УДК 575.112+616-097.1; 616-006

БИОИНФОРМАТИЧЕСКИЙ ПОДХОД НА ОСНОВЕ МЕТАМОДЕЛЕЙ ДЛЯ УЛУЧШЕНИЯ ПРОГНОЗИРОВАНИЯ ИММУНОГЕННОСТИ ОПУХОЛЕВЫХ ЭПИТОПОВ

И. С. Трусов, В. В. Гринев

Белорусский государственный университет,
Минск, Беларусь, ivan.trusau@yandex.by

В данной работе предложен и апробирован биоинформатический подход, основанный на метамоделях, комплексно оценивающих весь каскад событий от внутриклеточной обработки антигена до его распознавания иммунной системой. Проведённый в работе сравнительный анализ продемонстрировал преимущества метамоделей по сравнению с существующими аналогами в точности предсказания презентации и иммуногенности эпитопов.

Ключевые слова: неоантигены; прогнозирование иммуногенности; машинное обучение; метамодель.

BIOINFORMATIC APPROACH BASED ON METAMODELS FOR IMPROVING THE PREDICTION OF THE IMMUNOGENICITY OF TUMOUR EPITOPES

I. S. Trusau, V. V. Grinev

Belarusian State University,
Minsk, Belarus, ivan.trusau@yandex.by

In this paper, we propose and test a bioinformatics approach based on metamodels that comprehensively evaluate the entire cascade of events from intracellular processing of an antigen to its recognition by the immune system. A comparative analysis conducted in this work demonstrated the advantages of metamodels over existing analogues in terms of the accuracy of predicting immunogenicity and epitope presentation.

Keywords: neoantigens; immunogenicity prediction; machine learning; metamodel.

1. Введение

В современной онкологии иммунотерапия признана одним из наиболее перспективных и быстро развивающихся. Иммунотерапия открывает перспективы для пациентов с различными формами рака, которые раньше считались неизлечимыми. Клинический успех данного подхода во многом

обусловлен способностью нашей иммунной системы самостоятельно распознавать и уничтожать опухолевые клетки. Наиболее значимым элементом в этом механизме являются неоантигены, представляющие собой уникальные пептидные фрагменты, возникающие в результате соматических мутаций и отсутствующие в нормальных тканях организма. Их специфичность позволяет использовать их в качестве идеальных мишеней при лечении, минимизируя риск побочных эффектов и аутоиммунных реакций на здоровые клетки [5].

Для распознавания неоантигена иммунными клетками необходим целый ряд сложных и многоступенчатых реакций, называемых процессингом. Инициация процессинга антигена происходит на стадии протеолитического расщеплением белка-предшественника специализированными ферментными комплексами – протеасомами. Образующиеся пептидные фрагменты затем транспортируются в эндоплазматический ретикулум при участии транспортёра TAP, ассоциированного с процессингом антигена. В эндоплазматическом ретикулуме происходит укорочение пептидов специальными ферментами и их связывание с молекулами главного комплекса гистосовместимости (MHC) класса I. Пептид, связанный с MHC, также называют эпитопом. Такой комплекс впоследствии выносится на клеточную поверхность для презентации Т-лимфоцитам. Последние, благодаря своему Т-клеточному рецептору, способны распознать комплекс MHC-эпитоп и запустить иммунный ответ. Прохождение каждого этапа зависит от физико-химических свойств пептидов, поэтому комплексная оценка процесса крайне важна для определения способности неоантигена вызывать эффективный противоопухолевый иммунитет.

Несмотря на бурное развитие биоинформатических инструментов для прогнозирования потенциальных иммуногенных неоантигенов, их точность остаётся недостаточной, что ограничивает их широкое клиническое применение. Существующие модели зачастую фокусируются лишь на одном из этапов обработки антигена. Подобный узкоспециализированный подход не учитывает всю сложность процесса, что обусловливает появление множества ложных результатов, существенно снижающих предсказательную силу моделей и создавая серьёзное препятствие перед разработкой действительно персонализированных вакцин и иммунотерапии онкологических заболеваний.

