 © Авторы статьи / Authors of the article

МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ, БИОИНФОРМАТИКА/MATHEMATICAL BIOLOGY, BIOINFORMATICS

DOI: https://doi.org/10.60797/jbg.2026.33.3 EDN: FRRWIS

ВЛИЯНИЕ УЧЁТА ЭЛЕКТРИЧЕСКОГО ЗАРЯДА АТОМОВ НА ТОЧНОСТЬ ПРЕДСКАЗАНИЯ ВЗАИМОДЕЙСТВИЯ Т-КЛЕТОЧНЫХ РЕЦЕПТОРОВ С ЭПИТОПАМИ ПРОГРАММОЙ MULTIPASS

Научная статья

Смирнов А.С.[1, *], Рудик А.В.[2], Филимонов Д.А.[3], Лагунин А.А.[4]
[1] ORCID : 0000-0002-7510-1602;
[2] ORCID : 0000-0002-8916-9675;
[3] ORCID : 0000-0002-0339-8478;
[4] ORCID : 0000-0003-1757-8004;
[1, 4] Российский национальный исследовательский медицинский университет имени Н.И. Пирогова, Москва, Российская Федерация
[1] Медико-генетический научный центр имени академика Н.П. Бочкова, Москва, Российская Федерация
[2, 3, 4] Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича, Москва, Российская Федерация

* Корреспондирующий автор (anton.smirnov.9910[at]gmail.com)

Предложена: 17.05.2026; Принята: 21.09.2026; Опубликована: 25.09.2026

Аннотация

Цель: оценить влияние учёта электрического заряда атомов аминокислотных остатков пептидов на точность моделей связи «структура-активность» (SAR, structure-activity relationships) и расширить спектр прогнозируемых активностей разработанного ранее веб-сервиса TCR-Pred для предсказания взаимодействия Т-клеточного рецептора с эпитопом и главным комплексом гистосовместимости (MHC).

В качестве источника данных использовались обновленные базы данных IEDB, CEDAR, VDJdb, McPAS-TCR, PIRD от марта 2026 года. Классификационные модели SAR были созданы с использованием программы MultiPASS 2025. MultiPASS 2025 основан на модифицированном наивном Байесовском классификаторе и на описании структурных формул CDR3 (Complementarity-Determining Region 3) участков в виде дескрипторов многоуровневых атомных окрестностей (MNA, Multilevel Neighbourhoods of Atoms), как заряженных, так и незаряженных атомов.

Учёт заряда практически не повлиял на характеристики лучших моделей. Новые модели SAR для прогноза взаимодействия между Т-клеточными рецепторами и эпитопами/молекулами MHC имеют в сравнении с предыдущей версией более высокую точность (в среднем на 0,06 по величине ROC AUC) и расширенный спектр (в среднем в 2 раза) прогнозируемых активностей.

Ключевые слова: TCR-Pred, SAR, эпитопы, Т-клеточный рецептор, главный комплекс гистосовместимости.

THE INFLUENCE OF TAKING INTO ACCOUNT THE ELECTRIC CHARGE OF ATOMS ON THE ACCURACY OF PREDICTIONS OF T-CELL RECEPTORS' INTERACTIONS WITH EPITOPES USING THE MULTIPASS PROGRAMME

Research article

Smirnov A.S.[1, *], Rudik A.V.[2], Filimonov D.A.[3], Lagunin A.A.[4]
[1] ORCID : 0000-0002-7510-1602;
[2] ORCID : 0000-0002-8916-9675;
[3] ORCID : 0000-0002-0339-8478;
[4] ORCID : 0000-0003-1757-8004;
[1, 4] Pirogov Russian National Research Medical University, Moscow, Russian Federation
[1] Bochkov Medical and Genetic Research Center, Moscow, Russian Federation
[2, 3, 4] Orekhovich Institute of Biomedical Chemistry, Moscow, Russian Federation

* Corresponding author (anton.smirnov.9910[at]gmail.com)

Suggested: 17.05.2026; Accepted: 21.09.2026; Published: 25.09.2026

Abstract

Objective: to evaluate the impact of taking into account the electric charge of the amino acid residues in peptides on the accuracy of ‘structure-activity’ (SAR, structure-activity relationships) and to expand the range of predicted activities of the previously developed TCR-Pred web service for predicting the interaction between the T-cell receptor and the epitope and the major histocompatibility complex (MHC).

The updated IEDB, CEDAR, VDJdb, McPAS-TCR and PIRD databases from March 2026 were used as data sources. SAR classification models were created using the MultiPASS 2025 programme. MultiPASS 2025 is based on a modified naive Bayes classifier and on the description of CDR3 (Complementarity-Determining Region 3) structural formula segments in the form of descriptors of multilevel neighbourhoods of atoms (MNA, Multilevel Neighbourhoods of Atoms), for both charged and uncharged atoms.

 © Авторы статьи / Authors of the article

Taking charge into account had virtually no effect on the performance of the best models. The new SAR models for predicting interactions between T-cell receptors and epitopes/MHC molecules demonstrate higher accuracy (on average 0.06 higher in terms of ROC AUC) and a broader range (on average twice as wide) of predicted activities compared with the previous version.

Keywords: TCR-Pred, SAR, epitopes, T-cell receptor, major histocompatibility complex.

