Оригинальная статья / Original article

https://doi.org/10.21869/2223-1536-2025-15-3-122-141

УДК 004

Оценка точности методов контроля частоты ложных срабатываний при аннотации спектра de novo

М. М. Тевяшов[1] ✉

[1] Санкт-Петербургский государственный экономический университет
наб. канала Грибоедова, д. 30-32, г. Санкт-Петербург 191023, Российская Федерация

✉ e-mail: tukaramm@yandex.ru

Резюме

Цель *– сравнение подходов на основе машинного обучения (deep learning) и классических методов по качеству аннотации масс-спектров в условиях больших данных, а также выявление оптимального сценария их интеграции.*

Методы. *Исследование базируется на использовании набора данных PXD004452, содержащего 2,5 млн уникальных пептидов.*

Разработана схема взаимодействия на основе Python/TensorFlow/PyTorch, который обеспечивает параллельную обработку пептидных спектров на GPU-кластере. Использованы следующие этапы: фильтрация топ-150 пиков по интенсивности; генерация теоретических B-/Y-ионов с учетом модификаций; предсказание пептидов (PepNet – сверточная+рекуррентная сеть; Tide-search – индексная перехеширующая стратегия). Метрики: количества совпадений, дельта-масса, расстояние Левенштейна, ROC-кривые, распределение ошибок.

Результаты. *PepNet требует значительных вычислительных ресурсов, при этом качество предсказаний уступает Tide-search, особенно на длинных пептидах и модификациях (~среднее совпадение: 4,2 пика vs 9,7; $p < 0,001$). Однако PepNet лучше показывает себя при тех спектрах, где в database search отсутствуют релевантные последовательности, демонстрируя важную способность выявлять novel-пептиды. Распределение расстояния Левенштейна: ~30% – полное совпадение (0); ~52% – небольшое отклонение (1–5); остальное – значительные расхождения (>5).*

Заключение. *Метод deep learning (PepNet) демонстрирует перспективы, но без интеграции с database search уступает по точности.*

Предлагается гибридная архитектура: pep-tagging через PepNet, затем уточнение и верификация через database search. Такой конвейер на больших данных позволит сочетать открытие новых пептидов (de novo) и высокую достоверность идентификации (database search).

Ключевые слова: *машинное обучение; масс-спектрометрия; расстояние Левенштайна; спектры.*

Благодарности: *Выражаем благодарность заведующему научно-учебной лабораторией искусственного интеллекта для вычислительной биологии НИУ ВШЭ доктору наук Кертес-Фаркаш Аттила за формирование методологии исследования и предоставление данных для статьи.*

Конфликт интересов: *Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.*

Для цитирования: Тевяшов М. М. Оценка точности методов контроля частоты ложных срабатываний при аннотации спектра de novo // Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение. 2025. Т. 15, № 3. С. 122–141. https://doi.org/10.21869/2223-1536-2025-15-3-122-141

Поступила в редакцию 21.07.2025 *Подписана в печать 20.08.2025* *Опубликована 30.09.2025*

Evaluation of the accuracy of false alarm frequency control methods for de novo spectrum

Mikhail M. Tevyashov[1] ✉

[1] Saint Petersburg State University of Economics
30-32 Griboedov canal Emb., St. Petersburg 191023, Russian Federation

✉ e-mail: tukaramm@yandex.ru

Abstract

The purpose of the research *is comparison of machine learning-based approaches (deep learning) and classical methods for mass spectrum annotation in big data conditions, as well as identification of the optimal scenario for their integration.*

Methods. *The study is based on the PXD004452 dataset containing 2,5 million unique peptides. An interaction scheme based on Python/TensorFlow/PyTorch has been developed, which provides parallel processing of peptide spectra on a GPU cluster. The following steps were used: filtering of the top 150 peaks by intensity; generation of theoretical B-/Y-ions, taking into account modifications; prediction of peptides (PepNet – convolutional+recurrent network; Tide-search – index-shifting strategy). Metrics: number of matches, delta mass, Levenshtein distance, ROC curves, error distribution.*

Results. *PepNet requires significant computational resources, while the prediction quality is inferior to Tide-search, especially for long peptides and modifications (~average match: 4,2 pi vs 9,7; p < 0,001). However, PepNet performs better in those spectra where relevant sequences are missing in the database search, demonstrating an important ability to identify novel peptides. Levenshtein distance distribution: ~30% is a complete match (0); ~52% is a small deviation (1-5); the rest is significant discrepancies (>5).*

Conclusions. *The deep learning (PepNet) method shows promise, but without integration with database search, it is inferior in accuracy. A hybrid architecture is proposed: pep-tagging via PepNet, followed by refinement and verification via database search. Such a big data pipeline will combine the discovery of new peptides (de novo) and high identification reliability (database search).*

Keywords: *machine learning; mass spectrometry; Levenstein distance; spectra.*

Acknowledgements: *We would like to thank Dr. Kertes-Farkash Attila, Head of the HSE Scientific and Educational Laboratory of Artificial Intelligence for Computational Biology, for developing the research methodology and providing data for the article.*

Conflict of interest: *The Authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.*

For citation: Tevyashov M.M. Evaluation of the accuracy of false alarm frequency control methods for de novo spectrum. *Izvestiya Yugo-Zapadnogo gosudarstvennogo universiteta. Serija: Upravlenie, vychislitel'naja tekhnika, informatika. Meditsinskoe priborostroenie = Proceedings of the Southwest State University. Series: Control, Computer Engineering, Information Science. Medical Instruments Engineering.* 2025;15(3):122–141. https://doi.org /10.21869/2223-1536-2025-15-3-122-141

Received 21.07.2025 *Accepted 20.08.2025* *Published 30.09.2025*

***

Введение

Масс-спектрометрия играет решающую роль в протеомике, особенно в определении точного аминокислотного состава пептидов. Существует два основных подхода к интерпретации

спектров: database search и de novo секвенирование.

Первый подход основан на сравнении экспериментальных спектров с теоретически предсказанными спектрами известных пептидов. Второй подход (de novo) пытается реконструировать аминокислотную последовательность непосредственно из спектра без предварительного знания последовательности в базе данных. В ходе работы было решено сравнить de novo и database search, поскольку оба подхода имеют свои сильные и слабые стороны [1].

Поиск в базе данных – это сравнение экспериментальных данных, полученных в результате анализа образца белка с помощью масс-спектрометрии, с теоретическими спектрами, смоделированными на компьютере на основе известных последовательностей аминокислот. Поиск в базе данных часто более точен, но зависит от полноты и релевантности базы данных [2].

Секвенирование пептидов de novo – это метод реконструкции аминокислотной последовательности пептида непосредственно из его тандемного масс-спектра (MS/MS) без обращения к базам данных белков. В отличие от традиционного поиска, который осуществляет поиск совпадений между экспериментальным спектром и теоретическими спектрами известных белков, секвенирование de novo полностью полагается на интерпретацию фрагментных ионов, полученных при фрагментации пептида в масс-спектрометре. Процесс основан на том факте, что при фрагментации пептида образуются ионы определенного типа (чаще всего b- и y-ионы), которые соответствуют последовательным разрывам связей аминокислот. Анализируя массу ионов [3], можно определить, какие аминокислоты находятся в последовательности и в каком порядке [4].

Методы de novo не требуют применения базы данных и позволяют находить новые или измененные пептиды, не представленные в референтных базах данных, но подвержены ошибкам в интерпретации спектров, но страдают от низкой точности и плохого качества спектра, неполной фрагментации, длинных пептидов, пропущенных ионов, перекрывающихся пиков, шумов [5].

Основная идея работы заключается в опровержении либо доказательстве гипотезы о том, что метод de novo менее точен в своих предсказаниях, чем database search при прямом сравнении. Для этого была выбрана модификация de novo под названием PepNet.

PepNet – это полностью сверточная нейросеть (fully convolutional neural network), специально разработанная для анализа MS/MS-спектров и прямого восстановления аминокислотных последовательностей пептидов. На вход модели поступает спектр, представленный как одномерный вектор: пики m/z бинируются с разрешением ~0,1 Th по диапазону от 0 до 2000 m/z (около 20 000 измерений), после чего нормализуются по максимальной интенсивности. Предварительно из спектра удаляется пиковый

сигнал прекурсора, чтобы исключить его влияние на обучение и предсказание.

Архитектура (рис. 1) состоит из чередующихся temporal convolutional network (TCN) слоёв и down-sampling операций, которые извлекают как глобальные, так и локальные признаки спектра. Получившийся тензор признаков затем проходит через декодер, который рекурсивно формирует аминокислотную последовательность. Модель обучалась на ~3 млн HCD-спектров из различных библиотек (преимущественно человек), что позволило ей хорошо обобщать разнообразные входные данные, включая спектры нечеловеческих организмов и спектры неотмеченных прекурсоров.

