МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В.ЛОМОНОСОВА
ХИМИЧЕСКИЙ ФАКУЛЬТЕТ

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

ДИССЕРТАЦИЯ
на соискание учёной степени
кандидата химических наук

Научный руководитель:
кандидат химических наук
Самохин Андрей Сергеевич

Москва – 2026

Список используемых сокращений и обозначений

NIST (national institute of standards and technology) – национальный институт стандартов и технологий США

METLIN SMRT (METLIN dataset of small molecule retention times) – набор времен удерживания малых молекул METLIN

е.и.у. – единицы индексов удерживания

CDK (chemistry development kit) – химический набор разработчика

SMILES (simplified molecular input line entry system) – упрощенная система записи молекулярных структур

GNN (graph neural network) – графовая сверточная нейронная сеть

CNN (convolutional neural network) – сверточная нейронная сеть

MLP (multilayer perceptron, многослойный перцептрон) – полносвязная нейронная сеть

FCFP (fully connected fingerprint-based neural network) –полносвязная нейронная сеть, принимающая на вход молекулярные отпечатки пальцев как описание молекулы

FCD (fully connected descriptor-based neural network) – полносвязная нейронная сеть, принимающая на вход молекулярные дескрипторы как описание молекулы

RNN (recurrent neural network) – рекуррентная нейронная сеть

LSTM (long short-term memory) – рекуррентная нейронная сеть с учетом долгосрочных зависимостей

GRU (gated recurrent units) – рекуррентная нейронная сеть с управляемыми блоками

MEGNet (materials graph network) – графовая нейронная сеть для предсказания свойств материалов

PAGTN (path-augmented graph transformer network) – нейронная сеть типа графовый трансформер с информацией о цепочках атомов

ВЭЖХ – высокоэффективная жидкостная хроматография

ОФ-ВЭЖХ – обращенно-фазовая высокоэффективная жидкостная хроматография

ГХ – газовая хроматография

МС – масс-спектрометрия

EI (electron ionization) – электронная ионизация в масс-спектрометрии

RI (retention index) – индекс удерживания

RT (retention time) – время удерживания

MCMRT (multiple chromatographic methods-based retention time) – набор времен удерживания для разных хроматографических условий

InChI (international chemical identifier) – международный текстовый химический идентификатор

ADMET (absorption, distribution, metabolism, excretion, toxicity) – набор данных об адсорбции, распределении, метаболизме, токсичности молекул

AMDIS (automated mass spectral deconvolution and identification system) – автоматическая система деконволюции масс-спектров и идентификации компонентов

XML (extensible markup language) – расширяемый язык разметки

GNPS (global natural product social molecular networking) – глобальная система обмена масс-спектральными данными в биоинформатике

CFM (competitive fragmentation modelling) – моделирование конкурентной фрагментации; модель машинного обучения для предсказания масс-спектров

NEIMS (neural electron-ionization mass spectrometry) – нейросетевая модель для предсказания масс-спектров электронной ионизации

RASSP (rapid approximate subset-based spectra prediction) – быстрое предсказание масс-спектров с использованием поднаборов

MACCS (molecular access system) – вид бинарных молекулярных отпечатков пальцев

ReLU (rectified linear unit) – «выпрямленный линейный блок»; одна из базовых функций активации в искусственных нейронных сетях

GELU (gaussian error linear unit) – гладкий вариант ReLU с использованием нормального распределения

SiLU (sigmoid linear unit) – вариант ReLU с сигмоидным преобразованием

OHE (one-hot encoding) – бинарное кодирование информации

MAE (mean absolute error) – средняя абсолютная ошибка предсказания

MdAE (median absolute error) – медианная абсолютная ошибка предсказания

MPE (mean percentage error) – средняя относительная ошибка предсказания

ECFP (extended-connectivity fingerprint) – вид бинарных молекулярных отпечатков пальцев

CSV (comma-separated values) – табличный формат файлов со значениями, разделенными запятыми

AUC (area under curve) – площадь под кривой

ROC (receiver operating characteristic) – рабочая характеристика приемника или кривая ошибок

BCE (binary cross entropy) – бинарная перекрестная энтропия

Введение

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 алгоритмы, так и более комплексные подходы 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 для тривиальных, на первый взгляд, задач, например алгоритмов округления значений *m/z* в масс-спектрах до целочисленных или сравнения масс-спектров низкого разрешения далеко не всегда используют оптимальные подходы. Общепринятые и широко используемые решения 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 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 применяются 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.

Глава 1. Обзор литературы

Обзор литературы состоит из двух основных частей. Первая посвящена хроматографической части работы: индексам, временам удерживания, подходам к их предсказанию, а также к поиску ошибок в широко применяемых неразмеченных наборах данных. Вторая затрагивает вопросы, связанные с масс-спектрометрией электронной ионизации, связи масс-спектров со строением молекул, подходам, позволяющим решать как прямую, так и обратную задачи предсказания масс-спектра электронной ионизации по структуре молекулы.

1.1 Предсказание индексов удерживания газовой хроматографии

Индексы удерживания для газовой хроматографии широко применяются при нецелевом анализа и рассчитываются с использованием отношения разности времен удерживания аналита и ближайших к нему н-алканов. Благодаря использованию линейки н-алканов в качестве точки отсчета достигается хорошая воспроизводимость значений на разных приборах и хроматографических условиях эксперимента. Индексы удерживания были впервые предложены Ковачем в логарифмической форме для изотермических условий, затем доработаны Ван ден Дулом и Крацем [1] для программирования температуры. Эти два определения используются наиболее часто для соответствующих условий и численно близки для комбинации одной неподвижной фазы и вещества [2].

Уравнение для оригинальных логарифмических индексов удерживания Ковача [3]:

$$I = 100\left(\frac{log(t_s) - log(t_n)}{log(t_{(n+1)}) - log(t_n)} + n\right) \quad (1)$$

Линейные индексы удерживания рассчитывают следующим образом [1]:

$$I = 100\left(\frac{t_s - t_n}{t_{(n+1)} - t_n} + n\right) \quad (2)$$

где I – индекс удерживания (ИУ), t_s – время удерживания аналита, t_n и t_{n+1} – времена удерживания двух н-алканов, между которыми выходит аналит с n и $n+1$ атомов углерода, соответственно.

Несмотря на то, что оба варианта индексов удерживания нелинейно зависят от температуры, этими изменениями часто пренебрегают из-за их небольшой величины,

сравнимой со случайными экспериментальными погрешностями [2]. При схожих экспериментальных условиях индексы удерживания для большинства веществ воспроизводятся со средней ошибкой порядка 1-10 е.и.у [4] на разных системах. Однако в некоторых случаях при изменении температурной программы и других условий ошибка воспроизводимости даже на одном приборе может увеличиваться до 20 [5] и даже более 30 е.и.у [6]. Индексы удерживания часто используются при идентификации в хроматомасс-спектрометрии как дополняющая масс-спектр характеристика. Так, например, близкие структурные изомеры могут иметь очень похожие масс-спектры электронной ионизации из-за сходства путей фрагментации, но при этом значительно различаться по величине индекса удерживания, которые зависят от геометрической структуры молекулы. В этом случае информация об удерживании позволяет выявить более вероятного кандидата. При нецелевом анализе индексы удерживания могут также использоваться в качестве фильтра для отсечения наименее вероятных кандидатов.

Одна из наиболее часто используемых и полных баз данных индексов удерживания – NIST RI разных изданий. Наиболее актуальная версия NIST 23 RI содержит более 490 тысяч индексов удерживания для более чем 180 тысяч уникальных веществ. Однако это меньше половины от количества соединений, для которых доступны масс-спектры электронной ионизации и совсем малая доля от миллионов веществ, которые присутствуют в общехимических базах данных, таких как Chemical Abstracts [7], PubChem [8], ChemSpider [9]. Поэтому возникает задача предсказания индексов удерживания, которая зачастую решается методами машинного и глубокого обучения [10,11]. Существует большое количество работ, посвященных предсказанию индексов удерживания для небольших наборов данных (от десятков [12,13] до нескольких сотен веществ [14–17]). Для этого часто выбирают линейную регрессию и классические подходы машинного обучения, например регрессия на опорные вектора или случайный лес, так как они хорошо обучаются при небольшом количестве данных [18]. Другим подходом является обучение общей предсказательной модели для широкого круга веществ [19,20,18,21–25].В этом случае выбирают крупные наборы данных для обучения, которые позволяют задействовать более сложные предсказательные модели.

Алгоритмы предсказания совершенствуются – от простых полносвязных нейронных сетей и до ансамблей графовых трансформеров. Например, в подходе

MetExpert (2018) [19], одной из первых работ по предсказанию индексов удерживания с помощью нейронных сетей, использовали представление молекулы в виде 177 молекулярных дескрипторов из пакета Chemistry Development Kit (CDK) [26,27]. Полносвязная нейронная сеть содержала всего три слоя, при этом единственный скрытый слой содержал 88 нейронов, а в качестве функции активации использовали сигмоиду. Модель обучали на выборе из 2196 метаболитов из баз данных Golm Metabolome Database [28] и Adams Essential Oil Library [29]. В сверточной нейросетевой модели JCA19 (2019) [20], первой из обученных с использованием NIST RI, на вход подавали SMILES с бинарным кодированием: каждому символу в нотации SMILES был присвоен номер, каждая из SMILES записей была представлена в виде матрицы, где строки представляли возможные символы из набора, а столбцы – позиции в записи, с единицами на соответствующих пересечениях (по одной на каждый столбец). Архитектура состояла из двух сверточных слоев по 120 каналов с размером фильтра 6, слоя усреднения по каналам, а затем полносвязной части с 200 скрытых нейронов, активацией ReLU и выходным слоем.

Дальнейшим развитием стало применение ансамблей нейронных сетей и градиентного бустинга для предсказания индексов удерживания для неполярных и полярных неподвижных фаз с использованием NIST 17 RI (100 тысяч веществ) и меньших по размеру дополнительных баз данных (суммарно около 3 тысяч веществ) [18,21]. Градиентный бустинг обучали на дескрипторах из CDK, для полносвязной нейронной сети добавили молекулярные отпечатки пальцев. Для одномерной сверточной нейронной сети в составе ансамбля использовали бинарное кодирование SMILES строк, а для двухмерной сверточной – бинарное кодирование двухмерного представления молекулы. Такой подход позволил объединить предсказания четырех моделей для улучшения точности за счет разных входных данных и архитектур, улучшая способность к обобщению по сравнению с одиночными моделями [18]. Следующей была представлена двухмерная сверточная нейронная сеть DeepReI [23], которую обучали на выборке из NIST 14 RI с около 20 тысяч молекул. В этой работе не только усовершенствовали архитектуру нейронной сети добавлением остаточных связей, которые сглаживают градиенты в процессе обратного распространения ошибки при обучении, но и добавили графический интерфейс пользователя для уменьшения порога вхождения.

Наиболее актуальные архитектуры используют графовое представление молекул, которое наиболее полно отражает молекулярную структуру (атомы – вершины, а связи – ребра графа), а также позволяет значительно увеличить точность предсказания (2021) [22]. В архитектуру могут включать блоки MEGNet (MatErials Graph Network) [30] со сверточными графовыми и полносвязными слоями [22], которые позволяют учесть также и свойства всей молекулы. Также к графовым слоям применяют в том числе механизм внимания, напрмиер, в модели RIPred (2023) [24]. В модели AIRI (2024) [25] от сотрудников NIST применяется графовая архитектура с механизмом внимания PAGTN (Path-Augmented Graph Transformer Network) [31], которая позволяет учитывать длинные цепочки атомов. При этом используется ансамбль из 8 моделей с общей архитектурой, что позволяет дополнительно оценить неопределенность предсказания [25].

Как правило, в каждой следующей модели авторы улучшают точность предсказания, и ошибки предсказания наиболее актуальных моделей по величине сравнимы с экспериментальными погрешностями [32]. Во многих из них используют базу данных NIST RI из-за ее большого размера и широкой доступности. Так, например, согласно одному из наиболее актуальных независимых сравнений точности доступных универсальных подходов для предсказания индексов удерживания, 6 из 7 рассмотренных моделей обучали с использованием разных изданий NIST RI [32].

1.2 Предсказание времен удерживания жидкостной хроматографии

База данных METLIN является одной из наиболее широко применяемых для ВЭЖХ, доступный в ее составе набор времен удерживания для малых молекул (METLIN small molecule retention time, METLIN SMRT) [33] используется особенно широко. Это один из наиболее часто применяемых наборов данных для обучения предсказательных моделей времен удерживания для ВЭЖХ, как для обращенно-фазового варианта, так и для гидрофильной хроматографии (при использовании обучения с переносом). В отличие от NIST RI, METLIN SMRT находится в свободном доступе, что дополнительно увеличивает широту его распространения.

METLIN SMRT содержит времена удерживания для 80 038 молекул, их идентификаторы в PubChem, структуры и молекулярные отпечатки пальцев, рассчитанные при помощи программы Dragon 7 [33]. Все времена удерживания в этом

наборе были получены на ВЭЖХ хроматографе с квадрупольно-времяпролетным тандемным масс-спектрометром в качестве детектора (HPLC-Q-TOF) на колонке для обращенно-фазовой ВЭЖХ Zorbax Extend C-18 (2.1 x 50мм, 1.8 мкм) в одних и тех же хроматографических условиях для всех веществ.

С одной стороны, использование одного прибора, типа неподвижной фазы и условий проведения эксперимента значительно уменьшает количество переменных, которые нужно учитывать для предсказания времен удерживания. Это упрощает процесс предсказания и позволяет более глубоко изучить зависимость времени удерживания от структуры молекулы. С другой стороны, этот подход сильно сужает возможности для практического применения. При использовании моделей, обученных на данных METLIN SMRT, в других хроматографических условиях нужно проводить обучение с переносом (transfer learning), а значит – необходим собственный обучающий набор времен удерживания, содержащий несколько десятков, лучше сотен записей, полученных в целевых условиях. Любое изменение условий эксперимента потребует переобучения модели и создания нового тренировочного набора, что создает ограничения для практического применения.

До последнего времени у METLIN SMRT практически не было аналогов, которые бы могли быть использованы для обучения предсказательных моделей для ВЭЖХ. В начале 2024 был опубликован новый репозиторий времен удерживания ВЭЖХ – RepoRT [34]. Он содержит 373 набора данных, полученных в различных лабораториях на разных приборах, из которых 295 получены в обращенно-фазовых условиях и 71 – в условиях гидрофильной хроматографии. Также в состав входят несколько менее распространенных неподвижных фаз, например, пентафторфенол. METLIN SMRT также входит в состав RepoRT, в оставшейся части содержится всего 8 809 уникальных молекул.

Набор времен удерживания для нескольких хроматографических методов (Multiple chromatographic methods-based retention time, MCMRT) [35] также появился недавно и содержит 10 073 экспериментальных времени удерживания, каждое из которых является средним по трем параллельным измерениям для 343 молекул в 30 различных хроматографических условиях, полученных в УВЭЖХ с масс-спектрометрическим детектором типа орбитальной ионной ловушки.

В последние годы было опубликовано значительное число статей, посвященных предсказанию времен удерживания в ОФ-ВЭЖХ [36–39] и условиях гидрофильной хроматографии [40]. Наиболее точная из доступных моделей показала среднюю абсолютную ошибка предсказания меньше 30 секунд при использовании слоев с архитектурой «трансформер» вместе с графовыми слоями с механизмом внимания в модели RT-Transformer [41], что сравнимо с величиной экспериментальных погрешностей в тренировочном наборе. Однако предварительная очистка данных в этих работах была либо слишком упрощенной, либо не проводилась вообще.

1.3 Поиск ошибок в базах экспериментальных данных

Наиболее значительная часть исследований в области машинного обучения сфокусирована на улучшении методов предсказания [42]. В то же время широко известно, что предварительная подготовка и обработка данных требуют большого количества времени и усилий [43,44]. Очистка является одним из наиболее важных шагов в предварительной подготовке данных [45]. Несмотря на большую значимость этого этапа, его значение часто недооценивают, так как многие наборы данных считаются изначально свободными от ошибок [46]. Однако практика показывает, что даже широко используемые наборы данных, считающиеся «чистыми», такие как стандартный для области распознавания изображений ImageNet [47], могут содержать неточности [46]. В результате, в последние годы возрос интерес к подходам очистки данных [48].

В химии и смежных науках, работающих с молекулами, существует потребность в высококачественных наборах данных, содержащих разнообразные молекулярные свойства. Такие базы данных могут использоваться главным образом двумя способами: в качестве справочных в экспериментальной работе, а также для обучения предсказательный моделей. Размер наборов данных является наиболее критичным, однако необходимо также учитывать и качество данных. Несмотря на то, что модели машинного обучения могут частично компенсировать шум и ошибки в наборах данных благодаря обобщению (generalization), использование ошибочных данных в качестве справочных может приводить к неправильным экспериментальным результатам и интерпретациям. Поэтому, как разработчики баз данных, так и конечные пользователи

заинтересованы в развитии устойчивых и эффективных подходов к очистке от ошибочных записей.

Очистка набора данных молекулярных и химических свойств, таких как, например, хроматографические времена и индексы удерживания может рассматриваться как задача поиска неточных меток (label cleaning) и признаков (feature cleaning) одновременно. Ошибочными могут быть как молекулярные свойства, так и строковые представления молекул (SMILES, InChI), или молекулы и свойства могут быть перемешаны. Согласно одному из последних систематических исследований [48], рассматривающему 101 работу по поиску ошибок в нехимических данных, примерно 2/3 от всех работ посвящены признакам и меткам (32 и 36, соответственно) [48].

Примерно половина из работ по поиску некорректных меток (18) применяются к наборам изображений, но только 2 работы проводят фильтрацию ошибочных признаков для подобных данных. Менее четверти работ по поиску некорректных меток работали с табличными данными (8) и примерно 3/4 от работ по признакам (26). Насколько нам известно, вопросы очистки признаков и меток для наборов молекулярных свойств редко рассматриваются в научных публикациях. Наиболее близкие к представляемой работе исследования уменьшали неопределенность в данных об указанных в патентах реакциях (неверные реагенты, неточные условия эксперимента, некорректные литературные источники) при помощи комбинации из методов машинного обучения и больших языковых моделей [49], а также улучшали точность регрессии для предсказания адсорбции, распределений, метаболизма, токсичности (ADMET) для малых молекул, используя в качестве основы значения ошибки тренировочного набора, и феномен «катастрофического забывания» в трансформерах как основной инструмент [50]. Этот подход был также проверен на данных QM9 (“H298” и интервале “HOMO-LUMO”) [50].

Авторы упомянутого выше обзора [48] подчеркивают, что развитие подходов к очистке данных от ошибок сдерживается недостатком размеченных наборов. Создание баз данных высокого качества требует больших трудовых и временных затрат, даже в области информатики компьютерных технологий с расчетными данными, зачастую требует ручной разметки и, иногда, специальных знаний. Эта же задача становится еще более трудной для химии и смежных дисциплин, работающих с молекулами, потому что они опираются на экспериментальные данные. Исправление ошибок в таких

случаях требует не только ручной проверки, но также планирования и проведения дополнительных экспериментов, которые требуют участия квалифицированных сотрудников, использование дорогих инструментов и химических стандартов.

Два из наиболее часто используемых набора экспериментальных химических (хроматографических) свойств, которые используются в аналитической химии, – NIST RI [51] и METLIN SMRT [33]. Их основными достоинствами являются сравнительно большой размер (90 000 – 300 000 записей) и простота данных: как индексы, так и времена удерживания – скалярные величины, зависящие от структуры молекулы. Каждая запись включает в себя молекулярную структуру, условия эксперимента и сам хроматографический параметр удерживания. Молекулярные структуры могут быть закодированы в нескольких форматах: текстовые идентификаторы (SMILES, InChI), подходящие для сверточных сетей и трансформеров; молекулярные дескрипторы и отпечатки пальцев, которые можно рассчитать из текстовых идентификаторов и затем использовать в полносвязных нейронных сетях или в методах классического машинного обучения; и графовые представления, для графовых нейронных сетей.

Процесс наполнения первого издания базы данных индексов удерживания NIST RI в подробностях описан в литературе [52]. Алгоритм проверки качества записей был тщательно выверен и продуман, однако стоит отметить три проблемы. Во-первых, база данных на текущий момент содержит более 400 000 записей, что значительно затрудняет возможность ручной проверки даже сколько-нибудь значительной выборки. Во-вторых, подходы к предсказанию индексов удерживания, которые использовались для оценки правильности значений в 2007 году (линейная аддитивная схема – итоговое значение индекса является суммой «частичных» индексов функциональных групп в составе молекулы) были крайне неточными по сравнению с современными – средняя абсолютная ошибка понизилась с 45 единиц индекса удерживания (е.и.у) до 15 для неполярных неподвижных фаз [4,18]. В-третьих, литературные данные могут содержать ошибки [53] из-за разных экспериментальных условий [6], уникальных веществ, погрешностях в исходных литературных данных [54], а также отсутствия дублирующих записей для сравнения. Поэтому существует вероятность, что некорректные записи могли оказаться внутри базы данных, что отметили в нескольких работах сторонние исследователи [18,23] и даже сами разработчики [52]. Наиболее

вероятными источниками ошибок могут являться перепутанные экспериментальные данные, а также неправильно идентифицированные структуры веществ.

Используемые процедуры предварительной выборки и очистки хроматографических данных для машинного обучения обычно являются очень простыми, включающими выбор записей по молекулярным формулам, молекулярной массе, составу и т.п. Иногда упоминается [23,24] использование предсказательных моделей для предварительной очистки от потенциальных ошибок, однако процедура описана крайне кратко. Например, в одной из работ [23] на базе данных обучали небольшую нейронную сеть, после чего отбрасывали все записи, предсказанные с ошибкой больше 100 е.и.у. как ошибочные. Такой подход имеет право на существование для быстрой и грубой очистки перед обучением моделей, однако не подходит для серьезного обнаружения ошибок, так как не различает атипичные и ошибочные записи. При сравнении экспериментальных данных с базой предварительная очистка не проводится и доверие к литературным значениям остается на усмотрение исследователя в случае ручной проверки и не выполняется в случае автоматизации процесса.

По предварительным оценкам, около 80% соединений в NIST 17 RI имеют не более одного индекса удерживания для любой неподвижной фазы. Это не позволяет применить статистические подходы (например, гистограммы [55]) чтобы оценить достоверность значений. На эту проблему указывали и разработчики базы данных в статье, посвященной NIST 05 RI. По их оценкам, для 53% соединений не было дублирующих значений индексов удерживания [52]. Было также отмечено, что некоторые дублирующиеся значения были удалены при переносе данных из литературы. Записи с несколькими параллельными измерениями считали более достоверными по сравнению с имеющими только одно измерение, однако эта информация недоступна для конечного пользователя, поэтому различить записи с разной степенью достоверности невозможно. Более того, бывает, что добавленные дублирующие записи имеют один и тот же источник, например, статью [54]. Они могут быть получены при разных температурных программах на одном и том же оборудовании, поэтому не являются полностью независимыми и могут одновременно быть ошибочными из-за неверной идентификации.

Так же как и в NIST RI, большая часть соединений (80%) в METLIN SMRT не имеют дублирующих времен удерживания, поэтому найти такие потенциально ошибочные записи статистическими методами невозможно. Ручная проверка также не представляется возможной из-за большого размера базы данных. Авторы репозитория RepoRT также реализовали свой механизм очистки для контроля качества времен удерживания, попадающих в набор [34]. Они сравнивают экспериментальные значения с предсказаниями, полученными моделями градиентного бустинга (10-fold обучение), которые обучают на всех записях, полученных в одних хроматографических условиях.

Таким образом, ошибки в базах хроматографических данных могут возникать из-за неточностей в литературных источниках, при переносе данных, или в результате экспериментальных выбросов и неверно подобранных экспериментальных условий, неверной идентификации аналатов. До последнего времени многие из этих проблем не получали пристального внимания [6,23,24,52,53], а систематические исследования для выявления подобных аномалий не проводились.

1.4 Округление значений *m/z* в масс-спектрах низкого разрешения

Существует большое разнообразие производителей хроматомасс-спектрометрических комплексов, при этом каждый из них использует собственное программное обеспечение для обработки полученных данных. Например, ChemStation (Agilent), ChromaTOF (LECO), Xcalibur (Thermo). Кроме того, доступны бесплатные программы, такие как AMDIS (NIST), OpenChrom (Lablicate). При этом, зачастую, возможно произвести конвертацию данных из проприетарного формата, используемого производителем, в другой, доступный для обработки сторонним программным обеспечением, например NetCDF [56] или MZML [57]. Формат NetCDF широко используется для хранения и передачи научных данных, представленных в виде численных массивов. Наиболее часто его применяют для записи климатических данных (временные ряды температуры, влажности, направления и силы ветра и т.п.), однако он нашел место и для данных газовой хроматографии-масс-спектрометрии. В этом случае в виде отдельных массивов хранятся номера сканов, времена удерживания, значения *m/z*, значения интенсивности, также в файл записываются метаданные: тип прибора, температурная программа и другие экспериментальные условия. При этом

данные хранятся в двоичном (машиночитаемом) формате, то есть не могут быть прочитаны человеком напрямую через текстовый редактор, а метаданные могут записываться по-разному в зависимости от программы, в которой был сформирован файл. Однако в последнее время чаще используют формат MZML, базирующийся на XML разметке, который разработан специально для хроматомасс-спектральных данных, включает более полную и стандартизированную информацию об эксперименте [57], более широко распространен для данных жидкостной хроматомасс-спектрометрии и тандемной масс-спектрометрии.

В то время как NetCDF хранит данные в виде непрерывных массивов, которые нужно разделять на отдельные сканы, в MZML основной единицей является цельный масс-спектр, что повышает удобство работы с данными. Поскольку MZML является частично текстовым, а NetCDF полностью бинарным, то файлы MZML больше по размеру, медленнее считываются, однако с практической точки зрения эта разница невелика.

Выбор программы обычно обуславливается ее доступностью и личными предпочтениями исследователя, поэтому обработка одних и тех же данных несколькими программами происходит редко, в основном, при обучении сотрудников или сравнении программных пакетов. При этом, масс-спектры с целочисленными значениями *m/z*, полученные разными программами из одного и того же NetCDF файла, могут значительно различаться. Это явление вызвано разницей в алгоритмах, используемых для округления, поэтому полученные после импорта масс-спектры могут отличаться по набору *m/z* и их интенсивностей (вычитание фона не проводилось).

Округление значений *m/z* широко используется для масс-спектров как низкого, так и высокого разрешения в метаболомном и протеомном фингерпринтинге с новыми алгоритмами [58–62], машинном обучении (включая классификацию [63–65] и предсказание масс-спектров [66]), сравнении образов [67–69] (pattern matching), а такжев программах для обработки экспериментальных данных [70,71]. Использование округления уменьшает размерность данных и упрощает выравнивание масс-спектров. В задачах машинного обучения округление позволяет объединять данные в многомерные тензоры, что упрощает обработку и предсказания, а также поиск по масс-спектральным базам данных. Так, при обработке экспериментов в серии для создания

единого массива необходимо произвести выравнивание, чтобы объединить значения *m/z*, незначительно отличающиеся друг от друга из-за приборных погрешностей. Для этого нужно представить все возможные значения *m/z* в виде интервалов, для масс-спектров низкого разрешения их ширина обычно составляет 1 Да. Все значения, попадающие в один интервал, округляются до одного целочисленного *m/z,* а значения интенсивностей суммируются. Самым простым примером является математическое округление, которое используется наиболее часто. Однако некоторые программы (MetAlign [72], MSClust [73]) используют интервалы (bins) с другими границами для округления точных *m/z*. Программа MetAlign определяет интервал в 1 *m/z* [MZ – 0.35; MZ + 0.65] [70], а MSClust использует интервалы в 0.5 *m/z* ([MZ – 0.1; MZ + 0.4]; [MZ + 0.4; MZ + 0.9]) [71], где MZ – целочисленное значение *m/z*. Известно, что неоднозначное округление может приводить к ошибкам при обработке масс-спектров, например, нарушая разделение на интервалы в SpectraST [74]. Эта проблема также встречается в библиотеках MS^n [75].

1.5 Масс-спектры электронной ионизации в нецелевом анализе

Методы хроматомасс-спектрометрии часто используют для задач идентификации соединений или установления частичной структуры неизвестных в рамках скринингового или нецелевого анализа образцов сложного состава из-за высокой информативности метода. Полученные хроматомасс-спектральные данные при нецелевом анализе обычно обрабатывают двумя способами: проводят сравнение масс-спектров неизвестных веществ с базой экспериментальных (NIST MS [76], Wiley Registry of Mass Spectral Data [77], GNPS [78] и т.д.) или предсказанных данных; или проводят предсказание структуры, ее частей, функциональных групп или комбинаций атомов по масс-спектру.

1.5.1 Предсказание масс-спектров

Задача предсказания масс-спектров часто возникает, так как количество соединений, масс-спектры которых доступны в базах данных, на несколько порядков меньше, чем количество известных органических веществ. В настоящее время предложено несколько подходов для предсказания масс-спектров электронной ионизации по структуре молекулы с применением машинного обучения – CFM-EI (2016) [79], NEIMS (2019) [80], NEIMS-GNN (2020), RASSP (2023) [81], MSProject

(2025) [82]. Следует отметить, что наиболее актуальные архитектурные решения и приемы обучения нейросетевых предсказательных моделей, в последнее время, применяют для предсказания тандемных масс-спектров [83,84], которые значительно более информативные по сравнению с масс-спектрами электронной ионизации. Существуют также подходы, такие как QCxMS [85], QCEIMS [86] и QCxMS2 [87], основанные на использовании квантовохимических расчетов для оценки состояния ионов и методов молекулярной динамики для прогнозирования вероятности различных путей (деревьев) фрагментации. Однако они работают на несколько порядков медленнее моделей машинного обучения, поэтому применяются значительно реже.

CFM

Подход к предсказанию CFM (competitive fragmentation modelling) изначально был представлен в 2015 для тандемных масс-спектров (CFM-ESI) [88], годом позже был адаптирован для масс-спектров электронной ионизации (CFM-EI) [79], для обучения использовали NIST 14 MS. CFM – вероятностная модель, работающая в несколько шагов. На каждом из них предсказываются массы фрагментов и их интенсивности, при этом в каждом случае выбираются наиболее вероятные разрывы связей. Для оценки вероятности используется нейронная сеть, при этом все возможные варианты (включая отсутствие разрывов) конкурируют между собой. Для каждого из финальных фрагментов рассчитывается изотопное распределение, конечный предсказанный масс-спектр получается из этих значений с учетом вероятностей образования фрагментов.

Такой подход позволяет получать более физически обоснованные предсказания масс-спектров, с точными массами и пиками соответствующими конкретными фрагментами, однако является крайне медленным. Исходный код CFM написан на языке программирования C++ и зависит от библиотек Boost [89], RDKit [90] и LPSolve [91]. Для операционной системы Windows доступны предварительно скомпилированные исполняемые файлы, которые и были использованы в работе для сравнения результатов. Также доступны веса нейросетевой модели, предсказывающей вероятности образования фрагментов. Для запуска на других операционных системах необходима компиляция, исходный код для этого доступен в репозитории проекта. Существует также сторонний графический интерфейс, а также скомпилированные под Linux файлы в проекте SVEKLA [92].

