–31.
251. Schreiber J. bpnet-lite: This repository hosts a minimal version of a Python API for BPNet.– Github.
252. Ellington C.N., Sun N., Ho N., Tao T., Mahbub S., Li D., Zhuang Y., Wang H., Song L., Xing E.P. Accurate and General DNA Representations Emerge from Genome Foundation Models at Scale: biorxiv;2024.12.01.625444v1 // Bioinformatics.– bioRxiv, 2024.
253. Maslova A., Ramirez R.N., Ma K., Schmutz H., Wang C., Fox C., Ng B., Benoist C., Mostafavi S., Immunological Genome Project. Deep learning of immune cell differentiation // Proc. Natl. Acad. Sci. U. S. A.– Proceedings of the National Academy of Sciences, 2020.– Vol. 117, № 41.– P. 25655–25666.
254. Kaplan J., McCandlish S., Henighan T., Brown T.B., Chess B., Child R., Gray S., Radford A., Wu J., Amodei D. Scaling laws for neural language models // arXiv [cs.LG].– 2020.
255. Open Problems - Multimodal Single-Cell Integration [Electronic resource].– URL: https://kaggle.com/competitions/open-problems-multimodal (accessed: 11.11.2024).
256. Mitra S., Malik R., Wong W., Rahman A., Hartemink A.J., Pritykin Y., Dey K.K., Leslie C.S. Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis // Nat. Genet.– Springer Science and Business Media LLC, 2024.– Vol. 56, № 4.– P. 627–636.
257. DaSilva L.F., Senan S., Patel Z.M., Janardhan Reddy A., Gabbita S., Nussbaum Z., Valdez Córdova C.M., Wenteler A., Weber N., Tunjic T.M., Ahmad Khan T., Li Z., Smith C., Bejan M., Karmel Louis L., Cornejo P., Connell W., Wong E.S., Meuleman W., Pinello L. DNA-diffusion: Leveraging generative models for controlling chromatin accessibility and gene expression via synthetic regulatory elements // bioRxivorg.– 2024.– P. 2024.02.01.578352.
258. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv [cs.CL].– 2018.
259. Doersch C., Gupta A., Efros A.A. Unsupervised visual representation learning by context prediction // arXiv [cs.CV].– 2015.
260. OpenAI, Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I., Aleman F.L., Almeida D., Altenschmidt J., Altman S., Anadkat S., Avila R., Babuschkin I., Balaji S., Balcom V., Baltescu P., Bao H., Bavarian M., Belgum J., Bello I., Berdine J., Bernadett-Shapiro G., Berner C., Bogdonoff L., Boiko O., Boyd M., Brakman A.-L., Brockman G., Brooks T., Brundage M., Button K., Cai T., Campbell R.,

