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

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Москва - 2025

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте общей генетики им. Н. И. Вавилова Российской академии наук (ИОГен РАН) в лаборатории системной биологии и вычислительной генетики.

Научный руководитель - ***Кулаковский Иван Владимирович,*** *доктор биологических наук*

Официальные оппоненты - ***Храмеева Екатерина Евгеньевна*** *- доктор биологических наук, доцент Центра молекулярной и клеточной биологии Автономной некоммерческой образовательной организации высшего образования «Сколковский институт науки и технологий»*

Уткин Лев Владимирович *- доктор технических наук, профессор, главный научный сотрудник, профессор Высшей школы технологий искусственного интеллекта Института компьютерных наук и кибербезопасности Федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого»*

Орлов Юрий Львович *- доктор биологических наук, профессор РАН, профессор кафедры информационных технологий и обработки медицинских данных Центра цифровой медицины Института цифрового биодизайна и моделирования живых систем Федерального государственного автономного образовательного учреждения высшего образования Первый Московский государственный медицинский университет имени И.М. Сеченова Министерства здравоохранения Российской Федерации (Сеченовский Университет)*

Защита диссертации состоится 5 марта 2025 г. в 17:00 на заседании диссертационного совета МГУ.015.10 Московского государственного университета имени М.В. Ломоносова по адресу: 119234, Москва, Ленинские горы, дом 1, стр. 73, Факультет биоинженерии и биоинформатики, ауд. 221.
E-mail: dissovet@belozersky.msu.ru
С диссертацией можно ознакомиться в отделе диссертаций Научной библиотеки МГУ имени М.В. Ломоносова (Москва, Ломоносовский просп., д. 27) и на портале: https://dissovet.msu.ru/dissertation/3329 .

Автореферат разослан «___» февраля 2025 г.

Ученый секретарь диссертационного совета,
кандидат химических наук И.В. Шаповалова

Общая характеристика работы

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 (Bagger et al. 2024; Hawkes et al. 2024). 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 (Albert and Kruglyak 2015; Uffelmann et al. 2021; Mostafavi et al. 2023; Bagger et al. 2024). В то время как для аннотации замен в белок-кодирующих районах уже существуют общепринятые и хорошо себя зарекомендовавшие подходы (Cheng et al. 2023; Abramson et al. 2024), инструменты для анализа нуклеотидных замен в некодирующих областях генов, на которые приходится порядка 90% клинически значимых мутаций (Hindorff et al. 2009; Edwards et al. 2013; Farh et al. 2015; Khurana et al. 2016; Rojano et al. 2019; Walavalkar and Notani 2020), 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 (Sasse, Chikina, and Mostafavi 2024).

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 (Linder et al. 2023; Sasse, Chikina, and Mostafavi 2024). 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 (Sasse et al. 2023; Karollus, Mauermeier, and Gagneur 2023; C. Huang et al. 2023; Bajwa et al. 2023; Kathail et al. 2024). 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 (Patwardhan et al. 2009; Hiatt et al. 2010; White et al. 2013; Inoue and Ahituv 2015; van Arensbergen et al. 2019; Romanov and Laktionov 2023). 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.

Степень научной разработанности темы

Общей проблемой для биоинформатикческих методов приоритизации вариантов как картирование локусов количественных признаков (Quantitative Trait Loci, QTL)(Kearsey 1998; Albert and Kruglyak 2015) и полногеномных ассоциативных исследований (Genome-Wide Association Studies, GWAS) (Uffelmann et al. 2021) является 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" (Astle et al. 2016; van Arensbergen et al. 2019). 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 (Hindorff et al. 2009; Edwards et al. 2013; Farh et al. 2015; Khurana et al. 2016; Rojano et al. 2019; Walavalkar and Notani 2020). В свою очередь среди некодирующих вариантов наибольшая доля приходится на мутации в регуляторных областях, контролирующих транскрипцию – промоторах и энхансерах.

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 (Musunuru et al. 2010; Afanasyeva et al. 2017; Uvarova et al. 2023) или одновременное тестирование множества вариантов в массовых параллельных репортерных экспериментах (МПРЭ, van Arensbergen et al. 2019; Choi et al. 2020;

Weiss et al. 2021) и скринингах при помощи высокопроизводительных методов, основанных на технологии CRISPR (Bock et al. 2022; Morris et al. 2023; Ryu et al. 2024). 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 (Sasse, Chikina, and Mostafavi 2024; Zeiltinger et al. 2024).

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 (Stormo et al. 1982; Vorontsov, Kulakovskiy, and Khimulya 2015). Из подходов, основанных на классическом машинном обучении, популярность получил подход gkmSVM/deltaSVM(gapped-kmer/delta Support Vector Machine) (Lee et al. 2015), занявший первое место в нескольких открытых соревнованиях по предсказанию влияния однонуклеотидных вариантов на регуляцию экспрессии генов (Lee et al. 2015; Shigaki et al. 2019). 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.

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 (Zhou and Troyanskaya 2015; Kelley, Snoek, and Rinn 2016; Avsec et al. 2021; Kelley et al. 2018; Chen et al. 2022; Linder et al. 2023). 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 (Zhou and Troyanskaya 2015; Kelley et al. 2018; Avsec et al. 2021; Linder et al. 2023).

Нейросетевые модели демонстрируют хорошую согласованность между предсказаниями и данными насыщающего мутагенеза в промоторах, и успешно определяют некоторую часть причинных eQTL (Zhou and Troyanskaya 2015; Kelley, Snoek, and Rinn 2016; Avsec et al. 2021; Kelley et al. 2018; Chen et al. 2022; Linder et al. 2023). 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 (Sasse et al. 2023; Karollus, Mauermeier, and Gagneur 2023; C. Huang et al. 2023; Bajwa et al. 2023; Kathail et al. 2024).

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 (Sasse et al. 2023; Karollus, Mauermeier, and Gagneur 2023; C. Huang et al. 2023; Bajwa et al. 2023; Kathail et al. 2024). При этом, к сожалению, современные достижения в области дизайна архитектур нейронных сетей и их обучения практически не используются (Hu et al. 2017; Tan and Le 2021; Liu et al. 2022).

Помимо задачи оценки эффекта вариантов, широкое распространение получает применение нейронных сетей для задач генерации новых объектов в самых различных областях, включая задачи генетики и молекулярной биологии (Kadurin et al. 2017; Zhavoronkov et al. 2019; Wang et al. 2020; Vaishnav et al. 2022; Taskiran et al. 2024; Lal et al. 2024). В частности, модели на основе диффузионных процессов (Rombach et al. 2022; Bansal et al. 2022) 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.