В связи с вышеизложенным мы задались целью разработать комплексный биоинформатический подход, учитывающий все этапы процессинга неоантигенов и их распознавания иммунной системой. Данный подход, основанный на использовании мета-классификаторов, направлен на существенное повышение точности прогнозирования иммуногенности эпитопов неоантигенов, что, по нашему мнению, может повысить

эффективность и надёжность идентификации молекулярных мишеней для персонализированной иммунотерапии рака.

2. Материалы и методы

При выполнении данной работы был использован язык программирования Python 3.8. Для эффективной обработки данных и реализации алгоритмов машинного обучения применялись библиотеки NumPy, Pandas и Scikit-learn. Вся работа проводилась на компьютере с 16 Гб оперативной памяти, 8-ядерным 16-поточным процессором, видеокартой Nvidia GeForce GTX 1650 4 Гб и SSD.

Основные данные были получены из открытых баз IEDB и CEDAR. Данные содержали информация о пептидах, их связывании с молекулами MHC класса I и иммуногенности. Дополнительные данные были взяты из публикаций, где описывались программные пакеты ConvNeXt, TranspMHC, ImmuneApp, PRIME, BigMHC, ICERFIRE и IEPAPI. Собранные данные прошли оценку качества и фильтрацию, после чего они были сгруппированы в два класса – данные об иммуногенности и данные о презентации. Данные об иммуногенности мы разделили на два набора: 1) набор Neo, куда попали только данные о неоантигенах, и 2) набор IE, который включил пептиды инфекционной природы.

Для прогнозирования процессинга эпитопа использовались модели предсказания протеасомного расщепления (pepsickle), транспорта пептидов через TAP (DeepTAP), N-концевого тримминга ферментом ERAP1 (ERAMER) и связывания эпитопа с молекулой MHC (TranspMHC). Для оценки иммуногенности применялись модели, оценивающие секрецию интерлейкинов: IL2 (IL2pred), IL4 (IL4pred), IL6 (IL6pred), IL10 (IL10pred), IL13 (IL13pred), интерферон гамма (IFNepitope 2) и фактор некроза опухолей (TNFepitope).

В качестве физико-химических свойств были выбраны гидрофобность, алифатический индекс, индекс Бомана, изоэлектрическая точка и длина пептида, которые вычислялись при помощи библиотеки peptides [7]. Также проводилась оценка чужеродности пептида по отношению к человеческому протеому при помощи программы antigen.garnish [10].

Для создания метамоделей использовались алгоритмы случайного леса, реализованные при помощи библиотеки Scikit-learn. Оценка моделей осуществлялась с использованием стандартных метрик AUC-ROC и AUPRC. Для интерпретации результатов и анализа важности признаков применялся метод SHAP (SHapley Additive exPlanations) из библиотеки shap, позволяющий оценить вклад отдельных признаков в конечный результат прогнозов модели.

3. Результаты

3.1. Метамодель предсказания презентации эпитопа

Точная оценка презентации эпитопа является необходимым предварительным условием для последующего анализа иммуногенности, поскольку пептид может быть распознан иммунной системой только в случае его успешной презентации в комплексе с MHC на поверхности клетки. Процесс презентации антигена включает несколько последовательных внутриклеточных этапов процессинга и для его предсказания была разработана метамодель, объединяющая результаты работы нескольких моделей, каждая из которых предсказывает свой этап процессинга. Краеугольным камнем этой системы является модель глубокого обучения TranspMHC [2], которая оценивает финальный этап процессинга – аффинное связывание пептида с молекулой MHC. Данная модель, обученная на наборе данных из 191975 связанных и 191621 несвязанных пар эпитоп-MHC, продемонстрировала исключительно высокую производительность на тестовой выборке (ROC AUC 99,06%).