Cann A., Carey B., Carlson C., Carmichael R., Chan B., Chang C., Chantzis F., Chen D., Chen S., Chen R., Chen J., Chen M., Chess B., Cho C., Chu C., Chung H.W., Cummings D., Currier J., Dai Y., Decareaux C., Degry T., Deutsch N., Deville D., Dhar A., Dohan D., Dowling S., Dunning S., Ecoffet A., Eleti A., Eloundou T., Farhi D., Fedus L., Felix N., Fishman S.P., Forte J., Fulford I., Gao L., Georges E., Gibson C., Goel V., Gogineni T., Goh G., Gontijo-Lopes R., Gordon J., Grafstein M., Gray S., Greene R., Gross J., Gu S.S., Guo Y., Hallacy C., Han J., Harris J., He Y., Heaton M., Heidecke J., Hesse C., Hickey A., Hickey W., Hoeschele P., Houghton B., Hsu K., Hu S., Hu X., Huizinga J., Jain S., Jain S., Jang J., Jiang A., Jiang R., Jin H., Jin D., Jomoto S., Jonn B., Jun H., Kaftan T., Kaiser Ł., Kamali A., Kanitscheider I., Keskar N.S., Khan T., Kilpatrick L., Kim J.W., Kim C., Kim Y., Kirchner J.H., Kiros J., Knight M., Kokotajlo D., Kondraciuk Ł., Kondrich A., Konstantinidis A., Kosic K., Krueger G., Kuo V., Lampe M., Lan I., Lee T., Leike J., Leung J., Levy D., Li C.M., Lim R., Lin M., Lin S., Litwin M., Lopez T., Lowe R., Lue P., Makanju A., Malfacini K., Manning S., Markov T., Markovski Y., Martin B., Mayer K., Mayne A., McGrew B., McKinney S.M., McLeavey C., McMillan P., McNeil J., Medina D., Mehta A., Menick J., Metz L., Mishchenko A., Mishkin P., Monaco V., Morikawa E., Mossing D., Mu T., Murati M., Murk O., Mély D., Nair A., Nakano R., Nayak R., Neelakantan A., Ngo R., Noh H., Ouyang L., O'Keefe C., Pachocki J., Paino A., Palermo J., Pantuliano A., Parascandolo G., Parish J., Parparita E., Passos A., Pavlov M., Peng A., Perelman A., de Avila Belbute Peres F., Petrov M., de Oliveira Pinto H.P., Michael, Pokorny, Pokrass M., Pong V.H., Powell T., Power A., Power B., Proehl E., Puri R., Radford A., Rae J., Ramesh A., Raymond C., Real F., Rimbach K., Ross C., Rotsted