NEIMS

Подход NEIMS [80] был первым в своем роде с идеей предсказательной модели – «черного ящика», которая не основана на фрагментации, а преобразует представление молекулы в масс-спектр электронной ионизации напрямую. Нейронная сеть NEIMS состоит из трех основных частей: первая предсказывает непосредственно масс-спектр электронной ионизации, вторая – массы нейтральных потерь, третья – объединяет две предыдущие части, предсказывая, какая из них наиболее вероятна в каждом отдельном случае. В изначальном варианте используют молекулярные отпечатки пальцев для представления структуры молекулы. Архитектура состоит из полносвязных слоев, поэтому предсказывается фиксированное количество дискретных значений *m/z*. Позже, архитектуру дорабатывали третьи стороны, используя графовое представление молекулы [93] и увеличивая количество дискретных значений *m/z*, получая возможность предсказывать масс-спектры высокого разрешения.

RASSP

Подход RASSP (rapid approximate subset-based spectra prediction) состоит из двух независимых нейросетевых предсказательных моделей: SubsetNet и FormulaNet [81]. FormulaNet предсказывает вероятности для различных молекулярных формул. Используется исходный молекулярный граф, из которого комбинаторно генерируются все возможные комбинации химических брутто-формул, которые могут получиться при разрыве связей. Для каждого из таких фрагментов предсказывают вероятность образования, а также рассчитывают изотопное распределение. Для получения финального масс-спектра эти теоретические фрагментарные спектры суммируют с учетом вероятности образования каждого продукта.

В случае SubsetNet используются не формулы, а наборы атомов, которые получают имитацией разрывов связей и перегруппировок, для которых похожим образом рассчитывают вероятности образования и получают итоговый масс-спектр.

Авторы оригинального исследования обучали модели на NIST 17 MS.

1.5.2 Предсказание молекулярных отпечатков пальцев

Задача получения молекулярной структуры напрямую из масс-спектра является одной из самых актуальных в хемоинформатике и нецелевом анализе. Однако она до сих пор не решена полностью, то есть в редких случаях возможно получить

правильную структурную формулу молекулы только по масс-спектрам, как электронной ионизации, так и тандемным. Молекулярные отпечатки пальцев можно рассчитать по структуре молекулы с использованием широкого круга доступных библиотек: CDK [27] для Java, RDKit [90] и Scikit-fingerprints [94] для Python; и полноценных программ: PaDEL [95], Dragon [96], alvaDesc [97]. Многие из них также позволяют проводить сравнение структур между собой и искать похожие молекулы в базах данных.

Молекулярные отпечатки пальцев часто используют в качестве входных данных для моделей машинного и глубокого обучения для предсказания физико-химических свойств молекул: индексов удерживания Ковача [21], масс-спектров электронной ионизации [81], времен удерживания в обращенно-фазовой ВЭЖХ [36] и т.п.

Одни из наиболее распространенных видов молекулярных отпечатков пальцев – CDK [27], PubChem [98], Klekota-Roth [99], MACCS [100], ECFP или Моргана [101,102], функциональные FCFP [101], Daylight [103]. Молекулярные отпечатки PubChem, Klekota-Roth, MACCS имеют заданную длину – 4860, 881, 166 соответственно, каждый из них описывает наличие или отсутствие заранее заданной комбинации атомов или функциональной группы. Другие молекулярные отпечатки пальцев, такие как ECFP, CDK, FCFP могут иметь разную длину, так как основаны на расчете хэш-функции для выбранной комбинации атомов. Наиболее часто используют векторы длиной 1024. Если основываться на результатах поиска по Google Scholar, то молекулярные отпечатки ECFP применяются гораздо чаще всех других, за ними следуют PubChem и MACCS.

Существует несколько подходов к предсказанию молекулярных отпечатков пальцев по масс-спектрам. Например, CSI:FingerID [104] использует деревья фрагментации для аннотации экспериментального масс-спектра и предсказания молекулярной структуры. MetExpert [19] предсказывает 86 комбинаций атомов при помощи проекции на латентные структуры и дискриминантного анализа.

DeepEI

Подход DeepEI [105] является по-своему уникальным: каждый бит молекулярных отпечатков пальцев предсказывают при помощи отдельной полносвязной нейросетевой модели. Авторы исходной работы сгенерировали 8034 бит различных молекулярных отпечатков (CDK, PubChem, Klekota-Roth, MACCS, Estate и

ECFP) и выбрали 636, которые были сбалансированно представлены – от 10 до 90% молекул в NIST содержали этот отпечаток. Затем обучили 636 отдельных моделей, примерно 6.5 млн параметров в каждой, суммарное количество параметров составило более 4 млрд, что сравнимо с не самой маленькой современной большой языковой моделью.

Несмотря на то, что идея предсказания молекулярных отпечатков пальцев вполне популярна [106–109], многие авторы не использовали подход DeepEI непосредственно в своих работах. Некоторые взяли его за основу и создали свои инструменты и предсказательные модели [110], которые могли быть более узкоспециализированными [111], или использовали другие типы спектров [112,113]. Другие исследователи отметили, что планируют использовать DeepEI в будущих работах [114]. Можно предположить, что многие из них не стали использовать DeepEI напрямую из-за сложного процесса обучения и управления несколькими сотнями нейронных сетей, а также проблемами при интеграции этого подхода в свои отработанные алгоритмы обработки данных.

Более современные подходы, в основном, используют данные тандемной масс-спектрометрии и масс-спектры высокого разрешения. Так, MSNovelist [115] предсказывает молекулярные отпечатки и молекулярную структуру по тандемным масс-спектрам, как и Mass2SMILES [116]. В IDSL_MINT [110] применяют трансформер для этой же задачи. Поскольку новые статьи, в основном, сконцентрированы на более информативных данных тандемной масс-спектрометрии [117], масс-спектры электронной ионизации отошли на второй план. Несмотря на это, она не решена до конца, так как у подхода DeepEI множество ограничений.

Постановка задачи

Таким образом, при обработке хроматомасс-спектральных данных используется широкий круг вычислительных алгоритмов, методов машинного и глубокого обучения. При этом, не все из применяемых алгоритмов оптимальны, некоторые, например, для округления значений *m/z* в масс-спектрах низкого разрешения не опубликованы, так как используются в коммерческом программном обеспечении, а многие предсказательные подходы сложны для воспроизведения. Известны также и проблемы с базами данных: специализированные масс-спектральные библиотеки уступают общехимическим на

порядок и более по числу уникальных молекул, в хроматографических наборах существуют ошибки, которые затруднительно обнаружить статистическими методами и крайне затратно перепроверять экспериментально.

Это позволяет сформулировать основные задачи работы:

1. Предложить подход к поиску ошибок в базах хроматографических параметров удерживания, основанный на использовании предсказательных моделей.
2. Установить использующиеся алгоритмы для округления значений *m/z* в масс-спектрах низкого разрешения до целочисленных. Предложить подход, уменьшающий вероятность невоспроизводимого округления, зависящего от случайной приборной погрешности.
3. Усовершенствовать подходы к предсказанию масс-спектров электронной ионизации по структуре молекулы и к решению обратной задачи – получению молекулярных отпечатков пальцев по масс-спектрам.

Глава 2. Методы и алгоритмы

2.1 Архитектуры предсказательных моделей

2.1.1 Функции активации

Функции активации играют важную роль в глубоком обучении, потому что именно они обеспечивают нелинейное поведение моделей. Так, линейная комбинация любого количества полносвязных слоев может быть выражена в виде одного линейного преобразования [118]. При уменьшении размерности это преобразование превращается в метод главных компонент. Добавление же нелинейных функций позволяет описывать более сложные закономерности.

Одной из первых примененных нелинейных функций стала сигмоида (Sigmoid) [118]:

$$\sigma(x) = \frac{1}{1 + exp(-x)} \tag{3}$$

где x – входное значение или вектор. Ее поведение чем-то похоже на поведение биологических нейронов. На текущий момент эта функция часто применяется на выходе моделей в задачах, подобных двухклассовой регрессии, когда вывод нейронной сети преобразуется в интервал (0,1), крайние значения которого соответствуют выбранным классам.

Другая родственная функция – гиперболический тангенс [118]:

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{4}$$

где x – входное значение или вектор. Гиперболический тангенс и сигмоида связаны линейным преобразованием, однако при использовании в качестве функций активации в скрытых слоях демонстрируют разное поведение при обучении модели, так как веса слоев нейронной сети инициализируют разными значениями.

Обе функции ограничены в области больших отрицательных и положительных значений, что может приводить к затуханию градиентов при обучении, что, фактически, его останавливает.

Нелинейная функция активации ReLU является одной из наиболее часто применяемых и, фактически, стандартной для моделей глубокого обучения [118]. Она может быть представлена в следующем виде:

$$ReLU(x) = x^+ = max(0, x) \quad (5)$$

где x – входные значения. Ее простота и эмпирически проверенная эффективность привели к ее широчайшему применению, проблемы же могут возникать из-за негладкого перехода около точки 0 и полного обнуления отрицательных значений (может вызвать затухание градиентов) [118].

Для решения проблемы затухания градиентов в отрицательной области аргументов функции ReLU была предложена функция LeakyReLU:

$$LeakyReLU(x) = max(0, x) + \alpha min(0, x) \quad (6)$$

где x – входной вектор или значение. Второй член уравнения добавляет отрицательную обратную связь и решает проблему затухания градиентов.

Функция активации GELU лучше подходит по сравнению с ReLU для описания сложных функций благодаря свой немонотонности и кривизне в любой точке, достигнутой благодаря вогнутости. Кроме того, было добавлено случайное поведение умножением входных значений на кумулятивное нормальное распределение:

$$GELU(x) = x\Phi(x) = x * 0.5\left(1 + erf\left(x/\sqrt{2}\right)\right) \quad (7)$$

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt \quad (8)$$

где x – входное значение или вектор, $\Phi(x)$ – кумулятивная функция нормального распределения. При уменьшении значения x увеличивается вероятность, что это значение будет пропущено, как при использовании отдельного слоя случайного исключения [119].

SiLU (также известная как Swish) была предложена в той же работе, что и GELU [119], и охарактеризована как более быстрая и простая альтернатива, возможно, с чуть более скромными результатами. Входные значения x домножаются на сигмоиду, примененную к ним же:

$$SiLU(x) = x * \sigma(x) = x/(1 + e^{-x}) \quad (9)$$

Было показано, что даже простая замена ReLU на SiLU позволяет улучшить результаты для глубоких нейросетевых моделей [120]. Предполагается, что это является прямым следствием свойств самой функции: неограниченность при x>>1

позволяет избежать насыщения и уменьшения скорости обучения из-за маленького размера градиентов, ограничения при x << 0 выражаются в сильном регуляризующем влиянии на нейросетевую модель. Немонотонность SiLU улучшает распространение градиентов по сравнению с ReLU, а ее гладкость улучшает сходимость при оптимизации и способность нейросетевой модели к обобщению [120].

2.1.2 Полносвязная нейронная сеть

В работе использованы нейронные сети типа «многослойный перцептрон» (multilayer perceptron, MLP) или полносвязные (dense, fully-connected, FC) нейронные сети. Это самый простой вид нейронных сетей, каждый из слоев которой можно представить следующим уравнением (пример для одного образца) [121]:

$$\hat{y} = f(\overline{W} \times \overline{\mathrm{X}}^{\mathrm{T}} + \overline{\mathrm{B}}) = f(\sum_{i=1}^{d} w_i * x_i + b_i) \quad (10)$$

где $\hat{y}$ – предсказанное значение, $\bar{X}$ – вектор входных данных (x_i) размера d, $\overline{W}$ и w_i – обучаемые веса полносвязного слоя, b_i – обучаемый вектор смещения, $f\ (...)$ – функция активации. Благодаря соединению множества таких слоев и использованию нелинейных преобразований возможно эффективно аппроксимировать зависимости крайне сложной формы.

В данной диссертации полносвязные нейронные сети использовали для предсказания значений (индексов удерживания, времен удерживания, масс-спектров), связанных с молекулярной структурой. На вход такой модели поступают табличные данные, где каждый из образцов (молекул) описывается при помощи вектора значений. Наиболее часто используемые представления – молекулярные отпечатки пальцев [101,102], описывающие в виде длинного вектора нулей и единиц (Рис. 1) комбинации атомов и функциональных групп, присутствующих в молекуле, а также молекулярные дескрипторы – набор вещественных чисел, описывающих заданные свойства молекулы. Оба варианта представлений использовали в данной работе.

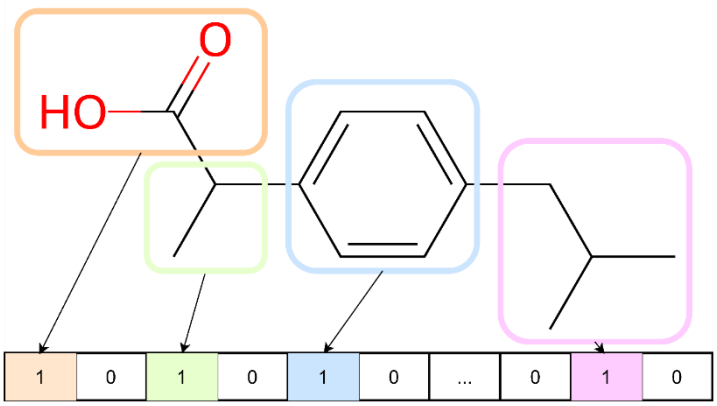


Рисунок 1. Схематичное представление функциональных групп молекулы в качестве значений (битов) в молекулярных отпечатках пальцев.

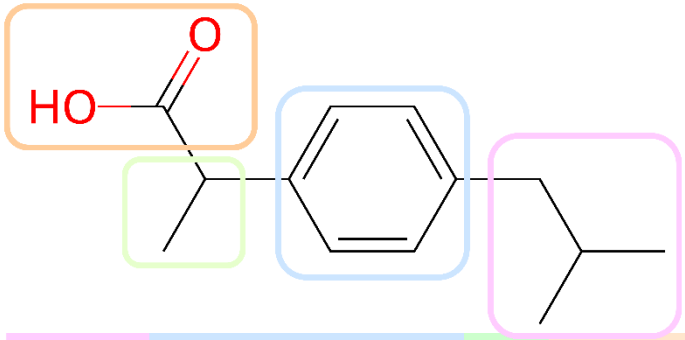
2.1.3 Сверточная нейронная сеть

В работе использовали сверточные нейронные сети (convolutional neural network, CNN). Каждый из слоев этой сети обладает следующими параметрами: размер фильтра (kernel size, k), смещение (stride, s), количество каналов (channels, c). Фильтр применяется через скалярное произведение последовательно к k значениям вектора входных параметров, каждый раз смещаясь на величину s, для получения выходного вектора $\hat{y}$ [118] . Эта операция повторяется с использованием различных значений фильтра для получения M выходных векторов. Так, для позиции t и заданного выходного канала m уравнение, описывающее свертку, будет выглядеть следующим образом:

$$\widehat{y_m}[t] = f\left(\sum_{c=1}^{C_{in}}\sum_{k=0}^{K-1} w_{m,c,k} * x_{c,t*s-k} + b_m\right) \quad (11)$$

В качестве входных данных для молекулярных структур обычно используют SMILES (Simplified Molecular Input Line Entry System) строки. Они отражают структуру молекулы, входящие в состав атомы, связи между ними, включая циклы и разветвления, а также может включать сведения о конфигурации двойной связи, при этом каждой молекуле соответствует больше одной строки SMILES. Для преобразования в машиночитаемый формат в данной работе использовалось бинарное (one-hot) кодирование. Каждый символ SMILES кодируется при помощи одной единицы, соответствующей номеру символа в векторе длиной, равной количеству всех доступных SMILES символов (Рис. 2). Такое кодирование может использоваться также

и для рекуррентных (RNN, LSTM, GRU) и трансформерных нейросетевых моделей, работающих с последовательностями символов или токенов.



CC(C)CC1=CC=C(C=C1)C(C)C(=O)O

C	C	(	C	)	...	(	=	O	)	O
0	0	0	0	0	0	0	0	0	0	0
1	**1**	0	**1**	0	0	0	0	0	0	0
0	0	**1**	0	0	0	**1**	0	0	0	0
0	0	0	0	**1**	0	0	0	0	**1**	0
0	0	0	0	0	0	0	**1**	0	0	0
0	0	0	0	0	0	0	0	**1**	0	**1**

Рисунок 2. Бинарное кодирование (one-hot) строки SMILES, отвечающей структуре молекулы

2.1.4 Графовая сверточная нейронная сеть

В работе использовали графовые сверточные нейронные сети (graph neural network, GNN). Они работают с графами – структурами, состоящими из вершин и ребер. Такие нейронные сети могут работать непосредственно со структурой молекулы – атомы являются вершинами, а связи, со всеми их параметрами, – ребрами такого графа. В параметрах вершин можно закодировать не только тип атома, но и другие параметры, такие как степень окисления, гибридизацию и т.п.

Самые простые реализации графовых сверточных нейронных сетей работают во многом аналогично сверточным нейронным сетям – операции свертки выполняются над соседями [121]. Например, операция обновления весов для вершины i может быть представлена так [122]:

$$x_i' = f(B_i * x_i + \sum_j e_{i,j} * x_j) \quad (12)$$

где x_i' – новые веса для вершины i, B_i– смещение, $e_{i,j}$ – вес связи между атомами i и j, x_j – веса вершины j, $f(...)$ – функция активации. Основное отличие состоит в том,

что у вершины графа может быть большее количество соседей, по сравнению с фиксированными двумя соседними символами в строке SMILES, поэтому полученные значения свертки нормируются на количество соседей.

2.1.5 Градиентный бустинг

В градиентном бустинге используют ансамбль «слабых» предсказателей, обычно, решающих деревьев, для получения конечного предсказания. При этом, процесс происходит итеративно – решающие деревья обучаются по очереди, каждая последующая модель стремится минимизировать разность между предсказанием всех предыдущих моделей и желаемый результат [123]. Для этого используют метод градиентного спуска.

Метод градиентного бустинга является одним из самых мощных классических в машинном обучении. Современные реализации, такие как XGBoost [124] и CatBoost [123] позволяют с небольшими усилиями достичь хорошей точности предсказания, сравнимой с нейросетевыми моделями. Табличные данные подаются на вход модели, соответственно, для представления структуры молекулы используют молекулярные отпечатки пальцев и дескрипторы, как и для полносвязных нейронных сетей.

2.1.6 Предсказание индексов удерживания для фильтрования NIST 17 RI

Для фильтрования NIST 17 RI использовали 5 независимых предсказательных моделей (Рис. 3): одномерную сверточную нейронную сеть (CNN(1D)), двухмерную сверточную нейронную сеть (CNN(2D)), полносвязную нейронную сеть (MLP), градиентный бустинг в реализации CatBoost и метод регрессии на опорные вектора (SVR). Для обучения моделей использовали среднюю абсолютную ошибку.

CNN(1D): SMILES представление молекулы закодированы в one-hot виде, 34 канала во входном сверточном слое, 2 одномерных сверточных слоя с 300 каналами на выходе, размер фильтра 6, смещение 1, затем общее усреднение, два полносвязных слоя с 600 скрытыми нейронами, в которые также поступала информация о неподвижной фазе, обучали в течение 150 эпох с шагом $1*10^{-4}$, размером батча 128 и оптимизатором AdamW.

CNN(2D): структура молекулы размещена в пространстве 65*65, тип атома или середина связи закодированы one-hot способом. Использовали 6 двухмерных

сверточных слоев (3 группы по 2 слоя), фильтры 3x3, смещение 1, со слоем максимальной выборки (пулинговый слой) между группами слоев, а также промежуточными соединениями через слой (residual connection) и всех сверточных слоев ко входу полносвязной части. Обучали в течение 70 эпох с шагом$1*10^{-4}$, размером батча 64 и оптимизатором AdamW.

MLP: молекула закодирована в виде 1604 дескрипторов из Mordred [125] и 166 MACCS запросов из RDKit [90] и 1024 молекулярных отпечатков пальцев ECFP4. В дескрипторной части – два полносвязных слоя с гиперболическим тангенсом в качестве функции активации, 300 скрытых нейронов в каждом. В части с молекулярными отпечатками пальцев один входной полносвязный слой, далее две группы по два полносвязных слоя с 1200 скрытых нейронов каждый, с промежуточными соединениями между группами. Две части объединяются группой из двух полносвязных слоев с 600 скрытых нейронов каждый. Обучали в течение 300 эпох с шагом $1*10^{-4}$, размером батча 512 и оптимизатором AdamW.

CatBoost: молекулу описали при помощи тех же дескрипторов и молекулярных отпечатков пальцев, что и для MLP. Обучали в течение 2000 итераций, с шагом 0.1 и среднеквадратичной функцией потерь.

SVR: использовали то же самое описание молекулы, что и в двух предыдущих подходах. Выбрали линейный вариант с $\varepsilon=0.0$, tol=0.0001, C=1.0, функцией потерь L1 и ограничением в 1000 итераций.

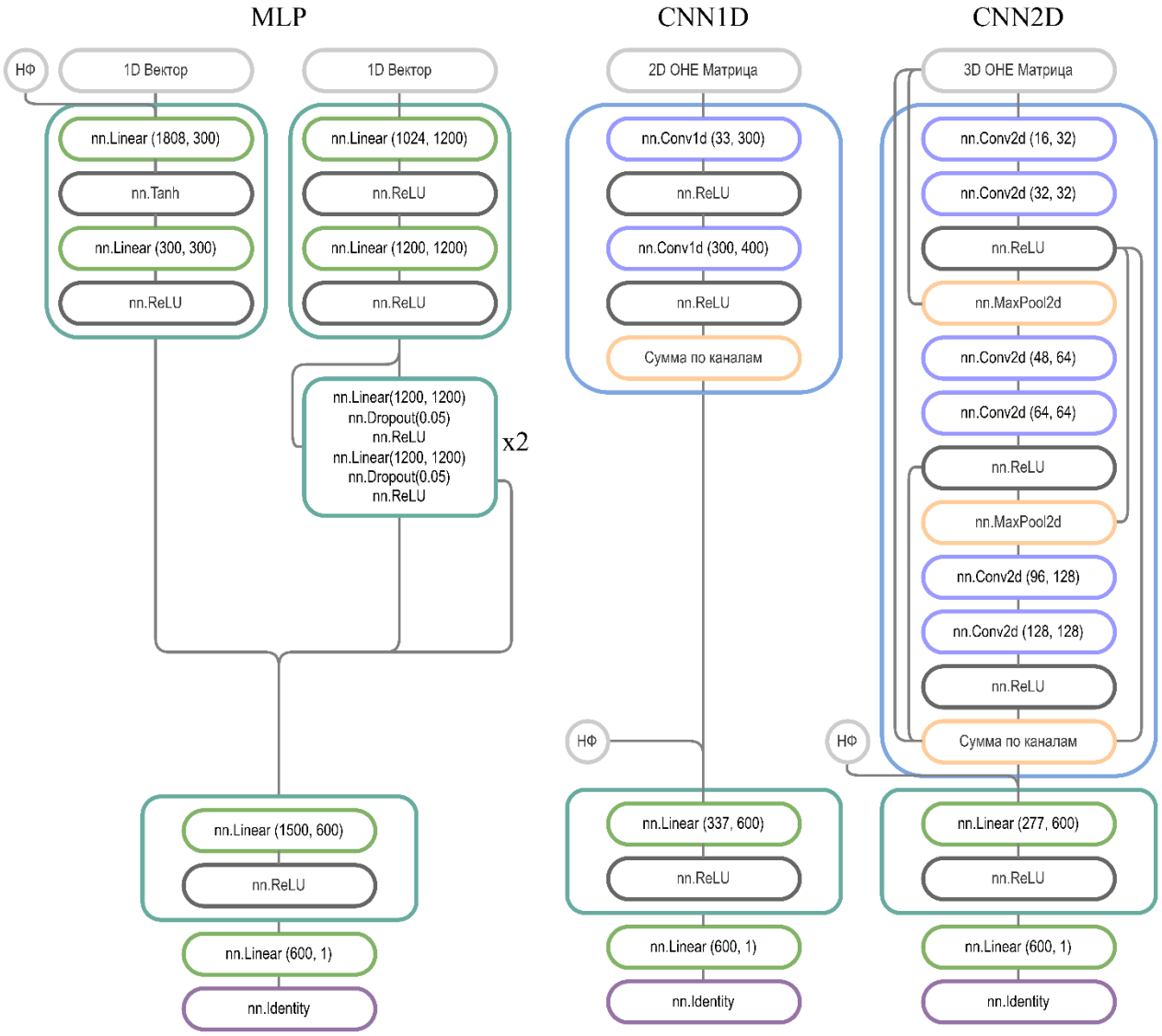


Рисунок 3. Архитектуры нейросетевых моделей для предсказания индексов удерживания, обученных на NIST 17 RI. НФ – информация о неподвижной фазе. OHE (one-hot-encoding) – бинарное кодирование.

2.1.7 Предсказание времен удерживания и поиск ошибок в METLIN SMRT

Использовали пять независимых предсказательных моделей (Рис. 4, Табл. 1): полносвязную нейронную сеть с дескрипторными входными данными (FCFP), полносвязную нейронную сеть с входными данными на основе молекулярных отпечатков пальцев (FCFP), одномерную сверточную нейронную сеть (CNN), графовую сверточную нейронную сеть (GNN), а также градиентный бустинг в виде CatBoost.

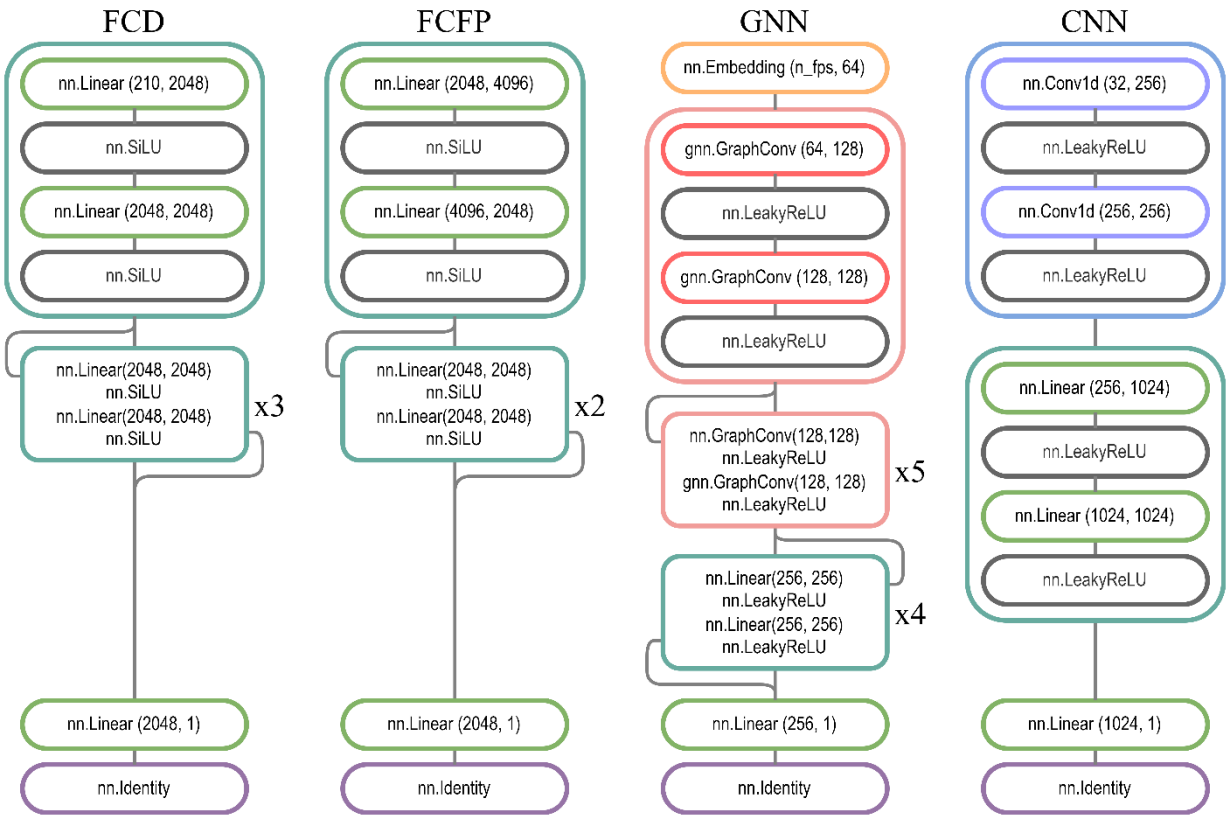


Рисунок 4. Архитектуры нейросетевых моделей для предсказания времен удерживания.

Таблица 1. Гиперпараметры предсказательных моделей для фильтрования METLIN SMRT

Модели → Гиперпараметры ↓	FCD	FCFP	GNN	CNN	Catboost
Шаг обучения	$3*10^{-4}$	$3*10^{-4}$	$3*10^{-4}$	$3*10^{-4}$	$5*10^{-3}$
Оптимизатор	Adam	Adam	Adam	Adam	—
Размера батча (группы)	512	512	128	512	—
Количество эпох/итераций	150	150	150	200	5000
Глубина (CatBoost)/ Число параметров (нейронные сети)	30 млн.	33 млн.	1 млн.	2 млн.	10

FCD: молекула закодирована с помощью 210 дескрипторов из RDKit [90], входной блок содержит входной и скрытый полносвязные слои с 2048 скрытых нейронов. Затем следуют три блока, состоящие из двух последовательных полносвязных слоев с 2048 скрытыми нейронами, в обход каждого из блоков идут промежуточные соединения. Функция активации – SiLU после каждого из

полносвязных слоев. Завершается архитектура полносвязным слоем с единственным выходным нейроном и функцией активации Identity.

FCFP: молекула закодирована при помощи молекулярных отпечатков пальцев RDKit и FCFP4, каждый длиной 1024 значения. Входной блок состоит из двух слоев с 4096 скрытыми нейронами. Далее следуют два блока по два линейных слоя с 2048 скрытыми нейронами, в обход каждого из блоков идут промежуточные соединения. Функция активации также SiLU, и завершается архитектура предсказательной модели аналогично предыдущей.

GNN: каждый из атомов молекулы описан при помощи основных параметров, таких как номер элемента, количество связей, формальная степень окисления и т.д., сама молекула представлена в виде графа. Входной блок состоит из слоя эмбеддинга, который переводит параметры атомов в вектор длиной 64, после чего идут два сверточных графовых слоя с 128 каналами. После этого следуют пять блоков по два сверточных графовых слоя с промежуточными соединениями в обход каждого из блоков. Далее следует функция суммирования всех значений по каждому из каналов, после которой идут четыре блока по два полносвязных слоя с 256 скрытых нейронов. Завершается архитектура выходным слоем с одним выходным нейроном и выходной функцией активации Identity. После всех других слоев нейронной сети используют функцию активации LeakyReLU.

CNN: как и в случае NIST 17 RI CNN1D, молекула была представлена в виде строки SMILES, зашифрованных в one-hot виде. Входной блок состоял из двух сверточных слоев с 256 каналами, за ними следовал блок с двумя полносвязными слоями с 1024 скрытыми нейронами, после чего следовал еще один полносвязный слой с одним выходным нейроном и функцией активации Identity. После каждого из других слоев использовали функцию активации LeakyReLU.

2.1.8 Предсказательные модели для фильтрования синтетических наборов данных

Архитектуру нейронных сетей и оптимальные гиперпараметры предсказательных моделей подбирали с использованием библиотеки Optuna [126]. Структурные мотивы и начальные параметры задавали в соответствии с предыдущими опубликованными работами [18,127,128]. При этом, оптимальные решения

подбирались для двух групп наборов данных по отдельности по следующим параметрам (когда присутствуют в архитектуре): число скрытых слоев, число каналов, размер сверточного фильтра, число скрытых нейронов, размер вектора эмбеддинга, число эпох или итераций и шаг обучения (Табл. 2).

