

АВТОНОМНАЯ НЕКОММЕРЧЕСКАЯ ОБРАЗОВАТЕЛЬНАЯ ОРГАНИЗАЦИЯ
ВЫСШЕГО ОБРАЗОВАНИЯ «СКОЛКОВСКИЙ ИНСТИТУТ НАУКИ И
ТЕХНОЛОГИЙ»

На правах рукописи



ОСИПЕНКО СЕРГЕЙ ВЛАДИМИРОВИЧ

**ПРОГНОЗИРОВАНИЕ ХРОМАТО-МАСС-СПЕКТРОМЕТРИЧЕСКИХ
ХАРАКТЕРИСТИК ХИМИЧЕСКИХ СОЕДИНЕНИЙ В НЕЦЕЛЕВОМ
АНАЛИЗЕ С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ**

Специальность – 1.4.2 Аналитическая химия

ДИССЕРТАЦИЯ

на соискание учёной степени

кандидата химических наук

Научный руководитель

д.х.н. Костюкевич Ю.И.

Москва – 2024

Оглавление

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ	5
ВВЕДЕНИЕ.....	6
ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ	13
1.1 Обзор библиотек, содержащих хромато-масс-спектрометрические характеристики низкомолекулярных соединений	13
1.2 Краткая характеристика основных методов машинного обучения	17
1.3 Способы оценки моделей машинного обучения	24
1.3.1 Компромисс «отклонение - дисперсия».....	24
1.3.2 Основные метрики, применяемые в машинном обучении	25
1.3.3 Способы валидации моделей машинного обучения.....	27
1.4 Способы представления молекул в машинном обучении	28
1.5 Примеры применения методов машинного обучения для предсказания аналитических характеристик низкомолекулярных соединений.....	32
1.5.1 Применение методов машинного обучения для предсказания характеристик удерживания низкомолекулярных соединений в газовой хромато-масс-спектрометрии	32
1.5.2 Применение методов машинного обучения для предсказания характеристик удерживания низкомолекулярных соединений в жидкостной хромато-масс-спектрометрии	36
1.5.3 Применение методов машинного обучения для предсказания масс-спектральных характеристик	40
1.5.4 Применение методов машинного обучения в спектрометрии ионной подвижности	42
1.6 Идентификации химических соединений в нецелевом хромато-масс-спектрометрическом анализе с применением характеристик, предсказанных с помощью машинного обучения	44
ГЛАВА 2. Оборудование, материалы, техника эксперимента.....	47
2.1 Оборудование и материалы	47
2.2 Выполнение анализа.....	48

2.2.1	Определение времен удерживания для получения обучающей и тестовой выборки в условиях разделения в нано-поточной хроматографии	48
2.2.2	Определение времен удерживания для получения внутрилабораторной обучающей выборки	49
2.2.3	Пробоподготовка образцов мочи для изучения селективности изотопного обмена $^{16}\text{O}/^{18}\text{O}$	49
2.3	Программное обеспечение.....	50
ГЛАВА 3. Применение машинного обучения для предсказания времен удерживания в жидкостной хромато-масс-спектрометрии.....		51
3.1	Предсказание времен удерживания в жидкостной хроматографии методом градиентного бустинга	51
3.1.1	Построение модели предсказания времен удерживания по данным библиотеки METLIN SMRT	51
3.1.2	Пересчет предсказаний на другие хроматографические условия	55
3.1.3	Фильтрация ложноположительных определений при идентификации химических соединений в нецелевом скрининге с помощью предложенного подхода	58
3.2	Предсказание времен удерживания в жидкостной хроматографии с использованием текстовых представлений молекул, глубоких нейронных сетей и обучения с переносом	60
3.2.1	Описание предложенного подхода.....	61
3.2.2	Результаты моделирования времен удерживания при использовании обучения с переносом.....	63
3.3	Предсказание времен удерживания с применением нейронных сетей с механизмом передачи сообщений (Message-Passing Neural Networks)	68
3.3.1	Описание предложенного подхода.....	69
3.3.2	Результаты предсказаний времен удерживания с помощью нейронных сетей с распространением сообщений	70
3.4	Сравнение предложенных подходов	75
ГЛАВА 4. Совместное применение методов предсказания времен удерживания и метода изотопного обмена для идентификации химических соединений в нецелевом скрининге		78
4.1	Определение селективности изотопного обмена изотопов кислорода $^{16}\text{O}/^{18}\text{O}$	80

4.2	Фильтрация ложноположительных определений с помощью изотопного обмена $^{16}\text{O}/^{18}\text{O}$	99
4.3	Предсказание времен удерживания	104
ГЛАВА 5. Предсказание индексов удерживания веществ, относящихся к спискам Конвенции по запрещению химического оружия.....		
108		
5.1	Оценка применимости методов глубокого обучения для предсказания индексов удерживания соединений из списков Конвенции по запрещению химического оружия	109
5.2	Повышение точности предсказания индексов удерживания за счет применения более специфичной модели, основанной на алгоритме градиентного бустинга	112
5.3	Инкрементный подход к моделированию индексов удерживания соединений из списков Конвенции по запрещению химического оружия	112
ГЛАВА 6. Предсказание масс-спектров электронной ионизации с помощью машинного обучения		
117		
6.1	Описание подхода к предсказанию масс-спектров электронной ионизации с применением машинного обучения	118
6.2	Предсказание спектра нейтральных потерь и усредненная модель	120
6.3	Применение разработанной модели для создания in-silica спектральных библиотек..	124
6.4	Сравнение предложенного подхода с квантово-химическими расчетами масс-спектров электронной ионизации.....	127
ЗАКЛЮЧЕНИЕ		130
ВЫВОДЫ		132
СПИСОК ЛИТЕРАТУРЫ.....		134

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

ГХ-МС	– газовая хромато-масс-спектрометрия
ЖХ-МС	– жидкостная хромато-масс-спектрометрия
ИУ	– индекс удерживания
ДИС	– Диссоциация, индуцируемая соударениями
СИП	– спектрометрия ионной подвижности
МО	– машинное обучение
SMRT	– Набор данных удерживания малых молекул (small molecule retention time dataset)
HMDB	– Human metabolome database
ГБ	– градиентный бустинг
ИНС	– искусственные нейронные сети
OCAD	– центральная аналитическая база данных Организации по запрещению химического оружия
CCS	– сечение столкновений
ROC	– рабочая характеристика приемника (receiver operation characteristic)
SMILES	– система упрощённого представления молекул в строке ввода (Simplified Molecular Input Line Entry System)
InChI	– международный текстовый химический идентификатор (International Chemical Identifier)
SMARTS	– SMILES arbitrary target specification
NEIMS	– Neural Electron–Ionization Mass Spectrometry
GBEIMS	– Gradient boosting based electron ionization mass spectra prediction
QCEIMS	– Quantum Chemistry Electron Ionization Mass Spectra
ПО	– Программное обеспечение

ВВЕДЕНИЕ

Актуальность темы. Газовая хромато-масс-спектрометрия (ГХ-МС) и жидкостная хромато-масс-спектрометрия (ЖХ-МС) являются наиболее информативными методами нецелевого анализа многокомпонентных природных и биологических образцов на содержание малых молекул (молекулярная масса которых не превышает 1500 Да). Одной из задач нецелевого хромато-масс-спектрометрического анализа является установление качественного состава многокомпонентных образцов, которая сводится к идентификации всех детектированных в образце химических соединений. Основной подход к решению данной задачи заключается в сопоставлении определенных в хромато-масс-спектрометрическом эксперименте параметров соединения (времени или индекса удерживания (ИУ), молекулярной массы, масс ионов, образующихся при диссоциации, индуцируемой соударениями (ДИС), а также их относительных интенсивностей) со справочными значениями возможных кандидатов, полученными в специализированных базах данных или измеренными с применением образцов сравнения известного состава. Необходимо отметить, что степень точности измерения массы с помощью масс-спектрометров высокого разрешения с времяпролетными масс-анализаторами или оснащенных орбитальной ионной ловушкой позволяет (с некоторыми допущениями) определение элементного состава ионов по точной измеренной массе. Это во многих случаях позволяет свести задачу идентификации к поиску по изомерным структурам, имеющим одинаковую брутто-формулу.

Основными ограничениями данного подхода являются низкая степень покрытия масс-спектральными базами и базами хроматографического удерживания химического разнообразия малых молекул, ограниченная доступность образцов сравнения, а также плохая воспроизводимость измеряемых параметров в различных условиях проведения эксперимента. Как результат, сигналу одного компонента образца может соответствовать несколько десятков или сотен изомерных молекул, и для однозначной идентификации потребуются встречный синтез всех возможных кандидатов.

Для сокращения пространства поиска и сужения списка кандидатов предлагаются различные подходы, как экспериментальные, так и вычислительные. Первые нацелены на разработку новых измеряемых параметров, характеристичных для определенных молекул, и которые могут быть измерены за счет модификации хромато-масс-спектрометрического оборудования (например, сечение столкновений (ССС) в спектрометрии ионной подвижности (СИП)), а также за счет селективной дериватизации компонентов образца (количество определенных функциональных групп). Последние позволяют оценивать значения измеряемых

параметров по структуре для наполнения баз хроматографического удерживания или масс-спектральных библиотек.

Среди экспериментальных методов необходимо отметить особое положение методов изотопного обмена, в первую очередь изотопов кислорода $^{16}\text{O}/^{18}\text{O}$, а также дейтериеводородного обмена. С одной стороны, их можно расценивать как разновидность химической дериватизации, позволяющей определять количество определенных функциональных групп по изменению измеряемой молекулярной массы, с другой, они имеют преимущество в сохранении других аналитических свойств молекул, в первую очередь, хроматографического удерживания, что существенно упрощает последующую интерпретацию данных. Несмотря на долгую историю изучения изотопного обмена и широкий набор вариаций его применения в сочетании с масс-спектрометрией, методология его применения в нецелевом хромото-масс-спектрометрическом анализе требует развития для уточнения его селективности, выбора условий проведения реакций при установлении качественного состава многокомпонентных образцов.

Среди параметров, предсказываемых вычислительными методами особое внимание уделяется характеристикам хроматографического удерживания, так как они дают дополнительную информацию для идентификации только при наличии справочных значений, в отличие от спектров фрагментации, которые могут быть интерпретированы непосредственно, для определения фрагментов определяемой структуры. Тем не менее, задача моделирования масс-спектров также является одной из ключевых для нецелевого анализа, ввиду того, что сопоставление экспериментальных масс-спектров со справочными значениями вносит определяющий вклад в идентификацию компонентов. Ограниченность библиотек, содержащих масс-спектры, сужает круг потенциально идентифицируемых веществ и снижает эффективность и достоверность идентификации.

Для моделирования хромото-масс-спектрометрических характеристик низкомолекулярных соединений применяются различные методы вычислительной химии, включая квантово-химические расчеты, методы молекулярной динамики. Однако наиболее перспективными представляются методы, основанные на алгоритмах машинного обучения (МО), которые хорошо зарекомендовали себя в смежных задачах предсказания различных молекулярных свойств, не требуют построения теоретических физико-химических моделей и существенно превосходят другие методы в производительности. Точность этих методов в основном ограничена доступным объемом обучающих выборок и эффективностью конкретных алгоритмов. Развитие методов машинного и глубокого обучения в совокупности с пополнением экспериментальных баз данных может существенно увеличить точность прогнозирования характеристик молекул, применяемых для определения состава многокомпонентных образцов.

Цель работы заключалась в разработке подходов к моделированию хромато-масс-спектрометрических характеристик молекул, применяемых при идентификации химических соединений в нецелевом анализе, методами МО.

Для достижения поставленной цели необходимо было решить следующие задачи:

- разработать основанные на методах МО подходы к предсказанию времен удерживания в жидкостной хроматографии, позволяющие моделировать удерживание для различных экспериментальных условий разделения;
- сравнить эффективность различных методов МО для моделирования времен удерживания;
- оценить эффективность фильтрации ложноположительных результатов при идентификации химических соединений в нецелевых исследованиях по предсказанным временам удерживания;
- оценить эффективность фильтрации ложноположительных результатов при идентификации химических соединений в нецелевых исследованиях при совместном применении методов предсказания времен удерживания и экспериментального метода изотопного обмена в сочетании с масс-спектрометрией;
- разработать подход к предсказанию ИУ для их использования для идентификации химических соединений при нецелевом анализе методом ГХ-МС;
- разработать подход к предсказанию масс-спектров электронной ионизации с применением методов МО для создания расчетных библиотек масс-спектров;
- оценить эффективность идентификации химических соединений при использовании расчетных библиотек.

Научная новизна

1. Для прогнозирования времен удерживания низкомолекулярных соединений в жидкостной хроматографии построены модели машинного обучения, основанные на алгоритмах градиентного бустинга, искусственных нейронных сетей с архитектурой Трансформер и графовых нейронных сетей с распространением сообщений, с использованием набора данных по удерживанию более 80 000 соединений в условиях обращенно-фазового разделения. Точность прогнозов характеризуется средним отклонением 32 с при общем времени разделения 23 мин.

2. Предложены новые способы оценки времен удерживания для различных экспериментальных систем в условиях ограниченных объемов доступной обучающей выборки с использованием разработанных моделей и метода обучения с переносом.
3. Установлены функциональные группы, которые способны вступать в реакцию изотопного обмена $^{16}\text{O}/^{18}\text{O}$; разработан подход к применению изотопного обмена $^{16}\text{O}/^{18}\text{O}$ в сочетании с хромато-масс-спектрометрией высокого разрешения для определения состава многокомпонентных образцов.
4. Для прогнозирования масс-спектров электронной ионизации использован алгоритм градиентного бустинга и разработано соответствующее программное обеспечение GBEIMS, которое превосходит по точности предсказаний известный метод прогнозирования масс-спектров электронной ионизации QCEIMS, основанный на квантово-химических расчетах.

Практическая значимость.

1. Предложены подходы, позволяющие предсказывать времена и индексы удерживания соединений, для которых получение экспериментальных значений затруднительно, ввиду отсутствия образцов сравнения известного состава. Продемонстрирована возможность фильтрации более 50% ложноположительных результатов по предсказанным временам удерживания при идентификации химических соединений в нецелевых исследованиях.
2. Разработан подход к применению метода изотопного обмена изотопов кислорода $^{16}\text{O}/^{18}\text{O}$ для анализа биологических образцов, включающий программные алгоритмы для использования экспериментальных данных при идентификации химических соединений и фильтрации изомерных структур. Продемонстрирована возможность фильтрации 75% ложноположительных результатов одновременно по предсказанным временам удерживания и данным, полученным с помощью изотопного обмена $^{16}\text{O}/^{18}\text{O}$ при идентификации лекарственных средств в модельном образце мочи человека.
3. Предложенные способы предсказания индексов удерживания в газовой хромато-масс-спектрометрии позволяют оценить значения индексов удерживания соединений, относящихся к Конвенции по запрещению химического оружия. Инкрементный подход с автоматическим поиском пар гомологов характеризуется средним отклонением до 5 ед для соединений, относящихся к гомологическим рядам. Способ предсказания на основе машинного обучения характеризуется средним отклонением в 16 единиц в режиме кросс-валидации с использованием данных библиотеки OCAD и

может быть применен для структурных аналогов соединений, входящих в эту библиотеку.

4. Предложенные подходы реализованы в виде программного обеспечения на языке Python с открытым исходным кодом или Web-приложений с графическим интерфейсом и могут быть использованы непосредственно или адаптированы под решение конкретных задач химического анализа.

Положения, выносимые на защиту.

1. Применение градиентного бустинга, искусственных нейронных сетей с архитектурой Трансформер, графовых искусственных нейронных сетей с распространением сообщений и обучающей выборки METLIN Small molecule retention dataset позволяет предсказывать времена удерживания низкомолекулярных соединений со средним отклонением 45.6, 57.0 и 31.5 с соответственно, что сопоставимо с прецизионностью измерений времен удерживания из обучающей выборки.
2. Применение кусочно-линейных функций пересчета или метода обучения с переносом позволяет использовать разработанные модели машинного обучения для предсказания времен удерживания в различных условиях хроматографического разделения.
3. Фильтрация ложноположительных определений по временам удерживания, полученным с использованием разработанных моделей, позволяет сократить пространство поиска среди изомерных структур, содержащихся в общехимических базах данных в среднем на 23-53%, в зависимости от условий разделения.
4. Изотопный обмен $^{16}\text{O}/^{18}\text{O}$ в сочетании с масс-спектрометрией высокого разрешения может быть использован для функционального анализа при нецелевом скрининге биологических образцов. Сопоставление определенного в эксперименте числа обменов с максимально возможным, рассчитанным по структуре, позволяет фильтровать ложноположительные определения, сокращая пространство поиска на 62%, совместное применение с фильтрацией по предсказанным временам удерживания увеличивает эффективность подхода до 75%.
5. Существующие универсальные модели машинного обучения для предсказания индексов удерживания позволяют предсказывать индексы удерживания соединений, относящихся к спискам Конвенции по запрещению химического оружия со средним отклонением 39.9-51.5 единиц. При применении специфичной модели градиентного бустинга, предложенной в работе, среднее отклонение составляет 16 единиц; при

применении инкрементного метода, предложенного в работе, среднее отклонение снижается до 4 единиц.

6. Предложенный в работе подход GBEIMS для моделирования масс-спектров электронной ионизации с помощью градиентного бустинга характеризуется высоким сходством предсказанных и экспериментально измеренных масс-спектров.

Степень достоверности.

Степень достоверности результатов проведенных исследований обеспечивалась применением современного хроматографического и масс-спектрометрического оборудования, реагентов высокой чистоты, современных методик проведения анализа и средств обработки результатов экспериментов.

Соответствие паспорту научной специальности.

Диссертационная работа соответствует паспорту специальности 1.4.2 Аналитическая химия по областям исследований:

- методы химического анализа (химические, физико-химические, атомная и молекулярная спектроскопия, хроматография, рентгеновская спектроскопия, масс-спектрометрия, ядерно-физические методы и др.);

- математическое обеспечение химического анализа;

Апробация результатов исследования.

Основные результаты, изложенные в работе, были представлены на следующих конференциях:

2023 г: IX Всероссийская конференция с международным участием «Масс-спектрометрия и ее прикладные проблемы», Москва, Россия, 30 октября – 03 ноября 2023 г; Международная конференция «Second Moscow International Conference on Multi-omics Technologies for Precision Medicine», Москва, Россия, 20-21 ноября 2023 г.

2022 г: Научно-практическая конференция «Медико-биологические аспекты обеспечения химической безопасности Российской Федерации», посвященная 60-летию федерального государственного унитарного предприятия «Научно-исследовательский институт гигиены, профпатологии и экологии человека» Федерального медико-биологического агентства, Санкт-Петербург, Россия, 27-28 апреля, 2022 г; Международная конференция «24th International Mass Spectrometry Conference», Маастрихт, Нидерланды, 27 августа – 2 сентября 2022 г;

2021 г: IX Всероссийская конференция с международным участием «Масс-спектрометрия и ее прикладные проблемы», Москва, Россия, 18-22 октября 2021 г.;

2020 г: Международная конференция 68th ASMS Conference on Mass Spectrometry and Allied Topics, онлайн, 1-12 июня, 2020 г.

Публикации.

По материалам работы опубликовано 60 печатных работ, в том числе 6 статей в рецензируемых научных изданиях, индексируемых международными базами данных (Web of Science, Scopus) и рекомендованных в диссертационном совете МГУ по специальности 1.4.2 Аналитическая химия.

Личный вклад автора.

Личный вклад автора заключался в формулировании цели исследования, постановке задач, систематизации литературных данных, планировании и проведении всех экспериментальных этапов исследования, обработке и интерпретации полученных результатов, разработке программного обеспечения, представлении полученных результатов на конференциях и подготовке материалов к публикации. Во всех опубликованных работах вклад автора является определяющим. Все исследования, представленные в работе, проводились автором лично или в сотрудничестве с коллегами.

Структура и объем работы.

Диссертационная работа состоит из введения, 6 глав, заключения, выводов, списка используемых сокращений и списка цитируемой литературы из 215 наименований. Полный объем диссертации составляет 163 страницы, включая 57 рисунков и 27 таблиц и одно приложение.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

1.1 Обзор библиотек, содержащих хромато-масс-спектрометрические характеристики низкомолекулярных соединений

Хромато-масс-спектрометрия является основным методом установления состава многокомпонентных образцов в нецелевых исследованиях ввиду непревзойденных аналитических характеристик метода. Параметры веществ, измеряемые в ходе хромато-масс-спектрометрического эксперимента, а именно время или индекс удерживания, измеренная молекулярная масса, спектр ДИС, могут использоваться для идентификации химических соединений, обнаруженных в образце[1]. Стандартный подход к идентификации химических соединений заключается в сравнении измеренных параметров со справочными значениями, которые могут быть получены из спектральных библиотек и библиотек хроматографического удерживания, или установлены по образцам сравнения известного состава[1-5].

К наиболее известным спектральным библиотекам можно отнести библиотеки, реализуемые Национальным институтом стандартов и технологий США (NIST) и компанией Wiley, масс-спектральную базу данных mzCloud[6], масс-спектральную библиотеку METLIN[7-9], а также публичные репозитории MassBank[10], MassBank of North America[11] и GNPS[12]. Характеристики удерживания в газовой хроматографии можно найти в библиотеке NIST Retention Index Library[13], а также в публичных репозиториях[10, 11, 14-17]. Данные по временам хроматографического удерживания аккумулируются в различных репозиториях[18, 19]. Нужно отметить, что времена удерживания сильно зависят от условий проведения эксперимента, описания которых в подобных репозиториях зачастую недостаточны для воспроизведения результатов.

В таблице 1 собрана информация о некоторых библиотеках масс-спектров электронной ионизации. Наиболее обширными масс-спектральными библиотеками являются Wiley Registry 12th Edition[20], содержащей масс-спектры электронной ионизации 668 000 соединений и NIST 20 Mass Spectral Library[13] содержащей масс-спектры более 300 000 соединений. Данные библиотеки являются коммерческими; объем библиотек, находящихся в открытом доступе существенно ниже. Тем не менее, даже коммерческие библиотеки покрывают менее 1% известных низкомолекулярных соединений, представленных в общехимических базах данных. Так, в библиотеках PubChem[21] и ChemSpider[22] содержатся сведения более чем о 100 миллионах различных соединений[21, 23]. Нельзя не отметить тот факт, что масс-спектры многих соединений не могут быть измерены ГХ-МС ввиду низкой летучести или термической нестабильности и отсутствуют в библиотеках. На рисунке 1 отражено пересечение библиотеки NIST 20 Mass Spectral Library и базы данных «Метаболом человека» (Human Metabolome

Database, (**HMDB**))[24, 25]. Можно видеть, что лишь незначительная часть метаболитов, представленных в библиотеке HMDB характеризуется доступными спектрами электронной ионизации (по крайней мере, в нативной форме).

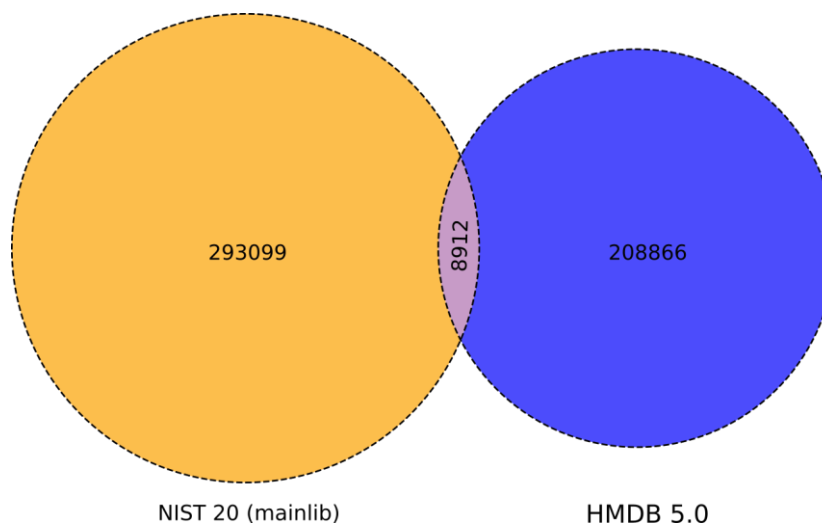


Рисунок 1. Пересечение библиотек NIST 20 Mass Spectral Library и Human Metabolome Database 5.0 (HMDB).

Таблица 1. Библиотеки, содержащие масс-спектры электронной ионизации

Библиотека	Количество масс-спектров	Количество уникальных соединений	Статус
Wiley Registry 12 th Edition[20]	817000	668000	Коммерческая
NIST 20 Mass Spectral Library[13]	350643	306869	Коммерческая
GOLM Metabolome Database[14]	1178	1157	Открытая
FiehnLib[26]	2112	>1000	Открытая
MassBank of North America[11]	18902	9762	Открытая
MassBank[27]	11810	11810	Открытая

Одним из преимуществ ГХ-МС с электронной ионизацией является информативность масс-спектров, обусловленная наличием фрагментных ионов, а также хорошая воспроизводимость масс-спектров при стандартизации энергии электронного пучка. Это позволяет использовать при идентификации не только положения ионов в масс-спектрах, но и относительные интенсивности их сигналов. При использовании «мягких» методов ионизации, в первую очередь ионизации электрораспылением, первичные масс-спектры мало информативны. Позволяя определять массу протонированных или депротонированных молекул, «мягкие»

методы ионизации обычно сочетаются с дополнительными методами диссоциации, для обеспечения возможности проведения структурного анализа. При анализе низкомолекулярных соединений наиболее распространена ДИС. В отличие от диссоциации молекулы при электронной ионизации в условиях достаточно высокой энергии электронов (70эВ), на масс-спектры ДИС влияет множество условий, в частности энергия соударений, конструкция ячейки соударений, газ, используемый для соударений. В результате, масс-спектры одного и того же вещества, полученные на различных приборах, могут отличаться не только соотношением сигналов ионов, образующихся при диссоциации, но и их качественным составом. Поэтому, при создании библиотек спектров ДИС, дополнительные усилия прикладываются для измерения спектров на масс-спектрометрах различных типов и разных производителей. Как результат, количество уникальных молекул в таких библиотеках существенно ниже, чем в библиотеках масс-спектров электронной ионизации, при сопоставимом общем количестве спектров (Таблица 2).

Таблица 2. Библиотеки вторичных масс-спектров диссоциации, индуцируемой соударениями

Библиотека	Количество масс-спектров	Количество уникальных соединений	Статус
NIST 20 Mass Spectral Library[13]	~1 300 000	~31 000	Коммерческая
mzCloud[6]	10080578	12083	Частично открытая
METLIN Gen2[28]	Нет данных	~860000	Коммерческая
METLIN [9]	Нет данных	14300	Открытая
MoNA [11]	145381	17174	Открытая
MassBank [10]	90471	15078	Открытая

Аналогичная ситуация складывается и в области создания библиотек хроматографического удерживания. Введение ИУ для нормализации времен хроматографического удерживания в газовой хроматографии позволило избежать зависимости от геометрии колонок и режима элюирования. С учетом ограниченного набора неподвижных фаз, традиционно применяемых в газовой хроматографии это позволило накопить обширную экспериментальную базу ИУ[13]. В то же время разнообразие подвижных и неподвижных фаз, применяемых в жидкостной хроматографии для анализа низкомолекулярных соединений ограничивает целесообразность создания подобных библиотек удерживания. Как результат, количество уникальных молекул в наборах данных, содержащих времена удерживания обычно не превышает 1000, что сопоставимо с размером коллекций образцов сравнения в среднестатистической аналитической лаборатории. Единственным известным исключением

является библиотека METLIN Small molecule retention dataset (SMRT)[29], насчитывающая времена удерживания более 80 000 молекул, измеренных в одних условиях разделения.

Пересечение библиотек удерживания с основными профильными библиотеками низкомолекулярных соединений также невелико, как и в случае с масс-спектральной информацией. На рисунке 2 показано количество молекул из баз данных DrugBank[30] и HMDB 4.0[24] для которых в библиотеке METLIN SMRT есть информация об удерживании. Можно сделать вывод, что применение имеющихся данных об удерживании весьма ограничено в нецелевых метаболомных исследованиях и скрининге лекарственных препаратов.

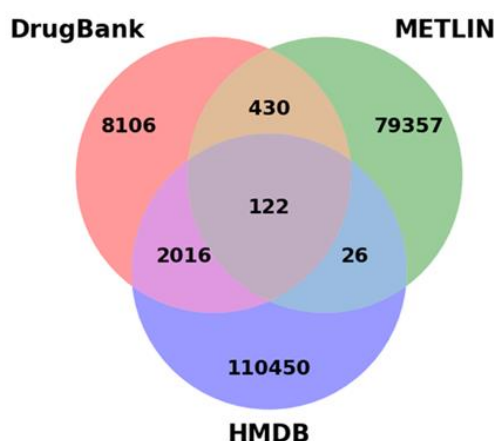


Рисунок 2. Пересечение библиотек HMDB 4.0 [24], DrugBank[30] и METLIN SMRT[29].

Пополнение библиотек новыми экспериментальными данными требует огромного объема ресурсов. Например, на создание библиотеки удерживания METLIN SMRT ушло более 5 лет, а динамика пополнения базы данных NIST (Рисунок 3) свидетельствует об ограниченных возможностях, даже в условиях работы в крупных государственных институтах. Хотя при пополнении библиотек авторы руководствуются распространенностью химических соединений, и стремятся включать вещества наиболее значимых классов, отсутствие соединения в библиотеке не позволит корректно аннотировать его сигнал при нецелевом анализе. Поэтому задача пополнения масс-спектральных библиотек и библиотек удерживания является одной из ключевых для качественного хромато-масс-спектрометрического анализа, и может быть частично решена с применением вычислительных методов.

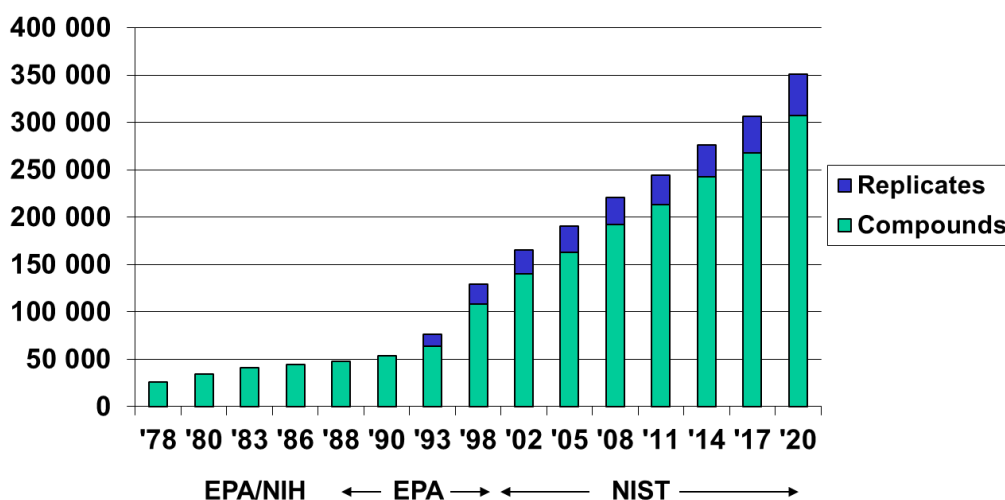


Рисунок 3. Динамика пополнения масс-спектральной библиотеки NIST EI Library[13].

1.2 Краткая характеристика основных методов машинного обучения

Машинное обучение включает методы искусственного интеллекта, позволяющие строить статистические модели по известным примерам, составляющим обучающую выборку, и использовать найденные закономерности для обработки новых входных данных. Методы МО подразделяются на методы обучения с учителем, требующие предварительно размеченные входные данные, и методы обучения без учителя, способные обрабатывать неразмеченные данные. Разметка данных заключается в предварительном разбиении обучающей выборки на известные классы, или определении значений моделируемой регрессионной характеристики для объектов обучающей выборки. Методы МО без учителя находят применение в задачах кластеризации данных, сокращения размерности, обобщения и выявления аномалий, но малоприменимы для предсказания задач предсказания дискретных и непрерывных переменных. Далее будут рассмотрены только методы обучения с учителем.

Описано множество алгоритмов МО с учителем для решения тех или иных задач. Однако, все эти алгоритмы сводятся к восстановлению неявной зависимости между векторами независимых переменных и вектором зависимой переменной по известному набору данных. Подбор параметров модели проводится путем поиска локального экстремума функции потерь, которая характеризует отклонение предсказанных моделью результатов от истинных значений. Методы МО с учителем можно классифицировать в соответствии с типом решаемой задачи, на методы классификации и регрессионные методы. К наиболее широко применяемым методам МО можно отнести линейную регрессию, метод опорных векторов, метод случайного леса, метод градиентного бустинга (ГБ), искусственные нейронные сети (ИНС).

Множественная линейная регрессия является простейшим алгоритмом МО, по нахождению зависимости между зависимой переменной Y и вектором независимых переменных X в виде линейной функции:

$$Y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n.$$

Подбор коэффициентов в уравнении регрессии осуществляется методом наименьших квадратов, основной принцип которого заключается в минимизации среднего квадратичного отклонения по обучающей выборке. Таким образом, задача сводится к поиску локального минимума функции среднего квадратичного отклонения от вектора параметров модели \bar{b} .

Метод частично наименьших квадратов основан на преобразовании исходного вектора независимых переменных в вектор признаков, объясняющий как можно больше ковариации между зависимыми и независимыми переменными. Далее этот вектор используется для построения линейной функции методом наименьших квадратов. Основным преимуществом метода частично наименьших квадратов является возможность работы с коррелированными признаками, а также сокращать размерность вектора независимых переменных.

Метод опорных векторов — еще один широко применяемый алгоритм классификации и регрессии[31]. Принцип метода заключается в поиске разделяющей гиперплоскости, на максимальном расстоянии от всех объектов разных классов (Рисунок 4). Существуют подходы к построению разделяющей гиперплоскости при линейной неразделимости классов, стремящиеся минимизировать число ошибок при классификации, или использующие нелинейные ядра.

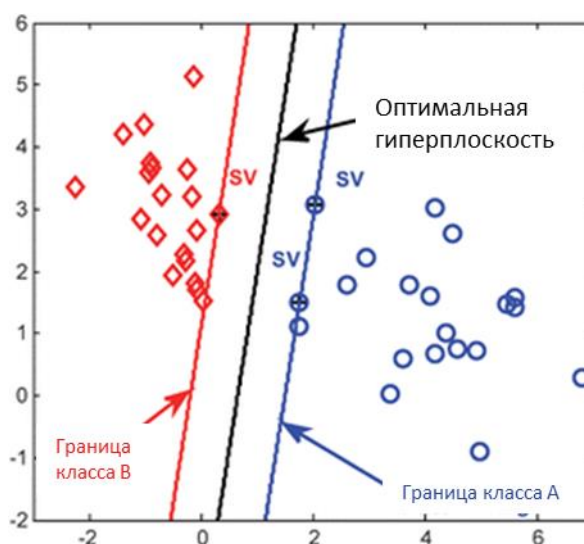


Рисунок 4. Принцип метода опорных векторов[31]. SV – опорные вектора.

Дерево решений — алгоритм классификации или регрессии, основанный на построении диаграмм, состоящих из «листьев» и «ветвей»[32]. В «ветвях» записывают значения независимых переменных (признаки), в «листьях» - значения зависимой переменной (целевой переменной).

Пример дерева решений для бинарной классификации по двум непрерывным признакам[33] представлен на рисунке 5.

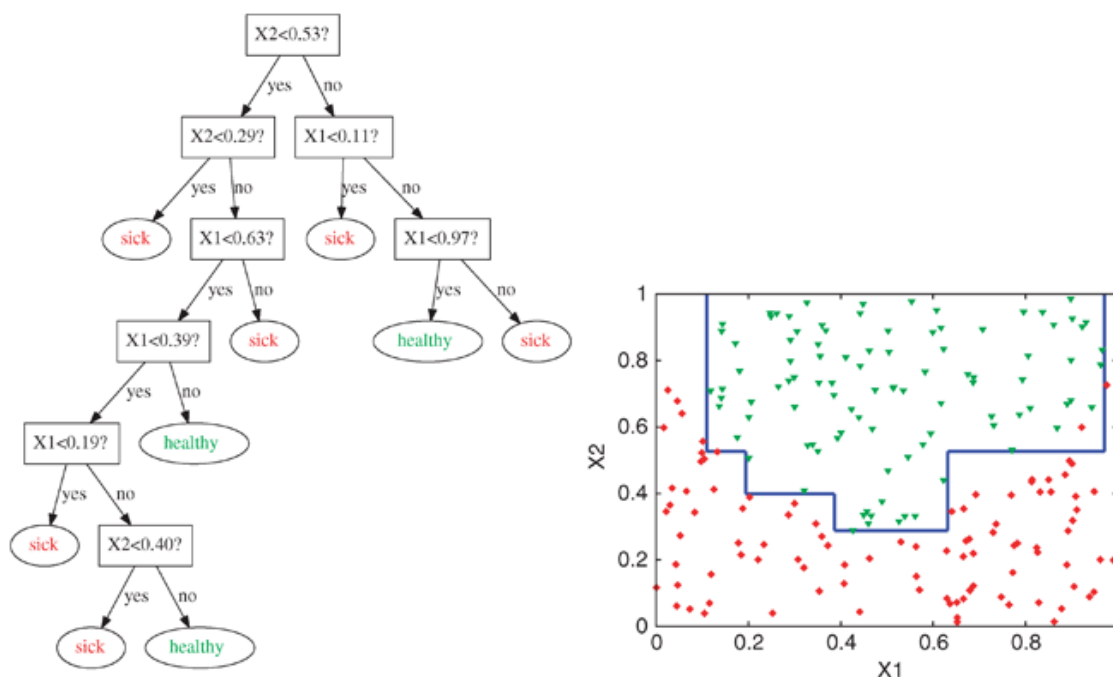


Рисунок 5. Пример дерева решений для бинарной классификации по двум непрерывным признакам X_1 и X_2 [33].

Для предсказания необходимо исходя из значений входного вектора независимых переменных дойти до листа, а для обучения предложены различные алгоритмы построения дерева по обучающей выборке. Основным преимуществом деревьев решений является их интерпретируемость и возможность визуализации. Однако, деревья решений склонны к переобучению, особенно при решении задач регрессии, и характеризуются невысокой точностью[34]. Хотя для преодоления этой проблемы предложены различные способы, деревья решений практически не применяются напрямую для моделирования молекулярных свойств, однако являются основой для более эффективных алгоритмов случайного леса и ГБ, и применяются в качестве вспомогательного инструмента для анализа данных[35].

Метод случайного леса (Рисунок 6) – алгоритм МО, основанный на обучении ансамбля деревьев решений[36]. Каждое дерево ансамбля обучается на случайной повторной подвыборке исходного обучающего набора данных, а результат предсказаний определяется «голосованием» каждого дерева. Построение дерева не зависит от построения других элементов ансамбля, т.е. обучение всех деревьев происходит параллельно. Метод случайного леса широко применялся для моделирования таких молекулярных свойств, как мутагенность, токсичность, растворимость [37-40], из-за возможности обрабатывать данные с большим количеством признаков.

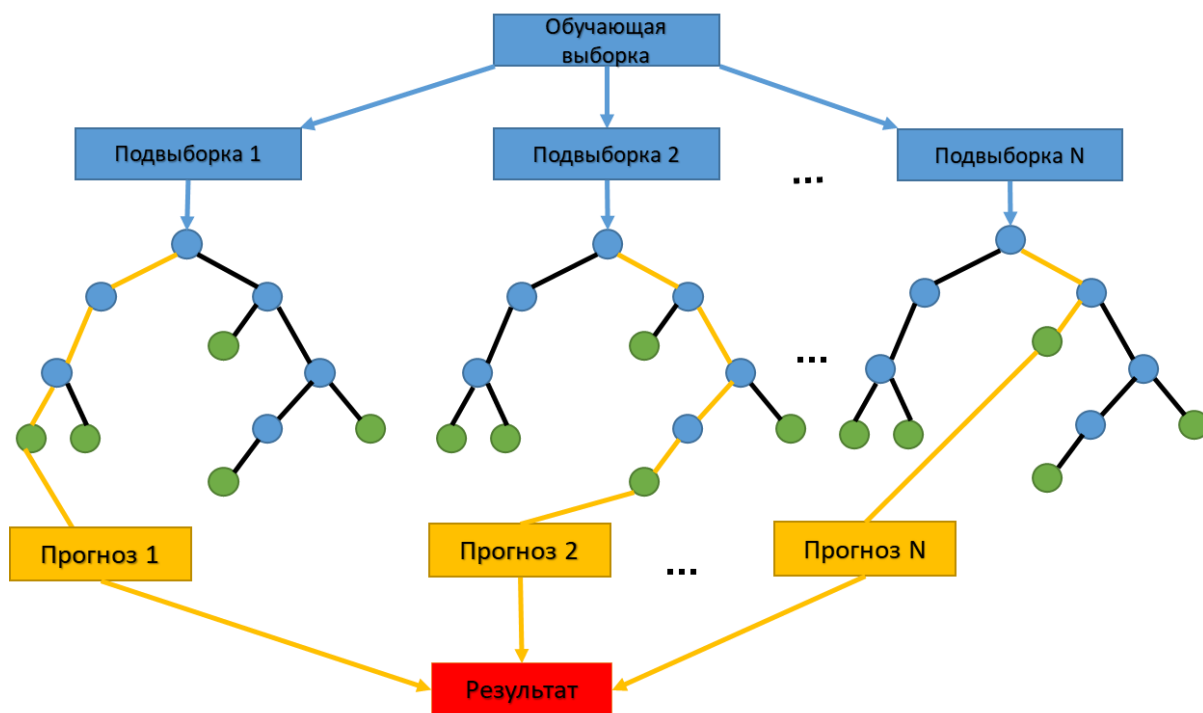


Рисунок 6. Принцип работы алгоритма случайного леса.

Метод бустинга, как и метод случайного леса, является ансамблевым методом, однако, в отличие от последнего, подразумевает последовательное, а не параллельное обучение индивидуальных моделей (Рисунок 7). Это позволяет последующей модели учитывать отклонения предсказаний, полученных предыдущей моделью, например, с помощью метода градиентного спуска. В качестве индивидуальных моделей как правило используются деревья решений. Модели, работающие по принципу адаптивного бустинга[41, 42], нашли применение для классификации органических соединений[43], однако, ранние реализации метода были сопоставимы по точности с моделями случайного леса[44]. Алгоритм адаптивного бустинга при обучении последующих деревьев увеличивал вклад обучающих примеров, которые были ошибочно классифицированы предыдущим деревом ансамбля.

Существенный прогресс был связан с появлением алгоритмов ГБ, где для коррекции весов последующих деревьев использовался метод градиентного спуска для поиска минимума функции потерь[45]. Существует несколько реализаций идеи ГБ, наиболее распространенными являются XGBoost[46], CatBoost[47, 48], LightGBM[49]. Они отличаются высокой скоростью обучения, особенно в реализациях для графических процессоров, и широко применяются, в том числе при моделировании химических свойств[50, 51].

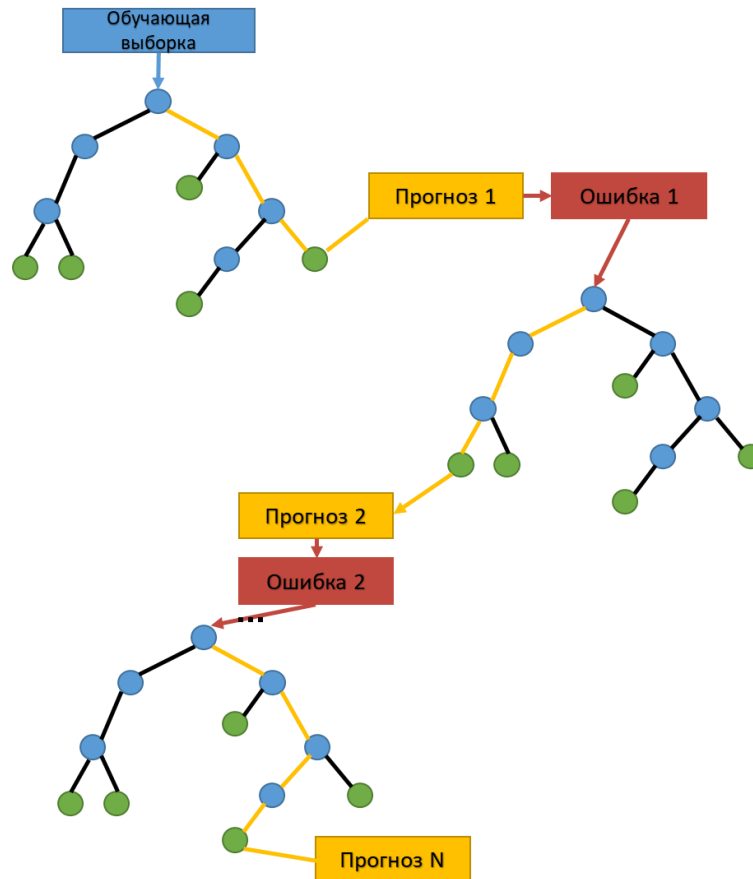


Рисунок 7. Принцип работы алгоритма градиентного бустинга

Искусственные нейронные сети, являясь одним из методов МО, положили начало новому быстроразвивающемуся направлению глубокого обучения. Искусственные нейронные сети появились в результате попыток смоделировать работу биологических нейронов[52]. Основными вехами в развитии ИНС была предложенная Розенблаттом модель перцептрона, которая моделировала поведение нейрона[53] и алгоритм обратного распространения ошибки[54], позволявший обучать ИНС из нескольких слоев перцептронов.

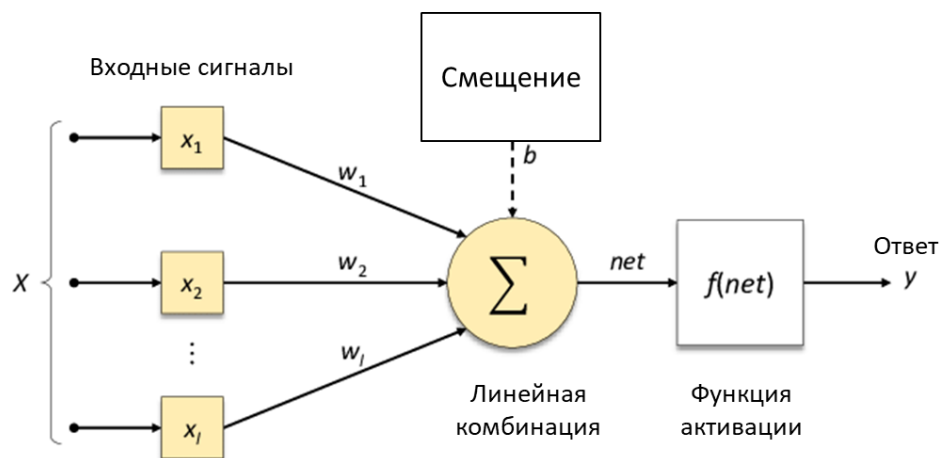


Рисунок 8 Схема работы перцептрона.

Схема работы перцептрона приведена на рисунке 8. Получив на вход вектор (x_1, \dots, x_n) независимых переменных (сигналов), система перемножает его на вектор случайных весов (w_1, \dots, w_n) . Сумма произведений входных сигналов и весов подается в нелинейную функцию активации, которая выдает ответ. Обучение перцептрона проводится методом коррекции ошибки, при котором в случае неверного ответа значения весов изменяют тем или иным образом.

Алгоритм обратного распространения ошибки позволил обучать многослойные перцептроны. Такое название алгоритм получил из-за того, что вычисления поправок к весам, например, методом градиентного спуска, проводят в направлении от выходного слоя к входному, т.е. в направлении, обратном распространению сигнала при нормальной работе сети. При этом, поправки к весам более низкого уровня выражаются через поправки более высокого уровня.