Цель и задачи работы

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

Были посталены следующие **задачи**:

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.

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.

Научная новизна

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 (D. D. Penzar et al. 2019).

В работе впервые в большом масштабе успешно применены методы классического машинного обучения для предсказания аллель-специфичного связывания факторов транскрипции (Abramov et al. 2021).

Разработан новый вычислительный метод на основе глубокого обучения специально оптимизированный для результатов высокопроизводительных МПРЭ (D. Penzar et al. 2023, Rafi et al. 2024). 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 (D. Penzar et al. 2023).

Разработанный метод успешно применен для работы с данными МПРЭ полученными для человека и предсказания событий аллель-специфичного связывания (Agarwal et al. 2025)

Теоретическая и практическая значимость

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.

Положения, выносимые на защиту

1. 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.
2. 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.
3. 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.
4. Архитектура LegNet при помощи подхода “холодная диффузия” успешно адаптирована для генерации регуляторных последовательностей с заданной экспрессией.

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.

Степень достоверности данных

Все экспериментальные данные, использовавшиеся в работе, находятся в открытом доступе и результаты их анализа могут быть воспроизведены. Код обучения и тестирования моделей машинного обучения выложены в открытый доступ. Результаты, представленные в работе переносимы между репликами одного эксперимента и независимыми экспериментами схожей природы. Обзор литературы и обсуждение подготовлены с использованием актуальной литературы.

Апробация результатов исследования

Результаты работы были представлены на 6 международных конференциях: RSG-DREAM (Лас-Вегас, США, 2022), Life of Genomes (Казань, Россия, 2022), AIPPA (Алматы, Казань, 2023), MCCMB-2023 (Москва, Россия, 2023), BGRS (Новосибирск, Россия, 2024), APBJC-2024 (Окинава, Япония, 2024)

Публикации по теме диссертации

По результатам исследования опубликовано 6 печатных работ, в том числе 6 статей в рецензируемых научных журналах, индексируемых в WoS и Scopus.

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.

Материалы и методы исследования

В данной работе использовались:

1) Данные массового параллельного репортерного эксперимента (МПРЭ) с мутагенезом насыщающей полимеразной цепной реакцией (ПЦР) из международного конкурса CAGI (Critical Assessment of Genome Interpretation) и из работ (Patwardhan et al. 2012; Kircher et al. 2019) для группы регуляторных элементов человека.
2) Данные МПРЭ об активности 6.7млн 80-нт промоторов в дрожжах *Saccharomyces cerevisiae*, культивировавшихся в виноградном соке сорта Шардоне, из конкурса “DREAM-2022: Predicting gene expression using millions of random promoter sequences”.
3) Данные МПРЭ об активности регуляторных последовательностей в двух средах из статьи (Vaishnav et al. 2022): YPD (стандартная полная среда, содержащая дрожжевой экстракт, пептон и декстрозу, 30 млн последовательностей) и SD-Ura (среда без урацила, 20 млн последовательностей), полученные по протоколу, аналогичному использовавшемуся в соревновании DREAM-2022.
4) Данные МПРЭ по измерению влияния некодирующих последовательностей длины 230 п.н. на транскрипцию репортерного гена в трех клеточных линиях HepG2 (клеточная линия гепатоцеллюлярной карциномы человека), K562 (клеточной линией иммортализованного миелогенного лейкоза) и WTC11 (индуцированные плюрипотентные стволовые клетки) из работы (Agarwal et al. 2023).

В качестве дополнительных признаков для обучение моделей машинного обучения использовались:

1) Признаки из нейронной сети DeepSEA, полученные по процедуре, описанной авторами (Zhou and Troyanskaya 2015).
2) Признаки из нейронной сети Enformer(Avsec et al. 2021), полученные согласно процедуре, описанной в (Agarwal et al. 2023).

3) Признаки, характеризующие связывание транскрипционных факторов (ТФ) с последовательностью, полученные при помощи базы мотивов ТФ HOCOMOCO [307] и программы PERFECTOS-APE(Vorontsov, Kulakovskiy, and Khimulya 2015).
4) Признаки, характеризующие аллель специфичную доступность хроматина, полученные в работе[308].

Результаты и их обсуждение

Утечка данных при обучении моделей по данным параллельных репортерных экспериментов с насыщающей ПЦР

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 нарушения экспрессии соответствующего гена (за исключением промотора гена теломеразы TERT, который был протестирован в двух клеточных линиях).

Модели команд, разделивших первое место в конкурсе (Shigaki et al. 2019) использовали признаки на основе нейронной сети DeepSEA (Zhou and Troyanskaya 2015), 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. 1 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 информаци, но по остальным свойствам напоминающие биологически релевантные.

Для этого помимо обычных признаков, характеризующих эффект каждой мутации, полученных согласно процедуре, предложенной авторами DeepSEA (Zhou and Troyanskaya 2015) был создан аналогичный набор признаков, 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.

Мы обучили случайный лес (Breiman 2001) 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. В результате мы получили, что модель, обученная на полностью нерелевантных признаках, не только показывает качество лучше случайного, но выступает наравне или на уровне с многими решениями конкурса (Shigaki et al. 2019) (**Рис. 1**)

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.

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 (**Таблица 1**), 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 модель, обученных или дообученных при помощи данных о насыщающем мутагенезе, оказывается под большим вопросом.