B., Roussez H., Ryder N., Saltarelli M., Sanders T., Santurkar S., Sastry G., Schmidt H., Schnurr D., Schulman J., Selsam D., Sheppard K., Sherbakov T., Shieh J., Shoker S., Shyam P., Sidor S., Sigler E., Simens M., Sitkin J., Slama K., Sohl I., Sokolowsky B., Song Y., Staudacher N., Such F.P., Summers N., Sutskever I., Tang J., Tezak N., Thompson M.B., Tillet P., Tootoonchian A., Tseng E., Tuggle P., Turley N., Tworek J., Uribe J.F.C., Vallone A., Vijayvergiya A., Voss C., Wainwright C., Wang J.J., Wang A., Wang B., Ward J., Wei J., Weinmann C.J., Welihinda A., Welinder P., Weng J., Weng L., Wiethoff M., Willner D., Winter C., Wolrich S., Wong H., Workman L., Wu S., Wu J., Wu M., Xiao K., Xu T., Yoo S., Yu K., Yuan Q., Zaremba W., Zellers R., Zhang C., Zhang M., Zhao S., Zheng T., Zhuang J., Zhuk W., Zoph B. GPT-4 Technical Report // arXiv [cs.CL].– 2023.

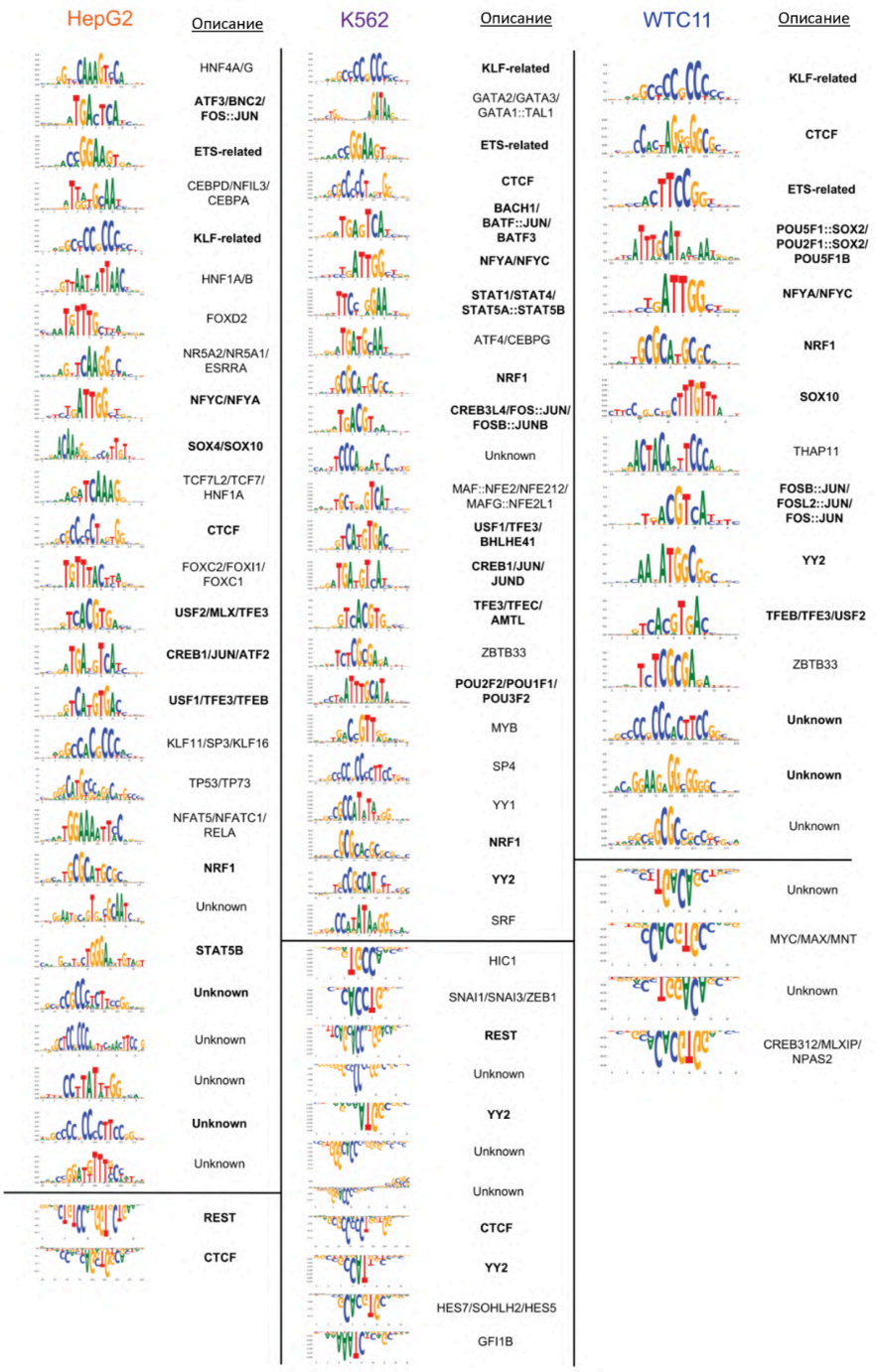
261. Caron M., Touvron H., Misra I., Jégou H., Mairal J., Bojanowski P., Joulin A. Emerging Properties in Self-Supervised Vision Transformers // arXiv [cs.CV].– 2021.
262. Liu Y., Zhang K., Li Y., Yan Z., Gao C., Chen R., Yuan Z., Huang Y., Sun H., Gao J., He L., Sun L. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models // arXiv [cs.CV].– 2024.
263. Frazer J., Notin P., Dias M., Gomez A., Min J.K., Brock K., Gal Y., Marks D.S. Disease variant prediction with deep generative models of evolutionary data // Nature.– 2021.– Vol. 599, № 7883.– P. 91–95.
264. Meier J., Rao R., Verkuil R., Liu J., Sercu T., Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function: biorxiv;2021.07.09.450648v2 // Synthetic Biology.– bioRxiv, 2021.
265. Teufel F., Almagro Armenteros J.J., Johansen A.R., Gíslason M.H., Pihl S.I., Tsirigos K.D., Winther O., Brunak S., von Heijne G., Nielsen H. SignalP 6.0 predicts all five types of signal peptides using protein language models // Nat. Biotechnol.– 2022.– Vol. 40, № 7.– P. 1023–1025.
266. Ji Y., Zhou Z., Liu H., Davuluri R.V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome // Bioinformatics.– 2021.– Vol. 37, № 15.– P. 2112–2120.