Во всех нейронных сетях использовали функцию активации ReLU, оптимизатор Adam, шаг обучения в $1*10^{-4}$, которые не изменяли под наборы данных. В остальном, архитектуры (Рис. 5) были максимально похожи на использованные для фильтрования METLIN SMRT, так как их точности предсказания было достаточно для решения всех поставленных задач.

Таблица 2. Гиперпараметры нейронных сетей и градиентного бустинга для поиска ошибок в синтетических наборах данных

Набор данных	Модели → Гиперпараметры ↓	CNN	GNN	FCFP	FCD	CB
Дескрипторный	Эмбеддинг Вход -> Выход	36 -> 36	1193 -> 128	—	—	—
	Число сверточных слоев	4	7	—	—	—
	Размер фильтра	5	—	—	—	—
	Число каналов	512	512	—	—	—
	Число полносвязных слоев	8	4	7	2	—
	Число нейронов в полносвязных слоях	1024	1024	512	2048	—
	Батч (группа)	512	128	512	512	—
	Число эпох	100	100	100	100	5000
	Шаг обучения	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	$6x10^{-3}$
	Глубина дерева	—	—	—	—	7
QM9	Эмбеддинг Вход -> Выход	27 -> 27	1980 -> 16	—	—	—
	Число сверточных слоев	3	9	—	—	—
	Размер фильтра	5	—	—	—	—
	Число каналов	512	1024	—	—	—
	Число полносвязных слоев	3	10	7	7	—
	Число нейронов в полносвязных слоях	1024	2048	4096	4096	—
	Батч (группа)	512	128	512	512	—
	Число эпох	200	200	250	250	10 000
	Шаг обучения	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	$1.2x10^{-3}$
	Глубина дерева	—	—	—	—	8

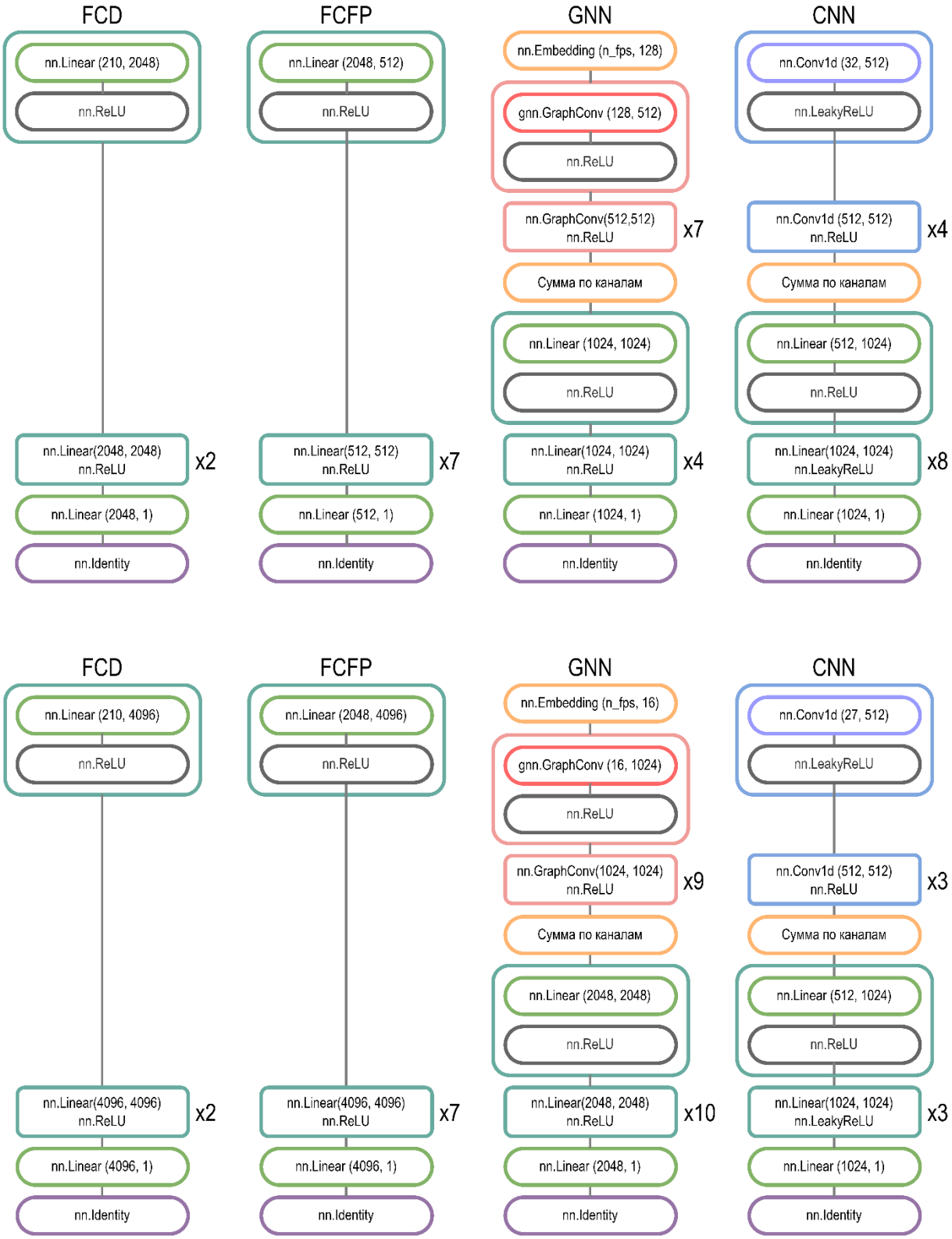


Рисунок 5. Архитектуры предсказательных моделей для синтетических наборов данных – дескрипторного (а) и QM9 (б).

2.1.9 Предсказание масс-спектров электронной ионизации по молекулярной структуре с NEIMS-PyTorch

В работе использовали базы экспериментальных масс-спектров электронной ионизации NIST 23 MS и NIST 17 MS. При проведении сравнения подходов к предсказанию масс-спектров рассматривали органические соединения, содержащие атомы C, H, N, O, P, S, Cl и F, что было обусловлено ограничениями алгоритма RASSP [81]. Тестовый набор содержал 5000 молекул, выбранных случайным образом из 47’044 соединений, присутствующих в основной библиотеке (mainlib) NIST 23 MS, но отсутствующих в NIST 17 MS (Рис. 6). Большая часть молекул обладала массой до 600 Да, с модой около 250 Да.

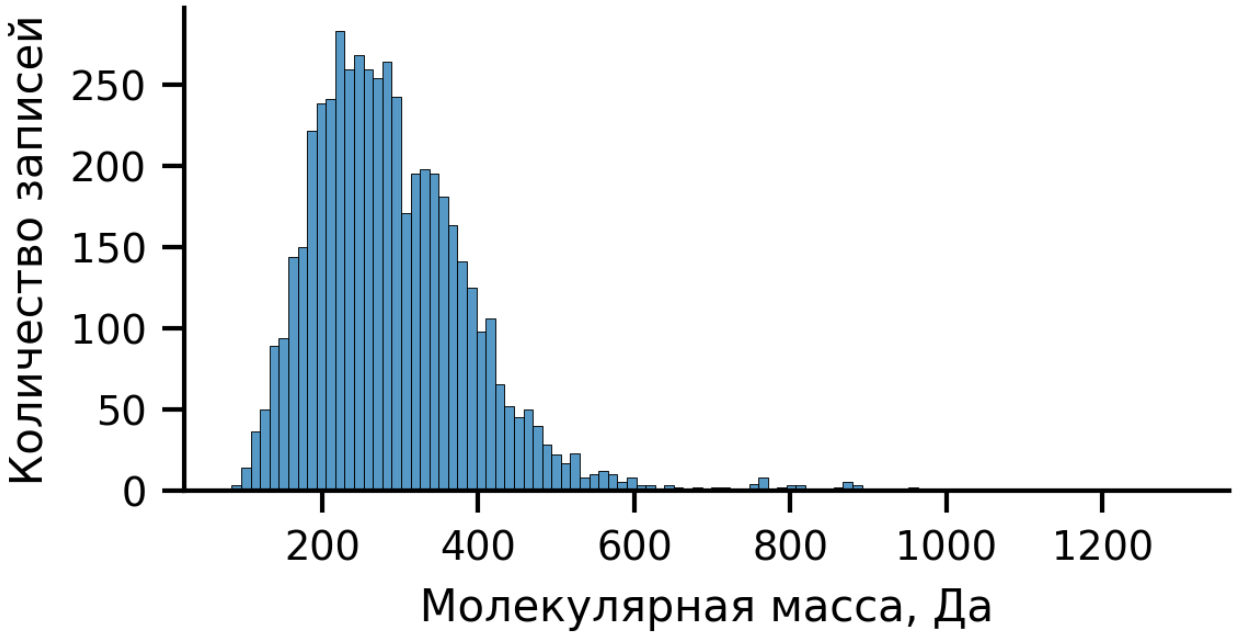


Рисунок 6. Распределение молекулярных масс для веществ из тестового набора.

Для обучения использовали базу экспериментальных масс-спектров NIST 17 MS (mainlib), разделив ее в соотношении 8:2 на тренировочный и валидационный наборы.

Использовали оптимизатор Adam, батч в 512 и шаг обучения $1*10^{-4}$. Оптимальную архитектуру реализации подбирали с использованием библиотеки Optuna [126]. Молекулы описывали при помощи вектора из молекулярных отпечатков пальцев ECFP6 и RDKit, оба длиной в 1024 значения. Как и оригинальная архитектура NEIMS [80], предложенная реализация состояла из двух частей: основной, преобразующей молекулу во внутреннее представление (2 скрытых полносвязных слоя размером 4096) и три предсказательных выхода: (1) для масс-спектров электронной

ионизации, (2) для нейтральных потерь, (3) вероятности вхождения в итоговый масс-спектр значения из (1) или (2). Выходы (1) и (2) содержали 4 полносвязных слоя по 4096 нейронов и по одному выходному с 750 значений и активацией SoftPlus. Вероятностный выход содержал 2 слоя по 4096 нейронов и один выходной в 750 с сигмоидой для получения значений в интервале (0,1) (Рис. 7).

Рисунок 7. Архитектура нейросетевой модели NEIMS-PyTorch для предсказания масс-спектров электронной ионизации. Ветка M предсказывает нормальные масс-спектры, ветка N – нейтральные потери, P – соотношение между M и N (значения от 0 до 1).

2.1.10 Предсказание молекулярных отпечатков пальцев по масс-спектрам электронной ионизации

Подход DeepEI воспроизвели в соответствии с оригинальной архитектурой (Рис. 14) [105]: использовали четыре полносвязных слоя (2000→1000→500→1 нейронов). В качестве финальной функции активации использовали сигмоиду вместо SoftMax, так как именно она чаще используется для задачи двухклассовой (0/1) классификации. Гиперпараметры оставили такими же, как и в оригинальной публикации: шаг обучения $1*10^{-3}$, размер батча 32, обучение в течение 8 эпох. Для каждого из значений ECFP6 и MACCS обучали отдельную нейронную сеть, как и предлагали авторы DeepEI [105].

Полная модель EI2FP:Full, предложенная в работе, состоит из 4 полносвязных слоев в основной части (Рис. 8) и 1 полносвязного слоя для каждого из вариантов молекулярных отпечатков пальцев – ECFP6 и MACCS, соответственно. Таким образом, к ней можно легко добавить дополнительные группы для предсказания других молекулярных отпечатков пальцев. SiLU использовали в качестве функции активации вместо ReLU, так как она позволила добиться лучшей точности предсказания и сделать графики функций потерь более гладкими как для тренировочного, так и для валидационного наборов. Слои случайного исключения (dropout) и нормирования по батчу (BatchNorm) использовали для предотвращения переобучения. Выходной функцией активации была сигмоида.

Облегченная модель EI2FP:Lite состояла из 3 полносвязных слоев в основной части (Рис. 8) и одного слоя в каждой из групп, предсказывающих конкретный вид молекулярных отпечатков пальцев. Остальные решения были аналогичны полной версии модели.

В обоих случаях использовали оптимизатор AdamW с шагом обучения $1*10^{-3}$, размером батча в 512. Шаг обучения подбирали в интервале от $1*10^{-4}$ до $1*10^{-2}$ с шагом множителя в $\sqrt{10}$.

Рисунок 8. Архитектуры нейросетевых моделей для предсказания молекулярных отпечатков пальцев по масс-спектрам электронной ионизации.

2.2 Базы и наборы данных

2.2.1 NIST RI (Retention Index) / NIST GC Database

В базе хроматографических индексов удерживания NIST 23 RI содержится 492 тысячи записей для 162 тысяч уникальных молекул, примерно 153 тысячи с масс-спектрами, из них 479 тысяч линейных и Ковача индексов удерживания, 12 тысяч других индексов удерживания (Ли и другие). Всего 8 строк SMILES с длиной, большей 256 символов и 35 некорректных структур молекул (ошибки при создании с RDKit). Из 478 тысяч корректно обработанных записей 467 содержат редко встречающиеся в

газохроматографических экспериментах типы атомов, например, металлы. Таким образом, после предварительной очистки базы данных от вышеперечисленных аномалий остается 477 тысяч записей.

Представлено несколько больших групп неподвижных фаз, представленных в капиллярных и набивных колонках: 119 тысяч записей – стандартные неполярные (standard nonpolar, Рис. 9, значения от 400 до 3500 е.и.у., мода около 1000 е.и.у.), 269 тысяч – полустандартные неполярные (semi-standard nonpolar, Рис. 10, значения от 400 до 4000, наиболее часто встречаются 1000-1500 е.и.у), 90 тысяч – полярные неподвижные фазы (standard polar & semi-standard polar, Рис. 11, значения от 500 до 3500, наиболее часто встречаются 1200-1700 е.и.у), 39 записей – другие типы колонок. Всего представлено 231 тысяча комбинаций молекула-конкретное коммерческое наименование неподвижной фазы (например, Agilent DB-5 или Varian VF-5, схожие по своим параметрам, но для этого подсчета являющиеся разными), примерно 80% веществ имеют не более 1 записи на конкретный тип неподвижной фазы.

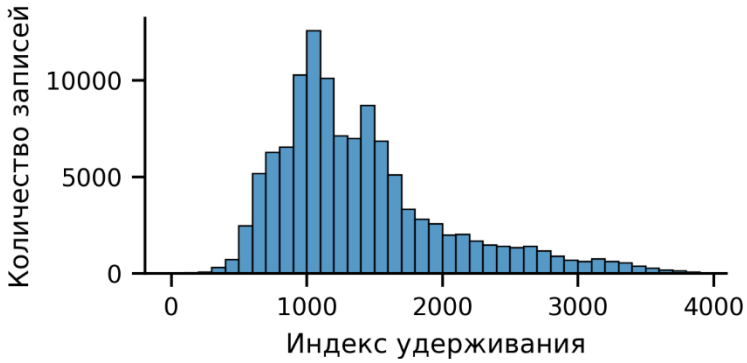


Рисунок 9. Распределение индексов удерживания для стандартных неполярных неподвижных фаз.

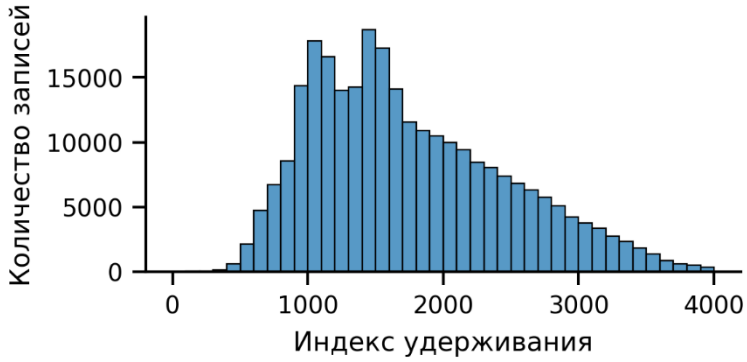


Рисунок 10. Распределение индексов удерживания для полустандартных неполярных неподвижных фаз.

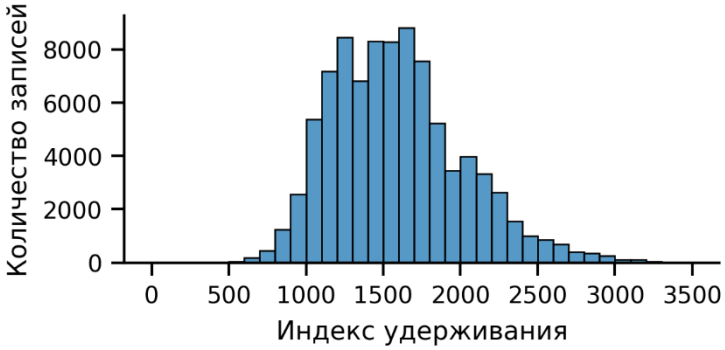


Рисунок 11. Распределение индексов удерживания для полярных неподвижных фаз.

Наиболее широко представлены различные виды полустандартных неполярных неподвижных фаз с 5%-фенилполидиметилсилоксаном (без коммерческого наименования – 43 тысячи, Agilent DB-5 – 35 тысяч, Restek Rxi-5Sil – 33 тысячи записей). Из стандартных неполярных фаз на основе полидиметилсилоксана без добавок наиболее часто встречаются Agilent DB-1 – 22 тысячи, Supelco SE-30 – 21 тысяча, Supelco OV-101 – 17 тысяч записей. Среди стандартных полярных фаз наиболее распространенная на основе полиэтиленгликоля – Agilent DB-Wax с 23 тысячами, GS-Tek Carbowax 20M с 13 тысячами записей.

Индексы удерживания в NIST RI получены из разных источников: литературные данные, работы сотрудников NIST MS DataCenter, а также других научных групп, в том числе и из России.

Для сравнения, в NIST 20 RI представлено примерно 447 тысяч записей для 140 тысяч уникальных веществ, а в NIST 17 RI – 404 тысячи индексов удерживания для 99 тысяч соединений [129].

2.2.2 METLIN SMRT (Small Molecules Retention Times)

METLIN SMRT содержит времена удерживания для 80 тысяч молекул, их идентификаторы в PubChem, структуры и молекулярные отпечатки пальцев, рассчитанные при помощи программы Dragon 7 [33]. Все времена удерживания (Рис. 12, значения от 500 до 1500 секунд, до 250с представлены неудерживаемые соединения) в этом наборе получены на ВЭЖХ хроматографе с квадрупольно-времяпролетным тандемным масс-спектрометром в качестве детектора (HPLC-Q-TOF) на колонке для обращенно-фазовой ВЭЖХ Zorbax Extend C-18 (2.1 x 50мм, 1.8 мкм) в одних и тех же хроматографических условиях для всех веществ.

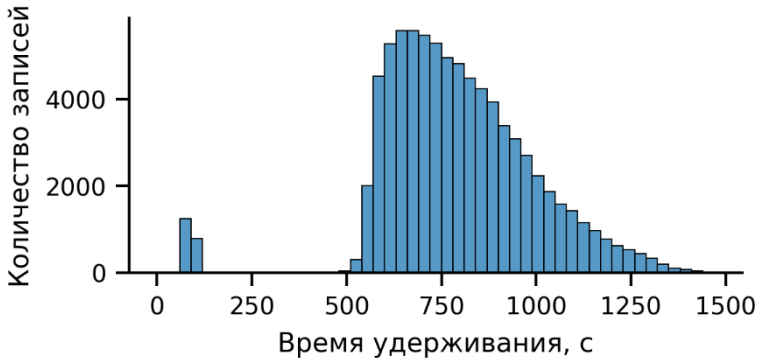


Рисунок 12. Распределение времен удерживания в наборе данных METLIN SMRT.

2.2.3 NIST MS (Mass Spectral library)

База масс-спектров электронной ионизации NIST 23 MS содержит 394 тысячи записей для 347 тысяч уникальных соединений. Наиболее качественные масс-спектры отбираются в основную библиотеку (mainlib, 341 тысяча записей, распределение молекулярных масс представлено на Рис. 13), дублирующие масс-спектры, полученные в других условиях, помещаются в библиотеку replib.

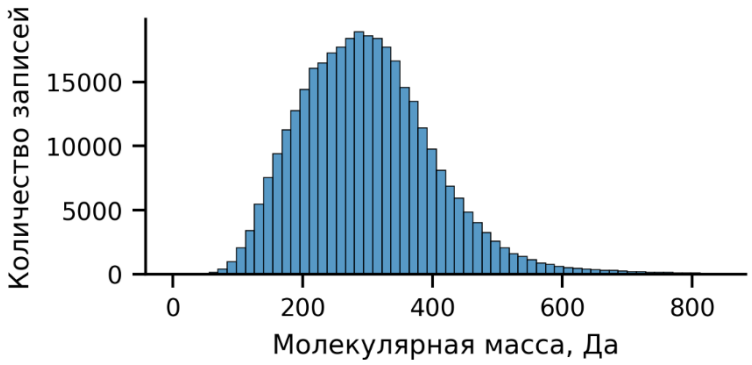


Рисунок 13. Распределение молекулярных масс в наборе данных NIST 23 MS (mainlib).

Масс-спектральная база данных NIST 17 MS содержит 306 тысяч записей для 267 тысяч уникальных молекул, в mainlib содержится 261 тысяча записей (распределение молекулярных масс представлено на Рис. 14).

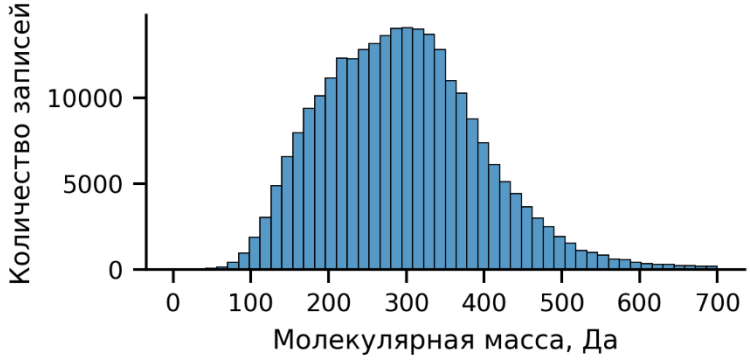


Рисунок 14. Распределение молекулярных масс в наборе данных NIST 17 MS (mainlib).

Из распределений молекулярных масс для NIST 17 MS (Рис. 14) и NIST 23 MS (Рис. 13) видно, что в базе данных, в основном, представлены масс-спектры для сравнительно небольших молекул с массой до 600 Да, и только в редких случаях встречаются большие значения. При этом максимум распределения приходится на область около 300 Да. Наиболее легкие органические молекулы, представленные в наборе данных, имеют молекулярную массу около 25-30 Да, однако их число крайне мало.

2.2.4 Наборы данных для предсказания молекулярных отпечатков пальцев

Для обучения предсказательных моделей использовали базу данных экспериментальных масс-спектров NIST 17 MS (mainlib). Ее разбивали в соотношении 9:1 для тренировочно-валидационного и тестового наборов. Затем тренировочный и валидационный наборы получали разбиением в соотношении 85:15. Валидационный набор использовали для оптимизации архитектуры и гиперпараметров моделей, а тестовый – только однажды, для сравнения точности предсказания между предложенными подходами и подходом DeepEI [105].

Молекулы описывали при помощи молекулярных отпечатков пальцев ECFP6 с радиусом 3 и длиной 1024, рассчитанных при помощи библиотеки RDKit [90]. Эти вектора были рассчитаны для всех молекул в наборе данных, после чего для сравнения с DeepEI были выбраны только те биты, которые были положительными для 10..90% молекул.

Аналогичную процедуру повторили для MACCS запросов, также доступных в RDKit. Для сравнения были выбраны 95 MACCS запросов.

2.2.5 Синтетические наборы данных для оценки эффективности фильтрования на основе дескрипторов и QM9

Два модельных сгенерированных набора данных были использованы: на основе автокорреляционных дескрипторов и на основе квантовомеханического набора QM9. В модифицированных наборах изменили заданную часть записей одним из двух способов: случайным перемешиванием записей и добавлением фиксированной ошибки.

Для дескрипторных наборов данных в качестве основы выбрали уникальные молекулярные структуры из METLIN SMRT (79 890 записей), для них сгенерировали автокорреляционные дескрипторы при помощи алгоритмов пакета Mordred [125] (поддерживаемый форк, пакет mordred-community [130]). Выбрали 428 дескрипторов, которые были корректно предсказаны для всех молекул, после чего применили к каждому из них автошкалирование – привели к нулевому среднему значению со стандартным отклонением 1. После этого взяли взвешенное среднее для всех выбранных дескрипторов со случайными равномерно распределенными коэффициентами из [0,1). Получившиеся значения были нормально распределенными (Рис. 15), что ожидается согласно центральной предельной теореме, со средним значением 0 и стандартным отклонением в 0.225. Таким образом, SMILES структуры молекул в сочетании с расчетными значениями назвали базовым дескрипторным набором данных.

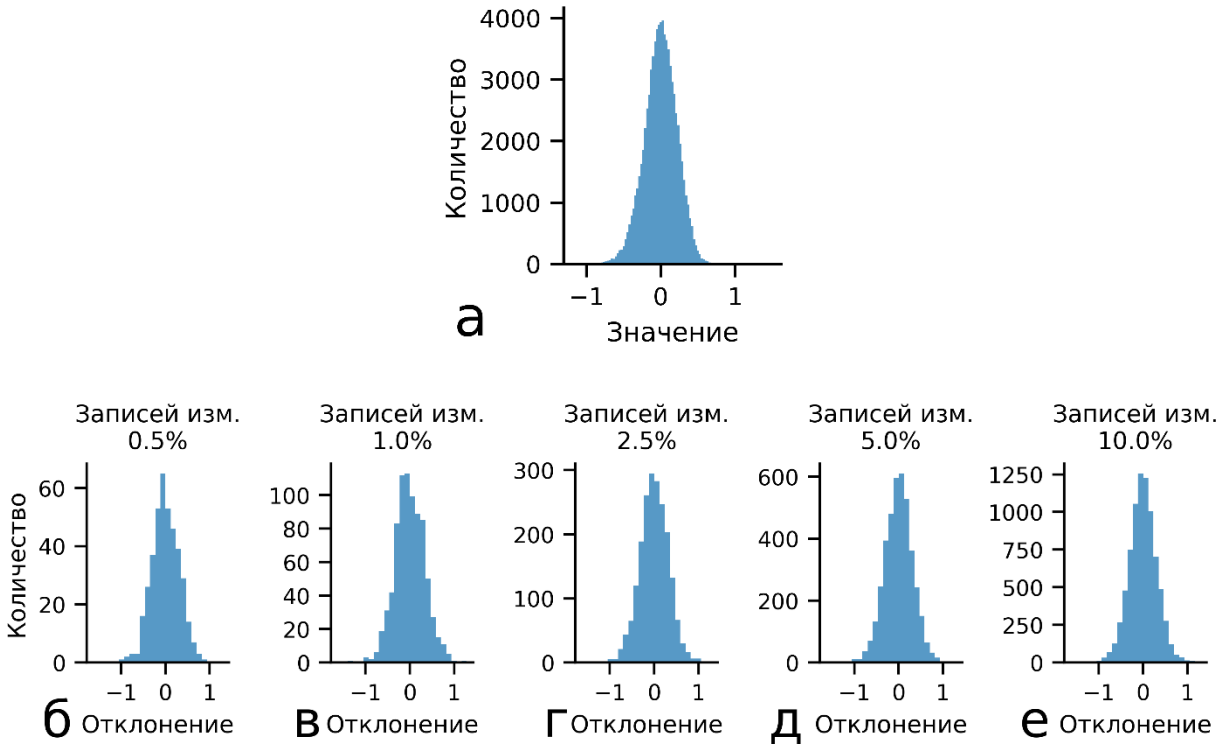


Рисунок 15. Распределения модельных значений в исходном (а) и отклонений от модельных значений: б – 0.5%, в – 1.0%, г – 2.5%, д – 5%, е – 10% в перемешанных дескрипторных наборах данных.

Для моделирования экспериментальных погрешностей к данным был добавлено небольшой нормально распределенный шум (0 среднее значение, 0.015 стандартное отклонение), полученный набор данных назвали зашумленным дескрипторным набором данных. Все модифицированные дескрипторные наборы данных основаны

именно на нем. Так, перемешанные дескрипторные наборы данных были получены перемешиванием значений свойства заданной доли (0.5, 1.0, 2.5, 5.0, 10.0%) случайно выбранных записей (Рис. 15). Это изменение имитирует часто встречающуюся экспериментальную ошибку неверной идентификации веществ. Другую группу наборов данных с добавленной ошибкой получили добавлением фиксированной ошибки (0.025, 0.04, 0.05, 0.075, 0.1, 0.15, 0.25) к заданной доле (0.5, 1.0, 2.5, 5.0, 10.0%) случайно выбранных записей. Таким образом, были сформированы 1 базовый, 1 зашумленный, 5 перемешанных, 35 с добавленной ошибкой дескрипторных наборов данных.

В качестве основы для наборов данных QM9 выбрали уникальные молекулы из квантовомеханического датасета QM9 [131,132] (132 398 записей). Энергия Гиббса атомизации при 298K (“g298_atomic”), разделенная на -1000, в сочетании со SMILES молекул сформировали базовый набор данных QM9. Полученные значения имели колоколообразное распределение со средним значением 1.63 (Рис. 16), стандартным отклонением в 0.220. На основе базового были сгенерированы перемешанные наборы данных QM9, в которых перемешали свойства заданной доли случайно выбранных записей (0.5, 1.0, 2.5, 5.0, 10.0%, Рис. 16)), полностью аналогично перемешанным дескрипторным наборам данных. Аналогично дескрипторным наборам, сгенерировали наборы QM9 с добавленной ошибкой (0.0075, 0.0125, 0.015, 0.0175, 0.02, 0.025, 0.05, 0.08, 0.15, 0.25) к заданной доле записей (0.5, 1.0, 2.5, 5.0, 10.0%). Таким образом, были сформированы 1 базовый, 5 перемешанных, 50 с добавленной ошибкой наборов данных QM9.