Для агрегации результатов моделей, предсказывающих этапы процессинга, и построения итоговой метамодели был выбран алгоритм случайного леса (Random Forest). Его выбор обусловлен лучшими показателями производительности на тестовом наборе данных (AUC 98,66% и F1 96,02%) в сравнении с такими альтернативными подходами, как градиентный бустинг (98,59% и 94,16%) и XGBoost (98,62% и 94,34%).

Разработанная метамодель прошла проверку путём сравнения с широко используемыми и авторитетными программами NetMHCpan [8], MHCflurry [9] и DeepMHCI [4]. Результаты сравнительного анализа, представленные на рис. 1, демонстрируют превосходство подхода на основе метамодели. Следует отметить, что размер тестового набора (N на графике) варьировал для каждой программы, так как из него были исключены все данные, пересекающиеся с их обучающими наборами.

Для сравнительной оценки вклада каждого компонента в итоговый результат был применён метод SHAP, результаты которого представлены на рис. 2. Анализ показал, что модель TranspMHC вносит наибольший вклад в предсказание. Вместе с тем, модели, описывающие другие этапы процессинга, также являются значимыми, повышая общую прогностическую способность. Таким образом, интеграция моделей всех этапов процессинга антигена позволяет получить более полную и надёжную оценку потенциала пептида к презентации, что является критически важным для дальнейшего прогнозирования его иммуногенности.

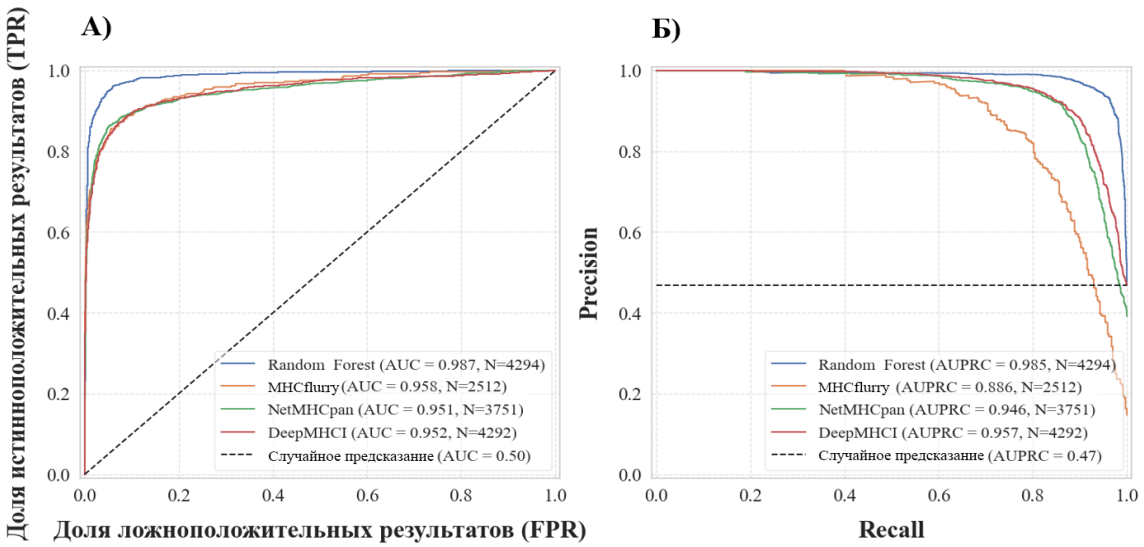

Рис. 1. ROC (А) и PR (Б) кривые сравнения производительности разработанной модели презентации на основе случайного леса с NetMHCpan, mhcflurry и DeepMHCI