267. Radford A., Narasimhan K. Improving language understanding by generative pre-training.– 2018.
268. Xu Z., Gupta R., Cheng W., Shen A., Shen J., Talwalkar A., Khodak M. Specialized foundation models struggle to beat supervised baselines // arXiv [cs.LG].– 2024.
269. Marin F.I., Teufel F., Horlacher M., Madsen D., Pultz D., Winther O., Boomsma W. BEND: Benchmarking DNA Language Models on biologically meaningful tasks // arXiv [q-bio.GN].– 2023.
270. Gordon M.G., Inoue F., Martin B., Schubach M., Agarwal V., Whalen S., Feng S., Zhao J., Ashuach T., Ziffra R., Kreimer A., Georgakopoulos-Soares I., Yosef N., Ye C.J., Pollard K.S., Shendure J., Kircher M., Ahituv N. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements // Nat. Protoc.– 2020.– Vol. 15, № 8.– P. 2387–2412.
271. Neumayr C., Pagani M., Stark A., Arnold C.D. STARR-seq and UMI-STARR-seq: Assessing Enhancer Activities for Genome-Wide-, High-, and Low-Complexity Candidate Libraries // Curr. Protoc. Mol.

Biol.– 2019.– Vol. 128, № 1.– P. e105.
272. de Almeida B.P., Reiter F., Pagani M., Stark A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers // Nat. Genet.– 2022.– Vol. 54, № 5.– P. 613–624.
273. Frömel R., Rühle J., Bernal Martinez A., Szu-Tu C., Pacheco Pastor F., Martinez Corral R., Velten L. Synthetic enhancers reveal design principles of cell state specific regulatory elements in hematopoiesis: biorxiv;2024.08.26.609645v1 // Molecular Biology.– bioRxiv, 2024.
274. Trauernicht M., Martinez-Ara M., van Steensel B. Deciphering gene regulation using massively parallel reporter assays // Trends Biochem. Sci.– Elsevier BV, 2020.– Vol. 45, № 1.– P. 90–91.
275. Movva R., Greenside P., Marinov G.K., Nair S., Shrikumar A., Kundaje A. Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays // PLoS One.– Public Library of Science (PLoS), 2019.– Vol. 14, № 6.– P. e0218073.
276. Gosai S.J., Castro R.I., Fuentes N., Butts J.C., Kales S., Noche R.R., Mouri K., Sabeti P.C., Reilly S.K., Tewhey R. Machine-guided design of synthetic cell type-specific cis-regulatory elements // bioRxiv.– 2023.
277. Hecker N., Kempynck N., Mauduit D., Abaffyová D., Vandepoel R., Dieltiens S., Sarropoulos I., González-Blas C.B., Leysen E., Moors R., Hulselmans G., Lim L., De Wit J., Christiaens V., Poovathingal S., Aerts S. Enhancer-driven cell type comparison reveals similarities between the mammalian and bird pallium // bioRxiv.– 2024.– P. 2024.04.17.589795.
278. Nair S., Shrikumar A., Schreiber J., Kundaje A. fastISM: performant in silico saturation mutagenesis for convolutional neural networks // Bioinformatics.– 2022.– Vol. 38, № 9.– P. 2397–2403.
279. Ribeiro M.T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.– New York, NY, USA: Association for Computing Machinery, 2016.– P. 1135–1144.
280. Tareen A., Kooshkbaghi M., Posfai A., Ireland W.T., McCandlish D.M., Kinney J.B. MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect // Genome Biol.– 2022.– Vol. 23, № 1.– P. 98.
281. Seitz E.E., McCandlish D.M., Kinney J.B., Koo P.K. Interpreting cis-regulatory mechanisms from genomic deep neural networks using surrogate models // Nature Machine Intelligence.– Nature Publishing Group, 2024.– Vol. 6, № 6.– P. 701–713.
282. Gou J., Yu B., Maybank S.J., Tao D. Knowledge Distillation: A Survey // Int. J. Comput. Vis.– 2021.– Vol. 129, № 6.– P. 1789–1819.