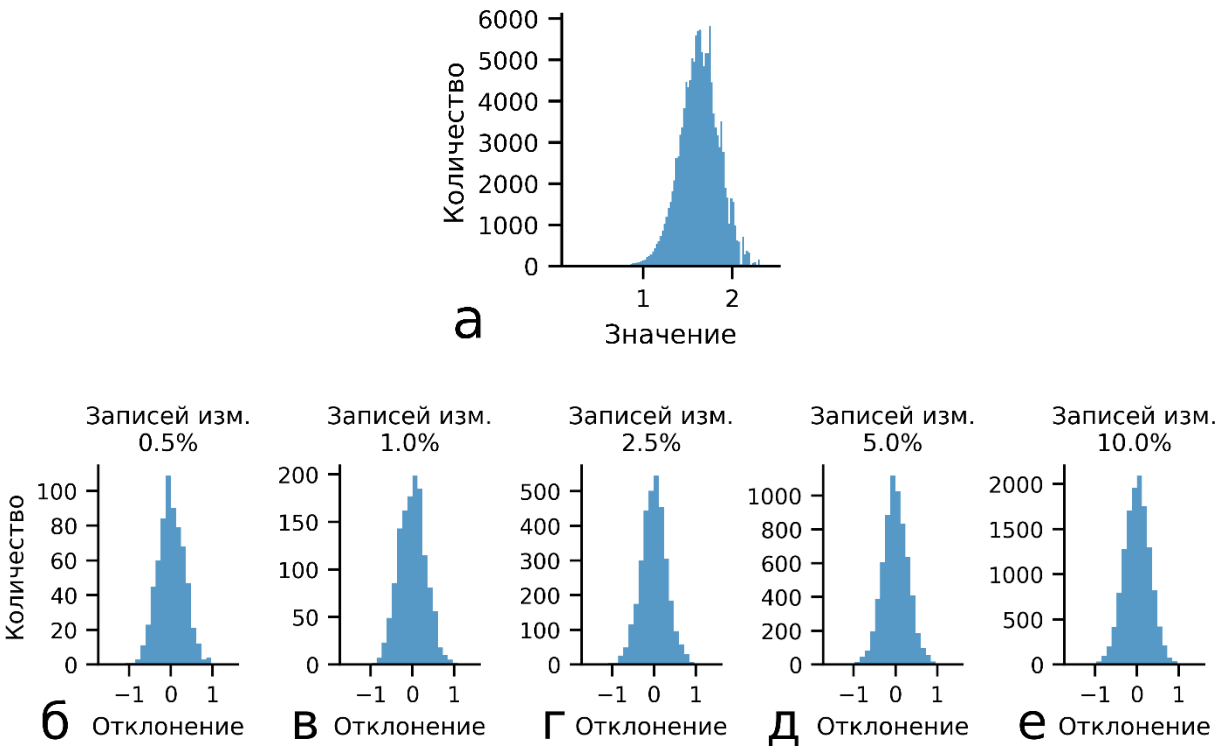


Рисунок 16. Распределения модельных значений в исходном (а) и отклонений от модельных значений: б – 0.5%, в – 1.0%, г – 2.5%, д – 5%, е – 10% в перемешанных QM9 наборах данных.

Для поиска ошибок в синтетических наборах данных использовали пять независимых моделей: сверточную нейронную сеть (CNN), сверточную графовую нейронную сеть (GNN), полносвязную нейронную сеть с дескрипторами для описания молекулы (FCD), полносвязную нейронную сеть с молекулярными отпечатками пальцев для представления молекулы (FCFP), а также CatBoost реализацию алгоритма градиентного бустинга.

2.3 Оборудование

Для всех вычислений и обучения моделей использовали несколько персональных компьютеров со следующими характеристиками:

1. CPU: Intel Xeon E5-2667 v2, RAM: 64 Gb DDR3 @ 1600 MT/s, GPU: Nvidia RTX 3060 12 Gb;
2. CPU: AMD Ryzen 7 7700, RAM: 64 Gb DDR5 @ 6400 MT/s, GPU: Nvidia RTX 4080 Super 16 Gb;

3. CPU: AMD Ryzen 7 5800X, RAM: 64 Gb DDR4 @ 3200 MT/s, GPU: Nvidia RTX 5070 Ti 16 Gb.

2.4 Программное обеспечение и библиотеки

Для расчетов использовали операционные системы Windows 10, Windows 11 24H2 и Linux Mint 22.1. Основной язык программирования – Python 3.9 – 3.12 со следующими библиотеками: NumPy [133], pandas [134], seaborn [135], RDKit [90], Matplotlib [136], PyTorch [137], CatBoost [123], PyTorch Geometric [138], Optuna [126], mordred-community [125,130], и scikit-learn [139]. Также использовались несколько вспомогательных библиотек: tqdm, optuna-dashboard, torchinfo, tensorboard, ipykernel.

2.5 Подход с использованием «желтых карточек»

Подход к фильтрованию с использованием «желтых карточек» [127,128] состоит из двух основных этапов: (i) обучение моделей, (ii) фильтрование. На первом шаге исходный набор данных разбивают на k частей (по-умолчанию, $k = 5$), как при k-fold кросс-валидации. Каждую из моделей обучают на $k-1$ частей набора, оставшуюся часть используют для получения предсказанных значений. Процесс повторяют k раз для каждой из N моделей, получая таким образом N предсказанных значений для каждой из исходных записей. На втором шаге выполняют фильтрование: выбирают метрику сравнения – абсолютную ошибку, относительную ошибку или обе одновременно, в зависимости от природы набора данных и ошибки предсказания. Также выбирают пороговое значение *threshold_% (t%).* Для каждой из моделей выбирают *threshold_% (t%)* наименее точно предсказанных записей и отмечают их «желтой карточкой». Так, каждая из записей может получить от 0 до N «желтых карточек». Записи в группе с 0 «желтых карточек» предсказываются точно, попавшие же в группу с N «желтых карточек», в основном, являются ошибочными, но туда также могут попадать небольшое количество уникальных молекул, для которых все модели предсказывают значения с большой ошибкой. Записи в группах с 1 по $N-1$ включительно представляют смесь нормальных и ошибочных записей, причем доля последних растет с увеличением номера группы.

2.6 Округление значений *m/z* до целочисленных

В качестве основного языка программирования использовали Python. Numpy и Scipy использовали для операций над массивами, а также для линейной алгебры, Pandas для таблиц, импорта и экспорта, Xarray для чтения файлов в формате NetCDF, Matplotlib для построения графиков. Исходный код, NetCDF и CSV файлы вместе с дополнительной информацией доступны в репозитории GitHub [140].

Набор NetCDF файлов использовали для реконструкции алгоритмов округления из программ AMDIS 2.73 (NIST), ChemStation E.01.00.237 (Agilent), ChromaTOF 4.72.0.0 (LECO), MS Search 2.3 (NIST) и OpenChrom 1.3 (Lablicate). Каждый из файлов открывали вышеуказанными программами, после чего сохраняли в виде CSV таблицы. Для MS Search использовали скрипт, автоматизирующий процесс обработки с опциями по-умолчанию. Интервалы для округления значений *m/z* до целочисленных находили минимизацией Евклидового расстояния между CSV таблицей и результатами обработки NetCDF файлов при помощи реализации алгоритмов на Python. Единственный алгоритм, который потребовал ручной доработки – из программы ChemStation, потому что он включает условный переход.

Для оценки распределения точных значений *m/z* (далее просто распределение) в каждом из интервалов были сгенерированы два модельных набора данных. Один состоял из 1 486 731 молекулярных формул (с молекулярной массой меньше 600), которые извлекли из общедоступной базы данных PubChem. В нем содержались только молекулярные формулы известных веществ и их смесей, все фрагменты исключили. Этот набор данных позволил получить оценку для нижней границы ширины распределения для каждого из интервалов. Второй набор данных был создан с использованием NIST 17. Он содержал 14 324 954 комбинаторно возможных фрагментов, включая исходные молекулярные формулы (с молекулярной массой меньше 600). Многие из этих фрагментов не существуют, поэтому этот набор позволил получить верхнюю оценку ширине распределения для каждого из интервалов. Во всех случаях учитывали только уникальные формулы, без повторов, и только ионы с единичным зарядом. Для всех формул рассчитали теоретические изотопные кластеры.

Существует несколько стратегий выбора одного из изотопологов для рассмотрения. С одной стороны, можно учитывать только наиболее интенсивный ион из каждого кластера, с другой – можно учесть все ионы с интенсивностью, большей

заданного порога (1% или 10% от максимума в кластере). Мы считаем, что первый подход позволяет получить более взвешенную оценку, так как каждая формула учитывается только один раз. Например, кластеры полихлорированных и полибромированных ионов содержат несколько интенсивных пиков в кластере, что увеличивает их вклад в конечный результат при использовании правила с граничным значением. Кроме того, у хлора и у брома отрицательные дефекты массы, что сдвинет ожидаемые распределения влево. Поэтому мы предпочли использовать только один наиболее интенсивный ион из каждого кластера.

Глава 3. Обработка хроматографических данных[1]

3.1 Подход к фильтрованию с использованием «желтых карточек»

В последнее время очистке данных уделяется все больше внимания, в том числе с использованием методов машинного и глубокого обучения. Однако не так много усилий сконцентрировано на базах экспериментально полученных химических данных. Задача дополнительно осложняется большой долей соединений, охарактеризованных всего одной записью. К этой категории в базах индексов удерживания NIST RI и времен удерживания METLIN SMRT относятся более 80% веществ. Это исключает возможность эффективного применения статистических подходов, основанных на анализе распределения экспериментальных значений, полученных для одного и того же соединения. Ручная проверка также крайне затруднительна из-за большого объема баз данных (сотни тысяч записей) и затратности экспериментальной перепроверки.

Для решения этой проблемы был предложен подход к обнаружению ошибочных записей в базах хроматографических данных, основанный на объединении предсказаний нескольких моделей путем голосования. Каждая из *N* моделей обучается на *k*-fold разбиении данных, где *k-1* частей используют для обучения, одну - для предсказания. Эту процедуру проводят для каждой модели, чтобы получить *N* предсказанных копий исходного набора данных. В каждой копии выбирают определенную долю записей (например, 5%), предсказанных хуже всего, и отмечают их «желтой карточкой». Записи, получившие «желтые карточки» от каждой из моделей, потенциально являются ошибочными. Стоит отметить также возможность дальнейшего развития и расширения подхода – в качестве дополнительных «желтых карточек» могут выступать расчетные характеристики, основанные на предсказаниях моделей, например, стандартное отклонение всех предсказаний или другие

---

1 При подготовке данной и последующих глав диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования:
***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. V. 1297. P. 342375. EDN: OPAIVM
***Khrisanfov M.**, Matyushin D., Samokhin A.* Finding potentially erroneous entries in METLIN SMRT // *Journal of Chromatography A*. 2025. P. 465761. EDN: IGUXUB

статистические параметры. Благодаря этому можно дополнительно гибко дополнять базовую часть алгоритма для увеличения точности поиска ошибок.

Подход применили к двум базам данных: экспериментальных времен (METLIN Small Molecule Retention Times) и индексов удерживания (NIST Retention Index). Для оценки эффективности предложенного подхода к фильтрованию сгенерировали две группы тестовых наборов данных – на основе автокорреляционных дескрипторов и квантовохимического набора QM9, в которые контролируемым образом добавили ошибки. Во всех рассмотренных случаях предложенный подход превзошел более простые способы фильтрования, например, основанные на использовании граничного значения абсолютной ошибки предсказания.

В ходе разработки подхода были сформулированы и подтверждены экспериментально 5 характеристик предлагаемого подхода [141]:

1. Предсказательные модели могут игнорировать небольшое количество ошибок в тренировочных наборах данных. Это поведение объясняется способностью к генерализации (обобщению) – выявлению наиболее общих закономерностей и отсеиванию редко встречающихся паттернов.
2. Ошибки предсказания моделей скоррелированы не полностью, что объясняется разными архитектурами нейронных сетей и типами используемых входных данных. Поэтому использование нескольких моделей улучшает эффективность фильтрования по сравнению даже с самой точной из индивидуальных моделей или ансамбля из моделей с одной архитектурой.
3. Последняя группа с максимальным количеством «желтых карточек» содержит, в основном, ошибочные записи и небольшое количество неточно предсказанных значений. В то же время группа 0 без «желтых карточек» содержит преимущественно корректно предсказанные записи. Промежуточные группы отвечают в основном корректно предсказанными записями и небольшим количеством ошибок. В следствие этого наблюдаются следующие закономерности: U-образное распределение записей и перевернутое U-образное распределение стандартного отклонения предсказаний моделей в зависимости от номера группы.
4. Первую производную кривой зависимости количества записей в группе с максимальным количеством «желтых карточек» от порогового параметра *t%*

можно использовать для оценки точности поиска ошибок на неразмеченных наборах данных и выбора оптимального значения *t%*.

5. Подход к фильтрованию с использованием «желтых карточек» работает лучше, чем более простые подходы, такие как отсечение по квантилю или по абсолютной ошибке.

3.1.1 Влияние ошибок в тренировочном наборе на точность предсказания

Известно, что предсказательные модели обладают свойством игнорировать небольшое количество ошибок в тренировочном наборе данных [142]. Этот принцип является общеизвестным в области машинного обучения, однако активно не обсуждается в статья, посвященных обработке химических данных.Так, модели, обучавшиеся на наборах данных, содержащих до 10% ошибочных записей, показывают одинаковые значения ошибок, рассчитанные раздельно как для исходных, так и для модифицированных записей (Рис. 17). Несмотря на то, что ошибки предсказания линейно возрастают вместе с долей модифицированных записей, угол наклона кривых небольшой. При этом, значения ошибок предсказания, рассчитанные относительно зашумленного перемешанного дескрипторного набора данных, лежат выше соответствующих точек, рассчитанных относительно базового набора без добавленного нормально распределенного шума. Это значит, что модели способны игнорировать не только значительные ошибки, но и небольшие погрешности, имитирующие экспериментальные.

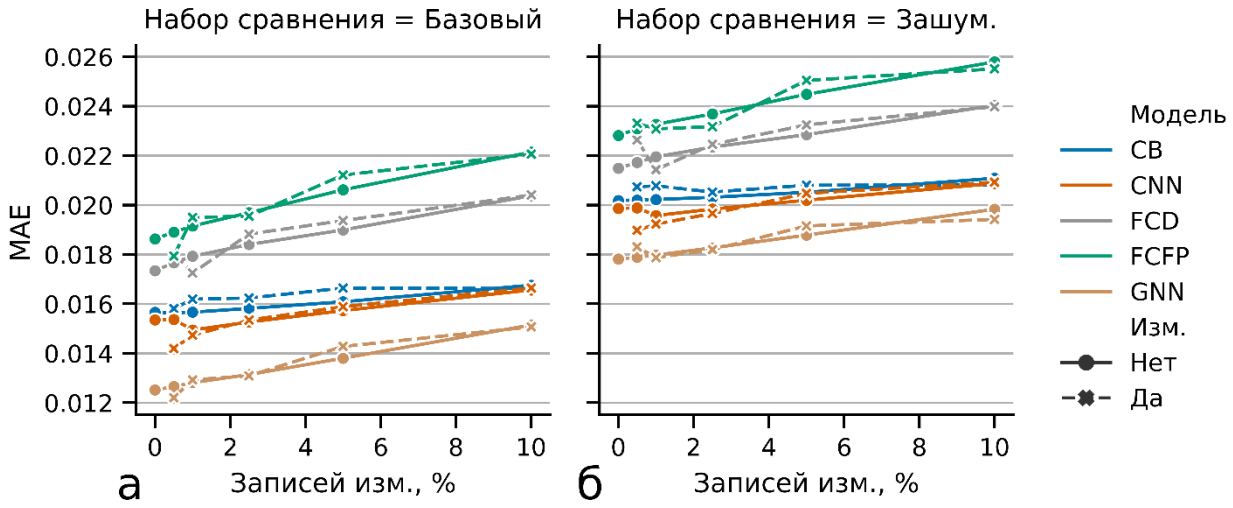


Рисунок 17. Средняя абсолютная ошибка предсказания (MAE) для 5 моделей на перемешанных дескрипторных наборах данных с 0.0-10.0% случайно перемешанных записей при расчете относительно: а – базового, б – зашумленного наборов. Сплошными линиями показаны изменения

значения, рассчитанные для исходных (немодифицированных) записей, а для ошибочных (измененных) – прерывистыми.

3.1.2 Корреляции между ошибками предсказания моделей

Все предсказательные модели, используемые в данной работе, демонстрируют умеренно высокую точность предсказания, поэтому ожидаемо, что значения, предсказанные любой парой моделей, будут в значительной степени скоррелированы. Вне зависимости от рассматриваемого набора данных предсказания моделей показали коэффициент корреляции не ниже 0.99. Однако в подходе к фильтрованию ошибочных записей большее значение имеет не прямая корреляции между предсказаниями, а корреляции между ошибками предсказания – различиями между предсказанной величиной и референсным значением (рассчитанным или экспериментальным).

При использовании в качестве референсных значений тех же величин, которые использовались для обучения (и включали как исходные, так и модифицированные записи) наблюдается устойчивая тенденция к росту коэффициентов корреляции между ошибками предсказания с увеличением доли модифицированных записей (Рис. 18). Абсолютные значения изменялись от умеренного 0.63 до крайне высокого 0.97 (Рис. 18), с некоторыми вариациями в зависимости от рассматриваемой пары моделей.

Пары FCFP-FCD и FCFP-CNN показывали более низкие коэффициенты корреляции чем другие сочетания для дескрипторного набора данных. Для QM9 пара FCFP-GNN характеризовалась наименьшими коэффициентами корреляции. С увеличением количества модифицированных записей и увеличением величины добавленной ошибки уменьшался разброс коэффициентов корреляции между разными парами моделей. Для экспериментальных наборов данных (NIST 17 RI, METLIN SMRT) коэффициенты корреляции между ошибками предсказания лежали в диапазоне 0.6...0.8.

Однако наблюдалась абсолютно другая ситуация, когда в качестве референсных значений использовали исходные (немодифицированные) значения. В этом случае коэффициенты корреляции между ошибками предсказания были существенно меньше и находились в диапазонах 0.4…0.7 и 0.2...0.6 для зашумленного или чистого наборов данных соответственно, а разброс между коэффициентами корреляции, рассчитанными для одной пары моделей на наборах данных с разной долей модифицированных записей, стал пренебрежимо мал (Рис. 18).

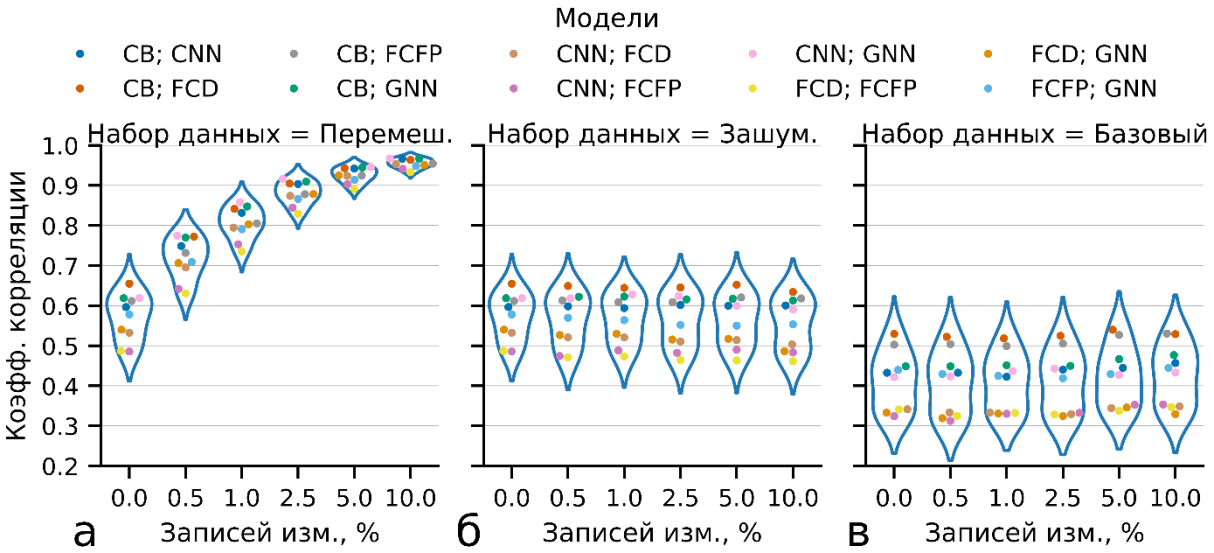


Рисунок 18. Коэффициенты корреляции между ошибками предсказания разных моделей на перемешанных дескрипторных наборах данных при расчете ошибок предсказания относительно разных референсных наборов: а – перемешанного, б – зашумленного, в – базового.

Наблюдаемые значения коэффициентов корреляции также зависят от рассматриваемого набора данных. Так, для QM9 наборов данных значения были в среднем ниже, чем для дескрипторных. Таким образом, несмотря на то что коэффициенты корреляции, рассчитанные с использованием наборов данных с ошибками в качестве референсных, могут быть высокими (0.6…0.8), при расчете с исходными референсными данными реальные значения оказываются ниже (0.2…0.6). Низкие реальные коэффициенты корреляции между ошибками предсказания указывают на перспективность объединения разнородных моделей в ансамбли.

3.1.3 Использование одной архитектуры для всех моделей в ансамбле

Умеренные коэффициенты корреляции между ошибками предсказаний в группах из 10 GNN (около 0.63 и 0.75) и 10 FCFP (около 0.725 и 0.770) моделей для чистого и зашумленного наборов, соответственно, свидетельствуют о том, что модели одного типа могут также быть объединены в ансамбль. Такой подход имеет значительные преимущества: значительно упрощается подготовка входных данных для моделей, устраняется необходимость подбора оптимальных гиперпараметров для нескольких архитектур. Однако наиболее очевидным недостатком является увеличенные значения коэффициентов корреляции между ошибками предсказания, что частично может быть компенсировано использованием большего количества моделей.

GNN и FCFP были выбраны для оценки эффективности фильтрования, так как первая архитектура показывает наиболее высокую точность предсказания, а вторая – наибольшую производительность и использует в качестве входных данных простые для расчета молекулярные отпечатки пальцев, которые доступны в RDKit.

Различие между моделями достигали за счет различного разбиения исходного набора данных на тренировочную и валидационную части в процессе 5-fold кросс валидации. Комбинации моделей для сравнения выбирали следующим образом: для каждого набора (1, 5, 10 моделей) выбирали комбинацию с наибольшей F_1 метрикой, рассчитанной для всех вариантов порогового значения *t%*. В качестве базового значения при сравнении использовали оригинальный ансамбль из пяти предсказательных моделей.

Сравнение эффективности фильтрования подходов «в целом» можно осуществить при помощи кривых precision-recall (Рис. 19а) и площади под ними (Рис. 19б), аналогично кривой ROC и площадью под ней (AUC). Кривые рассчитывали для пороговых значений *t%* от 0.01% до 100%. Более эффективные подходы к фильтрованию характеризуются кривыми, которые лежат выше и правее, и соответственно характеризуются большей площадью.

Наиболее эффективной одиночной моделью была GNN (Рис. 19), что ожидалось исходя из ее наибольшей точности предсказания. Лучшая из одиночных FCFP моделей показала значительно более скромные результаты. Даже объединение 5 и 10 FCFP моделей уступили лучшей из одиночных GNN моделей. При этом 10 GNN моделей показали эффективность фильтрования, крайне близкую к оригинальному ансамблю из 5 разнородных моделей.

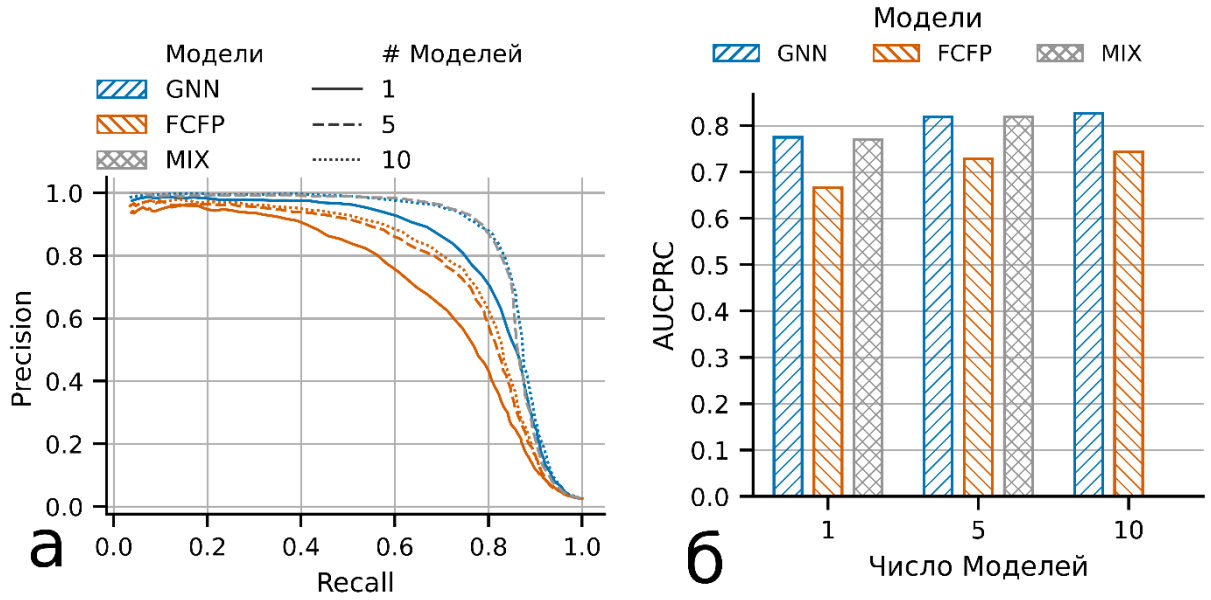


Рисунок 19. (а) кривые precision-recall (точность-полнота) для 1,5,10 GNN, 1,5,10 FCFP и изначального ансамбля из 5 моделей; (б) площадь под кривыми precision-recall (точность-полнота) для перечисленных ансамблей.

Таким образом, возможно сделать вывод, что использование нескольких моделей с одинаковой архитектурой для фильтрования является жизнеспособным подходом. В то же время использование нескольких разнородных моделей позволяет получать более устойчивые предсказания, так как неправильный выбор единичной архитектуры может значительно ухудшить точность поиска ошибок.

3.1.4 Распределение количества записей по группам

В наборах без добавленных ошибок нулевая группа (без «желтых карточек») на порядок превосходит по количеству записей все остальные, далее с увеличением номера группы количество записей в них монотонно убывает. Этот тренд характерен и для немодифицированных наборов дескрипторных и QM9 данных. Однако с появлением ошибочных записей форма распределения изменяется – она становится U-образной с увеличенным количеством записей в последней группе (Рис. 20). Такая форма наблюдалась как в случае синтетических, так экспериментальных наборов данных.

Анализ синтетических наборов данных показывает, что U-образная форма кривой сохраняется в широком диапазоне отношений порогового значения *t%* и действительного количества модифицированных записей в наборах данных. Появление такой формы распределения напрямую связано с размером последней группы. Более

ярко выраженная U-образная кривая обычно свидетельствует о более высокой точности фильтрования.

Использование синтетических наборов данных позволило не только определить точность и полноту фильтрования, но и отдельно рассмотреть распределения отдельно для модифицированных и немодифицированных записей. Во всех наборах данных количество немодифицированных записей убывало с увеличением номера группы, так что характерный изгиб объяснялся увеличением модифицированных записей в последней группе.

Эти результаты также подтверждают, что нулевая группа состоит, в основном, из немодифицированных записей (Рис. 20), а последняя – из модифицированных с небольшой долей немодифицированных. Состав групп определяется пороговым значением *t%*, долей модифицированных записей в наборе, типом внесенной ошибки.

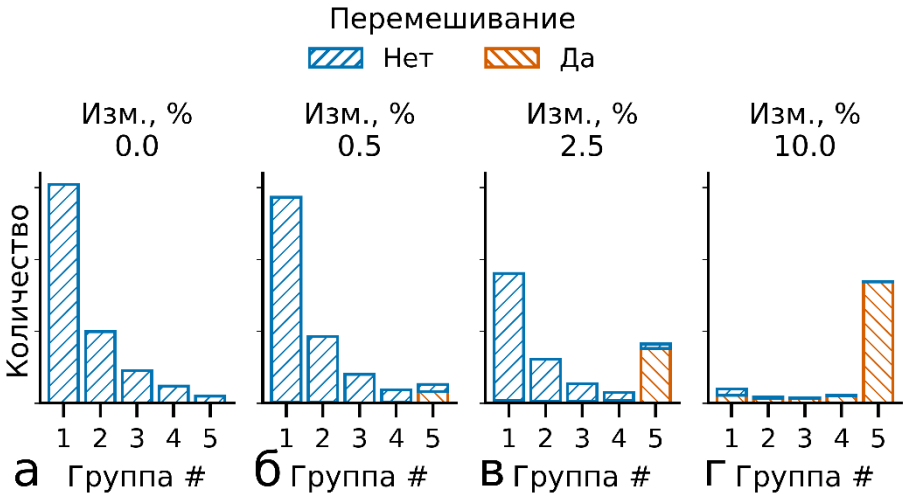


Рисунок 20. Распределение записей по группам для дескрипторных наборов данных с разной долей перемешанных записей (Изм. %): а – 0%, б – 0.5%, в – 1.0%, г – 2.5%, д – 10%. Синим отмечено количество исходных записей, оранжевым – перемешанных.

3.1.5 Зависимость стандартного отклонения предсказаний от номера группы

Стандартные отклонения предсказаний моделей возрастают с увеличением номера группы в наборах данных без ошибок, модели предсказывают некоторые записи менее точно и менее согласованно. Чем более уникальное соединение – тем больше ожидаемый разброс между предсказаниями. В то же время при наличии достаточного числа ошибочных записей наблюдается аномально низкое значение стандартного отклонения предсказаний для группы, отвечающей максимальному количеству «желтых карточек», обусловленное тем, что ошибочные записи предсказываются

моделями согласованно. Так как ошибочные записи преобладают в последней группе, то именно они определяют среднее значение стандартного отклонения. Таким образом в этих случаях наблюдается зависимость с максимумом (перевернутое U-образное распределение, Рис. 21).

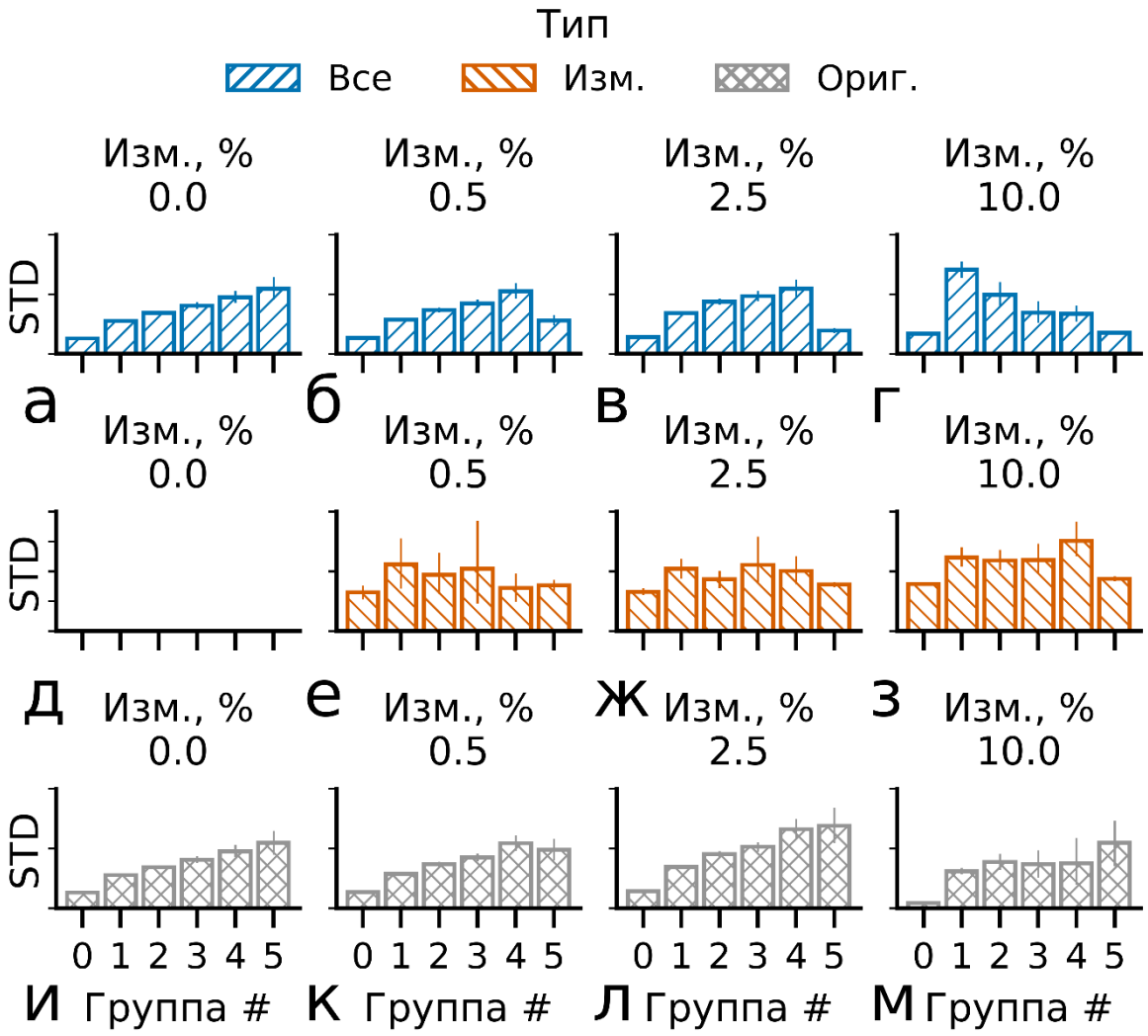


Рисунок 21. Стандартные отклонения для модифицированных (оранжевый, д – з), немодифицированных (серый, и – м) и всех записей (голубой, а – г) по группам 0...5 в наборах с разной долей модифицированных записей (Изм. %): а, д, и – 0%; б, е, к – 0.5%, в, ж, л – 2.5%, г, з, м – 10%.