Многослойные перцептроны (Рисунок 9), в которых все нейроны связаны между собой называются полносвязными ИНС[55]. Их основным недостатком считается очень большое количество весов, что негативно влияет на скорость обучения. Тем не менее, полносвязные ИНС широко применяются при моделировании молекулярных свойств[56, 57].

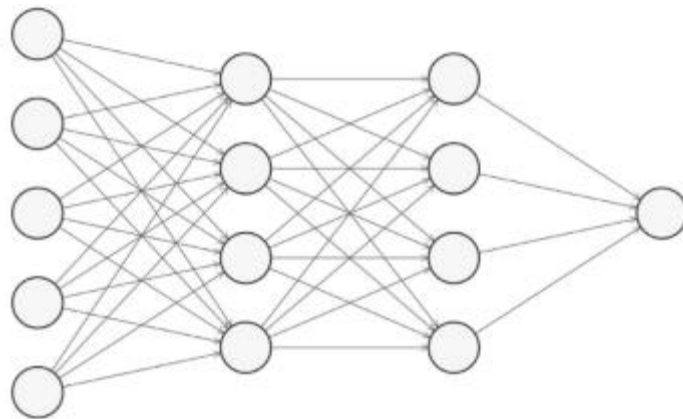


Рисунок 9. Схема полносвязной нейронной сети.

Сверточные ИНС были разработаны для обработки двумерной информации, в первую очередь изображений[58, 59]. Операция свертки заключается в перемножении участков исходного объекта, представленного в виде числовой матрицы на матрицу весов, называемую фильтром или ядром свертки (Рисунок 10). В сочетании с операцией подвыборки (Pooling), это позволяет существенно сократить число параметров (весов) сети, что положительно сказывается на ее склонности к переобучению. Операция подвыборки сводится к применению к матрице нелинейной функции. Например, на рисунке 11, изображен пример работы функции подвыборки с функцией максимума (MaxPooling). Другой широко применяемой операцией является функция усреднения (Average Pooling).

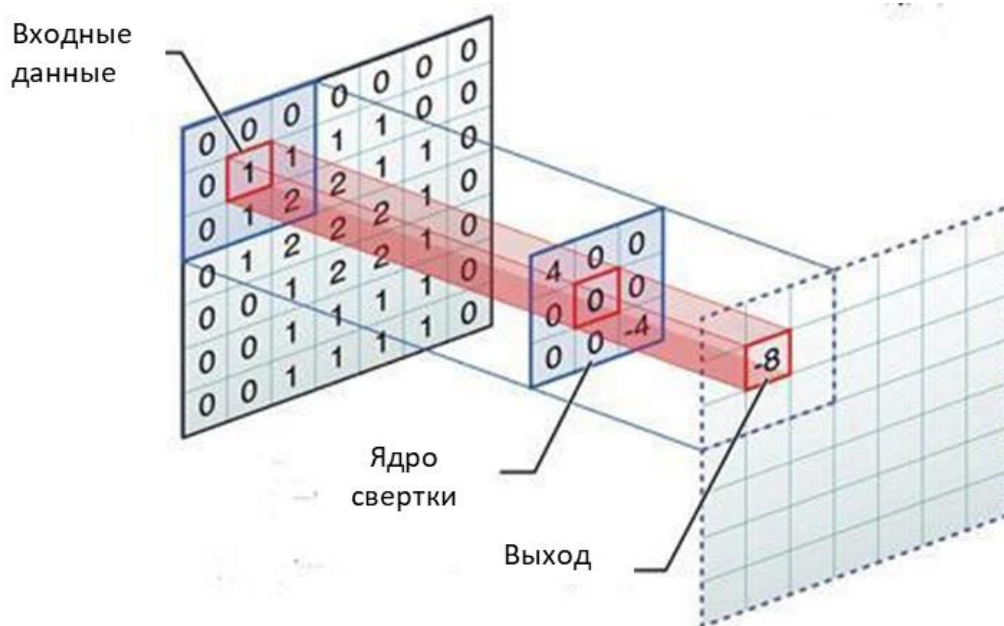


Рисунок 10. Принцип работы сверточных слоев.

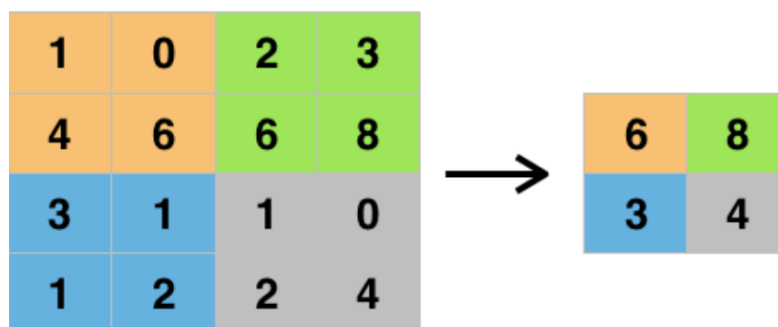


Рисунок 11. Пример операции подвыборки (с функцией максимума Max Pooling), применяемой в архитектуре сверточных нейронных сетей

Рекуррентные ИНС и впервые были предложены для обработки последовательных данных, в частности временных рядов. Появление рекуррентных ИНС с кратковременной и долгосрочной памятью[60] обеспечило существенный прогресс в области обработки текстов[61], распознавания речи.

Область глубокого обучения является одной из наиболее быстроразвивающихся, и появляются другие архитектуры ИНС, например, трансформеры[62, 63], основанные на механизме внимания. Разработанные в 2017 году для обработки текстовых данных, они уже нашли применение в моделировании молекулярных свойств[64-67].

Бурное развитие методов глубокого обучения связано с совокупным прогрессом в области вычислений на графических процессорах, который обеспечил ускорение обучения глубоких ИНС на несколько порядков, а также появление большого количества программных библиотек, для автоматического дифференцирования при вычислении градиентов. Это позволило использовать методы машинного и глубокого обучения исследователям из различных областей,

включая химические исследования. Помимо широко распространенных сред разработки TensorFlow[68], Keras[69], PyTorch, SciKit Learn[70], существует среда разработки для применения методов МО и разработки ИНС непосредственно для химических данных - DeepChem[71].

Одним из требований при моделировании с помощью ИНС является наличие обучающей выборки достаточного размера. Обладая большим количеством параметров ИНС более склонны к переобучению, и опережают традиционные методы МО по точности только при достаточном количестве обучающих примеров. Тем не менее, применяются различные подходы для работы с обучающими выборками ограниченного размера. Один из них – это аугментация[72, 73], за счет незначительного изменения исходных экспериментальных данных, или использование синтетических данных. Другой – так называемое «обучение с переносом», заключающееся в предварительном обучении всех, или нескольких слоев ИНС на выборке большого объема, имеющей косвенное отношение к решаемой задаче, с последующим до-обучением[74] ИНС на целевой выборке меньшего размера. При до-обучении, веса, полученные на первом этапе, сохраняются, или изменяются с небольшой скоростью.

Методы МО давно применяются в областях, связанных с проведением химического анализа[75-77]. Общая тенденция заключается в развитии методов глубокого обучения, применительно к задачам химического анализа, в виду их большого потенциала по выявлению закономерностей между структурами молекул, и молекулярными характеристиками, используемыми в аналитической химии[78]. Нужно отметить, что многие примеры применения машинного и глубокого обучения в химическом анализе, относятся к хромато-масс-спектрометрии белков и пептидов[79-85]. Моделирование свойств в исследованиях низкомолекулярных соединений, как правило, является более сложной задачей, ввиду широкого разнообразия структур и структурных фрагментов, а как следствие, высокой варибельности свойств, и условий хромато-масс-спектрометрического анализа[75].

1.3 Способы оценки моделей машинного обучения

1.3.1 Компромисс «отклонение - дисперсия»

Смещение — это погрешность оценки, возникающая в результате ошибочного предположения в алгоритме обучения. В результате большого смещения алгоритм может пропустить связь между признаками и выводом. Дисперсия — это ошибка чувствительности к малым отклонениям в тренировочном наборе. При высокой дисперсии алгоритм может как-то трактовать случайный шум в тренировочном наборе, а не желаемый результат. Модели с меньшим отклонением от имеющихся данных имеют более высокую дисперсию на новых данных (то есть подвержены переобучению), и наоборот. Разложение смещения-дисперсии —

это способ анализа ожидаемой ошибки обобщения алгоритма обучения для частной задачи сведением к сумме трёх членов — смещения, дисперсии и величины, называемой неустранимой погрешностью, которая является результатом шума в самой задаче. Разложение может быть выражено уравнением:

$$E[(y - \hat{f}(x))^2] = (\text{Bias}[\hat{f}(x)])^2 + \text{Var}[\hat{f}(x)] + \sigma^2, \text{ где}$$

$$\text{Bias}[\hat{f}(x)] = E[f(x) - \hat{f}(x)],$$

$$\text{Var}[\hat{f}(x)] = E[\hat{f}(x)^2] - (E[\hat{f}(x)])^2.$$

Здесь x — точка за пределами обучающей выборки, $y = f(x) + \varepsilon$ — истинная функция с шумом где шум ε имеет нулевое среднее и дисперсию σ^2 . $\hat{f}(x)$ — искомая функция, аппроксимирующая $f(x)$. Можно видеть, что общая ошибка складывается из трех компонентов, смещения, дисперсии и шума. В случае с высоким смещением и низкой дисперсией говорят о недообучении модели (или ее недостаточной сложности), в случае с низким смещением и высокой дисперсией о переобучении. Переобучение является одной из наиболее важных проблем в МО, так как в этом случае модели хорошо работают на примерах из обучающей выборки, но непредсказуемо на примерах за ее пределами. Поэтому контролю и управлению переобучением уделяется существенное внимание, особенно в глубоком обучении, где ИНС, характеризующиеся большим числом параметров, склонны к переобучению.

1.3.2 Основные метрики, применяемые в машинном обучении

Для задач регрессии, наиболее часто применяют следующие метрики:

Среднее (абсолютное) отклонение (Mean Absolute Error, MAE)

$$MAE = \frac{\sum_{i=1}^T |\hat{y}_i - y_i|}{T}$$

Среднеквадратичное отклонение (Root Mean Squared Error, RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^T (\hat{y}_i - y_i)^2}{T}}$$

Коэффициент детерминации (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^T (\hat{y}_i - y_i)^2}{\sum_{i=1}^T (y_i - \bar{y})^2}$$

Где T — число примеров, \hat{y}_i, y_i — вычисленные и экспериментальные значения, \bar{y} — среднее по экспериментальным значениям. Нужно отметить, что вместо средних значений метрик часто

используют медианные, как более устойчивые к выбросам. Для сравнения данных разного масштаба применяют также относительные метрики.

Нужно отметить различные метрики для многоцелевой регрессии, наиболее распространенной является косинусная мера сходства векторов, которая вычисляется через их скалярное произведение:

$$\text{cosine similarity} = \frac{A \cdot B}{\|A\| \|B\|}$$

В задачах бинарной классификации возможно 4 исхода. True positive (TP) – образец корректно классифицирован как положительный, true negative (TN) - образец корректно классифицирован как отрицательный, false positive (FP) – образец некорректно классифицирован как положительный, false negative (FN) - образец некорректно классифицирован как отрицательный. В зависимости от задачи в бинарной классификации используются следующие метрики:

Доля верно классифицированных ответов (Accuracy) $\frac{TP+TN}{TP+TN+FP+FN}$

Полнота (Recall, True positive rate, TPR) $\frac{TP}{TP+FN}$

Точность (Precision) $\frac{TP}{TP+FP}$

False positive rate (FPR) $\frac{FP}{FP+FN}$

Отдельно выделяют площадь под ROC-кривой (ROC – receiver operation characteristic, рабочая характеристика приемника). ROC-кривая представляет собой график TPR от FPR при варьировании определяющего порога. Площадь под кривой ROC определяет качество классификатора (Рисунок 12).

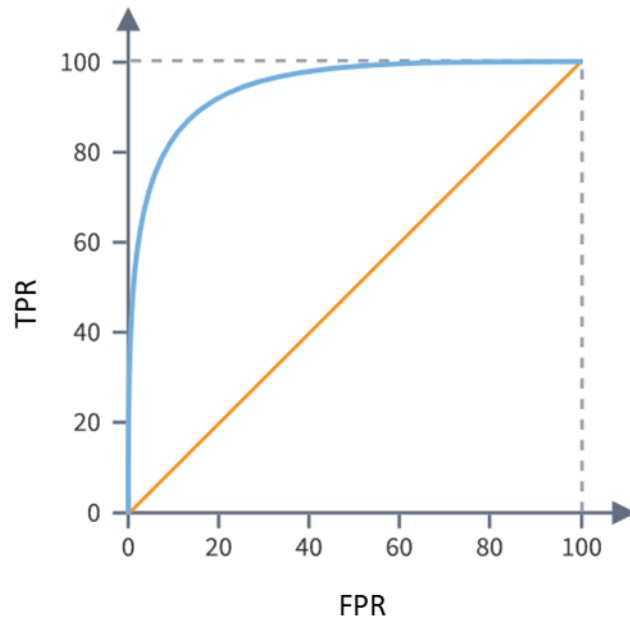


Рисунок 12. Пример ROC-кривой. Диагональная линия соответствует случайному классификатору.

1.3.3 Способы валидации моделей машинного обучения

Наиболее простым способом оценки модели является разбиение исходного данных на обучающую и тестовую выборки случайным образом. Несмотря на то, что однократное случайное разбиение обычно приводит к смещенной оценке, данный подход иногда используют для оценки очень сложных моделей, обучение которых требует существенных ресурсов.

Золотым стандартом является принцип кросс-валидации, изображенный на рисунке 13. При кросс-валидации общий набор данных разделяется на n равных частей случайным образом, одна из которых используется в качестве тестовой выборки, а оставшиеся $n-1$ в качестве обучающей выборки. Процедура повторяется N раз, в результате все образцы исходного набора один раз окажутся в тестовой выборке. Для обеспечения дополнительной уверенности можно использовать зарезервировать дополнительную тестовую выборку, которая вообще не будет участвовать в процессе обучения.

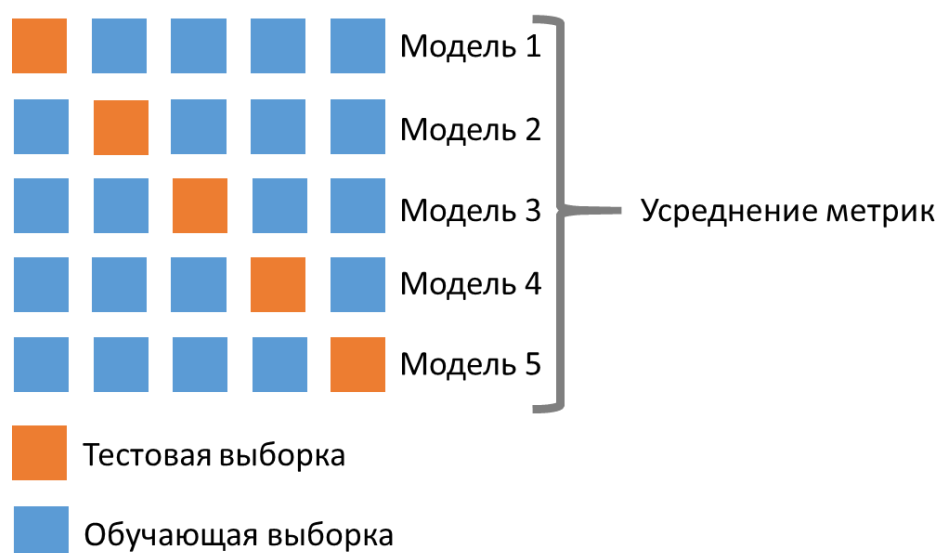


Рисунок 13. Принцип кросс-валидации ($n=5$).

В задачах моделирования молекулярных свойств, проблема ограниченности доступных данных является одной из ключевых. В разных исследованиях было показано, что наличие похожих молекул в обучающей и тестовой выборке может завышать результаты при кросс-валидации, при этом способность модели обобщать предсказания на молекулы с отличающейся структурой может оказаться ниже ожидаемой. Для достижения более адекватной оценки предложено вместо случайного разбиения формировать обучающую и тестовую выборку с учетом структуры молекулы[86].

1.4 Способы представления молекул в машинном обучении

Для применения методов МО к молекулам, последние должны быть тем или иным способом охарактеризованы числовым вектором. Методы глубокого обучения имеют преимущество в том, что этот вектор может быть просто числовым представлением молекулы в виде молекулярного графа, текста или даже изображения структурной формулы, из которых при обучении ИНС будут экстрагированы признаки, по которым можно строить предсказание. Для традиционных методов МО числовой вектор должен сразу содержать некоторый осмысленный набор признаков.

Исторически первым способом описания молекул для МО являлись молекулярные дескрипторы, которые могут быть рассчитаны по молекулярной структуре. Широко применяются топологические дескрипторы[87, 88], такие как индексы Винера[89] или Рандича[90], характеризующие молекулярный граф без учета химических особенностей молекул (кратности связей и типы атомов). Наиболее многочисленный класс физико-химических дескрипторов включает различные числовые характеристики физико-химических свойств, которые могут быть смоделированы и рассчитаны различными методами. Примерами таких

дескрипторов являются молекулярная масса, липофильность, молекулярные объемы. Квантово-химические дескрипторы являются характеристиками, полученными в результате квантово-химических расчетов, например, энергии граничных молекулярных орбиталей[91, 92]. Особняком стоят фрагментные дескрипторы, которые в англоязычной литературе зачастую выделяют в отдельный класс под названием «молекулярных фингерпринтов». Фрагментные дескрипторы — это числовая характеристика молекулы, показывающая присутствие в ее структуре определенного фрагмента. Вектор фрагментных дескрипторов может быть представлен в виде вектора логических переменных, что сокращает вычислительные ресурсы при их использовании. В тоже время довольно широко применяются аддитивные целочисленные фрагментные дескрипторы, учитывающие, сколько раз фрагмент встречается в структуре молекулы. Наиболее распространенными примерами фрагментных дескрипторов являются MACCS Keys, Extended-Connectivity Fingerprints[93].

При использовании таких алгоритмов, как множественная линейная регрессия или метод опорных векторов, перед исследователем стоит задача тщательного отбора дескрипторов. Использование широких наборов признаков не только увеличивает время обучения, но и приводит к зашумлению данных, переобучению и снижению общей точности предсказаний. При этом отбор признаков основывается на экспертной оценке, исходя из предварительных знаний о зависимости целевой молекулярной характеристики от тех или иных физико-химических свойств. Алгоритмы случайного леса и ГБ более устойчивы к использованию большого количества признаков, кроме того, в силу особенностей построения деревьев решений позволяют оценивать вклад того или иного признака в результат предсказаний. Широкие наборы молекулярных дескрипторов успешно применяются и для обучения глубоких ИНС.

В связи с появлением алгоритмов, способных устойчиво работать с большим набором молекулярных дескрипторов, последние обычно рассчитываются пакетом из нескольких сотен или тысяч параметров. Функционал для расчета таких пакетов зачастую встроен в программное обеспечение или библиотеки для работы с химическими данными (ACD Labs[94], RDKit[95], CDK[96]), или реализуется в виде отдельных библиотек для различных языков программирования (Mordred[97], PaDEL[98]).

Нейронные сети способны также работать с различными представлениями молекул, при этом ИНС обучается переводить эти представления в вектора латентного пространства признаков, которые по своей сути являются молекулярными дескрипторами, хотя и плохо интерпретируемыми[99, 100]. Модель для генерации таких дескрипторов может быть обучена на очень больших обучающих выборках из неразмеченных данных, а потому подобные

дескрипторы в дальнейшем могут быть применены для построения частных моделей регрессии и классификации[66, 67, 101].

Так, показано, что сверточная ИНС способна извлекать молекулярные дескрипторы из графических изображений структурной формулы молекулы[102]. При стандартизации изображений, удалось достичь довольно высокой степени точности при решении различных задач установления количественных отношений структура – свойство (QSPR).

В качестве текстовых представлений чаще всего используют представление молекулы в виде строк системы упрощённого представления молекул в строке ввода (Simplified Molecular Input Line Entry System) (**SMILES**) [103], которое переводится в числовой вектор или матрицу унитарным кодированием. Представление молекулы в виде строки SMILES производится по определенным правилам, которые могут отличаться в зависимости от применяемого ПО, кроме того, алгоритм построения допускает несколько форматов записи SMILES для одной и той же молекулы[104]. С одной стороны, это накладывает дополнительные требования к стандартизации и определению правил по записи канонической формы, с другой открывает возможность для аугментации данных. Другой широко применяемый формат представления молекулы в виде текста – международный химический идентификатор молекулы (International Chemical Identifier, **InChI**)[105], практически не применяется в МО ввиду того, что полученная строка плохо интерпретирует тип, и порядок связей атомов в молекуле, в отличие от SMILES (Рисунок 14).

Текстовые представления молекул оказались очень популярны при построении глубоких сетей для предсказания молекулярных свойств по нескольким причинам. Во-первых, повышенный интерес к нейролингвистическому программированию привел к созданию сложных архитектур ИНС для работы с текстовой информацией, решения задач машинного перевода, классификации текстовой информации. Подобные архитектуры могут быть обучены в том числе на размеченных данных. Например, в ходе обучения ИНС может учиться предсказывать следующее слово в предложении[106]. Такой подход можно применить и к молекулам, рассматривая SMILES как предложение, а каждый символ в строке — как слово. Тогда ИНС можно предварительно обучать на структурах молекул без какой-либо дополнительной информации[64, 107].

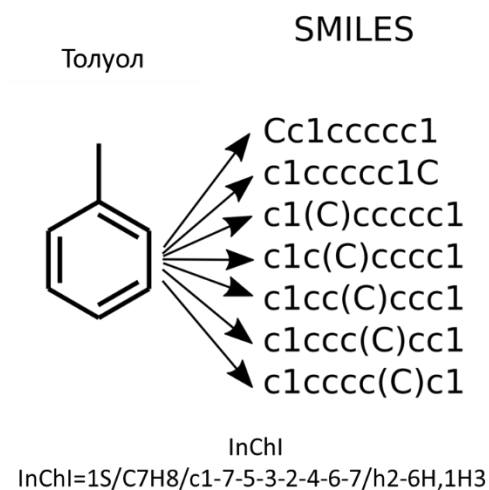


Рисунок 14. Текстовое представление молекулы толуола в виде SMILES и InChI.

Наиболее естественное представление молекул в виде графов широко применяется для обучения графовых ИНС различных архитектур[108-111]. При этом атомы (вершины графа) и связи (ребра графа) описываются числовыми векторами относительно простых признаков (например, атомный номер, валентность, молекулярный вес, кратность связи)[112]. На рисунке 15 приведен пример графовой ИНС, состоящей из нескольких основных операций. После агрегации состояний окружения атома, происходит обновление его вектора состояний, в результате чего получаются индивидуальные представления каждого атома в некотором латентном пространстве, которые затем считываются в вектор признаков всего графа. Далее этот вектор поступает на вход полносвязной ИНС.

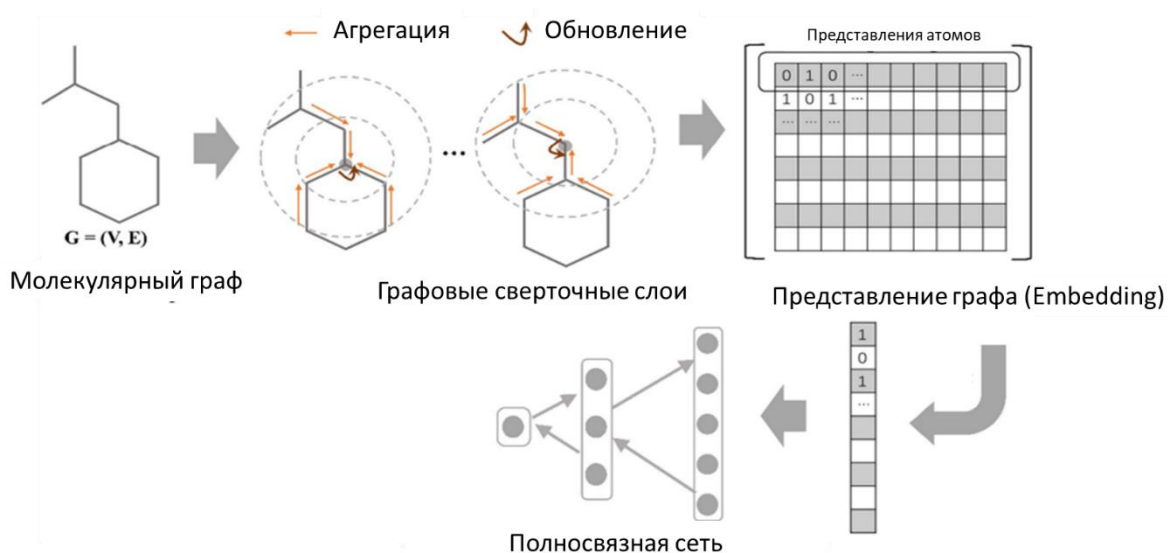


Рисунок 15. Пример графовой нейронной сети[113].

1.5 Примеры применения методов машинного обучения для предсказания аналитических характеристик низкомолекулярных соединений

1.5.1 Применение методов машинного обучения для предсказания характеристик удерживания низкомолекулярных соединений в газовой хромато-масс-спектрометрии

В ГХ-МС в качестве основной характеристики широко распространено применение ИУ. Использование относительной характеристики нивелирует влияние геометрических параметров колонки, и, с некоторыми допущениями, параметров элюирования, делая ИУ более универсальной характеристикой, чем абсолютные значения времен удерживания. Понятие ИУ впервые было введено в 1958 году Ковачем[114], который предложил для изотермических условий разделения нормировать времена удерживания на времена удерживания n-алканов, в соответствии с уравнением:

$$I_s = 100 \times \left(z + \frac{(\log X_s - \log X_z)}{(\log X_{z+1} - \log X_z)} \right) \quad (1),$$

где I_s — ИУ соединения, X_s , X_z , X_{z+1} — времена удерживания соединения, и n-алканов с количеством атомов углерода z и $z+1$, такие, что $X_z < X_s < X_{z+1}$. В 1963 году были предложены линейные ИУ для градиентного элюирования[115], рассчитываемые по уравнению:

$$I_s = 100 \times \left(z + \frac{(X_s - X_z)}{(X_{z+1} - X_z)} \right) \quad (2),$$

где I_s — ИУ соединения, X_s , X_z , X_{z+1} — времена удерживания соединения, и n-алканов с количеством атомов углерода z и $z+1$, такие, что $X_z < X_s < X_{z+1}$. В соответствии с определением индексы Ковача и линейные ИУ n-алкана с z атомами углерода равны $100 \times z$. Хотя индексы Ковача и линейные ИУ получили наибольшее распространение, были предложены и другие системы индексирования, основанные на различных молекулярных рядах, в первую очередь, для работы с полярными неподвижными фазами[116]. Среди них, чаще всего применяются индексы Ли, рассчитываемые по удерживанию полядерных ароматических углеводородов[117].

Хорошая воспроизводимость ИУ способствовала их накоплению в специализированных библиотеках. Так в библиотеке NIST 20 Retention index library[13], собрано более 440 000 экспериментально измеренных значений ИУ для почти 140 000 соединений. Отсутствующие значения могут быть предсказаны с применением методов МО.

На раннем этапе развития систем предсказания ИУ, основанных на методах МО, наибольшее распространение получил алгоритм множественной линейной регрессии, построенной по физико-химическим свойствам молекул[118, 119]. Подавляющее большинство

ранних работ по предсказанию ИУ отличались избирательностью по отношению к определенным классам органических соединений[120]. С одной стороны, это позволяло увеличить точность предсказаний внутри класса, с другой, существенно ограничивало универсальность модели и ее способность к обобщению. Были предложены подходы к моделированию ИУ флавоноидов[121], терпеноидов[122]. Более универсальные модели удалось построить с применением расширенных библиотек удерживания и наборов физико-химических дескрипторов[123]. Несмотря на большое количество работ, посвященных предсказанию ИУ [120, 124-128], наиболее применяемой долгое время являлась аддитивная модель[129], разработанная в 2007 г.

В рамках этой модели, ИУ молекулы рассматривается как сумма атомных и групповых инкрементов и поправки f , равной среднему отклонению предсказанных ИУ от экспериментальных значений для соединений данного ряда:

$$I = \sum_{i=1}^N \Delta I_i + f$$

Хотя данная аддитивная схема и характеризовалась высоким отклонением предсказанных значений от экспериментальных (медианное значение 46-60 единиц для различных типов неподвижных фаз), она была разработана с учетом информации об удерживании 35 000 молекул, и потому была достаточно универсальна. В тоже время модели, обученные на данных об удерживании молекул определенных классов, хотя и характеризовались более высокой точностью (в среднем стандартное отклонение составляло 38.4 единицы по результатам более 30 работ, рассмотренных в обзоре[119]).

Современная тенденция в моделировании ИУ заключается в построении универсальных моделей с широкой сферой применения. Появлению таких моделей способствовало развитие методов МО, в частности глубоких ИНС, и накопление экспериментальной информации о газохроматографическом удерживании. Сочетание больших обучающих выборок и продвинутых алгоритмов позволило достичь высокой точности предсказаний. Так, в одной из первых работ, использовалась модель ГБ [130], обученная на выборке из 42234 молекул из библиотеки NIST 17 с известными значениями ИУ, величина среднего отклонения находилась в диапазоне 46-58 единиц на трех независимых тестовых выборках. Входной вектор для данной модели формировался из 177 физико-химических дескрипторов, рассчитанных с помощью ПО Chemical Development Kit[96]. Дальнейшее развитие идея создания универсальных моделей получила в виде применения методов глубокого обучения, в частности, одномерных и двумерных сверточных ИНС[131, 132]. Глубокое обучение не требует формирования набора признаков, и допускает обучение модели на представлениях, с распознаванием признаков в ходе обучения.

Данные признаки, определенные в ходе обучения, формируют латентный вектор, который далее используется для формирования предсказания полносвязной ИНС с линейным слоем. В вышеупомянутых работах модели обучали на текстовых представлениях молекул SMILES[103, 104], представленных в виде вектора или матрицы унитарного кодирования (Рисунок 16)

	O	=	C	c	1	c	c	c	(O)	c	(O	C)	c	1	
C	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	0	1	0	1	1	1	0	0	0	1	0	0	0	0	0	1	0
O	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
=	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
)	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Рисунок 16. Представление строки SMILES в виде матрицы унитарного кодирования (one-hot encoding)[133].

Подход с применением двумерных сверточных ИНС позволил получить более высокую точность предсказаний со средним отклонением в 28.4 единицы, рассчитанным по независимой тестовой выборке, в то время как для подхода с применением одномерных сверточных ИНС среднее отклонение находилось в диапазоне 33.2–63.6 единиц для набора независимых тестовых выборок. Однако, нужно отметить, что прямое сравнение моделей по приведенным в статьях данным не всегда корректно, т.к. в этих работах использованы различные обучающие и тестовые выборки.

Другой подход к предсказанию ИУ методом глубокого обучения основан на представлении молекул в виде графов, вершинами которых являются атомы, а ребрами – межатомные связи. При применении ИНС с открытой архитектурой, включающей графовые слои, среднее отклонение составило 28 единиц при расчете по независимой тестовой выборке[134].

Наиболее высокой точностью отличается мультимодальный подход. Он подразумевает объединение результатов предсказаний различных моделей, обученных с применением различных представлений и/или молекулярных признаков дополнительной моделью. Так, для ИУ, было показано, что объединение результатов предсказаний одномерной и двумерной сверточной сети обученных на текстовых представлениях, а также полносвязной ИНС с остаточными связями и модели ГБ, обученных с использованием совокупного вектора фрагментных и физико-химических дескрипторов, позволяет предсказывать индексы удерживания со средним отклонением в диапазоне 10-93 единиц для различных независимых выборок[135]. Для получения объединенного предсказания результаты отдельных алгоритмов

использовались в качестве независимых переменных линейной модели, параметры которой оптимизировались при обучении для минимизации функции потерь. Схема данного подхода представлена на рисунке 17.

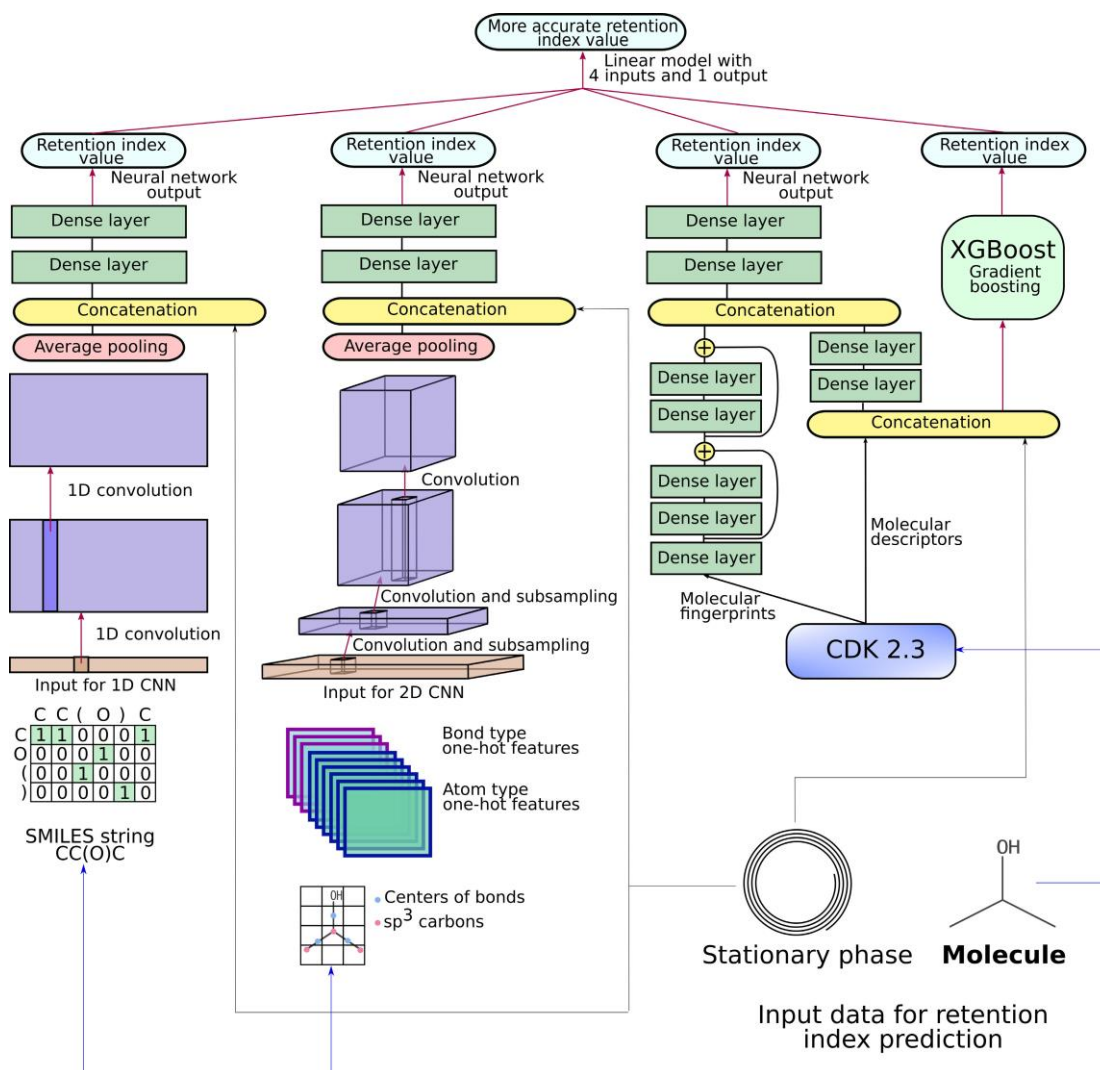


Рисунок 17. Предсказание индексов удерживание с использованием мультимодального подхода[135].

Нужно отметить, что в большинстве работ не делается различие между индексами Ковача и линейными ИУ [129], индексы Ли обычно пересчитывают в линейные по эмпирическим правилам. Кроме того, следуя классификации неподвижных фаз по полярности, представленной в библиотеке NIST, модели обучают по совокупным данным, полученным для неподвижных фаз одного класса. Однако, были предложены подходы к моделированию ИУ для конкретных сорбентов, и показано, что такой подход является более точным[136].

1.5.2 Применение методов машинного обучения для предсказания характеристик удерживания низкомолекулярных соединений в жидкостной хромато-масс-спектрометрии

Несмотря на многочисленные попытки введения систем ИУ в жидкостной хроматографии [137], основной характеристикой молекул остается абсолютное значение времени удерживания. Время удерживания в нецелевом ЖХ-МС анализе является дополнительной характеристикой, которая повышает достоверность идентификации по масс-спектрометрическим данным. Ввиду ограниченной доступности образцов сравнения, методы предсказания времен удерживания по структуре молекулы могут найти применение в аналитической практике. Механизмы удерживания довольно сложны, и прямое моделирование физико-химических процессов взаимодействия аналита и сорбента требует учета множества факторов, включая условия разделения, состав подвижной фазы, температуру. Методы МО требуют для моделирования только обучающую выборку достаточного размера, и потому наиболее распространены среди различных подходов к предсказанию времен удерживания.

Первые подходы к предсказанию времен удерживания с помощью МО были основаны на традиционных алгоритмах – множественной линейной регрессии, алгоритме случайного леса, ГБ. Ограниченное применение ИНС обуславливалось небольшими объемами обучающих выборок. В большинстве работ использовали небольшие наборы данных, число веществ в которых не превышало тысячи. Применение глубоких ИНС с обучением на представлениях в таких условиях неоправданно в виду эффекта переобучения, в то время как традиционные алгоритмы, в сочетании с небольшими наборами физико-химических или фрагментных дескрипторов для описания молекул, способны обеспечить приемлемую точность. Некоторые примеры работ, посвященных предсказанию времен удерживания низкомолекулярных соединений приведены в таблице 3.

Исследования различных алгоритмов при моделировании времен удерживания для различных условий разделения по открытым наборам данных не смогли выявить наиболее точный алгоритм для моделирования удерживания. В зависимости от состава и объема обучающей выборки, минимальные средние отклонения обеспечивали различные алгоритмы, хотя наблюдалась некоторая тенденция в пользу ГБ [138].

Таблица 3. Работы, посвященные предсказанию времен удерживания низкомолекулярных соединений

Метод МО	Размер обучающей выборки	Ссылка
Множественная линейная регрессия	532	[139]
Метод опорных векторов	442	[140]
Частичная регрессия наименьших квадратов, ИНС	260	[141]
Метод опорных векторов	201	[142]
Частичная регрессия наименьших квадратов	1383	[143]
Случайный лес	904	[144]
Метод опорных векторов	2138	[145]
Частичная регрессия наименьших квадратов	91	[146]
ИНС	550	[147]
ГБ	852	[148]

Отдельно нужно отметить, что подобный подход к моделированию на небольших внутрилабораторных обучающих наборах данных, измеренных в конкретных условиях разделения, не отличается универсальностью, и полученные модели обычно не находят широкого применения из-за различий в хроматографических условиях между лабораториями и решаемыми задачами [120]. Частично эти ограничения снимают разрабатываемые методы пересчета времен удерживания между похожими хроматографическими системами [149-151]. В работе [18] не только предложен алгоритм построения моделей пересчета времен удерживания между хроматографическими системами, но и разработан интернет ресурс PredRet для обмена данными по удерживанию. Эти данные широко используются для тестирования различных алгоритмов МО.

Принципиально другой подход заключается в предсказании порядка выхода соединений из хроматографической колонки вместо абсолютных значений времени удерживания, т.к. он более воспроизводим между различными системами (при условии одного механизма удерживания) [152, 153]. Кроме того, предложены подходы к предсказанию ИУ, например, рассчитанные относительно серии неразветвленных первичных алифатических аминов [137, 154].

Существенный прогресс в предсказании времен удерживания был обеспечен появлением библиотеки METLIN SMRT. Библиотека содержит времена удерживания 80038 соединений различных классов, измеренные в условиях обращенно-фазовой ВЭЖХ. В оригинальной

публикации было показано, что глубокая ИНС, обученная на молекулах из этой библиотеки, представленных в виде фрагментных дескрипторов характеризуется средним и медианным отклонением в 57 и 35 с соответственно для независимой тестовой выборки. С учетом того, что ширина хроматографического пика на полувысоте в условиях разделения, использованных при сборе данных, составляла порядка 20-30 с, точность полученной модели можно считать приемлемой. Авторы также продемонстрировали, что алгоритм случайного леса уступает ИНС по точности предсказаний. Для применения в реальной практике для идентификации химических соединений с учетом вариабельности времени удерживания вещества при изменении условий разделения предложено построение полиномиальных моделей пересчета времен удерживания, предсказанных для условий разделения при создании обучающей выборки на другие хроматографические условия. Показано, что для создания таких моделей достаточно данных по удерживанию набора молекул в целевых условиях разделения, при этом наличие данных по удерживанию этих молекул в библиотеке METLIN SMRT не требуется, т.к. они могут быть аппроксимированы предсказанными значениями. После пересчета медианное отклонение, в зависимости от метода, находилось в диапазоне 5.7 – 210 с[29].

Появление библиотеки спровоцировало развитие методов предсказания времен удерживания, как за счет применения более сложных архитектур, так и за счет новых подходов к пересчету предсказаний. Так, для повышения точности предсказаний предложено использование ИНС различных архитектур, использующих различные представления молекул, а для получения предсказаний для различных хроматографических систем – метод обучения с переносом.

Исследования, опирающиеся при построении модели на данные из библиотеки METLIN SMRT сведены в таблице 4. Можно отметить, что применение ИНС со сложными архитектурами позволило почти в 2 раза снизить погрешность предсказаний в условиях METLIN SMRT. Для получения предсказаний в других хроматографических условиях в основном использовали метод обучения с переносом. В таблице 4 приведены также опубликованные медианные отклонения предсказаний времен удерживания в условиях хроматографического разделения, которые встречаются в литературе[18] как «FEM_long»[155] (общее время элюирования 60 мин) и «LIFE_old»[156] (общее время элюирования 6 мин). Можно отметить, что точность предсказаний после пересчета тем или иным методом не всегда коррелируют с точностью исходной модели, обученной на METLIN SMRT.

Таблица 4. Сравнение методов моделирования времен удерживания с использованием библиотеки METLIN SMRT

Метод МО (архитектура ИНС)	Среднее отклонение, с (метод METLIN SMRT)	Медианное отклонение, с (метод FEM_long)	Медианное отклонение, с (метод LIFE_old)	Способ пересчета предсказаний на другие хроматографические системы	Ссылка
Полносвязная ИНС, 4 скрытых слоя	57 с	210	9.7	Полиномиальная функция пересчета, построенная по предсказанным значениям	[29]
Одномерные сверточные ИНС	35 с	Нет данных	11.8	Обучение с переносом	[157]
Мультимодальное обучение	39 с	203	9.9	Функция пересчета, построенная по предсказанным значениям	[158]
Графовая сверточная ИНС	29	Не применимо	Не применимо	Нет	[159]
Полносвязная ИНС, предобученная в режиме автоэнкодера	40	72	10.7	Обучение с переносом	[160]
Графовая ИНС	40	95	12.9	Обучение с переносом	[161]
Трехмерная ИНС	44	Нет данных	Нет данных	Обучение с переносом	[162]

Отдельно стоит отметить, что, хотя большая часть опубликованных работ посвящена предсказанию времен удерживания для обращенно-фазовой хроматографии, которая наиболее широко применяется при анализе низкомолекулярных соединений, существуют примеры успешного предсказания времен удерживания в условиях хроматографии гидрофильных взаимодействий (Hydrophilic interaction liquid chromatography, HILIC)[145, 148, 163-165]. Основной проблемой является недостаток данных для обучения сложных моделей. Так, авторы программного пакета Retip[148] для обучения моделей использовали набор данных по удерживанию 880 молекул, времена удерживания которых находились в диапазоне 1.5-10.5 мин, при этом наименьшее среднее отклонение составило порядка 47 с. Retip включает несколько методов МО — ГБ, ИНС, случайный лес.

Существенно повысить точность предсказаний удалось при использовании метода обучения с переносом, с предварительным обучением графовой ИНС на искусственном наборе 320 000 молекул, времена удерживания которых были предсказаны моделью, предложенной в работе[148]. Далее, полученная модель была до-обучена на экспериментальном наборе небольшого размера. В итоге, среднее отклонение снизилось до 39 с.

1.5.3 Применение методов машинного обучения для предсказания масс-спектральных характеристик

При идентификации низкомолекулярных химических соединений в нецелевом хромато-масс-спектрометрическом анализе решающую роль играет степень совпадения экспериментально измеренных масс-спектров компонентов с библиотечными спектрами молекул кандидатов. В ГХ-МС в основном проводится сопоставление первичных масс-спектров, полученных при электронной ионизации. В ЖХ-МС, где в основном используются «мягкие» методы ионизации, для структурного анализа используются вторичные масс-спектры, получаемые при диссоциации протонированных или депротонированных молекул вследствие соударений с молекулами инертного газа. Тем не менее, диссоциация, индуцируемая соударениями, находит применение и в газовой хромато-масс-спектрометрии для решения специфических задач. В любом случае, использование первичных или вторичных масс-спектров для идентификации требует наличия спектра, измеренного для образца сравнения известного состава. В разделе 1.1 показано, что существующие библиотеки не покрывают большую часть химического разнообразия низкомолекулярных соединений, что приводит к отсутствию идентификации до 98% компонентов образца, например в нецелевом метаболомном анализе[166]. Даже среди молекул, вовлеченных в известные метаболические пути, более половины не имеют измеренных спектров в библиотеках[167]. При этом, предсказание масс-спектров является более сложной задачей. Основными методами моделирования масс-спектров долгое время являлись квантово-химические расчеты, и статистические методы. Однако, с пополнением библиотек и развитием глубокого обучения, появились первые подходы, основанные на методах МО.

Предсказание масс-спектров электронной ионизации рассматривалось как задача многоцелевой регрессии, где вектор зависимой переменной представлял собой вектор интенсивностей сигнала при всех значениях m/z из диапазона сканирования. Данное представление масс-спектров электронной ионизации оправдано ввиду того, что подавляющее большинство библиотечных масс-спектров измерено на приборах с квадрупольными масс-анализаторами и имеют целочисленное представление m/z . В качестве вектора независимых переменных в работе[168] использовали аддитивные фрагментные дескрипторы, в работе[169]

представление молекулы в виде графа. Обучение моделей проводили по данным из библиотеки NIST различных версий. В качестве функции потерь использовали взвешенное среднее квадратичное отклонение.

Оценку качества предсказаний проводили с помощью взвешенной косинусной меры сходства масс-спектров. Показано, что с помощью предложенных моделей можно создавать расчетные библиотеки, поиск по которым экспериментальных масс-спектров с использованием взвешенной косинусной меры сходства масс-спектров более чем в 90% случаев помещает корректную молекулу в список первых десяти наиболее вероятных кандидатов.

Опосредовано МО использовалось в подходе к моделированию масс-спектров электронной ионизации через вероятностную модель Competitive Fragmentation Model (CFM-EI), где обучаемая ИНС позволяла оценить вероятность перехода между состояниями цепи Маркова, которой описывался процесс фрагментации. Хотя этот подход и уступает по точности, и производительности вышеописанным методам прямого предсказания масс-спектров, это был первый метод предсказания интенсивности сигналов в спектрах электронной ионизации, не использующий квантово-химические вычисления, и позволявший высокопроизводительное моделирование больших библиотек. Реализация CFM для моделирования вторичных масс-спектров является единственным примером применения МО для предсказания масс-спектров, полученных при диссоциации, индуцированной соударениями.