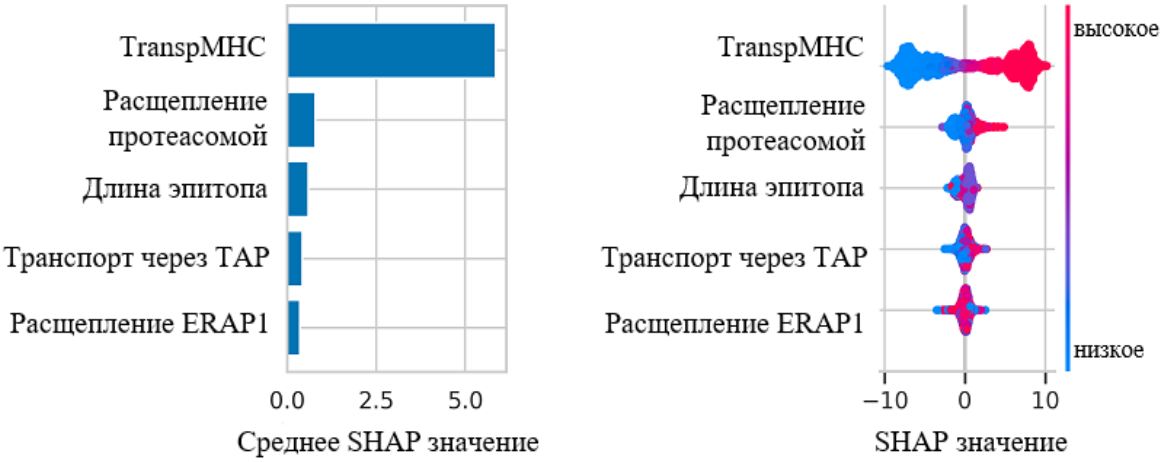

Рис. 2. SHAP-анализ для входных признаков метамодели презентации

3.2. Комплексная метамодель предсказания иммуногенности

Точная оценка презентации эпитопа является необходимым, но недостаточным условием для определения его иммуногенности. Далеко не каждый пептид, представленный на поверхности клетки, способен вызвать иммунный ответ. Поэтому следующим шагом стала разработка отдельной метамодедели, нацеленной на предсказание иммуногенности.

Учитывая, что модель TranspMHC показала себя как ключевой компонент в прогнозировании презентации, было принято решение проверить её потенциал и для оценки иммуногенности. Для этого TranspMHC была дополнительно обучена на двух наборах данных: Neo и IE, структура которых представлена на рис. 3. Обе полученные модели продемонстрировали высокую и сопоставимую производительность, с показателем AUC