Введение

Т-лимфоциты распознают клетки, содержащие признаки чужеродной генетической информации, с помощью специализированных молекул на своей мембране — Т-клеточных рецепторов (T-cell receptor, TCR). Т-клеточный рецептор взаимодействует с комплексом «эпитоп-главный комплекс гистосовместимости» (major histocompatibility complex, MHC) своими участками, определяющими комплементарность (complementarity determining regions, CDR). Взаимодействие «Т-клеточный рецептор-эпитоп-главный комплекс гистосовместимости» (TCR-peptide-MHC, TCR-pMHC) является одним из ключевых событий, приводящим к развитию иммунного ответа. Предсказание ответа иммунных клеток на специфические антигены является одной из ключевых задач иммуноинформатики. Решение этой проблемы позволяет ускорить разработку биологических препаратов, повысить их терапевтическую эффективность и снизить затраты на создание средств диагностики и лечения. Для этих целей широко применяются вычислительные методы, алгоритмы и прогностические модели, используемые при рациональном дизайне вакцин и иммунодиагностических систем [1], [2]. Экспериментально подтверждено, что наибольший вклад в распознавание эпитопов вносят участки CDR3 TCR [3]. На протяжении последних 30 лет научные коллективы по всему миру собирают экспериментальные данные о взаимодействии TCR с эпитопами. Эти данные агрегируются в специализированные базы данных: одна из крупнейших — Immune Epitope Database (IEDB) содержит больше 250000 записей о взаимодействии последовательностей CDR3 TCR с эпитопами [4].

Современные подходы для прогноза взаимодействия TCR-pMHC в основном опираются на модели машинного обучения, включая глубокие нейронные сети. Данное направление активно развивается благодаря прогрессу в области искусственного интеллекта [5], [6]. Однако, несмотря на достигнутые успехи, большинство современных моделей имеют относительно низкую точность и охватывают довольно узкий спектр эпитопов [7]. Так, NetTCR-2.2 позволяет делать прогноз для 26 эпитопов со средней величиной ROC AUC 0,85 [8]. TCRex показывает похожие результаты на выборке IMMPREP2023, при этом способен делать прогноз для 100 эпитопов [8], [9]. Доступные данные характеризуются очень низким видовым разнообразием, что ограничивает область применимости этих моделей [10].

Ранее нами был разработан веб-сервис TCR-Pred, который способен рассчитывать вероятность связывания TCR с пептидным эпитопом и молекулами MHC и который не уступает своим аналогам [11]. Наш подход использует модифицированный наивный Байесовский классификатор для построения моделей зависимости «структура-активность» (SAR) и основан на представлении белковой последовательности CDR3 в виде структурных формул, описанных набором дескрипторов многоуровневых атомных окрестностей (Multilevel Neighbourhoods of Atoms, MNA) [12]. Впервые для биологических макромолекул он был применен для предсказания сайтов фосфорилирования белков [13]. Предсказание взаимодействия TCR-pMHC и сайтов фосфорилирования белков послужило подтверждением возможности применения подобных методов хемоинформатики для решения задач, связанных с оценкой свойств биологических макромолекул. С использованием данного подхода ранее также были созданы классификационные модели для предсказания патогенных вариантов аминокислотных замен, приводящих к лекарственной устойчивости, сайтов ацетилирования и вторичных структур белка [14], [15], [16], [17]. Целью данного исследования являлось изучение влияния учёта электрического заряда атомов в описании структурных формул аминокислотных остатков пептидов на качество прогноза и обновление моделей TCR-Pred с учетом новых экспериментальных данных.

Методы и принципы исследования

Подготовка данных и построение моделей были автоматизированы с использованием Nextflow v. 25.04.06 и Python 3.10 [18]. Часть вычислений проводилась на базе инфраструктуры высокопроизводительного кластера Алдан3 вычислительного центра НИИ Трансляционной медицины ФГАОУ ВО РНИМУ им. Н.И. Пирогова (Пироговский Университет) [19]. Для построения моделей были использованы сведения, извлечённые из баз данных IEDB, CEDAR, VDJdb, McPAS-TCR, PIRD [4], [20], [21], [23]. Исходные выборки были скачаны 21 марта 2026 года. Были использованы следующие критерии фильтрации:

Общие правила: Наличие публикации (литературной ссылки). Результат взаимодействия между TCR и pMHC — положительный. Цепь CDR3 — альфа и/или бета. Корректные последовательности CDR3 (записанные в однобуквенном формате из 20 протеиногенных аминокислот). Организм-хозяин CDR3 – Homo sapiens (человек) или Mus musculus (мышь).

Данные о взаимодействии «Т-клеточный рецептор-эпитоп»: Корректная последовательность эпитопа. Эпитопы без химических модификаций.

Данные о взаимодействии «Т-клеточный рецептор — MHC аллель»: Первый или второй класс MHC (для моделей «Т-клеточный рецептор-эпитоп» эта информация может быть пропущена). Аллель MHC записана до конкретного белкового продукта (если была дополнительная информация, она была удалена).

Данные были сгруппированы в зависимости от типа цепи CDR3 (альфа или бета), организма-хозяина (человек или мышь) и типа прогнозируемой активности (эпитоп или аллель MHC) во всех возможных сочетаниях. При объединении данных были удалены записи-дубликаты. Аллели MHC приведены к общему номенклатурному виду с помощью библиотеки mhcgnomes [24]. Дубликаты выявлялись по совпадению столбцов со структурой (CDR3) и активностью (эпитоп или MHC). Получившиеся таблицы были преобразованы в Structure-Data Format (SDF) файлы с помощью специальной программы, разработанной авторами [25], [26]. В результате были сгенерированы SDF файлы,

 © Авторы статьи / Authors of the article

содержащие два типа структур: с полностью заряженными аминокислотными остатками и полностью нейтральными (такое представление было использовано при создании первой версии веб приложения TCR-Pred [11]).

Для построения моделей взаимодействия «Т-клеточный рецептор — эпитоп» и «Т-клеточный рецептор — МНС» была использована модифицированная версия программы MultiPASS (версия 2025 года) [27], [28]. MultiPASS работает на основе усовершенствованного алгоритма наивного Байесовского классификатора и использует представление структурных формул последовательностей CDR3 в виде дескрипторов многоуровневых атомных окрестностей (Multilevel Neighborhoods of Atoms, MNA). Принцип построения дескрипторов MNA представлен на рисунке 1. MultiPASS рассчитывает вероятности, что вещество будет активным (Pa) по отношению к некоторой мишени, и вероятность, что будет неактивно (Pi) по отношению к ней. Итоговый прогноз представляет собой разность этих вероятностей (Pa-Pi) и принимает значения в интервале от -1 до 1. Чем выше эта величина, тем выше вероятность, что исследуемое вещество обладает прогнозируемой активностью (взаимодействием с конкретным эпитопом или МНС аллели).