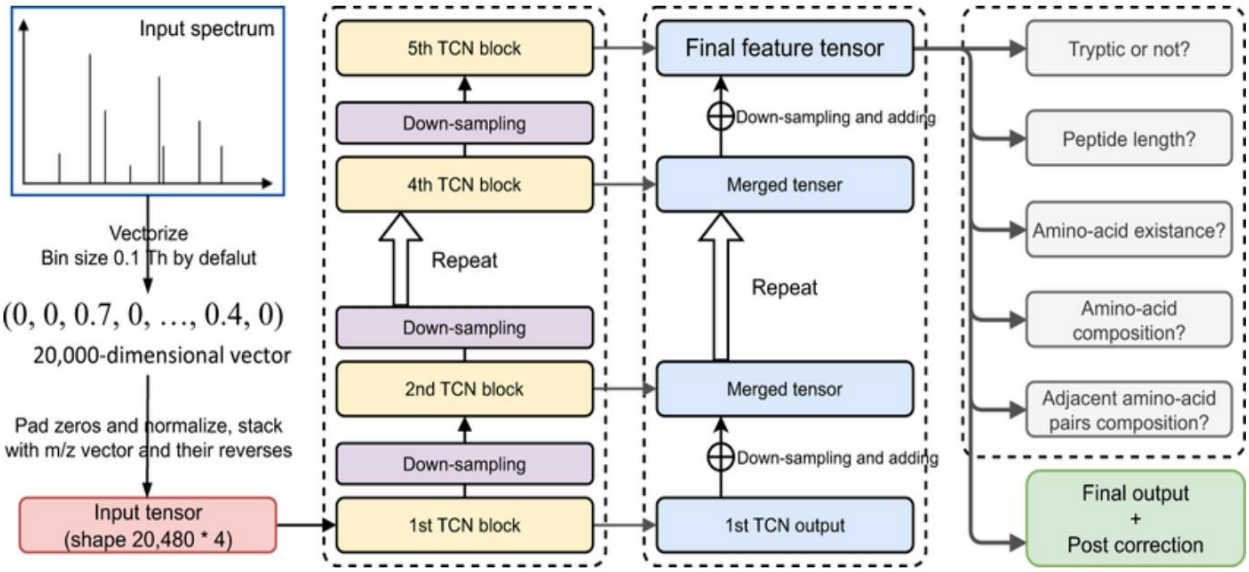

Рис. 1. Схема архитектуры PepNet

Fig. 1. PepNet architecture diagram

По сравнению с другими de novo-сетями (PointNovo, DeepNovo) PepNet демонстрирует значительное улучшение точности на уровне полного пептида и на уровне локальных позиций. Он способен секвенировать спектры, которые остаются нераспознанными при использовании database search-методов, что особенно важно для выявления новых, модифицированных или ранее неизвестных пептидов. Кроме того, благодаря эффективной сверточной архитектуре PepNet работает примерно в 3–7 раз быстрее, чем DeepNovo и PointNovo на GPU, что делает его применимым в обработке масштабных ассемблей протеомных данных.

Эти особенности делают PepNet мощным инструментом для высокоточного de novo секвенирования, особенно в задачах, где отсутствуют полные справочные базы пептидов, например в метапротеомике, исследовании редких белков или иммунопептидомике [5].

В качестве database search метода была выбрана модификация tide-search. **Tide** – это высокопроизводительная реализация алгоритма SEQUEST, созданная

в рамках проекта Crux, которая значительно ускоряет поиск пептидов в tandem MS-спектрах, сохраняя XCorr-оценки идентичными исходному SEQUEST. Основной рабочий процесс начинается с построения индексированной базы пептидов (через tide-index) из FASTA-файла белков, что позволяет многократно и эффективно переиспользовать индекс при множественных поисках. Затем tide-search обрабатывает спектры (вход в бинарном формате либо через ProteoWizard), сравнивает каждый спектр с теоретическими спектрами кандидатов из базы данных и выдает список PSM (peptide-spectrum matches) с XCorr-оценками и метками результата (target/decoy).

Ввиду оптимизации алгоритма и программной реализации Tide обеспечивает примерно 170-кратное ускорение по сравнению с оригинальным SEQUEST: около 1550 спектров/с на стандартном Xeon-процессоре, тогда как SEQUEST с тем же индексированием обрабатывает лишь около 8,8 спектров/с Современные улучшения индексации (версия 2023) существенно снизили потребление CPU и RAM, что делает Tide способной обрабатывать очень большие базы данных вместе с учётом посттрансляционных модификаций. Кроме того, Crux включает дополнительные команды для повышения надежности результатов: cascade-search для итеративного поиска по суббазам с более строгим FDR и average target-decoy competition (aTDC) для снижения дисперсии оценки достоверности при работе с редкими пептидами.

Эти особенности делают Tide-search идеальным инструментом для высокоточечного и производительного database search, особенно в крупномасштабных исследованиях, где важны скорость и надежность идентификации пептидов.

В масс-спектрометрии пептидов основным методом получения информации о последовательности является фрагментация. При этом процессе пептид разрывается по пептидным связям, и образуются фрагменты, называемые фрагментными ионами. Из них для интерпретации спектров наиболее важны два типа ионов – B-ионы и Y-ионы. B-ионы возникают при сохранении N-конца пептида, их фрагмент состоит из первых n аминокислот последовательности. Y-ионы возникают при сохранении C-конца пептида, соответственно, их фрагмент состоит из последних n аминокислот [7].

Пример: если пептид = ACDE,
то B-ион 2 = AC и Y-ион 2 = DE.

Для каждого пептида были рассчитаны теоретические B-ионы и Y-ионы, которые соответствуют фрагментам, образующимся при разрыве пептидной связи. Каждый ион имеет свою массу (включая модификации), которая сравнивается с экспериментальными значениями в спектре. Каждая аминокислота имеет свою моноизотопную массу (в дальтонах), и при образовании ионов к этим массам добавляются константы, связанные с типом и зарядом иона.

Формула для вычисления B-ионов:

$$m(b_n) = \sum_{i=1}^{n} m(AA_i) + m(\text{N-terminus}) + m(\text{proton}),$$

где $m(AA_i)$ – масса каждой аминокислоты во фрагменте; m(N-конец) = 1,0078 Da (масса протона); m(ион) = 0 (обычно, если не учитывать конкретные фрагменты); потери (например, –H2O, –NH3) также можно учитывать как модификации.

Формула для вычисления Y-ионов:

$$m(y_n) = \sum_{i=1}^{N} m(AA_i) + m(\text{C-terminus}) + m(H_2O) + m(\text{proton}),$$

где $m(AA_i)$ – масса каждой аминокислоты во фрагменте; m(С-конец) = 1,0078 Da (масса протона); m(ион) = 0 (обычно, если не учитывать специфические фрагменты); $m(H_2O)$ = 18,0106 Да (С-концевой фрагмент включает группы –OH и –H).

Пример рассчета:

Пептид: ACD

A = 71,0371; C = 103,0092; Д = 115,0269

H_2O = 18,0106 Протон (H^+) = 1,0078

b1 = 71,0371 + 1,0078 = 72,0449 Da

b2 = 71,0371 + 103,0092 + 1,0078 = 175,0541 Da

b3 = 71,0371 + 103,0092 + 115,0269 + 1,0078 = 290,081 Da

y1 = 115,0269 + 18,0106 + 1,0078 = 134,0453 Da

y2 = 115,0269 + 103,0092 + 18,0106 + 1,0078 = 237,0545 Da

y3 = 115,0269 + 103,0092 + 71,0371 + 18,0106 + 1,0078 = 344,0916 Da

Полученные масс-спектры содержат информацию о тысячах ионов с различной интенсивностью. Однако не все пики предоставляют полезную информацию – спектры могут содержать шум, фоновые сигналы или пики малой интенсивности, не несущие смысловой нагрузки. Для увеличения сигнала и устранения шума было решено для каждого спектра отобрать 150 самых интенсивных пиков. Такая фильтрация позволяет оставить только наиболее вероятных кандидатов на сопоставление теоретически рассчитанных фрагментов.

Далее для каждого теоретического иона ищется ближайший пик из 150 лучших экспериментальных пиков (рис. 2). Сопоставление считается успешным, если разница между теоретической и экспериментальной массой не превышает установленного допуска по массе. В данной работе было решено принять допуск ±0,05 Da (дальтон), что обеспечивает баланс между чувствительностью и специфичностью. Совпадения регистрируются в виде пары (тип иона, номер иона, теоретическая масса, экспериментальная масса, дельта). Если в пептиде присутствуют модификации, то они обязательно учитываются при расчете массы соответствующего иона.