При этом распределения стандартных отклонений для группы 0 и группы 5 значительно перекрываются (Рис. 22). Распределения стандартных отклонений для группы 0 и модифицированных записей из группы 5 очень похожи. Распределение немодифицированных записей, попавших в группу 5, имеет длинный хвост в области высоких значений. Поэтому в большинстве случаев невозможно выбрать граничное значение, позволяющее разделить модифицированные и немодифицированные записи на основе стандартного отклонения предсказаний моделей.

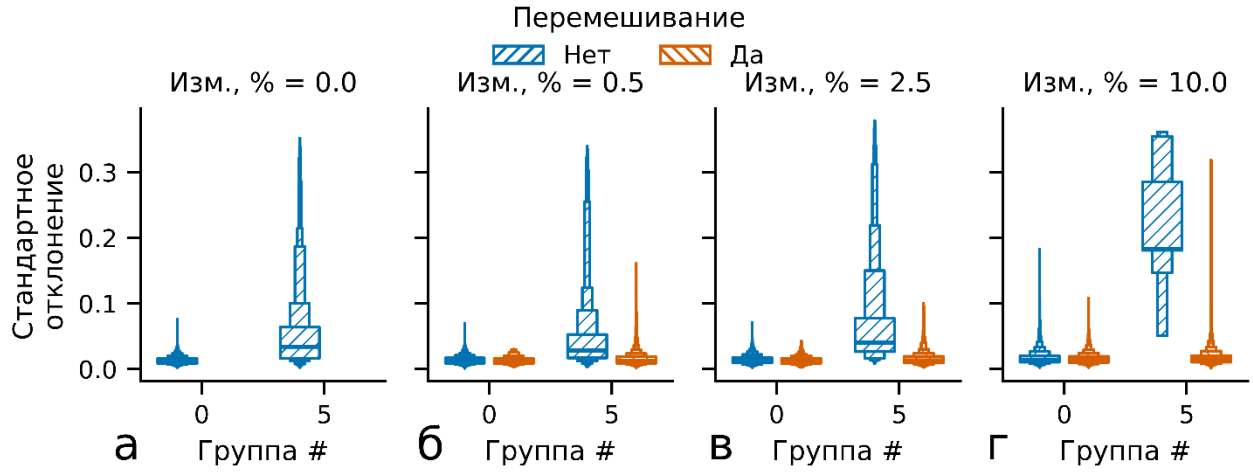


Рисунок 22. Распределения стандартных отклонений для групп 0 и 5 для перемешанных дескрипторных наборов данных: а – 0%, б – 0.5%, в – 2.5%, г – 10% модифицированных записей.

3.1.6 Сравнение эффективности фильтрования «желтых карточек» с более простыми подходами

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 (точности-полноты) для более простых методов поиска ошибок совпадают, так как каждое из значений квантиля соответствует определенной абсолютной ошибке для конкретного набора данных и предсказательных моделей. Тем не менее, на практике указанные методы применяются по-разному и требуют различных априорных знаний о наборе.

В этой работе для применения более простых подходов к фильтрованию рассчитывали медианное значение с использованием всех предсказаний моделей. Это позволило учесть вклад всех 5 моделей и сделало сравнение более корректным, чем использование только одной наиболее точной модели. Во всех рассмотренных случаях подход к фильтрованию с использованием «желтых карточек» превзошел более простые альтернативы, что видно из большей площади под кривыми precision-recall (точности-полноты) предлагаемого решения для перемешанных дескрипторных наборов данных (Рис. 23). Для наборов данных с большой долей модифицированных записей и большим значением добавленной ошибкой более простые подходы показали сравнимые с «желтыми карточками» результаты (Рис. 23).

Несмотря на то, что использование отсечения по абсолютной ошибке может казаться более естественным, этот подход опирается на априорные знания о природе записей и возможных ошибках в наборе данных, а также точности предсказания моделей. Фильтрование с использованием квантиля также предполагает наличие информации 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 определения числа ошибочных записей. Вместе с тем, использование ансамбля всех моделей делает их такими же вычислительно-затратными, как и использование «желтых карточек».

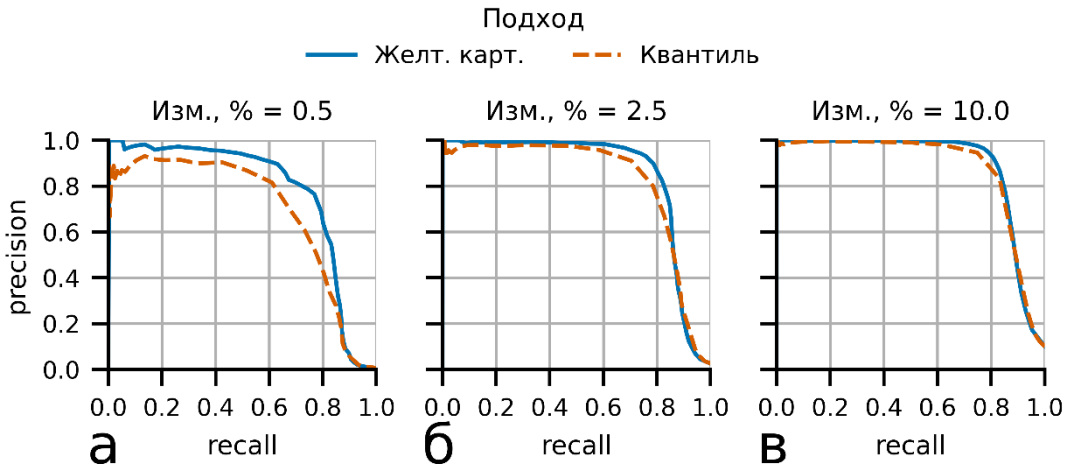


Рисунок 23. Графики precision-recall (точности-полноты) для метода квантильного отсечения и подхода «желтых карточек», примененного к набору данных на основе QM9 с различной долей перемешанных записей: а – 0.5%, б – 2.5%, в – 10%.

3.1.7 Выбор оптимального порогового значения t%

В отличие от более простых и часто применяемых способов очистки наборов данных от ошибочных записей, таких как квантильное отсечение, для подхода «желтых карточек» необязательно обладать априорной информацией о наборе данных для нахождения оптимального порогового значения *t%*. После получения предсказанных копий исходного набора данных становится возможным с минимальными вычислительными затратами варьировать параметр *t%,* что позволяет оценить интервал его значений, при котором точность поиска ошибок остается высокой.

При рассмотрении интегрального (кумулятивного) распределения количества записей в группе с максимальным количеством «желтых карточек» наблюдается возрастающая кривая (Рис. 24). При значении *t=0.0%* в группе с максимальным

количеством «желтых карточек» записи отсутствуют, при значении *t=100%* все записи в наборе данных отмечены как потенциально ошибочные. При этом для набора данных, не содержащих ошибочных записей, кривая не содержит явно выраженных перегибов, а в двух других рассматриваемых случаях наблюдается перегиб в начальной области, связанный с наличием модифицированных записей.

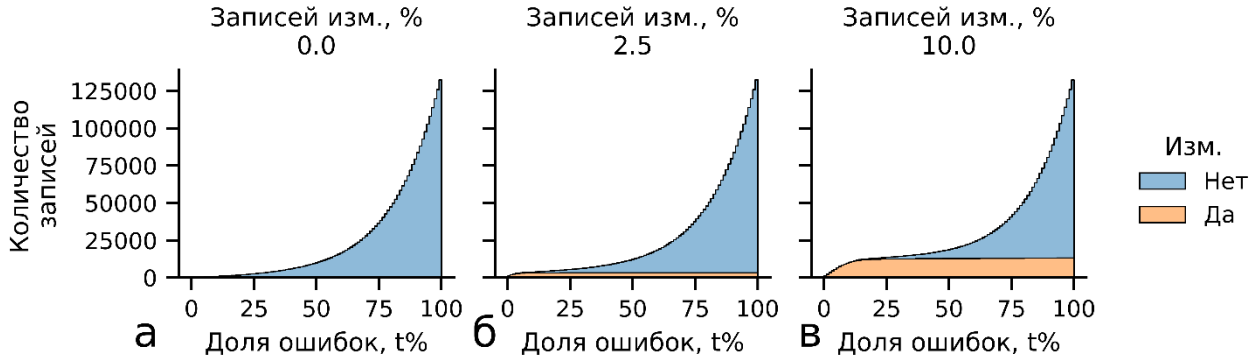


Рисунок 24. Зависимость количества потенциально ошибочных записей (в группе с максимальным количеством «желтых карточек») от порогового значения t%: а – 0%, б – 2.5%, в – 10% перемешанных записей.

Рассмотрение производной от этого графика (прироста количества записей) в зависимости от порогового параметра *t%* позволяет более явно выделить зону до точки перегиба (оранжевый пик на Рис. 25), в которой большая часть записей, отмеченных как потенциально ошибочные, является модифицированными. За этим пиком может следовать переходная зона с минимальным приростом количества записей в группе потенциально ошибочных записей, после чего рост продолжается за счет немодифицированных записей. При этом в случае набора данных без ошибок кривая является возрастающей и имеет экспоненциальный характер.

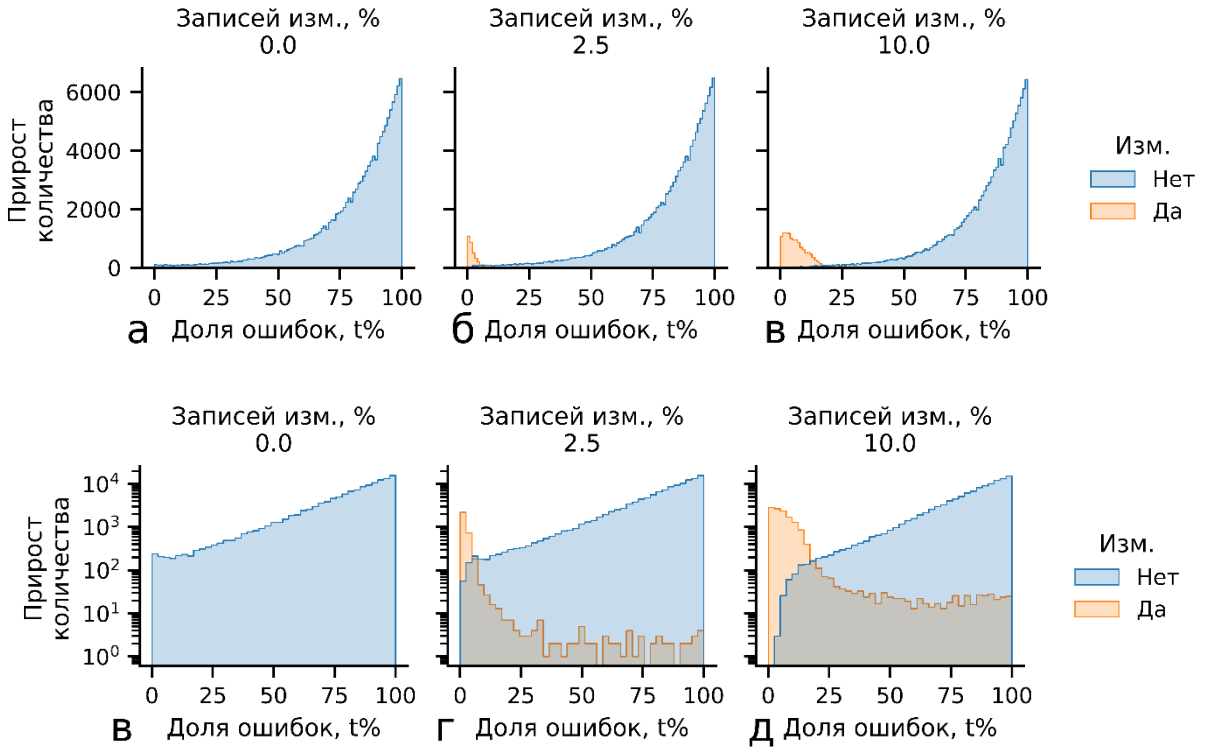


Рисунок 25. Зависимость прироста количества потенциально ошибочных записей (в группе с максимальным количеством «желтых карточек») от порогового значения t% для перемешанных QM9 наборов: а, г – 0%, б,д –2.5%, в,е – 10%. На рисунках (а – в) ось ординат в нормальных координатах, (г – е) – в логарифмических.

Стоит отметить, что для одного набора данных с разными типами и долей модифицированных записей правая часть кривой остается практически без изменений, так как в этой зоне прирост обеспечивается немодифицированными (корректными) записями, составляющих подавляющее большинство в любом наборе данных. При использовании логарифмической шкалы для оси ординат правая часть дифференциального распределения принимает форму, близкую к линейной (Рис. 25). Левый край кривой меняется в зависимости от типа и количества модифицированных записей, однако кривая для немодифицированных записей остается возрастающей.

Наличие пика в левой части графика позволяет оценить количество ошибочных записей. Так, наиболее консервативный вариант оценки основан на нахождении локального минимума на кривой прироста количества записей и проведении прямой, параллельной оси абсцисс, влево от этой точки (Рис. 26). Эта аппроксимация позволяет оценить снизу количество модифицированных (ошибочных) записей. Однако такая оценка может значительно занижать количество ошибочных записей, соответственно, завышая количество некорректно предсказанных. Это смещает отношение

модифицированных записей к немодифицированным, предполагаемая точность подхода может получиться заниженной.

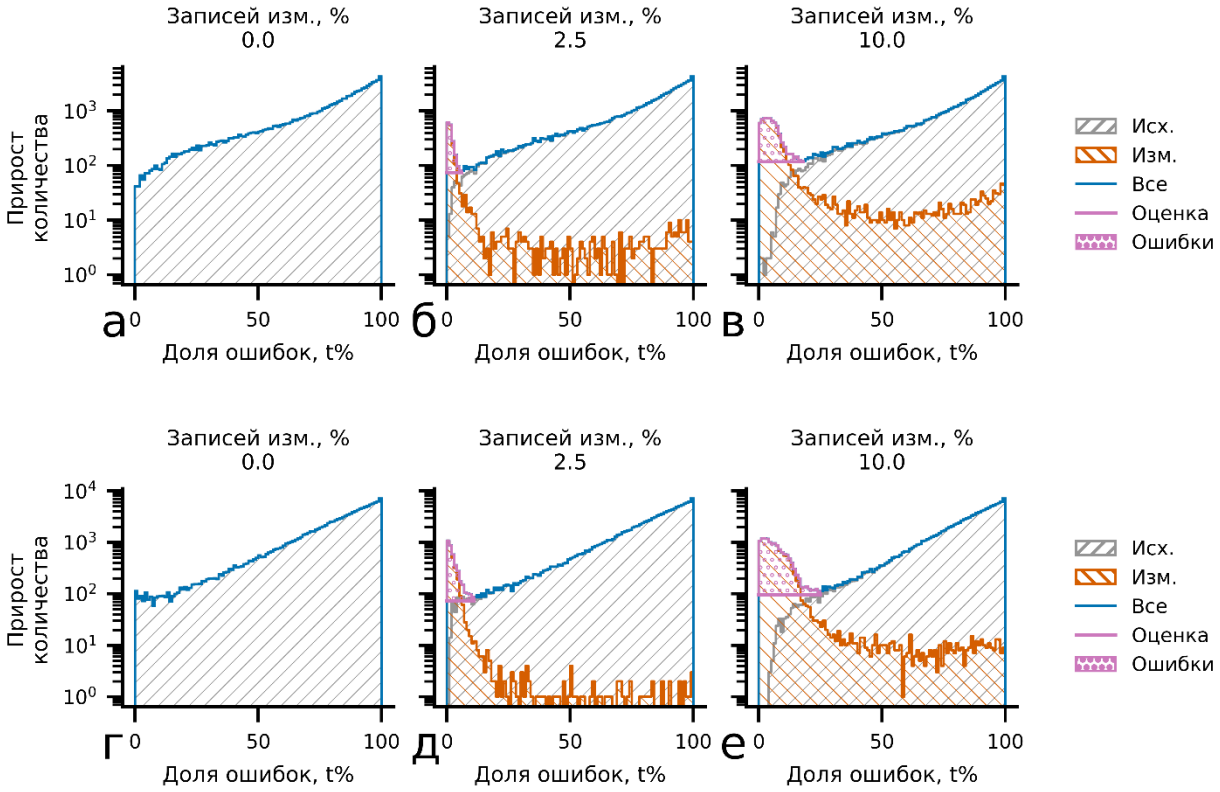


Рисунок 26. Аппроксимация зависимости прироста количества потенциально ошибочных записей (выделена фиолетовым, в группе с максимальным количеством «желтых карточек») от порогового значения t% для перемешанного дескрипторного (а – в) и QM9 (г – е) наборов данных: а, г – 0%, б, д – 2.5%, в, е – 10%. Синяя кривая отвечает всем записям, серая – исходным, оранжевая – перемешанным.

Использование аппроксимации кривой прироста количества записей от *t%* позволяет получить оценку точности поиска ошибок – главной характеристики подхода, что особенно полезно в случаях, когда подход не подходит для выбранных данных. Так, для рассмотренных синтетических наборов данных (дескрипторного и QM9) без ошибок оценочная точность не отличается от 0 (Рис. 27), это позволяет предположить, что в таком наборе нет ошибок, которые могли бы быть обнаружены с использованием подхода «желтых карточек». При этом, для перемешанных наборов с модифицированными записями подход дает достаточно точную оценку снизу (Рис. 27).

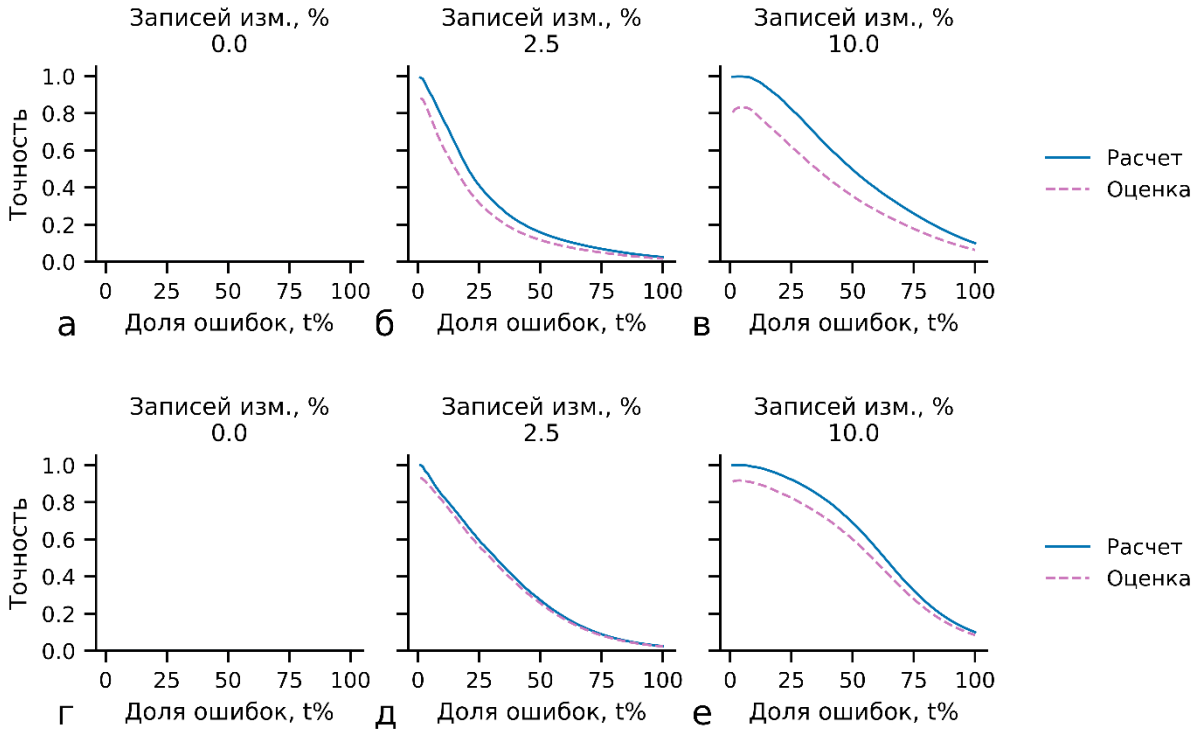


Рисунок 27. Оценка точности поиска ошибок с использованием аппроксимации зависимости прироста количества потенциально ошибочных записей (в группе с максимальным количеством «желтых карточек») от порогового значения t%. Фиолетовая кривая отвечает оценочной точности (на основе аппроксимации, Рис. 26), синяя – расчетной точности. На рисунке (а – в) – перемешанный дескрипторный набор данных, (г – е) – перемешанный QM9 набор данных: а, г – 0%, б, д – 2.5%, в, е – 10%.

В зависимости от поставленной задачи можно варьировать значение *t%*, увеличивая выборку потенциально ошибочных записей до достижения минимально допустимой точности. Например, для экспериментальной проверки стоит выбирать небольшое значение *t%* (оценочная точность поиска ошибок больше 0.8), а для наиболее полной очистки набора данных от ошибок может иметь смысл увеличивать *t%* пока точность не снизится до 0.5-0.6.

3.2 Поиск ошибок в NIST 17 RI

Поиск ошибок в NIST 17 RI производили с использованием одной из первых реализаций подхода «желтых карточек» [127]. Использовали пять независимых предсказательных моделей (Рис. 28) – одномерную и двухмерную сверточные нейронные сети (CNN1D и CNN2D, соответственно), полносвязную нейронную сеть (MLP), а также регрессию на опорные вектора (SVR), метод градиентного бустинга в реализации CatBoost. Архитектуры моделей глубокого обучения были адаптированы из

литературных источников [18] с небольшими изменениями для осуществления перехода с языка программирования Java на Python.

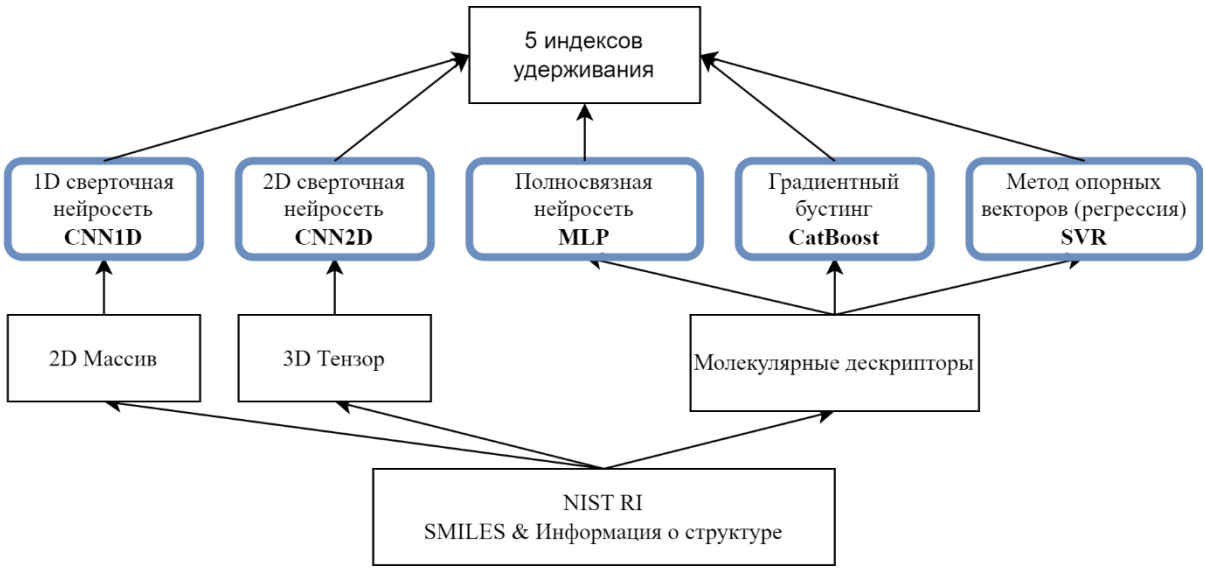


Рисунок 28. Пять независимых предсказательных моделей для индексов удерживания в NIST 17 RI.

Для обучения моделей использовали стандартные неполярные («standard non-polar», аналоги DB-1, 100%-полидиметилсилоксан) и полустандартные неполярные («semi-standard non-polar», аналоги DB-5, (5%-фенил)-полидиметилсилоксан) неподвижные фазы, как для капиллярных, так и для набивных колонок. Базу данных разбили на 5 частей, 4 из которых использовали для обучения, 1 – для предсказания таким образом, чтобы молекулы в двух разделах не пересекались. Процесс повторили 5 раз для каждой из моделей для получения полностью предсказанных копий исходной базы данных.

Полученные значения средней и медианной абсолютных ошибок предсказания (MAE и MdAE, соответственно) соответствовали уровню наиболее точных моделей на момент проведения исследований (Табл 3) и были сравнимы с нормальной экспериментальной погрешностью (до 20 е.и.у).

Таблица 3. Средняя (MAE), относительная (MPE) и медианная (MdAE) абсолютные ошибки предсказания индексов удерживания для пяти моделей

Модель	CNN1D	CNN2D	MLP	CatBoost	SVR
MAE, е.и.у	25.6	42.5	26.5	32.7	42.8
MdAE, е.и.у	12.0	24.1	13.7	20.0	27.6
MPE, е.и.у	1.7%	2.8%	1.9%	2.2%	2.9%

Провели оценку коэффициентов корреляции между ошибками предсказаний моделей, исключив наиболее некорректно предсказанные записи с ошибкой более 500 е.и.у, которые значительно искажали результат. Полученные коэффициенты корреляции не превышали 0.61, это позволяет использовать ансамбль этих моделей для поиска ошибок.

Таблица 4. Коэффициенты корреляции между абсолютными ошибками предсказания моделей

	Δ_{CNN1D}	Δ_{MLP}	$\Delta_{CatBoost}$	Δ_{CNN2D}	Δ_{SVR}
Δ_{CNN1D}	—				
Δ_{MLP}	0.60	—			
$\Delta_{CatBoost}$	0.61	0.58	—		
Δ_{CNN2D}	0.53	0.46	0.56	—	
Δ_{SVR}	0.53	0.44	0.57	0.51	—

Для каждой из моделей выбирали 5% записей с наибольшей как абсолютной, так и относительной ошибками одновременно и отмечали «желтой карточкой». На графике распределения количества записей от номера группы (Рис. 29) наблюдалась характерная для подхода «желтых карточек» аномальная форма кривой, описанная ранее, при которой в 5 группе содержалось больше записей, чем в 4. Также характерная для подхода форма кривой (с максимумом, Рис. 30) наблюдалась для зависимости стандартного отклонения от номера группы.

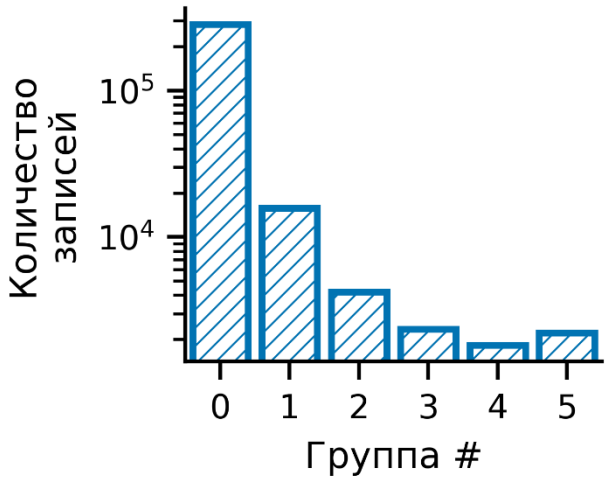


Рисунок 29. Распределение количества записей по группам с 0-5 «желтыми карточками».

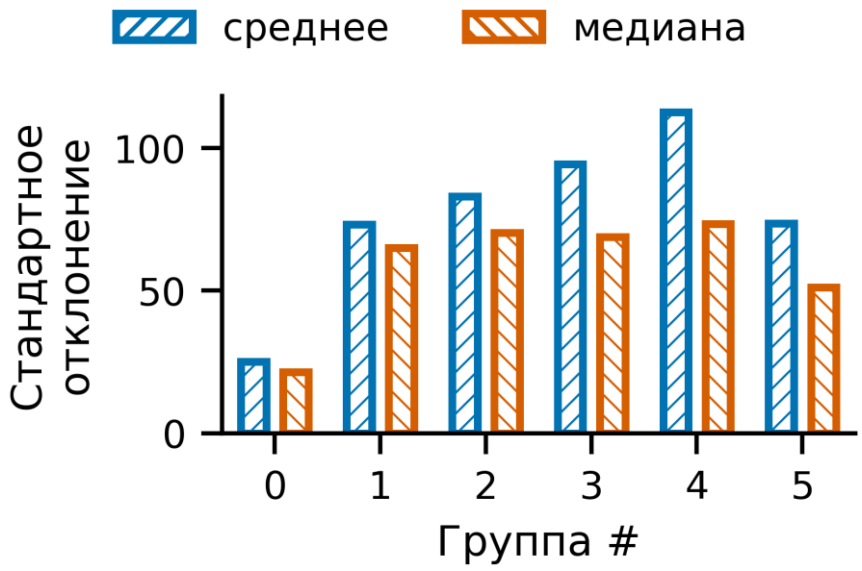


Рисунок 30. Изменение среднего и медианного стандартных отклонений в зависимости от количества «желтых карточек» в группе.

Следует отметить, что распределения стандартных отклонений предсказаний моделей для 0, 1 и 5 групп значительно перекрываются (Рис. 31), что ранее обсуждалось для синтетических данных. Это делает невозможным разделение плохо предсказанных и действительно ошибочных записей. Также заметны длинные правые хвосты у распределений из групп 1…5, что свидетельствует о существенном количестве записей, для которых согласованное предсказание индексов удерживания является проблематичным.