Среди других методов моделирования спектров нельзя не упомянуть методы, основанные на квантово-химических расчетах. Их разработка применительно к спектрам электронной ионизации началась фактически одновременно с широким внедрением приборов ГХ-МС, оснащенных источником электронной ионизации, но за последнее время эти методы были существенно усовершенствованы как с точки зрения точности предсказаний, так и с точки зрения вычислительной производительности, которая всегда была главным недостатком квантово-химических расчетов. Наиболее распространенным программным обеспечением для моделирования спектров электронной ионизации является пакет Quantum Chemistry Electron Ionization Mass Spectra (QCEIMS)[170], который был протестирован на относительно больших наборах данных, и показал высокое сходство между рассчитанными и экспериментальными спектрами. Одна из его модификаций позволяет предсказывать также вторичные масс-спектры в условиях диссоциации, индуцированной соударениями[171]. Тем не менее, основным недостатком квантово-химических методов по-прежнему является низкая производительность, на расчет одного спектра в среднем требуется несколько десятков часов.

Методы предсказания, основанные на эмпирических правилах, довольно широко распространены, хотя обычно доступны в составе коммерческого программного обеспечения

(ACD Labs, Mass Frontier), ввиду трудоемкости процесса выявления закономерностей по известным масс-спектрам. Комбинаторные методы оценивают все возможные пути фрагментации молекулы, с последующим ранжированием образующихся фрагментов по тем или иным критериям. Наиболее известное программное обеспечение, в котором реализован комбинаторный подход, MetFrag[172], ранжирует фрагменты по оценочной энергии разрываемой связи, кроме того, «поощряя» при фрагментации образование устойчивых нейтральных потерь.

1.5.4 Применение методов машинного обучения в спектрометрии ионной подвижности

Метод СИП основан на разделении ионов под действием электрического поля при их движении в среде инертного газа. Ускорение ионов в электрическом поле E уравновешивается их замедлением при столкновении с молекулами газа, и как результат, ионы движутся с постоянной скоростью v_d , пролетая при этом ячейку длиной l за время t_d . Ионная подвижность K при этом выражается как:

$$K = \frac{t_d}{lE}.$$

Хотя время t_d зависит в том числе от трехмерной структуры иона, оно будет отличаться при измерении на различных приборах. Поэтому, в качестве основной характеристики молекулярной структуры используется другой параметр, сечение столкновений, который меньше зависит от условий эксперимента, и связан с величиной ионной подвижности следующим уравнением:

$$K = \frac{3}{16} \sqrt{\frac{2\pi}{\mu k_b T}} \times \frac{ze}{N\Omega},$$

где $\mu = Mm/(M+m)$, m и M – молекулярные массы иона и газа, ze – заряд иона, T – температура, k_b – постоянная Больцмана, N – плотность газа, Ω – сечение столкновений.

Спектрометрия ионной подвижности давно известна как самостоятельный метод исследования, хотя и с ограниченной сферой применения. Однако, сочетание спектрометров ионной подвижности и масс-спектрометров, открывает дополнительные возможности для нецелевого анализа. Основной способ применения данных спектрометрии при идентификации химических соединений аналогичен применению времен удерживания, и сводится к сравнению экспериментально определенных значений CCS с библиотечными или измеренными для образцов сравнения известного состава.

Приборы, сочетающие спектрометры ионной подвижности и масс-спектрометры стали массово внедряться в аналитическую практику в последнее десятилетие, и потому доступность экспериментальных данных по CCS низкомолекулярных соединений существенно ниже, чем по временам удерживания[173], поэтому, методы предсказаний этих значений активно развиваются.

Было предложено использование метода опорных векторов для предсказаний CCS низкомолекулярных соединений для нецелевого метаболомного[174, 175] и липидомного анализа[176]. Предложенные модели MetCCS и LipidCCS, обученные с использованием выборок, содержащих несколько сотен экспериментальных значений, позволяют предсказывать CCS со средним относительным отклонением порядка 1%. Данные модели реализованы в виде Web-приложений с графическим интерфейсом, что позволяет использовать их широкому кругу специалистов. В работе [177] были предсказаны значения CCS для пестицидов, с применением ИНС. Несмотря на довольно высокое значение медианного относительного отклонения, определенного по независимой тестовой выборке, использование предсказанных значений позволило существенно повысить уверенность при идентификации 10 пестицидов в образцах шпината. Одновременное предсказание CCS и времен удерживания было проведено в работе [178] при обучении ИНС в режиме многоцелевой регрессии. Ожидалось, что такой подход позволит повысить точность предсказаний обоих параметров по сравнению с индивидуальными моделями [179, 180]. Однако, улучшить результат удалось только при моделировании времен удерживания. В целом, первые работы по моделированию CCS низкомолекулярных соединений можно охарактеризовать довольно высокой точностью (средние относительные отклонения составляли несколько процентов), однако они были обучены и валидированы на небольших выборках, зачастую без использования кросс-валидации, а потому вряд ли могут быть достаточно универсальными.

Как и в случае других характеристик, методы моделирования CCS развиваются в направлении создания универсальных моделей с применением глубокого обучения. Так, в работе [181] были применены сверточные ИНС, обученные на экспериментальных значениях CCS 2400 молекул. Полученная модель позволяла предсказывать CCS с медианным относительным отклонением в 2.7%. Как и в других задачах моделирования химических свойств, основной преградой является недостаток обучающих примеров. В работе [182] для решения этой задачи предложен подход с применением обучения с переносом. При разработке DarkChem использовали ИНС сложной архитектуры «вариативный авто кодировщик» (Variational Autoencoder, VAE) с дополнительным блоком регрессии, и трехступенчатая схема обучения. На первом этапе, ИНС, состоящая из блока кодирования и блока декодирования, обучалась на наборе SMILES представления молекул из библиотеки PubChem, и молекулярные массы этих

молекул, на втором этапе, предобученные блоки кодирования и регрессии дообучались на большом наборе данных, содержащих порядка 700 000 значений CCS, рассчитанных методами квантовой химии[183]. На последнем этапе веса модели подстраивали по обучающей выборке из 500 значений CCS, определенных экспериментально. Данный подход обеспечил возможность с высокой производительностью предсказывать CCS низкомолекулярных соединений, со средним относительным отклонением в 2.7%.

1.6 Идентификации химических соединений в нецелевом хромато-масс-спектрометрическом анализе с применением характеристик, предсказанных с помощью машинного обучения

При идентификации компонентов сложных проб в хромато-масс-спектрометрическом анализе измеренные характеристики сравниваются со значениями из библиотек или измеренными для образцов сравнения известного состава. Сравнение позволяет отфильтровать ложноположительные определения при установлении пороговых значений, либо ранжировать кандидатов по величине отличия. Установление пороговых значений проводят с учетом условий экспериментальной вариабельности характеристики, ответственности при ложном определении и других факторов. При использовании предсказанных величин вместо экспериментальных значений, вопрос установления порогов является довольно сложным, и требует учитывать не только оценочную точность предсказаний, но и разнообразие молекулярных структур в тестовой выборке, по которой проводили оценку. Другой вопрос, который нужно решить, это определение списка молекул, для которых будут предсказаны значения, по которым будет проведен библиотечный поиск. Очевидно, что при использовании узких предметных баз данных, например, о лекарственных средствах, метаболитах, пищевых добавках, количество изомерных кандидатов, которые необходимо ранжировать или отфильтровывать по пороговым значениям будет меньше, чем при поиске среди общехимических библиотек PubChem или ChemSpider. При этом, в первом случае, эффективность фильтрации ложноположительных определений будет выше, но выше и вероятность ложноотрицательного результата. Так, например в работе [29] показано, что при увеличении пространства поиска в 3 раза, количество корректных положительных результатов оказавшихся в первых трех местах при ранжировании по предсказанному времени удерживания сокращается с 70% до 45%.

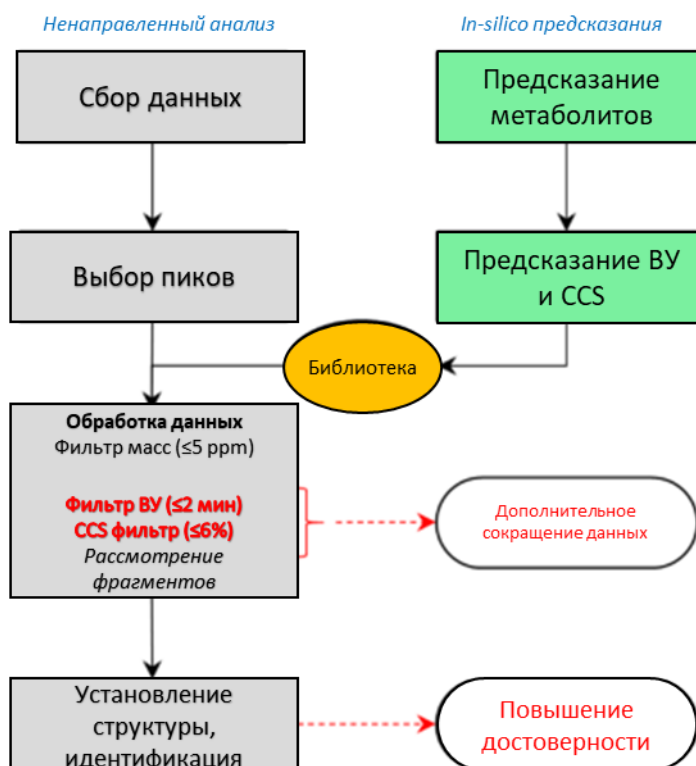


Рисунок 18. Схема использования предсказанных значений молекулярных характеристик при идентификации химических соединений в нецелевом анализе[184].

В работе [184] предложен подход по идентификации продуктов трансформации инсектицида пиримифос-метила в мясе лосося в нецелевом хромато-масс-спектрометрическом анализе с учетом предсказанных значений CCS и времен удерживания. Библиотека возможных продуктов трансформации была сгенерирована с помощью специализированного программного обеспечения, а пороговые значения были установлены с учетом оценки точности модели предсказания с помощью 95% доверительных интервалов. Общая схема изображена на рисунке 18. Данный подход позволил не только идентифицировать известные метаболиты пиримифос-метила с повышенной уверенностью, но и аннотировать несколько новых вероятных метаболитов.

Выводы из обзора литературы. Постановка цели и задач исследования.

Из проведенного обзора литературы, можно сделать вывод, что предсказание характеристик низкомолекулярных соединений методами МО для нецелевого хромато-масс-спектрометрического анализа представляет большой интерес и широко обсуждается в литературе. Методы МО уже применяли для моделирования времен и ИУ в жидкостной и газовой хроматографии, сечения столкновений в СИП, а также масс-спектров электронной ионизации. Однако, большинство работ посвящено созданию моделей с ограниченной сферой применения, обученных на небольших выборках соединений родственных классов. Хотя такой подход может

оказаться более точным при моделировании характеристик молекул одного ряда (например, гомологов), общая тенденция заключается в создании более универсальных методов предсказаний, в основном с применением глубоких ИНС. Этому способствует накопление экспериментальной информации в специализированных библиотеках. Тем не менее, точность существующих моделей пока ограничена, и потому в диссертационной работе ставилась цель разработать новые подходы к предсказанию молекулярных характеристик, используемых при идентификации химических соединений в нецелевом хромато-масс-спектрометрическом анализе.

ГЛАВА 2. Оборудование, материалы, техника эксперимента

2.1 Оборудование и материалы

В работе использовали следующее аналитическое оборудование:

- 1) ВЭЖХ-МСВР систему, состоящую из гибридного масс-спектрометрического детектора с квадрупольным масс-фильтром и орбитальной ионной ловушкой (Orbitrap) Thermo Scientific QExactive (Германия), оснащенного источниками электрораспылительной ионизации Thermo Scientific HESI-II Probe, Thermo Scientific Nanospray Flex Ion Source (США), а также системой Spectrograph MALDI/ESI Injector interface (США), и жидкостного хроматографа Thermo Scientific Ultimate 3000 RSLC Nano (США).
- 2) ВЭЖХ-МСВР систему, состоящую из гибридного масс-спектрометрического детектора с квадрупольным масс-фильтром и орбитальной ионной ловушкой (Orbitrap) Thermo Scientific QExactive (Германия), оснащенного источниками электрораспылительной ионизации Thermo Scientific HESI-II Probe, Thermo Scientific Nanospray Flex Ion Source (США), а также системой Spectrograph MALDI/ESI Injector interface (США), и жидкостного хроматографа Waters ACQUITY I-Class UPLC system (Великобритания).

Разделение проводили на следующих хроматографических колонках:

- 1) ACQUITY UPLC BEH C18, (100 × 2.1 мм), диаметр зерна сорбента 1.7 мкм (Waters Corp, США).
- 2) ACQUITY UPLC BEH C8, (100 × 2.1 мм), диаметр зерна сорбента 1.7 мкм (Waters Corp, США).
- 3) ACQUITY UPLC HSS T3, (100 × 0.075 мм), диаметр зерна сорбента 1.8 мкм (Waters Corp, США).
- 4) ACCLAIM RepMap C18 (150 × 0.075 мм), диаметр зерна сорбента 2 мкм (Thermo Scientific™, США).

В работе использовали следующее оборудование для пробоподготовки:

- 1) Взвешивание навесок проводили на весах Explorer Pro (Ohaus Corporation, США).
- 2) Инкубацию образцов проводили в программируемом твердотельном термостате «Гном» (ДНК-Технологии, Россия), и в программируемом горизонтальном шейкере TS-100 (Biosan, Латвия)

- 3) Для центрифугирования образцов использовали центрифугу 5804R (Eppendorf, Германия).
- 4) Для отбора точной аликвоты использовали автоматические дозаторы Ленпипет 10-100 мкл, 20-200 мкл, 100-1000 мкл с пределом допускаемой погрешности измерения не более $\pm 5\%$ (Thermo Scientific™, США) и наконечники необходимых объемов.

2.2 Выполнение анализа

2.2.1 Определение времен удерживания для получения обучающей и тестовой выборки в условиях разделения в нано-поточной хроматографии

Исходные растворы с концентрацией 10 мг/мл соединений из обучающей и тестовой выборки были приготовлены растворением соответствующих навесок в ДМСО. Далее исходные растворы смешивали и доводили ДМСО для получения смеси с концентрацией индивидуальных компонентов 200 мкг/мл. Далее эта смесь последовательно разбавлялась водой до концентрации индивидуальных компонентов 0.5 мкг/мл. Аликвоту смеси далее анализировали методом ВЭЖХ-МС.

ВЭЖХ-МС анализ для определения времен удерживания в нано-поточной хроматографии проводили в режиме предварительного онлайн концентрирования, с использованием концентрирующей колонки Acclaim PerMap C18 (20 × 0.1 mm) с диаметром сорбента 5 мкм. Хроматографическое разделение проводили на колонке ACCLAIM PerMap C18 (150 × 0.075 mm), с диаметром зерна сорбента 2 мкм, в градиентном режиме при постоянной скорости потока 300 нл/мин, термостатировании колонки при 30°C. В качестве подвижной фазы А использовали 2% раствор ацетонитрила в воде с добавкой 0.1% муравьиной кислоты, в качестве подвижной фазы В использовали 80% раствор ацетонитрила в воде с добавкой 0.1% муравьиной кислоты. Подвижная фаза А также использовалась в качестве загрузочного буфера. Разделение проводили в следующем градиенте: 0-7 мин, 5% В; 7-40 мин, увеличение от 5% до 90% В, 40-47 мин 90% В; 47-50 мин уменьшение от 90% до 5% В, 50-60 мин 5% В. Регистрацию ионов проводили в режиме детектирования полного ионного тока при разрешении 140 000, и в режиме информационно-зависимого сбора данных (Data dependent acquisition, DDA), с фрагментацией 5 наиболее интенсивных ионов со ступенчатой энергией соударений 10; 30; 45. Идентификацию соединений проводили по точной массе, с подтверждением путем сравнения измеренных спектров МС/МС с библиотечными, по библиотеке mzCloud. Образцы анализировали в двух повторностях, с усреднением полученных значений времен удерживания.

2.2.2 Определение времен удерживания для получения внутрилабораторной обучающей выборки

Для получения обучающей выборки для модели предсказания времен удерживания анализировали растворы индивидуальных веществ и их смеси с концентрацией в диапазоне 0.1-1 мкг/мл. Хроматографическое разделение проводили на колонке AQUITY UPLC BEH C18, (100 × 2.1 мм) с диаметром зерна сорбента 1.7 мкм в градиентном режиме при постоянной скорости потока 400 мкл/ мин, термостатировании колонки при 60°C. В качестве подвижной фазы А использовали 0.1% раствор муравьиной кислоты в воде, в качестве подвижной фазы В использовали 0.1% раствор муравьиной кислоты в ацетонитриле. Разделение проводили в следующем градиенте: 0-5 мин, 5% В; 5-25 мин, увеличение от 5% до 75% В, 25-26 мин, увеличение от 75% до 100% В; 26-33 мин 100% В, 33-35 мин уменьшение 100%-5% В, 35-40 мин 5%В. Объем вводимой пробы составлял 4 мкл. Масс-спектрометрический анализ проводили при ионизации электрораспылением в режиме образования положительных ионов с применением источника ионизации Thermo Scientific HESI-II Probe, при напряжении распыления 4.5 кВ, температуре испарителя 350°C. Расходы газа-распылителя и газа-осушителя были установлены на 20 и 45 единиц соответственно. Температура десольватирующего капилляра составляла 320°C, напряжение на S-линзе 60 В. Регистрацию ионов проводили в режиме детектирования полного ионного тока в диапазоне масс 100-1200 Да при разрешении 70 000, и в режиме информационно-зависимого сбора данных (Data dependent acquisition, DDA), с фрагментацией 10 наиболее интенсивных ионов со ступенчатой энергией соударений 10; 35; 60. Окно изоляции составляло 1.2 Да, разрешение в режиме MS/MS 17 500. Идентификацию соединений проводили по точной массе, с подтверждением путем сравнения измеренных спектров MS/MS с библиотечными, по библиотеке mzCloud. Образцы анализировали в двух повторностях, с усреднением полученных значений времен удерживания.

2.2.3 Пробоподготовка образцов мочи для изучения селективности изотопного обмена $^{16}\text{O}/^{18}\text{O}$

Для получения модельного образца для изучения селективности изотопного обмена $^{16}\text{O}/^{18}\text{O}$ к аликвоте 45 мкл мочи человека добавляли 5 мкл раствора смеси исследуемых соединений в ацетонитриле с концентрацией индивидуальных компонентов 4 мкг/мл и перемешивали на вортексе. К образцу мочи добавляли 150 мкл ацетонитрила, охлажденного до 4°C, перемешивали, и выдерживали при -20°C в течение 1 ч. После центрифугирования при 4°C и 10 000 об/мин в течение 15 мин отбирали 180 мкл надосадочной, упаривали досуха в вакуумном концентраторе без нагревания, и перерастворяли в 50 мкл 30% раствора ацетонитрила в воде или

H218O. Затем образец инкубировали в течение 24 ч при температуре 37°C или 95°C, после чего анализировали методом ВЭЖХ-МС.

Хроматографическое разделение проводили на колонке AQUITY UPLC BEH C18, (100 × 2.1 мм) с диаметром зерна сорбента 1.7 мкм в градиентном режиме при постоянной скорости потока 400 мкл/ мин, термостатировании колонки при 60°C. В качестве подвижной фазы А использовали 0.1% раствор муравьиной кислоты в воде, в качестве подвижной фазы В использовали 0.1% раствор муравьиной кислоты в ацетонитриле. Разделение проводили в следующем градиенте: 0-5 мин, 5% В; 5-25 мин, увеличение от 5% до 75% В, 25-26 мин, увеличение от 75% до 90% В; 26-33 мин 90% В, 33-35 мин уменьшение 100%-5% В, 35-40 мин 5%В. Объем вводимой пробы составлял 4 мкл. Масс-спектрометрический анализ проводили при ионизации электрораспылением в режиме образования положительных ионов с применением источника ионизации Thermo Scientific HESI-II Probe, при напряжении распыления 4.5 кВ, температуре испарителя 350°C. Расходы газа-распылителя и газа-осушителя были установлены на 20 и 45 единиц соответственно. Температура десольватирующего капилляра составляла 320°C, напряжение на S-линзе 60 В. Регистрацию ионов проводили в режиме детектирования полного ионного тока в диапазоне масс 100-1200 Да при разрешении 70 000, и в режиме информационно-зависимого сбора данных (Data dependent acquisition, DDA), с фрагментацией 10 наиболее интенсивных ионов со ступенчатой энергией соударений 10; 35; 60. Окно изоляции составляло 1.2 Да, разрешение в режиме MS/MS 17 500. Идентификацию соединений проводили по точной массе, с подтверждением путем сравнения измеренных спектров MS/MS с библиотечными, по библиотеке mzCloud.

2.3 Программное обеспечение

Регистрацию хроматограмм и обработку данных проводили при помощи программного обеспечения Thermo Scientific™ Xcalibur™ Software (версия 4.0) и Thermo Scientific Compound Discoverer 3.3.

Реализация предложенных подходов проводилась с использованием языков программирования Python 3 и R. Для построения кусочно-линейных функций использовался пакет *segmented*[185] для языка программирования R. Библиотека RDKit использовалась для работы молекулярными структурами, включая генерацию SMILES, InChI, InChiKey, расчета фрагментных дескрипторов, обработку шаблонов SMILES arbitrary target specification (SMARTS)[186]. Расчет физико-химических дескрипторов проводили средствами библиотеки MORDRED. Методы МО были реализованы с использованием библиотек *scikit-learn*, *XGBoost*. Методы глубокого обучения реализовывались с применением функционала библиотек *Keras*, *FastAI*, *TensorFlow*, *PyTorch*.

ГЛАВА 3. Применение машинного обучения для предсказания времен удерживания в жидкостной хромато-масс-спектрометрии¹

Целью первой части работы была разработка новых подходов по предсказанию времен удерживания в жидкостной хромато-масс-спектрометрии. Время удерживания в жидкостной хромато-масс-спектрометрии параметром, дополнительным к масс-спектральным характеристикам и может применяться для повышения уверенности в результатах идентификации или для фильтрации ложноположительных определений. Однако информация по удерживанию соединений в жидкостной хроматографии ограничена ввиду большого разнообразия применяемых в анализе малых молекул условий разделения и высокой стоимости образцов сравнения. Недостаток данных может быть отчасти компенсирован применением расчетных значений времен удерживания. Наиболее перспективные методы для моделирования времен удерживания основаны на алгоритмах МО, так как они не требуют установления физико-химических механизмов удерживания, характеризуются высокой производительностью, а при наличии разнообразных обучающих выборок – высокой точностью и универсальностью.

Разработка моделей МО обычно включает в себя несколько этапов, а именно подготовку обучающей выборки, выбор архитектуры модели, выбор гиперпараметров модели, обучение, кросс-валидацию. При моделировании времен удерживания важно также рассмотреть вопрос переносимости предсказаний между различными хроматографическими системами. В диссертационной работе предложено три новых подхода к предсказанию времен удерживания.

3.1 Предсказание времен удерживания в жидкостной хроматографии методом градиентного бустинга

3.1.1 Построение модели предсказания времен удерживания по данным библиотеки METLIN SMRT

Одной из задач, решаемых в диссертационной работе была разработка модели предсказания времен удерживания, обученной на библиотеке METLIN SMRT, которая будет более точной, чем ИНС, использованная в оригинальной публикации[29]. Повышение точности необходимо, т.к. для практического применения необходимо построение моделей пересчета на другие хроматографические системы. Для построения моделей пересчета необходимо измерить

¹ При подготовке данной главы диссертации использованы публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ отражены основные результаты, положения и выводы исследования. **Osipenko S.**, Bashkirova I., Sosnin S., Kovaleva O., Fedorov M., Nikolaev E., Kostyukevich Y. Machine learning to predict retention time of small molecules in nano-HPLC // Analytical and Bioanalytical Chemistry. – 2020. – Т. 412, № 28. – С. 7767-7776. (Импакт-фактор Web of Science – 4.478, Q1) 50%; **Osipenko S.**, Botashev K., Nikolaev E., Kostyukevich Y. Transfer learning for small molecule retention predictions // Journal of Chromatography A. – 2021. – Т. 1644. – С. 462119. (Импакт-фактор Web of Science – 4.601, Q1) 50%; **Osipenko S.**, Nikolaev E., Kostyukevich Y. Retention Time Prediction with Message-Passing Neural Networks // Separations. – 2022. – Т. 9, № 10. – С. 291 (Импакт-фактор Web of Science – 3.344, Q3) 75%.

времени удерживания набора молекул в обоих условиях разделения, и точность пересчета будет зависеть в том числе от размера данного набора. Однако, так как библиотека METLIN SMRT содержит не очень много веществ, распространенных в аналитических лабораториях, то крайне желательно повышать точность первого этапа. Для моделирования был выбран метод ГБ, реализованный в библиотеке XGBoost[46]. Эта реализация характеризуется высокой точностью, и скоростью обучения моделей при использовании графических процессоров.

В качестве основной обучающей выборки использовали набор данных METLIN SMRT. Этот набор содержит данные об удерживании 80 038 низкомолекулярных соединений в условиях обращено-фазового разделения.

В работе сравнивали два вида молекулярных дескрипторов. Фрагментные круговые дескрипторы[93] кодируют молекулу в виде бит-вектора, где 1 отвечает наличию в молекуле определенного фрагмента, а 0 – отсутствию. Фрагменты определяются следующим образом. Для каждого атома молекулы строится условный круг содержащий атомы, отстоящие от центрального на r , где r – радиус этого круга (Рисунок 19). При этом радиус и длина бит-вектора являются гиперпараметрами, однако обычно используют радиус 3-4 и максимально возможную длину вектора, чтобы избежать пересечений. Круговые фрагментные дескрипторы могут быть рассчитаны в библиотеке RDKit по алгоритму Моргана. Кроме того, в работе использовали библиотеку Mordred для вычисления большого набора из 1613 физико-химических дескрипторов[97].

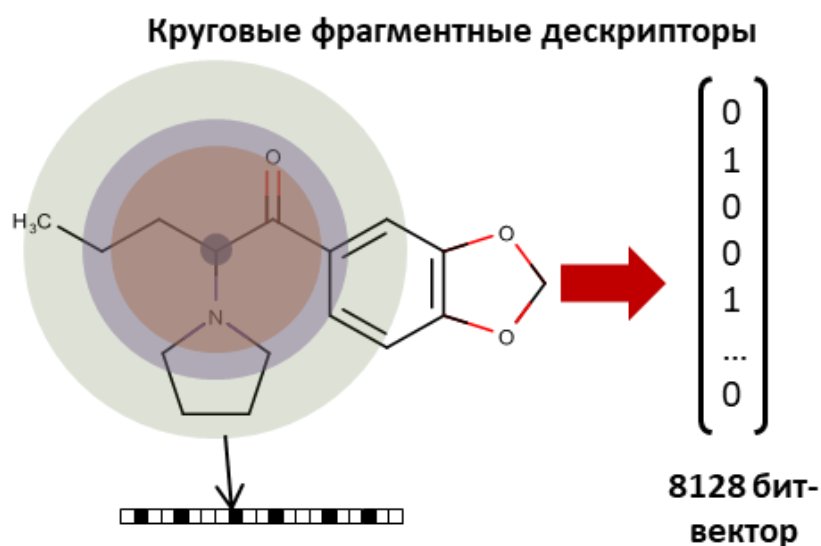


Рисунок 19. Схема определения круговых фрагментных дескрипторов

На рисунке 20 приведено распределение времен удерживания в наборе данных METLIN SMRT, которое является бимодальным, т.к. библиотека содержит неустойчивые молекулы. В оригинальной работе было показано, что включение неустойчивых молекул в обучающую и

тестовую выборки отрицательно влияет на общую точность предсказаний. Авторы предложили исключать неударживаемые молекулы из рассмотрения.

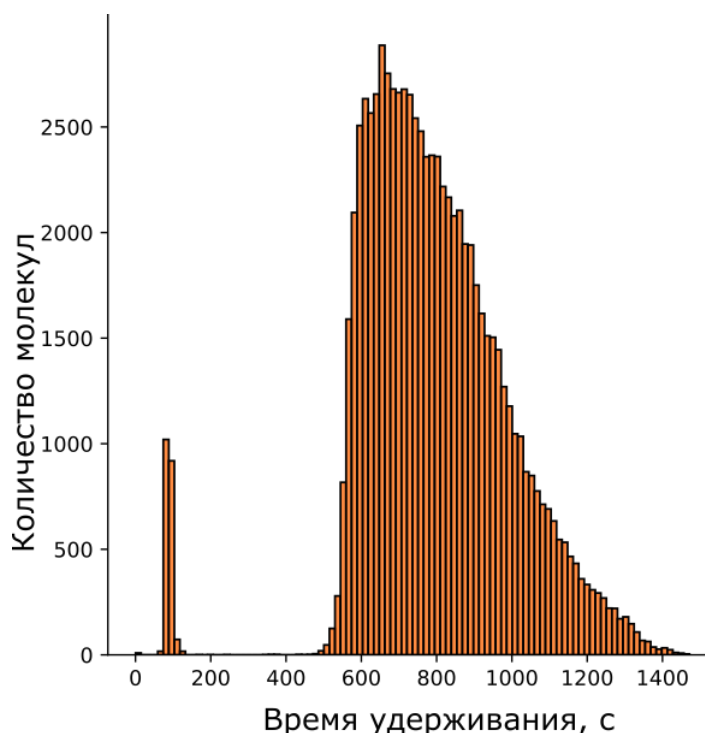


Рисунок 20. Распределение времен удерживания молекул из библиотеки METLIN SMRT.

В ходе предварительного исследования в диссертационной работе также было обнаружено, что при исключении из рассмотрения неударживаемых молекул среднее относительное отклонение снижается с 55% до 6.25%, хотя медианное отклонение, более устойчивое к выбросам, практически не изменилось. Однако, исключение из рассмотрения неударживаемых молекул может привести к тому, что модель будет систематически завышать предсказанные времена удерживания для неударживаемых молекул, и как результат, к ошибкам при идентификации. Это может быть существенной проблемой в нецелевом метаболомном анализе, где многие целевые аналиты полярны, и характеризуются слабым удерживанием в условиях обращенно-фазовой хроматографии. Кроме того, разрабатываемый подход планировалось применять в нано потоковой ВЭЖХ с предварительным концентрированием. В таких условиях неударживаемые молекулы принципиально не достигают детектора. Для учета этой проблемы, в работе предложен двухэтапный подход, включающий не только регрессионную модель предсказания времен удерживания, но и модель бинарной классификации молекул на ударживаемые и неударживаемые. Общая схема предложенного подхода представлена на рисунке 21.

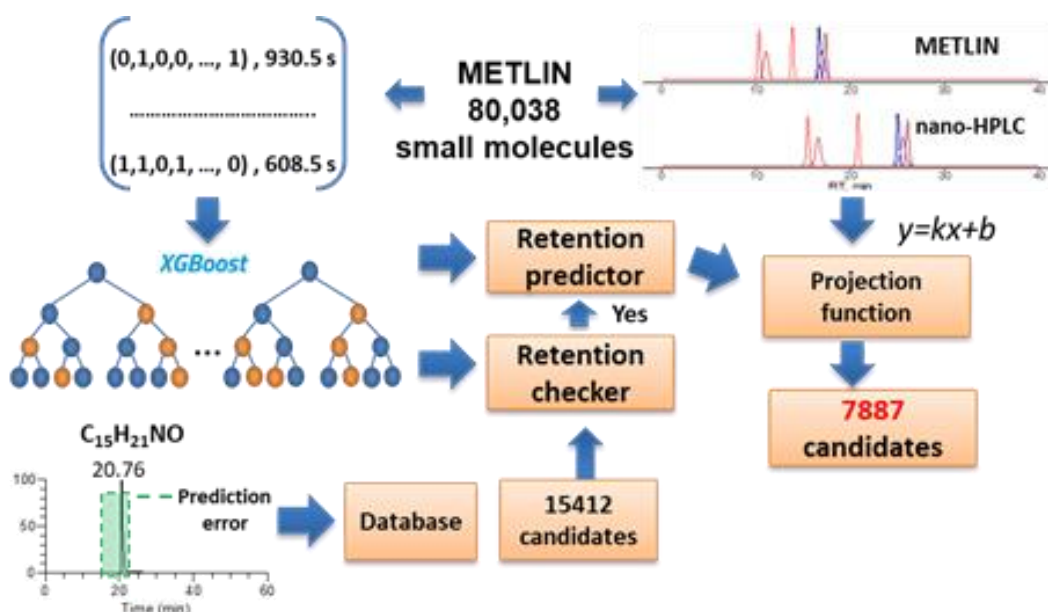


Рисунок 21. Общая схема предложенного подхода по использованию градиентного бустинга для предсказания времен хроматографического удерживания и идентификации химических соединений

Основными гиперпараметрами модели ГБ являются глубина дерева, количество деревьев и скорость обучения. Классы удерживаемых и недерживаемых молекул в библиотеке METLIN SMRT четко разделимы, и модели бинарной классификации, обученные со стандартными гиперпараметрами (глубина дерева 4, количество деревьев 1000, скорость обучения 1), показали хорошие результаты в режиме кросс-валидации. Значения параметра ROC AUC для моделей, обученных на фрагментных дескрипторах и на физико-химических дескрипторах из библиотеки Mordred составили 0.961 ± 0.004 и 0.966 ± 0.006 соответственно.

В случае регрессионных моделей требовался более тщательный подбор гиперпараметров. Для выбора максимальной глубины дерева и скорости обучения проводили подбор с помощью алгоритма полного перебора по сетке параметров (Grid Search), сетка параметров приведена в таблице 5.

Таблица 5. Сетка значений гиперпараметров при их выборе методом полного перебора (Grid Search)

Максимальная глубина дерева (Max tree depth)	Скорость обучения (Learning rate)
2	0.01
4	0.025
6	0.05
8	0.1
10	0.2

Выбранные гиперпараметры при использовании фрагментных дескрипторов и физико-химических дескрипторов из библиотеки Mordred представлены в таблице 6. После 1000

итераций все модели (и классификации, и регрессии) показывали лишь незначительное уменьшение функции потерь, поэтому для всех моделей использовали 1000 в качестве гиперпараметра общего количества деревьев.

Таблица 6. Выбранные гиперпараметры моделей градиентного бустинга

	Максимальная глубина дерева (Max tree depth)	Скорость обучения (Learning rate)
Фрагментные дескрипторы	6	0.2
Физико-химические дескрипторы Mordred	8	0.05

В результате кросс-валидации моделей ($n=5$) средние абсолютные отклонения составили 45.6 ± 0.4 с, и 47.1 ± 0.4 с при представлениях молекул физико-химическими и фрагментными дескрипторами соответственно. Тем не менее, нужно отметить, что, хотя дескрипторы из библиотеки Mordred позволяют достичь более высокой точности предсказаний, их расчет занимает намного больше времени, чем расчет фрагментных круговых дескрипторов. Это скажется не только на времени создания модели, но и на производительности работы моделей в реальной практике. Например, расчет фрагментных дескрипторов 13 000 молекул занял 5 с, в то время как расчет физико-химических дескрипторов из библиотеки Mordred занял более получаса на той же системе. С учетом того, что среднее абсолютное отклонение снижается всего на 3%, для задач, связанных с обработкой больших данных, модель, построенная по фрагментным дескрипторам, является более предпочтительной. Полученные результаты показывают, что модель на основе ГБ на 10 с точнее предложенной ранее модели на основе ИНС[29].

3.1.2 Пересчет предсказаний на другие хроматографические условия

Для оценки возможности переноса предсказаний между хроматографическими системами в условиях нано-поточной хроматографии были измерены времена удерживания 24 соединений из библиотеки METLIN SMRT и 20 соединений, отсутствующих в ней. Список последних и времена удерживания приведены в таблице 7. Этот набор данных далее обозначается как «nanoHPLC». Кроме того, из библиотеки PredRet были выбраны наборы данных содержащие времена удерживания не менее 100 молекул с валидными InChI, и имеющих пересечение не менее 20 молекул с библиотекой METLIN SMRT (Рисунок 22).

Таблица 7. Состав набора данных папо-HPLC, собранного в работе

Название	Экспериментально измеренное время удерживания, мин	Предсказанное время удерживания, мин	
		По фрагментным дескрипторам	По физико-химическим дескрипторам Mordred
Фенибут	19.96	20.65	22.64
Каптоприл	21.66	22.89	24.17
Фуразолидон	23.3	25.50	27.65
Напроксен	36.96	31.66	34.19
Толперизон	23.82	26.02	24.76
Оксиметазолин	25.99	24.63	27.36
Фенотерол	18.67	19.88	19.45
Ипратропий	20.98	30.17	26.27
Клемастин	32.12	30.13	35.47
Офлоксацин	21.66	21.01	20.23
Лоратадин	31.95	39.16	30.81
Дротаверин	29.1	38.56	32.19
Лоперамид	32.2	32.27	20.75
Ацетаминофен	18.67	19.52	20.96
Кофеин	20.09	25.64	25.08
Дексаметазон	32.71	26.74	29.26
Салициловая кислота	24.67	27.22	25.50
Мирамистин	37.93	38.68	37.55
Пантенол	16.03	22.48	20.40
Фурацилин	22.35	23.92	22.73

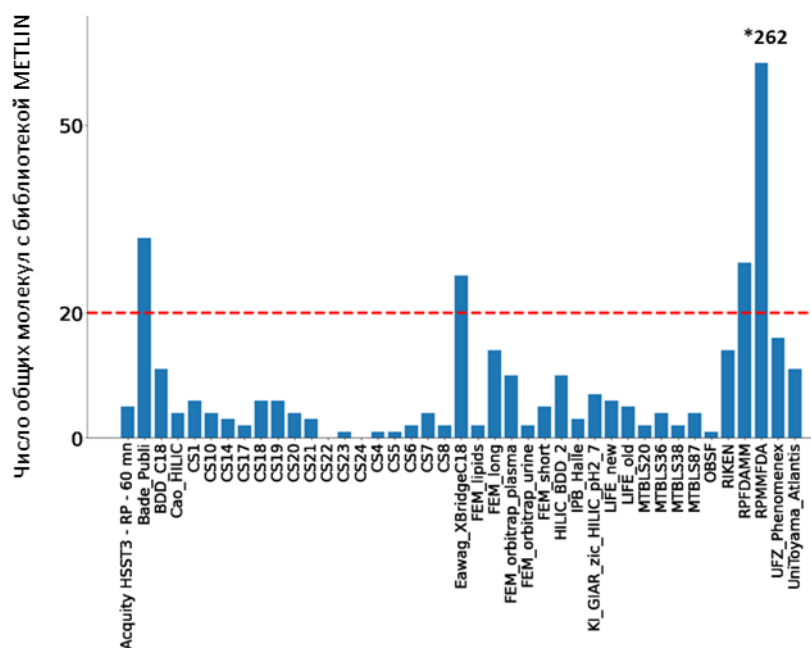


Рисунок 22. Наборы данных из репозитория PredRet[18], имеющие более 20 общих молекул с библиотекой METLIN SMRT[29]

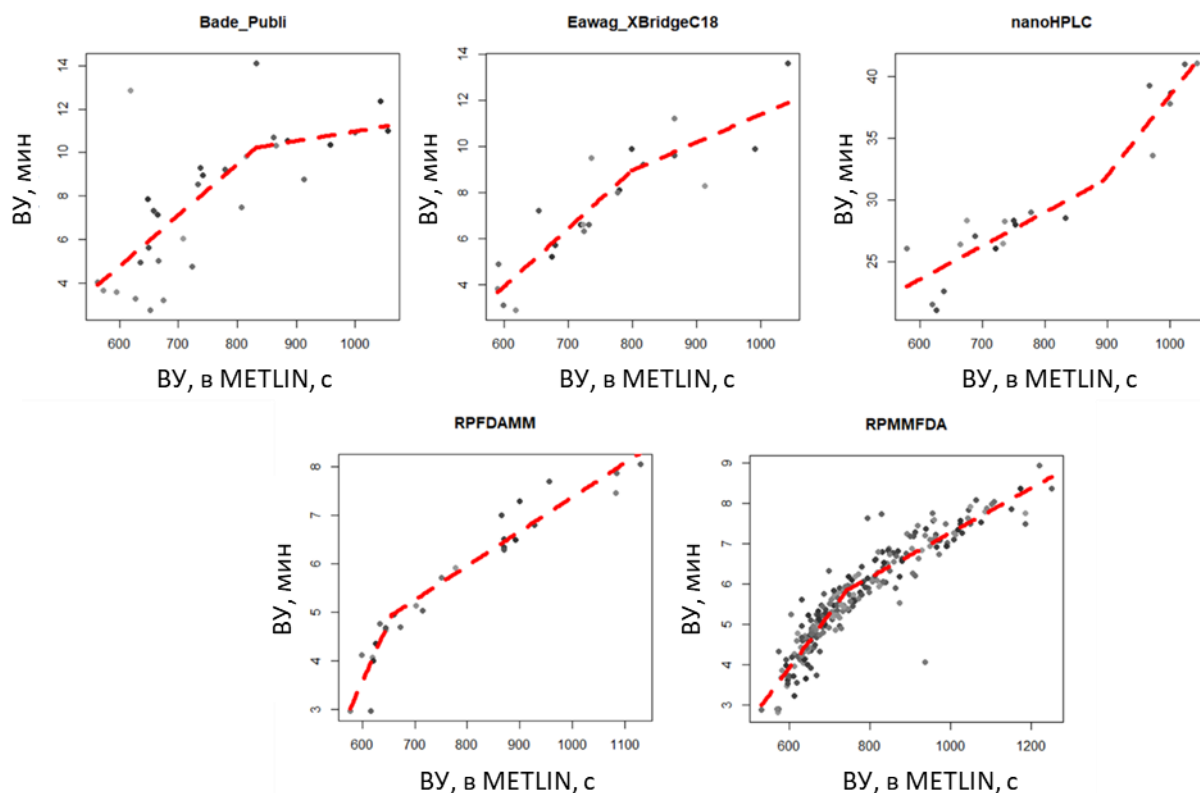


Рисунок 23. Графики кусочно-линейных функций пересчета предсказаний на различные условия разделения.

В качестве модели пересчета была выбрана кусочно-линейная функция с автоматическим выбором точек смены формул. Построение выполнялось с помощью функций пакета *segmented*

языка программирования R[185]. Общий вид моделей для 5 условий разделения приведен на рисунке 23.

Оценку общей точности проводили, оценивая времена удерживания молекул, отсутствующих в библиотеке METLIN SMRT по построенным функциям пересчета. Значения медианного абсолютного отклонения для рассмотренных наборов данных, приведены в таблице 8. Ожидаемо, точность предсказаний после пересчета между хроматографическими системами существенно снизилась. Сравнение результатов предсказаний предложенным методом с результатами индивидуальных моделей, для набора данных Eawag_XBridgeC18[138], свидетельствует о том, что предложенный подход с пересчетом предсказаний уступает моделированию напрямую по имеющимся данным. В первом случае медианное отклонение составило 92 с, во втором 61.3-81.8 с (в зависимости от модели). Однако, в случае с построением моделей пересчета, количество молекул в тестовой выборке существенно превышало 10% от общего количества в наборе данных, использовавшихся при валидации индивидуальных моделей.

Таблица 8. Медианные отклонения предсказанных значений от экспериментальных

Набор данных	Медианное отклонение, мин	
	Фрагментные дескрипторы	Физико-химические дескрипторы Mordred
RPMMFDA	0.85	0.66
RPFDA MM	0.79	0.59
Bade_Publi	1.60	1.23
Eawag_XBridge_C18	1.64	1.54
nanoHPLC	2.89	2.55

3.1.3 Фильтрация ложноположительных определений при идентификации химических соединений в нецелевом скрининге с помощью предложенного подхода

Основной целью предсказания времен удерживания являлось уменьшение пространства поиска при идентификации компонентов сложных образцов в нецелевом хромато-масс-спектрометрическом анализе. Для оценки эффективности предложенных моделей при идентификации химических соединений использовали внутри лабораторный набор данных, а также четыре набора данных из репозитория PredRet. Наборы были разделены на обучающие выборки, состоящие из молекул, общих с библиотекой METLIN SMRT, и тестовые выборки из уникальных молекул. Обучающие выборки использовали для построения моделей пересчета.

Модельную идентификацию химических соединений проводили исходя из допущения об использовании масс-спектрометрии высокого разрешения и возможности установления брутто-формулы по точной измеренной массе, т.е. идентификация сводилась к выбору среди изомерных кандидатов. Для поиска изомерных структур использовали библиотеку PubChem, для всех соединений из тестовых выборок из библиотеки PubChem были выгружены изомерные молекулы, и для всех предсказаны времена удерживания.

В диссертационной работе для установления порогового значения предложено построение ROC кривых. Для этого, для пороговой величины в диапазоне 0-200% с шагом 2.5% были определены доли истинно-положительных (TP), ложноположительных (FP), истинно-отрицательных (TN) и ложноотрицательных определений (FN). По построенным ROC кривым определяли значение порога, дающее максимальную долю истинно-положительных определений при минимальной доле ложноположительных определений. При фильтрации предсказанное время сравнивалось с экспериментальным, и если отличие превышало пороговое значение, то кандидат расценивался как ложноположительный. Площади под ROC кривыми находились в диапазоне 0.57-0.62. Определенные значения для внутри лабораторного набора данных, а также наборов Vade_Publi, RPMMFDA, RPFDAММ и Eawag_XBridgeC18 составили 27.5%, 30%, 25%, 12.5% соответственно. В среднем, при выбранных пороговых значениях, удалось отфильтровать 31%, 43%, 49%, 68% и 40% ложноположительных кандидатов от общего числа изомеров в PubChem. Однако, при выбранных пороговых значениях доля ложноотрицательных определений была высока, и для одного из наборов превысила 50%. Поэтому, пороговые значения были переопределены, с условием минимальной доли истинно-положительных определений в 80%. Обновленные пороговые значения составили 22.5%, 45%, 35%, 37.5% и 50%, и позволили отфильтровать 42%, 28%, 28%, 25%, 32% для внутри лабораторного набора данных, а также наборов Vade_Publi, RPMMFDA, RPFDAММ и Eawag_XBridgeC18 соответственно (Рисунок 24). Кроме того, можно отметить более высокую эффективность моделей удерживания при идентификации слабо или сильно удерживаемых соединений. В качестве примера можно привести результаты фильтрации изомеров оксиметазолина и парацетамола (Рисунок 25).

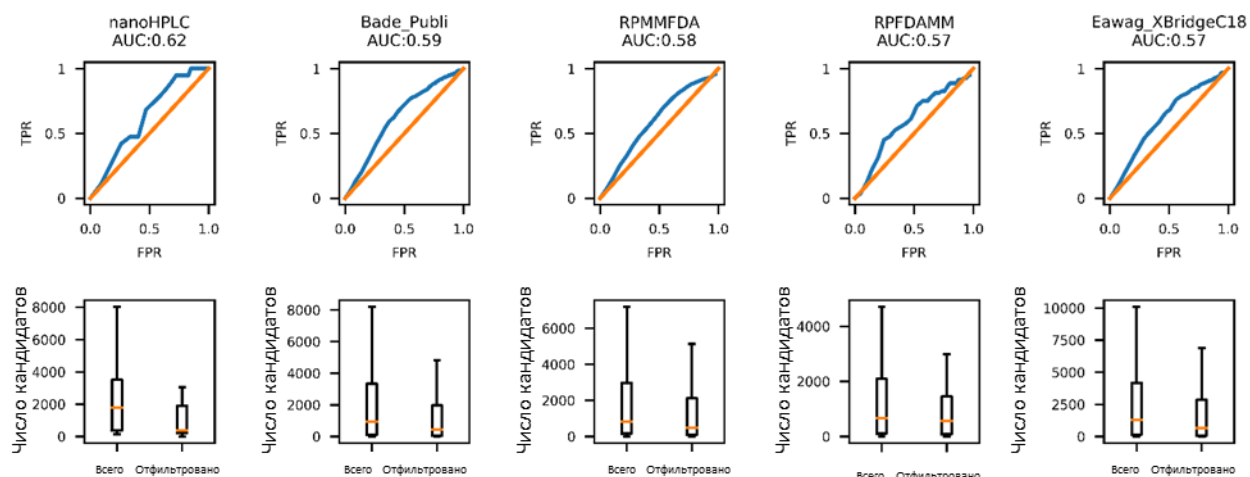


Рисунок 24. ROC кривые для определения пороговых значений (вверху) и эффективность фильтрации (внизу).