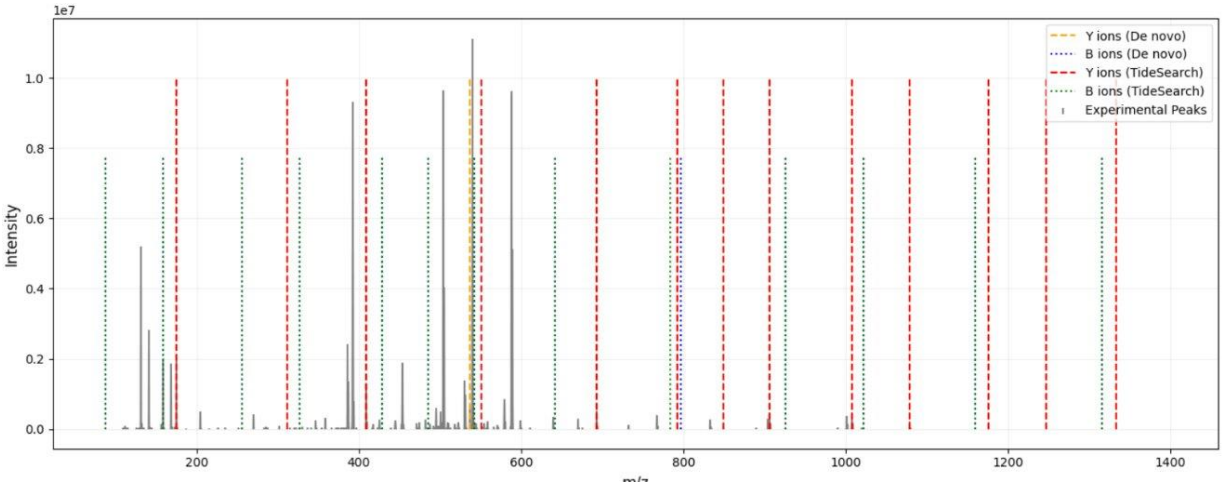

Рис. 2. 150 лучших пиков, а также ионы B и Y, рассчитанные с использованием методов De novo и Tide Search для сканирования 950

Fig. 2. Top 150 peaks, plus B-ions and Y-ions calculated using De novo and Tide Search methods for scan 950

После сравнения теоретических и экспериментальных пиков возникла необходимость количественно оценить, насколько хорошо предсказанные пептиды (например, из секвенирования de novo) соответствуют реальным спектрам. Одной из основных метрик точности является количество совпадений между теоретическими и экспериментальными пиками (рис. 3). Такие совпадения означают, что фрагмент (B-ион или Y-ион), предсказанный на основе последовательности пептида, фактически наблюдается в экспериментальном спектре. Эти значения позволяют напрямую сравнить, насколько хорошо работает метод de novo по сравнению с database search.

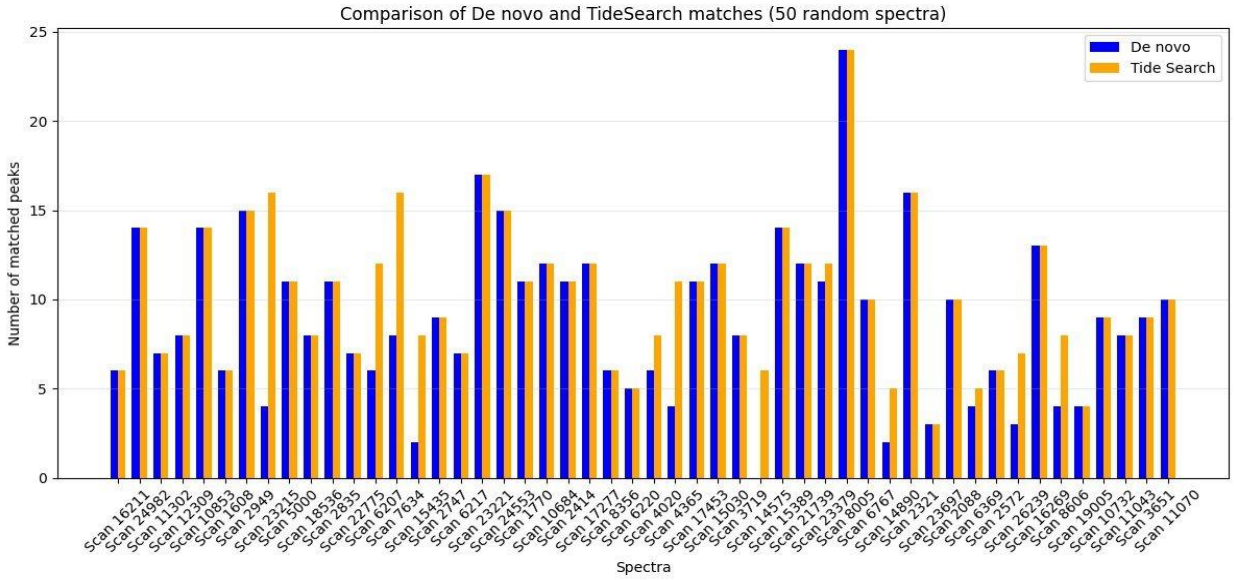

Рис. 3. Сравнение результатов De novo и Tide Search (50 случайных спектров)

Fig. 3. Comparison of De novo and Tide Search matches (50 random spectra)

Гистограммы, полученные на основе совпадений теоретических пиков (рис. 4 и 5), показывают, что для метода tide-search преобладают значения показателя от 6 до 15. В то же время для de novo результат несколько хуже, так как большее значение имеют случаи совпадения только от 1 до 5 пиков, а более высоких результатов – существенно меньше. Особенно явно видно различие при совмещении гистограмм (гистограмма 6). Определенные выводы можно сделать уже на этом этапе анализа результатов, но было решено углубиться в показатели точности обоих методов, и для этого была выбрана еще одна метрика.

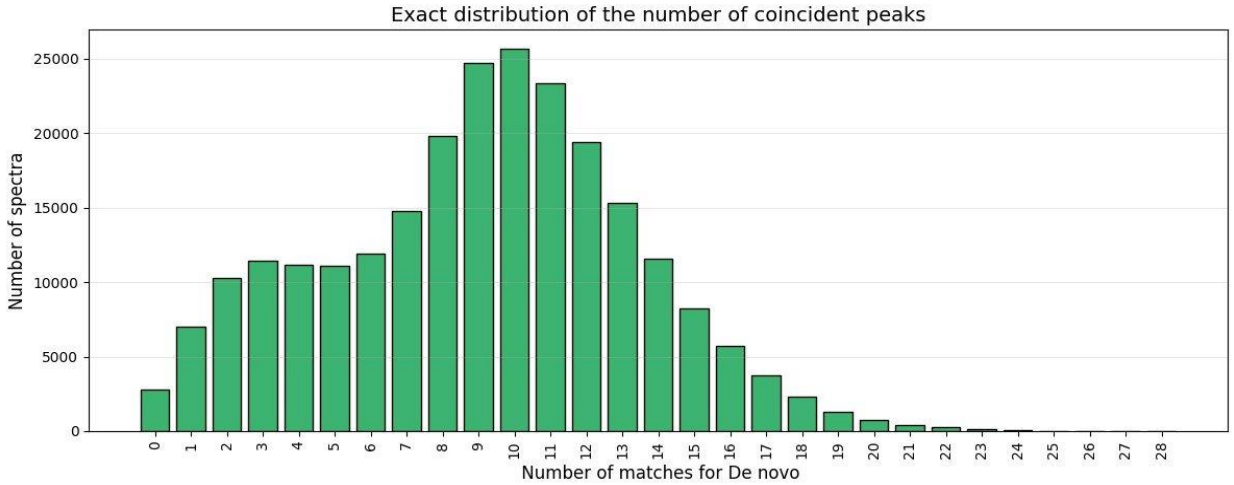

Рис. 4. Совпадения теоретически рассчитанных методом de novo пиков с экспериментальными