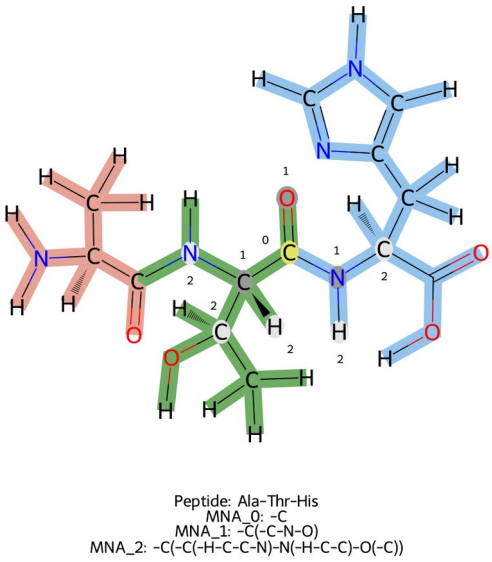

Рисунок 1 - Принцип генерации дескрипторов многоуровневых атомных окрестностей (MNA) для карбоксильного атома углерода треонина

DOI: https://doi.org/10.60797/jbg.2026.33.3.1

Примечание: выделен светло-жёлтым, номер «0»

Дескрипторы MNA являются линейной нотацией атомно-центрированных фрагментов молекулы, формируемой рекурсивно по следующему принципу. Дескриптором нулевого уровня (MNA_0) является символьное обозначение атома согласно периодической таблице Менделеева (например -C, атом углерода номер «0» в пептидной связи между остатками треонина и гистидина). Символ «-» добавляется к обозначению атомов, не входящих ни в какие циклы. Дескриптор любого другого уровня (MNA_N) формируется из дескриптора MNA_0 атома и заключенных в скобки дескрипторов предыдущего уровня (MNA_N-1) атомов, непосредственно связанных с ним ковалентной связью (MNA_1 – -C(-C-N-O), MNA_2 – -C(-C(-H-C-C-N)-N(-H-C-C)-O(-C)), и т.д.). Светло-серым выделены атомы, входящие в окрестность второго уровня. В MultiPASS применена специальная технология, благодаря которой дескрипторы MNA любого уровня состоят не более чем из 40 символов. Структура молекулы в MultiPASS представляется в виде бесповторного множества дескрипторов каждого из атомов молекулы от MNA_1 до MNA_N с заданным максимальным уровнем N. Розовым цветом выделен остаток аланина, зеленым — треонина, голубым — гистидина.

Модели SAR были созданы для каждого SDF файла с использованием максимального уровня (далее — просто «уровня») дескрипторов от 3 до 12. Оценка качества моделей производилась в процедуре перекрёстной проверки с разбиением на 5 частей с помощью метрики площади под кривой оперативной характеристики приемника (ROC AUC). Модели были отобраны по двум критериям. Во-первых, ROC AUC должен быть больше или равен 0,7. Второй критерий, минимальное количество пептидов участков CDR3 TCR в выборке, равен 3 последовательностям. Оптимальный уровень дескрипторов MNA равен минимальному уровню, при котором достигается максимальная медианная величина ROC AUC или очень близкая к нему. В случае затруднения в выборе лучшей модели по величине точности учитывали максимальное количество эпитопов или МНС аллелей, для которых можно было делать предсказание соответствующей моделью — широта спектра активностей. Таким образом, лучшими считались модели SAR, обладающие максимальными значениями как медианной ROC AUC, так и широты спектра. В случае незначительных отклонений между моделями с учетом зарядов атомов и без него приоритет отдавался последним, так как хранение заряда увеличивает размер файлов выборок и моделей, приводя к повышению нагрузки на работу сервера.

 © Авторы статьи / Authors of the article

Основные результаты

3.1. Характеристика данных

В таблице 1 представлены характеристики обучающих выборок старой (v.2023) и новой версии (v.2026) TCR-Pred. В новой версии в 1,5–2 раза больше CDR3 последовательностей, последовательностей эпитопов и аллелей MHC. Этот рост количества данных может быть связан как с обновлением баз, так и со смягчением критериев отбора. Так, из критериев фильтрации была исключена «каноничность» последовательности CDR3 — то есть требование наличия консервативного цистеина на N-конце и фенилаланина на C-конце. Наблюдается небольшое снижение числа последовательностей для человеческих бета-цепей CDR3. Это может быть связано с обновлением и уточнением информации в самих базах данных. Для определения оптимальных параметров моделирования итоговые обучающие выборки были разбиты в зависимости от типа цепи CDR3 и организма-хозяина CDR3 во всех возможных сочетаниях. Выборки по классам MHC делить не стали, так как из-за этого значительно падает общая точность прогноза вследствие уменьшения объема обучающей выборки.

Таблица 1 - Информация об обучающих выборках предыдущей и новой версии после обработки с разбиением по организму-хозяину и типу цепи

DOI: https://doi.org/10.60797/jbg.2026.33.3.2

Количество уникальных структур в моделях		v.2023				v.2026			
		Человек		Мышь		Человек		Мышь	
		α	β	α	β	α	β	α	β
CDR3 – эпитоп	CDR3	35616	141500	1604	3014	72630	134026	3401	6156
	Эпитоп	799	1257	96	123	1865	2052	148	171
CDR3 – MHC	CDR3	31066	47639	1462	2550	72630	134026	3401	6156
	MHC	76	88	11	12	121	146	11	14

Примечание: данные для моделей предсказания активности к эпитопам и MHC обрабатывались отдельно, поэтому у них отличается количество записей

Обращает на себя внимание, что активность Т-клеточных рецепторов изучена для малого разнообразия возможных эпитопов. На рисунке 2 представлено распределение объёма положительных примеров в обучающих выборках для конкретной активности (эпитопа/аллель MHC). Базы данных практически не содержат отрицательных примеров. MultiPASS принимает все примеры, которые не отмечены как положительные для данной активности, за отрицательные. Крайне мало эпитопов имеет большое количество (> 100) положительных примеров (взаимодействующих последовательностей CDR3), а половина всех имеющихся эпитопов/аллелей MHC имеет не более 2–3 положительных примеров.