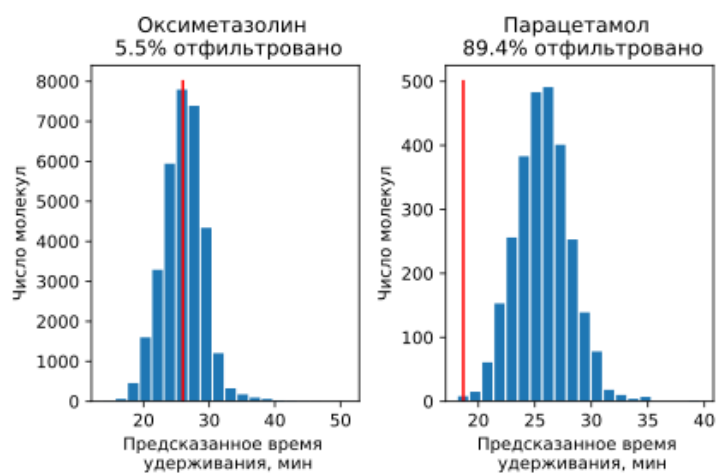


Рисунок 25. Распределение предсказанных времен удерживания изомеров оксиметазолина и парацетамола. Красная линия соответствует экспериментально измеренному времени удерживания оксиметазолина и парацетамола.

Несмотря на ограниченную способность предложенного подхода к сокращению пространства поиска, в абсолютных величинах количество отфильтрованных изомерных структур исчисляется сотнями и тысячами. Это позволит существенно облегчить дальнейшие действия при идентификации химических соединений, например, уменьшить временные затраты на идентификацию по вторичным масс-спектрам.

3.2 Предсказание времен удерживания в жидкостной хроматографии с использованием

текстовых представлений молекул, глубоких нейронных сетей и обучения с переносом

Основной проблемой предсказаний времен удерживания в жидкостной хроматографии является разнообразие условий разделения, как результат количество данных, полученных в одних условиях ограничено. В диссертационной работе предложено применение метода обучения с переносом. Данный подход применим в сочетании с методами глубокого обучения на

молекулярных представлениях. В общем случае, веса глубокой ИНС перед обучением инициализируются случайным образом. В процессе обучения они меняются для уменьшения функции потерь, однако при недостатке обучающих данных может потребоваться большое количество итераций, кроме того, высока вероятность переобучения. Метод обучения с переносом подразумевает предварительное обучение ИНС на данных, относящихся к родственной задаче, для которой есть обучающая выборка достаточно большого размера. Веса, полученные при предварительном обучении, далее используются при инициализации сети для обучения на целевом наборе данных.

Существует множество вариаций и реализаций подхода обучения с переносом. Возможно изменение как всех весов при до-обучении, так и фиксация весов некоторых (обычно первых) слоев ИНС неизменными при до-обучении. Кроме того, возможно изменение архитектуры ИНС, например, в конечную модель могут войти только некоторые слои исходной модели, новые слои при этом будут инициализированы случайным образом. Возможно ступенчатое до-обучение ИНС, применение градиентных изменений скорости обучения. Широкое применение метод обучения с переносом нашел в обработке естественного языка. При этом для предварительного обучения используется метод обучения с частичным привлечением учителя. Применительно к обработке естественного языка этот метод обычно реализуется как задача предсказания следующего слова в предложении (т.е. изначально неразмеченные текстовые данные могут быть размечены синтетически).

3.2.1 Описание предложенного подхода

Для повышения точности предсказаний времен удерживания в диссертационной работе предложен подход к предсказанию времен удерживания, общая схема которого представлена на рисунке 26. Предложенный подход включает предварительное обучение модели предсказания молекулярных структур, её промежуточное до-обучение на библиотеке METLIN SMRT (опционально) и до-обучение на целевом наборе данных.



Рисунок 26. Схема предложенного подхода к предсказанию времен удерживания с применением обучения с переносом.

В работе [107] применяли обучение с переносом для предсказания молекулярных свойств. Авторы предварительно обучали ИНС на обучающей выборке, состоящей из 1 000 000 молекул из библиотеки ChemBL[187] в режиме обучения с частичным привлечением учителя. В диссертационной работе также была использована эта обучающая выборка. Молекулы были представлены в виде строк SMILES, стандартизированных средствами библиотеки RDKit. Далее символы SMILES были преобразованы в числовые вектора. При этом каждый символ получал уникальный номер; символы химических элементов и ионов, состоящие из нескольких символов (Cl, Br, $[\text{NH}_4]^+$) рассматривались как один символ. Для моделирования использовали две архитектуры языковых моделей, реализованные в библиотеке FastAI[188]. Нейронная сеть AWD-LSTM[189], включает слой встраивания (Embedding layer), три слоя ячеек с долгосрочной и краткосрочной памятью (LSTM)[60] и слои регуляризации (Dropout)[190]. Другая архитектура, Transformer-XL[191], является одной из реализаций механизма внимания[62]. Обе ИНС использовали с установленными по умолчанию гиперпараметрами.

Для до-обучения последние слои получившихся моделей заменяли на полносвязную ИНС, веса которой инициализировали случайным образом. Далее проводили до-обучение на целевых наборах данных, содержащих времена удерживания молекул, измеренных в различных условиях хроматографического разделения. Для до-обучения выбраны наборы данных, полученные в условиях обращенно-фазовой хроматографии из репозитория PredRet[18], содержащие не менее 250 молекул (Eawag_XBridgeC18, Beck, Stravs, FEM_long). Сравнение предложенного подхода проводили с результатами применения традиционных методов МО, ранее опубликованными для

этих наборов [138]. Чтобы обеспечить корректное сравнение, при кросс-валидации использовали те же разбиения на обучающую и тестовую выборку, что и в этой работе. Дополнительно рассматривалась возможность моделирования времен удерживания в хроматографии гидрофильных взаимодействий (HILIC), для чего использовали набор данных Fiehn HILIC Library[148]. Для обращенно-фазовой хроматографии также оценивали эффективность трехступенчатого подхода, с первичным до-обучением на библиотеке METLIN SMRT[29].

3.2.2 Результаты моделирования времен удерживания при использовании обучения с переносом

В работе сравнивали архитектуры AWD-LSTM и Transformer-XL, оценивали влияние аугментации данных и предварительного обучения. На рисунке 27 приведено сравнение среднего абсолютного отклонения, определенного в режиме кросс-валидации (n=10) для рассмотренных обращенно-фазовых наборов данных. Можно видеть, что обе архитектуры дают схожие результаты. Хотя каждая из архитектур оказалась точнее на двух из четырех наборов данных, в абсолютных величинах точность Transformer-XL выше. Так как одной из задач данного этапа работы было создание универсального подхода, не требующего подбора архитектуры и гиперпараметров для различных условий разделения, то далее в работе использовалась архитектура Transformer-XL.

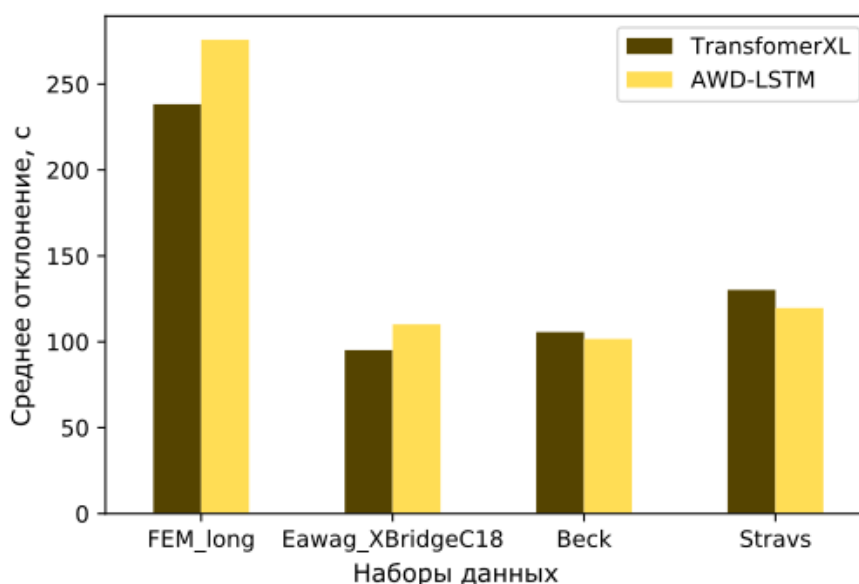
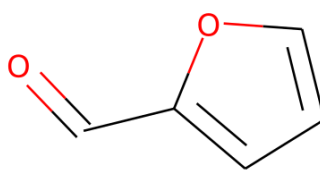


Рисунок 27. Сравнение архитектур AWD-LSTM и Transformer-XL.

Строки SMILES имеют преимущество по сравнению с другими представлениями молекул, т.к. позволяют провести аугментацию данных, варьируя атом молекулы с которого начинается запись строки (Рисунок 28). При этом нужно отметить, что одна строка SMILES соответствует

единственной молекуле. Кроме того, регрессионные характеристики могут быть аугментированы добавлением Гауссова шума. В работе максимальное значение σ установлено на уровне 5 с, как разумная величина отклонения времени удерживания при ширине пика в 30 с. Аугментация проводилась следующим образом: помимо канонической формы SMILES, генерировалась строка SMILES с начальным атомом, выбранным случайным образом. Если получившаяся строка ранее не была добавлена, она добавлялась к обучающей выборке, и ей приписывалось время удерживания $t_R + \varepsilon$, где ε случайное число в диапазоне $[-\sigma; \sigma]$, а t_R время удерживания молекулы из набора данных. Этот процесс повторялся N раз.



smiles	RT, s
c1ccoc1C=O	735.9
C(c1occc1)=O	736.3
O=Cc1ccco1	731.3
o1ccccc1C=O	729.9
c1(C=O)ccco1	737.5
...	...

Рисунок 28. Подход к аугментации данных.

В работе показано, что рост N приводит к увеличению точности предсказаний, поэтому в работе выбрано $N=800$ для молекул обучающей выборки (Рисунок 29). Аугментацию можно применять и непосредственно для предсказаний. Для этого делаются индивидуальные предсказания для нескольких SMILES строк одной и той же молекулы, после чего они усредняются. В работе показано, что такой подход повышает точность (Рисунок 30). Нужно отметить достаточно высокую воспроизводимость предсказаний для разных форм записи. Для молекул из набора FEM_long относительное стандартное отклонение предсказанных по разным записям SMILES времен удерживания составило в среднем 11.8%.

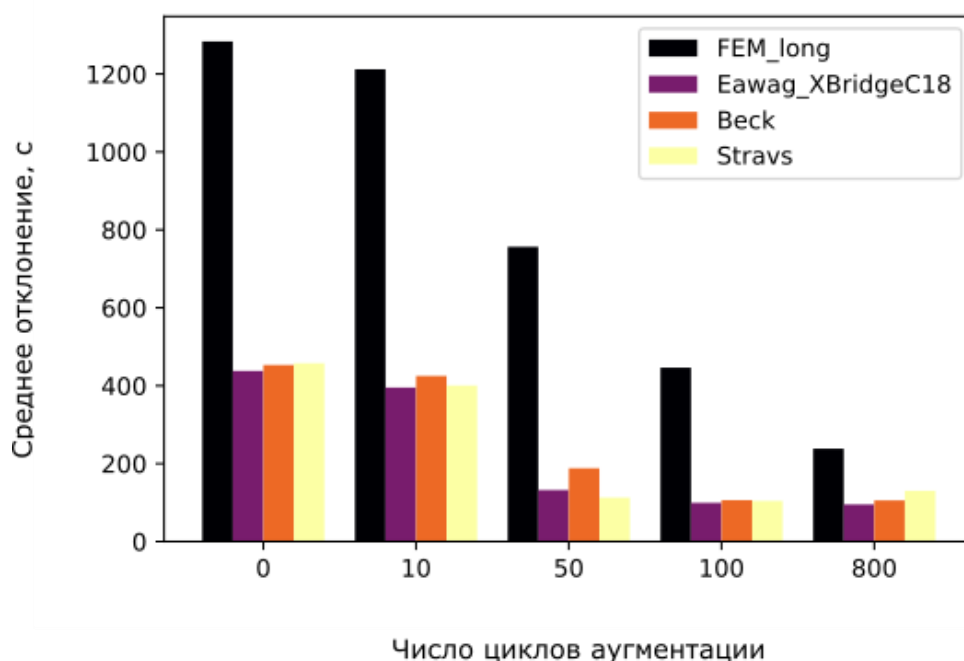


Рисунок 29. Влияние числа циклов аугментации на точность предсказаний.

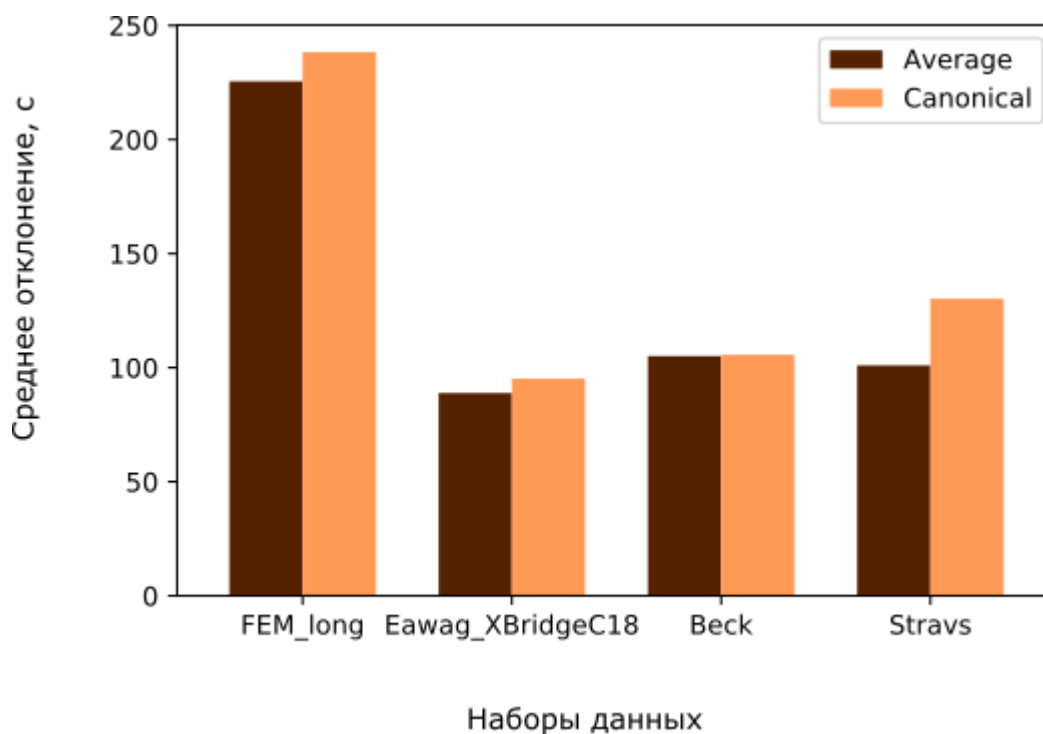


Рисунок 30. Увеличение точности предсказаний по набору SMILES.

Таким образом, для моделирования времен были выбраны следующие из рассмотренных параметров: архитектура Transformer-XL, аугментация обучающей выборки с 800 циклами, усреднение предсказаний по пяти случайно сгенерированным строкам SMILES. Результаты кросс-валидации приведены на рисунке 31 и в таблице, в сравнении с результатами, полученными традиционными методами МО.

Таблица 9. Результаты применения метода обучения с переносом для предсказания времен удерживания

Набор данных	Предсказания по канонической форме SMILES		Усредненные предсказания по набору SMILES		Лучший результат из работы [138]	
	Среднее отклонение, с	Медианное отклонение, с	Среднее отклонение, с	Медианное отклонение, с	Среднее отклонение, с	Медианное отклонение, с
FEM_long	238	145	226	120	248	117
Eawag_XBridgeC18	95	71	89	65	88	61
Beck	106	86	105	81	92	69
Stravs	130	107	101	84	110	76

Можно видеть, что предложенный подход сопоставим по точности с лучшими традиционными моделями МО, а на некоторых наборах данных его применение позволяет существенно снизить погрешность предсказаний. Пример зависимости предсказанных и экспериментальных времен удерживания для тестовой выборки набора данных FEM_long приведен на рисунке 32.

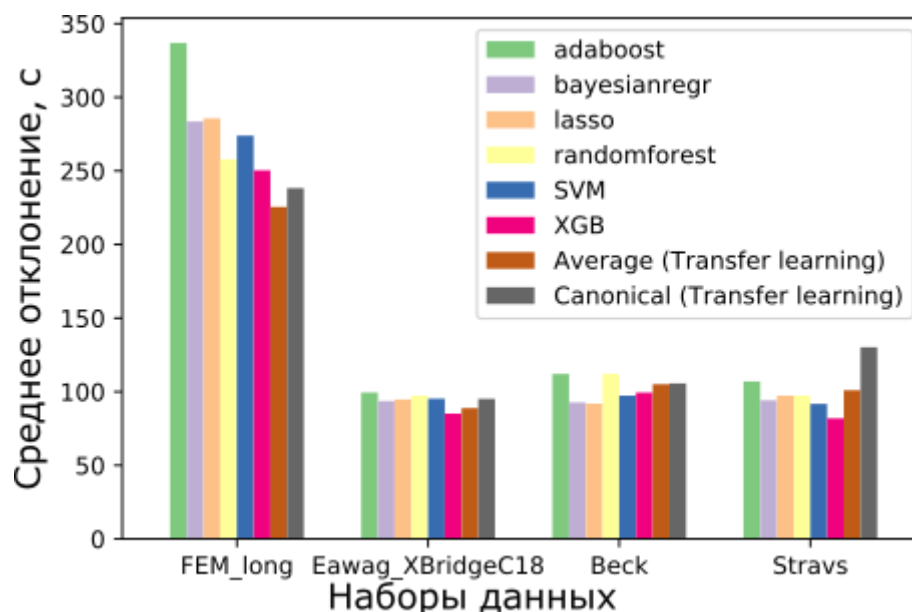


Рисунок 31. Сравнение предложенного подхода с традиционными методами машинного обучения[138].

Дополнительно, оценивалась эффективность промежуточного этапа до-обучения по библиотеке METLIN SMRT, по аналогии с подходом, предложенным для моделирования

CCS[182]. Приведенные в таблице 10 результаты демонстрируют, что такое промежуточное обучение способно дополнительно увеличить точность.

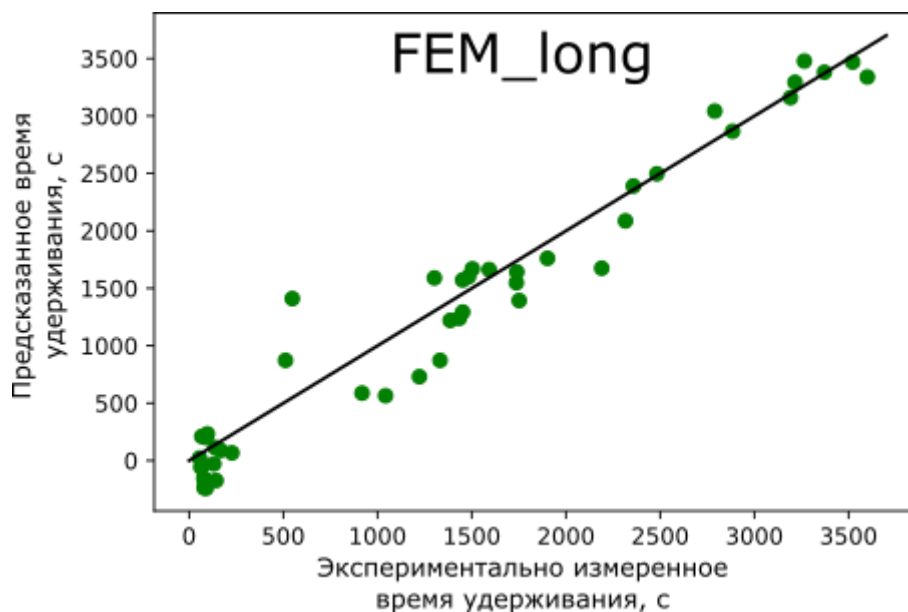


Рисунок 32. Соотношение предсказанных и экспериментальных времен удерживания в наборе данных FEM_long

Таблица 10. Результаты применения метода обучения с переносом с промежуточным обучением на данных библиотеки METLIN SMRT для предсказания времен удерживания

Набор данных	Предсказания по канонической форме SMILES		Усредненные предсказания по набору SMILES	
	Среднее отклонение, с	Медианное отклонение, с	Среднее отклонение, с	Медианное отклонение, с
FEM_long	260	157	223	127
Eawag_XBridge C18	101	76	88	66
Beck	104	84	96	69
Stravs	122	96	94	71

Чтобы убедиться в том, что повышение точности связано с предварительным обучением модели, был проведен следующий эксперимент. Регрессионные модели были инициализированы со случайными весами, и обучены (в режиме кросс-валидации) на наборах данных по удерживанию. На рисунке 33 приведены сравнительные диаграммы средних отклонений, полученных в режиме обучения с переносом и при инициализации со случайными весами. Можно видеть, что предварительное обучение значительно повышает точность.

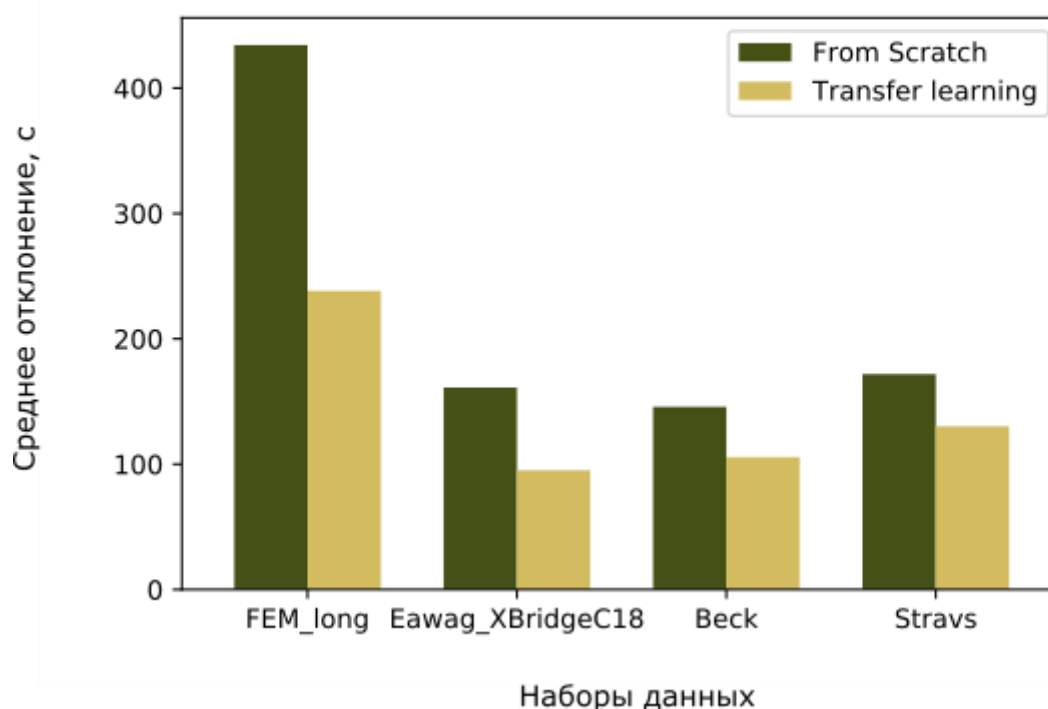


Рисунок 33. Сравнение средних отклонений, полученных в режиме обучения с переносом (Transfer learning) и при инициализации модели со случайными весами (From Scratch).

Отличительной особенностью предложенного подхода является тот факт, что при предварительном обучении не используется информация об удерживании молекул. Это позволяет использовать его не только для моделирования времен удерживания в обращенно-фазовой хроматографии, но и в других режимах разделения. Так, для набора данных Fiehn HILIC Library среднее абсолютное отклонение составило 1.1 мин, что хуже, чем результат индивидуальных моделей (0.78-1.11 мин). Однако, в отличие от индивидуальных моделей, предложенный подход использует универсальный набор гиперпараметров для всех моделей.

3.3 Предсказание времен удерживания с применением нейронных сетей с механизмом передачи сообщений (Message-Passing Neural Networks)

В диссертационной работе был предложен еще один подход к предсказанию времен удерживания с применением представления молекулы в виде графа и ИНС с распространением сообщения. Так как граф – наиболее естественное представление молекулы, ожидалось, что подобный подход позволит повысить точность предсказаний. Графовые ИНС с распространением сообщений изначально появились для решения задач связанных с моделированием молекулярных свойств[192]. Принцип работы таких ИНС изображен на рисунке 34. Предложено предварительное обучение ИНС на библиотеке METLIN SMRT с последующим до-обучением на других наборах данных по хроматографическому удерживанию, т. е. режим обучения с переносом.

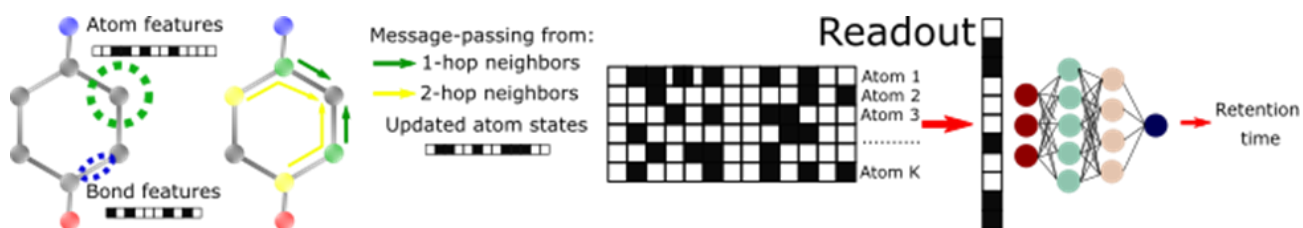


Рисунок 34. Принцип работы нейронных сетей с распространением сообщений

3.3.1 Описание предложенного подхода

В качестве выборки для предварительного обучения использовали библиотеку METLIN SMRT. На данном этапе неустойчивые молекулы с временем удерживания менее 2 мин были исключены из обучающей выборки. Ранее в диссертационной работе было показано, что бинарные классификаторы, основанные на алгоритме ГБ, справляются с задачей выявления неустойчивых молекул, поэтому на данном этапе этот вопрос не рассматривался. Итоговая обучающая выборка состояла из 77977 молекул. В качестве целевых наборов данных использовали наборы FEM_long, Eawag_XBridgeC18, LIFE_new, LIFE_old из репозитория PredRet[18], а также набор данных RIKEN Retip[148]. Молекулы описывались как ненаправленный граф, в котором вершинами являются атомы, а ребрами – связи. Атомы и связи описываются векторами свойств, перечень свойств и допустимых значений приведен в таблице 11.

Таблица 11. Свойства атомов и связей при описании молекулярного графа

Свойство	Допустимые значения
Свойства атомов	
Элемент	'Br', 'C', 'Cl', 'F', 'H', 'I', 'N', 'O', 'P', 'S', 'Si'
Число связанных атомов водорода	0, 1, 2, 3, 4
Валентность	0, 1, 2, 3, 4, 5, 6
Гибридизация	"s", "sp", "sp2", "sp3"
Является ли частью ароматической системы	True, False
Свойство связей	
Тип связи	"single", "double", "triple", "aromatic"
Является ли частью сопряженной системы	True, False

Архитектура ИНС с распространением сообщений включала блок кодирования признаков вершин и ребер графа, блоки распространения сообщений, в которых происходит агрегация состояний соседних атомов и связей, и обновление векторов состояний атомов с учетом

агрегированного состояния соседей. После матрица обновленных состояний переводится в латентный вектор признаков слоем `transformer`, и операцией подвыборки с функцией среднего (`Average Pooling`). Латентный вектор подается в полносвязную ИНС из трех слоев с `ReLU` активацией и линейным слоем в конце. Подробная архитектура использованной ИНС приведена на рисунке 35.

Обучение модели проводили со средним абсолютным отклонением в качестве функции потерь, размером батча 64, и начальной скоростью обучения 0.0002. Если значение функции потерь не снижалось за 10 эпох, то скорость обучения уменьшалась в 4 раза. Для предотвращения переобучения использовали механизм досрочной остановки, который прекращал обучение, если значение функции потерь не уменьшалось после 20 эпох. Все модели оценивали в режиме кросс-валидации ($n=5$). Дополнительно, были выделены независимые тестовые выборки, не вовлеченные в обучение и определение гиперпараметров, и содержащие 20% от общего количества данных. В режиме до-обучения веса всех слоев, кроме слоев полносвязной сети, фиксировались и не изменялись в ходе обучения. Размер батча был уменьшен до 8, другие гиперпараметры оставались неизменными.

3.3.2 Результаты предсказаний времен удерживания с помощью нейронных сетей с распространением сообщений

Результаты кросс-валидации предложенной модели после предварительного обучения представлены в таблице 12. На рисунке 36 приведены зависимость предсказанных и экспериментальных значений, а также распределение ошибок. Можно видеть, что подход с применением ИНС с распространением сообщений существенно превосходит по точности подход к предсказанию времен удерживания на основе модели ГБ. Кроме того, авторы библиотеки `METLIN SMRT` заявляют, что средние и медианные отличия времен удерживания при повторных измерениях составили 36 и 18 с, т.е. результаты модели дают результат, сопоставимый с экспериментальной погрешностью измерений времен удерживания. Результаты кросс-валидации в режиме обучения с переносом приведены в таблице 13, на рисунке 37 приведены распределения ошибок и соотношения предсказанных и экспериментально измеренных времен удерживания. Можно видеть, что предложенный подход отличается высокой точностью.

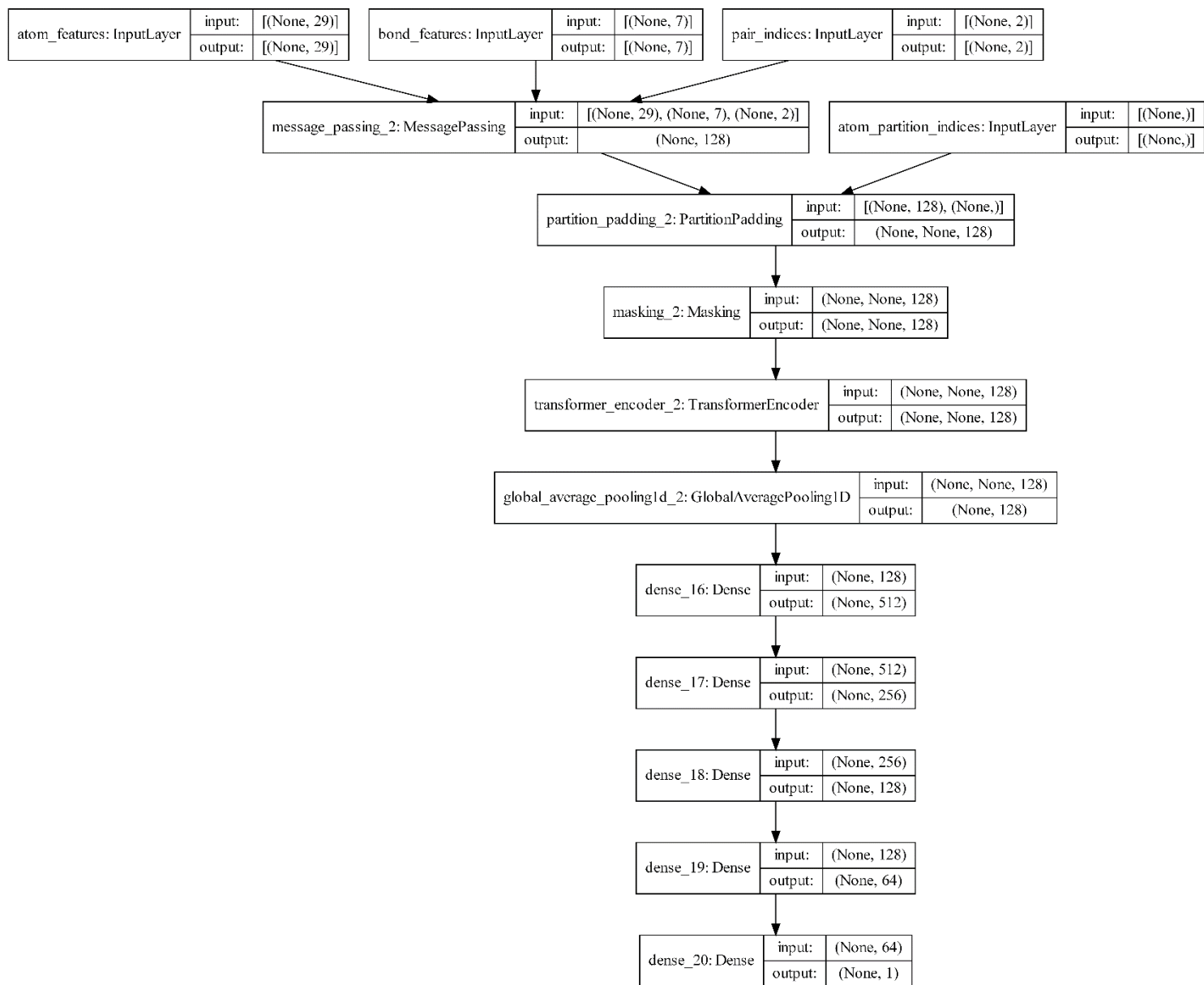


Рисунок 35. Архитектура нейронной сети с распространением сообщения.

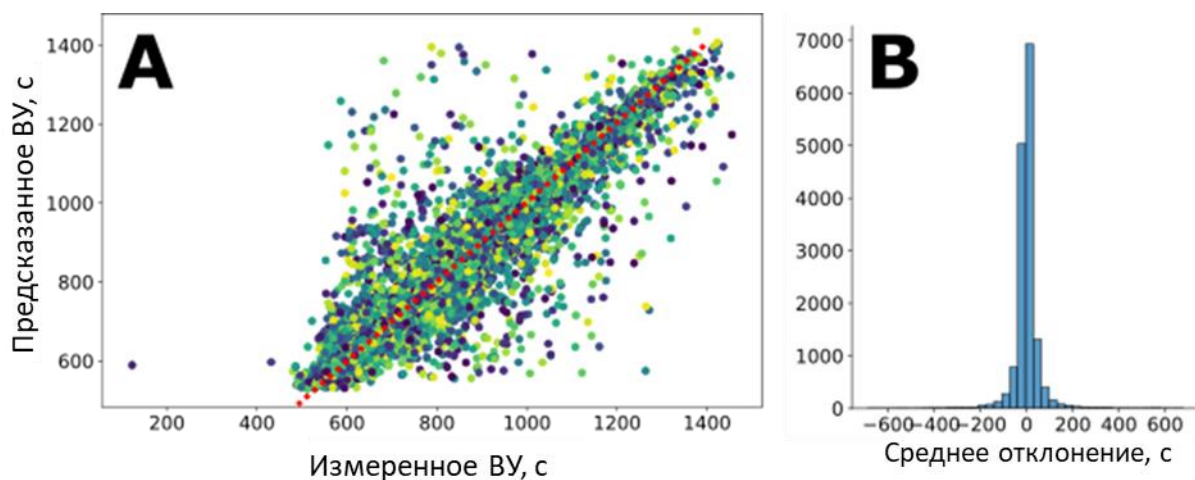


Рисунок 36 Распределение ошибок и соотношение предсказанных и экспериментально измеренных времен удерживания. Набор данных METLIN SMRT.

Таблица 12. Результаты кросс-валидации (n=5) нейронной сети с распространением сообщения на наборе данных METLIN SMRT

		Валидационная выборка	Тестовая выборка
Среднее отклонение, с	абсолютное	32.1±0.6	31.5±0.1
Медианное отклонение, с	абсолютное	16.2±0.2	16.0±0.2
Среднее отклонение, %	относительное	4.1±0.1	4.0±0.01
Среднее отклонение, с	квадратичное	62.8±1.9	60.5±0.3
R ²		0.872±0.008	0.879±0.001

Таблица 13. Результаты кросс-валидации (n=5) при моделировании времен удерживания для различных наборов данных. Представлены средние и медианные отклонения при моделировании в режиме обучения с переносом и при инициализации моделей со случайными весами

	RIKEN Retip		FEM_long		Eawag_XBridgeC18		LIFE_new		LIFE_old	
	Валидационная выборка	Тестовая выборка	Валидационная выборка	Тестовая выборка	Валидационная выборка	Тестовая выборка	Валидационная выборка	Тестовая выборка	Валидационная выборка	Тестовая выборка
MPNN Transfer Learning	34.7±3.3	38.2±3.2	214.4±25.6	204.6±23.0	79.5±10.3	80.9±6.5	23.3±3.8	22.1±3.3	16.9±2.9	16.9±2.0
MPNN From Scratch	56.2±13.6	57.2±13.2	299.4±51.0	317.7±25.8	137.1±24.7	135.8±14.1	37.5±17.2	39.1±15.2	27.2±8.3	23.2±2.1
MPNN Transfer Learning	22.2±2.4	25.0±3.5	93.5±14.9	125.2±12.0	56.0±4.7	57.4±9.3	12.2±1.1	9.7±3.6	9.9±3.1	9.5±0.7
MPNN From Scratch	38.0±9.5	40.8±12.4	162.4±51.2	193.1±41.3	105.3±20.7	115.1±18.1	20.4±21.5	24.0±22.2	18.11±8.6	19.5±3.6

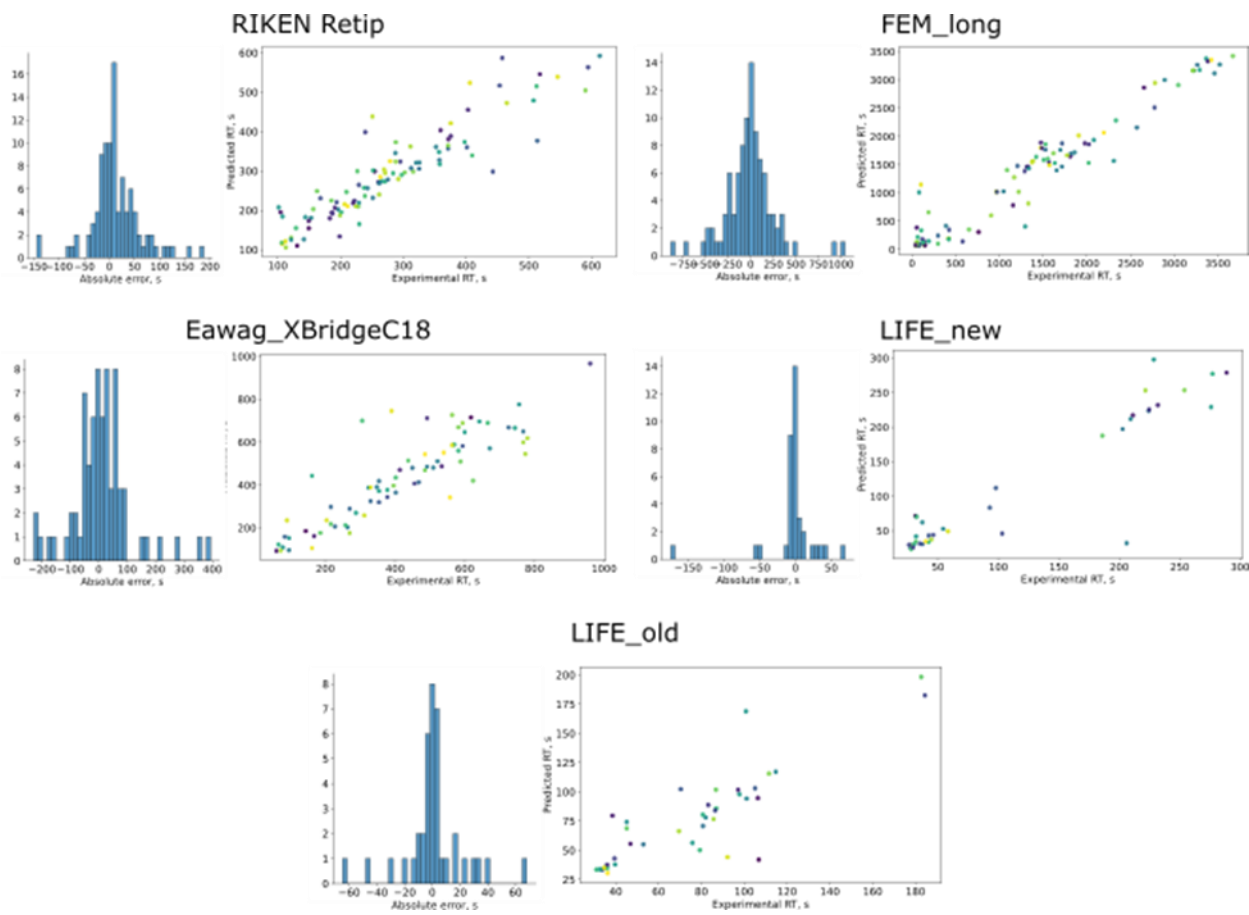


Рисунок 37. Распределения ошибок и соотношения предсказанных и экспериментально измеренных времен удерживания.

Для оценки применимости подхода для идентификации химических соединений в нецелевых исследованиях был проведен эксперимент по фильтрации ложноположительных определений среди изомерных кандидатов из базы данных PubChem, описанный в п. 3.1.3.,. Пороговые значения также были установлены по ROC-кривым, для наборов данных Retip, FEM_long, Eawag_XbridgeC18, LIFE_new and LIFE_old они составили 22.5, 10, 20, 17.5 and 20%, соответственно. В результате удалось отфильтровать в среднем 53, 23, 35, 33 and 31% ложноположительных определений (Рисунок 38). Хотя степень сокращения пространства поиска по предсказанным временам удерживания не позволяет прийти к однозначной идентификации, но способствует сокращению затрат при проведении идентификации в сочетании с другими методами.

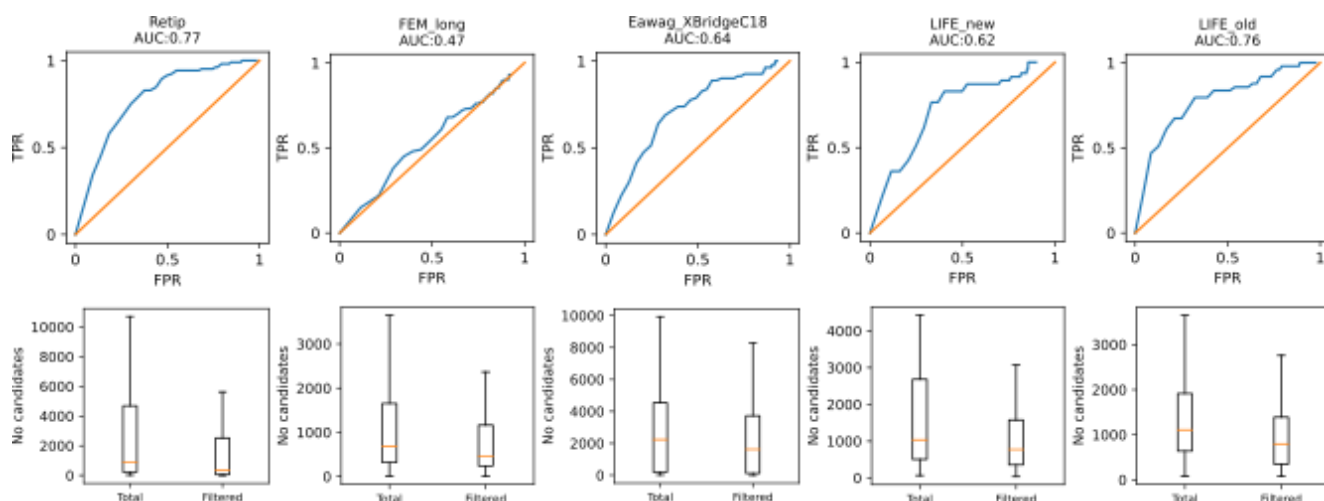


Рисунок 38. ROC кривые для определения порога при фильтрации ложноположительных определений и сокращение пространства поиска при использовании предсказанных значений времен удерживания.

3.4 Сравнение предложенных подходов

В работе предложено три подхода к моделированию времен удерживания с помощью методов машинного и глубокого обучения (Таблица 14). Все подходы решали задачу моделирования времен удерживания для различных хроматографических условий, в одном случае путем построения моделей пересчета между системами по экспериментальным данным, в двух других с применением метода обучения с переносом.

Таблица 14. Сравнение предложенных в работе подходов по моделированию времен удерживания

Алгоритм МО / переноса результатов предсказаний	Среднее отклонение, рассчитанное по независимой тестовой выборке, с	
	METLIN SMRT	Eawag_XBridgeC18
ГБ /кусочно-линейная функция	45.6±0.4	98.4
ИНС с архитектурой Трансформер / обучение с переносом	57.0±0.6	88.0±8.2
Графовая ИНС с передачей сообщений / обучение с переносом	31.5±0.1	79.5 ± 10.3

Первый подход обладает худшими показателями точности, однако может оказаться единственным решением при ограниченных возможностях накопления экспериментальных данных по удерживанию различных молекул для получения обучающей выборки. Подход к обучению с переносом при предварительном обучении модели в режиме частичного привлечения учителя на неразмеченных данных оказался менее точным, чем подход с предварительным обучением ИНС с распространением сообщений на данных удерживания из библиотеки METLIN SMRT. Однако, так как при обучении первичной модели не используется информация об удерживании, этот подход теоретически может применяться для моделирования удерживания в различных условиях разделения, например, в режиме хроматографии гидрофильных

взаимодействий HILIC. Хотя полученные в работе результаты для модельного набора данных уступают результатам индивидуальных моделей, увеличение размеров обучающей выборки может способствовать повышению точности обучения с переносом. Подход с предсказанием времен ИНС с распространением сообщений оказался наиболее точным, из предложенных в работе. Кроме того, он сопоставим, или превосходит наиболее точный на настоящий момент подход с применением одномерных сверточных сетей[157]. Кроме того, полученные отклонения сопоставимы с экспериментальной вариабельностью времен удерживания в библиотеке METLIN SMRT. Все разработанные модели, алгоритмы находятся в открытом доступе, и доступны по ссылкам:

<https://github.com/osv91/RTpredict> (дата обращения 14.01.2023)

<https://dx.doi.org/10.6084/m9.figshare.13315574> (дата обращения 14.01.2023)

<https://github.com/osv91/MPNN-RT> (дата обращения 14.01.2023).

Заключение к главе 3.

В работе предложено три различных подхода к моделированию времен удерживания с применением машинного и глубокого обучения. Все подходы направлены в том числе на решение задачи моделирования времен удерживания для различных хроматографических условий, в одном случае путем построения моделей пересчета между системами по экспериментальным данным, в двух других с применением метода обучения с переносом. Наилучших результатов удастся добиться при использовании ИНС с распространением сообщений, и обучении с переносом. Среднее отклонение при этом составило всего 31.5 с при обучении на большом наборе данных удерживания METLIN SMRT и является наилучшим результатом по сравнению с другими, ранее опубликованными или предложенными в работе подходами. Величина среднего отклонения сопоставима с вариабельностью времен удерживания на соответствующей хроматографической системе.

Предложенные подходы используют в качестве основной обучающей выборки обширный набор различных молекул, относящихся к различным химическим классам. Выборки меньшего размера, использованные в работе, также характеризуются разнообразием химических структур. Это отличает предложенные подходы от моделирования удерживания внутри определенных классов соединений. С одной стороны, результаты предсказаний для узких рядов схожих химических соединений хуже, чем при моделировании внутри ряда. С другой, предложенные модели оказываются более универсальными и имеют более высокий потенциал к экстраполяции. Нужно отметить, что подход с использованием кусочно-линейных функций пересчета будет хорошо работать только при условии сохранения порядка выхода веществ при изменении

условий разделения, в то время как подходы, использующие обучение с переносом, не имеют таких ограничений.