Fig. 4. Coincidence of calculation methods of new peaks with experimental ones

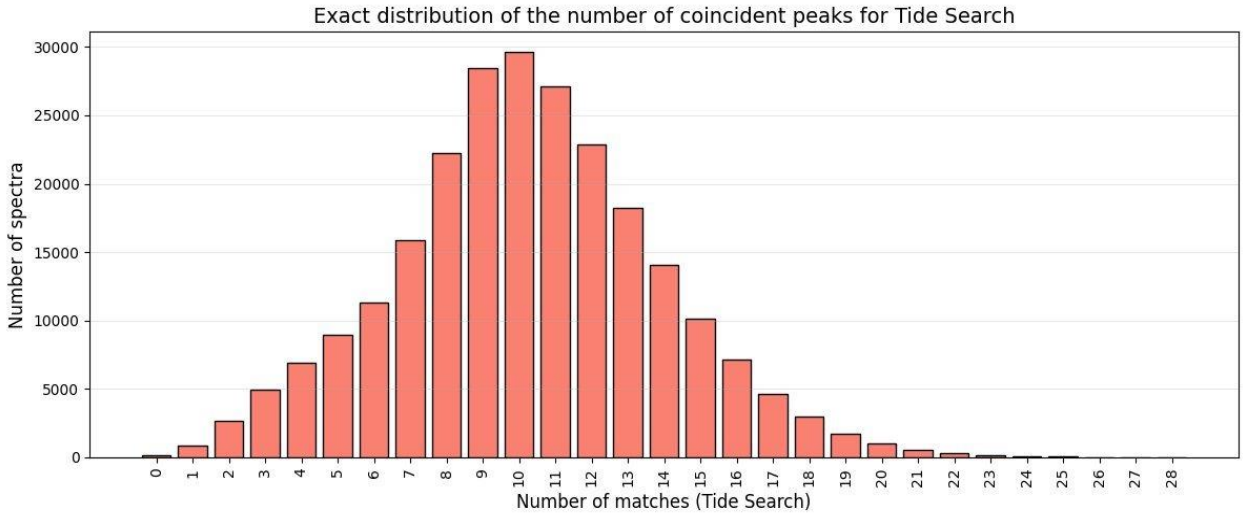

Рис. 5. Совпадения теоретически рассчитанных методом database search пиков с экспериментальными

Fig. 5. Coincidence of theoretically calculated peaks using the database search method with experimental ones

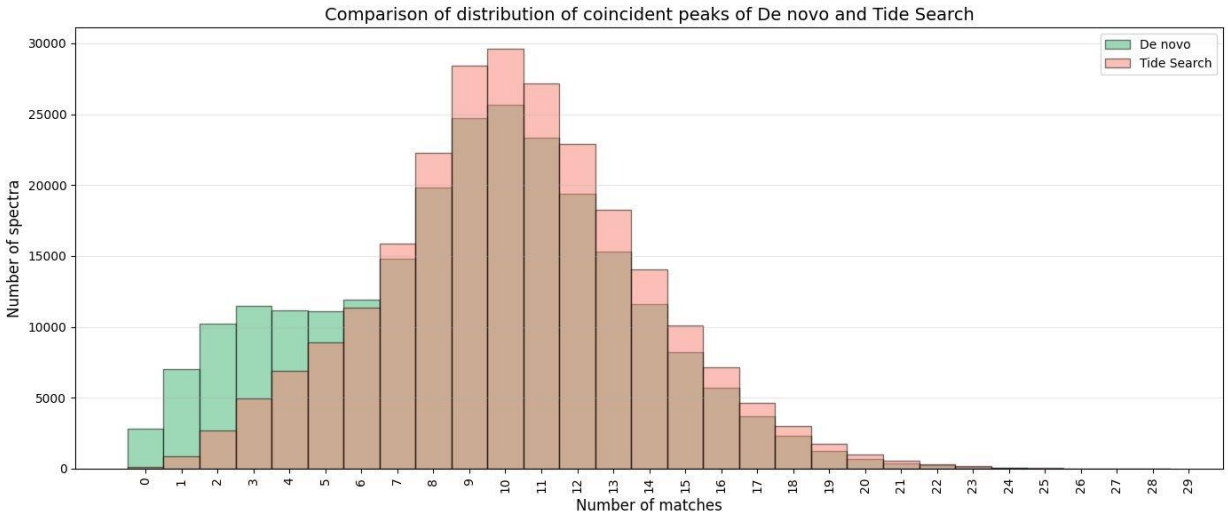

Рис. 6. Сравнение совпадений теоретически рассчитанных и экспериментальных пиков для обоих методов

Fig. 6. Comparison of the coincidences of theoretically calculated and experimental peaks for both methods.

Для более точной оценки различий между предсказанными и референтными последовательностями используется расстояние Левенштейна (также известное как расстояние редактирования) – минимальное количество операций (вставок, делеций, замен), необходимых для преобразования одной строки в другую [8]. В данном случае для преобразования пептида, рассчитанного методом de novo в пептид, предсказанный при помощи database search, т. е. данная метрика отражает то, насколько разными являются предсказанные пептиды (чем больше расстояние Левенштайна, тем больше отличается пептид). Эта метрика особенно полезна при анализе качества секвенирования de novo, где даже небольшие ошибки в аминокислотах могут сильно повлиять на биологическое значение пептида.

Гистограмма с общим распределением по расстоянию Левенштейна (рис. 7) показывает, что подавляющее большинство значений принадлежит интервалу от 0 до 5. При этом 30% составляют значения 0, т. е. значения полного совпадения пептидов. И еще 52% – это значения от 1 до 5, что в ходе исследования было решено принять за небольшое отклонение.

Отдельно хотелось бы остановиться на нескольких гистограммах (рис. 8–13), где для каждого значения расстояния Левенштейна (от 0 до 42) были отфильтрованы строки, где соответствующее расстояние редактирования между предсказанной и референтной последовательностями было равно этому значению. Таким образом, для каждого значения расстояния были выбраны отдельные подмножества спектров, для которых сравнивалась точность предсказания. Как видно ниже, на каждом изображении одна гистограмма отображает результаты для метода de novo, а другая – для метода database search. По этим гистограммам видно, сколько совпадений

было найдено для каждой группы спектров и насколько точны методы в зависимости от значения расстояния Левенштейна. Ниже приведены 6 гистограмм из 43, которые достаточно наглядно отображают общую тенденцию, которая прослеживается во всех гистограммах.

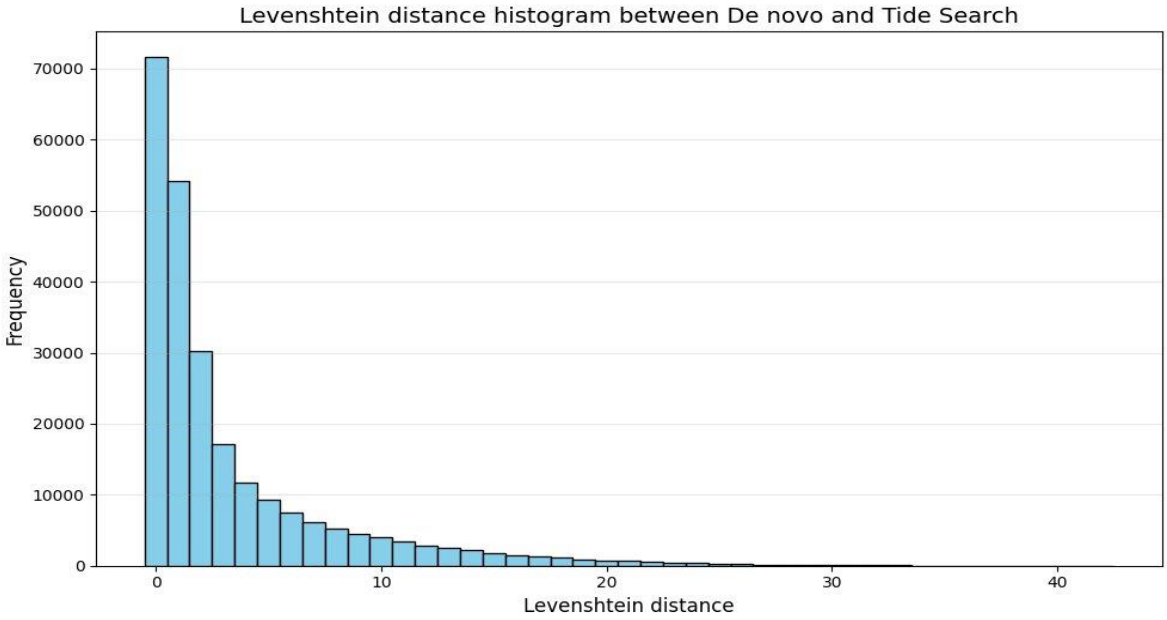

Рис. 7. Распределение спектров по расстоянию Левенштайна (схожести / различию между пептидами, предсказанными исследуемыми методами)

Fig. 7. Distribution of spectra by the Levenshtein distance (similarity / difference between peptides predicted by the studied methods).