283. Sasse A., Chikina M., Mostafavi S. Quick and effective approximation of in silico saturation mutagenesis experiments with first-order Taylor expansion // bioRxiv.– 2023.– P. 2023.11.10.566588.
284. Simonyan K., Vedaldi A., Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps // arXiv [cs.CV].– 2013.
285. Majdandzic A., Rajesh C., Koo P.K. Correcting gradient-based interpretations of deep neural networks for genomics // Genome Biol.– 2023.– Vol. 24, № 1.– P. 109.
286. Shrikumar A., Greenside P., Kundaje A. Learning Important Features Through Propagating Activation Differences // Proceedings of the 34th International Conference on Machine Learning / ed. Precup D., Teh Y.W.– PMLR, 06--11 Aug 2017.– Vol. 70.– P. 3145–3153.
287. Chen H., Lundberg S.M., Lee S.-I. Explaining a series of models by propagating Shapley values // Nat. Commun.– 2022.– Vol. 13, № 1.– P. 4512.
288. Sundararajan M., Taly A., Yan Q. Axiomatic Attribution for Deep Networks // Proceedings of the 34th International Conference on Machine Learning / ed. Precup D., Teh Y.W.– PMLR, 06--11 Aug 2017.– Vol. 70.– P. 3319–3328.
289. Balcı A.T., Ebeid M.M., Benos P.V., Kostka D., Chikina M. An intrinsically interpretable neural network architecture for sequence-to-function learning // Bioinformatics.– 2023.– Vol. 39, № 39 Suppl 1.– P. i413–i422.
290. Yin C., Hair S.C., Byeon G.W., Bromley P., Meuleman W., Seelig G. Iterative deep learning-design of human enhancers exploits condensed sequence grammar to achieve cell type-specificity: biorxiv;2024.06.14.599076v1 // Synthetic Biology.– bioRxiv, 2024.
291. Johnson L.A., Zhao Y., Golden K., Barolo S. Reverse-engineering a transcriptional enhancer: a case study in Drosophila // Tissue Eng. Part A.– Mary Ann Liebert Inc, 2008.– Vol. 14, № 9.– P. 1549–1559.
292. Vincent B.J., Estrada J., DePace A.H. The appeasement of Doug: a synthetic approach to enhancer biology // Integr. Biol. .– Oxford University Press (OUP), 2016.– Vol. 8, № 4.– P. 475–484.
293. Kotopka B.J., Smolke C.D. Model-driven generation of artificial yeast promoters // Nat. Commun.–

Springer Science and Business Media LLC, 2020.– Vol. 11, № 1.– P. 2113.
294. Katoch S., Chauhan S.S., Kumar V. A review on genetic algorithm: past, present, and future // Multimed. Tools Appl.– Springer Science and Business Media LLC, 2021.– Vol. 80, № 5.– P. 8091–8126.
295. Tripp A., Hernández-Lobato J.M. Genetic algorithms are strong baselines for molecule generation // arXiv [cs.NE].– 2023.
296. DeepDream: How Alexander Mordvintsev excavated the computer's hidden layers // The Artist in the Machine.– The MIT Press, 2019.– P. 59–70.
297. Schreiber J., Lu Y.Y., Noble W.S. Ledidi: Designing genomic edits that induce functional activity // bioRxiv.– bioRxiv, 2020.
298. Linder J., Seelig G. Fast activation maximization for molecular sequence design // BMC Bioinformatics.– Springer Science and Business Media LLC, 2021.– Vol. 22, № 1.– P. 510.
299. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative adversarial networks // Commun. ACM.– Association for Computing Machinery (ACM), 2020.– Vol. 63, № 11.– P. 139–144.
300. An Overview of Deep Generative Models in Functional and Evolutionary Genomics // Annual Review of Biomedical Data Science.
301. Zhang D., Zhang W., Zhao Y., Zhang J., He B., Qin C., Yao J. DNAGPT: A generalized pre-trained tool for versatile DNA sequence analysis tasks // arXiv [q-bio.GN].– 2023.