Основной сферой применения предложенных моделей может быть нецелевой анализ сложных объектов. В частности, в сравнительно новых направлениях, таких как нецелевая метаболомика разнообразие определяемых аналитов чрезвычайно велико, и количество известных веществ непрерывно растет. Так, например, количество соединений в библиотеке метаболитов человека HMDB (Human metabolome database) удвоилось за последние 4 года и составляет более 200 тысяч различных веществ. Еще более остро задача идентификации химических соединений стоит в области метаболомики растений, где вклад в химическое разнообразие вносят вторичные метаболиты.

При сужении круга возможных кандидатов, прогнозирование времен удерживания имеет ограниченную практическую значимость. Например, при решении задачи идентификации метаболитов новых лекарственных средств, набор возможных метаболитов определяется возможными путями. При этом возможные метаболиты являются близкими по структуре соединениями (в частности, позиционными изомерами), характеризующимися близкими значениями времен удерживания. Селективности предложенных универсальных моделей может не хватать для их разрешения.

При использовании предсказанных времен для фильтрации изомерных кандидатов, удастся существенно сократить пространство поиска при идентификации химических соединений в нецелевом хромато-масс-спектрометрическом анализе. Установление пороговых значений по ROC кривым позволяет управлять соотношением ложноположительных и ложноотрицательных определений в зависимости от конкретной задачи. Хотя использование такого фильтра не столь эффективно, как, например, экспертная интерпретация масс-спектров второго порядка, применение данного фильтра легко автоматизируется, сокращая ресурсы, необходимые для дальнейшей идентификации. Таким образом, предсказание времен удерживания хотя и не позволяет однозначно аннотировать идентифицированные компоненты, но может повышать эффективность других методов идентификации.

ГЛАВА 4. Совместное применение методов предсказания времен удерживания и метода изотопного обмена для идентификации химических соединений в нецелевом скрининге²

Использование предсказанных времен удерживания не позволяет однозначно установить компоненты сложных образцов, особенно, если при идентификации поиск возможных структур ведется по общехимическим библиотекам. Тем не менее, в диссертационной работе показано, что фильтрация ложноположительных определений по предсказанным временам удерживания может существенно сократить пространство поиска перед использованием других подходов к идентификации. Одним из таких подходов является сочетание изотопного обмена с хромато-масс-спектрометрическим анализом. Изотопный обмен, в первую очередь обмен протона на дейтерий (H/D), может быть использован для функционального анализа. Известно, что в реакции дейтериеводородного обмена, происходит замещение лабильных атомов водорода в молекуле на атомы дейтерия, что приводит к увеличению массы молекулы и отражается на масс-спектрах[193]. Так как наиболее лабильными атомами водорода в органических молекулах являются те атомы, которые связаны с гетероатомами, изменение молекулярной массы, регистрируемое масс-спектрометром, позволяет установить, как минимум количество соответствующих функциональных групп, содержащих такие атомы. При этом, реакция дейтериеводородного обмена может быть проведена в источнике ионизации масс-спектрометра, что позволяет проводить хроматографическое разделение с использованием недейтерированных растворителей[194].

Помимо дейтериеводородного обмена, известен обмен изотопов кислорода $^{16}\text{O}/^{18}\text{O}$ [195]. Например, в карбонильных соединениях возможен изотопный обмен вследствие обратимого присоединения к ним тяжелоокислородной воды H_2^{18}O (Рисунок 39).

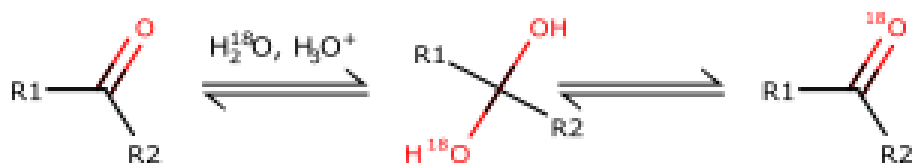


Рисунок 39. Механизм изотопного обмена $^{16}\text{O}/^{18}\text{O}$ в карбонильных соединениях.

² При подготовке данной главы диссертации использованы публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ отражены основные результаты, положения и выводы исследования. **Osipenko S.**, Zhrebker A., Rumiantseva L., Kovaleva O., Nikolaev E. N., Kostyukevich Y. Oxygen Isotope Exchange Reaction for Untargeted LC-MS Analysis // Journal of the American Society for Mass Spectrometry. – 2022. – Т. 33, № 2. – С. 390-398 (Импакт-фактор Web of Science – 3.262, Q1). 50%; **Osipenko S.**, Nikolaev E., Kostyukevich Y. Amine additives for improved in-ESI H/D exchange // Analyst. – 2022. – Т. 147, № 14. – С. 3180-3185. (Импакт-фактор Web of Science – 5.227, Q1). 50%.

Реакции обмена изотопов кислорода протекают медленно, и в целях химического анализа обычно проводились в жестких условиях (длительное кипячение образцов с водой H_2^{18}O) [196, 197]. Такой подход допустим при анализе стабильных сложных природных объектов (нефть, гуминовые вещества), однако совершенно неприменим при анализе чувствительных образцов, например, биологических жидкостей. Целью этого этапа диссертационной работы являлась оценка эффективности совместного применения фильтрации ложноположительных определений по предсказанным временам удерживания и результатам изотопного обмена.

В диссертационной работе было обнаружено, что интерпретация данных изотопного обмена H/D для идентификации может быть осложнена в связи с низкой степенью обмена в источнике ионизации атомов водорода в амидных группах. Например, в масс-спектрах ропивакаина не наблюдается замещение атома водорода амидной группы на дейтерий в предложенных ранее условиях с использованием D_2O в качестве дейтерирующего агента (Рисунок 40). Отчасти эта проблема может быть решена с помощью добавки алкиламинов в дейтерирующий агент. Другой проблемой при использовании H/D обмена в источнике ионизации является необходимость минимизации потоков элюента. В связи с этим, желательно использование нано- или микро-потоковой хроматографии, которая пока находит ограниченное применение для анализа низкомолекулярных соединений.

В связи с этим, в дальнейшем в работе использовали метод изотопного обмена $^{16}\text{O}/^{18}\text{O}$, который не требует модификации масс-спектрометрического оборудования. Для этого предварительно требовалось решение задач по выбору условий проведения реакции обмена изотопов кислорода, подходящих для анализа биологических проб, определению функциональных групп, вступающих в реакцию в этих условиях и разработке алгоритма к фильтрации изомерных кандидатов по результатам изотопного обмена.

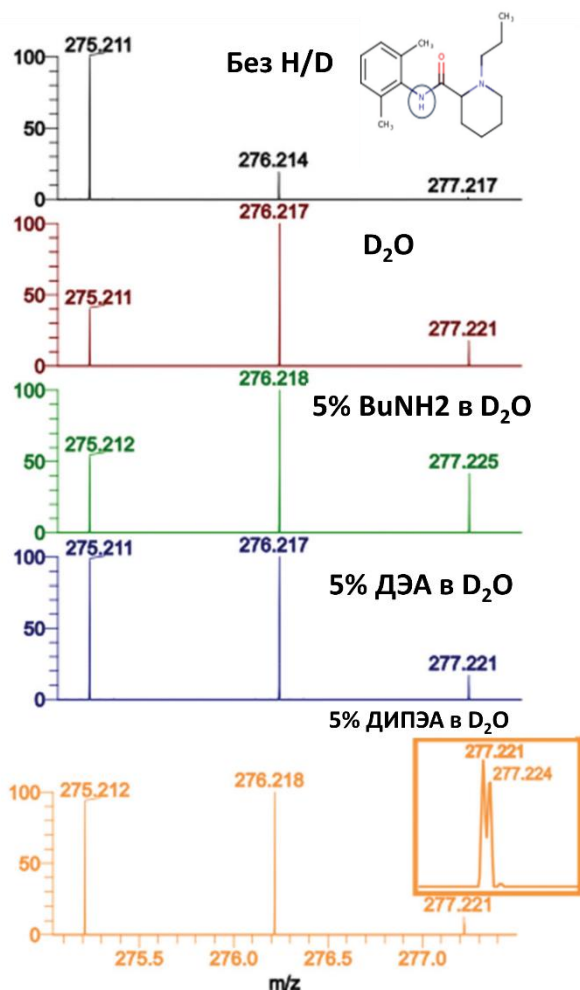


Рисунок 40. Масс спектры ропивакаина в условиях Н/Д обмена в источнике ионизации, в т.ч. с добавкой алкиламинов к дейтерирующему агенту (BuNH₂ – n-бутиламин, ДЭА – диэтиламин, ДИПЭА – диизопропилэтиламин).

4.1 Определение селективности изотопного обмена изотопов кислорода ¹⁶O/¹⁸O

Изучение селективности изотопного обмена ¹⁶O/¹⁸O проводили в «мягких» и «жестких» условиях. В первом случае образцы инкубировали при температуре 37°C, во втором при 95°C в течение 24 ч. Селективность изучали на модельной смеси соединений, в основном относящихся к различным классам лекарственных средств. В качестве биологической матрицы была выбрана моча человека. Для проведения изотопного обмена сухой остаток после пробоподготовки перерастворяли в 30% растворе ацетонитрила в тяжелоокислородной воде H₂¹⁸O, с добавкой 1% трифторуксусной кислоты. Таким образом расход тяжелоокислородной воды при анализе аликвоты мочи 50 мкл составил всего 35 мкл. Нужно отметить, что стоимость такого количества тяжелоокислородной воды не превышает стоимости картриджа для твердофазной экстракции. Анализ проводился методом жидкостной хромато-масс-спектрометрии, с применением масс-спектрометра высокого разрешения (140 000). Высокое разрешение важно при изучении реакции

обмена, особенно если ее эффективность не высока. В этом случае требуется разрешение спектральных пиков, соответствующих содержанию одного атома ^{18}O (+2.0042) и двух атомов ^{13}C (+2.0066). Примеры масс-спектров, измеренных после проведения изотопного обмена $^{16}\text{O}/^{18}\text{O}$ приведены на рисунке 41.

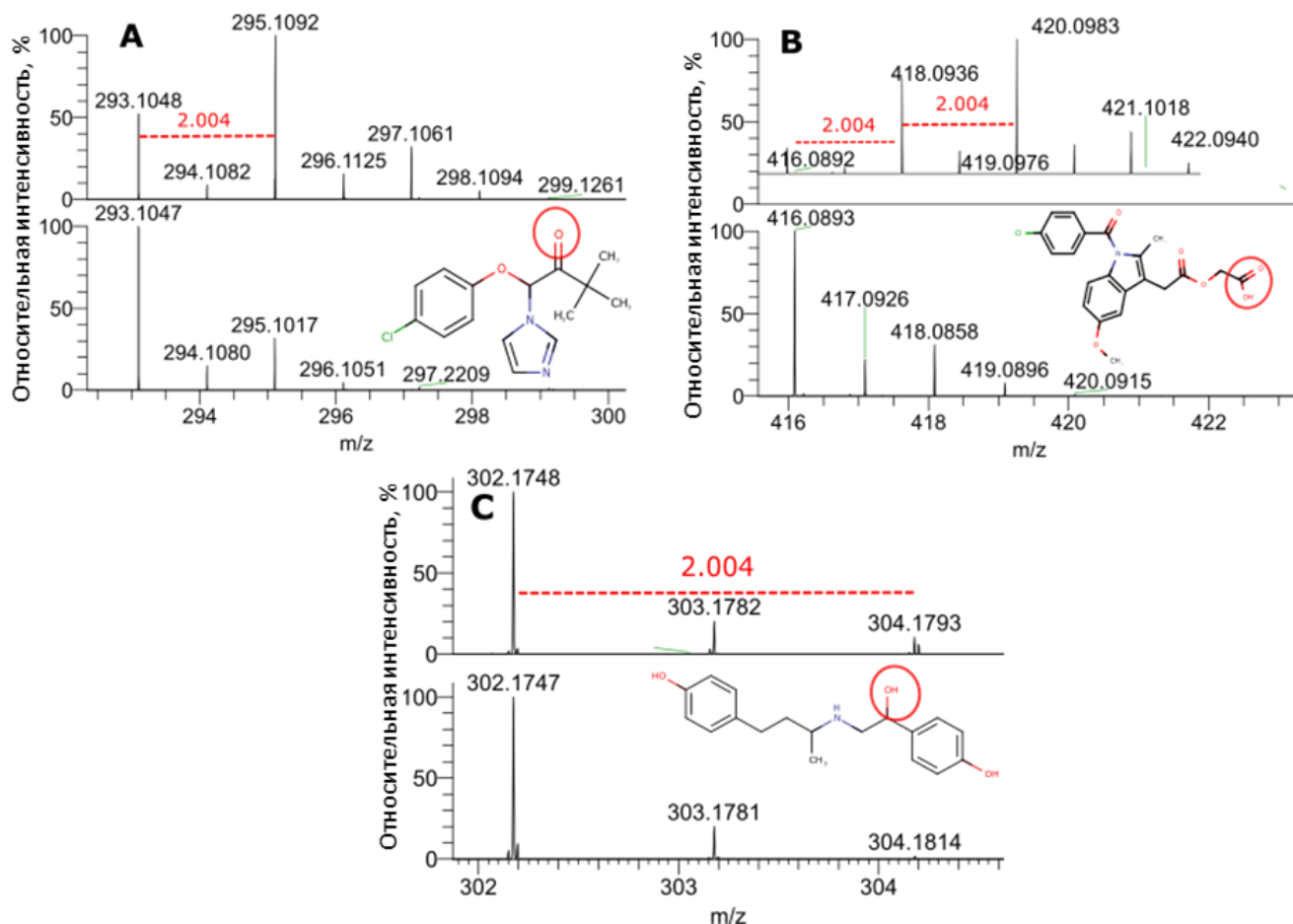


Рисунок 41. Наложение масс-спектров, измеренных до (внизу) и после (вверху) проведения реакции изотопного обмена $^{16}\text{O}/^{18}\text{O}$. А – Климбазол, В – ацететаин, С - рактопамин

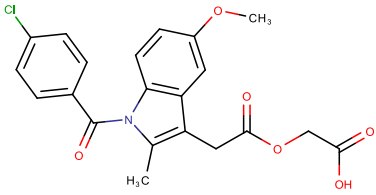
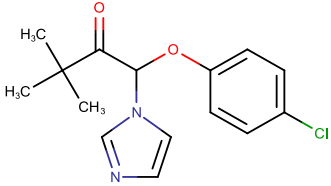
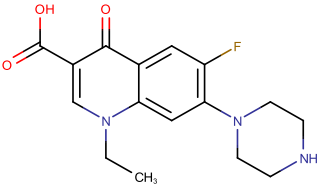
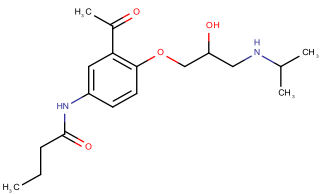
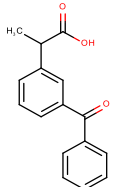
В таблице 16 приведены результаты определения числа обменов в изученных соединениях. Из представленных результатов можно сделать следующие наблюдения. Во-первых, не наблюдается обмен атомов кислорода эфирных (простых и сложных), и амидных групп. В карбоксильных группах наблюдался обмен одного или двух атомов кислорода, хотя в некоторых карбоксильных группах обмен вообще не наблюдался. Карбонильные группы в основном легко вступают в обмен, а гидроксильные – только в аллильном и бензильном положениях (Таблица 15).

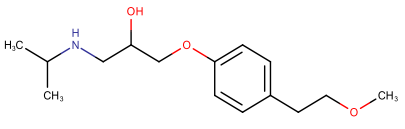
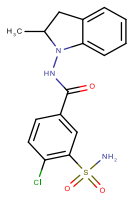
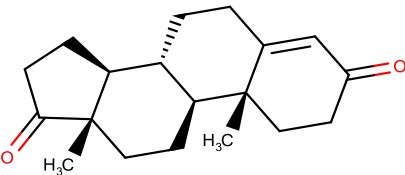
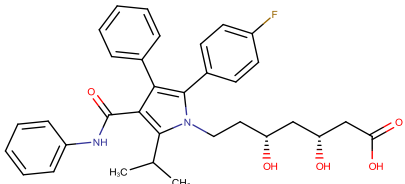
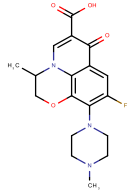
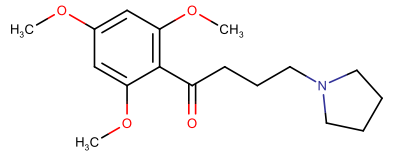
Таблица 15. Идентифицированные "обменные" и "необменные" группы

«Обменные» группы	«Необменные» группы
Карбонильная группа $R_1-C(=O)-R_2$	Нитро-группа $R-NO_2$
Карбоксильная группа $R-COOH$	Сульфоновая группа $R_1-SO_2-R_2$
Гидроксильная группа в аллильном положении $R_1-CH=CH-CH(OH)-R_2$	Амидная группа $R_1-N-C(=O)-R_2$
Гидроксильная группа в бензильном положении $Ar-CH(OH)-R$	Эфирная группа $R_1-C(=O)-O-R_2, R_1-O-R_2$
	Гидроксильная группа (включая фенолы) $R-OH, Ar-OH$

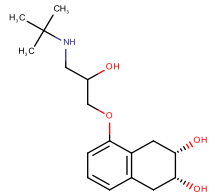
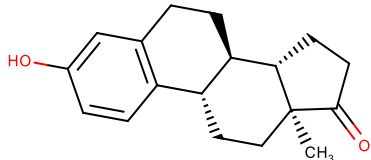
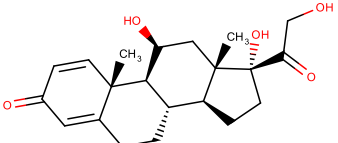
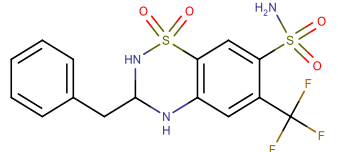
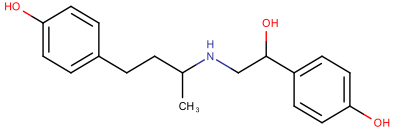
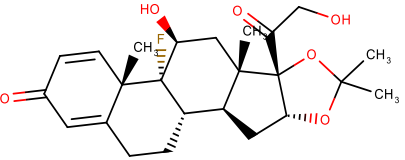
К сожалению, не удалось выявить закономерностей по изменению селективности в зависимости от температуры инкубации, хотя эффективность обмена ожидаемо увеличивалась при увеличении температуры. Для некоторых соединений это способствовало, например, детектированию второго обмена в карбоксильных группах при увеличении температуры реакции с 37°C до 95°C. Однако нужно отметить деградацию некоторых соединений при температуре 95°C и невозможность их обнаружения. В целом нужно отметить, что проведение реакции изотопного обмена приводит к повышению пределов обнаружения в хромато-масс-спектрометрическом анализе, как за счет распределения сигнала между пиками изотопного пакета, так и за счет частичной деградации аналитов при нагревании. Однако, обнаружение компонентов проводится при первичном анализе до проведения реакции изотопного обмена, и потому общее количество детектируемых компонентов не меняется при применении изотопного обмена, а пределы обнаружения соответствуют пределам обнаружения хромато-масс-спектрометрического анализа.

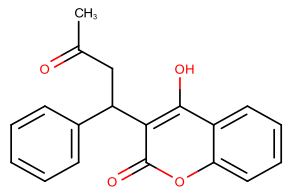
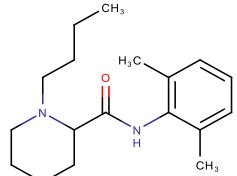
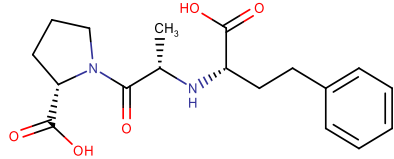
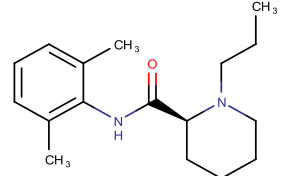
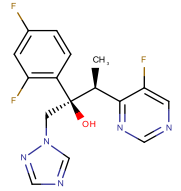
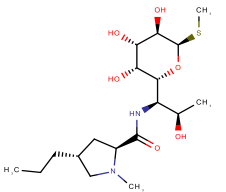
Таблица 16. Результаты определения числа изотопных обменов $^{16}\text{O}/^{18}\text{O}$ в кислородсодержащих органических соединениях

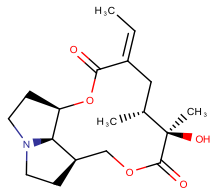
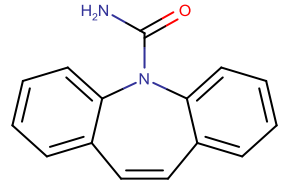
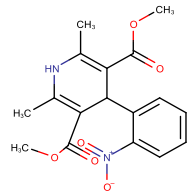
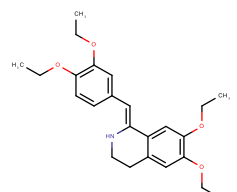
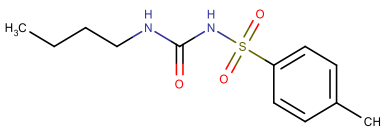
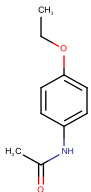
Название	Идентификатор в базе данных PubChem	Время удерживания, мин	Число наблюдаемых обменов $^{16}\text{O}/^{18}\text{O}$		Структура	Максимально возможное число обменов $^{16}\text{O}/^{18}\text{O}$	Число атомов кислорода
			При 37°C	При 95°C			
Ацеметацин	1981	17.82	2	2		2	6
Климбазол	37907	13.56	1	1		1	2
Норфлоксацин	4539	6.9	1	2		3	3
Ацебутолол	1978	8.94	1	1		1	4
Кетопрофен	3825	14.7	3	3		3	3

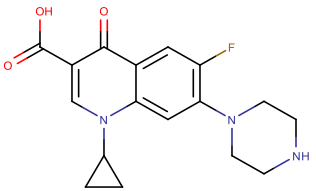
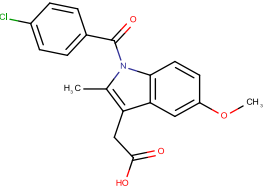
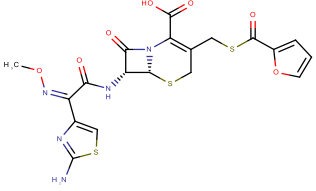
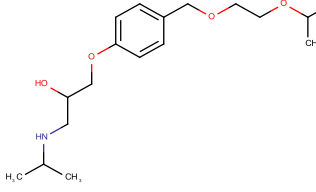
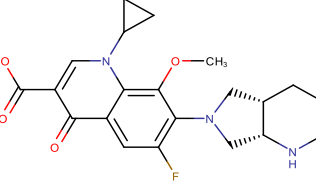
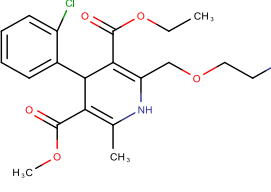
Метопролол	4171	9.06	0	0		0	3
Индапамид	3702	12.61	0	0		0	3
Андростендион	6128	15.59	2	2		2	2
Аторвастатин	60823	18.15	2	2		2	5
Офлоксацин	4583	7.09	1	1		3	4
Буфломедил	2467	9.39	1	1		1	4

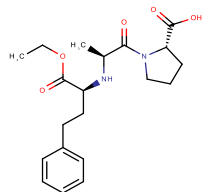
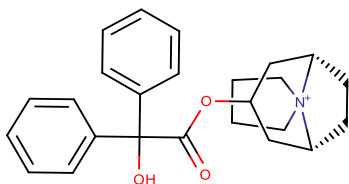
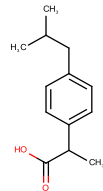
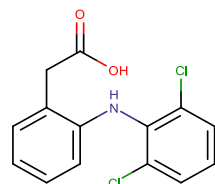
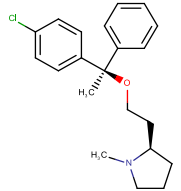
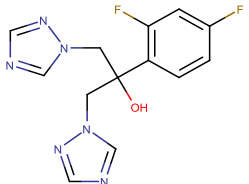
Фенофибрат	3339	22.23	1	1		1	4
Кофеин	2519	3.84	0	0		0	2
Тестостерон	6013	15.04	1	1		1	2
Бензбромарон	2333	20.88	1	1		1	3
Бензидамин	12555	12.99	0	0		0	1
Норгестрел	13109	16.5	1	1		1	2

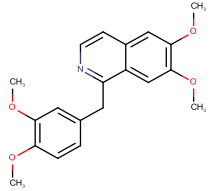
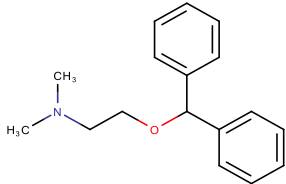
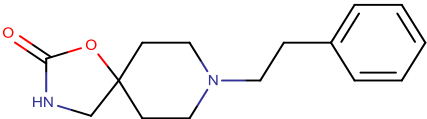
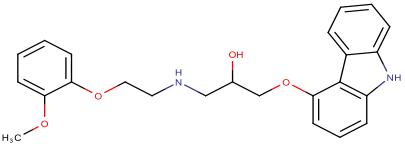
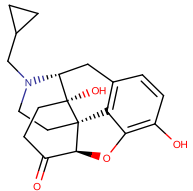
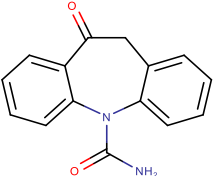
Надолол	39147	7.03	0	0		0	4
Эстрон	5870	15.28	1	1		1	2
Преднизолон	5755	16.91	2	2		2	5
Бендрофлуметиазид	2315	13.56	0	Дегградация соединения		0	4
Рактопамин	56052	7.46	1	1		1	3
Триамцинолон ацетонид	6436	13.79	1	2		2	6

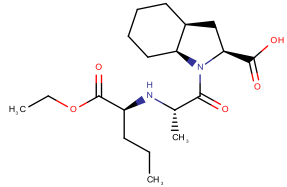
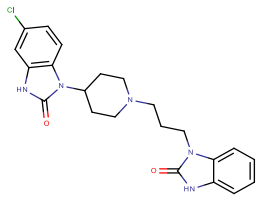
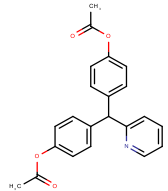
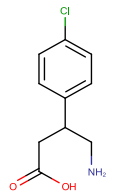
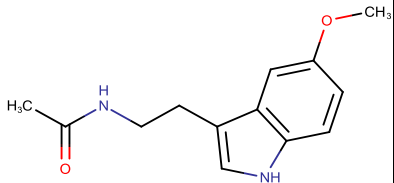
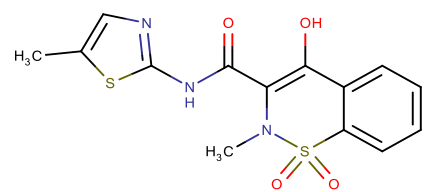
Варфарин	54678486	15.7	1	1		1	4
Бупивакаин	2474	10.88	0	0		0	1
Эналаприлат	5462501	7.86	2	3		4	5
Ропивакаин	175805	9.48	0	0		0	1
Вориконазол	71616	13.5	0	Дегградация соединения		1	1
Линкомицин	3000540	5.6	0	0		0	6

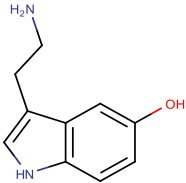
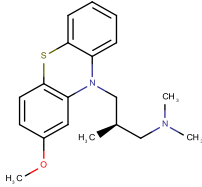
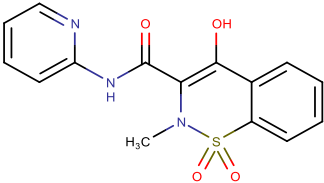
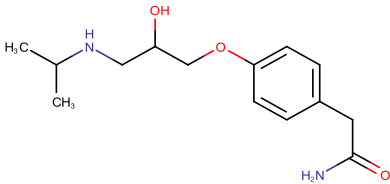
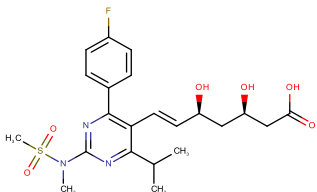
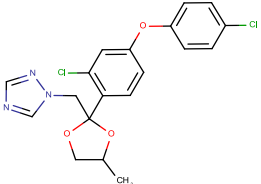
Платифиллин	5281742	8.25	0	0		0	5
Карбамазепин	2554	12.57	0	0		0	1
Нифедипин	4485	14.94	0	0		0	6
Дрогаверин	1712095	14.45	0	0		0	4
Толбутамид	5505	14.13	0	Дегградация соединения		0	3
Фенацетин	4754	9.72	0	0		0	2

Ципрофлоксацин	2764	7.33	1	1		3	3
Индометацин	3715	17.47	2	2		2	4
Цефтиофур	6328657	10.98	0	0		2	7
Бисопролол	2405	11.08	0	Дегградация соединения		0	4
Моксифлоксацин	152946	9.84	0	2		3	4
Амлодипин	2162	13.81	0	0		0	5

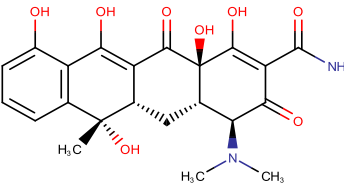
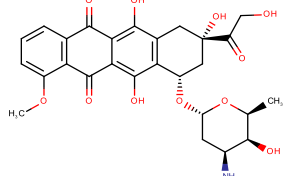
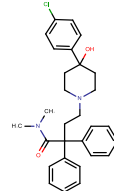
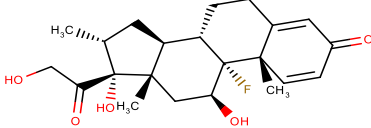
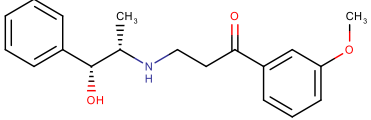
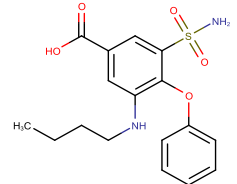
Эналаприл	5388962	11.46	2	2		2	5
Троспий	5284632	11.76	0	0		1	3
Ибупрофен	3672	8.57	0	0		2	2
Диклофенак	3033	17.25	2	2		2	2
Клемастин	26987	15.78	0	Дегградация соединения		0	1
Флуконазол	3365	8.49	0	0		1	1

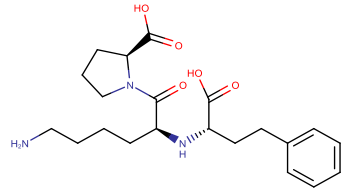
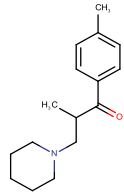
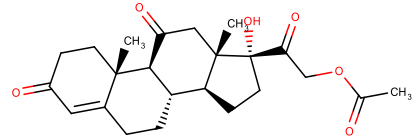
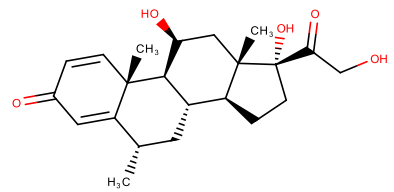
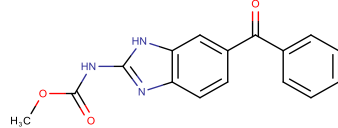
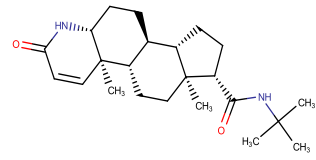
Папаверин	4680	10.38	0	0		0	4
Дифенгидрамин	3100	11.9	0	0		0	1
Фенспирид	3344	3.51	0	0		0	2
Карведилол	2585	12.83	0	0		0	4
Налтрексон	5360515	4.19	1	Дегградация соединения		1	4
Окскарбазепин	34312	11.23	1	1		1	2

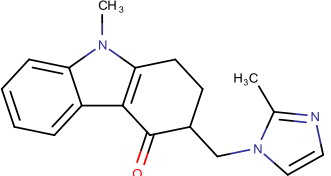
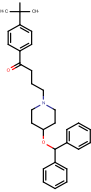
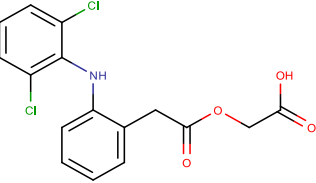
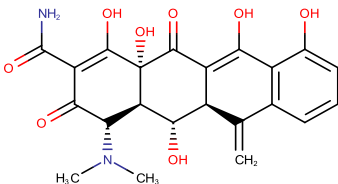
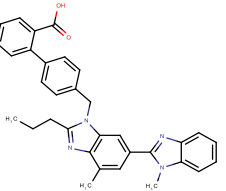
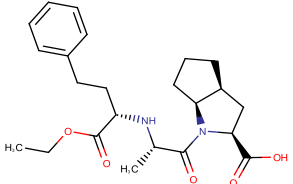
Периндоприл	107807	12.01	1	2		2	5
Домперидон	3151	10.76	0	0		0	2
Бисакодил	2391	14.25	0	Дегградация соединения		0	4
Баклофен	2284	4.03	2	2		2	2
Мелатонин	896	9.37	0	0		0	2
Мелоксикам	54677470	14.75	0	0		0	4

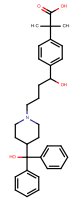
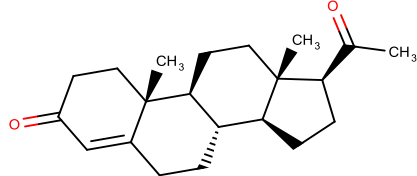
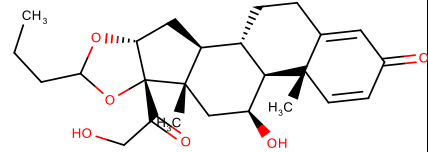
Серотонин	5202	0.82	0	0		0	1
Левомепромазин	72287	13.77	0	0		0	1
Пироксикам	54676228	12.59	0	0		0	4
Атенолол	2249	1.68	0	0		0	3
Розувастатин	446157	14.85	3	3		3	6
Дифенокназол	86173	19.38	0	0		0	3

Лозартан	3961	14.13	0	0		0	1
Итоприд	3792	8.65	0	0		0	4
Мебеверин	4031	13.51	0	0		0	5
Тримебутин	5573	11.83	0	0		0	5
Пентоксифиллин	4740	9.3	0	0		1	3
Верапамил	2520	13.84	0	0		0	4

Тетрациклин	54675776	7.57	1	Дегградация соединения		4	8
Доксорубицин	31703	11.33	1	Дегградация соединения		3	11
Лоперамид	3955	15.5	0	0		1	2
Дексаметазон	5743	13.24	2	2		2	5
Оксифедрин	5489013	11.19	1	Дегградация соединения		2	3
Буметанид	2471	15.46	0	2		2	5

Лизиноприл	5362119	10.66	1	1		4	5
Толперизон	5511	10.19	1	1		1	1
Кортизона ацетат	5745	14.57	3	3		3	6
Метилпреднизолон	6741	13.15	2	2		2	5
Мебендазол	4030	11.97	1	1		1	3
Финастерид	57363	16.04	0	0		0	2

Ондансетрон	4595	9.61	1	1		1	1
Эбастин	3191	19.13	1	Дегградация соединения		1	2
Ацеклофенак	71771	17.5	2	2		2	4
Метациклин	54675785	9.96	1	2		3	8
Телмисартан	65999	14.37	0	1		2	2
Рамиприл	5362129	13.32	1	2		2	5

Фексофенадин	3348	13.88	1	2		4	4
Прогестерон	5994	18.24	2	2		2	2
Будесонид	5281004	15.72	2	2		2	6

4.2 Фильтрация ложноположительных определений с помощью изотопного обмена $^{16}\text{O}/^{18}\text{O}$

По результатам изучения реакции обмена изотопов кислорода $^{16}\text{O}/^{18}\text{O}$ в индивидуальных соединениях были установлены «обменные» функциональные группы, обмен в которых наблюдался хотя в одном исследованном соединении, и «необменные» группы, обмен в которых не наблюдался вообще. Нужно отметить, что данное деление групп не является однозначным и основывается на изучении обмена в конкретных условиях, при рассмотрении хотя и довольно большого, но ограниченного набора молекул. Так, в работе гидроксильные группы считались «необменными», хотя известно, что в некоторых условиях атом кислорода в феноле может быть замещен его тяжелым изотопом. Тем не менее, например, в сложных эфирах и амидах, сложно предположить механизм обмена без разрыва соответствующей сложноэфирной или амидной связей, поэтому отсутствие наблюдаемых обменов в этих группах вполне ожидаемо.

В работе не удалось установить функциональные группы, атомы кислорода которых бы селективно и специфично вступали в изотопный обмен. Другими словами, наблюдение N обменов в эксперименте, не гарантирует, что в молекуле ровно N «обменных» групп. Поэтому в работе предложено использовать максимально возможное количество обменов, которое определяется по структуре молекулы с учетом результатов определения обменных групп, приведенных в таблице 15. Карбоксильная группа при этом учитывается дважды, так как содержит два атома кислорода, способных к обмену. Например, в молекуле мебендазола максимально возможное количество обменов составит 1, а в молекуле ципрофлоксацина 3. (Рисунок 42). При этом, в случае ципрофлоксацина экспериментально наблюдался только один обмен $^{16}\text{O}/^{18}\text{O}$. Точное соответствие экспериментально определенного и максимально возможного количества обменов наблюдалось для 77 из 96 исследованных соединений, что составило 80.2%. Отклонения в основном наблюдались в сопряженных структурах с делокализацией электронной плотности (например, в фторхинолонах), или в соединениях с гидроксильной группой в аллильном или бензильном положении, не вступивших в обмен.

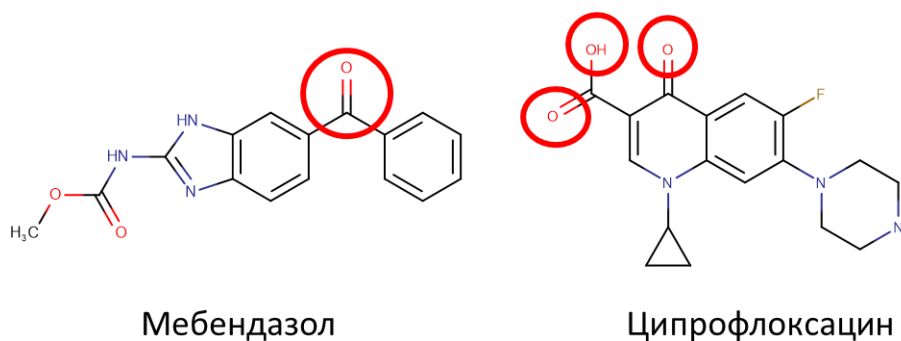


Рисунок 42. «Обменные» атомы кислорода в молекулах мебендазола и ципрофлоксацина.

Можно построить алгоритм фильтрации изомерных кандидатов исходя из того, что при наблюдаемом количестве обменов N , максимально возможное количество обменов, определенное по структуре не может быть меньше N . При этом, очевидно, что такой алгоритм будет применим только при идентификации химических соединений, в которых детектирован хотя бы один обмен, что является основным ограничением предложенного подхода. Для подсчета максимально возможного количества обменов, реализован алгоритм, основанный на применении SMARTS шаблонов.

Для оценки эффективности такой фильтрации, для соединений с экспериментально наблюдаемыми обменами были получены списки изомерных структур из библиотеки PubChem. Для всех изомеров по структуре были посчитаны максимально возможные количества обменов, и те, для которых это значение оказалось меньше экспериментально измеренного, были отнесены к ложноположительным определениям. Результаты такой фильтрации для некоторых соединений представлены на рисунке 43. Используя данные изотопного обмена $^{16}\text{O}/^{18}\text{O}$ удалось отфильтровать 9-92% ложноположительных определений (медианное значение составило 62%, $n=45$).

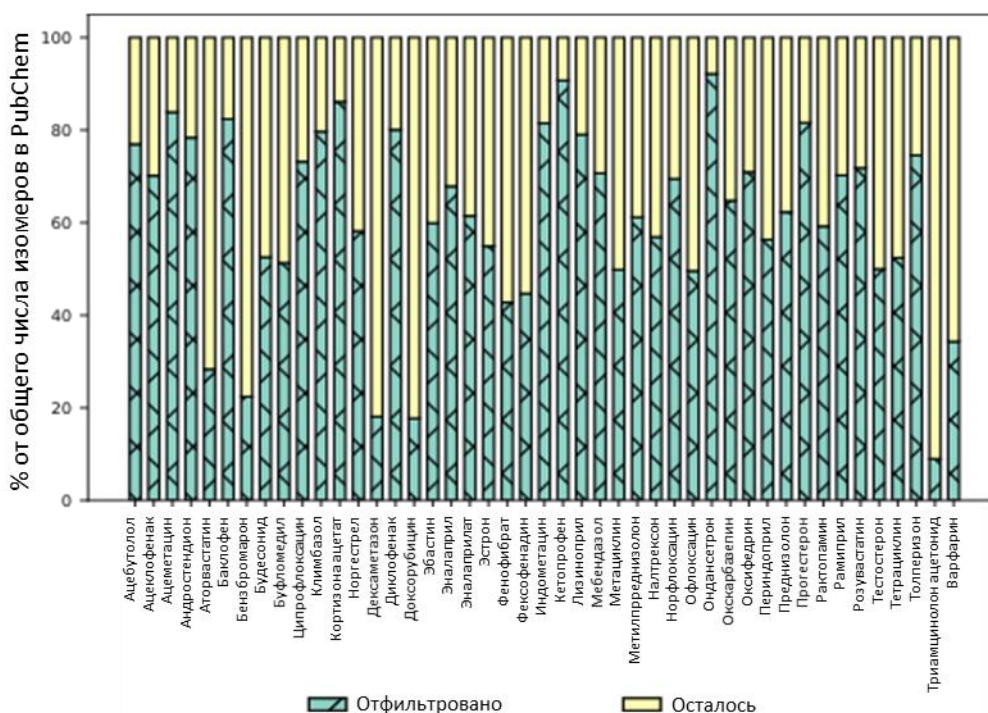


Рисунок 43. Фильтрация изомеров по данным $^{16}\text{O}/^{18}\text{O}$.

Дополнительно, в работе был предложен подход по сокращению пространства поиска среди изомеров с учетом данных тандемной масс-спектрометрии, в том числе с изотопным обменом. В отличие от H/D обмена, обмен $^{16}\text{O}/^{18}\text{O}$ приводит к внедрению устойчивой изотопной метки, которая сохраняется в фрагментах при диссоциации, индуцированной соударениями, и не

подвержена, например, явлению перескока (scrambling)[198]. Принцип подхода изображен на рисунке 44.

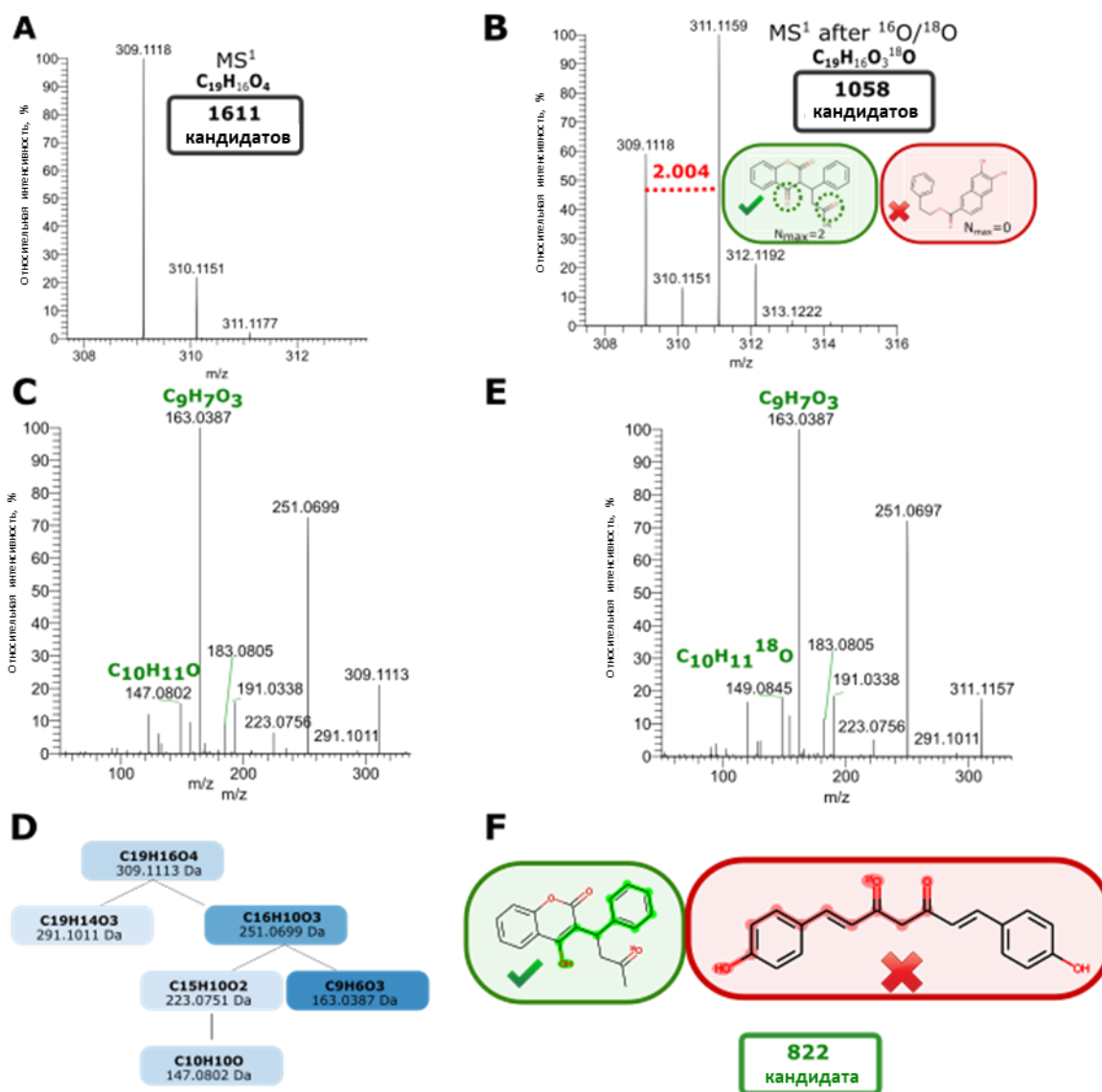


Рисунок 44. Схема предложенного подхода по фильтрации изомеров с учетом данных МС/МС и изотопного обмена.

На первом этапе необходимо аннотировать фрагментные ионы в масс-спектре брутто-формулами по точной измеренной массе фрагментов. В работе для этого использовали программное обеспечение SIRIUS 4[199]. Далее исходили из предположения (в общем случае неверного) об отсутствии перегруппировок и вторичных реакций. В этом случае, в молекулярном графе можно выделить связный подграф, соответствующий брутто-формуле фрагмента. Другими словами, фрагмент с определенной брутто-формулой можно «разместить» на структуре истинно-положительного кандидата. Если соответствующий фрагмент не удастся найти, то кандидат является ложноположительным. Для соответствующей проверки был разработан алгоритм, реализованный средствами библиотеки RDKit и SMARTS шаблонов. При этом на примере

варфарина показано, что введение тяжелого изотопа ^{18}O позволяет отфильтровать ложноположительные определения, которые без изотопной метки отфильтровать не удастся (Рисунок 45). Так как возможность перегруппировок при диссоциации, индуцируемой соударениями не исключена, для фильтрации имеет смысл рассматривать более тяжелые фрагменты, так как они меньше подвержены вторичным реакциям.

