ОТЗЫВ официального оппонента
на (о) диссертацию(и) на соискание ученой степени
кандидата биологических наук Пензара Дмитрия Дмитриевича
на тему: «Вычислительное предсказание эффектов мутаций в регуляторных районах генов»
по специальности 1.5.8 «Математическая биология, биоинформатика»

Диссертационная работа Д.Д. Пензара посвящена разработке новых вычислительных методов с использованием машинного обучения для предсказания эффектов однонуклеотидных замен в геноме человека на основе больших данных, получаемых при помощи современных высокопроизводительных подходов молекулярной биологии.

Мутации в регуляторных районах человеческого генома часто ассоциированы с предрасположенностью к различным заболеваниям, а вычислительные методы аннотации индивидуальных вариантов генома, в том числе, в участках, не кодирующих белки, являются критически важными для персонифицированной медицины. Значительный прогресс в методах глубокого машинного обучения, в частности, громкий успех AlphaFold2 в задаче предсказания структур белков, вкупе с развитием высокопроизводительных методов секвенирования, позволяет рассчитывать на то, что нейросетевые методы помогут найти решение и для других биоинформатических задач, включая анализ геномных вариантов, находящихся в некодирующих областях генов, что является темой представленной диссертации. В целом, тема исследования является актуальной, а исследование опирается как на достижения в наиболее современных экспериментальных методах молекулярной биологии, так и на успехи в развитии методов машинного обучения при работе с большими данными.

Диссертационная работа построена по традиционной схеме, включает введение, обзор литературы, описание материалов и методов исследования, результаты и их обсуждение, заключение, выводы и список цитируемой литературы, содержащий 334 источника. Работа изложена на 166 страницах,

проиллюстрирована 55 рисунками и 6 таблицами, содержит 1 приложение. Главы диссертации взаимосвязаны и логически последовательны.

Во введении автор убедительно обосновывает актуальность выбранной темы, формулирует цели и задачи, которые отражают логику исследования, реализация задач подробно изложена далее в содержательной части работы. Введение содержит и убедительно представляет необходимые формальные разделы, включая степень научной разработанности темы, теоретическую и практическую значимость, методологию исследования. В работе исчерпывающе охарактеризован личный вклад автора, перечислены положения, выносимые на защиту, и публикации по теме диссертации.

Обзор литературы достаточно, а иногда и избыточно подробно излагает состояние области исследований: рассмотрены и биологическая, и биоинформатическая стороны вопроса, методы машинного обучения, использующиеся в предметной области для предсказания активности регуляторных районов генома и эффектов регуляторных замен. Из 334 литературных ссылок, упомянутых в работе, более 30 ведут на статьи и препринты, опубликованные в 2024 году, что свидетельствует об отличном знании диссертантом современной ситуации в области проводимых исследований. Хочется выделить описание авторов проблем и сложностей использования современных подходов на основе нейронных сетей для решения задач в предметной области.

К обзору литературы есть несколько замечаний. Во-первых, достаточно скупо описаны особенности архитектуры моделей (Basenji и Basenji2 на стр. 56-57). Кроме того, сравнительно мало внимания уделено подробностям работы механизма внимания, дано лишь общее абстрактное описание, излишнее для специалистов в области, и недостаточное для неспециалистов. В то же время, на мой взгляд, избыточно подробно описаны методы интерпретации предсказаний моделей, тогда как в самой работе они используются в крайне ограниченной мере.

Раздел «Материалы и методы» написан достаточно подробно, хотя в некоторых случаях приходится обращаться к разделу результатов для того, чтобы полностью понять предлагаемые автором процедуры и их взаимосвязь. В то же время этот раздел показывает, что для решения поставленных задач автор скрупулёзно разбирался в нюансах применяемых их методов, что и позволило его разработкам превзойти качество имевшихся ранее решений, таких как DeepSEA и Enformer, используя значительно меньшее количество параметров. Кроме того, автор при помощи абляционных исследований выявил ключевые компоненты модели и предложил ее улучшенную версию.

Описание результатов, данное в последующей главе, в целом производит благоприятное впечатление, как по значимости полученных данных, так и по форме их изложения.

В первом подразделе результатов диссертант убедительно демонстрирует сложность проблемы использования данных массовых параллельных репортерных экспериментов с мутагенезом насыщающей ПЦР – обучение моделей машинного обучения на таких данных напрямую невозможно в связи с неизбежной утечкой информации между тренировочной и тестовой выборками.

Во втором подразделе автор показывает, что использование классических методов машинного обучения ("случайный лес") позволяет достигнуть приемлемого качества предсказания событий аллель-специфичного связывания. **В данной части работы хотелось бы видеть выводы автора о возможности дальнейшего применения и развития разработанной модели.**

В третьем подразделе автор описывает разработанный им нейросетевой метод LegNet, занявший первое место в международном конкурсе DREAM-2022. Данный метод использует набор подходов, которые находили ранее свое отражение в машинном обучении, однако не применялись в регуляторной геномике. В частности, используются заимствованные из архитектуры для классификации изображений EfficientNetV2 блоки сжатия-активации для

