МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В.ЛОМОНОСОВА

На правах рукописи

Хрисанфов Михаил Дмитриевич

Поиск ошибок в хроматографических базах данных и моделирование свойств идентифицируемых молекул методами машинного обучения

Специальность 1.4.2. Аналитическая химия

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата химических наук

Москва – 2026

Диссертация подготовлена на кафедре аналитической химии химического факультета МГУ имени М.В. Ломоносова

Научный руководитель: **Самохин Андрей Сергеевич**
кандидат химических наук

Официальные оппоненты:
Вирюс Эдуард Даниэлевич
доктор химических наук
Федеральный научный центр физической культуры и спорта, лаборатория биохимии, начальник лаборатории

Кирсанов Дмитрий Олегович
доктор химических наук
Санкт-Петербургский государственный университет, Институт химии, кафедра аналитической химии, профессор

Канатьева Анастасия Юрьевна
кандидат химических наук
Институт нефтехимического синтеза им. А.В. Топчиева РАН, лаборатория спектральных и хроматографических исследований, ведущий научный сотрудник

Защита диссертации состоится «09» сентября 2026 года в 15 часов 00 минут на заседании диссертационного совета МГУ.014.5 Московского государственного университета имени М.В. Ломоносова по адресу: 119991, Москва, ГСП-1, Ленинские горы, д.1, стр.3, МГУ имени М.В. Ломоносова, химический факультет, аудитория 446.
E-mail: dissovet02.00.02@mail.ru

С диссертацией можно ознакомиться в отделе диссертаций научной библиотеки МГУ имени М.В. Ломоносова (Ломоносовский просп., д. 27) и на сайте https://dissovet.msu.ru/dissertation/3970

Автореферат разослан «__» _____ 2026 г.

Ученый секретарь
диссертационного совета, Ананьева И.А.
кандидат химических наук

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

Различные варианты хроматографии (газовая и высокоэффективная жидкостная) в сочетании с масс-спектрометрическим детектированием широко используют как для исследовательских, так и для рутинных задач. Одним из важных достоинств таких гибридных систем является крайне высокая информативность за счет сочетания хроматографического разделения и масс-спектрометрической характеризации. Однако для получения конечного результата анализа необходимо не только провести эксперимент и собрать данные с прибора, но и обработать их. Большой объем данных и их разнородный характер вызывают необходимость в разработке и применении различных алгоритмов и компьютерных программ. В этой области используют как самые простые вычислительные подходы, так и более комплексные решения из областей хемометрики и хемоинформатики. В последнее десятилетие все более активно развиваются методы, использующие машинное и глубокое обучение на различных этапах обработки хроматомасс-спектральных данных.

Однако даже тривиальные, на первый взгляд, задачи, например округление значений *m/z* в масс-спектрах до целочисленных или сравнение масс-спектров низкого разрешения далеко не всегда решаются оптимально. Общепринятые и широко используемые алгоритмы являются частью коммерческого программного обеспечения, а некоторые эффективные подходы могут быть реализованы в рамках небольших проектов и оставаться неизвестными широкой публике.

В области идентификации двумя важными взаимодополняющими задачами являются предсказание хроматомасс-спектральных характеристик соединений и очистка экспериментальных баз данных от ошибочных записей. Поскольку число известных органических молекул на порядки больше, чем размер существующих масс-спектральных и хроматографических наборов данных, то предсказание становится особенно актуальным для решения задач нецелевого и скринингового анализа, например, объектов окружающей среды. При этом важную роль играет не только точность предсказания, но и скорость и простота применения подхода. Эти характеристики зачастую отходят на второй план, однако являются критическими для широкого применения разработанных решений сотрудниками химических лабораторий.

В то же время очистка уже существующих баз данных позволяет не только улучшить точность предсказаний, но и избежать возможного искажения результатов анализов при сравнениях экспериментально полученных и библиотечных данных. Оценка правильности записей вручную обычно затруднена из-за значительного размера наборов данных, а отсутствие дублирующих записей приводит к невозможности применить статистические инструменты. При всем разнообразии методов машинного обучения они редко используются для поиска ошибок не только в

случае химических баз данных, но и более популярных областях, таких как анализ изображений и обработка табличных данных.

Таким образом, разработка алгоритмов и программных инструментов для обработки и предсказания хроматомасс-спектральных данных, а также поиска и устранения ошибок в базах данных могут помочь решать как рутинные, так и исследовательские задачи более эффективно, увеличить надежность получаемых результатов, а также заменить часть экспериментальной работы вычислениями.

Цель и задачи работы

Усовершенствование существующих и разработка новых подходов к обработке и предсказанию хроматомасс-спектральных данных, в частности для поиска ошибок в базах времен и индексов удерживания, а также для предсказания масс-спектров соединений по их структуре и молекулярных отпечатков пальцев по масс-спектрам электронной ионизации.

Для достижения поставленной цели было необходимо решить следующие **задачи**:

1) Предложить подход к поиску ошибок в хроматографических базах данных, основанный на механизме голосования независимых предсказательных моделей. Сгенерировать синтетические наборы размеченных данных с контролируемым количеством и типом ошибок для проведения оценки эффективности поиска ошибок. Применить подход для поиска ошибок в базах данных NIST 17 RI (индексы удерживания) и METLIN SMRT (времена удерживания).
2) Предложить алгоритм округления значений *m/z* с плавающей запятой в целочисленный формат с целью уменьшения влияния случайной приборной погрешности на результаты обработки масс-спектральных данных.
3) Усовершенствовать существующие подходы к предсказанию масс-спектров электронной ионизации по структуре молекулы и структурных дескрипторов (например, молекулярных отпечатков пальцев) по масс-спектрам электронной ионизации.

Научная новизна:

1) Предложен подход к поиску ошибочных записей в хроматографических базах данных NIST RI и METLIN SMRT с использованием объединения предсказаний нескольких независимых моделей машинного и глубокого обучения путем голосования.
2) Показано с использованием синтетических наборов данных, что предложенный подход к поиску ошибок работает лучше более распространенных альтернатив, установлено влияние архитектуры и количества моделей на эффективность поиска ошибок.
3) Предложены способы оценки эффективности подхода к поиску ошибок и выбора оптимального значения его параметров для неразмеченных наборов данных.

4) Предложен алгоритм округления значений *m/z* масс-спектров низкого разрешения ($\Delta m_{50\%} \sim 0.5$) с плавающей запятой до целочисленных, позволяющий минимизировать влияние случайных приборных погрешностей на результаты.
5) Усовершенствованы архитектуры нейросетевых моделей и способы распространения подходов к предсказанию масс-спектров электронной ионизации по структуре молекулы и молекулярных отпечатков пальцев по масс-спектрам электронной ионизации.

Практическая значимость:

1) Предложенный подход к поиску ошибок в базах хроматографических данных распространяется в виде библиотеки для языка программирования Python с открытым исходным кодом вместе с пошаговой инструкцией.
2) Предложенный подход к фильтрованию позволил найти 2093 потенциально ошибочные записи в базе данных NIST 17 RI и 1544 в METLIN SMRT.
3) Предложенный алгоритм округления значений *m/z* до целочисленных позволяет минимизировать неопределенность результатов, зависящую от случайных приборных погрешностей, добавлен разработчиками в широко применяемый пакет OpenChrom.
4) Предложенные подходы с увеличенной точностью предсказания масс-спектров электронной ионизации по структуре молекулы и с возросшей скоростью обучения и предсказания молекулярных отпечатков пальцев по масс-спектрам электронной ионизации распространяются в открытом доступе, что позволяет использовать их в практических и исследовательских задачах сторонними группами.