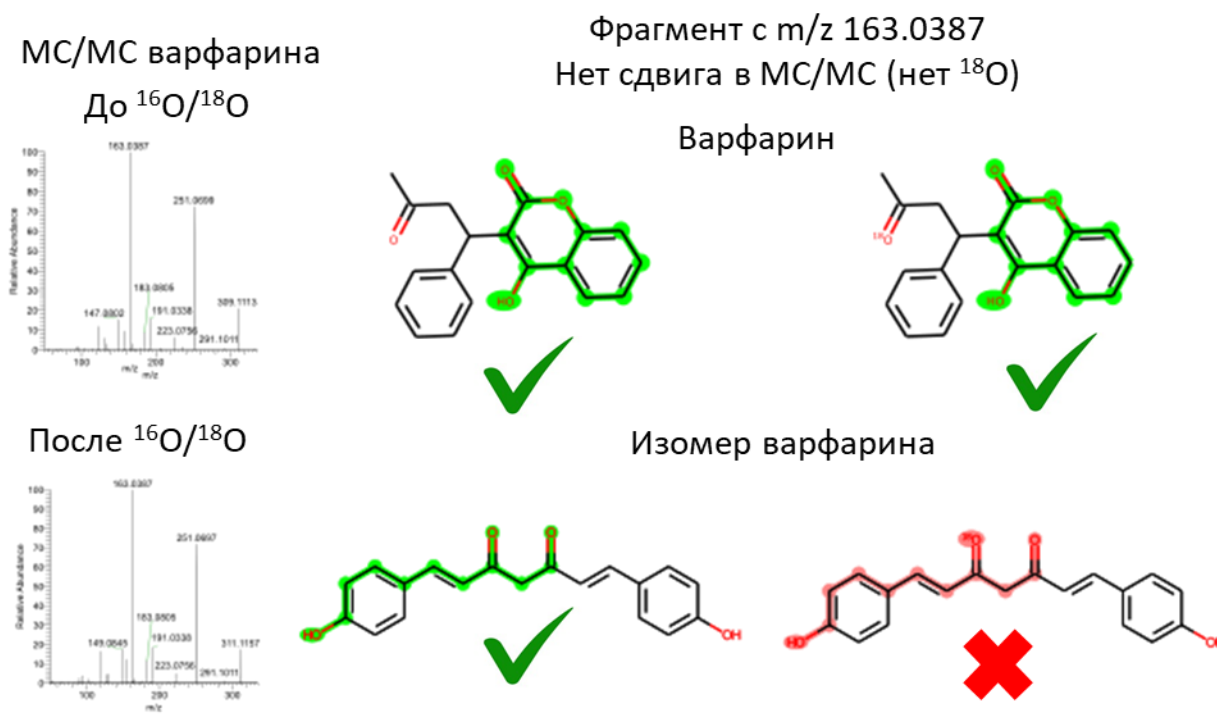


Рисунок 45. Фильтрация изомеров по данным МС/МС после изотопного обмена $^{16}\text{O}/^{18}\text{O}$ на примере изомеров варфарина.

Так, в случае мебендазола удалось отфильтровать 31.7% ложноположительных определений, при учете трех наиболее тяжелых фрагментов (Рисунок 46). Однако, при рассмотрении всех фрагментов был получен ложноотрицательный результат. Так, ион с m/z 95.0491 ($\text{C}_6\text{H}_7\text{O}^+$) не содержит изотопную метку, после проведения реакции изотопного обмена $^{16}\text{O}/^{18}\text{O}$ в молекуле мебендазола, и не может быть размещен на структуре мебендазола- ^{18}O без перегруппировки атомов кислорода. Однако, миграция кислорода из карбаматной группы в бензольное кольцо является крайне маловероятным сценарием. Поэтому, было сделано предположение о том, что этот ион образуется вследствие присоединения остаточной воды в ячейке соударений.

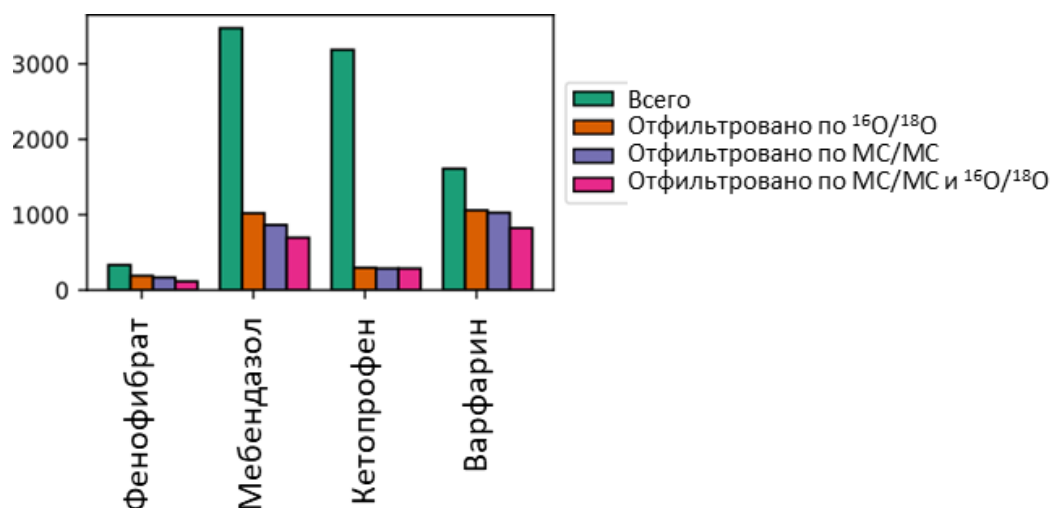


Рисунок 46. Пример сокращения пространства поиска при использовании предложенного подхода.

Для проверки этого предположения, был проведен гидролиз мебендазола в щелочных условиях, с образованием мебендазол-амина, с последующей реакцией изотопного обмена, содержащего только один атом кислорода. При этом в спектре мебендазол-амина содержится ион с m/z 95.0491 ($\text{C}_6\text{H}_7\text{O}^+$), и он также не содержит атом кислорода-18 после реакции изотопного обмена (Рисунок 47). Но так как в структуре мебендазол-амина нет атомов кислорода-16 после изотопного обмена, то атом кислорода в этом фрагменте очевидно появляется из среды, при вторичной реакции гидратации. При этом нужно отметить, что ион $\text{C}_6\text{H}_7\text{O}^+$ обнаруживается не только в спектрах, полученных в работе, но и в библиотечных спектрах мебендазола и мебендазол-амина. О возможности вторичных реакций с остаточной водой в ячейке соударений сообщалось и ранее, но проведенное исследование с использованием реакции обмена $^{16}\text{O}/^{18}\text{O}$ является примером подтверждения таких предположений[200].

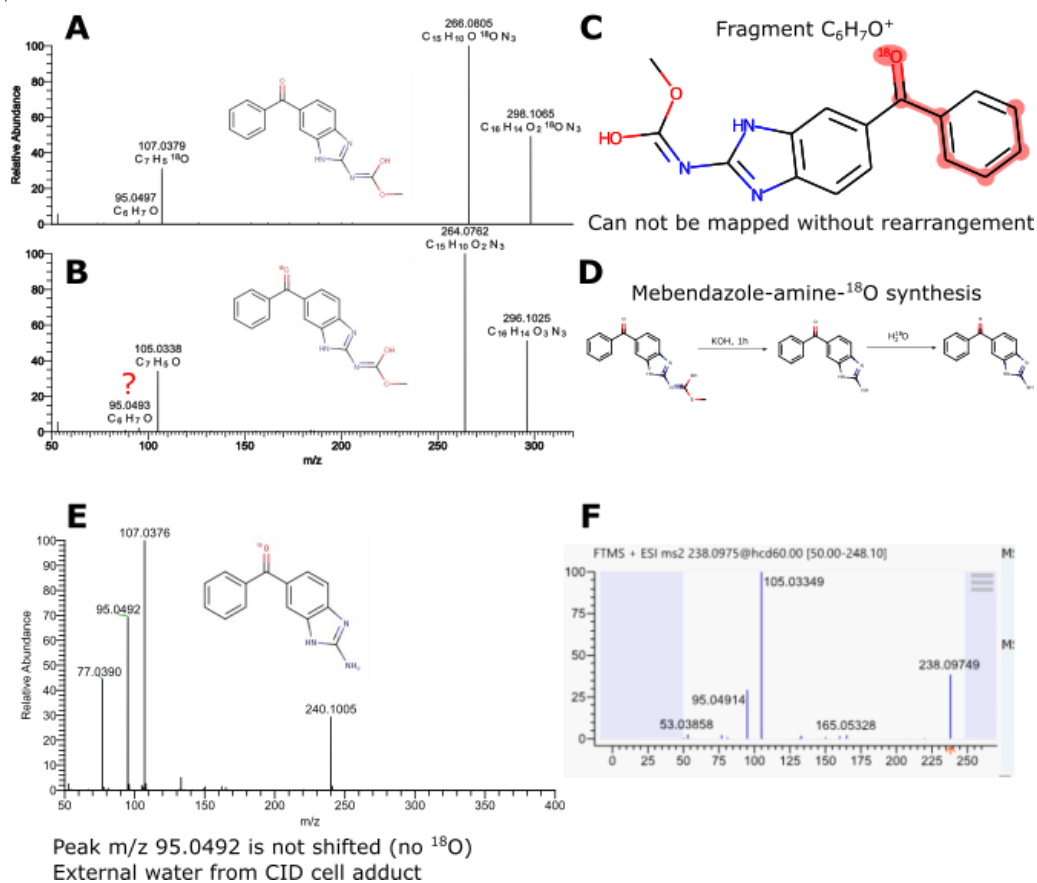


Рисунок 47. Аннотация иона с m/z 95.0492 в МС/МС мебендазола как продукта реакции гидратации внутри ячейки соударений.

Для работы с данными изотопного обмена $^{16}\text{O}/^{18}\text{O}$ разработано ПО на языке Python, обладающее следующими возможностями: определение «обменных» групп в молекуле, генерация всех изотопно-меченных вариантов молекулы, которые могут образоваться в реакции изотопного обмена $^{16}\text{O}/^{18}\text{O}$, с учетом экспериментально наблюдаемого числа обменов. Также ПО имеет функционал по учету спектров ДИС. Данное ПО реализовано в виде Web-приложения с графическим интерфейсом, и доступно по адресу: <https://oxygen-isotope-exchange.anvil.app> (дата обращения 14.01.2023 г.). Исходный код доступен по адресу https://github.com/osv91/16O-18O_isotope_exchange (дата обращения 14.01.2023 г.).

4.3 Предсказание времен удерживания

Для предсказания времен удерживания модель МО на основе ИНС с распространением сообщений была до-обучена в режиме обучения с переносом на внутрилабораторном наборе данных по удерживанию, измеренных в тех же условиях хроматографического разделения, которые использовались при анализе образцов по селективности изотопного обмена $^{16}\text{O}/^{18}\text{O}$. Соединения, в которых наблюдался изотопный обмен не включали в обучающую выборку, которая включала времена удерживания 461 молекулы и представлена в приложении.

Получившуюся модель оценивали в режиме кросс-валидации ($n=5$) и по отдельной независимой тестовой выборке, результаты кросс-валидации представлены в таблице 17.

Таблица 17. Результаты кросс-валидации ($n=5$) нейронной сети с распространением сообщения на внутрилабораторном наборе данных

		Валидационная выборка	Тестовая выборка
Среднее отклонение, с	абсолютное	124.2±12.4	120.2±5.2
Медианное отклонение, с	абсолютное	80.7±9.5	86.4±16.2
Среднее отклонение, с	квадратичное	178.9±22.0	169.0±6.2
R^2		0.727±0.08	0.732±0.02

Далее, для соединений, которые вступили в реакцию изотопного обмена $^{16}\text{O}/^{18}\text{O}$ были загружены списки изомеров из базы данных PubChem, и проведена их фильтрация по схеме, описанной ранее в разделе 3.1.3. В качестве порогового значения по ROC кривой выбрали 22.5%, площадь под ROC кривой равна 0.68. Далее, отфильтрованные списки изомеров подвергали дальнейшей фильтрации по результатам изотопного обмена $^{16}\text{O}/^{18}\text{O}$. В результате, при фильтрации только по данным кислородного обмена сокращение пространства поиска составило в среднем 29.9%, а при дальнейшей фильтрации по данным изотопного обмена в среднем 74.2%. Результаты по каждому соединению приведены в таблице 18.

Таблица 18. Результаты фильтрации изомеров с использованием данных изотопного обмена $^{16}\text{O}/^{18}\text{O}$ и предсказанных времен удерживания

Название	Номер в PubChem (CID)	Время удерживания, с	N	Всего изомеров	После фильтрации по времени удерживания	После фильтрации по $^{16}\text{O}/^{18}\text{O}$
Ацебутолол	1978	8.94	1	5851	4287	988
Ацеклофенак	71771	17.5	2	816	653	195
Ацеметацин	1981	17.82	2	455	310	50
Андростендион	6128	15.59	2	3846	2625	568
Аторвастатин	60823	18.15	2	145	120	86
Баклофен	2284	4.03	2	3768	2923	515
Бензбромарон	2333	20.88	1	76	58	45
Будесонид	5281004	15.72	2	1564	767	364
Буфломедил	2467	9.39	1	8371	6380	3107

Ципрофлоксацин	2764	7.33	1	2660	2001	537
Климбазол	37907	13.56	1	6115	5270	1072
Кортизона ацетат	5745	14.57	3	1632	821	114
Норгестрел	13109	16.5	1	3233	2272	950
Дексаметазон	5743	13.24	2	379	116	95
Диклофенак	3033	17.25	2	1489	1374	274
Доксорубицин	31703	11.33	1	266	62	51
Эбастин	3191	19.13	1	241	207	83
Эналаприл	5388962	11.46	2	4545	3179	1022
Эналаприлат	5462501	7.86	2	6176	4322	1666
Эстрон	5870	15.28	1	5726	4859	2191
Фенофибрат	3339	22.23	1	411	332	190
Фексофенадин	3348	13.88	1	580	414	229
Индометацин	3715	17.47	2	1736	1361	252
Кетопрофен	3825	14.7	3	3791	3187	296
Лизиноприл	5362119	10.66	1	1688	1216	255
Мебендазол	4030	11.97	1	3767	3475	1019
Метациклин	54675785	9.96	1	557	325	163
Метилпреднизолон	6741	13.15	2	2307	1146	445
Налтрексон	5360515	4.19	1	13239	9528	4106
Норфлоксацин	4539	6.9	1	2345	1733	529
Офлоксацин	4583	7.09	1	1353	1039	524
Ондансетрон	4595	9.61	1	7683	6854	538
Окскарбазепин	34312	11.23	1	4922	4623	1629
Оксифедрин	5489013	11.19	1	12847	9860	2867
Периндоприл	107807	12.01	1	778	490	214
Преднизолон	5755	16.91	2	2102	1182	446
Прогестерон	5994	18.24	2	3702	2091	385
Рактопамин	56052	7.46	1	9296	6978	2847
Рамиприл	5362129	13.32	1	1965	1293	385
Розувастатин	446157	14.85	3	203	117	33
Тестостерон	6013	15.04	1	4889	3263	1632
Тетрациклин	54675776	7.57	1	818	439	209
Толперизон	5511	10.19	1	15632	13805	3513

Триамцинолон ацетонид	6436	13.79	1	329	89	81
Варфарин	54678486	15.7	1	1895	1611	1058

Заключение к главе 4.

Изотопный обмен $^{16}\text{O}/^{18}\text{O}$ является эффективным способом выбора между изомерными структурами при идентификации химических соединений в нецелевом хромато-масс-спектрометрическом анализе. В работе проведено исследование селективности обмена для установления функциональных групп, способных вступать в обмен, эти группы определены, хотя селективность обмена невысока. Тем не менее в работе удалось предложить подход к фильтрации изомеров с применением данных изотопного обмена. Кроме того, показано, что совместное применение такой фильтрации и методов предсказания времен удерживания позволяет обеспечить существенное сокращение пространства поиска.

Реакция изотопного обмена может быть проведена в относительно мягких условиях (37°C , 24 ч), при повышении температуры степень обмена увеличивается. Хотя метод изотопного обмена не позволяет получить информацию о соединениях, которые нестабильны в условиях проведения реакции, на общее количество детектированных компонентов он не влияет. Данный метод можно рассматривать как разновидность химической дериватизации, не изменяющей основные свойства соединений, в первую очередь времена удерживания. Это в свою очередь существенно упрощает интерпретацию результатов, что особенно важно при изучении многокомпонентных образцов.

ГЛАВА 5. Предсказание индексов удерживания веществ, относящихся к спискам Конвенции по запрещению химического оружия³

Данная часть работы посвящена решению задачи предсказания ИУ в газовой хроматографии. История моделирования ИУ в газовой хроматографии насчитывает несколько десятилетий[118, 202]. Относительно новым направлением является создание универсальных моделей, основанных на машинном и глубоком обучении, с применением больших обучающих выборок, включающих молекулы различных классов. Несомненным преимуществом таких моделей является их универсальность, производительность и возможность предсказывать ИУ для соединений разных классов. Алгоритмы, используемые в таких моделях хорошо проработаны, и в диссертационной работе не удалось предложить универсальные подходы, характеризующиеся более высокой точностью при работе с библиотекой NIST Retention index library в качестве обучающей выборки.

В то же время, газовая хромато-масс-спектрометрия является основным методом в строго регулируемых областях таких как контроль за соблюдением Конвенции по запрещению химического оружия. Идентификация веществ из списков Конвенции по запрещению химического оружия проводится с использованием библиотечных масс-спектров электронной ионизации и ИУ. Ввиду особой важности задачи, справочные значения собираются в центральной аналитической базе данных Организации по запрещению химического оружия (ОСАД)[201], и перед включением в базу данных тщательно проверяются экспертной комиссией.

Однако база данных ОСАД ограничена по покрытию списков Конвенции по запрещению химического оружия. Так, база данных ОСАД версии 21 (2019 г.) содержит 5292 значения ИУ для 4482 химических веществ. В тоже время, только рассмотрение всех соединений, покрываемых списками 1.A.1– 1.A.12 Конвенции дает общее количество возможных структур, превышающее 1 300 000. Кроме того, списки веществ, относящихся к Конвенции, периодически пополняются. Так, в 2020 г добавлены четыре новых ряда соединений 1.A.13–1.A.16. Это еще сильнее усложняет процесс пополнения базы данных, и идентификацию соединений, относящихся к Конвенции.

³ При подготовке данной главы диссертации использована публикация, выполненная автором лично или в соавторстве, в которой, согласно Положению о присуждении ученых степеней в МГУ отражены основные результаты, положения и выводы исследования. Kireev A., **Osipenko S.**, Mallard G., Nikolaev E., Kostyukevich Y. Comparative Prediction of Gas Chromatographic Retention Indices for GC/MS Identification of Chemicals Related to Chemical Weapons Convention by Incremental and Machine Learning Methods // Separations. – 2022. – Т. 9, № 10. – С. 265 (Импакт-фактор Web of Science – 3.344, Q3) 50%.

Особенностью задач, относящихся к строго регулируемым областям (допинг-контроль, судебная и токсикологическая экспертиза, контроль за производством и применением химического оружия), является повышенная ответственность при принятии решения, и непоправимые последствия, особенно при ложноположительном результате. Поэтому, требования, предъявляемые к идентификации в этих областях существенно выше, чем при идентификации химических соединений, например, в нецелевом метаболомном анализе. Тем не менее, большое количество возможных соединений, определенных по комбинаторным правилам, при отсутствии образцов сравнения, и невозможности их синтеза, вынуждают обращаться к вычислительным подходам для оценки ИУ и масс-спектров электронной ионизации. Однако, к этим подходам предъявляются высокие требования по точности, а зачастую и интерпретируемости результатов предсказаний.

Следуя общепринятому стандарту, оценка универсальных моделей проводилась в режиме кросс-валидации и с применением независимых тестовых выборок. При этом, точность предсказаний ИУ для молекул, не имеющих структурных аналогов в обучающей выборке может быть существенно ниже той, что определена в режиме кросс-валидации со случайным разбиением данных. Учитывая, что соединения из списков Конвенции по запрещению химического оружия в основном относятся к фосфорорганическим соединениям, необходимо оценить точность предсказаний универсальных моделей для данного класса соединений.

5.1 Оценка применимости методов глубокого обучения для предсказания индексов

удерживания соединений из списков Конвенции по запрещению химического оружия

В рамках диссертационной работы, проводилась оценка существующих подходов к предсказанию ИУ на основе методов глубокого обучения, с целью проверить возможность использования предсказанных значений в экспертной практике. С помощью ранее опубликованных моделей одномерных[133] и двумерных сверточных ИНС[132], обученных по данным библиотеки NIST Retention index library, были предсказаны ИУ соединений, для которых в базе данных OCAD доступна экспериментальная информация по удерживанию. При этом, в виду ограничений этих нейросетевых подходов, часть молекул из базы данных OCAD не использовали при оценке. Так, при обучении одномерной сверточной сети, обучающая выборка была ограничена соединениями, содержащими в составе химические элементы Si, C, H, O, N, P, S, Cl, Br, F, I, и не могла предсказывать, например, ИУ селен содержащих соединений, присутствующих в OCAD.

Дополнительно, был оценен подход, на основе ИНС с архитектурой Transformer-CNN[67]. Эта ИНС, предварительно обученная в режиме обучения с частичным привлечением учителя, доступна для решения различных задач предсказания молекулярных свойств. Для предсказания

ИУ, в рамках диссертационной работы ИНС была до-обучена на данных библиотеки NIST 17 Retention Index Library. Были отобраны молекулы, содержащие только элементы C, N, O, F, Si, P, S, Cl, Se, H и As с приведенными значениями ИУ для неполярных (standard non-polar, semi-standard non-polar) неподвижных фаз. После чего, с помощью до-обученной модели также были предсказаны ИУ соединений из базы данных OCAD.

Отклонения, полученные при сравнении ИУ, предсказанных разными подходами, и экспериментальных ИУ из библиотеки OCAD представлены в таблице 19, а на рисунке 48 приведены распределения величины отклонения. Среднее отклонение составило 40-52 единицы, что существенно превышает значения отклонений, полученных при кросс-валидации этих моделей и опубликованных ранее. Можно отметить, что лучший результат получен при применении модели на основе одномерных сверточных ИНС.

Таблица 19. Результаты предсказания индексов удерживания соединений из центральной аналитической базы данных Организации по запрещению химического оружия

Модель	Среднее абсолютное отклонение, единиц	Медианное абсолютное отклонение, единиц	Среднее относительное отклонение, %	Медианное относительное отклонение, %
Одномерная сверточная ИНС	39.95	28.77	2.68	1.88
Двумерная сверточная ИНС	51.46	38.00	3.32	2.53
ИНС с архитектурой Transformer-CNN	48.11	33.45	3.23	2.17

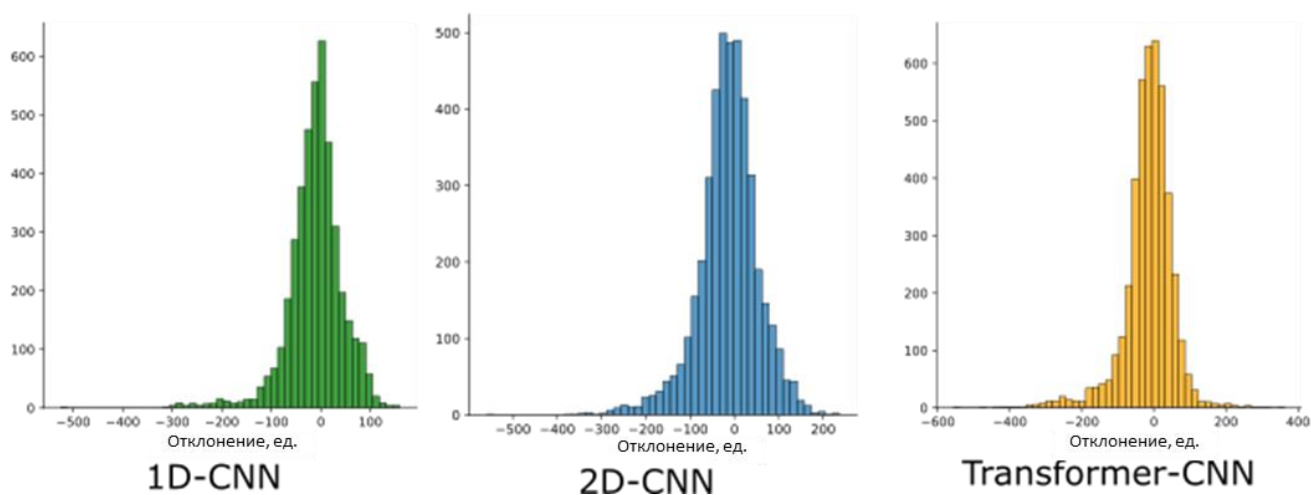


Рисунок 48. Распределение ошибок при предсказании индексов удерживания соединений из базы OCAD. 1D-CNN — одномерная сверточная сеть, 2D-CNN — двумерная сверточная сеть, Transformer-CNN — искусственная нейронная сеть с архитектурой Transformer-CNN.

В целом, для независимых тестовых выборок, авторы сообщали о погрешности, более высокой, чем полученной в режиме кросс-валидации. Это может быть связано со структурными отличиями молекул тестовой и обучающей выборки. Такие отличия можно оценить, например, по диаграммам, полученным с применением метода стохастического вложения соседей с t -распределением к фрагментным дескрипторам молекул (t-distributed Stochastic Neighbor Embedding)[203]. Из соответствующей диаграммы, построенной для молекул из библиотеки NIST17 и OCAD, можно видеть, что молекулы из списков Конвенции по запрещению химического оружия структурно отличаются от молекул из библиотеки NIST 17 RI (Рисунок 49).

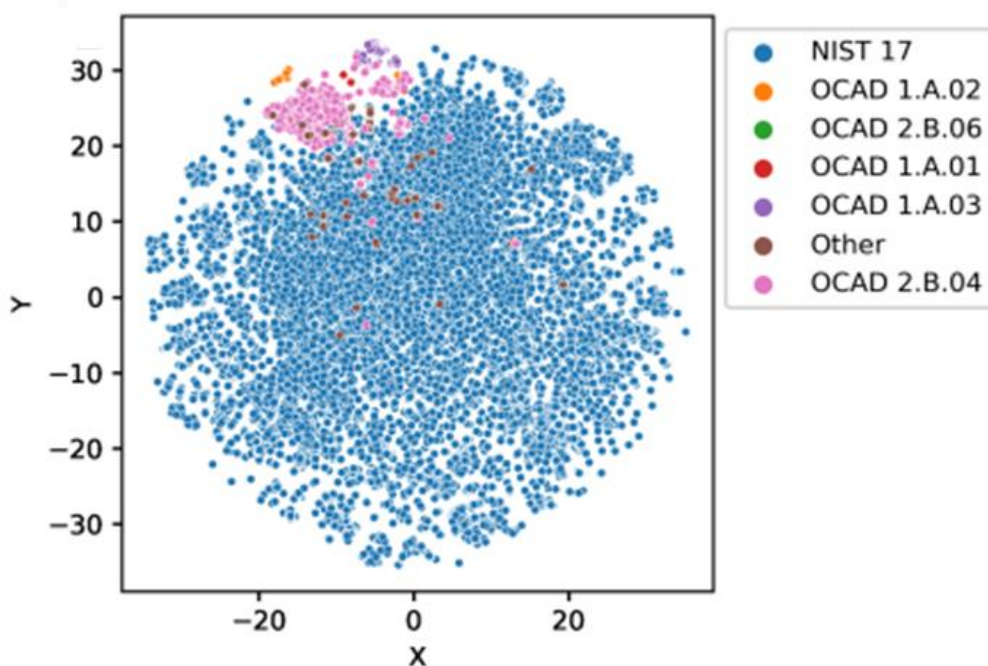


Рисунок 49. Проекция фрагментных дескрипторов молекул из библиотеки NIST17 и базы данных OCAD, построенная методом стохастического вложения соседей с t -распределением. Цветами отмечены соединения из разных списков Конвенции по запрещению химического оружия. X и Y соответствуют координатам, определяющим плоскость проекции преобразованного пространства фрагментных дескрипторов.

Экспериментальные отличия ИУ, измеренных в условиях повторяемости обычно составляют несколько единиц. Так, в исследовании межлабораторной воспроизводимости ИУ новых психоактивных веществ сообщалось об относительных отклонениях, не превышающих 0.35% [204]. В целом считается, что стандартное отклонение при измерении ИУ на неполярных неподвижных фазах составляет несколько единиц, и увеличивается для полярных подвижных фаз до 10-50 единиц [205]. Поэтому результаты, полученные с применением ИНС можно считать неудовлетворительными. Отчасти, ограниченную точность можно объяснить тем, что при обучении моделей использовали средние значения ИУ, измеренные в разных температурных

режимах, и зачастую для различных неподвижных фаз, в то время как данные в библиотеке OCAD стандартизованы с точки зрения условий газохроматографического разделения.

5.2 Повышение точности предсказания индексов удерживания за счет применения более специфичной модели, основанной на алгоритме градиентного бустинга

Для повышения точности предсказаний было предложено построение более специфичной модели, обученной на данных базы OCAD. Так как количество молекул в обучающей выборке не превышало 5000, ожидалось, что использование глубоких ИНС приведет к переобучению. Поэтому, был использован метод ГБ в сочетании с физико-химическими дескрипторами из библиотеки Mordred[97].

Для моделирования, молекулы из базы данных OCAD были представлены SMILES строками, которые были стандартизованы средствами библиотеки RDKit, при этом были отобраны молекулы, содержащие только элементы C, N, O, F, Si, P, S, Cl, Se, H и As. В основном были отфильтрованы изотопно-меченные молекулы, содержащие дейтерий. Гиперпараметры были выбраны методом полного перебора (Grid Search), сетка параметров при переборе представлена в таблице 20. Для параметров максимальной глубины дерева, числа деревьев, скорости обучения и регуляризации (γ) были выбраны значения 4, 1000, 0.05 и 1 соответственно. Для остальных параметров использовали значения по умолчанию.

Таблица 20. Сетка значений гиперпараметров при их выборе методом полного перебора (Grid Search)

Гиперпараметр	Значения
Максимальная глубина дерева (Max tree depth)	2,4,6,8,10
Число деревьев (Number of trees)	100, 200, 500, 1000, 2000, 5000
Скорость обучения (Learning rate)	0.01, 0.05, 0.075, 0.1, 0.2, 0.3
Регуляризация (Gamma)	0, 0.5, 1, 5, 10

В режиме кросс-валидации ($n=5$) среднее абсолютное отклонение составило 16.2 ± 0.9 единиц. Нужно отметить, что, хотя использование специфичного моделирования позволило снизить величину отклонения в среднем на 20 единиц, по сравнению с универсальными моделями, его нужно применять с осторожностью.

5.3 Инкрементный подход к моделированию индексов удерживания соединений из списков Конвенции по запрещению химического оружия

Особенностью списков Конвенции по запрещению химического оружия является организация соединений в виде гомологических рядов. В структурах молекул можно выделить ядро, и несколько боковых углеводородных цепей. При этом, можно отметить, что разница ИУ

между соединениями, отличающимися только одной боковой цепью, будет сохраняться в паре соединений, отличающихся другой боковой цепью (Рисунок 50). Данное наблюдение можно использовать для оценки ИУ одного соединения, если известны ИУ трех других. Такой подход, фактически является частным случаем инкрементного метода расчета ИУ, который широко применялся для предсказания ИУ, в том числе в программном обеспечении NIST[129]. Однако, в данном случае, задачу можно свести к поиску трех молекул-гомологов с известными ИУ, для оценки ИУ четвертой, т.е. без явного вычисления значений инкрементов.

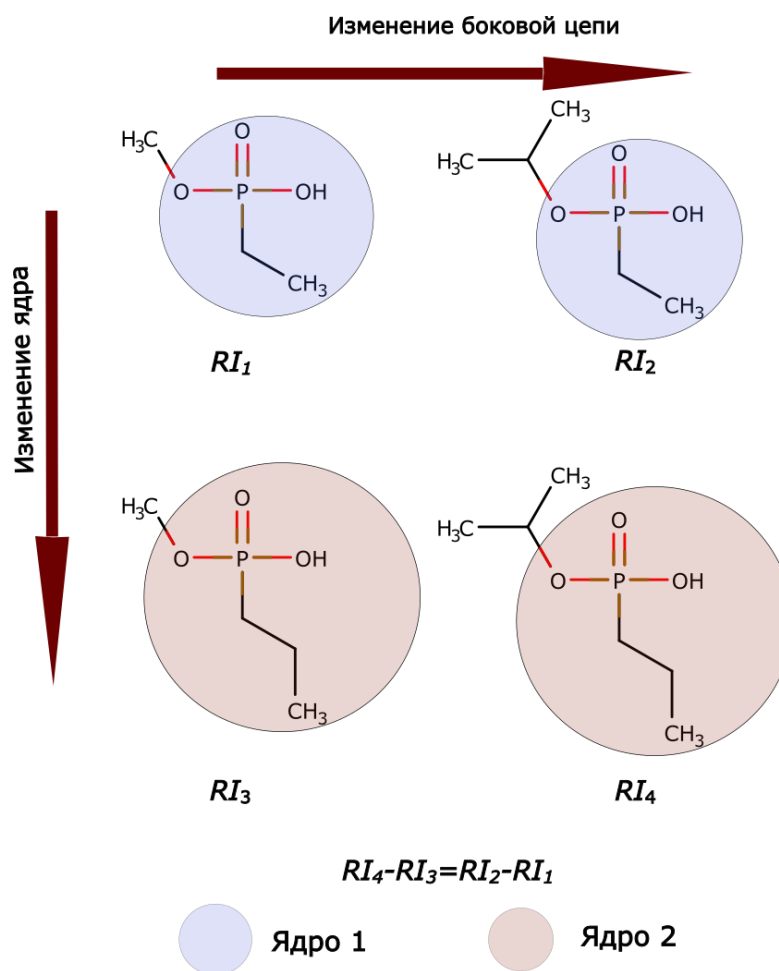


Рисунок 50. Принцип предложенного инкрементного метода по вычислению разностей индексов удерживания

В случае со списками Конвенции по запрещению химического оружия, решение этой задачи упрощается, так как молекулы уже организованы в ряды, в соответствии с ядром – ряд зарина, ряд VX, ряд зомана и т.д. При применении подхода к данным из OCAD, можно использовать организацию по названиям ИЮПАК, как представлено на рисунке 51.

Example data from OCAD

RO linked to phosphorus	Methyl phosphonofluoridate	Ethyl phosphonofluoridate	Isopropyl phosphonofluoridate	Propyl phosphonofluoridate
1-Propylcyclohexyl	1420	1517	Missing value 1	Missing value 2
1,2-Dimethylbutyl	1087	1180	1231	1264
Difference	333	337	335±3	335±3

$$\text{Missing value 1} = 1231 + 335 = 1566$$

$$\text{Missing value 2} = 1264 + 335 = 1599$$

Рисунок 51. Пример определения индекса удерживания предложенным инкрементным методом.

Ожидаемо, инкрементный метод оказался более точным, чем методы МО. В таблице 21 приведены средние отклонения между предсказанными и экспериментальными значениями списка 1.А.01 Конвенции по запрещению химического оружия для модели на основе одномерной свёрточной сети, специфичной модели на основе ГБ и инкрементного метода. Для объективного сравнения модель ГБ была обучена на базе данных OCAD с исключенным списком 1.А.01, который использовался как тестовая выборка.

Таблица 21 Средние отклонения, полученные при применении различных подходов к предсказанию индексов удерживания к соединениям из списка 1.А.01 Конвенции по запрещению химического оружия. Отдельно рассчитаны значения отклонений в рядах с метильным, этильным и пропильным радикалами при атоме фосфора

Выборка	Одномерная свёрточная ИНС	ГБ	Инкрементный метод
Метил – Р	35	39	4.0
Этил – Р	11	21	1.8
Пропил – Р	52	20	3.4
1.А.01	30	27	3.0

В работе было предсказаны ИУ некоторых соединений из списка 1.А.01 Конвенции по запрещению химического оружия, отсутствующие в базе данных OCAD. Их значения приведены в таблице 22.

Таблица 22. Рассчитанные индексы удерживания для некоторых соединений из списка 1.А.01 Конвенции по запрещению химического оружия

Радикал у атома кислорода	Пропил–Р	Изопропил–Р	Этил–Р	Метил–Р
1-пропилциклогексил	1600	1565	Данные есть в OCAD	Данные есть в OCAD
1-пропилгептил	1637	1602	Данные есть в OCAD	Данные есть в OCAD
2,2-диметилциклогексил	1494	1460	1409	Данные есть в OCAD
2,2-диметилгексил	1436	1402	1351	Данные есть в OCAD
2,3-диметилпентил	1390	1356	1305	Данные есть в OCAD

2,5-диметилциклогексил	1500	1465	Данные есть в OCAD	Данные есть в OCAD
------------------------	------	------	--------------------	--------------------

Для удобства поиска необходимых молекулярных пар, в работе был предложен алгоритм, реализованный средствами библиотеки RDKit с применением шаблонов SMARTS [186]. Теоретически, его можно применять для построения предсказаний не только в сочетании с базой данных OCAD, но и другими библиотеками. На первом этапе определяется ядро молекулы последовательным отщеплением от структуры молекулы концевых метильных групп, или алифатических углеводородных циклов (например, циклогексила). Оставшаяся структура далее рассматривается как ядро. Удалением ядра из молекулы определяются боковые цепи. Далее происходит организация базы данных: для каждой молекулы определяются ядро и боковые цепи.

Определяются ядро и боковые цепи молекулы, для которой необходимо предсказать ИУ. Далее алгоритм ищет молекулы, по которым можно построить предсказание, как изображено на рисунке 48. При нахождении минимум трех необходимых молекул, ИУ будет рассчитан через соответствующую разность. Если требуемые данные отсутствуют, предсказание будет построено с помощью описанной ранее модели ГБ.

На базе предложенного подхода разработано Web-приложение с графическим интерфейсом для поиска молекул, необходимых для предсказания инкрементным методом в базе данных OCAD. В случае их отсутствия приводится значение, предсказанное моделью ГБ. Приложение доступно по адресу <https://ri-cwc.anvil.app> (Последнее обращение 08 января 2023 г.).

Заключение к главе 5.

Проведена оценка возможности предсказаний ИУ соединений, относящихся к спискам Конвенции по запрещению химического оружия различными методами, основанными на глубоком обучении. Среднее отклонение от известных экспериментальных данных составило 40–52 единицы, что намного превышает экспериментальную вариабельность ИУ. Можно сделать вывод о том, что применение универсальных моделей, разработанных в последнее время нежелательно в областях с высокой степенью ответственности. В работе разработана более специфичная модель, основанная на алгоритме ГБ. Будучи обученной на данных из базы OCAD, она повышает точность предсказаний почти на 20 единиц, хотя и ограничена по области применения. Также предложена реализация инкрементного метода анализа через расчет разностей ИУ соединений-гомологов, и показано, что этот метод существенно точнее, чем методы МО. Отклонения рассчитанных этим способом ИУ от библиотечных значений составляют в среднем 3 ед., для соединений из списка 1.A.01 Конвенции по запрещению химического оружия., что сопоставимо с экспериментальной вариабельностью. Для применения

инкрементного метода реализован алгоритм автоматического поиска требуемых молекулярных пар, и разработано Web-приложение с графическим интерфейсом. Приложение доступно по адресу <https://ri-cwc.anvil.app> (Последнее обращение 08 января 2023 г.).

Необходимо отметить, что обучение и оценка предложенной модели ГБ для предсказания индексов удерживания проводилась в режиме кросс-валидации с использованием соединений из базы данных OCAD. Для оценки достоверности предсказаний при практическом применении может быть целесообразно проведение сравнения молекулярного подобия исследуемых молекул и молекул из базы данных OCAD, так как при отсутствии структурных аналогов исследуемой молекулы в обучающей выборке, точность предсказания может существенно снизиться.

ГЛАВА 6. Предсказание масс-спектров электронной ионизации с помощью машинного обучения

Масс-спектры электронной ионизации являются основной характеристикой молекул, используемой при идентификации химических соединений в нецелевом анализе методом газовой хромато-масс-спектрометрии[206]. Для повышения эффективности ионизации обычно применяется достаточно высокая энергия электронного пучка, что как правило вызывает диссоциацию молекулы по связям с наименьшей энергией, а также по возможности образования стабильных ионов. Это делает масс-спектры электронной ионизации относительно информативными и характеристичными, так как они отражают особенности структуры молекулы.

Стандартизация энергии электронов позволяет получать схожие масс-спектры на различных приборах. Причем, хорошей воспроизводимостью отличается не только качественный состав масс-спектров, но и соотношение интенсивностей сигналов, соответствующих различным фрагментным ионам. Существуют подходы к интерпретации масс-спектров электронной ионизации, для восстановления молекулярной структуры. Они включают аннотацию распространенных фрагментных ионов (например, m/z 77 соответствует наличию в молекуле одного или нескольких бензольных колец, m/z 91 — толуолу), ионов, содержащих галогены, и определяемых по характерному изотопному соотношению, распространенных нейтральных потерь (например, $M-18$ — отщепление воды). Кроме того, в ионном источнике образующиеся ионы могут вступать в различные реакции, среди которых наиболее известна перегруппировка Мак-Лафферти, характерная для соединений, содержащих кето-группу[207]. Известны и другие перегруппировки[208, 209]. Нужно отметить, что зачастую, степень диссоциации молекул при электронной ионизации настолько высока, что молекулярный ион отсутствует в масс-спектрах. Очевидно, что интерпретация масс-спектров для установления структуры молекул является трудоемким процессом, требующим высокой экспертной квалификации химика-аналитика.

Гораздо большей популярностью пользуется подход к идентификации химических соединений при помощи библиотечного поиска. Хорошая воспроизводимость масс-спектров электронной ионизации способствовала развитию библиотек, которые курируются государственными институтами, частными компаниями или экспертными сообществами. Для идентификации измеренный масс-спектр компонента сравнивается с масс-спектрами в библиотеке, из которых выбираются те, которые демонстрируют наибольшее сходство с экспериментальным масс-спектром по той или иной метрике. Далее, при необходимости, несколько наиболее вероятных кандидатов оцениваются экспертами, и итоговая структура подтверждается дополнительными методами и встречным синтезом. Такой подход имеет

существенный недостаток, заключающийся в отсутствии в библиотеках масс-спектров большинства молекул. Очевидно, что если масс-спектр соединения в библиотеке отсутствует, то данный подход не сможет идентифицировать это соединение в составе образца.

Эта проблема может быть отчасти решена при дополнении существующих библиотек расчетными данными. До настоящего времени активно развиваются квантово-химические методы моделирования масс-спектров электронной ионизации. Несмотря на то, что они достигли приемлемой точности предсказаний, по крайней мере для соединений определенных классов, они не применимы для создания больших расчетных библиотек, так как требуют очень существенных вычислительных ресурсов. В отличие от квантово-химических расчетов, методы МО требуют значительных ресурсов только при обучении моделей, при этом время, требуемое для вычислений с применением обученной модели пренебрежимо мало. Поэтому одной из задач диссертационной работы было создание основанного на МО подхода к моделированию масс-спектров электронной ионизации, который можно было бы применять для создания больших расчетных библиотек.

6.1 Описание подхода к предсказанию масс-спектров электронной ионизации с применением машинного обучения

Задача предсказания масс-спектров электронной ионизации рассматривалась как задача многоцелевой регрессии. Модель, получая на вход вектор признаков, описывающих молекулярную структуру, выдает на выходе спектральный вектор (I_1, I_2, \dots, I_n) , где I_k — интенсивность пика с $m/z = k$ в масс-спектре электронной ионизации, а n — верхняя граница диапазона сканирования (определяемая как масса наиболее тяжелого молекулярного иона в библиотеке с изотопами). Такая постановка задачи в случае масс-спектров электронной ионизации удобна, так как большая часть доступных масс-спектров электронной ионизации получена на приборах с квадрупольными масс-анализаторами низкого разрешения, и представлена в библиотеках с целочисленным представлением m/z .

В качестве вектора признаков использовали бинарный вектор круговых фрагментных дескрипторов, с длиной вектора 2048 и радиуса 4. В работе проводили предварительную оценку возможности использования физико-химических дескрипторов из библиотеки Mordred[97], однако, точность получившейся модели оказалась ниже. Вероятно, это связано с тем, что фрагментные дескрипторы связаны со структурой молекулы наиболее очевидным образом.

В качестве обучающей выборки использовали библиотеку NIST 20. Масс-спектры электронной ионизации в этой библиотеке представлены в двух базах данных — *mainlib*, содержащей спектры 306873 индивидуальных соединений, и *replib*, содержащей 43774 дополнительных спектра для молекул из *mainlib*. Масс-спектры из NIST 20 были преобразованы

с помощью скрипта на языке программирования Python, молекулярные структуры были преобразованы в InChI с помощью RDKit. Из обучающей выборки были исключены 4110 молекул, для которых InChIKey, сгенерированные в RDKit, отличались от InChIKey, приведенных в NIST 20. Кроме того, были исключены 4 молекулы, в спектрах которых были ионы с m/z более 1400, чтобы обеспечить адекватную длину выходного спектрального вектора. Наконец, 26717 масс-спектров, репликаты которых были представлены в *replib*, составили отдельную тестовую выборку. В итоге, 276042 масс-спектра из *mainlib* сформировали обучающую выборку.

Для построения модели использовали алгоритм ГБ, в реализации XGBoost[46], поддерживающей (начиная с версии 1.6.0) алгоритмы многоцелевой регрессии и классификации. Подбор некоторых гиперпараметров проводили в автоматическом режиме с помощью программного обеспечения Optuna[210], их значения приведены в таблице 23. Для остальных гиперпараметров использовали значения, установленные по умолчанию. В качестве функции потерь по умолчанию используется среднее квадратичное отклонение. Для предотвращения переобучения использовали механизм досрочной остановки, который прекращал обучение, если значение функции потерь не уменьшалось после 10 эпох. Все модели оценивали в режиме кросс-валидации ($n=5$).

Таблица 23. Выбранные гиперпараметры модели градиентного бустинга для предсказания масс-спектров электронной ионизации

Гиперпараметр	Значение
Скорость обучения (learning rate)	0.25
Выборка (Subsample)	0.75
Максимальная глубина дерева (Max tree depth)	15

Для оценки качества модели в режиме кросс-валидации определяли косинусную меру сходства векторов, и её взвешенный аналог, вычисляемые по формулам:

$$\text{Cosine similarity } (I_q, I_l) = \frac{\sum_{k=1}^M I_{qk} \cdot I_{lk}}{\left\| \sum_{k=1}^{M_q} I_{qk}^2 \right\| \left\| \sum_{k=1}^{M_l} I_{lk}^2 \right\|}$$

$$\text{Weighted cosine similarity } (I_q, I_l) = \frac{\sum_{k=1}^M m_k I_{qk}^{0.5} \cdot m_k I_{lk}^{0.5}}{\left\| \sum_{k=1}^{M_q} (m_k I_{qk}^{0.5})^2 \right\| \left\| \sum_{k=1}^{M_l} (m_k I_{lk}^{0.5})^2 \right\|},$$

где I_q и I_l вектора сравниваемых масс-спектров, m - вектор значений m/z . Веса, учитывающие положение спектрального пика при расчете косинусной меры сходства, увеличивают вклад ионов с более высокими m/z . Это особенно важно при выполнении библиотечного поиска, т.к. более тяжелые фрагменты являются наиболее характеристичными. Хотя существуют различные подходы к учету m/z при оценке сходства масс-спектров, в диссертационной работе использовали

наиболее широко применяемый подход, где в формулу сходства значения m/z входят с показателем степени 1, а значения интенсивности с показателем степени 0.5[211]. Данный подход реализован при вычислении сходства в программном обеспечении MS Search (NIST), которое является наиболее популярным средством библиотечного поиска.

