ОТЗЫВ официального оппонента
на диссертацию 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 представляет оригинальное фундаментальное исследование по разработке и применению современных методов машинного обучения в биоинформатике для предсказания эффектов мутаций в нуклеотидных последовательностях. Тема работы по использование методов искусственного интеллекта, в частности ансамблей деревьев решений и моделей глубокого обучения для вычислительного предсказания активности регуляторных областей генов **актуальна**, представляет современное, быстро развивающееся направление развития биоинформатики в мире.

Содержание диссертации Д.Д.Пензара соответствует специальности 1.5.8 «Математическая биология, биоинформатика», 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».

Работа Д.Д.Пензара основана на наиболее полных на сегодняшний день экспериментальных данных о влиянии мутаций в ДНК на изменения состояния хроматина и экспрессию генов, использовавшееся в международных соревнованиях по развитию предсказательных вычислительных моделей. Использование уникальных данных массовых параллельных репортерных экспериментов, и современных архитектур

нейронных сетей и вычислительных ресурсов, придает **оригинальность и масштабность** выполненной работе.

Разработка подхода Д.Д.Пензара имеет несомненную **научную новизну**. Автором **впервые** разработан сверточная нейронная сеть LegNet для предсказания активности регуляторных последовательностей и их влияния на экспрессию репортерных генов. **Впервые** на основе метода случайного леса с использованием геномных признаков разработана эффективная модель предсказания аллель-специфичного связывания факторов транскрипции.

Практическая значимость работы Д.Д. Пензара определяется тем, что предложенный метод и программный инструмент может быть использован для дальнейших разработок и улучшения качества решения в задачах предсказания активности регуляторных регионов генов. Код находится в открытом доступе.

Продолжающееся быстрое развитие высокопроизводительных постгеномных технологий, дает возможность ставить фундаментально новые задачи по анализу знаний о молекулярных механизмах регуляции экспрессии генов. Помимо необходимости обработки больших объемов данных анализ влияния мутаций в некодирующих участках на транскрипцию генов затруднен сложностью, неоднозначностью и многоуровневостью регуляторного кода.

Дмитрий Дмитриевич Пензар представляет новые методы на основе классического машинного обучения и искусственных нейронных сетей в биоинформатике - перспективную и развивающуюся тему современных постгеномных исследований. Д.Д. Пензар является соавтором работ, представивших лучшие на сегодняшний день вычислительные методы, получившие первые места в международных соревнованиях вычислительных моделей. Отметим, что такие протоколы обработки данных появились относительно недавно, бурно развиваясь в последние несколько лет.

Диссертационная работа Д.Д. Пензара построена по классическому признаку, включает оглавление, список сокращений и условных обозначений, разделы Введение, Обзор литературы, Материалы и методы, Результаты,

Заключение и выводы, список литературы и Приложение. Работа изложена на 167 страницах, иллюстрирована 55 рисунками и 6 таблицами. Список литературы состоит из 334 источников.

В литературном обзоре автор подробно описывает вычислительные методы для классификации нуклеотидных последовательностей с помощью нейронных сетей, методы машинного обучения.

В работе Д.Д. Пензара использовано большое количество экспериментальных данных, в том числе использовавшихся в международных соревнованиях, данные массовых параллельных репортерных экспериментов с мутагенезом, источники и характеристики которых подробно описаны в разделе «Материалы и методы», что определяет **масштабность** диссертационной работы. В качестве результатов автор представляет собственный вычислительный подход к анализу эффектов мутаций, доказывает его эффективность в сравнении с существующими программами (опубликованный результаты международных соревнований).

Научные положения в диссертационной работе Д.Д.Пензара **обоснованы**, опираются на экспериментальные данные. **Выводы - достоверны и обоснованы**. Рекомендации по построению признаков нуклетидных последовательностей для машинного обучения имеют практическое значение. Предложенный метод генерации последовательностей с заданной экспрессией может быть использован для рационального дизайна генноинженерных конструкций.

Результаты исследования Д.Д.Пензара всесторонне **опубликованы** в рецензируемых научных изданиях, в журналах с высоким импакт фактором, в том числе Nature, Nature Communications. Результаты обнародованы, представлены на серии всероссийских и международных конференций, метод и компьютерная программа участвовали в международных соревнованиях DREAM-2022, код программы представлен в открытом доступе.