Положения, выносимые на защиту

1) Подход, основанный на объединении предсказаний нескольких независимых моделей машинного и глубокого обучения путем голосования, позволяет эффективно обнаруживать ошибки в хроматографических базах данных NIST RI и METLIN SMRT, а также оценивать точность обнаружения на неразмеченных наборах данных.
2) Алгоритм округления значений *m/z* с плавающей запятой в масс-спектрах низкого разрешения ($\Delta m_{50\%} \sim 0.5$) до целочисленных позволяет уменьшить зависимость результатов от случайных приборных ошибок.
3) Оптимизированные архитектуры и гиперпараметры нейросетевых моделей позволяют увеличить точность предсказания масс-спектров электронной ионизации по структуре молекулы, а также ускорить обучение и улучшить правильность, точность и полноту предсказания молекулярных отпечатков по масс-спектрам электронной ионизации.

Степень достоверности

Достоверность полученных результатов на каждом этапе работ обеспечивалась применением общепринятых подходов и метрик к сравнению качества предсказания моделей, стандартных в области работы с данными библиотек для языка программирования Python, а также синтетических наборов данных, содержащих заданное количество ошибок, для валидации предложенных решений.

Соответствие паспорту научной специальности

Диссертационная работа соответствует паспорту специальности 1.4.2. Аналитическая химия по областям исследований: методы химического анализа (хроматография, масс-спектрометрия); методическое и математическое обеспечение химического анализа; анализ органических веществ и материалов; анализ объектов окружающей среды.

Апробация результатов исследования

Результаты работы представлены на 14 российских и международных конференциях: (1) XXVI Международная научная конференция студентов, аспирантов и молодых ученых "Ломоносов – 2019", секция «Химия», МГУ имени М.В.Ломоносова, Россия, 8-11 апреля 2019, (2) Девятый съезд ВМСО, VIII Всероссийская конференция с международным участием «Масс-спектрометрия и ее прикладные проблемы», Москва, Россия, 14-18 октября 2019, (3) Международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов-2021» секция Химия, МГУ им.М.В.Ломоносова химический факультет, Россия, 12-23 апреля 2021, (4) Десятый съезд ВМСО IX Всероссийская конференция с международным участием "Масс-спектрометрия и ее прикладные проблемы", Москва, Россия, 18-22 октября 2021, (5) XXIX Международная научная конференция студентов, аспирантов и молодых ученых "Ломоносов 2022", МГУ имени М.В. Ломоносова, Россия, 11-22 апреля 2022, (6) IV Съезд аналитиков России, Москва, Россия, 26 сентября - 30 октября 2022, (7) IV Всероссийская конференция по аналитической спектроскопии с международным участием, Краснодар, Россия, 24-30 сентября 2023, (8) X Всероссийская конференция с международным участием «Масс-спектрометрия и ее прикладные проблемы», г. Москва, Россия, 30 октября - 3 ноября 2023, (9) Fourteenth Winter Symposium on Chemometrics, Цахкадзор, Армения, 26 февраля - 1 марта 2024, (10) 53rd International Symposium on High Performance Liquid Phase Separations and Related Techniques - HPLC 2024, Далянь, Китай, 21-23 октября 2024, (11) Международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов-2025», секция «Химия», Москва, МГУ, Россия, 11-25 апреля 2025, (12) VII Всероссийский симпозиум «Разделение и концентрирование в аналитической химии и радиохимии» с международным участием, Краснодар, Россия, 21-27 сентября 2025, (13) XI Всероссийская конференция с международным участием «Масс-спектрометрия и ее прикладные проблемы», Москва, Россия, 13 – 17 октября 2025 года, (14) II Научная

конференция «Искусственный интеллект в химии и материаловедении», Москва, Россия, 17-21 ноября 2025.

Публикации

По теме диссертации опубликованы 4 печатные работы, включая 4 статьи в рецензируемых научных журналах, рекомендованных для защиты в диссертационном совете МГУ по специальности и отрасли наук.

Личный вклад автора

Представленные результаты исследования получены лично автором или в соавторстве. Автор провел поиск и критическое обобщение литературы по теме работы, реализовал все заявленные в работе алгоритмы и подходы машинного и глубокого обучения на языке программирования Python, обработал полученные результаты и оформил их в виде научных публикаций (при участии соавторов). Во всех опубликованных по теме диссертации работах вклад автора (а именно: формальный анализ, проведение исследования, разработка программного обеспечения, визуализация, написание черновика рукописи) является определяющим. Помощь в обсуждении подходов к предсказанию хроматомасс-спектральных данных, поиску ошибок, обучению моделей оказана Матюшиным Д.Д.

Структура и объем работы

Диссертационная работа состоит из списка основных обозначений и сокращений, введения, обзора литературы, методов и алгоритмов, обсуждения и результатов (в двух главах), заключения, выводов, списка цитируемой литературы из 158 наименований. Полный объем диссертации составляет 110 страниц, включая 39 рисунков, 12 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обозначена актуальность выбранной темы, сформулированы цели и задачи исследования, показаны научная новизна и практическая значимость работы, перечислены публикации, степень апробации работы, ее структура и объем.

Обзор литературы

В **главе 1** рассмотрены и обобщены наиболее актуальные подходы к предсказанию индексов и времен удерживания с использованием машинного и глубокого обучения. Описаны решения для поиска и исправления ошибок в крупных наборах данных, отмечены случаи их применения к химическим данным. Отдельно рассмотрены существующие подходы к поиску ошибок в хроматографических базах данных, которые зачастую очень упрощенные и редко применяются как для обучения моделей, так и в экспериментальной работе. В подразделе, посвященном масс-спектрометрическим данным и методам, описаны существующие подходы к округлению значений *m/z* в масс-спектрах электронной ионизации. Также приведены решения, позволяющие устанавливать структурные фрагменты молекулы по масс-спектру путем решения прямой и обратной задач. Так, рассмотрены подходы к

предсказанию молекулярных отпечатков пальцев по масс-спектрам электронной ионизации. Описаны наиболее актуальные модели для предсказания масс-спектров по структуре молекулы, в которых установление частичной структуры проводится поиском наиболее близкого совпадения по набору сгенерированных масс-спектров.

Методы и алгоритмы

В **главе 2** приведены использованные в исследовании методы, алгоритмы и оборудование. Кратко перечислены общие сведения об используемых предсказательных моделях – полносвязной, сверточной, графовой сверточной нейронной сетях, а также нейросетевых функциях активации, и методе градиентного бустинга. Далее рассмотрены использованные в работе наборы экспериментальных данных – базы индексов удерживания NIST RI, времен удерживания METLIN SMRT, масс-спектров электронной ионизации низкого разрешения NIST MS, двух групп сгенерированных наборов данных (на основе автокорреляционных дескрипторов и квантовомеханической базы QM9) и их основные характеристики. Перечислено использованное оборудование, программное обеспечение и основные библиотеки для языка программирования Python, использованные для расчетов и обучения предсказательных моделей. Затем в подразделах подробно описаны архитектуры и оптимальные гиперпараметры моделей машинного обучения, предложенные для предсказания индексов и времен удерживания, синтетических наборов данных, масс-спектров электронной ионизации по структуре молекулы и молекулярных отпечатков пальцев из масс-спектров. Кроме этого, рассмотрены наиболее актуальные известные решения для перечисленных задач, с которыми проводится сравнение. Описаны предложенные алгоритмы поиска ошибок путем объединения предсказаний нескольких независимых моделей через механизм голосования, и округления значений *m/z* в масс-спектрах электронной ионизации низкого разрешения до целочисленных.

Результаты и их обсуждение

Основные результаты исследования и их обсуждение представлены в **главах 3 и 4**. В **главе 3** предложен подход к поиску ошибок в химических базах данных, подробно рассмотрены различные аспекты его работы и применение к двум наборам экспериментальных данных: индексов удерживания NIST 17 RI и времен удерживания METLIN SMRT. В **главе 4** представлен алгоритм для округления значений *m/z* в масс-спектрах электронной ионизации низкого разрешения до целочисленных и проведено сопоставление с аналогами в коммерческих и открытых программных продуктах. Предложены подходы к предсказанию масс-спектров электронной ионизации по структуре молекулы и молекулярных отпечатков пальцев из масс-спектров, проведено сравнение с наиболее актуальными из представленных в литературе альтернативами.