Результаты оценки модели по определенным выше метрикам представлены в таблице 24. Можно отметить, что, хотя для валидационной и обучающей выборок среднее значение взвешенной косинусной меры сходства между библиотечными и предсказанными масс-спектрами было достаточно высоким, для независимой тестовой выборки среднее значение метрик немного превысило 0.6. Отчасти, такое различие между выборками может объясняться тем, что в качестве независимой тестовой выборки использовали масс-спектры молекул, для которых в *replib* присутствуют репликаты, а не случайную подвыборку. С другой, при выборе гиперпараметров, который проводился в режиме кросс-валидации, эта тестовая выборка не использовалась, и потому она дает более независимую оценку.

Таблица 24. Результаты кросс-валидации модели ($n=5$) предсказания масс-спектров электронной ионизации

	Взвешенная косинусная мера сходства	Косинусная мера сходства
Обучающая выборка	0.798±0.003	0.875±0.001
Валидационная выборка	0.798±0.003	0.875±0.001
Тестовая выборка	0.609±0.001	0.633±0.001

Полученные результаты можно признать неудовлетворительными, так как при идентификации химических соединений обычно используют пороговое значение 0.6 для сходства экспериментального и библиотечного спектра. Поэтому для генерации *in-silico* библиотек масс-спектров необходимо повышать точность модели.

6.2 Предсказание спектра нейтральных потерь и усредненная модель

Рассмотрение масс-спектров, предсказанных с помощью предложенной модели МО, показало, что, в целом справляясь с предсказанием интенсивностей легких ионов, модель плохо работает в правой части масс-спектра. Низкомолекулярные фрагменты, которые определяют левую часть спектра, являются довольно общими для широкого круга молекул. Например, ион с m/z 77, соответствующий иону бензольного кольца, присутствует в масс-спектрах большинства ароматических молекул. Ионы в правой части спектра, в основном образуются в результате нейтральных потерь, как правило, стабильных молекул. Например, нейтральная потеря с m/z 18 может соответствовать отщеплению воды, а уменьшение m/z на 35 – отщеплению хлора. В

результате одних и тех же нейтральных потерь образуются ионы с разными значениями m/z . Эти ионы редко встречаются в обучающей выборке, что негативно сказывается способности модели к обобщению и на общей точности. Для решения этой проблемы предложено предсказывать спектр нейтральных потерь, который получается при вычитании m/z ионов из молекулярной массы молекулы (Рисунок 52). Вектор спектра нейтральных потерь получался следующим образом. Интенсивность иона с m/z X помещалась в позицию $(M_w - X + \tau)$, где M_w соответствует молекулярной массе молекулы, а τ параметр сдвига, позволяющий учесть пики с m/z выше, чем M_w , например изотопные пики ^{13}C or ^{37}Cl . В работе параметр использовали $\tau=10$, а интенсивности пиков со значением m/z , превышающим молекулярную массу молекулы более чем на 10, приравнивались к нулю.

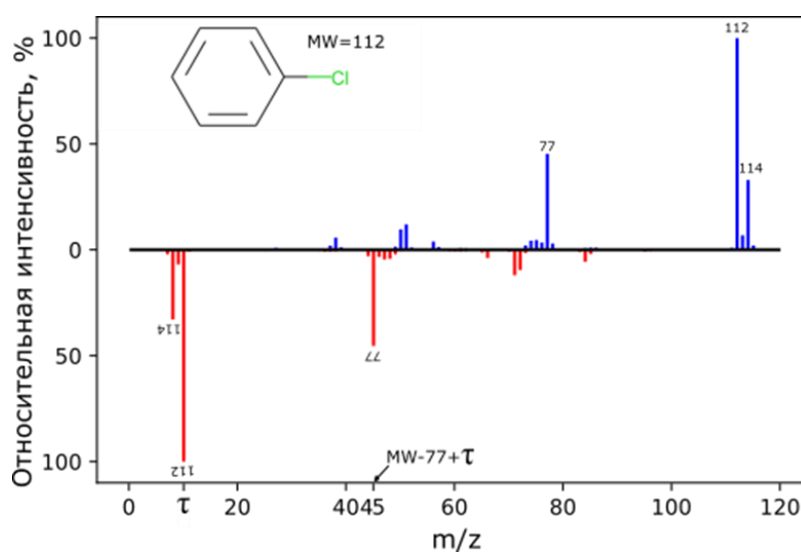


Рисунок 52. Масс-спектр электронной ионизации хлорбензола в стандартном виде (вверху), и в форме нейтральных потерь (внизу).

Такой подход был предложен ранее, в работах по предсказанию масс-спектров электронной ионизации с помощью ИНС [168, 169]. Однако, в этих работах предсказание масс-спектров в непосредственной форме и в форме нейтральных потерь проводилось одной ИНС, с обобщением первых слоев, и соответствующих признаков. При использовании алгоритма ГБ подобное обобщение невозможно. В работе была создана отдельная модель для предсказания спектра нейтральных потерь, для которой был подобран набор гиперпараметров (Таблица 25).

Таблица 25. Выбранные гиперпараметры модели градиентного бустинга для предсказания масс-спектров электронной ионизации в форме нейтральных потерь

Гиперпараметр	Значение
Скорость обучения (Learning rate)	0.25
Выборка (Subsample)	1
Максимальная глубина дерева (Max tree depth)	18

Использование модели нейтральных потерь закономерно повысило точность предсказаний при оценке по взвешенной метрике, однако, если судить по невзвешенной метрике, качество предсказанных масс-спектров немного ухудшилось (Таблица 26). Ожидается, модель нейтральных потерь хуже справлялась с предсказанием «легких» фрагментов.

Таблица 26. Результаты кросс-валидации модели предсказания спектров нейтральных потерь

	Взвешенная косинусная мера сходства	Косинусная мера сходства
Обучающая выборка	0.833±0.003	0.836±0.001
Валидационная выборка	0.833±0.003	0.836±0.002
Тестовая выборка	0.752±0.001	0.616±0.001

Наилучший результат дало усреднение результатов предсказаний прямой модели, и модели нейтральных потерь (после обратного преобразования спектра нейтральных потерь в традиционный вид). Примеры предсказанных масс-спектров прямой моделью, моделью нейтральных потерь и усредненных предсказаний представлены на рисунке 53.

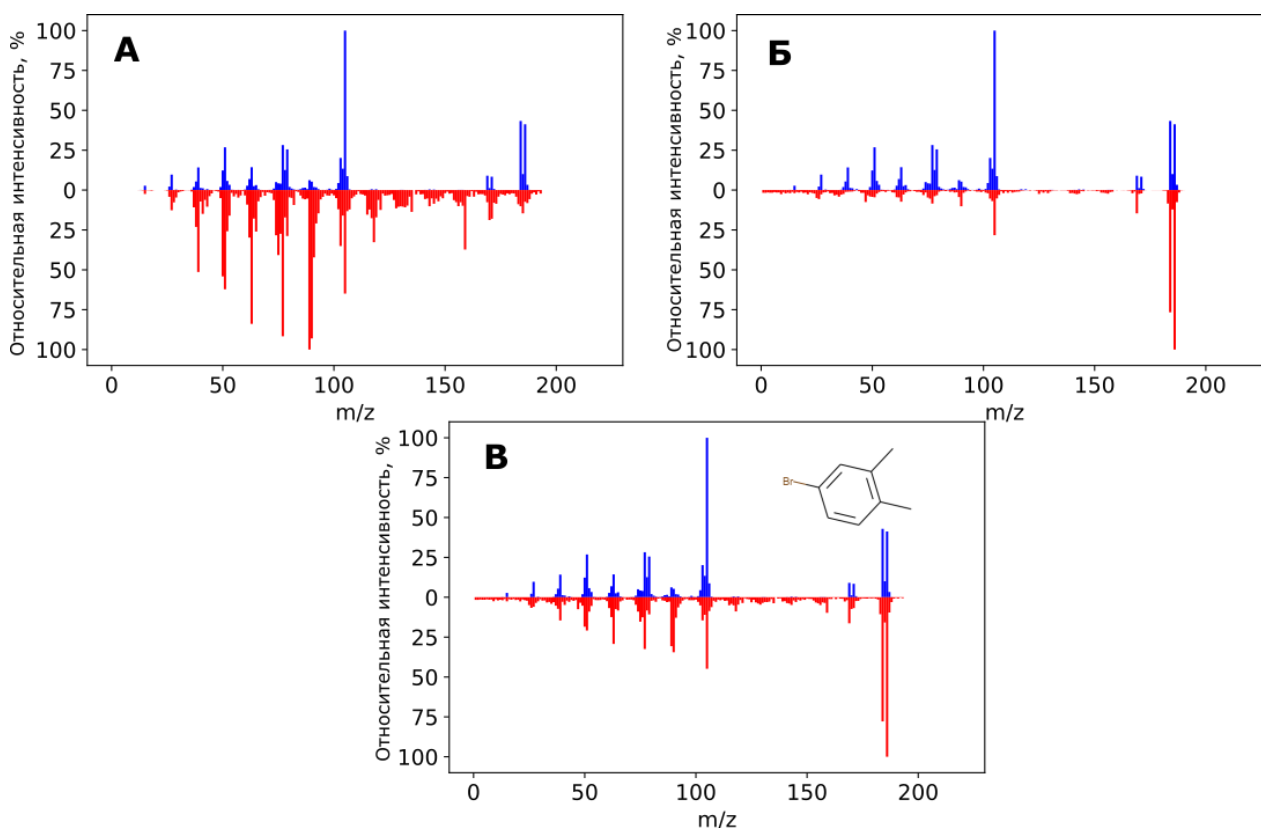


Рисунок 53. Примеры предсказания масс-спектров электронной ионизации 3-бром-орто-ксилола при использовании "прямой" модели (А), модели нейтральных потерь (Б), усредненных предсказаний (В).

Эксперименты по использованию усреднения с весовыми коэффициентами, применения линейных и нелинейных моделей для объединения результатов предсказаний прямой и обратной моделей не привели увеличению общей точности, при значительном повышении потребляемых вычислительных ресурсов. Поэтому, в качестве окончательного варианта был выбран подход с применением алгоритма ГБ для предсказания масс-спектра, спектра нейтральных потерь, и их усреднения, после обратного преобразования спектра нейтральных потерь в обычный вид. Данный подход был реализован в виде ПО GBEIMS (Gradient Boosting based Electron Ionization Mass Spectra prediction). Программное обеспечение GBEIMS написано на языке программирования Python 3, включает в себя веса обученных моделей, и для работы требует установки только свободно распространяемых библиотек. Работоспособность GBEIMS проверена в операционной системе Windows 10. Разработанное программное обеспечение доступно по ссылке <https://figshare.com/articles/software/GBEIMS/21538965> (Последнее обращение 07 января 2023 г.). Для предсказания масс-спектров необходим текстовый файл, содержащий InChI идентификаторы молекул (и, опционально, названия). На выходе будет сгенерирован текстовый файл, содержащий масс-спектры в формате MSP. Этот спектральный формат может быть использован с большинством программ библиотечного поиска. В работе была проверена поддержка генерируемых масс-спектров программой NIST MS Search версий 2.0 и 2.4.

Полученные при кросс-валидации средние метрики при использовании GBEIMS приведены в таблице 27. Можно видеть, что при использовании GBEIMS точность предсказаний возрастает почти на 25%. К сожалению, сравнить результаты предложенного подхода GBEIMS с результатами других подходов, основанных на методах МО по предложенным метрикам затруднительно, т.к. в публикациях по применению ИНС (NEIMS) и графовых сверточных ИНС для предсказаний масс-спектров электронной ионизации эти метрики не приведены. При их оценке основное внимание уделялось решению задачи создания искусственных библиотек, и возможности поиска по предсказанным спектрам. Оценка GBEIMS с этой точки зрения в диссертационной работе также была проведена, и сравнительные результаты приведены далее. Кроме того, в качестве обучающих выборок для этих подходов использовали другие версии библиотеки NIST. Также нужно отметить, что в случае графовых сверточных сетей, исходные коды и веса моделей не опубликованы.

Таблица 27. Результаты кросс-валидации усредненной модели предсказания масс-спектров электронной ионизации GBEIMS

	Взвешенная косинусная мера сходства	Косинусная мера сходства
Обучающая выборка	0.815±0.001	0.915±0.001
Валидационная выборка	0.851±0.001	0.915±0.005
Тестовая выборка	0.759±0.001	0.783±0.001

Распределение взвешенной косинусной меры сходства приведено на рисунке 54. Для 23% молекул взвешенная мера сходства превысила 0.85, и лишь для 10.6% оказалась менее 0.6. Учитывая, что в реальной практике сходство масс-спектров более 0.85 считается высоким, а при сходстве менее 0.6 кандидаты не рассматриваются, полученный результат можно считать вполне удовлетворительным.

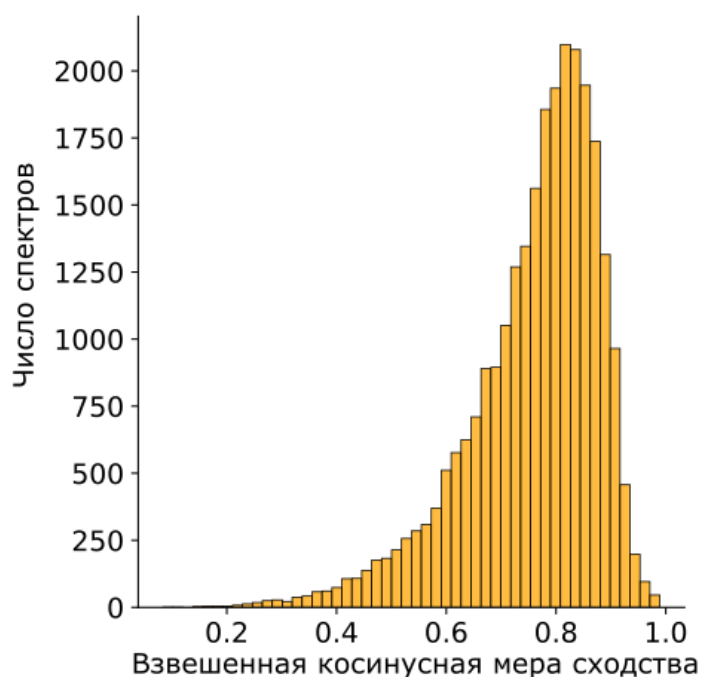


Рисунок 54. Распределение взвешенной косинусной меры сходства предсказанных и библиотечных масс-спектров молекул из тестовой выборки

6.3 Применение разработанной модели для создания in-silico спектральных библиотек

Основным направлением для применения разработанной модели предсказания масс-спектров электронной ионизации является пополнение существующих масс-спектральных библиотек расчетными масс-спектрами молекул, для которых получение экспериментальных данных по тем или иным причинам затруднительно. Поэтому, было необходимо удостовериться, что поиск экспериментально измеренных масс-спектров в нецелевом хромато-масс-

спектрометрическом анализе по таким синтетическим библиотекам будет корректно ранжировать истинно-положительные определения. В идеальном случае, истинно-положительное определение должно быть на первом месте в поисковой выдаче, и для оценки необходимо оценивать параметр полноты (Recall), определяемому как отношение количества истинно-положительных определений к общему количеству положительных определений. Однако, так как поиск по библиотекам дает лишь вероятностную идентификации химических соединений, и эксперты почти всегда изучают несколько первых позиций поисковой выдачи для принятия решения, может быть разумнее оценивать параметр Recall@N, т.е. процентное отношение истинно-положительных определений, попавших в первые N строк поисковой выдачи, к общему количеству положительных определений.

Для проведения такой оценки, экспериментальные масс-спектры из *mainlib* тех молекул, масс-спектры которых также есть в *replib*, были заменены на масс-спектры, предсказанные предложенной моделью. После чего были рассчитаны попарные взвешенные косинусные меры сходства всех масс-спектров из *replib*, и синтетической библиотеки, созданной на основе *mainlib* и включающей экспериментальные и расчетные масс-спектры. Для каждого спектра из *replib* был составлен список масс-спектров этой синтетической библиотеки, ранжированный по убыванию взвешенной косинусной меры сходства масс-спектров. Фактически, для всех масс-спектров из *replib* был проведен библиотечный поиск, и оценено количество истинно-положительных определений, расположенных в верхних N строках поисковой выдачи, при различных N.

Нужно заметить, что даже поиск масс-спектров из *replib* в библиотеке *mainlib*, не дает 100% значений Recall, и Recall@N. Это связано с тем, что масс-спектры в *replib* получены из разных источников, в некоторых случаях на разных приборах, и прошли различную подготовку перед включением в библиотеку. Даже несмотря на хорошую воспроизводимость масс-спектров электронной ионизации при энергии электронов в 70 эВ, возможна вариабельность и присутствие шума. Так Recall при поиске по библиотеке *mainlib* составил 80.9%, а по синтетической библиотеке 73.3%. Однако, более 90% истинно-положительных определений попали в верхние 5 строк поисковой выдачи, а Recall@10 составил 92.7% (Рисунок 55).

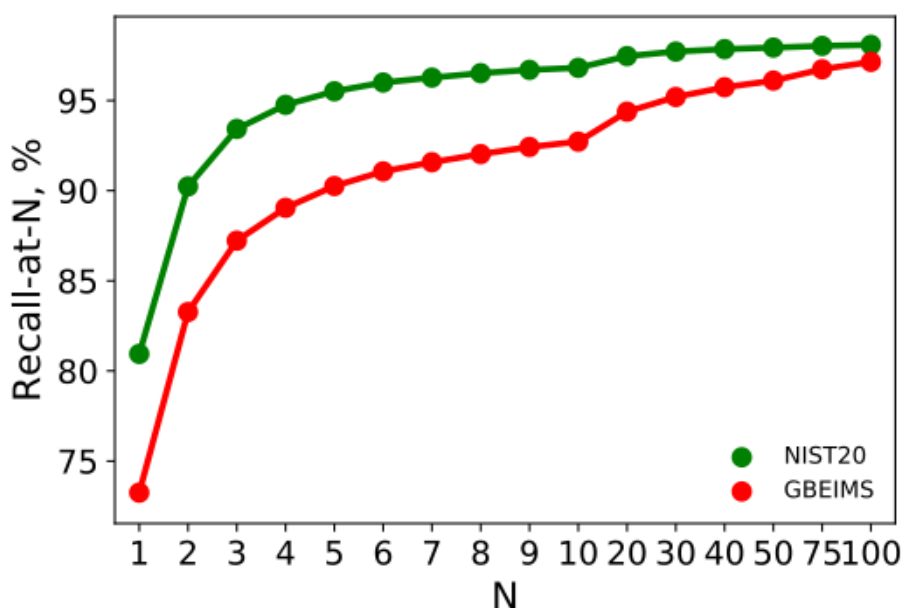


Рисунок 55. Процент истинно-положительных определений, попавших в верхние N строк поисковой выдачи (Recall-at-N) при поиске экспериментальных масс-спектров по библиотеке NIST20 и синтетической библиотеке, созданной с применением GBEIMS.

В сравнении с NEIMS[168], первым подходом к предсказанию масс-спектров электронной ионизации, предложенный подход демонстрирует сопоставимое значение Recall@10 (92% против 91.7% при оценке NEIMS), однако при оценке NEIMS при поиске использовался дополнительный фильтр по молекулярной массе. Авторы исходили из предположения, что молекулярную массу молекулы легко определить дополнительными методами исследования, и можно отфильтровывать кандидатов, у которых молекулярная масс сильно отличается от массы искомой молекулы. Однако, это не всегда так — зачастую определить молекулярную массу из масс-спектров электронной ионизации невозможно, и далеко не все соединения, определяемые в газовой хромато-масс-спектрометрии, можно определять жидкостной хромато-масс-спектрометрией. А без этого фильтра Recall@10 при использовании NEIMS оказывался ниже, чем при использовании предложенной модели и составлял 86%. Кроме того, при использовании предложенной модели, истинно-положительное определение чаще оказывалось на первом месте поисковой выдачи (73.3% против 54.3%). Отчасти это может быть связано с тем, что при выборе гиперпараметров NEIMS, авторы стремились максимизировать именно Recall@10, в то время как оптимизация гиперпараметров предложенной модели проводилась независимо от задачи библиотечного поиска.

6.4 Сравнение предложенного подхода с квантово-химическими расчетами масс-спектров электронной ионизации

Квантово-химические расчеты масс-спектров электронной ионизации долгое время являлись единственным универсальным методом моделирования масс-спектров электронной ионизации. Основным недостатком квантово-химических методов является высокое потребление вычислительных ресурсов, и как следствие, низкая производительность. Однако, за последнее десятилетие эти методы были существенно усовершенствованы, в том числе с точки зрения скорости вычислений. В частности, пакет QCEIMS[170] позволяет проводить вычисление масс-спектров электронной ионизации низкомолекулярных соединений в среднем за 42 часа[212, 213]. Это позволило оценивать их на относительно больших наборах данных.

В диссертационной работе проведено сравнение QCEIMS и предложенного подхода, основанного на алгоритме ГБ. Для этого из библиотеки Mass Bank of North America[11] были загружены масс-спектры электронной ионизации, рассчитанные с применением пакета QCEIMS. Расчетные масс-спектры 226 молекул, отсутствующих в обучающей выборке, и масс-спектры этих молекул, полученные с помощью разработанной модели, сопоставляли с экспериментальными, для определения средней косинусной меры сходства, которая составила 0.637 для QCEIMS и 0.854 для предложенного подхода (Рисунок 56). Результат оценки QCEIMS соответствует ранее опубликованным данным[214]. Нужно отметить, что для предсказания всех 226 масс-спектров методом МО потребовалось менее 1 минуты, т.е. МО позволяет не только получать более точные предсказания, но и существенно экономить вычислительные ресурсы, по сравнению с квантово-химическими расчетами.

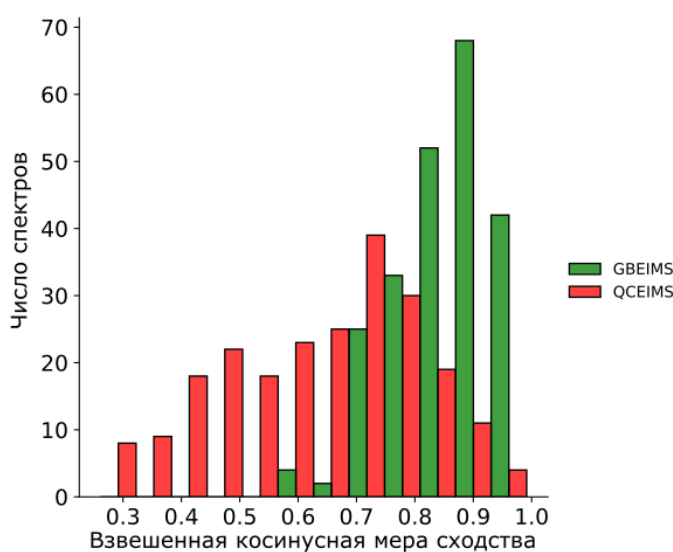


Рисунок 56. Сравнительное распределение взвешенной косинусной меры сходства библиотечных масс-спектров и масс-спектров, предсказанных квантово-химическим пакетом QCEIMS и предложенным подходом GBEIMS

В работе также было обнаружено, что квантово-химические расчеты с помощью пакета QCEIMS не могут интерпретировать некоторые пики в масс-спектрах. Например, в спектре зарина, рассчитанном в рамках диссертационной работы с помощью QCEIMS, пропущен пик с m/z 99, наблюдаемый в библиотечных спектрах. Его образование объясняется результатом перегруппировки Мак-Лафферти[215].

В отличие от квантово-химических расчетов, предложенный подход на основе МО использует информацию, полученную из обучающей выборки, и предсказывает ион, соответствующий перегруппировке Мак-Лафферти, вероятно потому, что это превращение широко представлено в обучающем наборе данных. Нужно отметить, что применение GBEIMS дает более шумные предсказанные масс-спектры, по сравнению с квантово-химическими расчетами пакетом QCEIMS (Рисунок 57). Отчасти эта проблема может быть решена применением фильтров по интенсивности. Так как подобные фильтры встроены в большинство программ для работы с масс-спектральными данными, этот вопрос подробно не рассматривался.

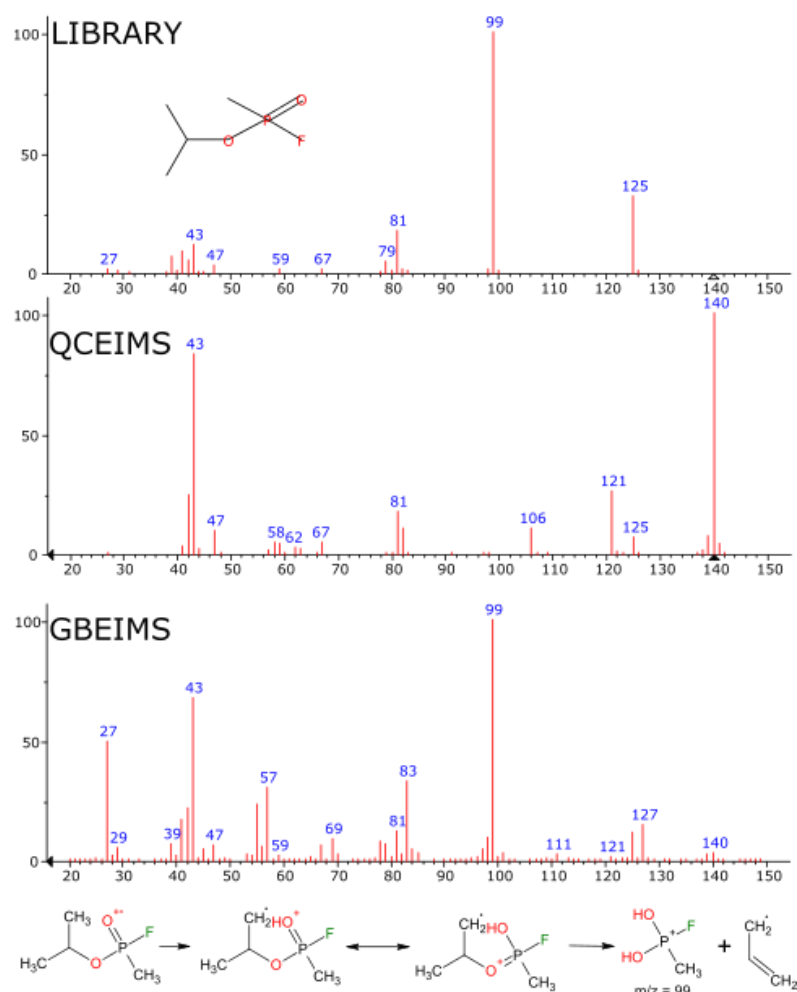


Рисунок 57. Сравнение библиотечного масс-спектра зарина (сверху), масс-спектра, предсказанного пакетом QCEIMS (в середине), и предложенным подходом GBEIMS (внизу). На рисунке также приведена схема образования иона с m/z 99

Заключение к главе 6

В работе предложен подход к моделированию масс-спектров электронной ионизации, с применением МО, а именно алгоритма ГБ. Задача предсказания масс-спектров рассматривалась как задача многоцелевой регрессии, где вектор зависимых переменных представлял собой вектор интенсивностей сигналов ионов в масс-спектре. Показано, что для повышения точности необходимо рассматривать масс-спектры не только в обычном виде, но и в форме масс-спектров нейтральных потерь. В работе предложено построение индивидуальных моделей предсказания масс-спектров и спектров нейтральных потерь. Показано, что при усреднении результатов предсказаний индивидуальных моделей, качество предсказанных масс-спектров повышается, среднее значение взвешенной косинусной мера сходства по тестовой выборке составляет для усредненной модели 0.76. Предложенный подход GBEIMS позволяет предсказывать масс-спектры электронной ионизации для пополнения существующих масс-спектральных библиотек. Показано, что поиск экспериментальных масс-спектров по таким библиотекам позволяет корректно определять молекулу более чем в 70% случаях, правильное определение попадает в первые 10 строк поисковой выдачи более чем в 90% случаях.

Необходимо отметить, что точность предсказаний масс-спектров электронной ионизации с помощью GBEIMS является недостаточной для проведения однозначной идентификации, которая требует совпадения масс-спектров со схожестью выше 800, а потому получаемые библиотеки могут быть использованы лишь для первичного скрининга возможных кандидатов. Тем не менее, предложенный подход по точности предсказаний превосходит существующие способы моделирования электронной ионизации, в частности квантово-химические расчеты с помощью пакета QCEIMS, и в перспективе может быть улучшен за счет дополнения известными эмпирическими правилами и закономерностями, а также учетом изотопных соотношений.

Разработанное программное обеспечение GBEIMS написано на языке Python и включает веса обученных моделей, а также скрипт для предсказания масс-спектров, принимающий InChI молекул в качестве входных данных. Разработанное программное обеспечение доступно по ссылке <https://figshare.com/articles/software/GBEIMS/21538965> (Последнее обращение 07 января 2023 г.).

ЗАКЛЮЧЕНИЕ

В результате проведенных исследований предложены новые подходы к моделированию аналитических характеристик низкомолекулярных соединений, используемых при идентификации химических соединений в нецелевом хромато-масс-спектрометрическом анализе, с применением методов МО. Разработаны модели предсказания времен удерживания с применением трех различных алгоритмов машинного и глубокого обучения — ГБ, ИНС с архитектурой трансформер, и ИНС с распространением сообщений. Для обучения моделей и получения предсказаний для описания молекул использовались фрагментные дескрипторы, текстовые представления молекул в виде строк SMILES, а также представления молекул в виде графа, соответственно. В качестве обучающей выборки использовали библиотеку времен удерживания METLIN SMRT. Среднее отклонение, определенное в режиме кросс-валидации ($n=5$) составило 32 с для наиболее точной из предложенных моделей, что сопоставимо с заявленной вариабельностью времен удерживания в этой библиотеке. Для практического применения разработаны подходы пересчета предсказанных времен удерживания для одних условий хроматографического разделения, на другие условия хроматографического разделения. Один подход основан на применении кусочно-линейных функций пересчета, другой — на технике обучения с переносом. Способы пересчета предсказаний оценивали с использованием доступных наборов данных по хроматографическому удерживанию в режиме кросс-валидации ($n=5$) при использовании обучения с переносом, или посредством независимой тестовой выборки. Также предложен подход к фильтрации изомерных кандидатов по предсказанным временам удерживания при идентификации химических соединений, с определением порогового значения по ROC-кривым. С использованием такого подхода удалось отфильтровать в среднем 23-53% ложноположительных результатов.

Также в работе оценивали эффективность совместного применения фильтрации по предсказанным временам удерживания и по данным изотопного обмена $^{16}\text{O}/^{18}\text{O}$. Предложены условия проведения реакции изотопного обмена и изучена селективность реакции на большом наборе кислородсодержащих молекул. Установлены функциональные группы, в которых возможен обмен, а также предложен алгоритм фильтрации изомерных кандидатов при идентификации химических соединений с учетом данных изотопного обмена. Кроме того, получен набор данных по удерживанию 472 соединений для обучения модели предсказания времен удерживания. Фильтрация по предсказанным временам удерживания позволила отфильтровать 29% кандидатов, а их совместное применение - 74% кандидатов.

Проведено сравнение существующих универсальных моделей предсказания ИУ в газовой хроматографии применительно к соединениям, относящимся к спискам Конвенции по

запрещению химического оружия. Показано, что средние отклонения предсказанных значений от экспериментальных достаточно велики и не позволяют применять такие модели в реальной практике по анализу в рамках Конвенции. В работе предложена более специфичная модель ГБ, обученная на соединениях из базы данных OCAD. Ее применение позволило снизить среднее отклонение до 16 единиц. Дальнейшее увеличение точности предсказаний возможно с применением инкрементного метода по предсказанию ИУ удерживания внутри гомологических серий. Для этого необходимо находить в базе данных молекулярные пары с определенными боковыми цепями, для реализации такого поиска в работе предложен алгоритм с использованием SMARTS шаблонов.

Заключительная часть работы посвящена моделированию масс-спектров электронной ионизации. Предложен подход к их предсказанию с применением алгоритма ГБ. Метрики, полученные в режиме кросс-валидации модели, показали высокое сходство экспериментальных и предсказанных масс-спектров. Продемонстрировано, что модель может быть использована для создания библиотек масс-спектров для применения в нецелевом анализе. Стандартный поиск по таким библиотекам позволяет корректно определять соединение более чем в 70% процентах случаях, более чем в 90% истинное определение находится в первых десяти строках поисковой выдачи.

Основные ограничения предложенных методов связаны с составом обучающих выборок, использованных при обучении моделей машинного обучения. Так, при прогнозировании времен удерживания в жидкостной хроматографии была применена библиотека METLIN SMRT, включающая преимущественно гетероциклические соединения и ароматические соединения. Предложенная для прогнозирования индексов удерживания модель обучена на данных библиотеки OCAD, содержащей информацию о структурно-схожих молекулах, относящихся к фосфорорганическим, сераорганическим, и галогенорганическим соединениям. Как правило, точность предсказаний с помощью машинного обучения зависит от наличия в обучающей выборке структурных аналогов исследуемого вещества. Выявление таких аналогов возможно с помощью изучения молекулярного подобия известными методами.

Таким образом, в работе разработан набор подходов к моделированию различных аналитических характеристик низкомолекулярных соединений, используемых при идентификации химических соединений, и показано, что предсказанные величины могут быть использованы там, где экспериментально определенные справочные значения недоступны, по крайней мере сокращая пространство поиска. Все подходы реализованы с использованием открытых библиотек для языков программирования Python и R, все алгоритмы также находятся в свободном доступе.

ВЫВОДЫ

В диссертационной работе были получены следующие научные результаты:

1. Предложено три подхода к предсказаниям времен хроматографического удерживания низкомолекулярных соединений в обращенно-фазовой жидкостной хроматографии с применением градиентного бустинга, нейронной сети с архитектурой Трансформер и графовой нейронной сети с распространением сообщений. Графовая нейронная сеть с распространением сообщений превосходит другие алгоритмы по точности предсказаний, среднее отклонение по обучающей выборке METLIN SMRT 31.5 с сопоставимо с прецизионностью измерений времен удерживания соединений из этой выборки.
2. Применение метода обучения с переносом позволяет прогнозировать времена удерживания химических соединений в различных условиях разделения с использованием обучающих выборок небольшого размера (несколько сотен соединений). Точность таких прогнозов характеризуется средним отклонением 9.5–205 с при общем времени хроматографического разделения 6-60 мин. Использование предсказанных значений для идентификации химических соединений в нецелевом анализе позволяет сократить пространство поиска до 50%.
3. Точность разработанной реализации инкрементного метода прогнозирования индексов удерживания в газовой хроматографии характеризуется средним отклонением 5 единиц при работе с соединениями-гомологами из списков Конвенции по запрещению химического оружия. В тех случаях, где предложенный инкрементный метод неприменим, можно использовать дополняющую модель машинного обучения, основанную на алгоритме градиентного бустинга, точность которой характеризуется средним отклонением 16 ед., полученным в режиме кросс-валидации с помощью библиотеки OCAD. Тем не менее, для молекул, структурно отличающихся от соединений из OCAD ожидается ухудшение точности предсказаний.
4. Установлены функциональные группы (карбонильная, карбоксильная группы, гидроксильная группа в аллильном и бензильном положении), способные вступать в реакцию изотопного обмена $^{16}\text{O}/^{18}\text{O}$ при инкубации в течение 24 ч при температуре 37°C и 95°C. Предложен подход к использованию данных изотопного обмена для идентификации химических соединений в нецелевом анализе. Совместное применение фильтров по предсказанным временам удерживания и данным изотопного

обмена повышает эффективность фильтрации ложноположительных определений до 75%.

5. Предложенный подход GBEIMS для предсказания масс-спектров электронной ионизации характеризуется средней взвешенной косинусной мерой сходства 0.759, определенной по независимой тестовой выборке, состоящей из соединений из библиотеки NIST 20, и превосходит по точности прогнозов методы моделирования масс-спектров с помощью пакета QCEIMS, основанного на квантово-химических расчетах. Хотя схожесть предсказанных и экспериментальных спектров не позволяет проводить однозначную идентификацию, поиск по расчетным библиотекам масс-спектров способствует определению списка возможных кандидатов.

СПИСОК ЛИТЕРАТУРЫ

1. Viant M. R., Kurland I. J., Jones M. R., Dunn W. B. How close are we to complete annotation of metabolomes? // *Current Opinion in Chemical Biology*. – 2017. – Т. 36. – С. 64-69.
2. Dunn W. B., Erban A., Weber R. J. M., Creek D. J., Brown M., Breitling R., Hankemeier T., Goodacre R., Neumann S., Kopka J., Viant M. R. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics // *Metabolomics*. – 2013. – Т. 9, № 1. – С. 44-66.
3. Sumner L. W., Amberg A., Barrett D., Beale M. H., Beger R., Daykin C. A., Fan T. W. M., Fiehn O., Goodacre R., Griffin J. L., Hankemeier T., Hardy N., Harnly J., Higashi R., Kopka J., Lane A. N., Lindon J. C., Marriott P., Nicholls A. W., Reily M. D., Thaden J. J., Viant M. R. Proposed minimum reporting standards for chemical analysis // *Metabolomics*. – 2007. – Т. 3, № 3. – С. 211-221.
4. Milman B. L. General principles of identification by mass spectrometry // *TrAC Trends in Analytical Chemistry*. – 2015. – Т. 69. – С. 24-33.
5. Milman B. L. Identification of chemical compounds // *TrAC Trends in Analytical Chemistry*. – 2005. – Т. 24, № 6. – С. 493-508.
6. mzCloud- Advanced Mass Spectral Database. – URL: <https://www.mzcloud.org> (дата обращения: 28 декабря 2022).
7. Smith C. A., Maille G. O., Want E. J., Qin C., Trauger S. A., Brandon T. R., Custodio D. E., Abagyan R., Siuzdak G. METLIN: A Metabolite Mass Spectral Database // *Therapeutic Drug Monitoring*. – 2005. – Т. 27, № 6.
8. Guijas C., Montenegro-Burke J. R., Domingo-Almenara X., Palermo A., Warth B., Hermann G., Koellensperger G., Huan T., Uritboonthai W., Aisporna A. E., Wolan D. W., Spilker M. E., Benton H. P., Siuzdak G. METLIN: A Technology Platform for Identifying Knowns and Unknowns // *Analytical Chemistry*. – 2018. – Т. 90, № 5. – С. 3156-3164.
9. Xue J., Guijas C., Benton H. P., Warth B., Siuzdak G. METLIN MS2 molecular standards database: a broad chemical and biological resource // *Nature Methods*. – 2020. – Т. 17, № 10. – С. 953-954.
10. Horai H., Arita M., Kanaya S., Nihei Y., Ikeda T., Suwa K., Ojima Y., Tanaka K., Tanaka S., Aoshima K., Oda Y., Kakazu Y., Kusano M., Tohge T., Matsuda F., Sawada Y., Hirai M. Y., Nakanishi H., Ikeda K., Akimoto N., Maoka T., Takahashi H., Ara T., Sakurai N., Suzuki H., Shibata D., Neumann S., Iida T., Funatsu K., Matsuura F., Soga T., Taguchi R., Saito K., Nishioka T. MassBank: a public repository for sharing mass spectral data for life sciences // *Journal of Mass Spectrometry*. – 2010. – Т. 45, № 7. – С. 703-714.
11. MoNA - MassBank of North America. – URL: <https://mona.fiehnlab.ucdavis.edu> (дата обращения: 28 декабря 2022 г.).

12. Wang M., Carver J. J., Phelan V. V., Sanchez L. M., Garg N., Peng Y., Nguyen D. D., Watrous J., Kapono C. A., Luzzatto-Knaan T., Porto C., Bouslimani A., Melnik A. V., Meehan M. J., Liu W.-T., Crüsemann M., Boudreau P. D., Esquenazi E., Sandoval-Calderón M., Kersten R. D., Pace L. A., Quinn R. A., Duncan K. R., Hsu C.-C., Floros D. J., Gavilan R. G., Kleigrewe K., Northen T., Dutton R. J., Parrot D., Carlson E. E., Aigle B., Michelsen C. F., Jelsbak L., Sohlenkamp C., Pevzner P., Edlund A., McLean J., Piel J., Murphy B. T., Gerwick L., Liaw C.-C., Yang Y.-L., Humpf H.-U., Maansson M., Keyzers R. A., Sims A. C., Johnson A. R., Sidebottom A. M., Sedio B. E., Klitgaard A., Larson C. B., Boya P C. A., Torres-Mendoza D., Gonzalez D. J., Silva D. B., Marques L. M., Demarque D. P., Pociute E., O'Neill E. C., Briand E., Helfrich E. J. N., Granatosky E. A., Glukhov E., Ryffel F., Houson H., Mohimani H., Kharbush J. J., Zeng Y., Vorholt J. A., Kurita K. L., Charusanti P., McPhail K. L., Nielsen K. F., Vuong L., Elfeki M., Traxler M. F., Engene N., Koyama N., Vining O. B., Baric R., Silva R. R., Mascuch S. J., Tomasi S., Jenkins S., Macherla V., Hoffman T., Agarwal V., Williams P. G., Dai J., Neupane R., Gurr J., Rodríguez A. M. C., Lamsa A., Zhang C., Dorrestein K., Duggan B. M., Almaliti J., Allard P.-M., Phapale P., Nothias L.-F., Alexandrov T., Litaudon M., Wolfender J.-L., Kyle J. E., Metz T. O., Peryea T., Nguyen D.-T., VanLeer D., Shinn P., Jadhav A., Müller R., Waters K. M., Shi W., Liu X., Zhang L., Knight R., Jensen P. R., Palsson B. Ø., Pogliano K., Lington R. G., Gutiérrez M., Lopes N. P., Gerwick W. H., Moore B. S., Dorrestein P. C., Bandeira N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking // *Nature Biotechnology*. – 2016. – T. 34, № 8. – С. 828-837.
13. Mass Spectrometry Data Center. – URL: <https://chemdata.nist.gov/> (дата обращения: 28 декабря 2022 г.
14. Kopka J., Schauer N., Krueger S., Birkemeyer C., Usadel B., Bergmüller E., Dörmann P., Weckwerth W., Gibon Y., Stitt M., Willmitzer L., Fernie A. R., Steinhauser D. GMD@CSB.DB: the Golm Metabolome Database // *Bioinformatics*. – 2005. – T. 21, № 8. – С. 1635-1638.
15. Hummel J., Strehmel N., Bölling C., Schmidt S., Walther D., Kopka J. Mass Spectral Search and Analysis Using the Golm Metabolome Database // *The Handbook of Plant Metabolomics*, 2013. – С. 321-343.
16. Identification of essential oil components by gas chromatography/mass spectrometry. / Adams R. P.: Allured publishing corporation Carol Stream, 2007.
17. Qualitative analysis of flavor and fragrance volatiles by glass capillary gas chromatography. / Jennings W.: Elsevier, 2012.
18. Stanstrup J., Neumann S., Vrhovsek U. PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems // *Analytical Chemistry*. – 2015. – T. 87, № 18. – С. 9421-9428.

19. Low D. Y., Micheau P., Koistinen V. M., Hanhineva K., Abrankó L., Rodriguez-Mateos A., da Silva A. B., van Poucke C., Almeida C., Andres-Lacueva C., Rai D. K., Capanoglu E., Tomás Barberán F. A., Mattivi F., Schmidt G., Gürdeniz G., Valentová K., Bresciani L., Petrásková L., Dragsted L. O., Philo M., Ulaszewska M., Mena P., González-Domínguez R., Garcia-Villalba R., Kamiloglu S., de Pascual-Teresa S., Durand S., Wiczkowski W., Bronze M. R., Stanstrup J., Manach C. Data sharing in PredRet for accurate prediction of retention time: Application to plant food bioactive compounds // *Food Chemistry*. – 2021. – Т. 357. – С. 129757.
20. KnowItAll Software & Spectral Libraries. – URL: <https://sciencesolutions.wiley.com> (дата обращения: 28 декабря 2022 г.)
21. Kim S., Chen J., Cheng T., Gindulyte A., He J., He S., Li Q., Shoemaker B. A., Thiessen P. A., Yu B., Zaslavsky L., Zhang J., Bolton E. E. PubChem in 2021: new data content and improved web interfaces // *Nucleic Acids Research*. – 2021. – Т. 49, № D1. – С. D1388-D1395.
22. Pence H. E., Williams A. ChemSpider: An Online Chemical Information Resource // *Journal of Chemical Education*. – 2010. – Т. 87, № 11. – С. 1123-1124.
23. Kim S., Chen J., Cheng T. J., Gindulyte A., He J., He S. Q., Li Q. L., Shoemaker B. A., Thiessen P. A., Yu B., Zaslavsky L., Zhang J., Bolton E. E. PubChem 2019 update: improved access to chemical data // *Nucleic Acids Research*. – 2019. – Т. 47, № D1. – С. D1102-D1109.
24. Wishart D. S., Feunang Y. D., Marcu A., Guo A. C., Liang K., Vazquez-Fresno R., Sajed T., Johnson D., Li C. R., Karu N., Sayeeda Z., Lo E., Assempour N., Berjanskii M., Singhal S., Arndt D., Liang Y. J., Badran H., Grant J., Serra-Cayuela A., Liu Y. F., Mandal R., Neveu V., Pon A., Knox C., Wilson M., Manach C., Scalbert A. HMDB 4.0: the human metabolome database for 2018 // *Nucleic Acids Research*. – 2018. – Т. 46, № D1. – С. D608-D617.
25. Wishart D. S., Guo A., Oler E., Wang F., Anjum A., Peters H., Dizon R., Sayeeda Z., Tian S., Lee Brian L., Berjanskii M., Mah R., Yamamoto M., Jovel J., Torres-Calzada C., Hiebert-Giesbrecht M., Lui Vicki W., Varshavi D., Varshavi D., Allen D., Arndt D., Khetarpal N., Sivakumaran A., Harford K., Sanford S., Yee K., Cao X., Budinski Z., Liigand J., Zhang L., Zheng J., Mandal R., Karu N., Dambrova M., Schiöth Helgi B., Greiner R., Gautam V. HMDB 5.0: the Human Metabolome Database for 2022 // *Nucleic Acids Research*. – 2022. – Т. 50, № D1. – С. D622-D631.
26. Kind T., Wohlgemuth G., Lee D. Y., Lu Y., Palazoglu M., Shahbaz S., Fiehn O. FiehnLib: Mass Spectral and Retention Index Libraries for Metabolomics Based on Quadrupole and Time-of-Flight Gas Chromatography/Mass Spectrometry // *Analytical Chemistry*. – 2009. – Т. 81, № 24. – С. 10038-10048.
27. Horai H., Arita M., Kanaya S., Nihei Y., Ikeda T., Suwa K., Ojima Y., Tanaka K., Tanaka S., Aoshima K., Oda Y., Kakazu Y., Kusano M., Tohge T., Matsuda F., Sawada Y., Hirai M. Y., Nakanishi H., Ikeda K., Akimoto N., Maoka T., Takahashi H., Ara T., Sakurai N., Suzuki H., Shibata D., Neumann S., Iida T., Tanaka K., Funatsu K., Matsuura F., Soga T., Taguchi R., Saito K., Nishioka T. MassBank:

- a public repository for sharing mass spectral data for life sciences // *Journal of Mass Spectrometry*. – 2010. – Т. 45, № 7. – С. 703-714.
28. METLIN Gen2. – URL: <https://massconsortium.com> (дата обращения: 28 декабря 2022 г.)
29. Domingo-Almenara X., Guijas C., Billings E., Montenegro-Burke J. R., Uritboonthai W., Aisporna A. E., Chen E., Benton H. P., Siuzdak G. The METLIN small molecule dataset for machine learning-based retention time prediction // *Nature Communications*. – 2019. – Т. 10.
30. Wishart D. S., Knox C., Guo A. C., Shrivastava S., Hassanali M., Stothard P., Chang Z., Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration // *Nucleic Acids Research*. – 2006. – Т. 34. – С. D668-D672.
31. Brereton R. G., Lloyd G. R. Support Vector Machines for classification and regression // *Analyst*. – 2010. – Т. 135, № 2. – С. 230-267.
32. Decision trees for decision making. / Magee J. F.: Harvard Business Review Brighton, MA, USA, 1964.
33. Geurts P., Irrthum A., Wehenkel L. Supervised learning with decision tree-based methods in computational and systems biology // *Molecular BioSystems*. – 2009. – Т. 5, № 12. – С. 1593-1605.
34. Myles A. J., Feudale R. N., Liu Y., Woody N. A., Brown S. D. An introduction to decision tree modeling // *Journal of Chemometrics*. – 2004. – Т. 18, № 6. – С. 275-285.
35. Hammann F., Drewe J. Decision tree models for data mining in hit discovery // *Expert Opinion on Drug Discovery*. – 2012. – Т. 7, № 4. – С. 341-352.
36. Breiman L. Random Forests // *Machine Learning*. – 2001. – Т. 45, № 1. – С. 5-32.
37. Svetnik V., Liaw A., Tong C., Culberson J. C., Sheridan R. P., Feuston B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling // *Journal of Chemical Information and Computer Sciences*. – 2003. – Т. 43, № 6. – С. 1947-1958.
38. Svetnik V., Liaw A., Tong C., Wang T. Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules // *Multiple Classifier Systems / Под ред. Roli F. и др.* – Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. – С. 334-343.
39. Palmer D. S., O'Boyle N. M., Glen R. C., Mitchell J. B. O. Random Forest Models To Predict Aqueous Solubility // *Journal of Chemical Information and Modeling*. – 2007. – Т. 47, № 1. – С. 150-158.
40. Zhang Q.-Y., Aires-de-Sousa J. Random Forest Prediction of Mutagenicity from Empirical Physicochemical Descriptors // *Journal of Chemical Information and Modeling*. – 2007. – Т. 47, № 1. – С. 1-8.
41. Schapire R. E. A brief introduction to boosting. – Т. 99 –Citeseer. – С. 1401-1406.
42. Freund Y., Schapire R., Abe N. A short introduction to boosting // *Journal-Japanese Society For Artificial Intelligence*. – 1999. – Т. 14, № 771-780. – С. 1612.