Таблица 1. 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

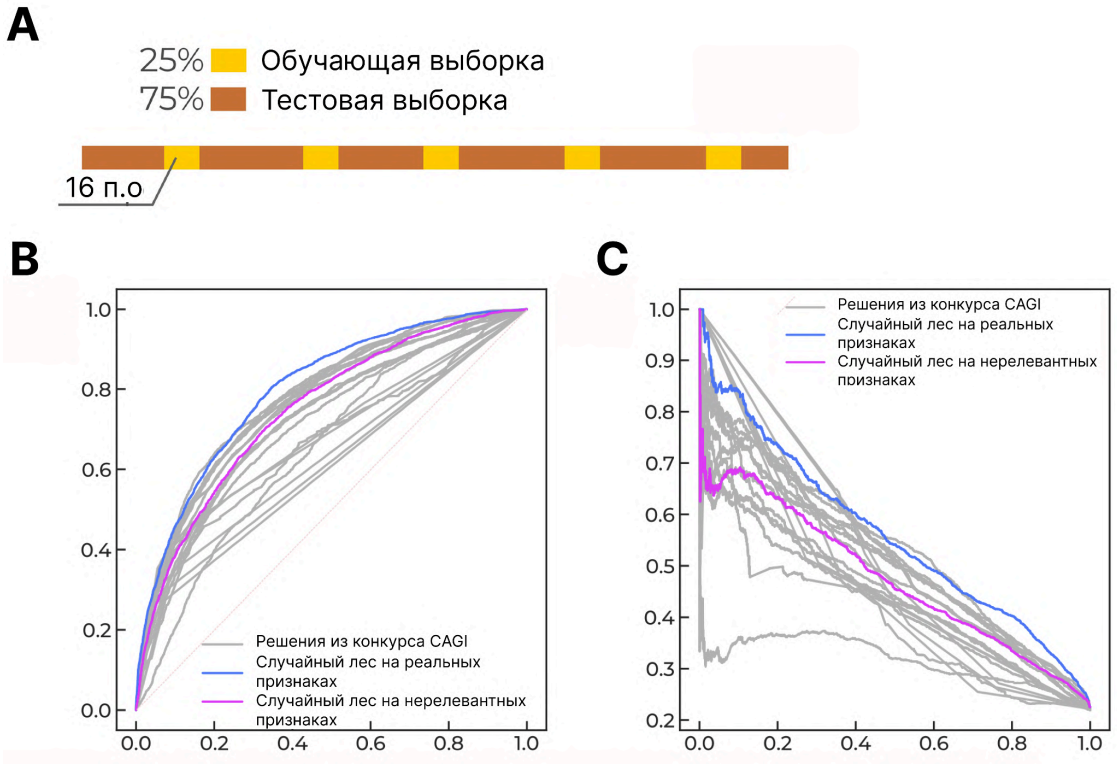

Рисунок. 1. A. Схема разбиения данных CAGI5, предложенное авторами конкурса. Для каждого репортера, обучающая выборка SNV (в сумме составляющая 25%) состоит из множества блоков длины 16 распределенных по координатам репортера. **B-C.** Синим показана модель, обученная на настоящих признаках DeepSEA, 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.

Предсказания событий аллель-специфичного связывания

Для того чтобы определить, насколько возможно обучение модели с использованием признаков DeepSEA на данных аллель-специфичного связывания (Abramov et al. 2021), мы обучили модели

случайного леса с использованием признаков DeepSEA и признаков, основанных на мотивах связывания транскрипционных факторов (Ambrosini et al. 2020).

Задача предсказания аллель-специфичного связывания может быть формализована двумя путями: 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 оценивалась качество модели на отложенной хромосоме, и вычислялись средние значения ROC-AUC и PR-AUC.

Для первой подзадачи (**Рис. 2 A-B**) 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. Для второй подзадачи (**Рис. 2 C-D**) для каждого ТФ и каждого типа клеток была обучена отдельная модель. 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. 2 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 (Maurano et al. 2015; Shi et al. 2016). В случае предсказания сайтов аллель-специфичного связывания отдельных ТФ, признаки на основе мотивов также оказались полезными для моделей (**Рис. 2 F**).

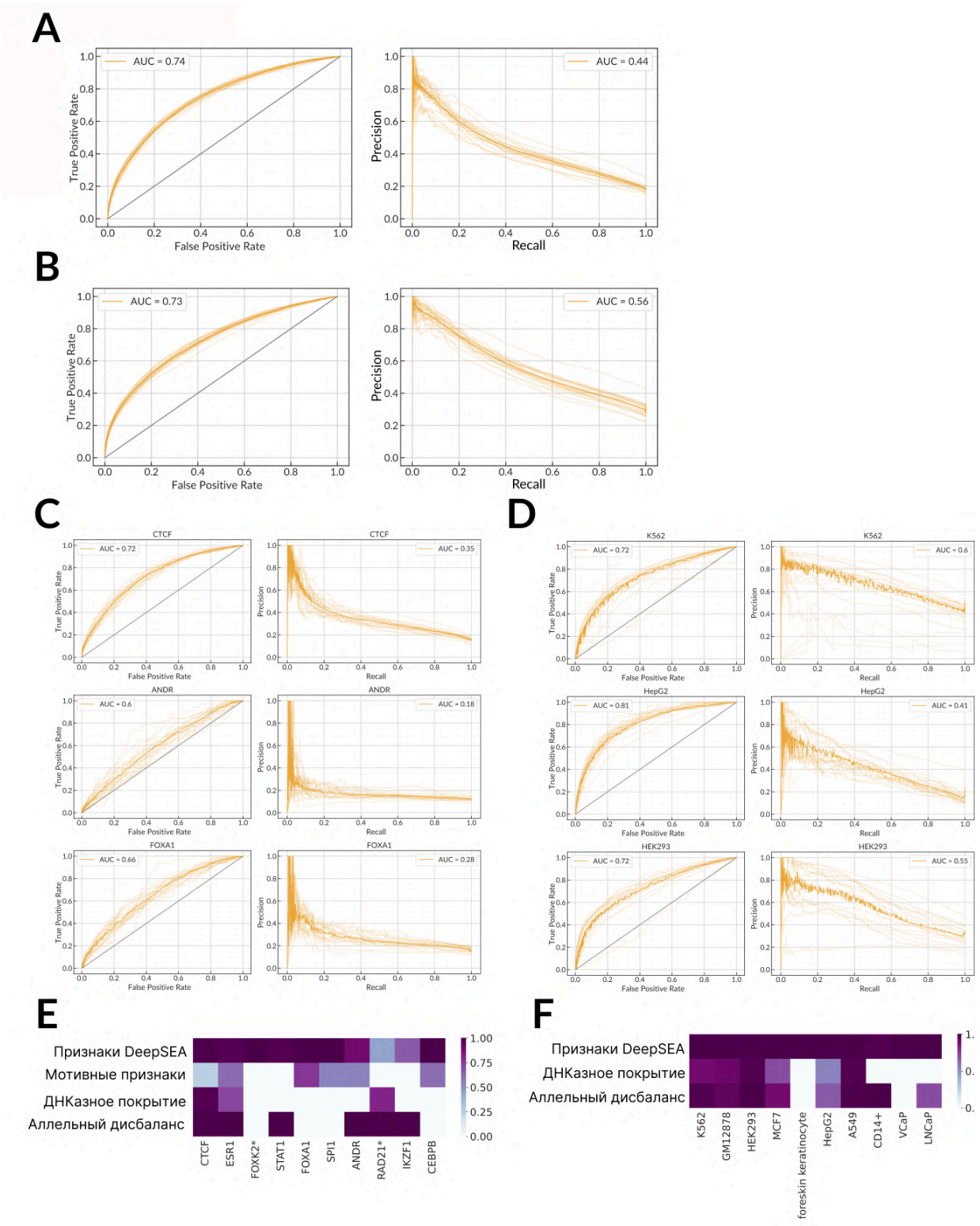

Рисунок 2. A-B. ROC-кривые и PR-кривые для задачи предсказания события аллель-специфичного связывания на общем наборе данных из 10 ТФ/10 типах клетках с наибольшим числом аллель-специфичных событий. **C-D**. ROC-кривые и PR-кривые для отдельных классификаторов для 3 транскрипционных факторов и типов 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.