 © Авторы статьи / Authors of the article

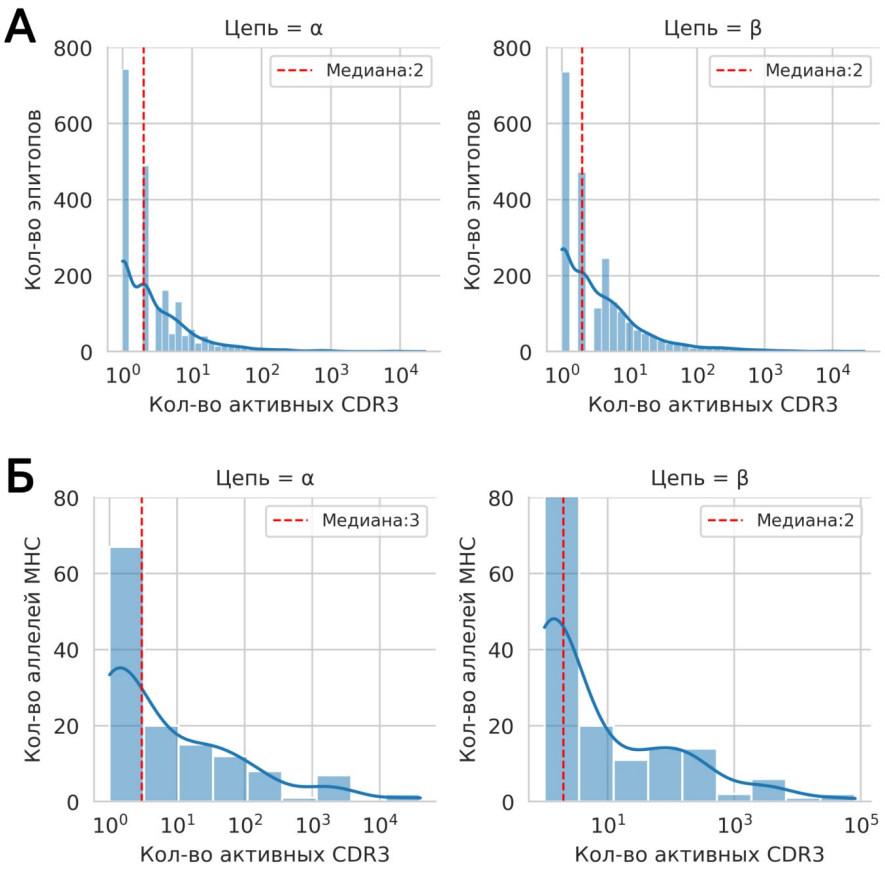

Рисунок 2 - Гистограммы распределения количества последовательностей CDR3, взаимодействующих с конкретным эпитопом/аллелем MHC в обучающих выборках с разбиением по цепи CDR3:
А – для эпитопов; *Б* – для аллелей MHC; *ось ординат* – количество эпитопов/аллелей MHC; *ось абсцисс* – число CDR3-последовательностей на эпитоп/аллель MHC; *синяя линия* – ядерная оценка плотности распределения; *красная линия* – отметка медианы распределения
DOI: https://doi.org/10.60797/jbg.2026.33.3.3

3.2. Проверка влияния учёта электрического заряда атомов аминокислотных остатков на точность моделей SAR

Перед построением моделей SAR была выдвинута гипотеза, что учёт электрического заряда атомов аминокислотных остатков в структурных формулах последовательностей CDR3 должен улучшить медианную точность моделей SAR. Для проверки этой гипотезы были сформированы два вида SDF файлов, с указанием заряда атомов в структурах аминокислотных остатков и без указания заряда. Для обоих видов SDF файлов были проведены одинаковые процедуры обучения, валидации и отбора лучших моделей SAR. Моделям SAR были присвоены названия заглавными буквами на английском языке в соответствии с использованной выборкой в следующем формате: *цепь_активность_вид*, например BETA_EPITOPE_MOUSE. Результаты сравнения представлены на рисунке 3. Из приведенных на рисунке 3 данных видно, что значимого улучшения медианной точности предсказываемых эпитопов или MHC аллелей, так же как и увеличения их количества, не наблюдается при добавлении информации о заряде молекул. Диаграммы на рисунке 3 построены по активностям, для которых удалось провести процедуру перекрёстной проверки. Для моделей SAR ALPHA_EPITOPE_HUMAN, BETA_EPITOPE_HUMAN количество эпитопов, прошедших перекрёстную проверку, не возрастало монотонно. Для дальнейших экспериментов использовались модели SAR без учёта зарядов атомов, потому что файлы таких выборок и моделей SAR занимают меньше места на диске. Следует отметить, что на диаграммах показаны результаты пятикратной кросс-валидации, поэтому количество активностей, которые могут предсказывать модели SAR несколько выше. По имеющимся диаграммам были выбраны оптимальные уровни дескрипторов MNA. Для каждого уровня MNA дескрипторов была проведена проверка статистической значимости различий между распределениями значений ROC AUC был использован тест Манна-Уитни из библиотеки scipy для языка программирования Python с параметрами по умолчанию. Этот критерий был выбран, так как закон распределения ROC AUC в нашем случае неизвестен и не симметричен. Из-за большого количества тестирований поправка Бенджамини-Хохберга, реализованная в этой же библиотеке, с параметрами по умолчанию была использована для корректировки p-значений. Ни на одном из уровней MNA дескрипторов, ни для одной из моделей не было найдено статистически значимых отличий между заряженными и нейтральными дескрипторами. Подробная информация приложена в дополнительных материалах, таблица 1.