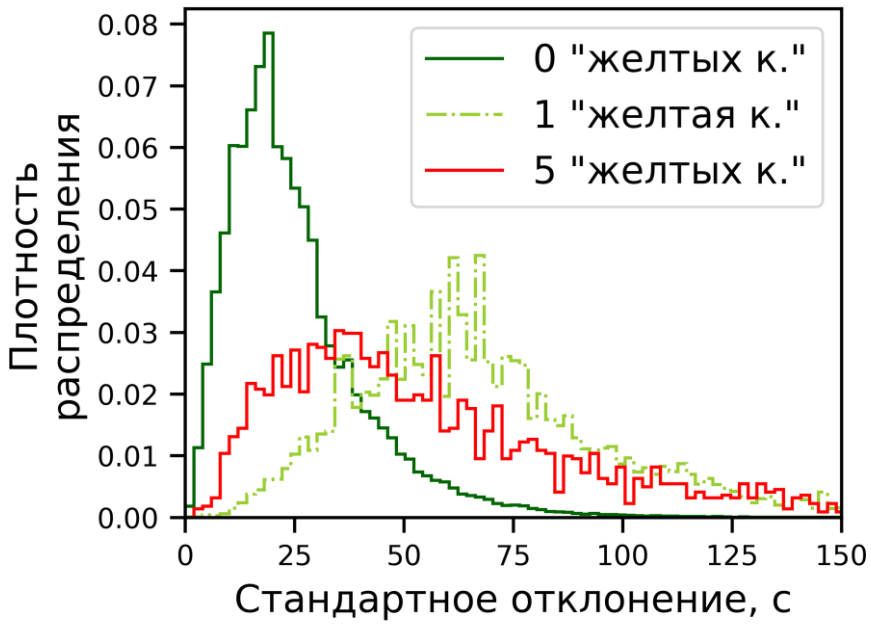


Рисунок 31. Распределение стандартных отклонений предсказаний индексов удерживания, полученных при использовании пяти моделей.

В результате найдено 2093 потенциально ошибочные записи (0.71% от общего числа). Рассмотрели распределение этих записей по группам и лабораториям, предоставивших наиболее значительное количество записей в базу данных (из доступных метаданных, описывающих источник происхождения индексов удерживания). Две наиболее крупные лаборатории – NIST MS Datacenter с примерно 13 000 записей и группа В.Г. Заикина (ИНХС РАН) с примерно 42 000 записей.

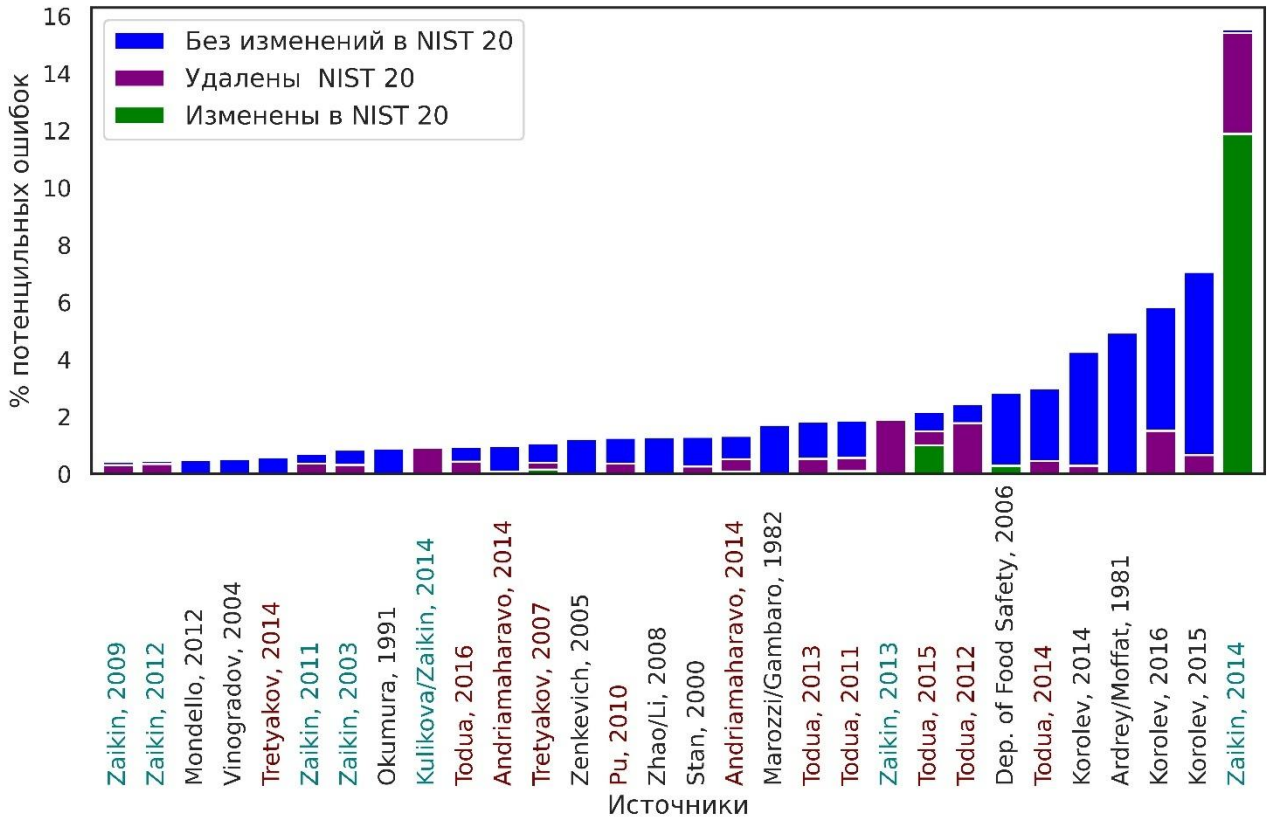


Рисунок 32. Выборка источников индексов удерживания с наибольшей долей потенциально ошибочных записей. Зеленым цветом отмечены исправленные записи, фиолетовым – удаленные, синим – оставленные без изменений в NIST 20 RI по сравнению с NIST 17 RI.

Также построили распределение ошибок по научным группам и лабораториям (Рис. 32). В среднем, в каждом из наборов индексов удерживания, зарегистрированных в определенной лаборатории, среднее количество потенциальных ошибок составляло около 1%, что крайне немного. Две вышеописанные группы также вошли и в этот список. В одном из указанных источников – индексах удерживания, предоставленных группой В.Г. Заикина в 2014 году, содержалось на порядок больше потенциальных ошибок по сравнению со средним значением. Сравнение с базой данных NIST 20 RI показало, что 219 из этих ошибочных значений был исправлены, 65 удалены и только 2 записи остались в неизмененном виде. Скорее всего, ошибки в этом источнике

появились в процессе переноса исходных значений в базу данных. Крупные правки и удаления были также отмечены в других источниках из списка. Можно предположить, что в NIST использовали более точную модель предсказания индексов удерживания, ставшую доступной после публикации NIST 17 RI [22], для обнаружения и исправления ошибок. Всего из 2093 потенциально ошибочных записей, обнаруженных нами в NIST 17 RI, в NIST 20 RI удалили 328, исправили 238.

3.3 Поиск ошибок в METLIN SMRT

Поиск ошибок в METLIN SMRT проводили с использованием усовершенствованного подхода «желтых карточек» [128]. Использовали 5 независимых предсказательных моделей: полносвязные нейронные сети с входными данными в виде молекулярных отпечатков пальцев (FCFP) и молекулярных дескрипторов (FCD), сверточную нейронную сеть (CNN), графовую сверточную нейронную сеть (GNN), градиентный бустинг в реализации CatBoost.

В полносвязных нейронных сетях использовали функцию активации SiLU, которая позволила улучшить точность предсказания и избежать переобучения без использования других подходов регуляризации. Сверточная и графовая нейронные сети использовали функцию активации LeakyReLU, что позволило немного увеличить точность по сравнению с ReLU, а с SiLU для этих архитектур получались не лучшие результаты. Все модели продемонстрировали приемлемую точность предсказания времен удерживания (Табл. 5), сопоставимую с наиболее актуальными подходами, известными в литературе [36,41] и экспериментальной погрешностью (оценочно, до 36 секунд средней и 18 секунд медианной [33]). Коэффициенты корреляции между ошибками моделей для адекватно предсказанных соединений также были умеренно низкими (Табл. 6).

Таблица 5. Средняя (MAE), относительная (MPE) и медианная (MdAE) абсолютные ошибки предсказания времен удерживания в METLIN SMRT. В скобках указаны стандартные отклонения

Метрика	GNN	CNN	FCFP	FCD	CatBoost
MAE, с	30.49 ± 0.08	37.44 ± 0.11	36.13 ± 0.05	40.30 ± 0.12	39.72 ± 0.05
MdAE, с	15.30 ± 0.06	21.30 ± 0.14	20.10 ± 0.04	23.50 ± 0.17	22.30 ± 0.11
MPE, %	3.74 ± 0.01	4.71 ± 0.02	4.45 ± 0.01	4.97 ± 0.01	4.91 ± 0.01

Таблица 6. Коэффициенты корреляции абсолютных ошибок предсказания моделей (стандартные отклонения указаны в скобках)

	ΔGNN	ΔCNN	ΔFCFP	ΔFCD	ΔCB
ΔGNN	—				
ΔCNN	0.408 ± 0.004	—			
ΔFCFP	0.455 ± 0.005	0.436 ± 0.001	—		
ΔFCD	0.343 ± 0.011	0.348 ± 0.012	0.386 ± 0.019	—	
ΔCB	0.392 ± 0.005	0.478 ± 0.007	0.563 ± 0.005	0.493 ± 0.011	—

Для получения предсказанных копий исходного набора данных использовали такой же подход, как и ранее, разбив список всех соединений на 5 частей, 4 из которых использовали для обучения, 1 для предсказания. Для каждой из моделей процесс повторили 5 раз и выбрали 5% записей с наибольшей абсолютной и относительной ошибками предсказания, после чего присвоили им по «желтой карточке». Как и ранее, каждая запись могла получить от 0 до 5 «желтых карточек».

Проверили устойчивость подхода к небольшим изменениям между повторениями, не связанными с изменением записей в тренировочных и предсказательных наборах, а именно: случайная инициализация параметров моделей, порядок записей внутри тренировочного набора и т.п. Получили 1550 ± 7 (стандартное отклонение, $n = 3$) потенциально ошибочных записей. Из них 1361 или 88% присутствовали во всех трех наборах одновременно.

После этого изменили разбиения на тренировочный и тестовый наборы, также три параллельных эксперимента. Среднее значение найденных потенциально ошибочных записей составило 1534 ± 14, причем 1299 или 85% из них встречались во всех трех наборах одновременно.

Эти два эксперимента показывают устойчивость подхода как к изменениям случайных параметров при обучении модели, так и к изменению разбиения на тренировочный и предсказательный наборы. Разница в наборах потенциально ошибочных записей обусловлена соединениями, для которых предсказания являются некорректными, небольшая доля таких соединений присутствует в группе с максимальным количеством «желтых карточек».

Как и ранее, стандартные отклонения предсказаний моделей могут служить индикатором для обнаружения потенциально ошибочных записей: при небольшом стандартном отклонении модели предсказывают согласованно друг с другом, соответственно, велика вероятность, что такие записи в группе с максимально возможным числом «желтых карточек» являются ошибочными. Если же стандартное отклонение предсказаний большое, то молекула достаточно уникальная, и результаты предсказания могут быть неточными.

В группе с 0 «желтых карточек» распределение стандартных отклонений (Рис. 33) имеет максимум в районе 20 секунд. Большая часть записей здесь предсказаны точно. При рассмотрении групп с большим числом «желтых карточек» соответствующее распределение сдвигается вправо и становится значительно шире (Рис. 33). Несмотря на это, распределение для 5 группы очень похожа на распределение для группы 0, с несколько более тяжелым правым хвостом, который вызван, скорее всего, наличием неточно предсказанных записей.

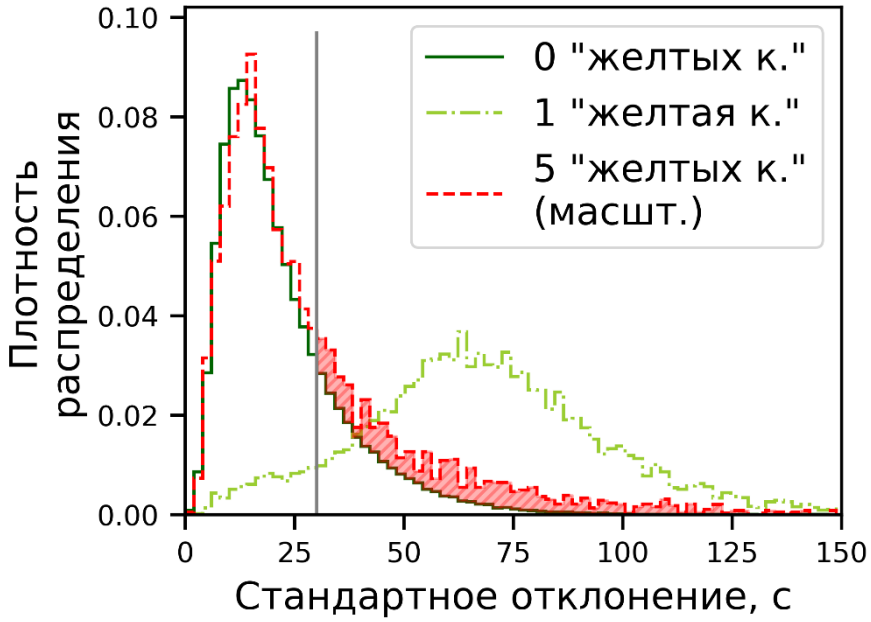


Рисунок 33. Распределения стандартных отклонений предсказаний пяти моделей для групп с 0, 1, 5 «желтыми карточками». Закрашенная часть отмечает предположительно некорректно предсказанные записи в группе потенциально ошибочных. Оценка получена сравнением распределения группы с 5 «желтыми карточками» с группой без «желтых карточек».

С использованием распределений для 0 и 5 группы можно оценить количество ошибочных записей в 5 группе. Для этого выбрали область распределения до 30 секунд стандартного отклонения, где большая часть веществ должна быть точно предсказана

(Рис. 33). Эту часть распределений использовали для нормировки, чтобы выбранные области совпадали, после чего вычли из отнормированного распределения 5 группы распределение для 0 группы. Провели обратное масштабирование разницы, что позволило оценить число неточно предсказанных записей в 5 группе. Получили оценку в 303 ± 11 (стандартное отклонение) некорректно предсказанные записи и, соответственно 1231 ± 11 потенциально ошибочные записи.

Другой метод оценки может быть основан на распределении числа записей по группам. Для этого зависимость для групп 0…4 аппроксимировали при помощи логарифмической функции (Рис. 34):

$$log2(\#compounds) = a + b/(\#group + c) \quad (13)$$

Логарифмическую форму зависимости выбрали для достижения приемлемого качества аппроксимации в условиях, когда группа 0 на порядок больше всех остальных. Для группы 5 экстраполяция полученной зависимости позволила получить оценку в 417 ± 23 (стандартное отклонение) неточно предсказанных записей и, соответственно, 1116 ± 23 ошибочных записей.

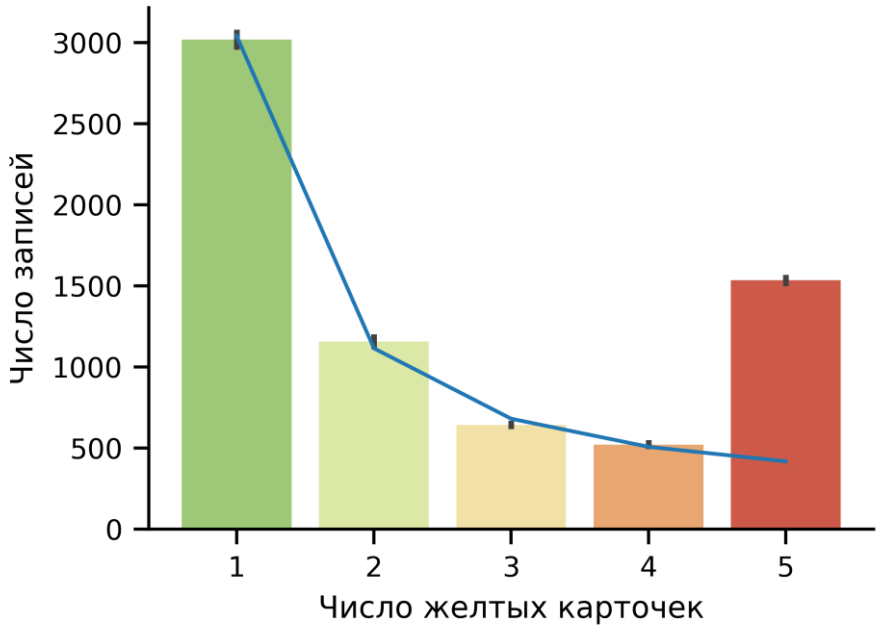


Рисунок 34. Распределение записей по группам с разным числом «желтых карточек». Кривая аппроксимирует значения для групп 0-4 и позволяет оценить количество некорректно предсказанных записей в группе с 5 «желтыми карточками» путем экстраполяции.

Два независимых способа оценки показали очень близкие результаты, которые также совпадают с пересечением из списков ошибок для 3 повторений. Таким образом,

набор потенциально ошибочных времен удерживания содержит около 1200 записей (1.5%), а размер группы с 5 «желтыми карточками» составляет примерно 1500 записей.

Выводы из главы 3

Предложен подход к поиску ошибок в химических базах данных, основанный на объединении предсказаний нескольких независимых моделей машинного и глубокого обучения при помощи механизма голосования. Эффективность фильтрования ошибочных записей оценили на синтетических наборах данных с контролируемо добавленными ошибками. Продемонстрировали, что предложенное решение работает лучше общепринятых альтернативных вариантов. Предложили алгоритм выбора ключевого варьируемого параметра – порогового *t%* записей, отмечаемых потенциально ошибочными. Разработанный подход применили для поиска потенциально ошибочных записей в базах NIST 17 RI и METLIN SMRT.

Глава 4. Обработка масс-спектральных данных[2]

Хроматомасс-спектральные данные активно используются в нецелевом анализе, при этом очевидно, что большую часть информации о структуре содержат масс-спектры, а не времена удерживания. Отчасти это объясняется тем, что масс-спектры являются векторами и как минимум по этой причине содержат больше информации, чем скалярные величины (времена удерживания). Это добавляет дополнительную сложность при обработке, сравнении и предсказании.

В результате хроматомасс-спектрального эксперимента генерируется большой объем данных. Один из самых простых способов уменьшения размера, особенно для приборов низкого разрешения, – округление значений *m/z* с плавающей запятой до целочисленных. Как и любая другая экспериментально полученная величина, значение *m/z* измеряется с некоторой случайной погрешностью, вызванной как аппаратными, так и программными особенностями приборов. При этом, для дальнейшей обработки не так важна дробная часть значений *m/z*, однако важно постоянство: значения *m/z*, принадлежащие одному фрагменту, должны, по возможности, округляться к одному целому.

4.1 Округление значений *m/z* до целочисленных

Проблема округления масс-спектров кажется тривиальной на первый взгляд, однако она слабо освещена в литературе, в большинстве случаев для обработки используют коммерческое программное обеспечение, алгоритмы работы которого неизвестны. Для того, чтобы установить алгоритмы округления, применяемые в коммерческом программном обеспечении, использовали модельные масс-спектральные файлы в распространённом универсальном формате NetCDF, которые конвертировали в табличный CSV. Доступные и установленные путем анализа результатов конвертации модельных файлов подходы к округлению значений *m/z*

---

2 При подготовке данной и последующих глав диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования:
***Khrisanfov M.**, Samokhin A.* A general procedure for rounding m/z values in low-resolution mass spectra // *Rapid Communications in Mass Spectrometry*. 2022. V. 36, № 11. EDN: PANTTY
***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. V. 15, № 4. P. 178–185. DOI: 10.5478/MSL.2024.15.4.178

приведены в Таблице 7. Как можно увидеть, в интервале значений *m/z* [*MZ* – 0.3; *MZ* + 0.4], где *MZ* – целое число, все программы демонстрируют одинаковое поведение. Разница в подходах к округлению проявляется в оставшемся интервале значений, где одно и то же точное значение *m/z* может быть округлено к разным целочисленным. Такое поведение видно на следующем примере: в масс-спектре электронной ионизации гексадецилового эфира октадекановой кислоты ($C_{34}H_{68}O_2$) из NIST WebBook [143] присутствует пик молекулярного иона. Интенсивность максимального пика в кластере превышает 10%, точное значение – 508.5214. Используя алгоритмы из Таблицы 7, можно установить, что с использованием AMDIS и ChromaTOF это значение будет округлено до 508, с MS Search и OpenChrom – до 509, а с ChemStation не войдет при экспорте в табличный CSV файл. Стоит отметить, что пользователи MS Search могут изменять алгоритм округления, что приведет к другим результатам.

Таблица 7. Алгоритмы округления значений m/z до целочисленных в распространенном программном обеспечении

Программа	Правило округления	Интервал округления к целому MZ	Интенсивность двух совпадающих пиков (округляющихся к одному MZ)
OpenChrom	MZ = ⌈mz - 0.505⌉	[MZ - 0.495; MZ + 0.505]	Сохраняется, суммируется
AMDIS	MZ = ⌈mz - 0.649⌉	[MZ - 0.351; MZ + 0.649]	
ChromaTOF	MZ = ⌈mz - 0.700⌉	[MZ - 0.300; MZ + 0.700]	
ChemStation*	$\lvert mz - \lfloor mz \rfloor \rvert \leq 0.400$ <=> MZ = ⌈mz - 0.500⌉	[MZ - 0.400; MZ + 0.400]	Сохраняется только пик с максимальной интенсивностью

Влияние неоднозначного округления на результаты анализа данных, ранее не исследовалось. В настоящей работе данная проблема была рассмотрена для данных газовой хроматмасс-спектрометрии. Для большинства анализируемых таким методом веществ молекулярная масса не превышает 600 а.е.м. Например, в базе данных индексов удерживания NIST RI 17 только 1.7% веществ обладают большей молекулярной массой. Более того, в случаях, когда молекулярная масса превышает 600 а.е.м., пики молекулярного иона могут отсутствовать в масс-спектре электронной

ионизации из-за интенсивной фрагментации [144]. Было найдено, что именно так происходит с 40% веществ, представленных в масс-спектральной базе данных NIST. Поэтому в этом исследовании рассматривали только однозарядные ионы с отношением *m/z* 600.

Плотность распределения значений *m/z* с плавающей запятой внутри одного интервала имеет форму зависимости с максимумом. Каждый такой пик состоит из нескольких частей: основная часть ([MZ – 0.3; MZ + 0.4]), отвечающая примерно 90% всей площади, и хвосты ([MZ + 0.4; MZ + 0.7]). Распределения, полученные для набора данных, основанного на базе данных NIST (все комбинаторно возможные фрагменты), шире и с более выраженными хвостами, по сравнению с молекулами из PubChem. Такую форму можно объяснить наличием большого количества фрагментов, которые не существуют в реальности, и обладают большим по модулю дефектом масс. Пики для распределения, полученного из данных PubChem, более узкие и с менее выраженными хвостами, потому что для существующих соединений составы молекул гораздо более ограничены и имеют меньший по модулю дефект масс.

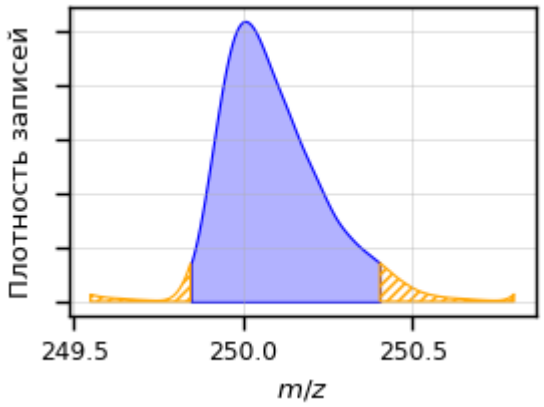


Рисунок 35. Распределение точных m/z возможных фрагментов, основная часть площади отмечена синим, зона минимальной плотности – оранжевым.

Для каждого из пиков рассчитали границы, отвечающие 95% от всей площади (Рис. 35). Очевидно, что надежный алгоритм округления значений *m/z* до целочисленных должен разделять интервалы в областях с наименьшим количеством *m/z*. Такой подход позволяет снизить влияние случайной погрешности на результаты округления. Как видно из Рисунка 36, такие значения для границы интервалов изменяются вместе со значением *m/z*.

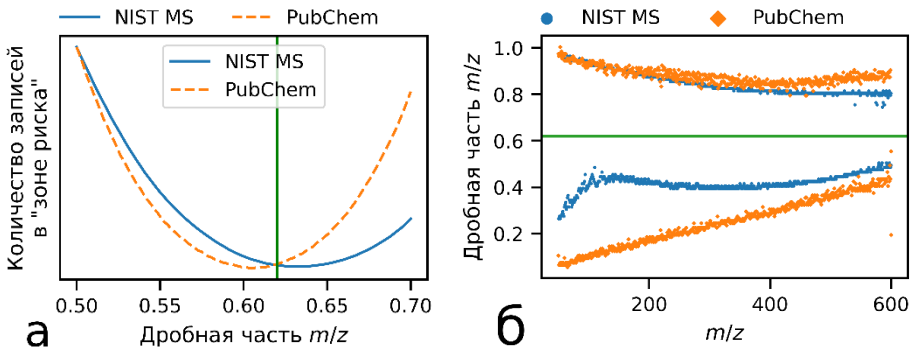


Рисунок 36. Зависимость количества фрагментов, попадающих в зону неопределенного округления от границы разделения пиков (а). Левая и правая границы оптимальной зоны границы округления (б). Оранжевые кривые для PubChem и синие – для NIST. Зеленым цветом отмечено предлагаемое правило для округления значений m/z до целочисленных (0.62).

Также построили зависимости количества фрагментов, попадающих в зону неопределенного округления, обусловленного случайными погрешностями, от граничного значения, разделяющего соседние пики *m/z* (Рис. 36). Для фрагментов NIST MS оптимальное значение составило 0.635, а для молекул из PubChem – 0.605. Средним арифметическим является значение в 0.620. При усреднении границ для пиков в диапазоне [500; 600] а.е.м. по набору данных, основанном на NIST, было также получено значение в 0.620. Для более низких значений *m/z* это граничное значение также хорошо подходит, так как в этих случаях еще меньше значений *m/z* с плавающей запятой лежит в этой области. Таким образом, для простоты выбрали одно значение, учитывающее распределения легких и тяжелых фрагментов как в PubChem, так и NIST MS. Все значения внутри интервала [*MZ* – 0.38; *MZ* + 0.62] округляют к целочисленному *MZ*.

Производители современных квадрупольных масс-спектрометров заявляют точность и воспроизводимость измерения значений *m/z* на уровне 0.1 или лучше. Были рассмотрены два критических значения ошибок измерения, когда разница между теоретическим из измеренным значением *m/z* была меньше или равна 0.05 или 0.1. Это позволило оценить вероятность, что точные значения *m/z* будут округлены до целочисленных случайным образом. Полученные результаты отражены в Таблице 8. Алгоритм, применяемый в AMDIS, позволяет получить наименьшую из всех рассмотренных существующих программ вероятность неоднозначного округления, вызванную ошибками измерения. Стандартное математическое правило округления, используемое в OpenChrom (версии 1.3), работает значительно хуже. Для фрагментов с

большими значениями *m/z*, характерных для жидкостной хроматографии – масс-спектрометрии, такое правило работает еще хуже.

Таблица 8. Доля фрагментов из модельных наборов, основанных на PubChem (только соединения) и NIST MS (все комбинаторно возможные фрагменты), попадающих в зону неопределенного округления, зависящего от случайных погрешностей.

Алгоритмы	Граница округления x: [MZ-1+x; MZ+x] →MZ	Δm/z = 0.05 Да		Δm/z = 0.1 Да	
		PubChem,%	NIST,%	PubChem,%	NIST,%
ChemStation	0.4*	2.33	6.57	4.95	10.53
OpenChrom	0.5	0.37	1.46	1.02	3.58
Предлагаемый	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

Наибольшая разница между двумя программами, рассмотренными в этом исследовании, была обнаружена между OpenChrom и ChromaTOF. Точные значения масс были округлены по-разному в 4.0% и 1.2% случаев для данных, основанных на NIST и PubChem, соответственно. Кроме того, ChemStation отбрасывает до 3.6% (NIST) и 1.0% (PubChem) пиков от всех значений *m/z*. Несмотря на то, что эти значения могут казаться незначительными, они могут повлиять на конечный результат анализа. Поэтому использование одинакового правила для округления могло бы улучшить все программные продукты. Оптимальным решением является использование алгоритма округления, реализованного в программе AMDIS, поскольку:

- AMDIS – одна из наиболее популярных и доступных программ для обработки данных газовой хроматографии масс-спектрометрии
- Разница между предложенным правилом округления и используемым в AMDIS очень мала
- Некоторые другие программы (MetAlign) используют почти тот же самый алгоритм.

Также целесообразно рассмотреть достоинства и недостатки округления значений *m/z* при использовании библиотечного поиска. Первые масс-спектрометры, способные передавать данные в цифровом формате, использовали целочисленные значения *m/z* из-за ограничений собственной компонентной базы и также небольшой вычислительной мощности компьютеров. Кроме того, первые компьютеризированные системы библиотечного поиска появились в начале 1970х и использовали только несколько наиболее интенсивных пиков в масс-спектре, а также дополнительно обрабатывали данные перед поиском [145–148]. Полные масс-спектры стали использовать для поиска больше чем через десятилетие [149]. С 1990х такой подход стал обычным, чему существенно поспособствовало создание цифровых версий библиотек NIST и Wiley [150]. В то время наиболее широко использовались два алгоритма: вероятностное сравнение (probability-based matching, PBM), разработанное группой под руководством МакЛафферти [151,152] и алгорим Composite/Identity, разработанный Штейном и Скоттом [153]. Алгоритм Identity используется в настоящее время в программе MS Search. Более того, многие программные продукты, такие как ChromaTOF, Xcalibur, и т.д. используют этот алгоритм через динамическую библиотеку, предоставляемую NIST MS. Алгоритм PBM доступен только для пользователей ChemStation. Недавно было показано, что Identity лучше справляется с задачей ранжирования, чем PBM [154].

В 2000х квадрупольные масс-спектрометры и компьютеры стали достаточно производительными, чтобы обрабатывать значения *m/z* как величины с плавающей запятой. MS Search, начиная с версии 2.2 позволяет импортировать масс-спектры с дробными значениями *m/z*, однако сравнение происходит с базой данных масс-спектров электронной ионизации NIST, в которой все данные являются целочисленными. Кроме того, процесс импорта масс-спектров недостаточно полно описан в инструкции к NIST MS Search.

Более того, округление значений *m/z* до целочисленных может вызвать проблемы с интерпретацией масс-спектров. Несмотря на то, что ионы с двойным зарядом нечасто встречаются при электронной ионизации, они могут образовываться для определенных классов органических веществ. Например, в масс-спектре высокого разрешения карбазола (получен на Leco PegasusHT) присутствуют пики с *m/z* 82.5284 и 83.5362. Из-за высокой точности определения масс (<1 ppm) при наложении жестких ограничений

на возможные молекулярные формулы удается однозначно установить состав фрагментов: $^{12}C_{12}{}^{1}H_{7}{}^{14}N^{++}$ и $^{12}C_{12}{}^{1}H_{9}{}^{14}N^{++}$. Мы сравнили масс-спектры карбазола, представленные в NIST и MassBank. В трех случаях эти значения *m/z* были 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, а 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 наиболее современных выпусках масс-спектральных библиотек.