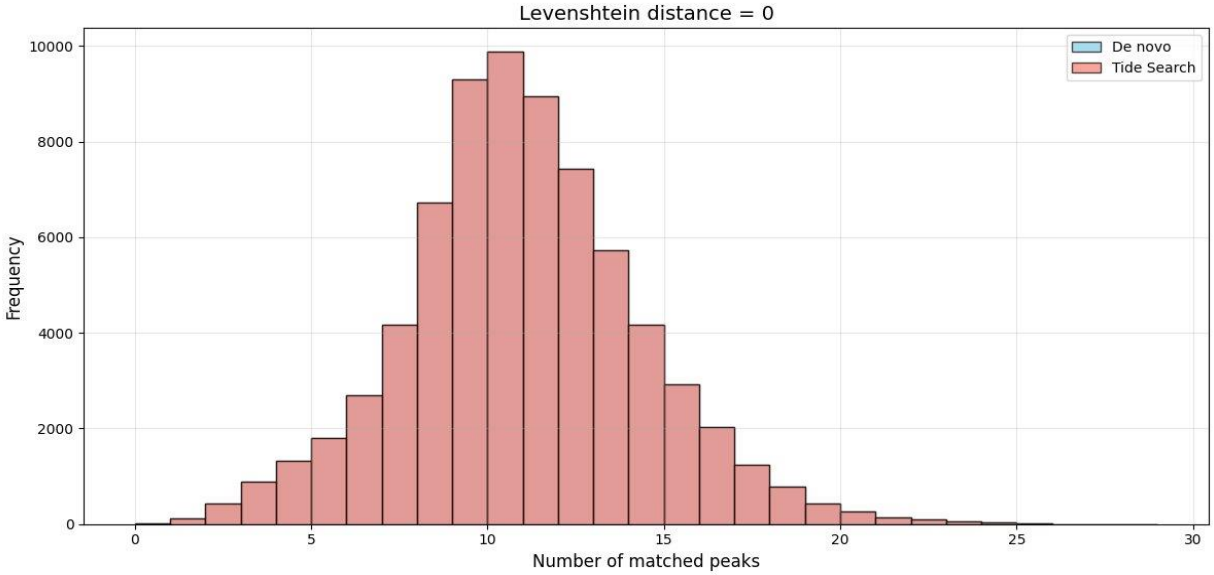

Рис. 8. Сравнение распределения совпадающих пиков для de novo и database search с расстоянием Левенштейна, равным 0

Fig. 8. Comparison of the distribution of matching peaks for de novo and database search with Levenshtein distance equal to 0

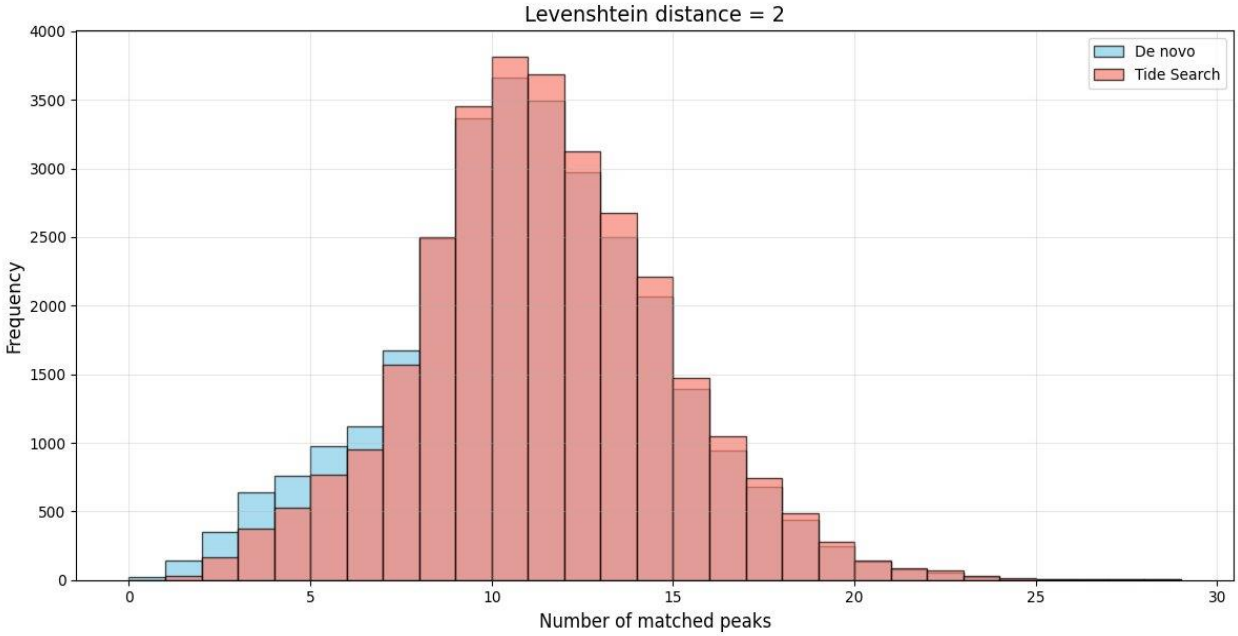

Рис. 9. Сравнение распределения совпадающих пиков для de novo и database search с расстоянием Левенштейна, равным 2

Fig. 9. Comparison of the distribution of matching peaks for de novo and database search with Levenshtein distance equal to 2

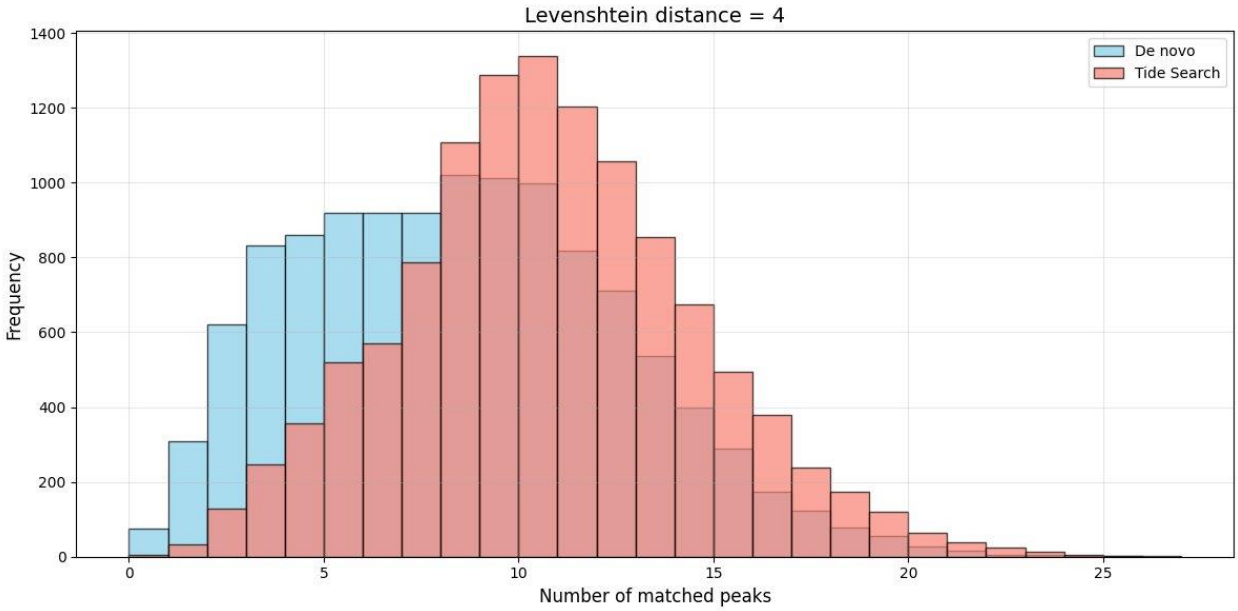

Рис. 10. Сравнение распределения совпадающих пиков для de novo и database search с расстоянием Левенштейна, равным 4

Fig. 10. Comparison of the distribution of matching peaks for de novo and database search with Levenshtein distance equal to 4