Архитектура LegNet и ее применение к данным DREAM-2022

В ходе соревнования DREAM-2022 нами была разработана модель LegNet (**Рис. 3**) для предсказания активности регуляторных участков в дрожжевых клетках. За основу была взята архитектура EfficientNetV2 (Tan and Le 2021), 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 (G. Huang et al. 2016), улучшает сходимость нейросети и помогает работать с шумными данными (**Рис. 3 А**).

Как было показано нами в работах (D. Penzar et al. 2023; Rafi et al. 2024), существенный вклад в итоговое качество принес не только выбор архитектуры, но и современный выбранный способ обучения – на основе OneCycleLearningRate (“1-Cycle Schedule” 2024) и учет биологической природа данных. А именно, данные были получены в результате проточной цитофлюориметрии и предоставленные участникам активности последовательностей представляли собой усредненные номера групп светимости, куда попадала каждая последовательность. Поэтому 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 то, что в работе (de Boer et al. 2020) показано, что наблюдаемое значение экспрессии является реализацией нормально распределенной величины. По этой причине каждому значению экспрессии мы ставим в соответствие нормальное распределение:

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

где e - значение измеренной экспрессии.

Во время обучения нейронная сеть училась предсказывать вероятности попадания последовательности в каждый бин, посчитанные на основании этого распределения. Во время тестирования мы умножали предсказанные вероятности на номер бина и суммировали, получая итоговое предсказание.

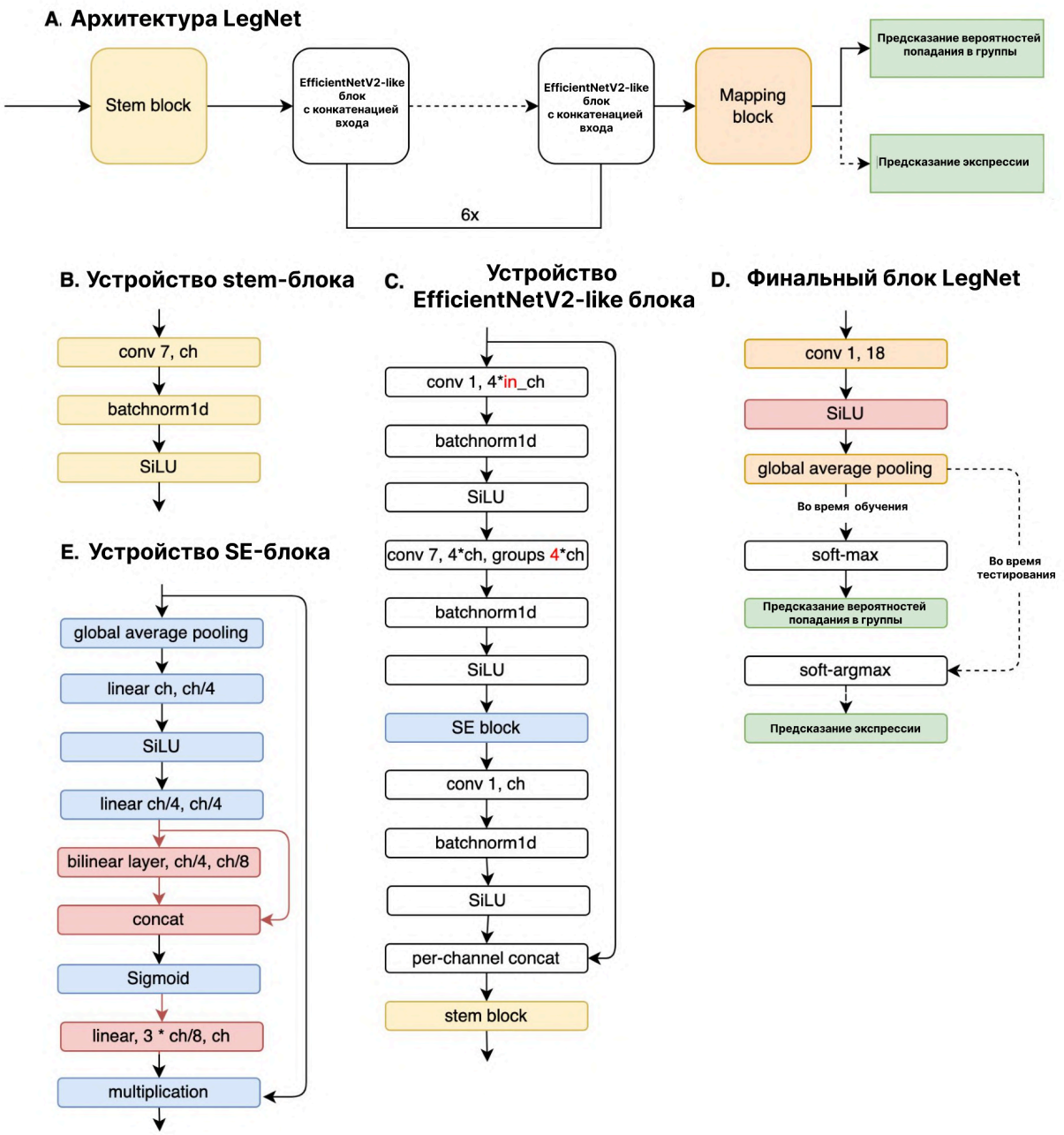

Рисунок 3. Схема архитектуры LegNet. **A** - обзор архитектуры. **B** - структура блока Stem. **C** - сверточный блок адаптированный из EfficientNetV2. **D** - последний слой нейросети. Пунктирные линии соответствуют сценариям при валидации и тесте. **E** - структура SE-блока. Красные блоки соответствуют блокам, которые были удалены в ходе постконкурсного анализа решения. Global average pooling – поканальное усреднение. mulitplication – поточечное произведение.

Представленное нами решение заняло первое место во всех номинациях, устойчиво превосходя решения всех других участников (**Рис. 4**). Что особенно важно, наша модель превзошла остальные решения в задаче предсказания эффектов однонуклеотидных мутаций.