Поиск ошибок в химических базах данных

В последнее время очистке данных уделяется все больше внимания, в том числе с использованием методов машинного и глубокого обучения. В химии и смежных науках вопрос стоит особенно остро, так как существует потребность в высококачественных наборах данных. При этом записи в таких базах могут быть уникальны, что исключает возможность применения статистических методов для оценки надежности значений. Проверка записей вручную неэффективна и зачастую требует проведения экспериментов, что тратит время сотрудников, приборное время и дорогостоящие реагенты.

Для решения этой проблемы разработан подход к обнаружению ошибочных записей в базах химических (хроматографических) данных, основанный на объединении предсказаний нескольких независимых моделей путем голосования «желтыми карточками». Каждая из N моделей обучается на разбиении данных на k частей (k-fold), где $k-1$ частей используют для обучения, а одну - для предсказания. Эту процедуру проводят для каждой модели, чтобы получить N предсказанных копий исходного набора данных. В каждой копии выбирают определенную долю записей $t\%$ (например, 5%), предсказанных хуже всего, и отмечают их «желтой карточкой». Этот варьируемый параметр является основным в подходе. Записи, получившие «желтые карточки» от всех моделей одновременно, являются потенциально ошибочными.

Для оценки эффективности предложенного подхода к фильтрованию сгенерированы две группы тестовых наборов данных – на основе автокорреляционных молекулярных дескрипторов и квантовохимического набора QM9, в которые контролируемым образом добавлены ошибки.

Предсказания, полученные для моделей, обучавшихся на модифицированных наборах данных с добавленными ошибками близки к немодифицированным значениям (Рисунок 1). Несмотря на то, что ошибки предсказания линейно возрастают вместе с долей модифицированных записей, угол наклона небольшой. Более того, точность предсказания для модифицированных и немодифицированных записей отличается незначительно. Таким образом, предсказательные модели не повторяют ошибки в тренировочных наборах данных и корректно предсказывают целевые величины для всех записей.

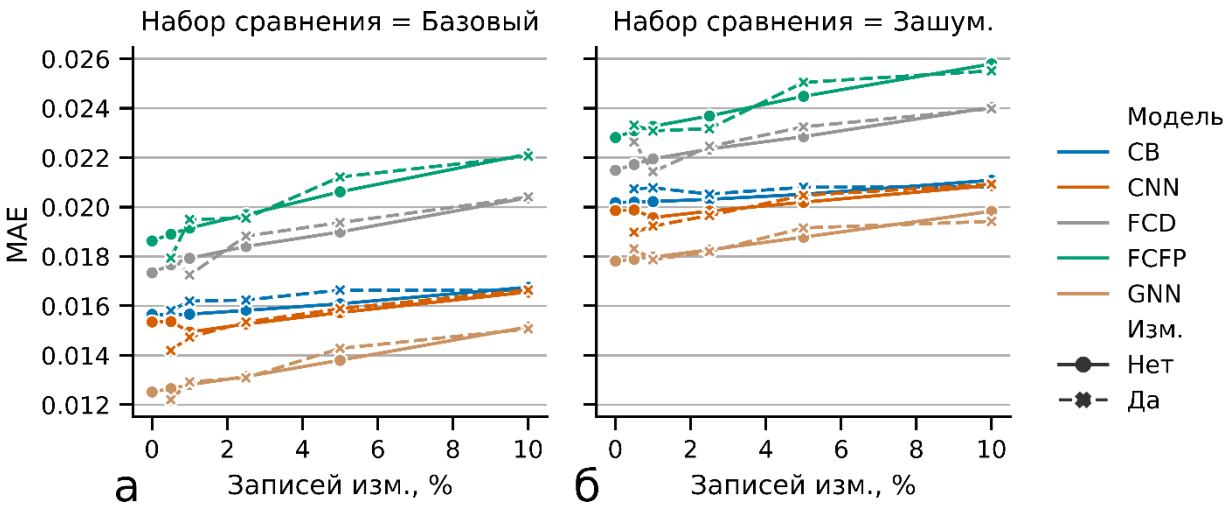


Рисунок 1. Средняя абсолютная ошибка предсказания (MAE) для 5 моделей на перемешанных дескрипторных наборах данных с 0.0-10.0% случайно перемешанных записей при расчете относительно: а – базового и б – зашумленного (с добавленным нормально распределенным шумом) наборов. Сплошными линиями показаны изменения значения, рассчитанные для исходных (немодифицированных) записей, а для ошибочных (измененных) – прерывистыми

Все предсказательные модели, используемые в данной работе, демонстрируют умеренно низкую ошибку предсказания, поэтому ожидается, что предсказания всех пар независимых моделей сильно скоррелированы. Действительно, вне зависимости от рассматриваемого набора данных предсказания любых двух моделей показывают коэффициент корреляции порядка 0.99. Однако в подходе к фильтрованию с «желтыми карточками» большее внимание уделяется не корреляции между предсказаниями, а корреляции между ошибками предсказания – небольших различиях между предсказанной величиной и референсным значением (рассчитанным или экспериментальным).

При расчете коэффициентов корреляции между ошибками предсказания разных моделей с использованием модифицированных наборов данных (тех же, которые использовались для обучения) наблюдается устойчивая тенденция к росту значений с увеличением доли модифицированных записей (Рис. 2). Абсолютные значения варьируются от умеренного 0.63 до крайне высокого 0.97, с некоторыми изменениями в зависимости от рассматриваемой пары моделей. С увеличением количества модифицированных записей и увеличением размера добавленной ошибки уменьшается разброс коэффициентов корреляции между разными парами моделей. Коэффициенты корреляции, полученные для QM9 набора данных, в среднем ниже, чем для дескрипторного набора данных.

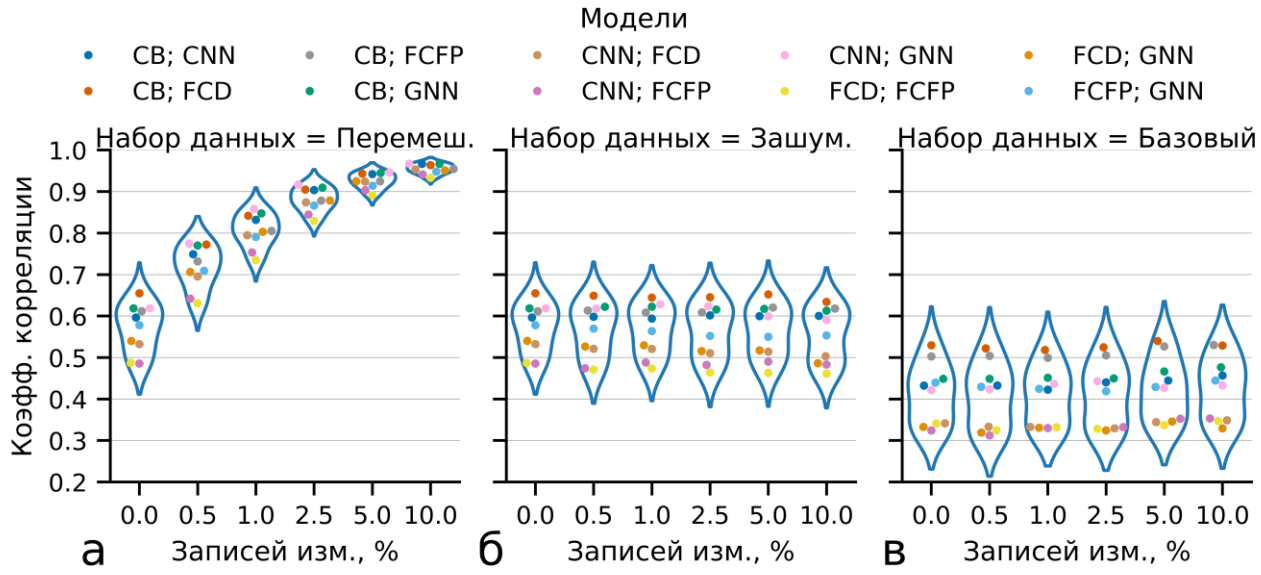


Рисунок 2. Коэффициенты корреляции между ошибками предсказания моделей для дескрипторного набора данных, рассчитанные относительно:
а – перемешанных, б – зашумленного и в – базового наборов

Однако коэффициенты корреляции значительно меняются при расчете ошибок предсказания относительно зашумленного (без ошибок, но с добавленным нормально распределенным шумом, имитирующим экспериментальные погрешности) или чистого (без ошибок и шума) наборов данных (рис. 2). В этих случаях значения лежат в диапазонах 0.4…0.7 и 0.2…0.6 соответственно, а разброс между коэффициентами корреляции, рассчитанными для одной пары моделей на разных наборах данных, становится пренебрежимо мал.