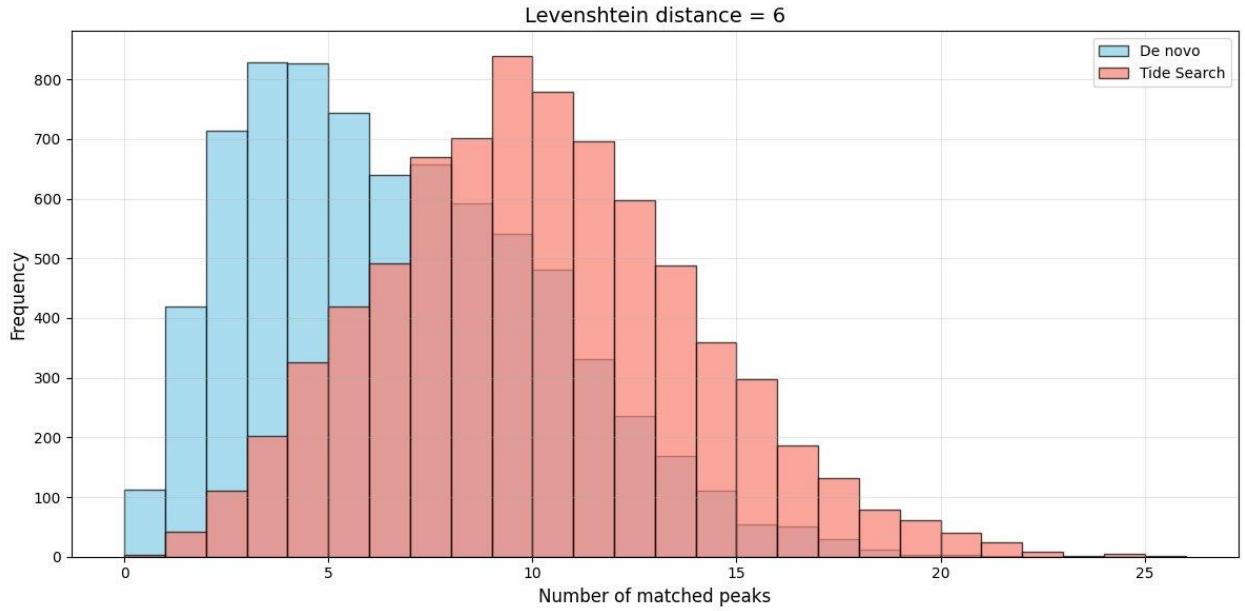

Рис. 11. Сравнение распределения совпадающих пиков для de novo и database search с расстоянием Левенштейна, равным 6

Fig. 11. Comparison of the distribution of matching peaks for de novo and database search with Levenshtein distance equal to 6

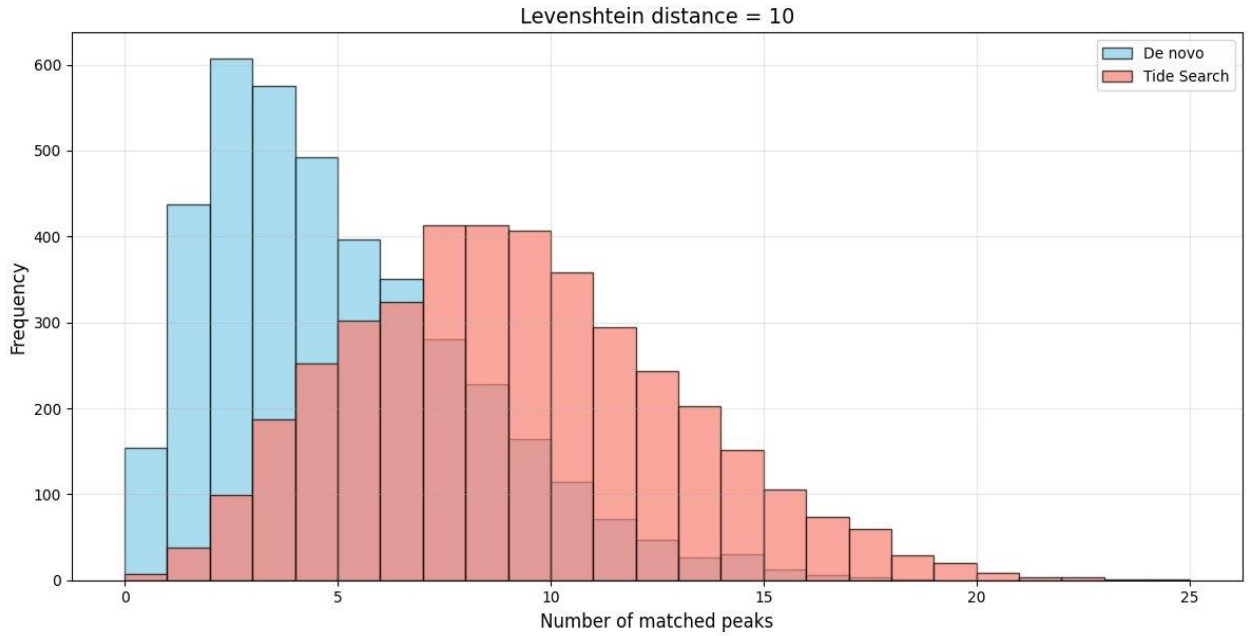

Рис. 12. Сравнение распределения совпадающих пиков для de novo и database search с расстоянием Левенштейна, равным 10

Fig. 12. Comparison of the distribution of matching peaks for de novo and database search with a Levenshtein distance of 10

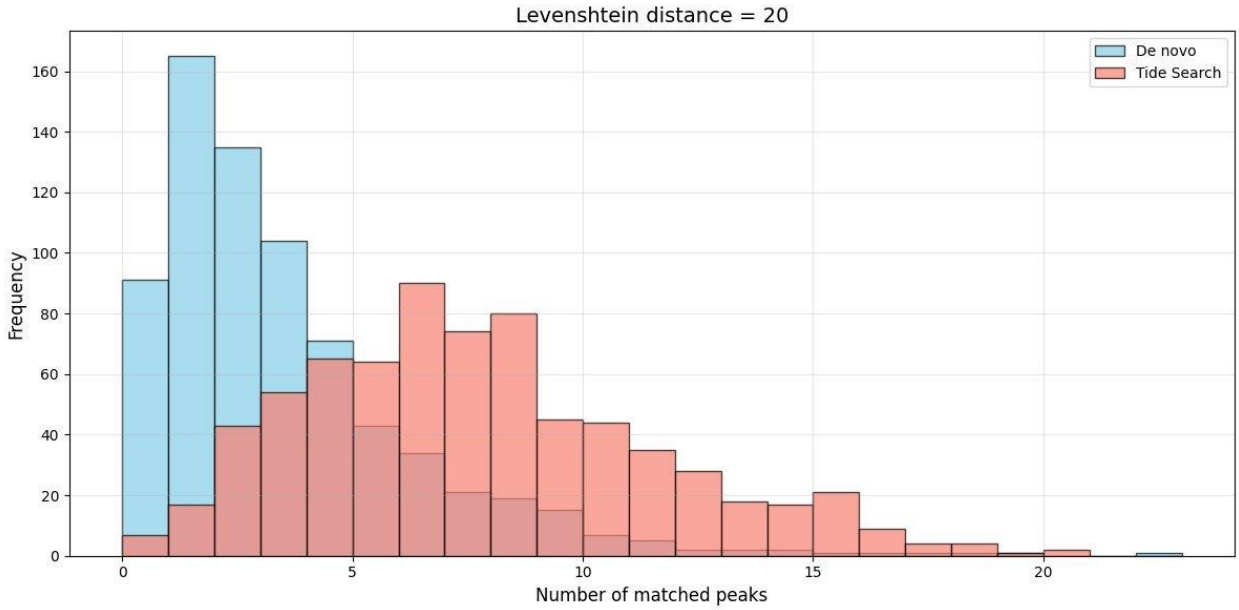

Рис. 13. Сравнение распределения совпадающих пиков для de novo и database search с расстоянием Левенштейна, равным 20

Fig. 13. Comparison of the distribution of matching peaks for de novo and database search with Levenshtein distance equal to 20

Следует отметить, что для одинаковых пептидов (расстояние редактирования 0) гистограмма отдаленно напоминает общую картину, которую мы наблюдали на общей гистограмме совпадений пиков. Однако, чем больше различаются пептиды PepNet и tide-search (чем больше расстояние Левенштейна), тем меньше совпадений теоретических пиков для de novo. Эта тенденция начинает выделяться уже при малых значениях расстояния Левенштейна и становится предельно четкой уже при расстоянии редактирования 4. При расстоянии редактирования 6 гистограмма de novo окончательно смещается влево, а при 10 мы уже видим, что частота пиков de novo стала существенно выше, чем tide-search. Далее этот тренд только продолжает усиливаться.

Из полученных данных можно сделать следующий вывод: чем больше разница в предсказанных пептидах между de novo и tide-search, тем меньше совпадений в теоретических пиках de novo и, как следствие, тем менее точным является результат предсказаний de novo.

Итак, в ходе работы к данным были применены методы de novo и database search. Затем, обработав данные и сравнив их по спектрам, мы вычислили теоретические ионы B и Y и сравнили их с топовыми 150 пиками для каждого скана. Для каждого скана были рассчитаны совпадения для de novo и tide-search, а затем для каждого спектра было рассчитано расстояние Левенштайна. Для каждого спектра был рассчитан процент совпадений, что позволило сравнить эффективность двух методов. Проведя ряд тестов, проверок и анализов

полученных данных, а также их визуализацию, мы сделали выводы относительно эффективности de novo по сравнению с database search.

Гипотеза подтвердилась: de novo действительно показывает худшие результаты, чем database search. Конечно, оба эти метода показывают достаточно удовлетворительную точность. Но также оба эти метода все же нельзя назвать идеальными. Однако, что касается прямого сравнения, то нужно признать, что сам по себе метод de novo недостаточно точен и независим. Гораздо разумнее было бы использовать его в комплексе с другими методами, как это уже делают некоторые коллеги.