 © Авторы статьи / Authors of the article

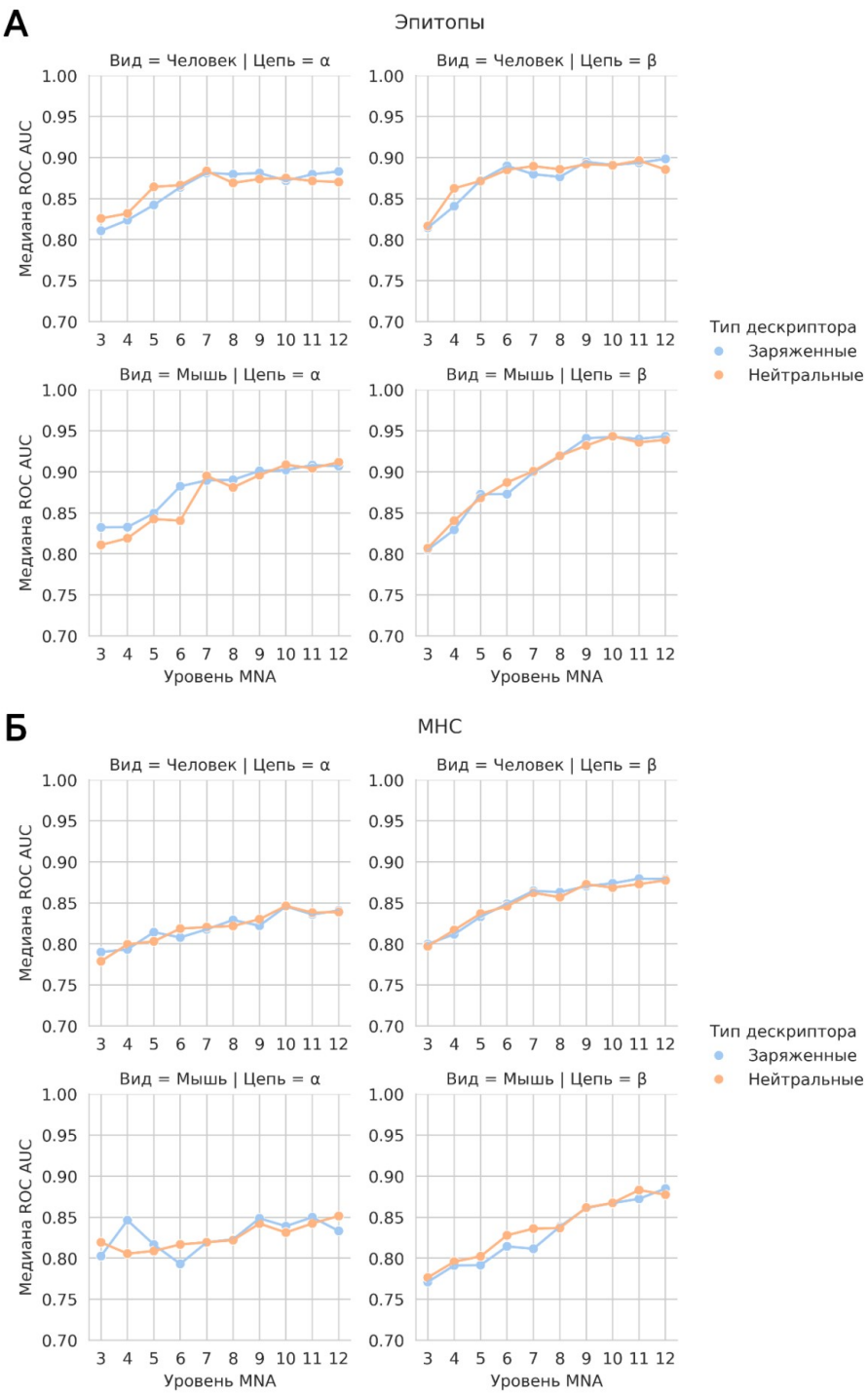

Рисунок 3 - Диаграммы зависимости медианной точности прогноза для эпитопов (*А*) и MHC (*Б*) по результатам пятикратной кросс-валидации от уровня дескрипторов MNA для моделей SAR с учётом зарядов атомов и без него
DOI: https://doi.org/10.60797/jbg.2026.33.3.4

3.3. Сравнение качества моделей SAR версий v.2023 и v.2026

Для первоначального сравнения моделей SAR разных версий мы отобрали только те активности, которые присутствуют в обеих версиях. Результаты такого сравнения представлены на рисунке 4. Новые версии лучших моделей SAR по медианной точности при пятикратной кросс-валидации превзошли все старые, за исключением ALPHA_EPITOPE_MOUSE. Для общего сравнения качества моделей SAR было проведено сравнение версий на всем спектре активностей. Характеристики моделей SAR при оптимальном значении уровня дескрипторов MNA представлены в таблице 2. Из приведенных в таблице 2 данных видно, что новые версии моделей (v.2026) превосходят старые (v.2023) по спектру активностей и медианной точности. Исключение составляет модель ALPHA_EPITOPE_MOUSE, где наблюдается расхождение, однако мы считаем, что уменьшение медианного AUC на 0,05 не так значимо, как расширение спектра активностей в 2 раза, с 13 до 28 MHC аллелей. Для некоторых моделей SAR, таких как BETA_EPITOPE_HUMAN, наблюдается рост количества предсказываемых эпитопов в 6 раз. Для большинства моделей SAR получилось добиться значительного прироста точности. Статистически значимые отличия

 © Авторы статьи / Authors of the article

в распределениях ROC AUC найдены в 6 из 8 типах моделей (исключая ALPHA_EPITOPE_MOUSE, BETA_MHC_MOUSE) в 36 точках (из 80). Для BETA_EPITOPE_HUMAN, BETA_EPITOPE_MOUSE различия наблюдаются практически на каждом уровне MNA дескрипторов. Подробная информация приложена в дополнительных материалах, таблица 2.