Умеренные коэффициенты корреляции, полученные и в группах из 10 графовых (GNN) и полносвязных (с молекулярными отпечатками пальцев в качестве входных данных, FCFP) моделей, свидетельствуют о том, что возможно применение моделей одного типа вместо ансамбля из нескольких архитектур. Эти модели выбраны для оценки эффективности фильтрования, так как первая показывает наиболее высокую точность предсказания, а вторая – наибольшую скорость и использует простые для расчета молекулярные отпечатки пальцев, которые рассчитываются функциями из библиотеки RDKit. Такой подход имеет значительные преимущества: значительно упрощается подготовка входных данных для моделей и устраняется необходимость подбора оптимальных гиперпараметров для нескольких моделей. Наиболее очевидный недостаток – увеличенные значения коэффициентов корреляции между ошибками предсказания моделей – может быть компенсирован использованием большего их количества.

Модели обучены на наборе данных, по-разному разбитом на тренировочную и валидационную части. Комбинации моделей для сравнения выбраны следующим образом: для каждого набора (1, 5, 10 моделей) использована комбинация с наибольшей F_1 метрикой, рассчитанная для всех вариантов порогового значения *t%* (доля наименее

точно предсказанных записей, считающихся ошибочными для каждой из моделей). Результаты сопоставлены с оригинальным ансамблем из пяти предсказательных моделей.

Сравнение эффективности фильтрования подходов в настоящей работе осуществлено при помощи кривых precision-recall (рис. 3а) и площади под ними (рис. 3б). Кривые рассчитаны для пороговых значений *t%* от 0.01% до 100%. Более эффективные подходы к фильтрованию характеризуются кривыми, которые лежат выше и правее менее эффективных подходов.

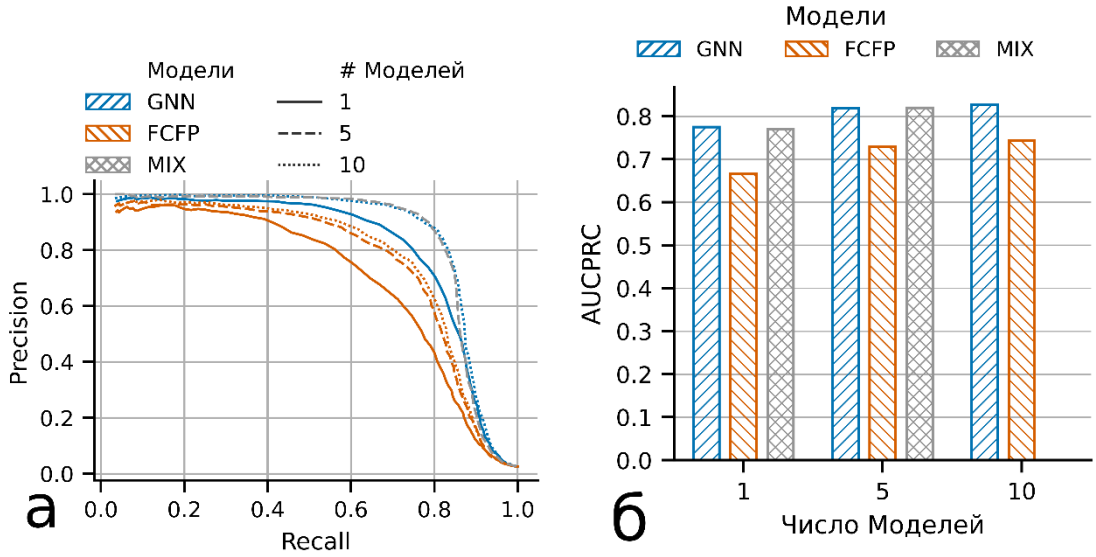


Рисунок 3. Оценка эффективности поиска ошибок разными ансамблями предсказательных моделей. (а) кривые precision-recall (точность-полнота) для 1,5,10 GNN, 1,5,10 FCFP и изначального ансамбля из 5 моделей; (б) площади под кривыми (AUC) precision-recall (точность-полнота)

Наиболее эффективной одиночной моделью является GNN, что ожидаемо, исходя из ее наибольшей точности предсказания. Лучшая из одиночных FCFP моделей показывает значительно более скромные результаты. Даже объединение 5 и 10 FCFP моделей уступает лучшей из одиночных GNN моделей. При этом 10 GNN моделей показывают эффективность фильтрования, крайне близкую к оригинальному ансамблю из 5 разнородных моделей.

Таким образом, возможно использовать несколько моделей с одинаковой архитектурой для фильтрования для значительного упрощения подхода. В то же время использование нескольких разнородных моделей является более устойчивым подходом, так как неправильный выбор единичной архитектуры может значительно ухудшить эффективность фильтрования.

Два наиболее распространенных и простых подхода к фильтрованию – отсечение по абсолютной ошибке (все записи с ошибкой предсказания больше заданной считаются ошибочными) или по квантилю/процентилю (определенная доля наименее точно предсказанных записей считается ошибочными). Кривые precision-recall для более простых методов одинаковые, так как каждое из значений квантиля соответствует определенной абсолютной ошибке для конкретного набора данных и предсказательных

моделей. Тем не менее, на практике часто выбирают один или другой подход исходя из личных предпочтений и представлений о возможных ошибочных записях в наборе.

В этой работе для применения квантильного подхода к фильтрованию рассчитывали медианное значение с использованием всех предсказательных моделей. Это позволило учесть вклад всех 5 моделей и сделало сравнение более корректным, чем использование только одной из моделей. Во всех рассмотренных случаях подход к фильтрованию с использованием «желтых карточек» превосходит более простые альтернативы (рис. 4). Для наборов данных с большой долей модифицированных записей и добавленной ошибкой более простые подходы показали сравнимые с «желтыми карточками» результаты.