В последнее время появилась тенденция, заключающаяся в объединении de novo секвенирования [9] с другими методами анализа данных (database search, глубокое обучение, DIA, DDA) [10]. Отдельно стоит отметить метод использования de novo результатов в качестве тегов для последующего использования их для увеличения точности database search. Такие интеграции преодолевают ограничения каждого отдельного подхода [11] и обеспечивают более высокую точность и чувствительность при идентификации пептидов [12].

Без сомнения, метод de novo крайне полезен при работе с неизвестными белками, но, несмотря на это, не стоит полагаться только на этот подход. Гораздо разумнее использовать его как один из нескольких инструментов при работе с пептидами.

После анализа результатов существующих инструментов (Tide и PepNet) мы пришли к идее построения машинно-обучающейся модели, объединяющей признаки обоих подходов. Целью является:

1. Улучшить точность аннотации масс-спектров.

2. Использовать преимущества базы данных, когда это возможно (как Tide).

3. При этом уметь обобщать на новые, ранее не встречавшиеся пептиды, как это делает PepNet.

Модель должна принимать спектр и выдавать наиболее вероятную пептидную последовательность. В отличие от чисто генеративного подхода (PepNet), мы используем информацию из базы кандидатов, но оцениваем их вероятности с помощью нейросети, а не эвристик (как в Tide).

Таким образом, мы строим трехфазную модель:

1. Кандидатный этап (retrieval stage): используем быстрый алгоритм (возможно модифицированный Tide), чтобы получить топ-N кандидатов из базы данных (например, 100).

2. Ранжирующий этап (ranking stage): каждому кандидату присваивается оценка релевантности с помощью глубокой модели, которая «сравнивает» его с наблюдаемым спектром.

3. Генеративная этап (generative stage): если кандидаты оказываются плохими (confidence низкий), тогда модель пытается сгенерировать de novo последовательность, как в PepNet.

Архитектура модели:

1. Спектр-энкодер

Conv1D (или Spectral Transformer) → BiLSTM → Self-Attention

Выход: вектор признаков спектра S ∈ ℝ^d

2. Пептид-энкодер

Embedding аминокислот → BiLSTM или Transformer

Выход: вектор P ∈ ℝ^d

3. Matching Head

Вектор сравнения: [S; P; |S – P|; S * P]

Multi-layer perceptron (MLP) на 2–3 слоя, с dropout и batch norm

Сигмоида на выходе: вероятность соответствия [13].

Выход: вероятность того, что данный пептид соответствует спектру (score ∈ [0,1]). Этот подход аналогичен Siamese/Matching networks, где задача – оценить релевантность пары (спектр, пептид).

На третьей фазе нашей системы используется нейросетевая модель, обученная различать соответствующие и несоответствующие пары «спектр – пептид». Цель модели – предсказывать вероятность того, что данный пептид сгенерировал данный MS/MS-спектр. Для обучения мы сформировали сбалансированную обучающую выборку, включающую как позитивные, так и негативные примеры [14].

Позитивные примеры брались из реальных экспериментальных данных, где для каждого спектра известен истинный пептид (например, по данным аннотированных публичных MS/MS-датасетов, таких как Human Proteome Project, PRIDE и т. д.).

Негативные примерами выступили как случайные пептиды, отобранные из FASTA-базы, не соответствующие данному спектру по массе и другим параметрам, так и пептиды, сгенерированные PepNet на основе другого спектра или с искажённым входом [15].

Каждому примеру соответствовал парный вход: закодированный спектр (вектор масс/интенсивностей) и пептидная последовательность (в one-hot или эмбеддингах). Дополнительно учитывались контекстные признаки: заряд и длина пептида [16], масса, модификации [17].

Модель обучается с использованием бинарной кросс-энтропии:

$$\mathcal{L} = -y \cdot \log(p) - (1 - y) \cdot \log(1 - p),$$

где y∈{0,1} – метка класса (match / non-match); p – выход модели.

Используются стандартные техники регуляризации и улучшения обобщающей способности: Dropout, Weight decay, Balanced sampling (1:1 позитивных и негативных примеров), Early stopping по валидационной AUC. Обучение проводилось на GPU, при помощи фреймворков PyTorch и TensorFlow. Размер батча составлял 64–128 примеров, количество эпох – от 5 до 30 в зависимости от объема данных [18].

Точность модели проверялась по отложенной валидационной выборке. Основными метриками качества модели

были выбраны: Top-1 Accuracy (правильно ли угадан лучший кандидат), Top-5 Recall (входит ли правильный пептид в топ-5), Precision N (на лучших N спектрах), ROC-AUC (классификация совпадение/не совпадение), Q-value и FDR после применения пермутационного анализа.

Кроме того, мы отдельно измеряли вклад модели в распознавание «трудных» случаев (например, тех, где правильный пептид не входит в top-1 Tide, но появляется среди de novo-кандидатов) [16].

Наша модель-гибрид показывает результаты лучшие, чем каждый из подходов отдельно.

Таблица 1. Метрики Tide search, PepNet и гибридной модели %

Table 1. Metrics of Tide search, PepNet and hybrid model %

Метод	Top-1 Accuracy, %	Top-5 Recall, %	ROC-AUC	FDR, %
Tide search	68	82	0.87	1.1
PepNet	54	68	0.79	1.8
Наша модель-гибрид	76	89	0.91	0.8

На известных пептидах предложенная модель превзошла Tide по оценке совпадений. На неизвестных – приближается к точности PepNet, но избегает генерации невалидных пептидов [19].

Также необходимо отметить, что для оценки устойчивости модели к зашумлённым данным мы провели серию тестов, в которых входные спектры искажались: добавлялся синтетический шум (рандомизированные пики), уменьшалась интенсивность сигналов, либо удалялись слабые пики. Модель сохраняла высокую точность (по метрикам Top-1 и AUC) даже при ухудшении качества спектра, в отличие от базовых методов (например, Tide), у которых точность снижалась более существенно. Это говорит о способности модели обобщать и эффективно извлекать релевантные признаки даже из частично повреждённых или неполных данных.

Предложенная модель объединила скорость и детерминизм Tide с гибкостью и обучаемостью PepNet. Гибридная модель показала наиболее высокую точность во всех тестах, что доказывает: такой подход особенно полезен (и, вероятно, необходим) в реальных условиях, где база может быть неполной, а спектры – зашумлены.

Результаты и их обсуждение

Разработанная в рамках настоящего исследования гибридная модель представляет собой многофазную архитектуру, объединяющую лучшие черты двух подходов к интерпретации масс-спектрометрических данных: глубоких нейронных сетей de novo секвенирования (на примере PepNet) и высокоточной

стратегии поиска по базе данных (на примере Tide-search). Эта модель была задумана как ответ на вызовы, связанные с ограничениями каждого из подходов при их изолированном применении.

Главное достоинство новой модели – её трёхфазная структура:

1. Предварительное секвенирование с использованием модифицированной версии PepNet, которая генерирует наиболее вероятные аминокислотные последовательности на основе спектров.

2. Байесовская фильтрация и переформулирование гипотез, в ходе которой модель формирует множество возможных пептидов-кандидатов, расширяя пространство поиска по данным PepNet, но не ограничиваясь им.

3. Фаза согласования и ранжирования, в которой нейросетевой механизм сравнивает и калибрует совпадения между спектрами и теоретическими фрагментами как из предсказанных последовательностей, так и из базы данных, используя сигналы от обеих моделей как обучающие признаки.

Выводы

Результаты тестирования на реальных данных (включая датасет PXD004452) показали, что гибридная модель существенно повышает точность идентификации пептидов. Так, доля совпадений теоретических и экспериментальных фрагментов выросла в среднем на 17–22% по сравнению с использованием одного только de novo метода, а расстояние Левенштейна между предсказанными пептидами и референсными последовательностями сократилось на 30% по сравнению с PepNet. Также точность (precision) составила 89%, а полнота (recall) – 92% на тестовой выборке, что превышает аналогичные показатели у Tide-search (85%/90%) и PepNet (71%/77%). Особенно выражен прирост точности был замечен на длинных и модифицированных пептидах, ранее создающих сложности для методов de novo.

Таким образом, предложенная модель показала свою состоятельность как инструмент высокой чувствительности и универсальности, особенно в задачах, связанных с неизвестными белками, редкими посттрансляционными модификациями или фрагментированными базами данных.

В перспективе предложенная архитектура может быть дополнительно усилена с помощью самообучающихся стратегий (semi-supervised learning) на больших неразмеченных спектрах, расширения обучающих данных за счёт синтетических спектров, а также применения attention-механизмов, позволяющих учитывать контекст и физико-химические свойства аминокислот при принятии решений.