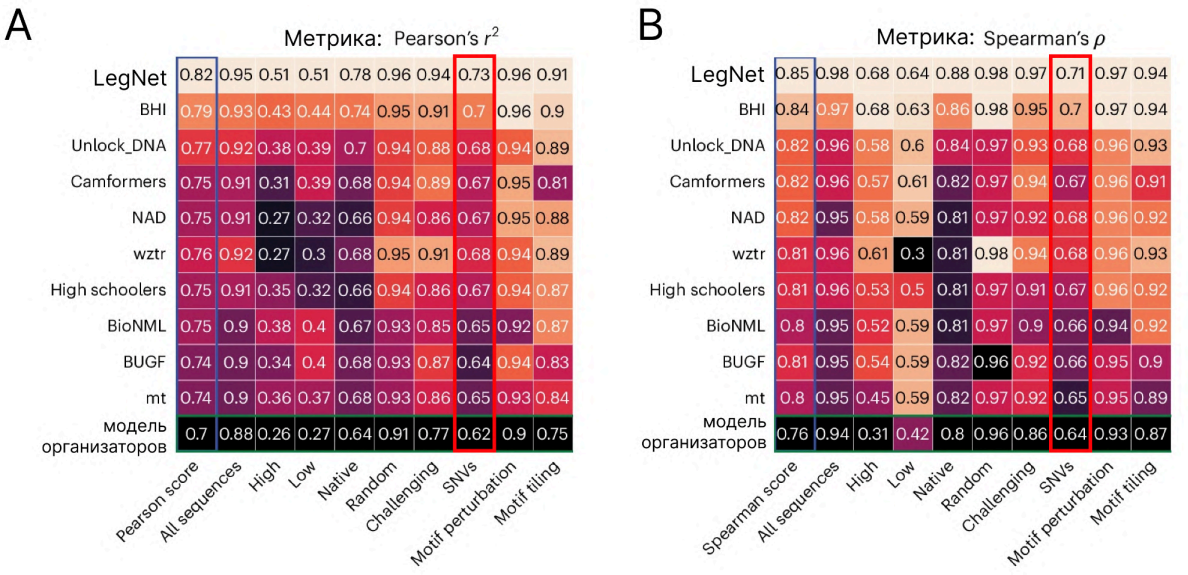

Рисунок 4. C-D. Модель LegNet (autosome.org) значимо превосходит все остальные решения в обоих конкурсных метриках. Во всех подкатегориях задач наша модель также превосходит остальные решения, что особенно важно, в задаче предсказания эффектов однонуклеотидных мутаций (SNV, красная рамка).

Далее мы провели тренировку LegNet на опубликованных данных экспрессии репортерного гена в дрожжевых клетках для промотеров с известной последовательностью в среде YPD (комплексная среда: дрожжевой экстракт, пептон и глюкоза) и среде SD-Ura (синтетическая среда, не содержащая урацил). 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. 5**). Во всех сценариях LegNet продемонстрировал значительное улучшение метрик предсказания по сравнению с моделью, предложенной авторами.

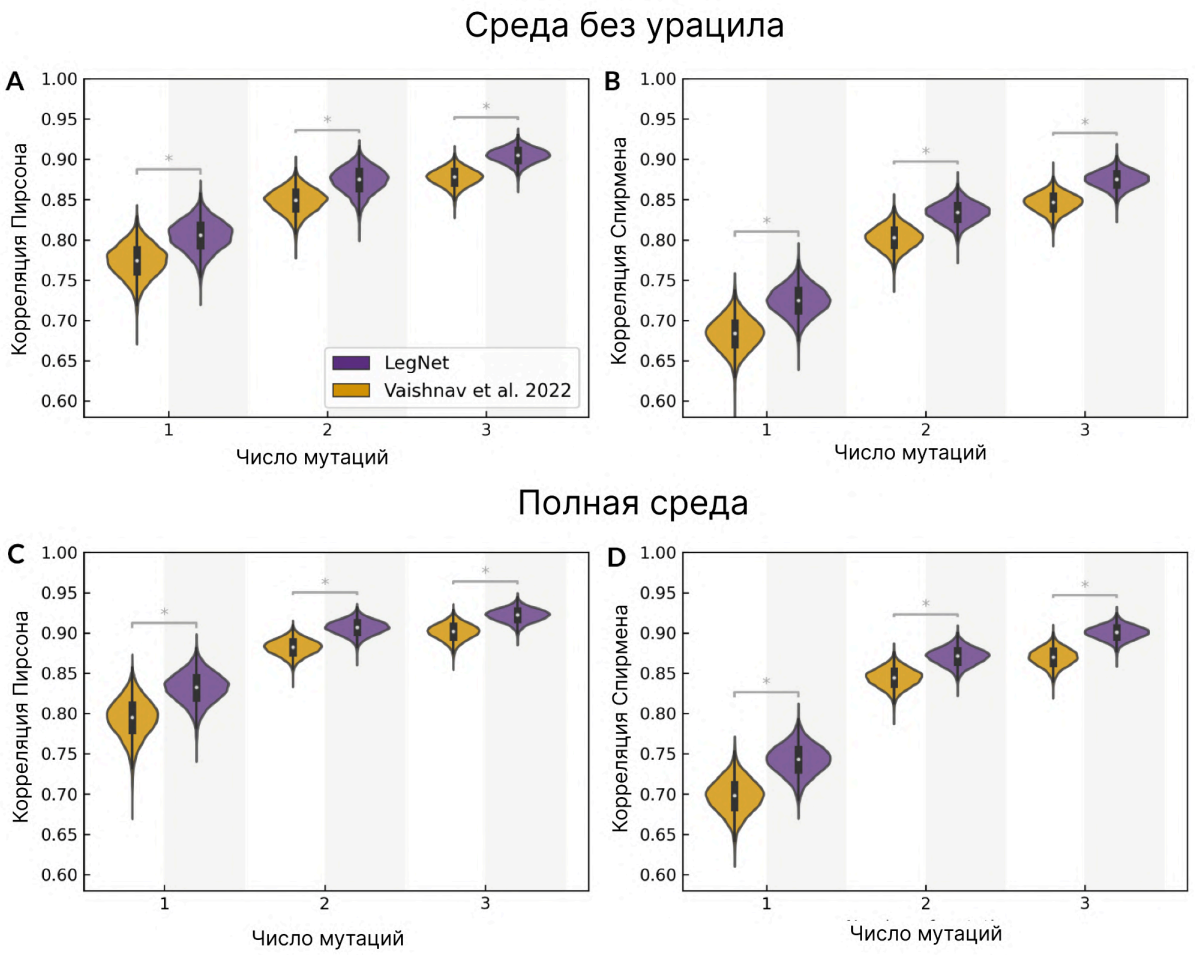

Рисунок 5. Предсказание влияния замен моделью LegNet по последовательности промотора на открытых данных для дрожжей, выращенных в среде SD-Ura - Defined (**A-B**) и YPD - Complex (**C-B**). Графики показывают распределение корреляции Пирсона (**A, C**) и корреляции Спирмена (**B, D**) между предсказанными и истинными значениями. Распределения метрик получены процедурой бутстрепа с n=10,000. *p < 0.001 из теста зависимых корреляций Сильвера [32].