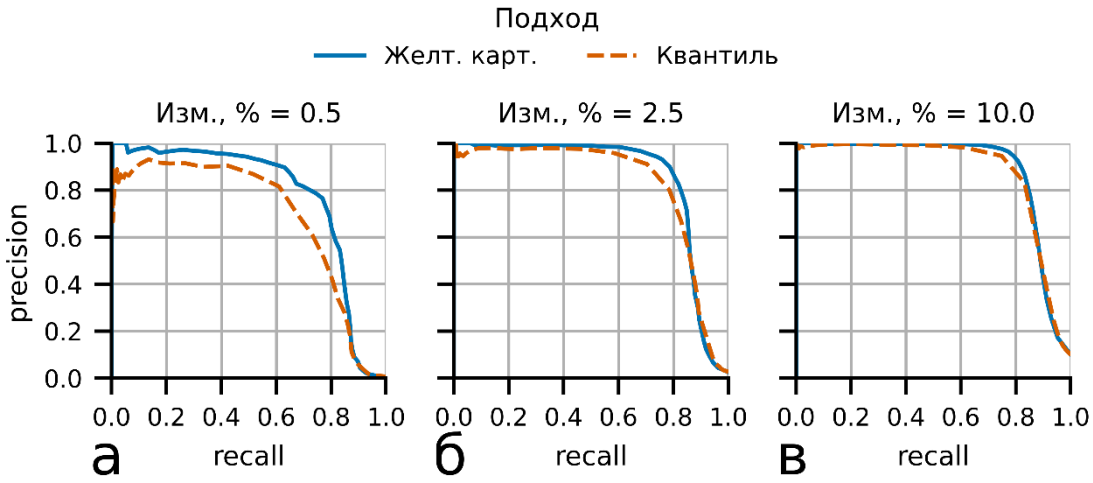


Рисунок 4. Кривые precision-recall (точность-полнота) для квантильного отсечения (синяя сплошная линия) и для подхода «желтых карточек» (прерывистая оранжевая) на примере перемешанных дескрипторных наборов данных: а – 0.5%, б – 2.5%, в – 10.0% перемешанных записей

Несмотря на то, что использование отсечения по абсолютной ошибке может казаться более естественным, этот подход опирается на априорные знания о природе записей и возможных ошибок в наборе данных, а также точности предсказания моделей. Фильтрование с использованием квантиля также предполагает наличие априорного знания о количестве ошибок в наборе данных, в противном случае результаты могут быть крайне неточными. Кроме того, в отличие от подхода с использованием «желтых карточек» более простые варианты фильтрования не позволяют контролировать процесс, так как не предлагают индикаторов для оценки результатов. Вместе с тем, использование ансамбля из всех моделей делает их такими же вычислительно-затратными, как и использование «желтых карточек».

Основным параметром, определяющим работу подхода «желтых карточек», является пороговое значение *t%*. Такая доля наименее точно предсказанных записей для каждой из моделей отмечается «желтыми карточками». Записи, отмеченные «желтыми карточками» для всех моделей одновременно, являются потенциально ошибочными. Именно использование нескольких независимых предсказаний приводит

к нелинейному поведению. При увеличении параметра *t%* от 0 до 100% на каждом шаге в группу потенциально ошибочных добавляется некоторое количество записей (рис. 5).

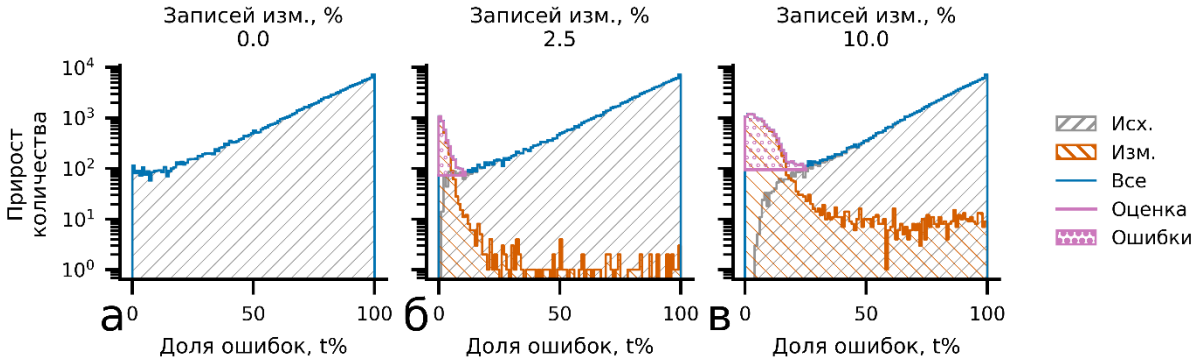


Рисунок 5. Зависимость количества записей, впервые отмеченных ошибочными, от порогового параметра t% для дескрипторных наборов данных с разными долями перемешанных записей в логарифмических координатах: а – 0%, б – 2.5%, в – 10%. В подходе «желтых карточек» t% наименее точно предсказанных записей для всех моделей одновременно считают ошибочными. Синей кривой показаны все записи в группе с максимальным количеством «желтых карточек», серым – исходные записи, оранжевым – перемешанные, фиолетовым – оценка количества перемешанных записей

Для набора данных без добавленных ошибок эта кривая имеет вид возрастающей экспоненциальной функции. Для наборов данных с ошибками в начале кривой появляется заметный пик, содержащий, в основном, ошибочные записи (рис. 5). Именно его присутствие является индикатором корректной работы подхода, а также позволяет оценить точность результатов фильтрования и количество ошибок.

Для оценки полноты и точности фильтрования необходимо отделить ошибочные записи от неточно предсказанных. При наличии ошибок в распределении количества записей от *t%* присутствует минимум, большая часть ошибок лежит слева, остальные записи – справа. Кривую количества записей без ошибок правее минимума аппроксимировали при помощи прямой, параллельной оси абсцисс. Такой алгоритм позволяет оценить как количество ошибок в наборе данных – площадь над аппроксимирующей функцией слева от минимума (заполнена фиолетовым цветом на рис. 5), так и точность фильтрования – соотношение площади над аппроксимирующей функцией ко всей площади до минимума. Этот подход позволяет получить достаточно точную оценку этих величин снизу (рис. 6). Это позволяет проверить работоспособность подхода на новых данных и выбрать оптимальное значение *t%* такое, что ожидаемая точность остается высокой.

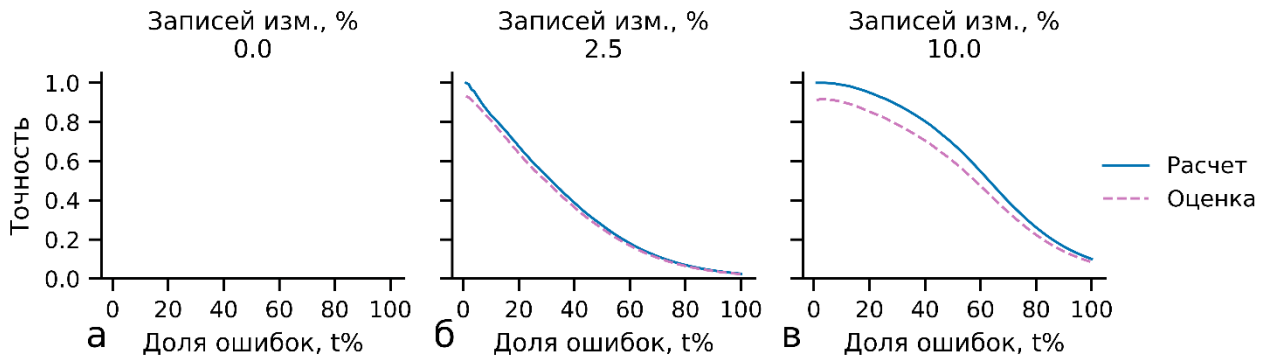


Рисунок 6. Зависимость точности нахождения ошибок от порогового параметра t% для QM9 наборов данных с разными долями перемешанных записей: а – 0%, б – 2.5%, в – 10%. В подходе «желтых карточек» t% наименее точно предсказанных записей для всех моделей одновременно считают ошибочными. Фиолетовая кривая – оценка, синяя – истинные значения точности

Алгоритм округления значений m/z до целочисленных

Значения *m/z* в масс-спектрах электронной ионизации даже низкого разрешения регистрируются в виде чисел с плавающей запятой, что обусловлено явлением дефекта массы и особенностями работы приборов. При обработке данных в программных пакетах, обучении предсказательных моделей и даже поиске по масс-спектральным базам данных используют округленные целые значения *m/z*. При этом, в разных программных пакетах округление может происходить по-разному, что может привести к разным результатам. Так, значения *m/z*, отвечающие одному иону, могут быть округлены к разным целым в зависимости от используемого алгоритма и случайных погрешностей.