дополнительной нормализации каналов сверточных блоков, а также функция активации SiLU. Для обучения сети используется специальное расписание шага обучения – One-Cycle Learning Rate. Помимо этого, автор не забывает и о конкретике решаемой проблемы и специальным образом модифицирует оптимизируемую метрику для того, чтобы она лучше отражала эксперимент. Диссертант демонстрирует, что предложенный им подход существенно превосходит другие решения как на данных конкурса в целом (для оценки так называемой "активности" последовательности), так и в конкретной задаче предсказания эффектов однонуклеотидных мутаций. Помимо этого демонстрируется, что разработанная модель является экономной по числу параметров. К недостаткам раздела можно отнести то, что автор не обсуждает, были ли апробированы другие архитектуры и методы в ходе разработки итоговой модели. Например, хотя среди моделей конкурса и встречается модель на основе механизма внимания, остается вопрос – **улучшает ли добавление блоков на основе данного механизма поведения предложенного автором решения? Аналогично, проверялось ли качество моделей на основе длинных конволюций (HyenaDNA) и моделей на основе пространств состояний (Mamba)?**

В четвертом подразделе рассказывается об адаптации архитектуры LegNet к задаче генерации последовательностей с заданной активностью при помощи метода "холодной диффузии". Это является первым примером применения методов на основе диффузии к задаче генерации участков ДНК, управляющих активностью генов. Возникает искушение использовать модель на практике для генерации новых регуляторных участков ДНК, но генеративная модель провалидирована только вычислительно, и предсказать ее поведение на реальных задачах невозможно. **Предпринимались ли попытки проверить подход экспериментально?**

В последнем подразделе результатов автор описывает адаптацию модели LegNet к данным об активности регуляторных последовательностей генома человека. Показано, что модель автора превосходит по качеству

имеющиеся решения, включая большие полногеномные модели, предобученные больших экспериментальных данных. Продемонстрировано, что авторская модель выучивает биологически важные признаки такие как наличия участков ДНК-белкового взаимодействия, а также умеет предсказывать сложные комбинаторные взаимодействия различных белков с ДНК. Помимо этого, демонстрируется хорошая корреляция предсказанных моделью результатов однонуклеотидных замен в регуляторных элементах с экспериментальными данными. В работе демонстрируется, что качество модели зависит от числа объектов в обучающей выборке согласно общепринятому в области машинного обучения закону масштабирования. Наконец, автор показывает, что добавление в LegNet предсказаний популярной предобученной модели позволяет дополнительно улучшить результат, особенно на маленьких наборах данных, что имеет практическую важность в связи с трудоемкостью и стоимостью измерения активности большого числа регуляторных элементов в одном эксперименте.

Вместе с тем возникает несколько вопросов по поводу данной части работы:

1) **Не вполне понятно, почему экспериментальные признаки, использовавшиеся для линейной модели, названы в работе «биохимическими»;**
2) **В работе говорится о том, что модель была уменьшена (стр. 132), однако не указывается точное число параметров модели и сравнение этого числа с размером исходной модели;**
3) **«Биохимическая» модель представляет собой линейную регрессию. Не поможет ли использование более сложных моделей с теми же признаками улучшить качество, в какой-то мере догнав качество моделей глубокого обучения?**

В целом же результаты понятны, хорошо проиллюстрированы и адекватно обсуждены.

Высказанные в отзыве замечания носят рекомендательный характер и не умаляют ценности интересного и важного исследования, проведенного Пензаром Д.Д.

Достоверность и значимость полученных диссертантом результатов обусловлена грамотным использованием современных методов машинного обучения, также адекватной и взвешенной оценкой полученных результатов, и, судя по тексту работы, хорошим знакомством соискателя с состоянием исследований в предметной области.

Выводы диссертационной работы Пензара Д.Д. обоснованы, хорошо соответствуют поставленным задачам и подтверждены полученными результатами. В целом, результаты диссертационной работы характеризуются высокой степенью новизны и имеют существенную теоретическую и научно-практическую значимость.

Необходимо отметить публикации Пензара Д.Д – 6 статей опубликовано в международных журналах первого квартиля, включая известные в области журналы Nature Publishing Group.

Содержание диссертации Пензара Д.Д соответствует специальности 1.5.8 «Математическая биология, биоинформатика» (по биологическим наукам), а именно следующим ее направлениям «Компьютерная системная биология», «Разработка и применение новых вычислительных алгоритмов для анализа экспериментальных данных в биологии и медицине», «Разработка и применение методов машинного обучения и искусственного интеллекта для анализа и прогнозирования свойств биологических объектов на основе анализа больших биомедицинских данных», а также критериям, определенным пп. 2.1-2.5 Положения о присуждении ученых степеней в Московском государственном университете имени М.В. Ломоносова, а также оформлена согласно требованиям Положения о совете по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук Московского государственного университета имени М.В. Ломоносова.

Таким образом, соискатель Пензар Дмитрий Дмитриевич заслуживает присуждения ученой степени кандидата биологических наук по специальности 1.5.8 «Математическая биология, биоинформатика».

Официальный оппонент:

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»

Уткин Лев Владимирович *подпись*

20.02.2025

Контактные данные:

тел.: +7(921) 344-63-90, e-mail: utkin_lv@spbstu.ru
Специальность, по которой официальным оппонентом защищена диссертация:
05.13.18 – Математическое моделирование, численные методы и комплексы программ

Адрес места работы:

195251, Санкт-Петербург, Политехническая, ул., д. 29
Санкт-Петербургский политехнический университет Петра Великого (СПбПУ), Институт компьютерных наук и кибербезопасности

Тел.: 8(800) 707-18-99; e-mail: office@spbstu.ru