Решением могло бы стать включение в современные библиотеки значений *m/z* с хотя бы одной значащей цифрой после запятой. Для совместимости с более старыми масс-спектрами и алгоритмами можно использовать единое правило для округления, снижающее вероятность неоднозначного округления.

Выводы

Таким образом, AMDIS, ChemStation, ChromaTOF, MS Search и OpenChrom используют различные алгоритмы для округления нецелочисленных значений *m/z*. ChemStation при экспорте в CSV исключает из рассмотрения пики в интервале [MZ + 0.4; MZ + 0.6], что может вызывать сложности с интерпретацией для двухзарядных ионов. В этой работе сгенерировали два модельных набора данных на основе молекул из PubChem и всех комбинаторно возможных фрагментов из NIST 17 MS, что позволило предложить более устойчивый алгоритм для округления значений из отрезка [MZ – 0.38, MZ + 0.62] до целочисленного MZ, который очень близок к используемому в программе AMDIS [MZ – 0.351, MZ + 0.649].

4.2 Предсказание масс-спектров электронной ионизации по структуре молекулы

Как упоминалось в обзоре литературы, масс-спектральные библиотеки широко используются в нецелевом анализе. При этом их размер значительно уступает общехимическим базам данных, поэтому для сокращения этого разрыва используют

подходы к предсказанию масс-спектров по структуре молекулы. На текущий момент существует не так много общедоступных подходов, решающих эту задачу, при этом не все из опубликованных методов можно с одинаковой легкостью использовать на практике (Табл. 9). Так, например, для наиболее «актуального» алгоритма MSProject (2025) [82] недоступны не только веса обученной модели, но и исходный код и даже подробное описание. Для чуть более старого подхода RASSP (2023) [81] авторы достаточно подробно описывают оба используемых алгоритма в статье, также доступен исходный код и веса обученных моделей. Однако, как выяснилось при тестировании, в коде присутствуют ошибки, мешающие его корректной работе под операционной системой Windows, а использование предобученных весов приводит к получению крайне неточных предсказаний.

Таблица 9. Сравнение основных параметров доступных подходов к предсказанию масс-спектров электронной ионизации

Подход	Язык программирования	Доступные ресурсы	Набор данных
CFM-EI (2016) [79]	C++	Скомпилированные файлы для Windows, исходный код, веса модели	NIST14
NEIMS (2019) [80]	Python 3.6	Исходный код, веса обученных моделей	NIST17
RASSP (2023) [81]	Python 3.7	Исходный код, веса моделей (низкое качество предсказания)	NIST17
MSProject (2025) [82]	Python 3.8	X	X

При этом для двух наиболее ранних из доступных современных подходов – CFM-EI (2015) [79] и NEIMS (2019) [80] представлен как исходный код, так и предварительно обученные веса моделей, а результаты, заявленные в статьях, воспроизводятся при независимом тестировании. К сожалению, и эти подходы не лишены недостатков. Так, авторы CFM-EI предоставили исполняемые файлы только для системы Windows, компиляция для Linux и Mac возможна, но значительно затруднена из-за необходимости использовать сторонние библиотеки. Для Linux проблему решили независимые разработчики в программе SVEKLA [92], в которой также был реализован графический интерфейс пользователя. Вместе с тем, расчет одного масс-спектра на современном процессоре с использованием CFM-EI составляет несколько секунд на молекулу, что на

три порядка больше по сравнению с нейросетевыми подходами, такими как NEIMS. Алгоритм CFM-EI задействует всего 1 поток для вычислений, несмотря на возможность использовать 4…16 на современных системах. Эти проблемы делают применение CFM-EI затруднительным для генерирования больших наборов данных.

При использовании алгоритма NEIMS [80] проблемы могут возникнуть на этапе установки, так как использованная авторами версия окружения Python 3.6 и соответствующие библиотеки устарели к 2026 году. Несмотря на это NEIMS показывает хорошую точность предсказания масс-спектров и высокое быстродействие, на расчет одного масс-спектра требуется всего несколько миллисекунд. Однако из-за использования устаревших библиотек не поддерживается ускорение вычислений с использованием графического адаптера, активно применяемое в настоящее время.

Предложенная оптимизированная реализация архитектуры NEIMS (NEIMS-PyTorch) базируется на идеях оригинальной статьи [80], однако использует более современную связку Python 3.12 и PyTorch 2.8, которая будет поддерживаться еще долгое время. Кроме того, библиотека PyTorch позволяет использовать видеокарту для значительного ускорения расчетов на Linux и Windows.

Оригинальная структура, основанная на нейронной сети с тремя выходами, была сохранена. Основными изменениями стали переход на библиотеку PyTorch для глубокого обучения, изменение функции активации с классической ReLU на более современную SiLU, а также использование библиотеки Optuna для подбора гиперпараметров, таких как число нейронов в скрытых слоях, количество слоев и значения случайного исключения (dropout). Были также опробованы несколько альтернативных вариантов архитектуры, использующие комбинации слоев, не представленные в оригинальной публикации, другое количество выходов, а также применение графовой сверточной сети в основной части предсказательной модели. Все опробованные альтернативные подходы, кроме использования графовых слоев, показали более низкую точность предсказания на валидационном наборе, поэтому далее не рассматривались.

Точность предсказания моделей в значительной мере зависит от тренировочного набора данных – чем больше в нем записей, чем они более разнообразные по своей природе – тем лучше итоговая точность предсказания для новых молекул, которые в этот набор не входили. Все модели, используемые в работе, были обучены на версиях

базы данных NIST MS, выпущенных не позже 2017 года (2017 для RASSP, NEIMS, NEIMS-PyTorch, 2014 для нейросетевой части CFM-EI). Использование в качестве тестового набора масс-спектров соединений, добавленных в релизах NIST 20 и NIST 23, позволяет провести корректное сравнение подходов без необходимости сравнивать тренировочный набор каждой из моделей с тестовым.

В качестве количественной меры для оценки качества предсказания масс-спектров использовали степень совпадения между предсказанным и экспериментальным масс-спектрами, рассчитываемую как взвешенный косинусный коэффициент (weighted cosine similarity, WCS или Similarity):

$$Similarity(\mathrm{W_{pred}}, \mathrm{W_{lib}}) = \frac{\mathrm{W_{pred}} * \mathrm{W_{lib}}}{(\| \mathrm{W_{pred}} \| * \| \mathrm{W_{lib}} \|)} \quad (14)$$

где W = m√I, а ‖W‖ - норма вектора.

Реализация NEIMS-PyTorch показала точность предсказания немного лучше оригинального алгоритма NEIMS, эти два варианта реализации одной идеи превзошли все другие подходы (Рис. 35). В 40% случаев для оригинальной NEIMS и 50% для NEIMS-PyTorch степень совпадения превышала 0.8, то есть предсказанные масс-спектры были очень похожи на экспериментальные. При этом доля записей с Similarity > 0.9, наиболее близких к экспериментальным (при сравнении с использованием Identity это значение соответствует «отличному, достоверному совпадению» согласно инструкции NIST MS Search [76]), для NEIMS составила 5%, а для NEIMS-PyTorch – 28%.

CFM-EI несмотря на то, что был предложен почти 10 лет назад, существенно превзошел алгоритм RASSP (Рис. 37). Хотя CFM-EI значительно уступает NEIMS по точности предсказания, CFM-EI присваивает каждому из пиков молекулярную формулу, что может быть полезно при ручной интерпретации результатов предсказания. Обе части алгоритма RASSP - SubsetNet и FormulaNet с предобученными параметрами моделей, предоставленными авторами, не смогли достичь заявленных в оригинальной публикации результатов [81], что также подтверждено другой группой исследователей [155].

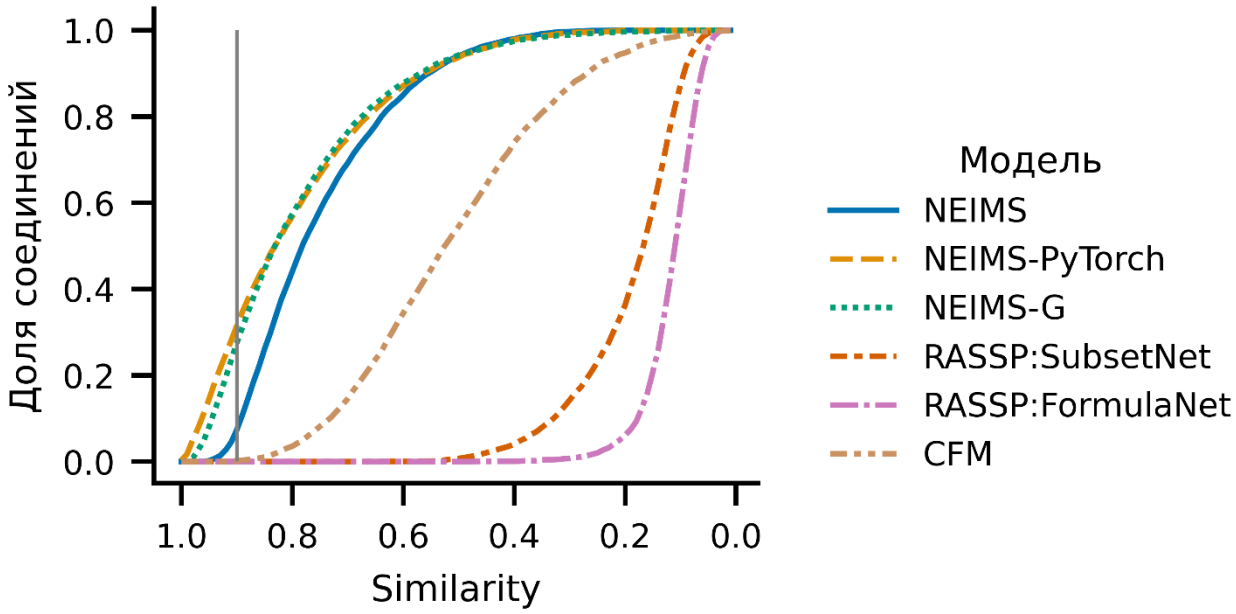


Рисунок 37. Кумулятивные распределения степени совпадения, рассчитанной между предсказанным и экспериментальным масс-спектрами для 5000 пар масс-спектров.

Для распространения подхода NEIMS-PyTorch была выбрана комбинированная схема, предоставляющая удобный доступ для исследователей с разными уровнями навыков в области программирования и системного администрирования. Все необходимые инструкции и код для запуска и обучения нейронной сети были выложены в репозитории на GitHub [156] в виде библиотеки для языка программирования Python, которая может быть легко установлена вместе со всеми необходимыми зависимостями в несколько команд. Параметры обученной модели доступны в репозитории HuggingFace [157]. Для улучшения совместимости в будущем, все необходимые зависимости и веса модели были также упакованы в Docker контейнер и размещены на DockerHub [158], который обеспечит легкий запуск даже через несколько лет, независимо от устаревания использованной версии Python и соответствующих библиотек.

Выводы

Таким образом, были усовершенствованы не только архитектура предсказательной модели и точность предсказания масс-спектров, но и способы распространения обученной модели, которую теперь проще использовать благодаря реализации с использованием актуальной версии Python и соответствующих библиотек. Упаковка всех зависимостей в Docker контейнер значительно улучшает

простоту и удобство запуска в долгосрочной перспективе даже при отсутствии поддержки разработчиками.

4.3 Предсказание структурных фрагментов по масс-спектрам электронной ионизации

Альтернативным сравнению масс-спектров подходом является предсказание молекулярных отпечатков пальцев по масс-спектрам, а затем их сравнение с рассчитанными по структуре молекулы. Это также позволяет использовать обширные общехимические библиотеки в нецелевом анализе. Наиболее актуальным из доступных решений является DeepEI [105].

Для сравнения с уже существующим подходом DeepEI были выбраны ECFP6 и MACCS молекулярные отпечатки пальцев, так как именно их использовали в оригинальной публикации [105], кроме того, они принадлежат к разным типам – ECFP6 рассчитывают при помощи хэш-функции, а MACCS используют предопределенные SMARTS запросы для отражения структурной информации о молекуле. Стоит отметить, что оба вида отпечатков можно сгенерировать при помощи RDKit. При воспроизведении оригинальной публикации DeepEI были обнаружены несоответствия на этапе выбора количества молекулярных отпечатков пальцев для сравнения. Так, в оригинальной публикации предсказывали только отдельные значения (биты) молекулярных отпечатков пальцев, которые принимали положительные значения для 10…90% молекул в NIST MS. При этом, число выбранных MACCS запросов, приведенное в оригинальном тексте (44), значительно отличается от числа, приведенного в дополнительных материалах (93) [105]. В этой работе были отобраны 95 MACCS значений. Такие же проблемы возникли при определении длины вектора ECFP6: 67 упомянуто в статье, 61 – в дополнительных материалах. В работе было выбрано 62 молекулярных отпечатка пальцев ECFP6.

В настоящей работе для проведения корректного сравнения оригинальный подход DeepEI [105] был реализован самостоятельно. При оценке значений функции потерь для тренировочного и валидационного наборов данных выявили, что все нейросетевые модели в составе DeepEI значительно переобучаются. Каждая из моделей обучается до оптимального состояния за 4 эпохи, после чего значения функции потерь начинают увеличиваться.

В этой работе также были предложены и опробованы альтернативные варианты архитектуры без использования регуляризации – слоев BatchNorm (групповое нормирование) или Dropout (случайное исключение), которые также переобучались. Добавление вышеперечисленных слоев решило проблему переобучения без уменьшения точности предсказания. Это также позволило сделать архитектуру модели более глубокой и получить меньшее значение функции потерь для валидационного набора.

Изначально в архитектурах предсказательных моделей использовали функцию активации ReLU, которая, фактически, является стандартной для полносвязных нейронных сетей. Однако в ходе разработки были опробованы несколько альтернативных недавно предложенных функций: Leaky ReLU, GELU, SiLU. Leaky ReLU согласно литературным данным должна решать проблему затухающих градиентов, однако в использованных архитектурах эта проблема не проявляется, поэтому значительных улучшений замечено не было. GELU добавляет случайное поведение и функцию случайного исключения к функции ReLU путем умножения входных значений на кумулятивное нормальное распределение. Эту функцию тестировали в нескольких промежуточных вариантах архитектуры, после чего заменили на SiLU.

Было показано, что даже простая замена ReLU на SiLU позволяет улучшить результаты для глубоких нейросетевых моделей [120]. Предполагается, что это является прямым следствием свойств самой функции: неограниченность при x>>1 позволяет избежать насыщения и уменьшения скорости обучения из-за маленького размера градиентов, ограничения при x<<0 выражаются в сильном регуляризующем влиянии на нейросетевую модель. Немонотонность SiLU улучшает распространение градиентов по сравнению с ReLU, а ее гладкость улучшает сходимость при оптимизации и способность нейросетевой модели к обобщению [120]. Даже при использовании в сравнительно неглубоком варианте модели в этой работе были отмечены улучшения при переходе на SiLU. Уменьшились значения функции потерь на валидационном наборе данных, а также уменьшилась разница значений потерь между тренировочным и валидационным наборами, что свидетельствует об улучшении обобщающих способностей модели. Кроме того, графики функции потерь стали более гладкими как для тренировочного, так и для валидационного наборов данных.

В подходе DeepEI каждая из нейросетевых моделей предсказывает только один бит молекулярных отпечатков пальцев [105]. Из-за этого сильно страдает скорость предсказания, а также значительно увеличивается сложность процесса из-за работы с десятками-сотнями моделей. В связи с этим одной из наших задач было улучшить точность предсказания, использовав при этом только одну модель для предсказания сразу всех отпечатков пальцев одновременно (ECFP6 и MACCS). Также эта модель должна быть достаточно быстрой с относительно небольшим числом весов для упрощения процедуры встраивания в более сложные программные продукты и давать возможность проводить предсказания на среднестатистическом лабораторном компьютере без задействования видеокарты.

Было выбрано два направления для оптимизации архитектуры, основное из которых минимизация значений ошибок предсказания на валидационном наборе. Результатом стала модель EI2FP:Full. Второе направление было направлено на разработку более легкой модели EI2FP:Lite, незначительно уступающей по точности, но с меньшим размером. Один из крайне эффективных способов понизить количество гиперпараметров в модели – использовать сверточные слои. Кажется, что такой подход должен хорошо работать для масс-спектров электронной ионизации, в которых существуют кластеры пиков, каждый из которых можно обработать одним фильтром независимо от других кластеров. К сожалению, предварительное тестирование показало, что применение сверточных нейросетевых моделей не позволило сохранить точность предсказания и уменьшить количество параметров одновременно. Только сравнимые по количеству параметров и скорости работы с полносвязными версии со сверточными слоями показывали сопоставимую точность предсказания. Поэтому приняли решение сконцентрироваться на оптимизации полносвязной архитектуры.

На этапе поиска облегченной архитектуры модели параллельно применяли следующие подходы: (i) увеличивали количество слоев с одновременным уменьшением количества нейронов в слое, (ii) уменьшали количество слоев с увеличением количества нейронов в слое, затем последовательно уменьшали количество нейронов до тех пор, пока разница в точности предсказания по сравнению с полной моделью не станет значительной. Итоговая архитектура, подобранная для модели EI2FP:Lite характеризуется на 9% более высокими значениями ошибок BCELoss для значений с плавающей запятой (которые не всегда напрямую увеличивают ошибку предсказания

бинарных молекулярных отпечатков) на валидационном наборе данных по сравнению с EI2FP:Full, но при этом работает примерно в 3.5 раза быстрее (Табл. 10).

Несмотря на то, что асамбль моделей DeepEI показывает удовлетворительное качество предсказания (точность, полноту, F_1 метрику), замена нескольких сотен моделей на одну позволяет значительно сократить общее количество параметров и ускорить выполнение. Так суммарный размер моделей DeepEI, рассматриваемых в этой работе, сравнительно небольшой – 3.8 ГиБ, однако размер модели EI2FP:Full составляет всего 133 МиБ с возможностью предсказания 1024 ECFP6 и 166 MACCS.

Кроме того, разделение битов для предсказания позволяет получить большее количество нейронов для предсказания каждого из них, но устраняет возможность перекрестного обучения. Когда поверхностные слои нейросетевой модели преобразуют информацию из масс-спектра в структуру молекулы, они могут использоваться несколькими битами отпечатков пальцев, соответственно, более эффективно обучаться.

Еще одним фактором относительно невысокой скорости работы исходной модели является размера группы (batch size) в 32 образца. Такой небольшой размер группы ограничивает скорость обучения, так как требует большего количества операций перемножения маленьких матриц по сравнению с большим размером группы и меньшим количеством операций перемножения матриц большего размера. Из-за этого даже с использованием ускорения расчетов на видеокарте обучение занимает примерно 10 часов для выбранных 65 ECFP6 и 92 MACCS при всего 8 эпохах для каждой из моделей. Для сравнения, EI2FP:Lite и EI2FP:Full обучаются за 8 и 20 минут соответственно с использованием ускорения на видеокарте (100 эпох). Предсказание на центральном процессоре для DeepEI также выполняется медленно в сравнении с предлагаемыми подходами: для примерно 27 000 молекул предсказание 62 ECFP6 и 92 MACCS заняло 3 и 5 минут, соответственно. Lite и Full версии предлагаемых моделей были в 130 и 34 раза быстрее, соответственно (Табл. 10).

Таблица 10 Сравнение скоростей моделей DeepEI с вариантами моделей EI2FP

Модель	ЦП предсказание, с (≈ 26 000 молекул)	ГП обучение, с (≈ 200 000 молекул)	Увеличение скорости по сравнению с DeepEI
DeepEI	480	36 000 / 8 эпох	1x
EI2FP:Full	13	1200 / 100 эпох	34x
EI2FP:Lite	3	360 / 100 эпох	130x

Предложенные модели превосходят, хоть и незначительно, DeepEI в правильности, точности, полноте предсказания как для ECFP6 и MACCS по отдельности, так и вместе (Табл. 11, Рис. 38).

Таблица 11. 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

Значения правильности достигают наибольших значений по сравнению с другими метриками, поскольку использованная для обучения функция BCELoss наиболее близко соответствует именно этой метрике.

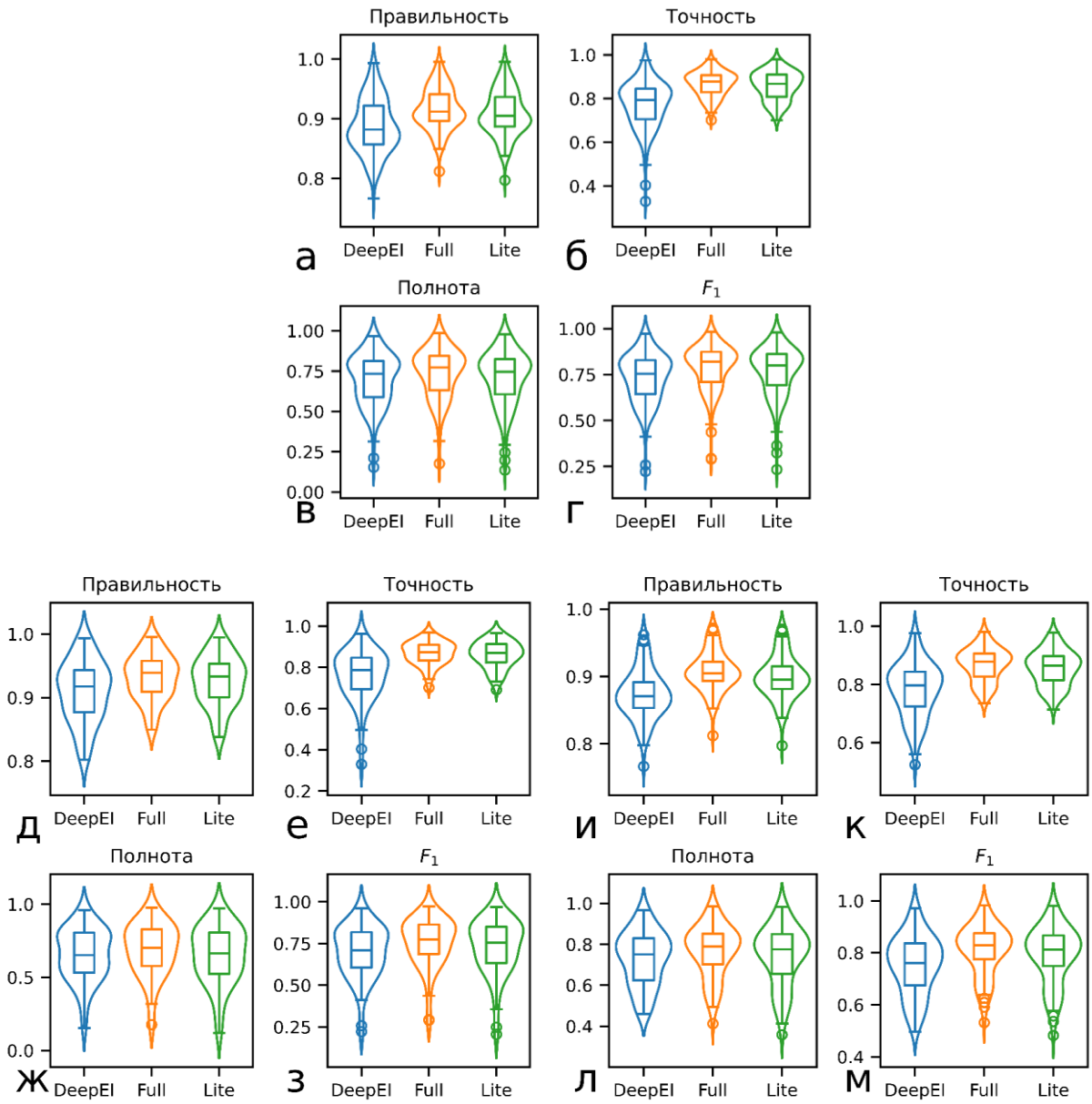


Рисунок 38. Сравнение статистических характеристик предсказательной способности моделей для выбранных значений 95 MACCS + 62 ECFP6 (а – г), 62 ECFP6 (д – з), 95 MACCS (и – м). Правильность предсказания (а, д, и), точность предсказания – (б, е, к), полнота предсказания – (в, ж, л), F_1-метрика – (г, з, м).

Предсказанные значения молекулярных отпечатков пальцев округляли до целочисленных с использованием математического округления. Полученные метрики хорошо воспроизводились для Lite модели при использовании разных вариантов перемешивания набора данных перед разбиением на тренировочную, валидационную и тестовую части (Рис. 39).

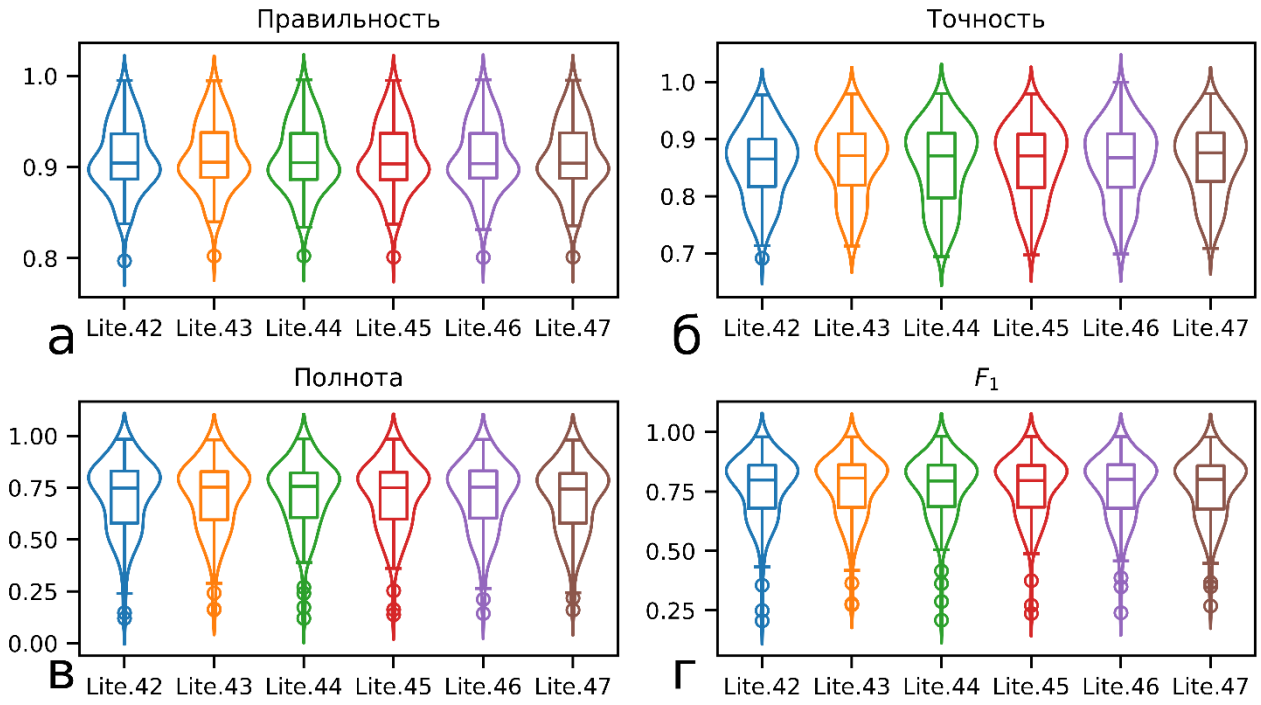
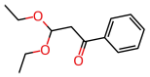
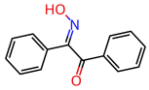
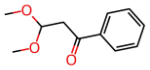
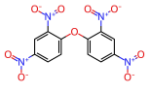
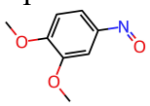
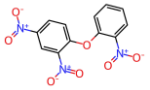


Рисунок 39. Сравнение статистических характеристик предсказания (а – правильность, б – точность, в – полнота, г – F_1 метрика предсказания) молекулярных отпечатков пальцев моделью EI2FP:Lite при разных составах тренировочных и валидационных наборов данных.

Применение молекулярных отпечатков пальцев наиболее оправданно в тех случаях, когда неизвестное соединение отсутствует в доступных специализированных базах данных. Первый шаг для использования молекулярных отпечатков пальцев при идентификации – формирование списка SMILES возможных кандидатов, который можно получить из общехимических баз данных, таких как PubChem или ChemSpider, а также расширить алгоритмически (например, заменив активные протоны на триметилсилильную группу). Молекулярные отпечатки пальцев для всех кандидатов могут быть достаточно просто и быстро рассчитаны с использованием доступных библиотек, таких как RDKit. Затем для неизвестного соединения по масс-спектру электронной ионизации предсказываются молекулярные отпечатки пальцев, например, с использованием предложенной модели EI2FP:Lite, как это сделано в данной работе. После этого производят поиск по сгенерированной базе данных с использованием, например, функции молекулярной близости Танимото или любой другой меры сравнения векторов по выбору.

Таблица 12. Примеры поиска с использованием молекулярных отпечатков пальцев, сравнение с часто применяющимся подходом, основанном на алгоритме 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 

Для примера были выбраны экспериментальные масс-спектры электронной ионизации 3,3-диэтоксипропиофенона и бис(2,4-динитрофенил)ового эфира из базы данных MassBank, которые отсутствуют в NIST MS mainlib (основная библиотека). Эти масс-спектры искали по библиотеке NIST с использованием алгоритма Similarity (рекомендуется для соединений, отсутствующих в базе данных). В обоих случаях найденные кандидаты значительно отличаются от структур «неизвестных» (Табл. 12). Альтернативным вариантом стало предсказание молекулярных отпечатков пальцев моделью EI2FP:Lite, затем поиск по базе сгенерированных молекулярных отпечатков пальцев для всех соединений из NIST MS. Наиболее вероятные кандидаты в этом случае гораздо ближе по структуре к «неизвестным».

Выводы

В этой работе разработали два варианта полносвязных нейросетевых предсказательных моделей: EI2FP:Lite и EI2FP:Full, которые способны предсказывать несколько типов молекулярных отпечатков пальцев одновременно по масс-спектрам электронной ионизации. Обе модели превосходят представленный раньше подход DeepEI по удобству использования и точности предсказания. Также, они на один-два порядка быстрее как при обучении, так и при предсказании. В ходе поиска оптимальной модели были рассмотрены возможности применения различных функций активации,

способов регуляризации, а также сверточной архитектуры для повышения качества предсказания. Показали, что EI2FP:Lite версия модели показывает устойчивые результаты вне зависимости от разбиения набора данных на тренировочный-валидационный-тестовый. Полученные модели могут быть использованы в составе более комплексного пакета для обработки ГХ/МС данных. Предсказанные по масс-спектрам электронной ионизации молекулярные отпечатки пальцев позволяют проводить поиск неизвестных по общехимическим базам данных, таким как PubChem и также могут применяться в алгоритмах ранжирования для сокращения числа возможных кандидатов.

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.