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 могут 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 решение 2го и 3го места (**Рис. 6 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. 6 B**). Это лишний раз подчеркивает важность оптимизации решений моделей машинного обучения и учета биологической составляющей решаемой проблемы.

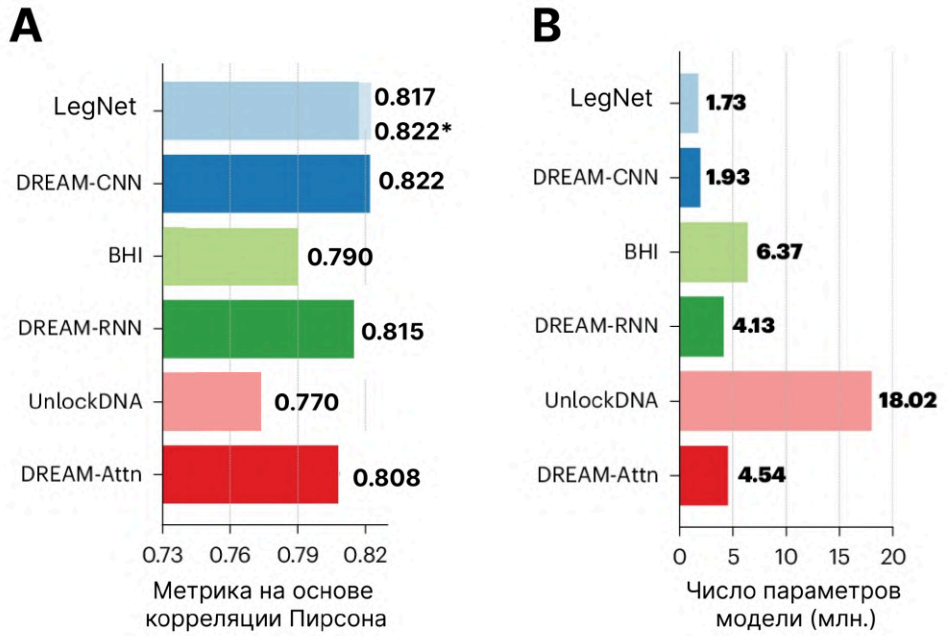

Рисунок 6. Постконкурсная оптимизация моделей участников конкурса DREAM-2022 при помощи пакета PrixFixe. Модель участников и результат их оптимизации при помощи пакета расположены парами (LegNet – DREAM-CNN, BHI – DREAM-RNN, UnlockDNA – DREAM-Attn) **A.** Качество моделей на задаче конкурса. Указано 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.

Генерация промоторных последовательностей с заданной активностью

Для того чтобы адаптировать нашу модель к задаче генерации регуляторных последовательностей в качестве основы мы решили использовать подход, основанный на методе **холодной диффузии** (Bansal et al. 2022). В качестве источника шума мы использовали внесение одиночной замены (**рис. 7 A**) в случайное место последовательности . 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.

Мы изменили исходную архитектуру таким образом, чтобы она принимала на вход зашумленную последовательность, число шагов зашумления и экспрессию исходной последовательности, а на выходе предсказывала исходную последовательность.

Генерация последовательностей с заданной активностью при помощи диффузионной модели осуществлялась по следующей схеме (**рис. 7 B**):

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.

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.

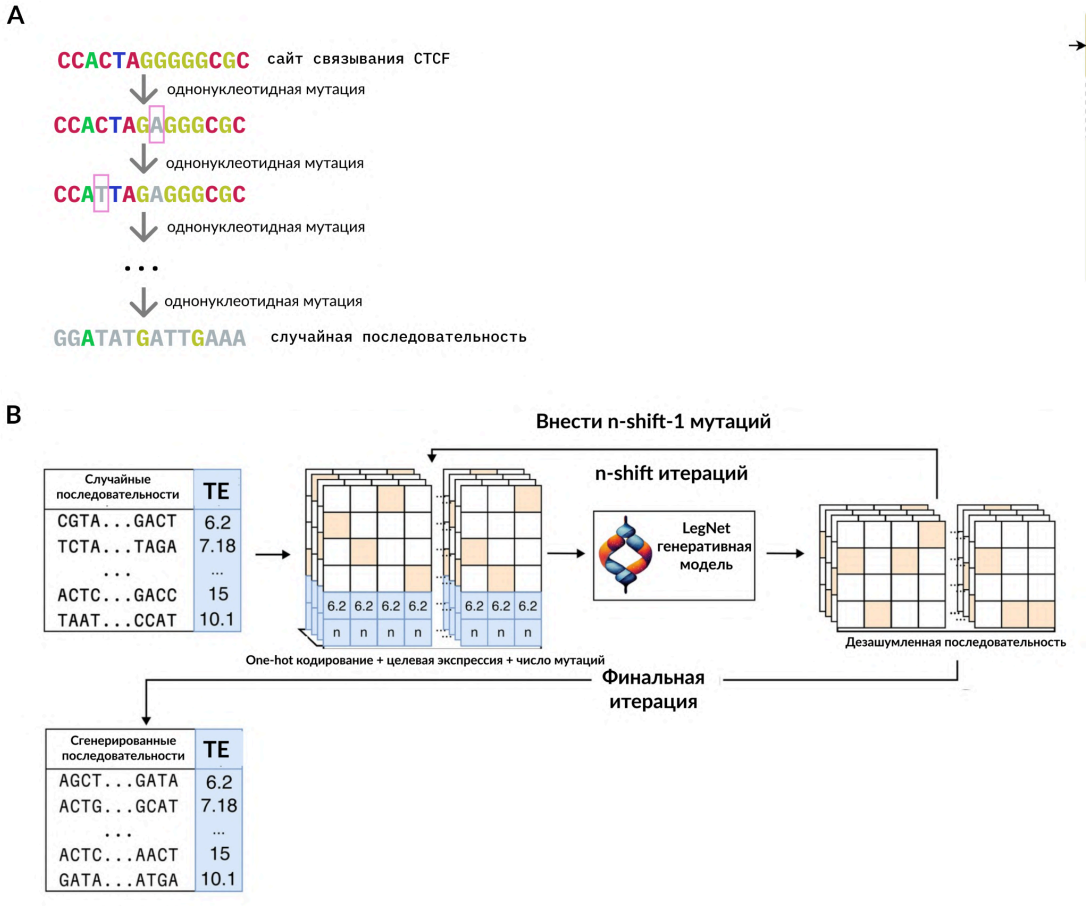

Рисунок 7. A. Предложенный нами зашумляющий процесс для использования в холодной диффузии. **B.** Схема генерации последовательностей с заданной экспрессией при помощи холодной диффузионной модели. TE – целевая экспрессия.