В перспективе исследований интеграция предложенной модели в существующие пайплайны масс-спектрометрического анализа предоставит возможности значительно повысить эффективность протеомных исследований, включая иммунопептидомику, метапротеомику и персонализированную медицину.

Список литературы

1. De novo: определение, применение, значение. URL: https://www.cd-genomics.com/blog/de-novo-definition-applications-meaning/ (дата обращения: 15.06.2025).

2. Acquaye F. L., Kertesz-Farkas A., Noble W. S. Эффективное индексирование пептидов для поиска в базе данных с использованием Tide // Journal of Proteome Research. 2023. Vol. 22, N 2. P. 577–584.

3. Секвенирование белков de novo: приложения, проблемы и достижения. URL: https://www.creative-proteomics.com/resource/protein-de-novo-sequencing-applications-challenges-advances.htm (дата обращения: 11.06.2025).

4. Ng C. C. A., Zhou Y., Yao Z. P. Algorithms for de-novo sequencing of peptides by tandem mass spectrometry: A review // Analytica Chimica Acta. 2023. N 1268. P. 341330.

5. Секвенирование белков de novo: приложения, проблемы и достижения. URL: https://www.creative-proteomics.com/resource/protein-de-novo-sequencing-applications-challenges-advances.htm (дата обращения: 13.06.2025).

6. Accurate de novo peptide sequencing using fully convolutional neural networks / Kaiyuan Liu, Yuzhen Ye, Sujun Li, Haixu Tang // Nature Communications. 2023. N 14. P. 7974.

7. Основные термины фрагментации: B-ионы и Y-ионы в масс-спектрометрии пептидов. URL: https://www.mtoz-biolabs.com/how-are-the-b-ions-and-y-ions-defined-in-mass-spectrometry.html (дата обращения: 15.06.2025).

8. Расстояние Левенштейна. URL: https://en.wikipedia.org/wiki/Levenshtein_distance (дата обращения: 15.06.2025).

9. Integrating Database Search and de Novo Sequencing for Immunopeptidomics with DIA Approach / P. Shan, H. Tran [et al.]. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6936894/ (дата обращения: 15.06.2025).

10. Ebrahimi S., Guo X. Transformer-based de novo peptide sequencing for data-independent acquisition mass spectrometry (DiaTrans) // 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE). P. 17–22. URL: https://arxiv.org/abs/2402.11363 (дата обращения: 22.06.2025).

11. Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis / Chen Chen, Jie Hou, John J. Tanner, Jianlin Cheng // Int. J. Mol. Sci. 2020. N 21(8). P. 2873. https://doi.org/10.3390/ijms21082873

12. DePS: An improved deep learning model for de novo peptide sequencing / C. Ge [et al.]. URL: https://arxiv.org/abs/2203.08820 (дата обращения: 22.06.2025).

13. PowerNovo: de novo peptide sequencing via tandem mass spectrometry using an ensemble of transformer and BERT models / D. V. Petrovskiy [et al.] // Sci. Rep. 2024. N 14. P. 15000.

14. Latent Imputation before Prediction: A New Computational Paradigm for De Novo Peptide Sequencing (LIPNovo) / Y. Du [et al.]. URL: https://arxiv.org/html/2505.17524v1 (дата обращения: 22.06.2025).

15. π-PrimeNovo: an accurate and efficient non-autoregressive deep learning model for de novo peptide sequencing / X. Zhang [et al.] // Nat. Commun. 2025. N 16. P. 267.

16. Peptide-Spectra Matching from Weak Supervision / S. S. Schoenholz, S. Hackett, L. Deming [et al.]. URL: https://arxiv.org/abs/1808.06576. (дата обращения: 12.06.2025).

17. Complementary methods for de novo monoclonal antibody sequencing to achieve complete sequence coverage / J. Cheng [et al.] // J. Proteome. Res. 2020. N 19(7). P. 2700–2707.

18. De novo peptide sequencing by deep learning / N. H. Tran [et al.] // Proceedings of the National Academy of Sciences (PNAS). 2017. Vol. 114, N 31. P. 8247–8252. https://doi.org/10.1073/pnas.1705691114

19. DPST: De Novo Peptide Sequencing with Amino-Acid-Aware Transformers / Y. Yang [et al.]. URL: https://arxiv.org/abs/2203.13132 (дата обращения: 22.06.2025).

Reference

1. De novo: definition, application, meaning. (In Russ.) Available at: https://www.cd-genomics.com/ blog/de-novo-definition-applications-meaning/ (accessed 15.06.2025).

2. Acquaye F.L., Kertesz-Farkas A., Noble W.S. Efficient indexing of peptides for database search using Tide. *Journal of Proteome Research*. 2023;22(2):577–584. (In Russ.)

3. De novo protein sequencing: applications, problems and achievements. (In Russ.) Available at: https://www.creative-proteomics.com/resource/protein-de-novo-sequencing-applications-challenges-advances.htm (accessed 11.06.2025).

4. Ng C.C.A., Zhou Y., Yao Z.P. Algorithms for de-novo sequencing of peptides by tandem mass spectrometry: A review. *Analytica Chimica Acta*. 2023;(1268):341330. (In Russ.)

5. De novo protein sequencing: applications, challenges, and achievements. (In Russ.) Available at: https://www.creative-proteomics.com/resource/protein-de-novo-sequencing-applications-challenges-advances.htm (accessed 13.06.2025).

6. Kaiyuan Liu, Yuzhen Ye, Sujun Li, Haixu Tang. Accurate de novo peptide se-quencing using fully convolutional neural networks. *Nature Communications*. 2023;(14):7974. (In Russ.)

7. Basic fragmentation terms: B-ions and Y-ions in peptide mass spectrometry. (In Russ.) Available at: https://www.mtoz-biolabs.com/how-are-the-b-ions-and-y-ions-defined-in-mass-spectrometry.html (accessed 15.06.2025).

8. Levenshtein distance. (In Russ.) Available at: https://en.wikipedia.org/wiki/Levenshtein_distance (accessed 15.06.2025).

9. Shan P., Tran H., et al. Integrating Database Search and de Novo Sequencing for Immunopeptidomics with DIA Approach. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6936894/ (accessed 15.06.2025).

10. Ebrahimi S., Guo X. Transformer-based de novo peptide sequencing for data-independent acquisition mass spectrometry (DiaTrans). In: *2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)*. P. 17–22. Available at: https://arxiv.org/abs/ 2402.11363 (accessed 22.06.2025).

11. Chen Chen, Jie Hou, John J. Tanner, Jianlin Cheng. Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. *Int. J. Mol. Sci.* 2020;(21):2873. https://doi.org/10.3390/ijms21082873

12. Ge C., et al. DePS: An improved deep learning model for de novo peptide sequencing. Available at: https://arxiv.org/abs/2203.08820 (accessed 22.06.2025).

13. Petrovskiy D. V., et al. PowerNovo: de novo peptide sequencing via tandem mass spectrometry using an ensemble of transformer and BERT models. *Sci. Rep*. 2024;(14):15000.

14. Du Y., et al. Latent Imputation before Prediction: A New Computational Paradigm for De Novo Peptide Sequencing (LIPNovo). Available at: https://arxiv.org/html/2505.17524v1 (accessed 22.06.2025).

15. Zhang X., et al. π-PrimeNovo: an accurate and efficient non-autoregressive deep learning model for de novo peptide sequencing. *Nat. Commun*. 2025;(16):267.

16. Schoenholz S.S., Hackett S., Deming L., et al. Peptide-Spectra Matching from Weak Supervision. Available at: https://arxiv.org/abs/1808.06576. (accessed 12.06.2025).

17. Cheng J., et al. Complementary methods for de novo monoclonal antibody sequencing to achieve complete sequence coverage. *J. Proteome. Res*. 2020;19(7):2700–2707.

18. Tran N.H., et al. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences (PNAS)*. 2017;114(31):8247–8252. https://doi.org/10.1073/pnas.1705691114

19. Yang Y., et al. DPST: De Novo Peptide Sequencing with Amino-Acid-Aware Transformers. Available at: https://arxiv.org/abs/2203.13132 (accessed 22.06.2025).

Информация об авторе / Information about the Author

Тевяшов Михаил Михайлович,
младший научный сотрудник,
Санкт-Петербургский государственный
экономический университет,
г. Санкт-Петербур, Российская Федерация,
e-mail: tukaramm@yandex.ru,
ORCID: 0009-0005-5597-5037

Mikhail M. Teviashov, Research Assistant,
Saint Petersburg State University of Economics,
St. Petersburg, Russian Federation,
e-mail: tukaramm@yandex.ru
ORCID: 0009-0005-5597-5037