Для поиска оптимального алгоритма округления сгенерированы два набора данных: на основе брутто-формул веществ из PubChem (2018) (1 480 811 записей) и всех комбинаторно возможных фрагментов на основе базы данных масс-спектров NIST 17 MS (14 269 691 запись). В исследовании рассматривали соединения с молекулярной массой меньше 600 Да. В каждом случае использовали только значение *m/z*, отвечающее наиболее интенсивному изотопному пику. Набор данных PubChem позволяет получить оценку снизу, так как не учитывает образование ионов, состав которых отличается от существующих молекул. Набор NIST MS позволяет получить оценку сверху, так как во внимание принимаются ионы, которые не могут образоваться в реальном эксперименте.

Построили распределение фрагментов для каждого из целочисленных масс-спектральных пиков, затем рассчитали зону, в которую попадает 95% от всего их количества (рис. 7а). Очевидно, что оптимальная граница округления лежит за границами этой зоны, внутри оставшегося интервала с минимальным количеством фрагментов. Эта граница изменяется в зависимости от самого значения *m/z* и выбранного набора – для PubChem эта зона шире, чем для NIST MS (рис. 7б).

Видно, что оптимальное правило округления зависит от значения m/z (рис. 7б) и изменяется нелинейно, однако для простоты и устойчивости алгоритма целесообразно

использовать одно значение – 0.62. Данное критическое значение получается усреднением границ для пиков в интервале [500;600], так как для меньших значений *m/z* плотность фрагментов в этой области меньше. Таким образом все фрагменты со значениями *m/z* в интервале [*MZ* - 0.38; *MZ* + 0.62] округляются до целочисленного *MZ*.

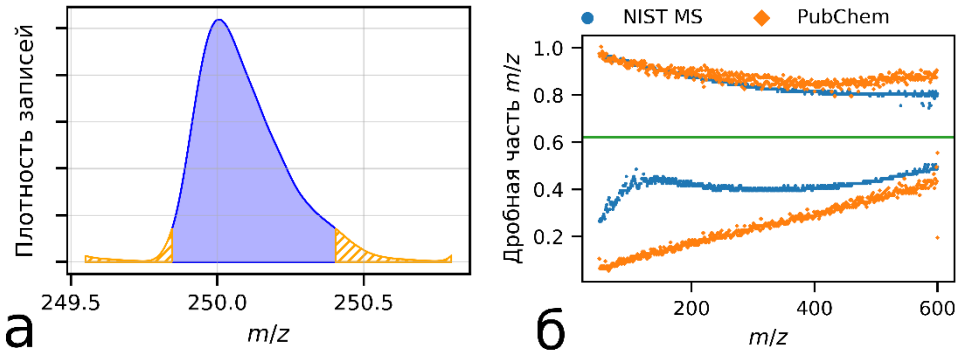


Рисунок 7. А – распределение точных m/z возможных фрагментов, основная часть площади отмечена синим, зона минимальной плотности – оранжевым. Б – Левая и правая границы оптимальной зоны границы округления для PubChem (оранжевый) и для NIST (голубой). Зеленым цветом отмечено предлагаемое правило для округления значений m/z до целочисленных (0.62).

Анализ результатов обработки масс-спектральных данных в доступных программах позволил установить алгоритмы округления, использующиеся в коммерческом и открытом программном обеспечении, и сравнили с предлагаемым подходом к округлению (табл. 1) при условии, что прибор позволяет определять значение m/z с точностью 0.05 и 0.1 Да. Предлагаемый подход позволяет минимизировать случаи, когда на результаты округления могут влиять случайные погрешности. Наиболее близкие результаты показывает программа AMDIS. Наихудшие результаты демонстрирует пакет ChemStation, который отбрасывает фрагменты с m/z в интервале [*MZ* + 0.4; *MZ* + 0.6], что приводит к значительному искажению результатов (табл. 1).

*Таблица 1. Вероятность округления значения m/z до целого, зависящая исключительно от случайных аппаратных погрешностей (Δm/z). Результаты представлены для нескольких программ. Значение x – такая дробная часть m/z, что все значения m/z с дробной частью меньшей x округляются в меньшую сторону, с большей – в большую. *ChemStation использует интервал [MZ - 0.4; MZ + 0.4] → MZ, при этом все значения m/z, не входящие в интервал, отбрасываются*

Программа	Граница округления x: [*MZ-1+x; MZ+x*] →*MZ*	*Δm/z* = 0.05 Да		*Δm/z* = 0.1 Да	
		PubChem, %	NIST, %	PubChem, %	NIST, %
ChemStation	n/a*	2.33	6.57	4.95	10.53
OpenChrom	0.5	0.37	1.46	1.02	3.58
Предлагаемая	0.62	0.11	0.50	0.34	1.18
AMDIS	0.649	0.15	0.50	0.43	1.20
ChromaTOF	0.7	0.3	0.67	0.88	1.69

Примерами случайного округления могут служить пики в масс-спектре высокого разрешения карбазола с *m/z* 82.5284 и 83.5362 фрагментов: $^{13}C_{12}{}^{1}H_{7}{}^{14}N^{++}$ и $^{13}C_{12}{}^{1}H_{9}{}^{14}N^{++}$

соответственно. В трех случаях эти значения *m/z* в масс-спектрах, представленных в NIST MS и MassBank, округлены в большую сторону, и один раз – в меньшую. Если бы эти масс-спектры были обработаны при помощи AMDIS или ChromaTOF, то значения *m/z* были бы округлены к меньшему значению, с ChemStation – исключены из рассмотрения, с OpenChrom – округлены случайным образом, в зависимости от ошибки определения массы. Несмотря на то, что этот пример неоднозначного округления вряд ли может вызвать значительные проблемы с идентификацией (рассмотренные пики являются низкоинтенсивными), он показывает, что проблема присутствует даже в современных масс-спектральных библиотеках.

Предсказание молекулярных отпечатков пальцев по масс-спектрам электронной ионизации

В нецелевом хроматомасс-спектрометрическом анализе желаемым результатом является идентификация неизвестного аналита. При этом размер масс-спектральных баз данных, с которыми можно сравнивать экспериментально полученный масс-спектр, на несколько порядков меньше, чем общехимических. Поэтому возникает задача установления структурных фрагментов непосредственно из масс-спектра. До широкого применения методов хемоинформатики ее частично решали вручную, однако это требует большого опыта и занимает длительное время. Для рутинной обработки больших объемов данных важно ускорить этот процесс, для чего применяют предсказательные модели машинного и глубокого обучения.

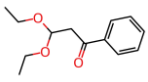
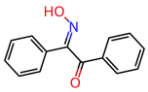
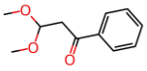
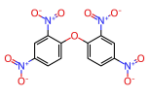
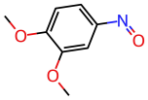
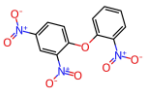
Структурные фрагменты молекулы могут быть описаны при помощи различных машиночитаемых молекулярных отпечатков пальцев. DeepEI является наиболее актуальной известной моделью для решения этой предсказательной задачи по экспериментальным масс-спектрам и основана на использовании небольших полносвязных нейронных сетей. Для каждого из выбранных значений молекулярных отпечатков пальцев авторы обучали отдельную модель, что было очень время- и трудозатратно и существенно затрудняло воспроизведение и использование подхода. В этой работе разработаны полносвязные нейросетевые модели EI2FP (Full – полная и Lite – облегченная) с оптимизированной архитектурой и гиперпараметрами, которые обучаются в 34 и 130 раз быстрее DeepEI, соответственно. В соответствии с оригинальной статьей DeepEI, сравнение производилось для 157 выбранных значений (95 MACCS из 166 + 62 ECFP6 из 1024) молекулярных отпечатков. Использовали 157 отдельных моделей DeepEI и по одной EI2FP:Full и EI2FP:Lite, каждая из которых предсказывала все 1024 ECFP6 и 166 MACCS за один проход, а не только выбранные по алгоритму авторов DeepEI. Архитектура EI2FP:Full выбрана таким образом, чтобы она могла предсказывать и другие молекулярные отпечатки пальцев одновременно, а EI2FP:Lite оптимизирована для предсказания ECFP6 + MACCS. При этом, модель EI2FP:Full превосходит, а EI2FP:Lite не уступает DeepEI по точности, правильности и полноте предсказания выбранных молекулярных отпечатков пальцев (табл. 2).