Предсказание активности регуляторных элементов человека

Для работы с регуляторными участками человека в изначальную архитектуру нейронной сети LegNet были внесены изменения 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. Финальная модифицированная архитектура для данной задачи получила название MPRALegNet.

На независимых данных из статьи (Agarwal et al. 2023) предложенная нами архитектура MPRALegNet показала лучшее качество в двух из трех типов клеток (**рис. 8 A,C**)**.** Анализ, проведенный на подвыборках исходных датасетов показал, что качество модели MPRALegNet может быть улучшено при увеличении выборки (**рис 8. B**), причем качество модели зависит от данных по отрицательному степенному закону, что может соответствовать “закону шкалирования”, экспериментально показанному для многих моделей машинного обучения (Kaplan et al. 2020).

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 (Shrikumar et al. 2018) 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. 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 8. D)**.

Далее для каждой клеточной линии были выбраны топ-10 матриц ТФ, выделенных TF-MoDISCO (Shrikumar et al. 2018). Каждая регуляторная последовательность промотора и энхансера из независимой библиотеки была проверена при помощи FIMO (Grant, Bailey, and Noble 2011) 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. При этом модель MPRALegNet смогла корректно воспроизвести зависимость и для сложных случаев, где наблюдалось насыщение эффекта, оказываемого добавлением нового сайта ТФ, на экспрессию, например, STAT1/4/5A/5B (**рис. 9. А**).

Далее для каждой возможной пары из ТФ, отобранных при помощи TF-MoDISco (Shrikumar et al. 2018) на прошлом этапе, были взяты только последовательности, содержащие не более одного сайта связывания каждого из ТФ пары и не содержащие сайтов связывания других ТФ. На примере этих последовательностей также было показано, что MPRALegNet видит взаимодействие разных ТФ. Например, на **рис. 9, В** приведен пример субмультипликативного взаимодействия ТФ HNF4A/G и ТФ NFYA/C, а на **рис 9, С** – сверхмультпликативного между факторами NRF1 и FOS::JUN.

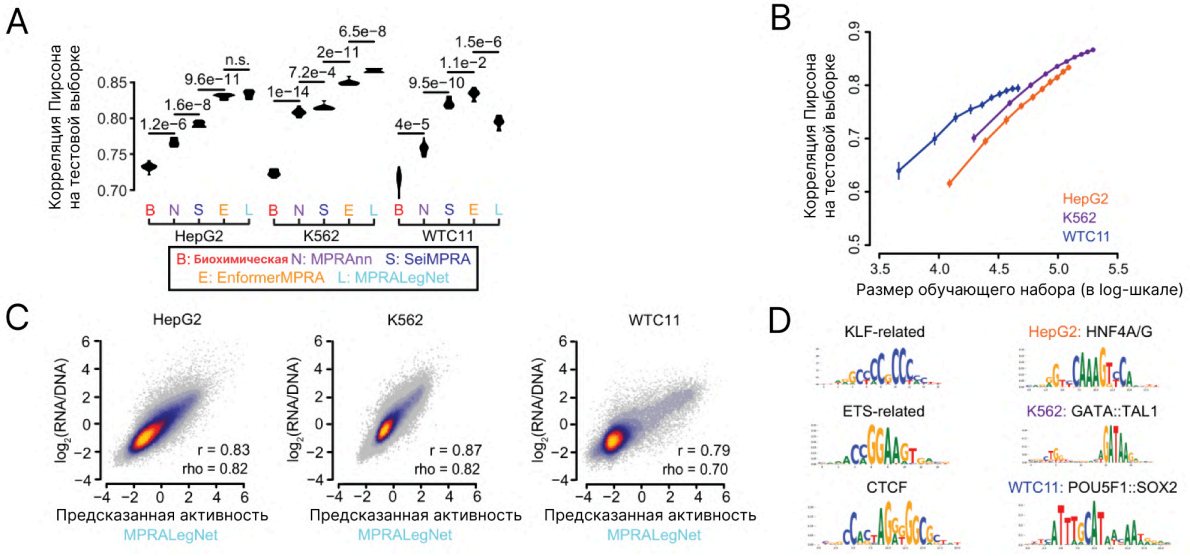

Рисунок 8. А. Скрипичные диаграммы, показывающие качество моделей, основанных на последовательностях, и биохимических моделей на десяти фолдах кросс-валидации, с улучшением по сравнению с предыдущей моделью, оценённым с помощью одностороннего парного t-теста. **В.** Влияние размера обучающего набора 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). **С.** 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.