Из положительных сторон исследования стоит отметить детальность

исследования, новизну используемых биоинформационных методов, фундаментальность подхода к оценке зависимостей в нуклеотидных последовательностях, ограничивающих применение методов машинного обучения.

К работе есть **ряд технических замечаний**. В тексте присутствуют некоторые англицизмы и жаргонизмы.

К недостаткам оформления диссертации следует отнести перегруженность обзорной части работы, составляющей до половины объема текста. В то же время детальный обзор современных методов машинного обучения и существующих решений применительно к анализу нуклеотидных последовательностей имеет самостоятельную ценность – такой обзор очень востребован для биоинформатики, образования, выбора приоритетных направлений разработки собственных методов и вычислительных моделей.

Недостатком является и избыточное цитирование (три и более публикаций вместе, без деталей - например, [9–13], [14–20] на стр.8, [26–31] на стр.9 и стр.11). Список литературы и так перегружен (более 300 наименований). Стоит или убрать лишние ссылки, ли добавить детали, перефразировать, пояснить чем отличаются цитируемые работы.

Текст включает англицизмы, никак не связанные с терминологией, например «инкрементальные темпы» - можно написать «возрастающие», или «ускоряющиеся темпы». Такие слова как «продвинутый» вводят в заблуждение, можно написать «усовершенствованный». Не нужно писать повторяющиеся по смыслу слова - «экспериментальная верификация» - достаточно «экспериментальная проверка», или просто «верификация».

Термин «утечка данных» (утечка информации) стоит прокомментировать с самого начала, как термин именно машинного обучения. Под утечкой данных понимают ситуацию, когда один или несколько входных признаков, использующихся в процессе обучения модели, оказываются недоступными при ее практическом применении. Результатом утечки данных является ухудшение точности модели относительно ее оценки, полученной на

тестовых данных, о чем и идет речь в работе. Без этого можно считать, что была утечка (скрытие данных) использовавшихся в конкурсе вычислительных моделей нуклеотидных последовательностей, а речь идет о признаках.

Оглавление диссертации избыточно детально, представлено четыре уровня нумерации - 3.4.4.1. Многие разделы занимают меньше страницы. Стоило бы объединить их в более крупные разделы, уменьшить нумерации уровней для ясности.

В разделе Материалы и методы идет описание собственной программы LegNet, как одной из участвующих в соревновании, без упоминания что это работа автора. То есть, это описание должно относиться к разделу Результаты.

Список работ автора в тексте диссертации смешан с общим списком литературы, сбито форматирование. При этом работы автора 33 и 35 в общем списке литературы должны входить в отдельный список работ автора, по ним даются пояснения, описан личный вклад. Стоило сделать отдельную нумерацию публикаций автора по теме диссертации явным образом, отдельным списком. Отметим, что список корректно представлен в автореферате, но не в самом тексте диссертации.

Вместе с тем, указанные замечания не умаляют значимости диссертационного исследования. Диссертация отвечает требованиям, установленным Московским государственным университетом имени М.В.Ломоносова к работам подобного рода. Содержание диссертации соответствует специальности 1.5.8 «Математическая биология, биоинформатика» (по биологическим наукам), а также критериям, определенным пп. 2.1-2.5 Положения о присуждении ученых степеней в Московском государственном университете имени М.В.Ломоносова, диссертация 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 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)

Орлов Юрий Львович

21.02.2025г.

Контактные данные:

тел.: +7(495) 609-14-00, e-mail: y.orlov@seche
Специальность, по которой официальным оп
защищена диссертация:
03.01.09 –Математическая биология, биоинф

Адрес места работы:

119991, Москва, ГСП-1, ул.Трубецкая, д.8, стр.2
Федеральное государственное автономное образовательное учреждение высшего образования Первого Московского государственного медицинского университета имени И.М. Сеченова Министерства здравоохранения Российской Федерации (Сеченовский Университет), Институт цифрового биодизайна и моделирования живых систем, Кафедра информационных технологий и обработки медицинских данных, Центр цифровой медицины
Тел.: +7(495) 609-14-00; e-mail: rectorat@staff.sechenov.ru