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 литературы

1. *van Den Dool H.*, *Dec. Kratz P.* A generalization of the retention index system including linear temperature programmed gas—liquid partition chromatography // J. Chromatogr. A. 1963. V. 11. P. 463–471.
2. *Castello G.*, *Moretti P.*, *Vezzani S.* Retention models for programmed gas chromatography // J. Chromatogr. A. 2009. V. 1216. № 10. P. 1607–1623.
3. *Kováts E.* Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone // Helv. Chim. Acta. 1958. V. 41. № 7. P. 1915–1932.
4. *Stein S. E.*, *Babushok V. I.*, *Brown R. L.*, *Linstrom P. J.* Estimation of Kováts Retention Indices Using Group Contributions // J. Chem. Inf. Model. 2007. V. 47. № 3. P. 975–980.
5. *Tarján G., Nyiredy Sz., Györ M., et al.* Thirtieth anniversary of the retention index according to Kováts in gas-liquid chromatography // J. Chromatogr. A. 1989. V. 472. P. 1–92.
6. *Matyushin D. D.*, *Sholokhova A. Yu.* Large-scale statistical study of the dependence of retention index on heating rate in temperature-programmed gas chromatography // J. Chromatogr. A. 2024. V. 1732. P. 465223.
7. *Lipkus A. H., Watkins S. P., Gengras K., et al.* Recent Changes in the Scaffold Diversity of Organic Chemistry As Seen in the CAS Registry // J. Org. Chem. 2019. V. 84. № 21. P. 13948–13956.
8. *Kim S., Chen J., Cheng T., et al.* PubChem 2025 update // Nucleic Acids Res. 2025. V. 53. № D1. P. D1516–D1525.
9. *Pence H. E.*, *Williams A.* ChemSpider: An Online Chemical Information Resource // J. Chem. Educ. 2010. V. 87. № 11. P. 1123–1124.
10. *Ayres L. B., Gomez F. J. V., Linton J. R., et al.* Taking the leap between analytical chemistry and artificial intelligence: A tutorial review // Anal. Chim. Acta. 2021. V. 1161. P. 338403.
11. *Joshi P. B.* Navigating with chemometrics and machine learning in chemistry // Artif. Intell. Rev. 2023.
12. *Acimovic M., Pezo L., Jeremic J. S., et al.* QSRR Model for Predicting Retention Indices of Geraniol Chemotype of Thymus serpyllum Essential Oil // J. Essent. Oil Bear. Plants. 2020. V. 23. № 3. P. 464–473.
13. *Aćimović M., Pezo L., Tešević V., et al.* QSRR Model for predicting retention indices of Satureja kitaibelii Wierzb. ex Heuff. essential oil composition // Ind. Crops Prod. 2020. V. 154. P. 112752.
14. *Zhang J., Fang A., Wang B., et al.* iMatch: A retention index tool for analysis of gas chromatography–mass spectrometry data // J. Chromatogr. A. 2011. V. 1218. № 37. P. 6522–6530.
15. *Dossin E., Martin E., Diana P., et al.* Prediction Models of Retention Indices for Increased Confidence in Structural Elucidation during Complex Matrix Analysis: Application to Gas Chromatography Coupled with High-Resolution Mass Spectrometry // Anal. Chem. 2016. V. 88. № 15. P. 7539–7547.
16. *Marrero-Ponce Y.*, *Barigye S. J.*, *Jorge-Rodríguez M. E.*, *Tran-Thi-Thu T.* QSRR prediction of gas chromatography retention indices of essential oil components // Chem. Pap. 2018. V. 72. № 1. P. 57–69.
17. *Kumar P., Kumar A., Lal S., et al.* CORAL: Quantitative Structure Retention Relationship (QSRR) of flavors and fragrances compounds studied on the stationary

phase methyl silicone OV-101 column in gas chromatography using correlation intensity index and consensus modelling // J. Mol. Struct. 2022. V. 1265. P. 133437.
18. *Matyushin D. D.*, *Buryak A. K.* Gas Chromatographic Retention Index Prediction Using Multimodal Machine Learning // IEEE Access. 2020. V. 8. P. 223140–223155.
19. *Qiu F.*, *Lei Z.*, *Sumner L. W.* MetExpert: An expert system to enhance gas chromatography–mass spectrometry-based metabolite identifications // Anal. Chim. Acta. 2018. V. 1037. P. 316–326.
20. *Matyushin D. D.*, *Sholokhova A. Yu.*, *Buryak A. K.* A deep convolutional neural network for the estimation of gas chromatographic retention indices // J. Chromatogr. A. 2019. V. 1607. P. 460395.
21. *Matyushin D. D.*, *Sholokhova A. Y.*, *Buryak A. K.* Deep Learning Based Prediction of Gas Chromatographic Retention Indices for a Wide Variety of Polar and Mid-Polar Liquid Stationary Phases: 17 // Int. J. Mol. Sci. 2021. V. 22. № 17. P. 9194.
22. *Qu C., Schneider B. I., Kearsley A. J., et al.* Predicting Kováts Retention Indices Using Graph Neural Networks // J. Chromatogr. A. 2021. V. 1646. P. 462100.
23. *Vrzal T.*, *Malečková M.*, *Olšovská J.* DeepReI: Deep learning-based gas chromatographic retention index predictor // Anal. Chim. Acta. 2021. V. 1147. P. 64–71.
24. *Anjum A., Liigand J., Milford R., et al.* Accurate prediction of isothermal gas chromatographic Kováts retention indices // J. Chromatogr. A. 2023. V. 1705. P. 464176.
25. *Geer L. Y.*, *Stein S. E.*, *Mallard W. G.*, *Slotta D. J.* AIRI: Predicting Retention Indices and Their Uncertainties Using Artificial Intelligence // J. Chem. Inf. Model. 2024. V. 64. № 3. P. 690–696.
26. *Steinbeck C., Han Y., Kuhn S., et al.* The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics // J. Chem. Inf. Comput. Sci. 2003. V. 43. № 2. P. 493–500.
27. *Willighagen E. L., Mayfield J. W., Alvarsson J., et al.* The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching // J. Cheminformatics. 2017. V. 9. № 1. P. 33.
28. *Hummel J.*, *Selbig J.*, *Walther D.*, *Kopka J.* The Golm Metabolome Database: a database for GC-MS based metabolite profiling // Metabolomics: A Powerful Tool in Systems Biology / ed. Nielsen J., Jewett M. C. 2007. P. 75–95.
29. *Adams R. P.* Identification of essential oil components by gas chromatography/quadrupole mass spectroscopy. 3rd ed. Carol Stream, Ill: Allured Pub. Corporation. 2001. 456 p.
30. *Chen C., Ye W., Zuo Y., et al.* Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals // Chem. Mater. 2019. V. 31. № 9. P. 3564–3572.
31. *Chen B.*, *Barzilay R.*, *Jaakkola T.* Path-Augmented Graph Transformer Network: arXiv:1905.12712. 2019.
32. *Шолохова А. Ю.*, *Матюшин Д. Д.* A comparison of published in 2018-2024 general-purpose models for predicting gas chromatographic retention indices: 5 // Сорбционные И Хроматографические Процессы. 2024. Т. 24. № 5. С. 711–722.
33. *Domingo-Almenara X., Guijas C., Billings E., et al.* The METLIN small molecule dataset for machine learning-based retention time prediction // Nat. Commun. 2019. V. 10. № 1. P. 5811.
34. *Kretschmer F., Harrieder E.-M., Hoffmann M. A., et al.* RepoRT: a comprehensive repository for small molecule retention times // Nat. Methods. 2024. V. 21. № 2. P. 153–155.

35. *Zhang Y., Liu F., Li X. Q., et al.* Retention time dataset for heterogeneous molecules in reversed–phase liquid chromatography // Sci. Data. 2024. V. 11. № 1. P. 946.
36. *Fedorova E. S., Matyushin D. D., Plyushchenko I. V., et al.* Deep learning for retention time prediction in reversed-phase liquid chromatography // J. Chromatogr. A. 2022. V. 1664. P. 462792.
37. *Obradović D., Stavrianidi A., Fedorova E., et al.* A comparative study of the predictive performance of different descriptor calculation tools: Molecular-based elution order modeling and interpretation of retention mechanism for isomeric compounds from METLIN database // J. Chromatogr. A. 2024. V. 1719. P. 464731.
38. *Osipenko S.*, *Nikolaev E.*, *Kostyukevich Y.* Retention Time Prediction with Message-Passing Neural Networks: 10 // Separations. 2022. V. 9. № 10. P. 291.
39. *Vik D., Pii D., Mudaliar C., et al.* Performance and robustness of small molecule retention time prediction with molecular graph neural networks in industrial drug discovery campaigns // Sci. Rep. 2024. V. 14. № 1. P. 8733.
40. *Stienstra C.*, *Nazdrajić E.*, *Hopkins W. S.* From Reverse Phase Chromatography to HILIC: Graph Trans-formers Power Method-Independent Machine Learning of Retention Times. 2024.
41. *Xue J.*, *Wang B.*, *Ji H.*, *Li W.* RT-Transformer: retention time prediction for metabolite annotation to assist in metabolite identification // Bioinformatics. 2024. V. 40. № 3. P. btae084.
42. A Chat with Dr. Andrew Ng on MLOps: From Model-centric to Data-centric AI. 2021.
43. *Whang S. E.*, *Roh Y.*, *Song H.*, *Lee J.-G.* Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective: arXiv:2112.06409. 2022.
44. *Agrawal A., Chatterjee R., Curino C., et al.* Cloudy with high chance of DBMS: A 10-year prediction for Enterprise-Grade ML: arXiv:1909.00084. 2019.
45. *Zha D., Bhat Z. P., Lai K.-H., et al.* Data-centric Artificial Intelligence: A Survey: arXiv:2303.10158. 2023.
46. *Northcutt C.*, *Jiang L.*, *Chuang I.* Confident Learning: Estimating Uncertainty in Dataset Labels // J. Artif. Intell. Res. 2021. V. 70. P. 1373–1411.
47. *Russakovsky O., Deng J., Su H., et al.* ImageNet Large Scale Visual Recognition Challenge // Int. J. Comput. Vis. 2015. V. 115. № 3. P. 211–252.
48. *Côté P.-O., Nikanjam A., Ahmed N., et al.* Data cleaning and machine learning: a systematic literature review // Autom. Softw. Eng. 2024. V. 31. № 2. P. 54.
49. *Lee C., Chen S., Ong K. T., et al.* Noise Analysis and Data Refinement for Chemical Reactions from US Patents via Large Language Models. 2024.
50. *Adrian M.*, *Chung Y.*, *Cheng A. C.* Denoising Drug Discovery Data for Improved Absorption, Distribution, Metabolism, Excretion, and Toxicity Property Prediction // J. Chem. Inf. Model. 2024.
51. NIST Standard Reference Database 1A // NIST. 2014.
52. *Babushok V. I., Linstrom P. J., Reed J. J., et al.* Development of a database of gas chromatographic retention properties of organic compounds // J. Chromatogr. A. 2007. V. 1157. № 1. P. 414–421.
53. *Matyushin D. D.*, *Karnaeva A. E.*, *Sholokhova A. Yu.* Critical evaluation of the NIST retention index database reliability with specific examples // Anal. Bioanal. Chem. 2024. V. 416. № 28. P. 6181–6186.
54. *Sholokhova A. Y.*, *Borovikova S. A.*, *Matyushin D. D.* Refinement of Retention Indices in Gas Chromatography for a Number of Substituted Phenols: 4 // Analytica. 2024. V. 5. № 4. P. 641–653.

55. *Zenkevich I. G., Babushok V. I., Linstrom P. J., et al.* Application of histograms in evaluation of large collections of gas chromatographic retention indices // J. Chromatogr. A. 2009. V. 1216. № 38. P. 6651–6661.
56. *Rew R., Davis G., Emmerson S., et al.* DOI.org (Datacite): Unidata NetCDF . 1989. URL: http://www.unidata.ucar.edu/software/netcdf/ (дата обращения: 17.02.2026).
57. *Martens L., Chambers M., Sturm M., et al.* mzML—a Community Standard for Mass Spectrometry Data * // Mol. Cell. Proteomics. 2011. V. 10. № 1.
58. *Dettmer K.*, *Aronov P. A.*, *Hammock B. D.* Mass spectrometry-based metabolomics // Mass Spectrom. Rev. 2007. V. 26. № 1. P. 51–78.
59. *Chung C. R., Wang H. Y., Lien F., et al.* Incorporating Statistical Test and Machine Intelligence Into Strain Typing of Staphylococcus haemolyticus Based on Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry // Front. Microbiol. 2019. V. 10. № September. P. 1–14.
60. *Govorukhina N. I., Reijmers T. H., Nyangoma S. O., et al.* Analysis of human serum by liquid chromatography–mass spectrometry: Improved sample preparation and data analysis // J. Chromatogr. A. 2006. V. 1120. № 1–2. P. 142–150.
61. *Gaspari M.*, *Verhoeckx K. C. M.*, *Verheij E. R.*, *Van Der Greef J.* Integration of two-dimensional LC-MS with multivariate statistics for comparative analysis of proteomic samples // Anal. Chem. 2006. V. 78. № 7. P. 2286–2296.
62. *Chan L. W., Wang F., Meng F., et al.* Multi-scale representation of proteomic data exhibits distinct microRNA regulatory modules in non-smoking female patients with lung adenocarcinoma // Comput. Biol. Med. 2018. V. 102. № April. P. 51–56.
63. *Dittwald P., Vu T. N., Harris G. A., et al.* Towards automated discrimination of lipids versus peptides from full scan mass spectra // EuPA Open Proteomics. 2014. V. 4. P. 87–100.
64. *Hollemeyer K.*, *Altmeyer W.*, *Heinzle E.* Identification and quantification of feathers, down, and hair of avian and mammalian origin using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry // Anal. Chem. 2002. V. 74. № 23. P. 5960–5968.
65. *Arroyo-Manzanares N., Markiv B., Hernández J. D., et al.* Head-space gas chromatography coupled to mass spectrometry for the assessment of the contamination of mayonnaise by yeasts // Food Chem. 2019. V. 289. № November 2018. P. 461–467.
66. *Allen F.*, *Pon A.*, *Greiner R.*, *Wishart D.* Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification // Anal. Chem. 2016. V. 88. № 15. P. 7689–7697.
67. *Deng F.*, *Wang L.*, *Liu X.* An efficient algorithm for the blocked pattern matching problem // Bioinformatics. 2015. V. 31. № 4. P. 532–538.
68. *Hansen B. T.*, *Jones J. A.*, *Mason D. E.*, *Liebler D. C.* Salsa: A pattern recognition algorithm to detect electrophile-adducted peptides by automated evaluation of CID spectra in LC - MS - MS analyses // Anal. Chem. 2001. V. 73. № 8. P. 1676–1683.
69. *Park C. Y., Klammer A. A., Käli L., et al.* Rapid and accurate peptide identification from tandem mass spectra // J. Proteome Res. 2008. V. 7. № 7. P. 3022–3027.
70. *Vorst O., de Vos C. H. R., Lommen A., et al.* A non-directed approach to the differential analysis of multiple LC-MS-derived metabolic profiles // Metabolomics. 2005. V. 1. № 2. P. 169–180.
71. *Badaloni E., Barbarino M., Cabri W., et al.* Efficient organic monoliths prepared by γ-radiation induced polymerization in the evaluation of histone deacetylase inhibitors by

capillary(nano)-high performance liquid chromatography and ion trap mass spectrometry // J. Chromatogr. A. 2011. V. 1218. № 25. P. 3862–3875.
72. *Lapierre N., Alser M., Eskin E., et al.* Metalign: Efficient alignment-based metagenomic profiling via containment min hash // Genome Biol. 2020. V. 21. № 1. P. 1–15.
73. *Tikunov Y. M., Laptenok S., Hall R. D., et al.* MSClust: A tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data // Metabolomics. 2012. V. 8. № 4. P. 714–718.
74. *Mrzic A., Lermyte F., Vu T. N., et al.* InSourcerer: a high-throughput method to search for unknown metabolite modifications by mass spectrometry // Rapid Commun. Mass Spectrom. 2017. V. 31. № 17. P. 1396–1404.
75. *Milman B. L.*, *Zhurkovich I. K.* Towards a full reference library of MS n spectra. II: A perspective from the library of pesticide spectra extracted from the literature/Internet // Rapid Commun. Mass Spectrom. 2011. V. 25. № 24. P. 3697–3705.
76. *Stein S. E.* NIST/EPA/NIH Mass Spectral Library-PC Version, NIST Standard Reference Database 1A. 1977.
77. Wiley Registry of Mass Spectral Data 2023 [Электронный ресурс] // Wiley Science Solutions. URL: https://sciencesolutions.wiley.com/solutions/technique/gc-ms/wiley-registry-of-mass-spectral-data/ (дата обращения: 23.03.2026).
78. *Wang M., Carver J. J., Phelan V. V., et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking // Nat. Biotechnol. 2016. V. 34. № 8. P. 828–837.
79. *Allen F.*, *Pon A.*, *Greiner R.*, *Wishart D.* Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification // Anal. Chem. 2016. V. 88. № 15. P. 7689–7697.
80. *Wei J. N.*, *Belanger D.*, *Adams R. P.*, *Sculley D.* Rapid Prediction of Electron–Ionization Mass Spectrometry Using Neural Networks // ACS Cent. Sci. 2019. V. 5. № 4. P. 700–708.
81. *Zhu R. L.*, *Jonas E.* Rapid Approximate Subset-Based Spectra Prediction for Electron Ionization–Mass Spectrometry // Anal. Chem. 2023. V. 95. № 5. P. 2653–2663.
82. *Zhang X.-Y.*, *Gong X.-Q.* Comprehensive and Explainable Fragmentation: A Machine Learning Approach for Fast and Accurate Mass Spectrum Prediction // J. Phys. Chem. A. 2025. V. 129. № 15. P. 3552–3559.
83. *Chen L., Xia B., Wang Y., et al.* CMSSP: A Contrastive Mass Spectra-Structure Pretraining Model for Metabolite Identification // Anal. Chem. 2024. V. 96. № 42. P. 16871–16881.
84. *Goldman S.*, *Bradshaw J.*, *Xin J.*, *Coley C.* Prefix-Tree Decoding for Predicting Mass Spectra from Molecules // Adv. Neural Inf. Process. Syst. 2023. V. 36. P. 48548–48572.
85. *Grimme S.* Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules // Angew. Chem. Int. Ed. 2013. V. 52. № 24. P. 6306–6312.
86. *Bauer C. A.*, *Grimme S.* Elucidation of Electron Ionization Induced Fragmentations of Adenine by Semiempirical and Density Functional Molecular Dynamics // J. Phys. Chem. A. 2014. V. 118. № 49. P. 11479–11484.
87. *Gorges J.*, *Grimme S.* QCxMS2 – a program for the calculation of electron ionization mass spectra via automated reaction network discovery // Phys. Chem. Chem. Phys. 2025. V. 27. № 14. P. 6899–6911.
88. *Allen F.*, *Greiner R.*, *Wishart D.* Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification // Metabolomics. 2015. V. 11. № 1. P. 98–110.

89. GitHub: boostorg/boost : C++. 2025. URL: https://github.com/boostorg/boost (дата обращения: 02.10.2025).
90. *Landrum G., Tosco P., Kelley B., et al.* Zenodo: rdkit/rdkit: 2024_03_5 (Q1 2024) Release . 2024. URL: https://zenodo.org/records/12782092 (дата обращения: 24.08.2025).
91. GitHub: lp-solve/lp_solve : Answer Set Programming. 2025. URL: https://github.com/lp-solve/lp_solve (дата обращения: 02.10.2025).
92. *Matyushin D.* GitHub: mtshn/svekla : Java. 2025. URL: https://github.com/mtshn/svekla (дата обращения: 30.09.2025).
93. *Zhang B., Zhang J., Xia Y., et al.* Prediction of electron ionization mass spectra based on graph convolutional networks // Int. J. Mass Spectrom. 2022. V. 475. P. 116817.
94. *Adamczyk J.*, *Ludynia P.* Scikit-fingerprints: Easy and efficient computation of molecular fingerprints in Python // SoftwareX. 2024. V. 28. P. 101944.
95. *Yap C. W.* PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints // J. Comput. Chem. 2011. V. 32. № 7. P. 1466–1474.
96. *Mauri A.*, *Consonni V.*, *Pavan M.*, *Todeschini R.* DRAGON software: An easy approach to molecular descriptor calculations // MATCH Commun. Math. Comput. Chem. 2006. V. 56. P. 237–248.
97. *Mauri A.* alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints // Ecotoxicological QSARs / ed. Roy K. 2020. P. 801–820.
98. *PubChem*. Data Specification [Электронный ресурс]. URL: https://pubchem.ncbi.nlm.nih.gov/docs/data-specification (дата обращения: 16.07.2024).
99. *Klekota J.*, *Roth F. P.* Chemical substructures that enrich for biological activity // Bioinformatics. 2008. V. 24. № 21. P. 2518–2525.
100. *Durant J. L.*, *Leland B. A.*, *Henry D. R.*, *Nourse J. G.* Reoptimization of MDL Keys for Use in Drug Discovery // J. Chem. Inf. Comput. Sci. 2002. V. 42. № 6. P. 1273–1280.
101. *Rogers D.*, *Hahn M.* Extended-Connectivity Fingerprints // J. Chem. Inf. Model. 2010. V. 50. № 5. P. 742–754.
102. *Morgan H. L.* The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. // J. Chem. Doc. 1965. V. 5. № 2. P. 107–113.
103. Daylight Theory: Fingerprints [Электронный ресурс]. URL: https://www.daylight.com/dayhtml/doc/theory/theory.finger.html (дата обращения: 16.07.2024).
104. *Dührkop K., Shen H., Meusel M., et al.* Searching molecular structure databases with tandem mass spectra using CSI:FingerID // Proc. Natl. Acad. Sci. 2015. V. 112. № 41. P. 12580–12585.
105. *Ji H.*, *Deng H.*, *Lu H.*, *Zhang Z.* Predicting a Molecular Fingerprint from an Electron Ionization Mass Spectrum with Deep Neural Networks // Anal. Chem. 2020. V. 92. № 13. P. 8649–8653.
106. *Fan X., Xu Z., Zhang H., et al.* Fully automatic resolution of untargeted GC-MS data with deep learning assistance // Talanta. 2022. V. 244. P. 123415.
107. *de Jonge N. F., Mildau K., Meijer D., et al.* Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools // Metabolomics. 2022. V. 18. № 12. P. 103.

108. *Xue X., Sun H., Yang M., et al.* Advances in the Application of Artificial Intelligence-Based Spectral Data Interpretation: A Perspective // Anal. Chem. 2023. V. 95. № 37. P. 13733–13745.
109. *Seifrid M., Pollice R., Aguilar-Granda A., et al.* Autonomous Chemical Experiments: Challenges and Perspectives on Establishing a Self-Driving Lab // Acc. Chem. Res. 2022. V. 55. № 17. P. 2454–2466.
110. *Baygi S. F.*, *Barupal D. K.* IDSL_MINT: a deep learning framework to predict molecular fingerprints from mass spectra // J. Cheminformatics. 2024. V. 16. № 1. P. 8.
111. *Galati S., Di Stefano M., Martinelli E., et al.* VenomPred: A Machine Learning Based Platform for Molecular Toxicity Predictions: 4 // Int. J. Mol. Sci. 2022. V. 23. № 4. P. 2105.
112. *Liebal U. W., Phan A. N. T., Sudhakar M., et al.* Machine Learning Applications for Mass Spectrometry-Based Metabolomics: 6 // Metabolites. 2020. V. 10. № 6. P. 243.
113. *Hu G.*, *Qiu M.* Machine learning-assisted structure annotation of natural products based on MS and NMR data // Nat. Prod. Rep. 2023.
114. *Herrera-Rocha F., Fernández-Niño M., Duitama J., et al.* FlavorMiner: A Machine Learning Platform for Extracting Molecular Flavor Profiles from Structural Data. 2024.
115. *Stravs M. A.*, *Dührkop K.*, *Böcker S.*, *Zamboni N.* MSNovelist: de novo structure generation from mass spectra // Nat. Methods. 2022. V. 19. № 7. P. 865–870.
116. *Elser D.*, *Huber F.*, *Gaquerel E.* Mass2SMILES: deep learning based fast prediction of structures and functional groups directly from high-resolution MS/MS spectra. 2023. P. 2023.07.06.547963.
117. *Gao S., Chau H. Y. K., Wang K., et al.* Convolutional Neural Network-Based Compound Fingerprint Prediction for Metabolite Annotation: 7 // Metabolites. 2022. V. 12. № 7. P. 605.
118. *Bishop C. M.*, *Bishop H.* Deep Learning: Foundations and Concepts. Cham: Springer International Publishing. 2024.
119. *Hendrycks D.*, *Gimpel K.* Gaussian Error Linear Units (GELUs): arXiv:1606.08415. 2023.
120. *Ramachandran P.*, *Zoph B.*, *Le Q. V.* Searching for Activation Functions: arXiv:1710.05941. 2017.
121. *Aggarwal C. C.* Neural Networks and Deep Learning: A Textbook. Cham: Springer International Publishing. 2023.
122. *Morris C., Ritzert M., Fey M., et al.* Weisfeiler and Leman Go Neural: Higher-order Graph Neural Networks: arXiv:1810.02244. 2021.
123. *Prokhorenkova L., Gusev G., Vorobev A., et al.* CatBoost: unbiased boosting with categorical features: arXiv:1706.09516. 2019.
124. *Chen T.*, *Guestrin C.* XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 785–794.
125. *Moriwaki H.*, *Tian Y.-S.*, *Kawashita N.*, *Takagi T.* Mordred: a molecular descriptor calculator // J. Cheminformatics. 2018. V. 10. № 1. P. 4.
126. *Akiba T., Sano S., Yanase T., et al.* Optuna: A Next-generation Hyperparameter Optimization Framework // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019. P. 2623–2631.
127. *Khrisanfov M. D.*, *Matyushin D. D.*, *Samokhin A. S.* A general procedure for finding potentially erroneous entries in the database of retention indices // Anal. Chim. Acta. 2024. V. 1297. P. 342375.

128. *Khrisanfov M.*, *Matyushin D.*, *Samokhin A.* Finding potentially erroneous entries in METLIN SMRT // J. Chromatogr. A. 2025. V. 1745. P. 465761.
129. NIST 23 Gas Chromatography (GC) Database with Search Software and Retention Indicies (RI) (2023/2020/2017) [Электронный ресурс]. URL: https://www.sisweb.com/software/nist-gc-library.htm (дата обращения: 22.12.2025).
130. *Burns J.* GitHub: JacksonBurns/mordred-community : Python. 2025. URL: https://github.com/JacksonBurns/mordred-community (дата обращения: 24.08.2025).
131. *Anatole von Lilienfeld O.* Quantum chemistry structures and properties of 134 kilo molecules. 2019.
132. *Wu Z., Ramsundar B., N. Feinberg E., et al.* MoleculeNet: a benchmark for molecular machine learning // Chem. Sci. 2018. V. 9. № 2. P. 513–530.
133. *Harris C. R., Millman K. J., van der Walt S. J., et al.* Array programming with NumPy // Nature. 2020. V. 585. № 7825. P. 357–362.
134. *The pandas development team*. Zenodo: pandas-dev/pandas: Pandas . 2025. URL: https://zenodo.org/records/16918803 (дата обращения: 24.08.2025).
135. *Waskom M. L.* seaborn: statistical data visualization // J. Open Source Softw. 2021. V. 6. № 60. P. 3021.
136. *Hunter J. D.* Matplotlib: A 2D Graphics Environment // Comput. Sci. Eng. 2007. V. 9. № 3. P. 90–95.
137. *Paszke A., Gross S., Massa F., et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library // Advances in Neural Information Processing Systems. 2019. V. 32.
138. *Fey M., Sunil J., Nitta A., et al.* PyG 2.0: Scalable Learning on Real World Graphs: arXiv:2507.16991. 2025.
139. *Pedregosa F., Varoquaux G., Gramfort A., et al.* Scikit-learn: Machine Learning in Python // J. Mach. Learn. Res. 2011. V. 12. № 85. P. 2825–2830.
140. *Khrisanfov M.*, *Samokhin A.* mkhrisanfov/Procedure_For_Rounding [Электронный ресурс]. URL: https://github.com/mkhrisanfov/Procedure_For_Rounding (дата обращения: 14.01.2022).
141. *Khrisanfov M.*, *Matyushin D.*, *Samokhin A.* Towards robust databases: an ensemble-based workflow for error detection applied to chemical data. 2025.
142. *Aggarwal C. C.* Neural Networks and Deep Learning: A Textbook. 2nd Edition. Springer International Publishing. 2023.
143. Cetyl stearate [Электронный ресурс]. URL: https://webbook.nist.gov/cgi/inchi/InChI%3D1S/C34H68O2/c1-3-5-7-9-11-13-15-17-19-20-22-24-26-28-30-32-34(35)36-33-31-29-27-25-23-21-18-16-14-12-10-8-6-4-2/h3-33H2%2C1-2H3 (дата обращения: 09.01.2022).
144. *Samokhin A. S.*, *Revelsky I. A.* Intensity of molecular ion peak in electron ionization mass spectra // J. Anal. Chem. 2012 6714. 2012. V. 67. № 14. P. 1066–1068.
145. *Grotch S. L.* Matching of mass spectra when peak height is encoded to one bit // Anal. Chem. 1970. V. 42. № 11. P. 1214–1222.
146. *Knock B. A., Smith I. C., Wright D. E., et al.* Compound Identification by Computer Matching of Low Resolution Mass Spectra // Anal. Chem. 1970. V. 42. № 13. P. 1516–1520.
147. *Hertz H. S.*, *Hites R. A.*, *Biemann K.* Identification of mass spectra by computer-searching a file of known spectra // Anal. Chem. 1971. V. 43. № 6. P. 681–691.
148. *Grotch S. L.* Computer identification of mass spectra using highly compressed spectra codes // Anal. Chem. 1973. V. 45. № 1. P. 2–6.

149. *Rasmussen G. T.*, *Isenhour T. L.* The Evaluation of Mass Spectral Search Algorithms // J. Chem. Inf. Comput. Sci. 1979. V. 19. № 3. P. 179–186.
150. *Heller S.* The history of the NIST/EPA/NIH mass spectral database // Today's Chem. Work. 1999. V. 8. № 2. P. 45–46.
151. *McLafferty F. W.*, *Hertel R. H.*, *Villwock R. D.* Probability based matching of mass spectra. Rapid identification of specific compounds in mixtures // Org. Mass Spectrom. 1974. V. 9. № 7. P. 690–702.
152. *McLafferty F. W.*, *Zhang M. Y.*, *Stauffer D. B.*, *Loh S. Y.* Comparison of algorithms and databases for matching unknown mass spectra // J. Am. Soc. Mass Spectrom. 1998. V. 9. № 1. P. 92–95.
153. *Stein S. E.*, *Scott D. R.* Optimization and testing of mass spectral library search algorithms for compound identification // J. Am. Soc. Mass Spectrom. 1994. V. 5. № 9. P. 859–866.
154. *Samokhin A.*, *Sotnezova K.*, *Lashin V.*, *Revelsky I.* Evaluation of mass spectral library search algorithms implemented in commercial software // J. Mass Spectrom. 2015. V. 50. № 6. P. 820–825.
155. *Hájek A.*, *Hecht H.*, *Price E. J.*, *Křenek A.* SpecTUS: Spectral Translator for Unknown Structures annotation from EI-MS spectra: arXiv:2502.05114. 2025.
156. *Khrisanfov M.* GitHub: mkhrisanfov/neims-pytorch . 2025. URL: https://github.com/mkhrisanfov/neims-pytorch (дата обращения: 01.10.2025).
157. mkhrisanfov/neims-pytorch · Hugging Face [Электронный ресурс]. URL: https://huggingface.co/mkhrisanfov/neims-pytorch (дата обращения: 19.11.2025).
158. mkhrisanfov/neims-pytorch general | Docker Hub [Электронный ресурс]. URL: https://hub.docker.com/repository/docker/mkhrisanfov/neims-pytorch/general (дата обращения: 19.11.2025).