302. Nguyen E., Poli M., Durrant M.G., Thomas A.W., Kang B., Sullivan J., Ng M.Y., Lewis A., Patel A., Lou A., Ermon S., Baccus S.A., Hernandez-Boussard T., Re C., Hsu P.D., Hie B.L. Sequence modeling and design from molecular to genome scale with Evo // bioRxiv.– 2024.
303. Ho J., Jain A., Abbeel P. Denoising Diffusion Probabilistic Models // arXiv [cs.LG].– 2020.
304. Avdeyev P., Shi C., Tan Y., Dudnyk K., Zhou J. Dirichlet diffusion score model for biological sequence generation // arXiv [cs.LG].– 2023.
305. Sarkar A., Tang Z., Zhao C., Koo P. Designing DNA with tunable regulatory activity using Discrete Diffusion // bioRxiv.– 2024.
306. Denoising Diffusion Models: A Generative Learning Big Bang [Electronic resource] // Denoising Diffusion-based Generative Modeling: Foundations and Applications.– URL: https://cvpr2023-tutorial-diffusion-models.github.io/ (accessed: 10.12.2024).
307. Dhariwal P., Nichol A. Diffusion models beat GANs on image synthesis // arXiv [cs.LG].– 2021.
308. Ho J., Salimans T. Classifier-Free Diffusion Guidance // NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications.– 2021.
309. Watson J.L., Juergens D., Bennett N.R., Trippe B.L., Yim J., Eisenach H.E., Ahern W., Borst A.J., Ragotte R.J., Milles L.F., Wicky B.I.M., Hanikel N., Pellock S.J., Courbet A., Sheffler W., Wang J., Venkatesh P., Sappington I., Torres S.V., Lauko A., De Bortoli V., Mathieu E., Ovchinnikov S., Barzilay R., Jaakkola T.S., DiMaio F., Baek M., Baker D. De novo design of protein structure and function with RFdiffusion // Nature.– Springer Science and Business Media LLC, 2023.– Vol. 620, № 7976.– P. 1089–1100.
310. Weiss T., Mayo Yanes E., Chakraborty S., Cosmo L., Bronstein A.M., Gershoni-Poranne R. Guided diffusion for inverse molecular design // Nat. Comput. Sci.– Springer Science and Business Media LLC, 2023.– Vol. 3, № 10.– P. 873–882.
311. Stark H., Jing B., Wang C., Corso G., Berger B., Barzilay R., Jaakkola T. Dirichlet flow matching with applications to DNA sequence design // ArXiv.– 2024.
312. Patwardhan R.P., Hiatt J.B., Witten D.M., Kim M.J., Smith R.P., May D., Lee C., Andrie J.M., Lee S.-I., Cooper G.M., Ahituv N., Pennacchio L.A., Shendure J. Massively parallel functional dissection of mammalian enhancers in vivo // Nat. Biotechnol.– 2012.– Vol. 30, № 3.– P. 265–270.
313. Ambrosini G., Vorontsov I., Penzar D., Groux R., Fornes O., Nikolaeva D.D., Ballester B., Grau J., Grosse I., Makeev V., Kulakovskiy I., Bucher P. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study // Genome Biol.– Springer Science and Business Media LLC, 2020.– Vol. 21, № 1.– P. 114.
314. Kulakovskiy I.V., Vorontsov I.E., Yevshin I.S., Sharipov R.N., Fedorova A.D., Rumynskiy E.I., Medvedeva Y.A., Magana-Mora A., Bajic V.B., Papatsenko D.A., Kolpakov F.A., Makeev V.J. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis // Nucleic Acids Res.– 2018.– Vol. 46, № D1.– P. D252–D259.
315. Maurano M.T., Haugen E., Sandstrom R., Vierstra J., Shafer A., Kaul R., Stamatoyannopoulos J.A. Large-scale identification of sequence variants influencing human transcription factor occupancy in

vivo // Nat. Genet.– Springer Science and Business Media LLC, 2015.– Vol. 47, № 12.– P. 1393–1401.