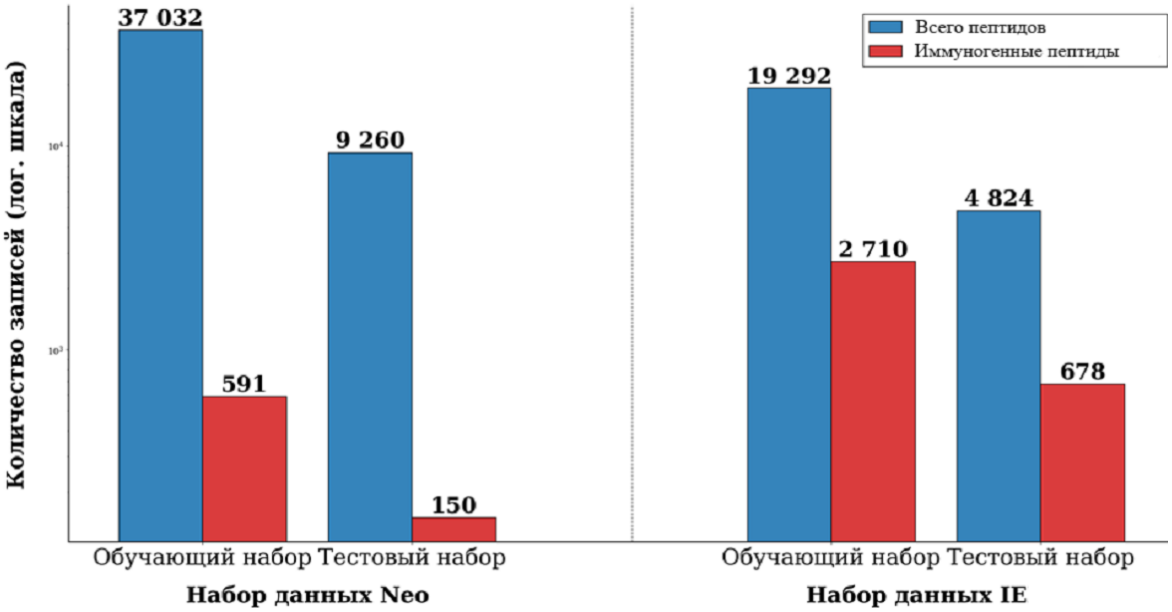

Рис. 3. Структура наборов для предсказания иммуногенности моделью TranspMHC

94,59% на наборе Neo и 94,11% на IE. Однако модель, обученная на данных IE, показала вдвое большую чувствительность (83,43%) по сравнению с Neo (44,67%), что отражает различия в этих наборах данных.

Затем для создания итоговой комплексной метамодели, названной IMM (ImmunoMetaModel), был применён подход, аналогичный построению метамодели презентации. В качестве входных признаков были объединены: предсказание метамодели презентации, предсказания обеих версий TranspMHC (Neo и IE), предсказания моделей, оценивающих секрецию цитокинов, и физико-химические свойства пептидов.

Для определения наиболее значимых признаков был проведён анализ с использованием метода SHAP, результаты которого представлены на рис. 4. Исходя из полученных результатов, наибольший вклад в прогнозирование вносят предсказание метамодели презентации, предсказания обеих моделей TranspMHC, индекс чужеродности пептида, а также прогнозы индукции секреции интерферона-гамма и фактора некроза опухолей альфа.

На основе вклада отобранных признаков было создано три варианта итоговой IMM для определения оптимального соотношения производительности и количества признаков: imm_plus – использует все 16 признаков, imm – использует топ-6 признаков, imm_mini – использует топ-5 признаков без индекса чужеродности. По результатам даже самая компактная модель IMM продемонстрировала высокую производительность – AUC 92,7%, что всего на 2,5% меньше, чем у самой полной версии модели.

Для подтверждения эффективности разработанного подхода все три варианта IMM прошли сравнительный анализ с существующими