Таблица 2 - Сравнительная характеристика двух версий TCR-Pred

DOI: https://doi.org/10.60797/jbg.2026.33.3.5

Модель SAR	Версия	Оптимальный уровень дескрипторов MNA	Количество активностей	Медиана ROC AUC
ALPHA_EPITOPE_HUMAN	v.2023	10	97	0,8054
	v.2026	**10**	**540**	**0,8750**
BETA_EPITOPE_HUMAN	v.2023	9	196	0,7908
	v.2026	**9**	**803**	**0,8919**
ALPHA_MHC_HUMAN	v.2023	7	20	0,8233
	v.2026	**10**	**44**	**0,8464**
BETA_MHC_HUMAN	v.2023	10	28	0,7849
	v.2026	**9**	**53**	**0,8727**
ALPHA_EPITOPE_MOUSE	**v.2023**	**5**	**13**	**0,9583**
	v.2026	10	**28**	0,9086
BETA_EPITOPE_MOUSE	v.2023	9	42	0,8463
	v.2026	**10**	**45**	**0,9434**
ALPHA_MHC_MOUSE	v.2023	6	7	0,7865
	v.2026	**9**	**8**	**0,8423**
BETA_MHC_MOUSE	v.2023	8	7	0,7539
	v.2026	**11**	**9**	**0,8831**

 © Авторы статьи / Authors of the article

Примечание: жирным шрифтом отмечены значения параметров, превосходящие такие же в другой версии для конкретной модели SAR

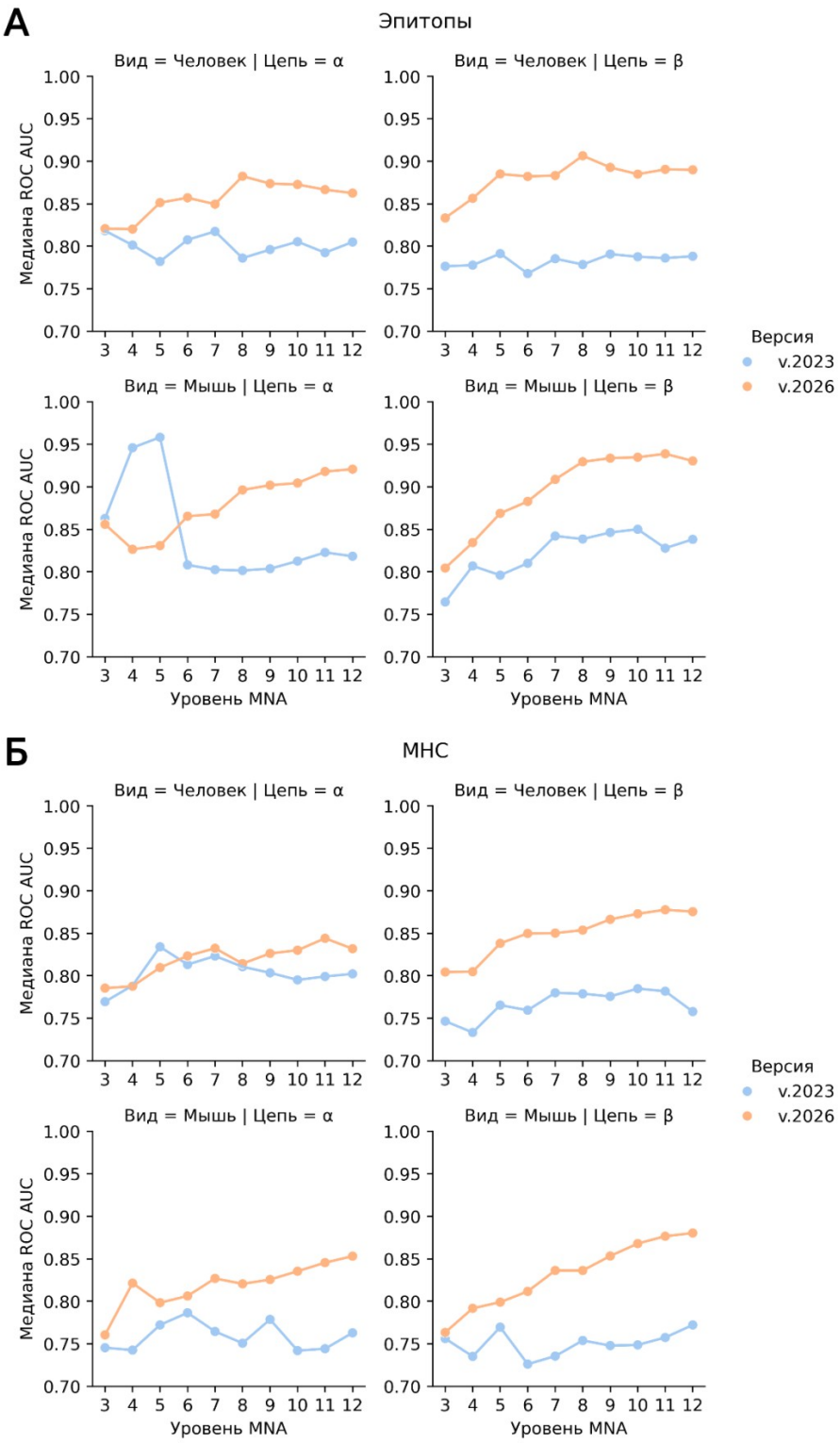

Рисунок 4 - Диаграммы зависимости медианной точности прогноза по результатам пятикратной кросс-валидации от уровня дескрипторов MNA для старой и новой версии моделей SAR:
А – для эпитопов; *Б* – для MHC
DOI: https://doi.org/10.60797/jbg.2026.33.3.6

Примечание: диаграммы построены с использованием только тех активностей, которые являются общими для двух версий