316. Fortin F.-A., Rainville F., Gardner M.-A., Parizeau M., Gagné C. DEAP: evolutionary algorithms made easy // J. Mach. Learn. Res.– 2012.– Vol. 13.– P. 2171–2175.
317. Smith L.N., Topin N. Super-convergence: Very fast training of neural networks using large learning rates // arXiv [cs.LG].– 2017.
318. Howard J., Gugger S. fastai: A Layered API for Deep Learning // arXiv [cs.LG].– 2020.
319. Smith L.N. A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay // arXiv [cs.LG].– 2018.
320. Loshchilov I., Hutter F. Decoupled weight decay regularization // arXiv [cs.LG].– 2017.
321. Chen X., Liang C., Huang D., Real E., Wang K., Liu Y., Pham H., Dong X., Luong T., Hsieh C.-J., Lu Y., Le Q.V. Symbolic discovery of optimization algorithms // arXiv [cs.LG].– 2023.
322. 1-Cycle Schedule [Electronic resource] // DeepSpeed.– 2024.– URL: https://www.deepspeed.ai/tutorials/one-cycle/ (accessed: 28.10.2024).
323. Rana A. Handling the Woes of Training // Aditya Rana Blog.– 2021.
324. Smith R.P., Taher L., Patwardhan R.P., Kim M.J., Inoue F., Shendure J., Ovcharenko I., Ahituv N. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model // Nat. Genet.– Springer Science and Business Media LLC, 2013.– Vol. 45, № 9.– P. 1021–1028.
325. Ernst J., Melnikov A., Zhang X., Wang L., Rogov P., Mikkelsen T.S., Kellis M. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions // Nat. Biotechnol.– Springer Science and Business Media LLC, 2016.– Vol. 34, № 11.– P. 1180–1190.
326. Inoue F., Kreimer A., Ashuach T., Ahituv N., Yosef N. Identification and massively parallel characterization of regulatory elements driving neural induction // Cell Stem Cell.– Elsevier BV, 2019.– Vol. 25, № 5.– P. 713–727.e10.
327. Breiman L. Random Forests // Mach. Learn.– 2001.– Vol. 45, № 1.– P. 5–32.
328. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É. Scikit-learn: Machine Learning in Python // J. Mach. Learn. Res.– 2011.– Vol. 12, № Oct.– P. 2825–2830.
329. Shi W., Fornes O., Mathelier A., Wasserman W.W. Evaluating the impact of single nucleotide variants on transcription factor binding // Nucleic Acids Res.– 2016.– Vol. 44, № 21.– P. 10106–10116.
330. Sauvalle B., de La Fortelle A. Unsupervised multi-object segmentation using attention and soft-argmax // arXiv [cs.CV].– 2022.
331. Huang G., Liu Z., van der Maaten L., Weinberger K.Q. Densely connected convolutional networks // arXiv [cs.CV].– 2016.
332. Silver N.C., Hittner J.B., May K. Testing dependent correlations with nonoverlapping variables: A Monte Carlo simulation // J. Exp. Educ.– Informa UK Limited, 2004.– Vol. 73, № 1.– P. 53–69.
333. Gopalakrishnan V. REST and the RESTless: in stem cells and beyond // Future Neurol.– Informa UK Limited, 2009.– Vol. 4, № 3.– P. 317–329.
334. Grant C.E., Bailey T.L., Noble W.S. FIMO: scanning for occurrences of a given motif // Bioinformatics.– Oxford University Press (OUP), 2011.– Vol. 27, № 7.– P. 1017–1018.

Приложения



Приложение 1. Мотивы связывания факторов транскрипции, учитываемые MPRALegNet. b обнаруженных с помощью TF-MoDISco для каждого из трёх оцененных типов клеток. Связывающие сайты транскрипционных факторов (TFBS), ассоциированные с подавлением транскрипции (например, REST), ориентированы вверх ногами и показаны ниже горизонтальных линий. TFBS, обнаруженные как минимум в двух типах клеток (т.е., вероятно, связанные с общими транскрипционными факторами), выделены полужирным шрифтом.