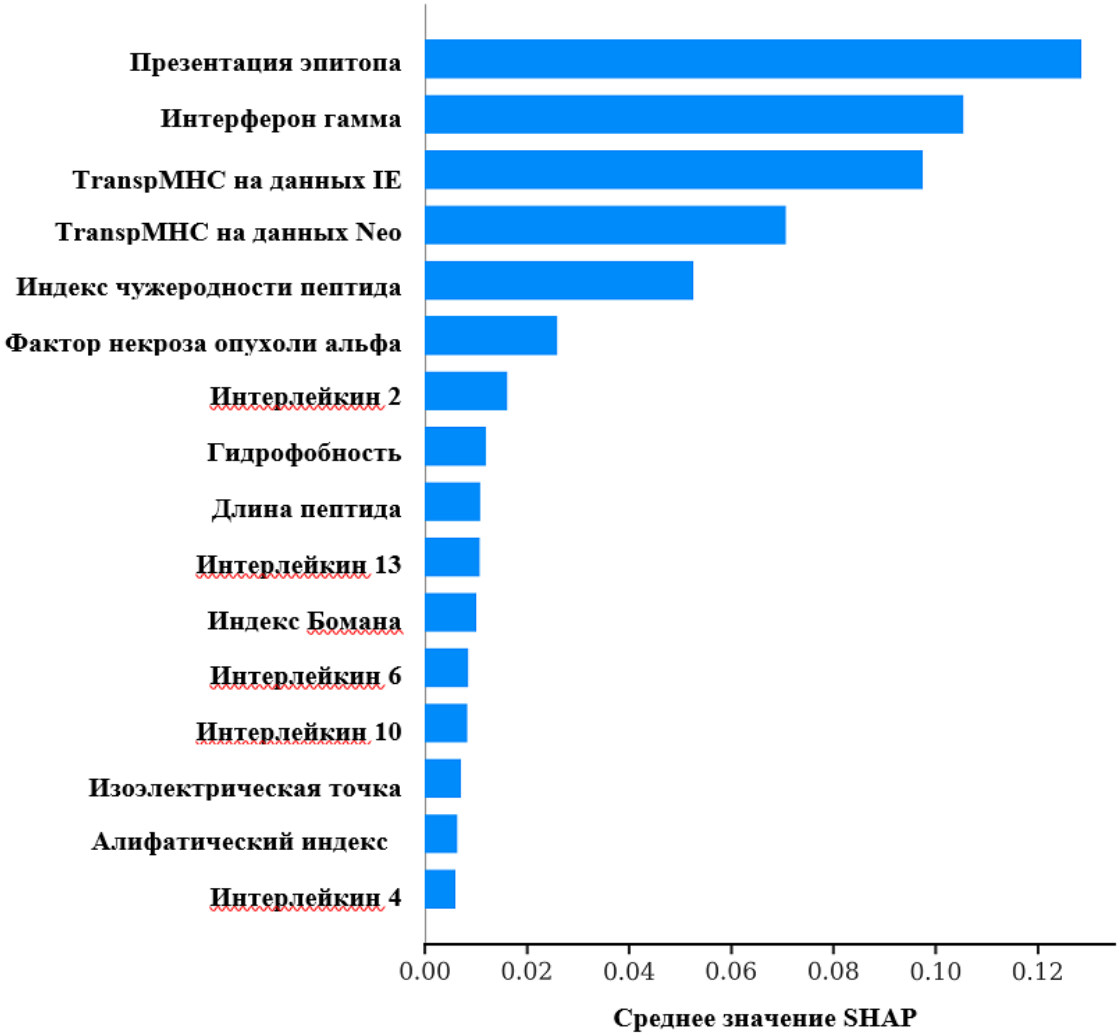

Рис. 4. SHAP-анализ для входных признаков метамодели иммуногенности

программными решениями DeepHLApan [1], DeepImmuno [3] и IEPAPI [6]. Результаты, представленные на рис. 5, демонстрируют значительное превосходство разработанных IMM. Разница в производительности между imm_plus и моделью IEPAPI составила 21,1% по метрике AUC и 56,1% по AUCPR. Высокий показатель AUCPR особенно важен, так как он свидетельствует о способности метамодели эффективно идентифицировать иммуногенные пептиды даже в условиях сильного дисбаланса классов, что является критически важным для решения задач, связанных с поиском неоантигенов в персонализированной иммунотерапии.

4. Заключение

В ходе работы были созданы две метамодели: для предсказания презентации эпитопа на клеточной поверхности и для определения его иммуногенности. Полученные метамодели были сравнены с известными