 © Авторы статьи / Authors of the article

Обсуждение

Предсказание взаимодействия рецепторов адаптивного иммунитета с эпитопами в настоящий момент является трудной задачей. Экспериментально эти взаимодействия изучены для малого количества эпитопов, причем для большинства из них известно небольшое количество положительных примеров последовательностей, взаимодействующих с ними CDR3 участков Т-клеточных последовательностей, при этом нет информации об однозначно отрицательных примерах, что затрудняет построение классификационных моделей. За три года произошел существенный рост количества экспериментальных данных, что позволило в новой версии TCR-Pred построить отдельные модели для человека и мыши. Увеличить количество данных также помогла их менее строгая фильтрация. В то же время, данное исследование показало, что учёт дополнительной информации, такой как электрический заряд атомов аминокислотных остатков фрагментов белков, не приводит к повышению статистически значимого количества более точных моделей. Такой результат является неожиданным, потому что электростатические взаимодействия играют важную роль при белок-белковых взаимодействиях. Вполне вероятно, что в текущем виде MNA дескрипторы уже неявно учитывают электростатические взаимодействия. Кроме того, теоретические расчеты указывают на то, что гидрофобные взаимодействия между Т-клеточным рецептором и эпитопом играют более важную роль, чем кулоновские взаимодействия [29]. Несмотря на то, что с точки зрения медианной точности прогноза разница между моделями, созданными с учетом заряда аминокислотных остатков и без их учета несущественна, для отдельных эпитопов и MHC аллелей использование зарядов атомов аминокислотных остатков фрагментов белков приводило к большей точности моделей. Это можно использовать для дальнейшего совершенствования веб-приложения в будущем.

Отдельно стоит отметить, что прогноз взаимодействия Т-клеточных рецепторов с эпитопами, которые отсутствовали в обучающей выборке, остаётся нерешённой задачей. Хемоинформатический подход в представлении белковых последовательностей в виде молекулярных дескрипторов имеет большой потенциал для обхода этого ограничения, а также для решения обратной задачи — подбор последовательностей рецептора под конкретный эпитоп. В дальнейшем мы продолжим совершенствовать TCR-Pred в направлении увеличения широты предсказываемых активностей, решения проблемы «неизвестных эпитопов» и создания REST API интерфейсов для Python и R.

Заключение

В данной работе было показано, что учёт информации о заряде атомов в описании структурных формул аминокислотных остатков CDR3 TCR практически не влияет на точность прогноза активности Т-клеточных рецепторов по отношению к эпитопам и аллелям MHC.

Также нами было обновлено веб приложение TCR-Pred, в котором значительно расширился спектр предсказываемых эпитопов и MHC аллелей, а также для большинства моделей повысилась медианная точность.

5.1. Доступность данных и исходного кода

Исходный код подготовки и анализа данных доступен по адресу github.com/SmirnygaTotoshka/TCRpred. Программа MultiPASS доступна по обоснованному запросу. Базы данных, используемые в работе, являются открытыми и общедоступными. Веб приложение TCR-Pred является свободно доступным (way2drug.com/TCR-Pred/).

Дополнительные материалы

Дополнительные материалы доступны на онлайн-странице статьи.

Финансирование

Работа выполнена при финансовой поддержке Российского научного фонда № 25-25-01171, https://rscf.ru/project/25-25-01171/.

Конфликт интересов

Не указан.

Рецензия

Яковлев И.В., Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, Казань Российская Федерация
DOI: https://doi.org/10.60797/jbg.2026.33.3.7

Supplementary materials

Supplementary materials are available online on the article's webpage.

Funding

This work was supported by the Russian Science Foundation (project № 25-25-01171, https://rscf.ru/project/25-25-01171/).

Conflict of Interest

None declared.

Review

Iakovlev I.V., Kazan National Research Technical University named after A.N. Tupolev – KAI, Kazan Russian Federation
DOI: https://doi.org/10.60797/jbg.2026.33.3.7

Список литературы на английском языке / References in English

1. Schaap-Johansen A.-L. T Cell Epitope Prediction and Its Application to Immunotherapy / A.-L. Schaap-Johansen, M. Vujović, A. Borch [et al.] // Frontiers in Immunology. — 2021. — Vol. 12. — URL: https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2021.712488/full (accessed: 11.09.2025).
2. Ahmad T.A. B-cell epitope mapping for the design of vaccines and effective diagnostics / T.A. Ahmad, A.E. Eweida, S.A. Sheweita // Trials in Vaccinology. — 2016. — Vol. 5. — P. 71–83.
3. Krogsgaard M. How T cells "see" antigen / M. Krogsgaard, M. M. Davis // Nature Immunology. — 2005. — Vol. 6. — № 3. — P. 239–245.
4. Vita R. The Immune Epitope Database (IEDB): 2018 update / R. Vita, S. Mahajan, J. A. Overton [et al.] // Nucleic Acids Research. — 2019. — Vol. 47. — № D1. — P. D339–D343.
5. Yadav S. TCR-ESM: Employing protein language embeddings to predict TCR-peptide-MHC binding / S. Yadav, D.S. Vora, D. Sundar [et al.] // Computational and Structural Biotechnology Journal. — 2024. — Vol. 23. — P. 165–173.

 © Авторы статьи / Authors of the article

6. Jokinen E. Predicting recognition between T cell receptors and epitopes with TCRGP / E. Jokinen, J. Huuhtanen, S. Mustjoki [et al.] // PLOS Computational Biology. — 2021. — Vol. 17. — № 3. — P. e1008814.

7. Lu T. Deep learning-based prediction of the T cell receptor–antigen binding specificity / T. Lu, Z. Zhang, J. Zhu [et al.] // Nature Machine Intelligence. — 2021. — Vol. 3. — № 10. — P. 864–875.

8. Jensen M.F. Enhancing TCR specificity predictions by combined pan- and peptide-specific training, loss-scaling, and sequence similarity integration / M.F. Jensen, M. Nielsen // eLife. — 2024. — Vol. 12. — P. RP93934.