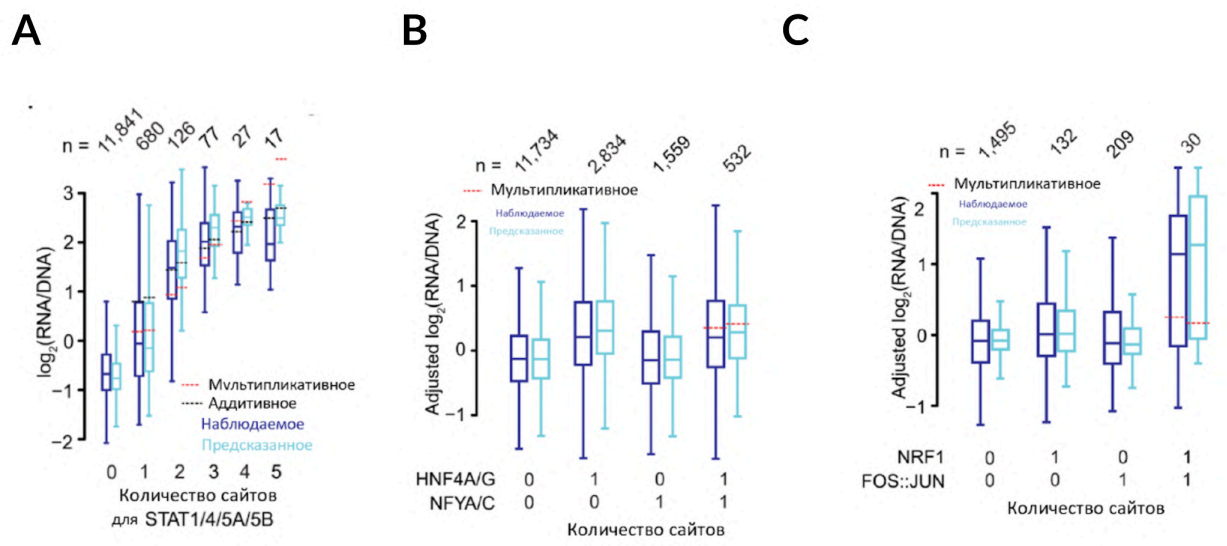

Рисунок 9. A. Зависимость экспериментальной и предсказанной активности регуляторной последовательности в клеточной линии K562 от количества предсказанных сайтов связывания ТФ STAT1/4/5A/5B. **B.** Субмультипликативный характер взаимодействия факторов HNF4A/G и NFYA/C. **C.** Сверхмультипликативный характер взаимодействия факторов NRF1 и FOS::JUN.

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 (Kircher et al. 2019). Данные регуляторные последовательности использовались так как измерение эффектов однонуклеотидных мутаций для них производилось в клеточных типах, на которых производилось обучение модели.

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. 10**).

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, 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.

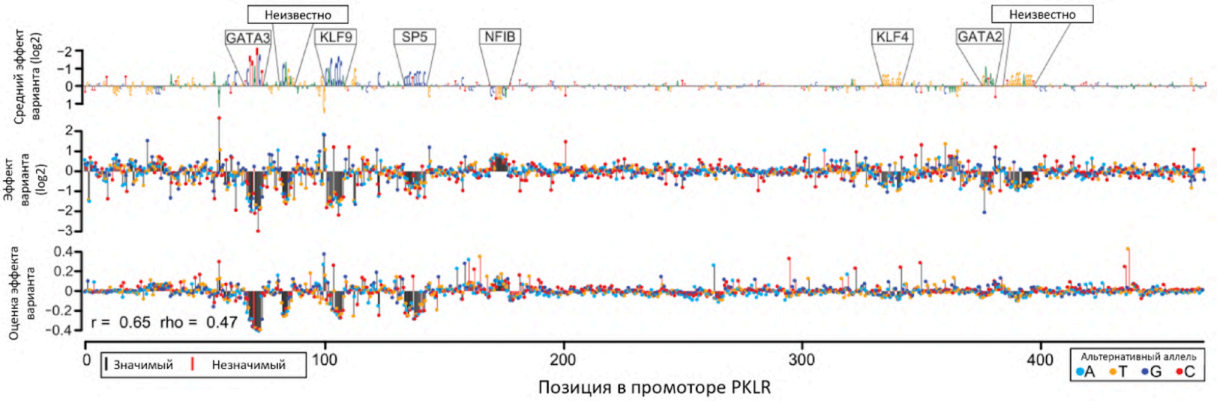

Рисунок 10. Данные насыщенного мутагенеза для промотора PKLR. Верхняя строка представляет собой референсную последовательность, масштабированную по среднему размеру эффекта среди всех альтернативных мутаций, с аннотацией значимых сайтов связывания транскрипционных факторов (TFBS), которые соответствуют 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).

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

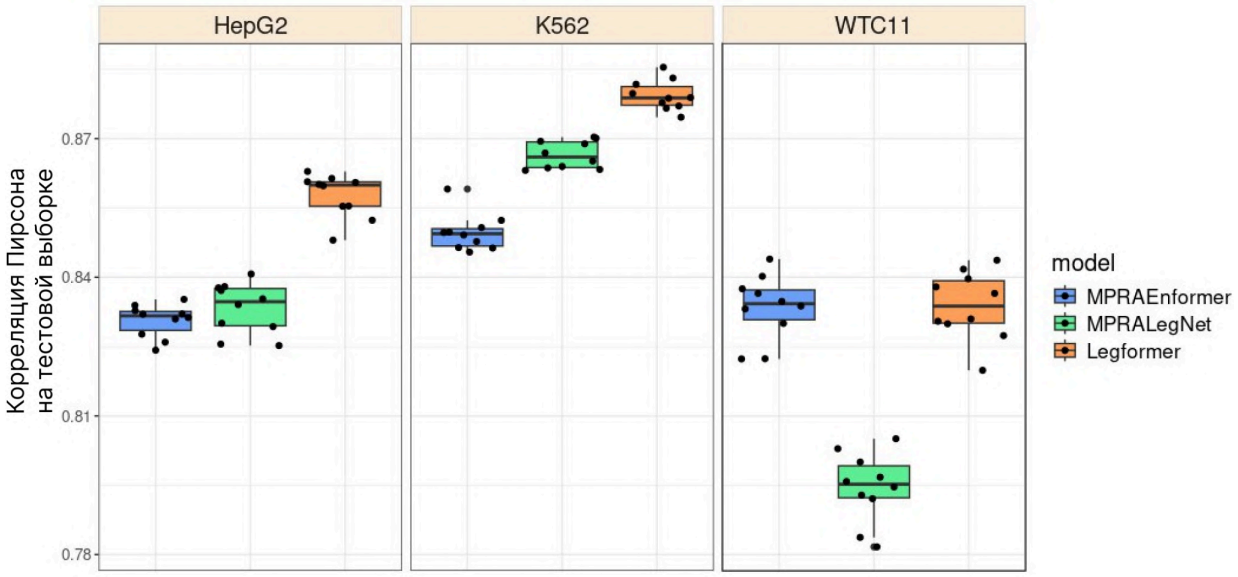

Рисунок 11. Сравнение качества (корреляции Пирсона) моделей MPRAEnformer, MPRALegNet и Legformer на десяти фолдах кросс-валидации на независимых библиотеках регуляторных участков

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.

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[1]

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. // Nature Biotechnology – 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 // Nature Communications – 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 Biology – 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 // Frontiers in Genetics– 2019.– Vol. 10.– P. 1078. doi: 10.3389/fgene.2019.01078. JIF (для WoS) = **2.8,** (0.70/0.40)

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