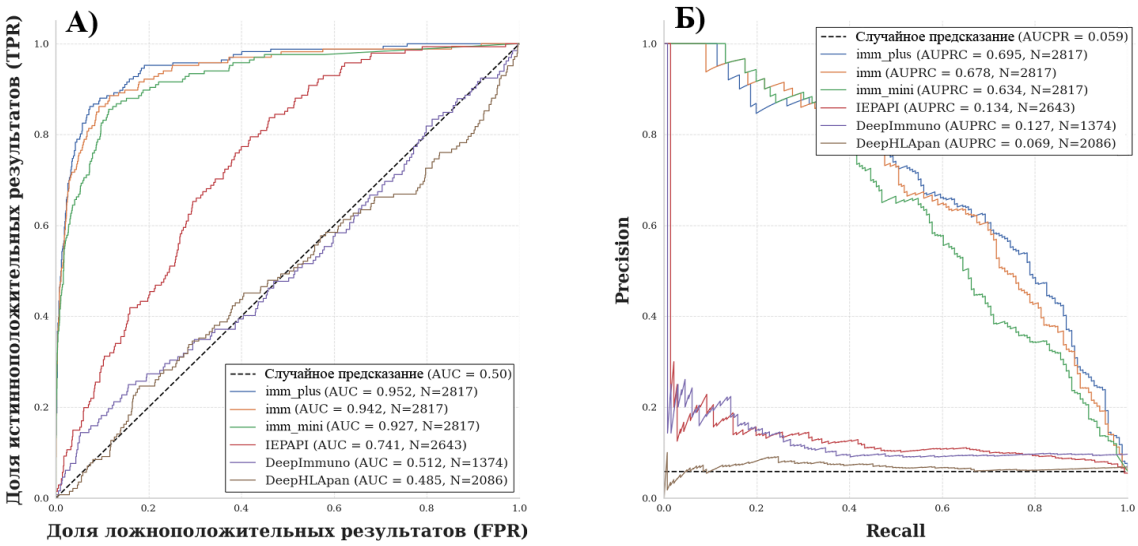

Рис. 5. ROC (А) и PR (Б) кривые сравнения производительности модели иммуногенности на основе случайного леса с IEPAPI, DeepHLApan и DeepImmuno

аналогами. По результатам сравнения на независимых наборах данных обе метамодели продемонстрировали наилучшую производительность в своих категориях, превзойдя ближайших конкурентов в оценке иммуногенности более чем на 21% по метрике AUC и на 56% по AUCPR.

Таким образом, результаты данной работы демонстрируют, что преодоление существующих ограничений в прогнозировании иммуногенности неоантигенов требует перехода от узкоспециализированных моделей к комплексному и многоуровневому анализу. Разработанный подход доказывает, что объединение признаков, связанных с презентацией и узнаванием иммунной системой неоантигена, при помощи метамоделей является оправданным и позволяет достигать высокой предсказательной точности.

Дальнейшее развитие этого подхода может основываться на включении дополнительных факторов, таких как особенности опухолевого микроокружения, что будет способствовать открытию прямых перспектив для ускорения перехода персонализированной иммунотерапии из области научных изысканий в стандартную клиническую практику.

Библиографические ссылки

1. A deep learning approach for neoantigen prediction considering both HLA-peptide binding and immunogenicity / J. Wu [et al.] // Front. Immunol. 2019. Vol. 10. Article no. 2559.

2. Attention-aware differential learning for predicting peptide-MHC class I binding and T cell receptor recognition / R. Niu [et al.] // Brief. Bioinform. 2024. Vol. 26. Article no. bbaf038.

3. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity / G. Li [et al.] // Brief. Bioinform. 2021. Vol. 22. Article no. bbab160.

4. DeepMHCI: an anchor position-aware deep interaction model for accurate MHC-I peptide binding affinity prediction / W. Qu [et al.] // Bioinformatics. 2023. Vol. 399. Article no. btad551.

5. Driver mutation landscape of acute myeloid leukemia provides insights for neoantigen-based immunotherapy / P. Jin [et al.] // Cancer Lett. 2025. Vol. 611. Article no. 217427.

6. IEPAPI: a method for immune epitope prediction by incorporating antigen presentation and immunogenicity / J. Deng [et al.] // Brief. Bioinform. 2023. Vol. 24. Article no. bbad171.

7. Meta learning for mutant HLA class I epitope immunogenicity prediction to accelerate cancer clinical immunotherapy / L. Xu [et al.] // Brief. Bioinform. 2024. Vol. 26. Article no. bbae625.

8. NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data / V. Jurtz [et al.] // J. Immunol. Baltim. Md 1950. 2017. Vol. 199. P. 3360–3368.

9. *O'Donnell T. J., Rubinsteyn A., Laserson U.* MHCflurry 2.0: Improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing // Cell Syst. 2020. Vol. 11. P. 42–48.

10. *Richman L. P., Vonderheide R. H., Rech A. J.* Neoantigen dissimilarity to the self-proteome predicts immunogenicity and response to immune checkpoint blockade // Cell Syst. 2019. Vol. 9. P. 375–382.