9. Gielis S. Identification of Epitope-Specific T Cells in T-Cell Receptor Repertoires / S. Gielis, P. Moris, W. Bittremieux [et al.] // Bioinformatics for Cancer Immunotherapy : Methods in Molecular Biology / Ed. by S. Boegel. — New York : Springer US, 2020. — Vol. 2120. — P. 183–195. — URL: http://link.springer.com/10.1007/978-1-0716-0327-7_13 (accessed: 12.01.2026).

10. Hudson D. Can we predict T cell specificity with digital biology and machine learning? / D. Hudson, R. A. Fernandes, M. Basham [et al.] // Nature Reviews Immunology. — 2023. — Vol. 23. — № 8. — P. 511–521.

11. Smirnov A.S. TCR-Pred: A new web-application for prediction of epitope and MHC specificity for CDR3 TCR sequences using molecular fragment descriptors / A.S. Smirnov, A.V. Rudik, D.A. Filimonov [et al.]. — URL: https://onlinelibrary.wiley.com/doi/10.1111/imm.13641 (date accessed: 22.05.2025).

12. Lagunin A. PASS: prediction of activity spectra for biologically active substances / A. Lagunin, A. Stepanchikova, D. Filimonov [et al.] // Bioinformatics. — 2000. — Vol. 16. — № 8. — P. 747–748.

13. Karasev D.A. Application of molecular descriptors for recognition of phosphorylation sites in amino acid sequences / D.A. Karasev, P.I. Savosina, B.N. Sobolev [et al.] // Biomeditsinskaya Khimiya. — 2017. — Vol. 63. — № 5. — P. 423–427.

14. Zadorozhny A. Prediction of pathogenic single amino acid substitutions using molecular fragment descriptors / A. Zadorozhny, A. Smirnov, D. Filimonov [et al.] // Bioinformatics (Oxford, England). — 2023. — Vol. 39. — № 8. — P. btad484.

15. Porfireva E.S. Sequence-structure based prediction of pathogenicity for amino acid substitutions in proteins associated with primary immunodeficiencies / E.S. Porfireva, A.D. Zadorozhny, A.V. Rudik [et al.] // Frontiers in Immunology. — 2025. — Vol. 16. — P. 1492751.

16. Zakharov O.S. Prediction of Protein Secondary Structures Based on Substructural Descriptors of Molecular Fragments / O.S. Zakharov, A.V. Rudik, D.A. Filimonov [et al.] // International Journal of Molecular Sciences. — 2024. — Vol. 25. — № 23. — P. 12525.

17. Lebedev N.V. Prediction of New Nε-Acetylation Sites in the Human Proteome Based on Molecular Multilevel Neighborhoods of Atom Descriptors / N.V. Lebedev, D.A. Filimonov, V.V. Poroikov [et al.] // Molecular Biology. — 2025. — Vol. 59. — № 5. — P. 836–848.

18. Di Tommaso P. Nextflow enables reproducible computational workflows / P. Di Tommaso, M. Chatzou, E.W. Floden [et al.] // Nature Biotechnology. — 2017. — Vol. 35. — № 4. — P. 316–319.

19. Kotliar V. Building a Computer Cluster for Bioinformatics and Biomedical Research from Scratch to Production / V. Kotliar, D. Bolotin, D. Syrko // Physics of Particles and Nuclei. — 2024. — Vol. 55. — № 3. — P. 474–476.

20. Koşaloğlu-Yalçın Z. The Cancer Epitope Database and Analysis Resource (CEDAR) / Z. Koşaloğlu-Yalçın, N. Blazeska, R. Vita [et al.] // Nucleic Acids Research. — 2022. — Vol. 51. — № D1. — P. D845–D852.

21. Goncharov M. VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2 / M. Goncharov, D. Bagaev, D. Shcherbinin [et al.] // Nature Methods. — 2022. — Vol. 19. — № 9. — P. 1017–1019.

22. Tickotsky N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences / N. Tickotsky, T. Sagiv, J. Prilusky [et al.] // Bioinformatics. — 2017. — Vol. 33. — № 18. — P. 2924–2929.

23. Zhang W. PIRD: Pan Immune Repertoire Database / W. Zhang, L. Wang, K. Liu [et al.] // Bioinformatics. — 2020. — Vol. 36. — № 3. — P. 897–903.

24. Rubinsteyn A. mhcgnomes: Parsing MHC nomenclature in the wild / A. Rubinsteyn, T. O'Donnell, N. Wheeler.

25. BIOVIA CTFILE FORMATS. — URL: https://discover.3ds.com/sites/default/files/2020-08/biovia_ctfileformats_2020.pdf (accessed: 14.08.2025).

26. Dalby A. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited / A. Dalby, J.G. Nourse, W.D. Hounshell [et al.] // Journal of Chemical Information and Computer Sciences. — 1992. — Vol. 32. — № 3. — P. 244–255.

27. Zhuravleva S.I. Prediction of Amino Acid Substitutions in ABL1 Protein Leading to Tumor Drug Resistance Based on «Structure-Property» Relationship Classification Models / S.I. Zhuravleva, A.D. Zadorozhny, B.V. Shilov [et al.] // Life (Basel, Switzerland). — 2023. — Vol. 13. — № 9. — P. 1807.

28. Filimonov D.A. Computer-aided prediction of biological activity spectra for chemical compounds: opportunities and limitation / D.A. Filimonov, D.S. Druzhilovskiy, A.A. Lagunin [et al.] // Biomedical Chemistry: Research and Methods. — 2018. — Vol. 1. — № 1. — P. e00004.

29. Rollins Z.A. Using molecular dynamics simulations to interrogate T cell receptor non-equilibrium kinetics / Z.A. Rollins, R. Faller, S.C. George // Comput. Struct. Biotechnol. J. — 2022. — Vol. 20. — P. 2124–2133.