Таблица 2. Сравнение статистических параметров эффективности предсказания молекулярных отпечатков пальцев моделями DeepEI и EI2FP

Метрика		DeepEI	EI2FP:Full	EI2FP:Lite
Среднее	Правильность	0.89	0.92	0.91
	Точность	0.77	0.87	0.86
	Полнота	0.70	0.73	0.71
	F_1	0.73	0.79	0.77
Медиана	Правильность	0.88	0.91	0.91
	Точность	0.79	0.88	0.87
	Полнота	0.73	0.77	0.75
	F_1	0.75	0.82	0.80

Применение модели EI2FP:Lite продемонстрировано (табл. 3) на экспериментальных масс-спектрах электронной ионизации 3,3-диэтоксипропиофенона и бис(2,4,-динитрофенил)ового эфира из базы данных MassBank, которые отсутствуют в NIST MS.

Таблица 3. Примеры поиска с использованием молекулярных отпечатков пальцев, сравнение с часто применяющимся подходом Similarity

Исходное соединение	Лучший кандидат (Similarity)	Лучший кандидат (молекулярные отпечатки)
3,3-диэтоксипропиофенон 	β-Бензилоксим 	3,3-диметоксипропиофенон 
бис(2,4,-динитрофенил)овый эфир 	1,2-диметокси-4-нитрозобензол 	2,4-динитро-1-(о-нитрофенокси)бензол 

Стандартный подход к получению информации о структуре соединений, отсутствующих в базах данных заключается в проведении библиотечного поиска с

использованием алгоритма Similarity. Результаты применения такого подхода и модели EI2FP:Lite (предсказание молекулярных отпечатков и их поиск по затем поиск по базе сгенерированных молекулярных отпечатков для всех соединений из NIST MS) представлены в Таблице 3. Из таблицы видно, что применение модели EI2FP:Lite позволяет найти структуры в базе данных, которые гораздо ближе по структуре к «неизвестным».

Предсказание масс-спектров электронной ионизации по структуре молекулы

Как было отмечено ранее, в общехимических базах данных представлено на несколько порядков больше соединений, чем в масс-спектральных. При этом в нецелевом анализе возможно провести сравнение масс-спектров не только с экспериментально полученными, но и со сгенерированными, например, при помощи предварительно обученной модели, что позволяет сократить этот разрыв. В этой работе сравнили наиболее актуальные подходы к предсказанию масс-спектров электронной ионизации низкого разрешения моделями машинного и глубокого обучения – CFM-EI (2015), NEIMS (2019), RASSP (2023), MSProject (2025). Решения, основанные на квантовомеханических и молекулярнодинамических расчетах не рассматривали из-за большой длительности предсказаний, которая не позволяет сформировать сколько-нибудь большой набор синтетических масс-спектров.

Для сравнения точности предсказания использовали выборку из 5000 молекул, добавленных в NIST 23 MS, но отсутствующих в предыдущих изданиях. В качестве функции близости масс-спектров использовали наиболее широко применяющуюся для таких задач Similarity. Это позволило корректно сравнивать модели разных лет выпуска, так как вещества из выборки точно не попадали в соответствующие тренировочные наборы. Обнаружили, что подход NEIMS несмотря на то, что не является самым современным показывает наилучшие результаты, CFM-EI занимает второе место (рис. 8). Для более современных подходов не удалось независимо подтвердить утверждения авторов оригинальных публикаций. Так, обе модели RASSP показали крайне низкую точность предсказания с предобученными весами, предосавленными авторами. Не удалось воспроизвести решение MSProject, так как описание в оригинальной публикации недостаточно подробное, исходный код в открытом доступе отсутствует, так же как и готовая для запуска модель.

Были предложены и реализованы несколько модификаций NEIMS с использованием актуальной версии Python и библиотек. Предлагаемая реализация сохраняет основную идею оригинальной работы, заключающуюся в использовании архитектуры нейронной сети, состоящей из трех частей для предсказания: (1) масс-спектра, (2) спектра нейтральных потерь, (3) весов для объединения двух спектров. Структуру молекулы представляли в виде графа (вариант NEIMS-G) или молекулярных отпечатков пальцев (NEIMS-PyTorch), подбирали функции активации (ReLU, LeakyReLU, сигмоида), количество слоев и скрытых нейронов в них, гиперпараметры

обучения. Итоговый вариант NEIMS-PyTorch с оптимизированной полносвязной архитектурой превосходит оригинал по точности на выбранном тестовом наборе (рис. 8). Основное улучшение связано с увеличением доли точно предсказанных масс-спектров с (Similarity > 0.9). Так, для NEIMS это 5%, для предложенного подхода – 28%.

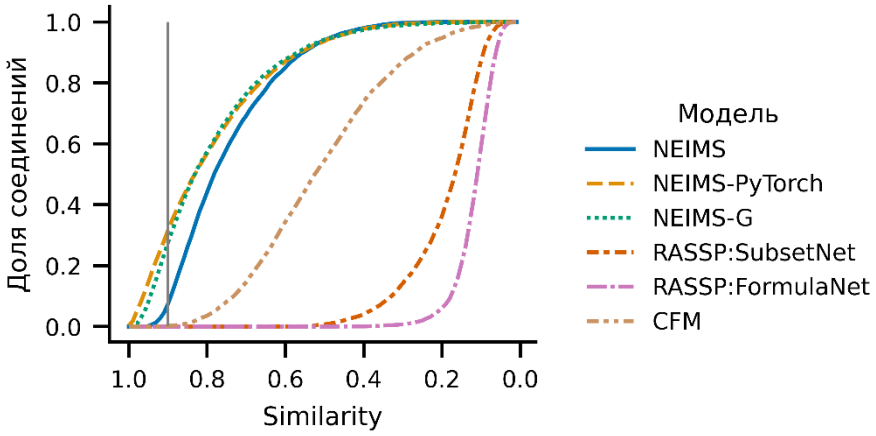


Рисунок 8. Кумулятивные распределения степени совпадения, рассчитанной между предсказанным моделями и экспериментальным масс-спектрами для 5000 пар масс-спектров

Кроме того, рассмотрели различные варианты распространения обученных предсказательных моделей с целью обеспечения кроссплатформенной и «временной» совместимости. Несмотря на то, что обычно наибольшее внимание уделяется повышению точности предсказания, простота воспроизведения подхода и применения готовой программы на практике являются не менее важными его для широкого применения в химических лабораториях. В случае решений, написанных с использованием компилируемых языков программирования, таких как CFM-EI (язык C++), открытый доступ к исполняемым файлам в сочетании с исходным кодом позволяет запускать проект даже через несколько лет после окончания развития, хоть и с некоторыми сложностями. Для большинства решений, реализованных на языке программирования Python, наиболее простым способом распространения является упаковка в библиотеку с известными требованиями к окружению и отдельное распространение весов. Сборка программы вместе весами модели, интерпретатором Python и всеми необходимыми зависимостями в Docker контейнер является оптимальным вариантом, так как позволяет применить подход в несколько простых шагов. Предложенное решение NEIMS-PyTorch распространяется именно таким образом, исходный код в виде библиотеки и веса обученной модели также находятся в свободном доступе.

ЗАКЛЮЧЕНИЕ

В ходе проведенных исследований предложен новый подход к поиску ошибок в химических базах данных. Подход основан на механизме голосования – ошибочными считаются записи, которые наименее точно предсказываются всеми моделями одновременно. Разработанное решение применено для поиска ошибок в базах индексов удерживания в газовой хроматографии (NIST RI) и времен удерживания в высокоэффективной жидкостной хроматографии (METLIN SMRT). В результате в NIST 17 RI найдено 2093 потенциально ошибочных записи, а в METLIN SMRT – 1544.

Для оценки эффективности и дальнейшей доработки подхода к поиску ошибок сгенерированы две группы синтетических наборов данных: на основе комбинации автокорреляционных дескрипторов и квантовомеханической базы QM9, в которые контролируемым образом добавлены ошибки разной природы и величины. Показано, что подход к поиску ошибок с голосованием моделей работает лучше более простых алгоритмов, известных ранее, например, сравнения значений из базы данных с предсказаниями одной модели или даже медианы предсказаний нескольких.

Подтверждено, что предсказательные модели способны правильно предсказывать даже ошибочные записи в химических наборах данных благодаря способности к обобщению. Исследована зависимость эффективности фильтрования от количества и архитектуры используемых моделей: модели с большей ошибкой предсказания, ожидаемо, демонстрируют более низкую точность и полноту по сравнению с более точными моделями. Показано, что наиболее консервативным и эффективным вариантом является использование нескольких предсказательных моделей с разной архитектурой.

Предложена расчетная величину – зависимость прироста количества записей, отмеченных ошибочными, от основного параметра подхода – доли наименее точно предсказанных записей для каждой из моделей. С использованием этой зависимости, сформулирован алгоритм выбора диапазона допустимых значений основного параметра, а также оценки снизу количества ошибочных записей и точности фильтрования. Это позволяет уменьшить количество ложноположительных результатов и проверить, насколько подход применим к выбранному набору данных. Подход реализован в виде библиотеки для языка программирования Python и выложен в открытый доступ.

В части работы, посвященной масс-спектрометрии, установлены алгоритмы округления значений *m/z* до целочисленных в широко используемых коммерческих и бесплатных программных пакетах OpenChrom, ChemStation, AMDIS, ChromaTOF. С использованием базы данных молекул PubChem и сгенерированного набора всех комбинаторно возможных фрагментов, основанного на NIST 17, проведено сравнение этих алгоритмов и предложен алгоритм округления, минимизирующий влияние случайных ошибок регистрации значений *m/z* на итоговый результат.

Экспериментальные масс-спектры электронной ионизации могут использоваться для поиска по базам данных не только напрямую, сравнением неизвестного с библиотечными или предсказанными по структуре молекулы масс-спектрами, но и предсказанием молекулярной структуры в виде молекулярных отпечатков пальцев методами машинного обучения. Более того, при таком подходе становится возможным сравнение с молекулами из общехимических библиотек PubChem, Chemical Abstracts и т.п., для которых молекулярные отпечатки пальцев можно рассчитать. Актуальный подход для решения этой задачи – DeepEI – медленно обучается и сложен для практического применения. В работе предложена архитектура полносвязной нейросетевой модели EI2FP:Full, которая более чем в 30 раз быстрее существующей модели DeepEI, а также превосходит ее по точности, правильности и полноте предсказания. Исходный код модели, упакованный в библиотеку для языка Python, предобученные веса, переменные окружения выложены в открытый доступ, обеспечивая простоту воспроизведения и практического применения.

Задача предсказания масс-спектра по структуре молекулы также актуальна для нецелевого анализа. В работе проведено сравнение не только по точности, но и по простоте использования доступных подходов к предсказанию масс-спектров с использованием набора из 5000 экспериментальных масс-спектров электронной ионизации, не входящих в тренировочные наборы предсказательных моделей. Показано, что NEIMS – один из первых подходов к решению этой задачи – до сих пор показывает наилучшую точность предсказания. Предложена усовершенствованная архитектура нейронной сети, позволяющая увеличить точность предсказания, решение также упаковано в виде библиотеки для языка программирования Python и выложено в открытый доступ вместе с переменными окружения и предобученными весами. Все необходимое для воспроизведения подхода также свободно распространяется в виде готового к запуску Docker контейнера, что упрощает воспроизведение подхода даже через несколько лет после окончания поддержки авторами и устаревания используемой версии Python и библиотек-зависимостей.

Полученные в ходе выполнения диссертационной работы научные результаты позволяют сделать следующие **выводы**:

1. Предложен подход к нахождению ошибок в хроматографических базах данных, основанный на механизме голосования нескольких независимых предсказательных моделей машинного и глубокого обучения. Сгенерированы синтетические наборы размеченных данных, проведена оценка эффективности поиска ошибок. Предложенный подход использован для обнаружения 2093 и 1544 потенциально ошибочных записей в базах данных NIST RI (индексы удерживания) и METLIN SMRT (времена удерживания) соответственно.
2. Предложен подход к округлению значений *m/z* до целочисленных, позволяющий уменьшить зависимость результатов от случайных приборных ошибок на примере

масс-спектров низкого разрешения ($\Delta m_{50\%} \sim 0.5$). Предложенный алгоритм округления основан на анализе большого числа органических соединений, представленных в базах данных, а также их возможных фрагментных ионов.

3. Улучшены известные подходы к предсказанию масс-спектров электронной ионизации по структуре молекулы и к предсказанию молекулярных отпечатков пальцев по масс-спектрам электронной ионизации. Доля точно (Similarity > 0.9) предсказанных масс-спектров увеличена с 5% (NEIMS) до 28% благодаря оптимизации архитектуры. Предложенный подход к предсказанию молекулярных отпечатков пальцев более чем в 30 раз быстрее обучается и работает по сравнению с известными аналогами при лучшей точности, правильности и полноте предсказания. Применение технологии контейнеризации позволило существенно упростить использование предсказательных моделей, а также обеспечить кроссплатформенную и «временную» совместимость.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

Научные статьи, опубликованные в рецензируемых научных журналах, рекомендованных для защиты в диссертационном совете МГУ по специальности и отрасли наук:

1) ***Khrisanfov M.**, Samokhin A.* A general procedure for rounding m/z values in low-resolution mass spectra // *Rapid Communications in Mass Spectrometry*. 2022. Vol. 36. № 11. e9294. EDN: PANTTY – 0.69 п.л. Вклад автора 60%. Импакт-фактор **2.0** (JIF).
2) ***Khrisanfov M. D.**, Matyushin D. D., Samokhin A. S.* A general procedure for finding potentially erroneous entries in the database of retention indices // *Analytica Chimica Acta*. 2024. Vol. 1297. 342375. EDN: OPAIVM – 0.81 п.л. Вклад автора 50%. Импакт-фактор **6.0** (JIF).
3) ***Khrisanfov M. D.**, Matyushin D. D., Samokhin A. S., Buryak A. K.* EI2FP: Efficient prediction of molecular fingerprints from electron ionization mass spectra // *Mass Spectrometry Letters*. 2024. Vol. 15. № 4. pp. 178-185. DOI: 10.5478/MSL.2024.15.4.178 – 0.92 п.л. Вклад автора 50%. Импакт-фактор **0.7** (JIF).
4) ***Khrisanfov M.**, Matyushin D., Samokhin A.* Finding potentially erroneous entries in METLIN SMRT // *Journal of Chromatography A*. 2025. Vol. 1745. 465761. EDN: IGUXUB – 0.69 п.л. Вклад автора 50%. Импакт-фактор **4.0** (JIF).

БЛАГОДАРНОСТИ

Автор выражает благодарность научному руководителю к.х.н. Самохину Андрею Сергеевичу; коллективу лаборатории «умных» методов химического анализа ИФХЭ РАН: Матюшину Дмитрию Дмитриевичу, к.х.н. Шолоховой Анастасии Юрьевне, к.х.н. Боровиковой Светлане Александровне; и д.х.н. Ревельскому Александру Игоревичу за помощь в выполнении работы и плодотворные обсуждения; д.х.н. Ставрианиди Андрею Николаевичу за ценные советы по тематике работы;

Автор выражает искреннюю признательность и благодарность семье, друзьям и коллегам за поддержку.

Автор выражает благодарность Министерству науки и высшего образования Российской федерации за финансовую поддержку выполненных исследований.