43. He P., Xu C.-J., Liang Y.-Z., Fang K.-T. Improving the classification accuracy in chemistry via boosting technique // *Chemometrics and Intelligent Laboratory Systems*. – 2004. – Т. 70, № 1. – С. 39-46.
44. Svetnik V., Wang T., Tong C., Liaw A., Sheridan R. P., Song Q. Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling // *Journal of Chemical Information and Modeling*. – 2005. – Т. 45, № 3. – С. 786-799.
45. Friedman J. H. Stochastic gradient boosting // *Computational Statistics & Data Analysis*. – 2002. – Т. 38, № 4. – С. 367-378.
46. Chen T. Q., Guestrin C., Assoc Comp M. XGBoost: A Scalable Tree Boosting System // *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*. – 2016.10.1145/2939672.2939785. – С. 785-794.
47. CatBoost. – URL: <https://catboost.ai> (дата обращения: 29 декабря 2022).
48. Dorogush A. V., Ershov V., Gulin A. CatBoost: gradient boosting with categorical features support // *arXiv preprint arXiv:1810.11363*. – 2018.
49. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y. Lightgbm: A highly efficient gradient boosting decision tree // *Advances in neural information processing systems*. – 2017. – Т. 30.
50. Zhang J. h., Liu Z. m., Liu W. r. QSPR study for prediction of boiling points of 2475 organic compounds using stochastic gradient boosting // *Journal of Chemometrics*. – 2014. – Т. 28, № 3. – С. 161-167.
51. Sheridan R. P., Wang W. M., Liaw A., Ma J., Gifford E. M. Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships // *Journal of Chemical Information and Modeling*. – 2016. – Т. 56, № 12. – С. 2353-2360.
52. McCulloch W. S., Pitts W. A logical calculus of the ideas immanent in nervous activity // *The bulletin of mathematical biophysics*. – 1943. – Т. 5, № 4. – С. 115-133.
53. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain // *Psychological review*. – 1958. – Т. 65, № 6. – С. 386.
54. Rumelhart D. E., Hinton G. E., Williams R. J. Learning representations by back-propagating errors // *Nature*. – 1986. – Т. 323, № 6088. – С. 533-536.
55. Delashmit W. H., Manry M. T. Recent developments in multilayer perceptron neural networks – .
56. Tang W., Chen J., Wang Z., Xie H., Hong H. Deep learning for predicting toxicity of chemicals: a mini review // *Journal of Environmental Science and Health, Part C*. – 2018. – Т. 36, № 4. – С. 252-271.
57. Goh G. B., Hodas N. O., Vishnu A. Deep learning for computational chemistry // *Journal of Computational Chemistry*. – 2017. – Т. 38, № 16. – С. 1291-1307.

58. LeCun Y., Bottou L., Bengio Y., Haffner P. Gradient-based learning applied to document recognition // Proceedings of the IEEE. – 1998. – Т. 86, № 11. – С. 2278-2324.
59. Zhang Q., Zhang M., Chen T., Sun Z., Ma Y., Yu B. Recent advances in convolutional neural network acceleration // Neurocomputing. – 2019. – Т. 323. – С. 37-51.
60. Hochreiter S., Schmidhuber J. Long short-term memory // Neural Computation. – 1997. – Т. 9, № 8. – С. 1735-1780.
61. Cho K., van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation // arXiv e-prints. – 2014. – С. arXiv:1406.1078.
62. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention is All you Need // NIPS –, 2017. –.
63. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // . – 2018. – URL: <https://ui.adsabs.harvard.edu/abs/2018arXiv181004805D> (дата обращения: 14 ноября 2023).
64. Payne J., Srouji M., Ang Yap D., Kosaraju V. BERT Learns (and Teaches) Chemistry // . – 2020 (дата обращения: 1 июля 2020).
65. Jablonka K. M., Schwaller P., Smit B. Is GPT-3 all you need for machine learning for chemistry? – . –.
66. Irwin R., Dimitriadis S., He J., Bjerrum E. J. Chemformer: a pre-trained transformer for computational chemistry // Machine Learning: Science and Technology. – 2022. – Т. 3, № 1. – С. 015022.
67. Karpov P., Godin G., Tetko I. V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation // Journal of Cheminformatics. – 2020. – Т. 12, № 1.
68. Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., Citro C., Corrado G. S., Davis A., Dean J., Devin M. Tensorflow: Large-scale machine learning on heterogeneous distributed systems // arXiv preprint arXiv:1603.04467. – 2016.
69. Chollet F. c., et al. Keras // Book Keras / Editor, 2015.
70. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. – 2011. – Т. 12. – С. 2825-2830.
71. Ramsundar B. Molecular machine learning with DeepChem // Abstracts of Papers of the American Chemical Society. – 2018. – Т. 255. – С. 1.
72. Van Dyk D. A., Meng X.-L. The art of data augmentation // Journal of Computational and Graphical Statistics. – 2001. – Т. 10, № 1. – С. 1-50.

73. Zhang Y., Wang L., Wang X., Zhang C., Ge J., Tang J., Su A., Duan H. Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes // *Organic Chemistry Frontiers*. – 2021. – T. 8, № 7. – C. 1415-1423.
74. Pan S. J., Yang Q. A. A Survey on Transfer Learning // *Ieee Transactions on Knowledge and Data Engineering*. – 2010. – T. 22, № 10. – C. 1345-1359.
75. Liebal U. W., Phan A. N. T., Sudhakar M., Raman K., Blank L. M. Machine Learning Applications for Mass Spectrometry-Based Metabolomics // *Metabolites*. – 2020. – T. 10, № 6.
76. Puthongkham P., Wirojsaengthong S., Suea-Engam A. Machine learning and chemometrics for electrochemical sensors: moving forward to the future of analytical chemistry // *Analyst*. – 2021. – T. 146, № 21. – C. 6351-6364.
77. Cui F., Yue Y., Zhang Y., Zhang Z., Zhou H. S. Advancing biosensors with machine learning // *ACS sensors*. – 2020. – T. 5, № 11. – C. 3346-3364.
78. Debus B., Parastar H., Harrington P., Kirsanov D. Deep learning in analytical chemistry // *TrAC Trends in Analytical Chemistry*. – 2021. – T. 145. – C. 116459.
79. Goloborodko A. A., Levitsky L. I., Ivanov M. V., Gorshkov M. V. Pyteomics-a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics // *Journal of the American Society for Mass Spectrometry*. – 2013. – T. 24, № 2. – C. 301-304.
80. Levitsky L. I., Klein J. A., Ivanov M. V., Gorshkov M. V. Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework // *Journal of Proteome Research*. – 2019. – T. 18, № 2. – C. 709-714.
81. Ma C. W., Ren Y., Yang J. R., Ren Z., Yang H. M., Liu S. Q. Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning // *Analytical Chemistry*. – 2018. – T. 90, № 18. – C. 10881-10888.
82. Meyer J. G. Deep learning neural network tools for proteomics // *Cell Reports Methods*. – 2021. – T. 1, № 2. – C. 100003.
83. Moruz L., Staes A., Foster J. M., Hatzou M., Timmerman E., Martens L., Kall L. Chromatographic retention time prediction for posttranslationally modified peptides // *Proteomics*. – 2012. – T. 12, № 8. – C. 1151-1159.
84. Moruz L., Kall L. PEPTIDE RETENTION TIME PREDICTION // *Mass Spectrometry Reviews*. – 2017. – T. 36, № 5. – C. 615-623.
85. Wen B., Zeng W. F., Liao Y., Shi Z., Savage S. R., Jiang W., Zhang B. Deep learning in proteomics // *Proteomics*. – 2020. – T. 20, № 21-22. – C. 1900335.
86. Wu Z., Ramsundar B., Feinberg E. N., Gomes J., Geniesse C., Pappu A. S., Leswing K., Pande V. MoleculeNet: A Benchmark for Molecular Machine Learning // *arXiv e-prints*. – 2017. – C. arXiv:1703.00564.

87. Gozalbes R., Doucet J. P., Derouin F. Application of topological descriptors in QSAR and drug design: history and new trends // *Current Drug Targets-Infectious Disorders*. – 2002. – Т. 2, № 1. – С. 93-102.
88. Klein D. J. *Topological Indices and Related Descriptors in QSAR and QSPR* Edited by James Devillers & Alexandru T. Balaban. Gordon and Breach Science Publishers: Singapore. 1999. 811 pp. 90-5699-239-2. \$198.00 // *Journal of Chemical Information and Computer Sciences*. – 2002. – Т. 42, № 6. – С. 1507-1507.
89. Wiener H. Structural determination of paraffin boiling points // *Journal of the American chemical society*. – 1947. – Т. 69, № 1. – С. 17-20.
90. Randic M. Characterization of molecular branching // *Journal of the American Chemical Society*. – 1975. – Т. 97, № 23. – С. 6609-6615.
91. Karelson M., Lobanov V. S., Katritzky A. R. Quantum-chemical descriptors in QSAR/QSPR studies // *Chemical reviews*. – 1996. – Т. 96, № 3. – С. 1027-1044.
92. Wang L., Ding J., Pan L., Cao D., Jiang H., Ding X. Quantum chemical descriptors in quantitative structure–activity relationship models and their applications // *Chemometrics and Intelligent Laboratory Systems*. – 2021. – Т. 217. – С. 104384.
93. Rogers D., Hahn M. Extended-Connectivity Fingerprints // *Journal of Chemical Information and Modeling*. – 2010. – Т. 50, № 5. – С. 742-754.
94. ACD Labs. – URL: Chemistry Software (acdlabs.com) (дата обращения: 30 декабря 2022 г.)
95. RDKit: Open-source cheminformatics. – URL: <http://www.rdkit.org> (дата обращения: 30 декабря 2022 г.)
96. Willighagen E. L., Mayfield J. W., Alvarsson J., Berg A., Carlsson L., Jeliaskova N., Kuhn S., Pluskal T., Rojas-Chertó M., Spjuth O., Torrance G., Evelo C. T., Guha R., Steinbeck C. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching // *Journal of Cheminformatics*. – 2017. – Т. 9, № 1. – С. 33.
97. Moriwaki H., Tian Y. S., Kawashita N., Takagi T. Mordred: a molecular descriptor calculator // *Journal of Cheminformatics*. – 2018. – Т. 10.
98. Yap C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints // *Journal of Computational Chemistry*. – 2011. – Т. 32, № 7. – С. 1466-1474.
99. Jaeger S., Fulle S., Turk S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition // *Journal of Chemical Information and Modeling*. – 2018. – Т. 58, № 1. – С. 27-35.
100. Hodas N., Siegel C., Vishnu A., Goh G. SMILES2vec: An interpretable general-purpose deep neural network for predicting chemical properties // *Abstracts of Papers of the American Chemical Society*. – 2018. – Т. 256. – С. 1.

101. Deng D., Chen X., Zhang R., Lei Z., Wang X., Zhou F. XGraphBoost: Extracting Graph Neural Network-Based Features for a Better Prediction of Molecular Properties // *Journal of Chemical Information and Modeling*. – 2021. – T. 61, № 6. – C. 2697-2705.
102. Goh G. B., Siegel C., Vishnu A., Hodas N. O., Baker N. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models // *arXiv preprint arXiv:1706.06689*. – 2017.
103. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules // *Journal of Chemical Information and Computer Sciences*. – 1988. – T. 28, № 1. – C. 31-36.
104. Weininger D., Weininger A., Weininger J. L. SMILES. 2. Algorithm for generation of unique SMILES notation // *Journal of Chemical Information and Computer Sciences*. – 1989. – T. 29, № 2. – C. 97-101.
105. Heller S., McNaught A., Stein S., Tchekhovskoi D., Pletnev I. InChI - the worldwide chemical structure identifier standard // *Journal of cheminformatics*. – 2013. – T. 5, № 1. – C. 7-7.
106. Howard J., Ruder S. Universal Language Model Fine-tuning for Text Classification // . – 2018. – URL: <https://ui.adsabs.harvard.edu/abs/2018arXiv180106146H> (дата обращения: January 01, 2018).
107. Li X. H., Fourches D. Inductive transfer learning for molecular activity prediction: Next-Gen QSAR Models with MolPMoFiT // *Journal of Cheminformatics*. – 2020. – T. 12, № 1.
108. Withnall M., Lindelöf E., Engkvist O., Chen H. Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction // *Journal of Cheminformatics*. – 2020. – T. 12, № 1. – C. 1.
109. Wu Z., Pan S., Chen F., Long G., Zhang C., Philip S. Y. A comprehensive survey on graph neural networks // *IEEE transactions on neural networks and learning systems*. – 2020. – T. 32, № 1. – C. 4-24.
110. Xiong J., Xiong Z., Chen K., Jiang H., Zheng M. Graph neural networks for automated de novo drug design // *Drug Discovery Today*. – 2021. – T. 26, № 6. – C. 1382-1393.
111. Wieder O., Kohlbacher S., Kuenemann M., Garon A., Ducrot P., Seidel T., Langer T. A compact review of molecular property prediction with graph neural networks // *Drug Discovery Today: Technologies*. – 2020. – T. 37. – C. 1-12.
112. Reiser P., Neubert M., Eberhard A., Torresi L., Zhou C., Shao C., Metni H., van Hoesel C., Schopmans H., Sommer T. Graph neural networks for materials science and chemistry // *Communications Materials*. – 2022. – T. 3, № 1. – C. 1-18.
113. Jiang D., Wu Z., Hsieh C.-Y., Chen G., Liao B., Wang Z., Shen C., Cao D., Wu J., Hou T. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models // *Journal of Cheminformatics*. – 2021. – T. 13, № 1. – C. 12.

114. Kováts E. Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone // *Helvetica Chimica Acta*. – 1958. – Т. 41, № 7. – С. 1915-1932.
115. van Den Dool H., Dec. Kratz P. A generalization of the retention index system including linear temperature programmed gas—liquid partition chromatography // *Journal of Chromatography A*. – 1963. – Т. 11. – С. 463-471.
116. Castello G. Retention index systems: alternatives to the n-alkanes as calibration standards // *Journal of Chromatography A*. – 1999. – Т. 842, № 1. – С. 51-64.
117. Lee M. L., Vassilaros D. L., White C. M. Retention indices for programmed-temperature capillary-column gas chromatography of polycyclic aromatic hydrocarbons // *Analytical Chemistry*. – 1979. – Т. 51, № 6. – С. 768-773.
118. Kaliszan R. Quantitative structure-retention relationships // *Analytical Chemistry*. – 1992. – Т. 64, № 11. – С. 619A-631A.
119. Жохов А., Лоскутов А., Рыбальченко И. МЕТОДИЧЕСКИЕ ПОДХОДЫ К ВЫЧИСЛЕНИЮ И ПРОГНОЗИРОВАНИЮ ИНДЕКСОВ УДЕРЖИВАНИЯ В КАПИЛЛЯРНОЙ ГАЗОВОЙ ХРОМАТОГРАФИИ // *Журнал аналитической химии*. – 2018. – Т. 73, № 3. – С. 163-180.
120. Héberger K. Quantitative structure–(chromatographic) retention relationships // *Journal of Chromatography A*. – 2007. – Т. 1158, № 1. – С. 273-305.
121. Payares P., Díaz D., Olivero J., Vivas R., Gómez I. Prediction of the gas chromatographic relative retention times of flavonoids from molecular structure // *Journal of Chromatography A*. – 1997. – Т. 771, № 1. – С. 213-219.
122. Hemmateenejad B., Javadnia K., Elyasi M. Quantitative structure–retention relationship for the Kovats retention indices of a large set of terpenes: A combined data splitting-feature selection strategy // *Analytica Chimica Acta*. – 2007. – Т. 592, № 1. – С. 72-81.
123. Mihaleva V. V., Verhoeven H. A., de Vos R. C. H., Hall R. D., van Ham R. C. H. J. Automated procedure for candidate compound selection in GC-MS metabolomics based on prediction of Kovats retention index // *Bioinformatics*. – 2009. – Т. 25, № 6. – С. 787-794.
124. Zhokhov A. K., Loskutov A. Y., Rybal'chenko I. V. Methodological Approaches to the Calculation and Prediction of Retention Indices in Capillary Gas Chromatography // *Journal of Analytical Chemistry*. – 2018. – Т. 73, № 3. – С. 207-220.
125. Zenkevich I. G., Makarov A. A., Schrader S., Moeder M. A new version of an additive scheme for the prediction of gas chromatographic retention indices of the 211 structural isomers of 4-nonylphenol // *Journal of Chromatography A*. – 2009. – Т. 1216, № 18. – С. 4097-4106.

126. Farkas O., Héberger K., Zenkevich I. G. Quantitative structure–retention relationships XIV: Prediction of gas chromatographic retention indices for saturated O-, N-, and S-heterocyclic compounds // *Chemometrics and Intelligent Laboratory Systems*. – 2004. – Т. 72, № 2. – С. 173-184.
127. Babushok V. I., Linstrom P. J., Reed J. J., Zenkevich I. G., Brown R. L., Mallard W. G., Stein S. E. Development of a database of gas chromatographic retention properties of organic compounds // *Journal of Chromatography A*. – 2007. – Т. 1157, № 1. – С. 414-421.
128. Babushok V. I., Linstrom P. J., Zenkevich I. G. Retention indices for frequently reported compounds of plant essential oils // *Journal of Physical and Chemical Reference Data*. – 2011. – Т. 40, № 4.
129. Stein S. E., Babushok V. I., Brown R. L., Linstrom P. J. Estimation of Kováts Retention Indices Using Group Contributions // *Journal of Chemical Information and Modeling*. – 2007. – Т. 47, № 3. – С. 975-980.
130. Matyushin D. D., Sholokhova A. Y., Buryak A. K. Gradient boosting for the prediction of gas chromatographic retention indices // *Сорбционные и хроматографические процессы*. – 2019. – Т. 19, № 6. – С. 630-635.
131. Matyushin D. D., Sholokhova A. Y., Buryak A. K. A deep convolutional neural network for the estimation of gas chromatographic retention indices // *Journal of Chromatography A*. – 2019. – Т. 1607. – С. 460395.
132. Vrzal T., Malečková M., Olšovská J. DeepReI: Deep learning-based gas chromatographic retention index predictor // *Analytica Chimica Acta*. – 2021. – Т. 1147. – С. 64-71.
133. Matyushin D. D., Sholokhova A. Y., Buryak A. K. A deep convolutional neural network for the estimation of gas chromatographic retention indices // *Journal of Chromatography A*. – 2019. – Т. 1607.
134. Qu C., Schneider B. I., Kearsley A. J., Keyrouz W., Allison T. C. Predicting Kováts Retention Indices Using Graph Neural Networks // *Journal of Chromatography A*. – 2021. – Т. 1646. – С. 462100.
135. Matyushin D. D., Buryak A. K. Gas Chromatographic Retention Index Prediction Using Multimodal Machine Learning // *Ieee Access*. – 2020. – Т. 8. – С. 223140-223155.
136. 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 // *International Journal of Molecular Sciences*. – 2021.
137. Samaraweera M. A., Hall L. M., Hill D. W., Grant D. F. Evaluation of an Artificial Neural Network Retention Index Model for Chemical Structure Identification in Nontargeted Metabolomics // *Analytical Chemistry*. – 2018. – Т. 90, № 21. – С. 12752-12760.
138. Bouwmeester R., Martens L., Degroeve S. Comprehensive and Empirical Evaluation of Machine Learning Algorithms for Small Molecule LC Retention Time Prediction // *Analytical Chemistry*. – 2019. – Т. 91, № 5. – С. 3694-3703.

139. Bruderer T., Varesio E., Hopfgartner G. The use of LC predicted retention times to extend metabolites identification with SWATH data acquisition // *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences*. – 2017. – T. 1071. – C. 3-10.
140. Wolfer A. M., Lozano S., Umbdenstock T., Croixmarie V., Arrault A., Vayer P. UPLC-MS retention time prediction: a machine learning approach to metabolite identification in untargeted profiling // *Metabolomics*. – 2016. – T. 12, № 1.
141. Eugster P. J., Boccard J., Debrus B., Bréant L., Wolfender J.-L., Martel S., Carrupt P.-A. Retention time prediction for dereplication of natural products (C_xH_yO_z) in LC-MS metabolite profiling // *Phytochemistry*. – 2014. – T. 108. – C. 196-207.
142. Aicheler F., Li J., Hoene M., Lehmann R., Xu G. W., Kohlbacher O. Retention Time Prediction Improves Identification in Nontargeted Lipidomics Approaches // *Analytical Chemistry*. – 2015. – T. 87, № 15. – C. 7698-7704.
143. Falchi F., Bertozzi S. M., Ottonello G., Ruda G. F., Colombano G., Fiorelli C., Martucci C., Bertorelli R., Scarpelli R., Cavalli A., Bandiera T., Armirotti A. Kernel-Based, Partial Least Squares Quantitative Structure-Retention Relationship Model for UPLC Retention Time Prediction: A Useful Tool for Metabolite Identification // *Analytical Chemistry*. – 2016. – T. 88, № 19. – C. 9510-9517.
144. Broeckling C. D., Ganna A., Layer M., Brown K., Sutton B., Ingelsson E., Peers G., Prenni J. E. Enabling Efficient and Confident Annotation of LC-MS Metabolomics Data through MS1 Spectrum and Time Prediction // *Analytical Chemistry*. – 2016. – T. 88, № 18. – C. 9226-9234.
145. Aalizadeh R., Nika M. C., Thomaidis N. S. Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants // *Journal of Hazardous Materials*. – 2019. – T. 363. – C. 277-285.
146. Randazzo G. M., Tonoli D., Hambye S., Guillarme D., Jeanneret F., Nurisso A., Goracci L., Boccard J., Rudaz S. Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification // *Analytica Chimica Acta*. – 2016. – T. 916. – C. 8-16.
147. Bade R., Bijlsma L., Miller T. H., Barron L. P., Sancho J. V., Hernandez F. Suspect screening of large numbers of emerging contaminants in environmental waters using artificial neural networks for chromatographic retention time prediction and high resolution mass spectrometry data analysis // *Science of the Total Environment*. – 2015. – T. 538. – C. 934-941.
148. Bonini P., Kind T., Tsugawa H., Barupal D. K., Fiehn O. Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics // *Analytical Chemistry*. – 2020. – T. 92, № 11. – C. 7515-7522.
149. Abate-Pella D., Freund D. M., Ma Y., Simon-Manso Y., Hollender J., Broeckling C. D., Huhman D. V., Krokhin O. V., Stoll D. R., Hegeman A. D., Kind T., Fiehn O., Schymanski E. L., Prenni J. E., Sumner L. W., Boswell P. G. Retention projection enables accurate calculation of liquid

- chromatographic retention times across labs and methods // *Journal of Chromatography A*. – 2015. – T. 1412. – C. 43-51.
150. Boswell P. G., Schellenberg J. R., Carr P. W., Cohen J. D., Hegeman A. D. A study on retention "projection" as a supplementary means for compound identification by liquid chromatography-mass spectrometry capable of predicting retention with different gradients, flow rates, and instruments // *Journal of Chromatography A*. – 2011. – T. 1218, № 38. – C. 6732-6741.
151. Boswell P. G., Schellenberg J. R., Carr P. W., Cohen J. D., Hegeman A. D. Easy and accurate high-performance liquid chromatography retention prediction with different gradients, flow rates, and instruments by back-calculation of gradient and flow rate profiles // *Journal of Chromatography A*. – 2011. – T. 1218, № 38. – C. 6742-6749.
152. Bach E., Szedmak S., Brouard C., Bocker S., Rousu J. Liquid-chromatography retention order prediction for metabolite identification // *Bioinformatics*. – 2018. – T. 34, № 17. – C. 875-883.
153. Liu J. J., Alipuly A., Baczek T., Wong M. W., Zuvela P. Quantitative Structure-Retention Relationships with Non-Linear Programming for Prediction of Chromatographic Elution Order // *International Journal of Molecular Sciences*. – 2019. – T. 20, № 14.
154. Wen Y. V., Amos R. I. J., Talebi M., Szucs R., Dolan J. W., Pohl C. A., Haddad P. R. Retention Index Prediction Using Quantitative Structure-Retention Relationships for Improving Structure Identification in Nontargeted Metabolomics // *Analytical Chemistry*. – 2018. – T. 90, № 15. – C. 9434-9440.
155. Theodoridis G., Gika H., Franceschi P., Caputi L., Arapitsas P., Scholz M., Masuero D., Wehrens R., Vrhovsek U., Mattivi F. LC-MS based global metabolite profiling of grapes: solvent extraction protocol optimisation // *Metabolomics*. – 2012. – T. 8, № 2. – C. 175-185.
156. Barri T., Holmer-Jensen J., Hermansen K., Dragsted L. O. Metabolic fingerprinting of high-fat plasma samples processed by centrifugation- and filtration-based protein precipitation delineates significant differences in metabolite information coverage // *Analytica Chimica Acta*. – 2012. – T. 718. – C. 47-57.
157. Fedorova E. S., Matyushin D. D., Plyushchenko I. V., Stavrianidi A. N., Buryak A. K. Deep learning for retention time prediction in reversed-phase liquid chromatography // *Journal of Chromatography A*. – 2022. – T. 1664. – C. 462792.
158. García C. A., Gil-de-la-Fuente A., Barbas C., Otero A. Probabilistic metabolite annotation using retention time prediction and meta-learned projections // *Journal of Cheminformatics*. – 2022. – T. 14, № 1. – C. 33.
159. Kensert A., Bouwmeester R., Efthymiadis K., Van Broeck P., Desmet G., Cabooter D. Graph Convolutional Networks for Improved Prediction and Interpretability of Chromatographic Retention Data // *Analytical Chemistry*. – 2021. – T. 93, № 47. – C. 15633-15641.

160. Ju R., Liu X., Zheng F., Lu X., Xu G., Lin X. Deep Neural Network Pretrained by Weighted Autoencoders and Transfer Learning for Retention Time Prediction of Small Molecules // *Analytical Chemistry*. – 2021. – T. 93, № 47. – C. 15651-15658.
161. Yang Q., Ji H., Lu H., Zhang Z. Prediction of Liquid Chromatographic Retention Time with Graph Neural Networks to Assist in Small Molecule Identification // *Analytical Chemistry*. – 2021. – T. 93, № 4. – C. 2200-2206.
162. Zaretskii M., Bashkirova I., Osipenko S., Kostyukevich Y., Nikolaev E., Popov P. 3D chemical structures allow robust deep learning models for retention time prediction // *Digital Discovery*. – 2022.10.1039/D2DD00021K.
163. Gorynski K., Bojko B., Nowaczyk A., Bucinski A., Pawliszyn J., Kaliszan R. Quantitative structure-retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: Endogenous metabolites and banned compounds // *Analytica Chimica Acta*. – 2013. – T. 797. – C. 13-19.
164. Creek D. J., Jankevics A., Breitling R., Watson D. G., Barrett M. P., Burgess K. E. V. Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography-Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction // *Analytical Chemistry*. – 2011. – T. 83, № 22. – C. 8703-8710.
165. Cao M., Fraser K., Huege J., Featonby T., Rasmussen S., Jones C. Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics // *Metabolomics*. – 2015. – T. 11, № 3. – C. 696-706.
166. da Silva R. R., Dorrestein P. C., Quinn R. A. Illuminating the dark matter in metabolomics // *Proceedings of the National Academy of Sciences*. – 2015. – T. 112, № 41. – C. 12549-12550.
167. Frainay C., Schymanski E. L., Neumann S., Merlet B., Salek R. M., Jourdan F., Yanes O. Mind the Gap: Mapping Mass Spectral Databases in Genome-Scale Metabolic Networks Reveals Poorly Covered Areas // *Metabolites*. – 2018.
168. Wei J. N., Belanger D., Adams R. P., Sculley D. Rapid Prediction of Electron-Ionization Mass Spectrometry Using Neural Networks // *ACS Central Science*. – 2019. – T. 5, № 4. – C. 700-708.
169. Zhang B., Zhang J., Xia Y., Chen P., Wang B. Prediction of electron ionization mass spectra based on graph convolutional networks // *International Journal of Mass Spectrometry*. – 2022. – T. 475. – C. 116817.
170. Grimme S. Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules // *Angewandte Chemie International Edition*. – 2013. – T. 52, № 24. – C. 6306-6312.
171. Koopman J., Grimme S. From QCEIMS to QCxMS: A Tool to Routinely Calculate CID Mass Spectra Using Molecular Dynamics // *Journal of the American Society for Mass Spectrometry*. – 2021. – T. 32, № 7. – C. 1735-1751.

172. Ruttkies C., Neumann S., Posch S. Improving MetFrag with statistical learning of fragment annotations // *Bmc Bioinformatics*. – 2019. – T. 20. – C. 14.
173. Zheng X. Y., Aly N. A., Zhou Y. X., Dupuis K. T., Bilbao A., Paurus V. L., Orton D. J., Wilson R., Payne S. H., Smith R. D., Baker E. S. A structural examination and collision cross section database for over 500 metabolites and xenobiotics using drift tube ion mobility spectrometry // *Chemical Science*. – 2017. – T. 8, № 11. – C. 7724-7736.
174. Zhou Z. W., Shen X. T., Tu J., Zhu Z. J. Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry // *Analytical Chemistry*. – 2016. – T. 88, № 22. – C. 11084-11091.
175. Zhou Z. W., Xiong X., Zhu Z. J. MetCCS predictor: a web server for predicting collision cross-section values of metabolites in ion mobility-mass spectrometry based metabolomics // *Bioinformatics*. – 2017. – T. 33, № 14. – C. 2235-2237.
176. Zhou Z. W., Tu J., Xiong X., Shen X. T., Zhu Z. J. LipidCCS: Prediction of Collision Cross-Section Values for Lipids with High Precision To Support Ion Mobility-Mass Spectrometry-Based Lipidomics // *Analytical Chemistry*. – 2017. – T. 89, № 17. – C. 9559-9566.
177. Bijlsma L., Bade R., Celma A., Mullin L., Cleland G., Stead S., Hernandez F., Sancho J. V. Prediction of Collision Cross-Section Values for Small Molecules: Application to Pesticide Residue Analysis // *Analytical Chemistry*. – 2017. – T. 89, № 12. – C. 6583-6589.
178. Mollerup C. B., Mardal M., Dalsgaard P. W., Linnet K., Barron L. P. Prediction of collision cross section and retention time for broad scope screening in gradient reversed-phase liquid chromatography-ion mobility-high resolution accurate mass spectrometry // *Journal of Chromatography A*. – 2018. – T. 1542. – C. 82-88.
179. Sosnin S., Karlov D., Tetko I. V., Fedorov M. V. Comparative study of multitask toxicity modeling on a broad chemical space // *Journal of chemical information and modeling*. – 2018. – T. 59, № 3. – C. 1062-1072.
180. Sosnin S., Vashurina M., Withnall M., Karpov P., Fedorov M., Tetko I. V. A survey of multi-task learning methods in chemoinformatics // *Molecular informatics*. – 2019. – T. 38, № 4. – C. 1800108.
181. Plante P. L., Francovic-Fontaine E., May J. C., McLean J. A., Baker E. S., Laviolette F., Marchand M., Corbeil J. Predicting Ion Mobility Collision Cross-Sections Using a Deep Neural Network: DeepCCS // *Analytical Chemistry*. – 2019. – T. 91, № 8. – C. 5191-5199.
182. Colby S. M., Nunez J. R., Hodas N. O., Corley C. D., Renslow R. R. Deep Learning to Generate in Silico Chemical Property Libraries and Candidate Molecules for Small Molecule Identification in Complex Samples // *Analytical Chemistry*. – 2020. – T. 92, № 2. – C. 1720-1729.
183. Colby S. M., Thomas D. G., Nuñez J. R., Baxter D. J., Glaesemann K. R., Brown J. M., Pirrung M. A., Govind N., Teegarden J. G., Metz T. O., Renslow R. S. ISiCLE: A Quantum Chemistry Pipeline

- for Establishing in Silico Collision Cross Section Libraries // *Analytical Chemistry*. – 2019. – Т. 91, № 7. – С. 4346-4356.
184. Bijlsma L., Berntssen M. H. G., Merel S. A Refined Nontarget Workflow for the Investigation of Metabolites through the Prioritization by in Silico Prediction Tools // *Analytical Chemistry*. – 2019. – Т. 91, № 9. – С. 6321-6328.
185. Muggeo V. M. R. Estimating regression models with unknown break-points // *Statistics in Medicine*. – 2003. – Т. 22, № 19. – С. 3055-3071.
186. SMARTS - A Language for Describing Molecular Patterns. – URL: <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>.
187. Gaulton A., Bellis L. J., Bento A. P., Chambers J., Davies M., Hersey A., Light Y., McGlinchey S., Michalovich D., Al-Lazikani B., Overington J. P. ChEMBL: a large-scale bioactivity database for drug discovery // *Nucleic Acids Research*. – 2012. – Т. 40, № D1. – С. D1100-D1107.
188. Howard J., Gugger S. Fastai: A Layered API for Deep Learning // *Information*. – 2020. – Т. 11, № 2.
189. Merity S., Shirish Keskar N., Socher R. Regularizing and Optimizing LSTM Language Models // – 2017. – URL: <https://ui.adsabs.harvard.edu/abs/2017arXiv170802182M> (дата обращения: August 01, 2017).
190. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting // *J. Mach. Learn. Res.* – 2014. – Т. 15, № 1. – С. 1929–1958.
191. Dai Z., Yang Z., Yang Y., Carbonell J., Le Q. V., Salakhutdinov R. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context // – 2019. – URL: <https://ui.adsabs.harvard.edu/abs/2019arXiv190102860D> (дата обращения: January 01, 2019).
192. Gilmer J., Schoenholz S. S., Riley P. F., Vinyals O., Dahl G. E. Neural Message Passing for Quantum Chemistry // *arXiv e-prints*. – 2017. – С. arXiv:1704.01212.
193. Kostyukevich Y., Acter T., Zhrebker A., Ahmed A., Kim S., Nikolaev E. Hydrogen/deuterium exchange in mass spectrometry // *Mass Spectrometry Reviews*. – 2018. – Т. 37, № 6. – С. 811-853.
194. Kostyukevich Y., Kononikhin A., Popov I., Nikolaev E. Simple Atmospheric Hydrogen/Deuterium Exchange Method for Enumeration of Labile Hydrogens by Electrospray Ionization Mass Spectrometry // *Analytical Chemistry*. – 2013. – Т. 85, № 11. – С. 5330-5334.
195. Бродский А. Химия изотопов, издание 2 // М., Изд-во АН СССР. – 1957.
196. Kostyukevich Y., Kononikhin A., Zhrebker A., Popov I., Perminova I., Nikolaev E. Enumeration of non-labile oxygen atoms in dissolved organic matter by use of O-16/O-18 exchange and Fourier transform ion-cyclotron resonance mass spectrometry // *Analytical and Bioanalytical Chemistry*. – 2014. – Т. 406, № 26. – С. 6655-6664.

197. Kostyukevich Y., Osipenko S., Rindin K., Zhrebker A., Kovaleva O., Rumiantseva L., Borisova L., Borisova N., Vlaskin M. S., Nikolaev E. Analysis of the Bio-oil Produced by the Hydrothermal Liquefaction of Biomass Using High-Resolution Mass Spectrometry and Isotope Exchange // *Energy & Fuels*. – 2021. – T. 35, № 15. – C. 12208-12215.
198. Zheng S.-J., Zheng J., Xiong C.-F., Xiao H.-M., Liu S.-J., Feng Y.-Q. Hydrogen–Deuterium Scrambling Based on Chemical Isotope Labeling Coupled with LC–MS: Application to Amine Metabolite Identification in Untargeted Metabolomics // *Analytical Chemistry*. – 2020. – T. 92, № 2. – C. 2043-2051.
199. Dührkop K., Fleischauer M., Ludwig M., Aksenov A. A., Melnik A. V., Meusel M., Dorrestein P. C., Rousu J., Böcker S. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information // *Nature Methods*. – 2019. – T. 16, № 4. – C. 299-302.
200. Neta P., Farahani M., Simón-Manso Y., Liang Y., Yang X., Stein S. E. Unexpected peaks in tandem mass spectra due to reaction of product ions with residual water in mass spectrometer collision cells // *Rapid Communications in Mass Spectrometry*. – 2014. – T. 28, № 23. – C. 2645-2660.
201. Nyanyira C. The OPCW Central Analytical Database // *Chemical Weapons Convention Chemicals Analysis*, 2005. – C. 133-149.
202. Erdey L., Takács J., Szalanczy E. Contribution to the theory of the retention index system: I. Retention indices using programmed-temperature gas chromatography // *Journal of Chromatography A*. – 1970. – T. 46. – C. 29-32.
203. van der Maaten L. Accelerating t-SNE using Tree-Based Algorithms // *Journal of Machine Learning Research*. – 2014. – T. 15. – C. 3221-3245.
204. Kelly K., Bell S. Evaluation of the reproducibility and repeatability of GCMS retention indices and mass spectra of novel psychoactive substances // *Forensic Chemistry*. – 2018. – T. 7. – C. 10-18.
205. Zellner B. d. A., Bicchi C., Dugo P., Rubiolo P., Dugo G., Mondello L. Linear retention indices in gas chromatographic analysis: a review // *Flavour and Fragrance Journal*. – 2008. – T. 23, № 5. – C. 297-314.
206. Kind T., Fiehn O. Advances in structure elucidation of small molecules using mass spectrometry // *Bioanalytical Reviews*. – 2010. – T. 2, № 1. – C. 23-60.
207. McLafferty F. W. Mass spectrometric analysis. Molecular rearrangements // *Analytical chemistry*. – 1959. – T. 31, № 1. – C. 82-87.
208. Brown P., Djerassi C. Electron-Impact Induced Rearrangement Reactions of Organic Molecules // *Angewandte Chemie International Edition in English*. – 1967. – T. 6, № 6. – C. 477-496.
209. Beynon J. H., Lester G. R., Williams A. E. Some specific molecular rearrangements in the mass spectra of organic compounds // *The Journal of Physical Chemistry*. – 1959. – T. 63, № 11. – C. 1861-1868.

210. Akiba T. a. S. S. a. Y. T. a. O. T. a. K. M. Optuna: A Next-Generation Hyperparameter Optimization Framework // Book Optuna: A Next-Generation Hyperparameter Optimization Framework / Editor Association for Computing Machinery, 2019. – C. 2623–2631 , numpages = 9.
211. Stein S. E., Scott D. R. Optimization and testing of mass spectral library search algorithms for compound identification // Journal of the American Society for Mass Spectrometry. – 1994. – T. 5, № 9. – C. 859-866.
212. Lee J., Kind T., Tantillo D. J., Wang L.-P., Fiehn O. Evaluating the Accuracy of the QCEIMS Approach for Computational Prediction of Electron Ionization Mass Spectra of Purines and Pyrimidines // Metabolites. – 2022.
213. Wang S., Kind T., Bremer P. L., Tantillo D. J., Fiehn O. Quantum Chemical Prediction of Electron Ionization Mass Spectra of Trimethylsilylated Metabolites // Analytical Chemistry. – 2022. – T. 94, № 3. – C. 1559-1566.
214. Spackman P. R., Bohman B., Karton A., Jayatilaka D. Quantum chemical electron impact mass spectrum prediction for de novo structure elucidation: Assessment against experimental reference data and comparison to competitive fragmentation modeling // International Journal of Quantum Chemistry. – 2018. – T. 118, № 2. – C. e25460.
215. Riches J. Chapter 7 - Analysis of Organophosphorus Chemicals // Best Synthetic Methods / Timperley C. M. – Oxford: Academic Press, 2015. – C. 721-752.

Приложение 1.

Таблица П1. Внутрिलाбораторный набор данных по удерживанию

№	Название	Идентификатор в PubChem	Время удерживания, с
1	2-(Метиламино)-1-фенилбутан-1-ол	46260	874
2	4-амино-3-фенилбутановая кислота	14113	95
3	N-дезметил-офлоксацин	11725233	458
4	Агомелатин	82148	846
5	Азоназол	43233	870
6	Азаметинос	71482	748
7	Азилсартан	135415867	920
8	Азиннос-метил	2268	972
9	Азиннос-этил	17531	1114
10	Азитромицин	447043	657
11	Азоксистеробин	3034285	1036
12	Акарифлор	13218777	1368
13	Аланикарб	5484171	1136
14	Альдикарб	5353395	628
15	Альфа-пирролидинвалерофенон	11148955	572
16	Амантадин	2130	425
17	Амбазон	1549158	773
18	Амидосульфурон	91777	811
19	Аминокарб	16247	41
20	Аминопирин	6009	170
21	Амитриптилин	2160	853
22	Амлодипин	2162	854
23	Амоксициллин	33613	95
24	Амфетамин	3007	236
25	Анастрозол	2187	801
26	Антипирин	2206	489
27	Апиксабан	10182969	785
28	Арбидол	131411	955
29	Атенолол	2249	115
30	Атропин	174174	563
31	Ацетаминофен	1983	102
32	Ацетамиприд	213021	578
33	Ацефат	1982	41
34	Ацикловир	135398513	50
35	Беналаксил	51369	1165
36	Бендрофлуметиазид	2315	857

37	Бензидамин	12555	807
38	Бензоилэкгонин	448223	479
39	Бензокаин	2337	688
40	Бенфуракарб	54886	1331
41	Бенциклан	2312	932
42	Бетаксоллол	2369	740
43	Бэфлубутамид	6451159	1180
44	Бикалутамид	2375	986
45	Биклотимол	71878	1566
46	Биластин	185460	745
47	Бипериден	2381	808
48	Бисакодил	2391	860
49	Бисопролол	2405	677
50	Битертанол	91656	1120
51	Боскалид	213013	1050
52	Бриналдикс	12492	629
53	Бромгексин	2442	791
54	Бромуконазол	3444	1064
55	Бромфенак	60726	1017
56	Буметанид	2471	960
57	Бупивакаин	2474	673
58	Бупиримат	38884	962
59	Бупрофезин	50367	1085
60	Буспирон	2477	678
61	Бутокарбаксим	5360962	628
62	Вамидотион	560193	539
63	Венлафаксин	5656	660
64	Верапамил	2520	879
65	Вилдаглиптин	6918537	179
66	Вориконазол	71616	856
67	Гексаконазол	66461	1098
68	Гемцитабин	60750	738
69	Гидрохлортиазид	3639	149
70	Гимекромон	5280567	582
71	Гистидин	6274	33
72	Глибурид	3488	1086
73	Гликлазид	3475	957
74	Дапсон	2955	439
75	Дезлоратадин	124087	614
76	Десмедифам	24743	972
77	Джозамицин	5282165	1154
78	Диазинон	3017	1199

79	Дидрогестерон	9051	1061
80	Диеногест	68861	818
81	Дилтиазем	39186	781
82	Диметахлор	39722	914
83	Диметиримол	135424353	552
84	Диметоат	3082	533
85	Диметоморф	5889665	997
86	Димоксистробин	9797414	1128
87	Диниконазол	6436605	1130
88	Дисульфирам	3117	1180
89	Диурон	3120	853
90	Дифенгидрамин	3100	743
91	Дифеноконазол	86173	1189
92	Дифлубензурон	37123	1081
93	Дифлуфеникан	91735	1268
94	Дихлорвос	3039	727
95	Диэтилтолуамид	4284	860
96	Доксазозин	3157	739
97	Доксиламин	3162	412
98	Доксициклин	54671203	650
99	Домперидон	3151	676
100	Донепезил	3152	1571
101	Дротаверин	1712095	905
102	Енилконазол	37175	824
103	Зофеноприл	92400	1120
104	Зуклопентиксол	5311507	912
105	Ибупрофен	3672	1097
106	Ивабрадин	132999	669
107	Идебенон	3686	1151
108	Изокарбафос	90479	841
109	Изоксабен	73672	1081
110	Изоксафлутол	84098	1010
111	Изопротиолан	39681	1081
112	Изофенфос-метил	127394	1214
113	Имидаклоприд	86418	527
114	Имипрамин	3696	832
115	Индапамид	3702	788
116	Индоксакарб	107720	1295
117	Инозин	135398641	50
118	Инозитол	5353356	275
119	Йодантипирин	8522	660
120	Ипидакрин	604519	433

121	Ипратропий	657309	514
122	Ипроваликарб	10958189	1032
123	Итоприд	3792	545
124	Итраконазол	55283	1198
125	Кандесартан цилексетил	2540	1317
126	Каптоприл	44093	919
127	Карбамазепин	2554	778
128	Карбендазим	25429	191
129	Карбидопа	34359	92
130	Карбоксин	21307	816
131	Карбофуран	2566	779
132	Карведилол	2585	801
133	Карфентразол-этил	86222	1163
134	Кафедрин	5489638	536
135	Кверцетин	5280343	541
136	Кветиапин	5002	716
137	Кларитромицин	84029	901
138	Клемастин	26987	977
139	Клетодим	135616187	1299
140	Клозапин	135398737	671
141	Кломипрамин	2801	915
142	Кломифен	2800	1082
143	Клопидогрел	60606	985
144	Клофентизин	73670	1216
145	Кодеин	5284371	200
146	Коргард	39147	454
147	Кофеин	2519	280
148	Крезоксим-метил	5483874	1169
149	Ксилометазолин	5709	799
150	Кумафос	2871	1216
151	Ламотриджин	3878	507
152	Ландрин	17592	879
153	Латанопрост	5311221	1127
154	Левитерацетам	5284583	164
155	Леводопа	6047	49
156	Левокабастин	54385	784
157	Левомепромазин	72287	853
158	Левотироксин	5819	872
159	Ленацил	16559	773
160	Лерканидипин	65866	1069
161	Летрозол	3902	787
162	Лефлуномид	3899	906

163	Линкомицин	3000540	385
164	Линурон	9502	974
165	Лозартан	3961	863
166	Лоперамид	3955	958
167	Лоратадин	3957	911
168	Лорноксикам	54690031	791
169	Луфенурон	71777	1348
170	Малаоксон	15415	797
171	Малатион	4004	1087
172	Мандипропамид	11292824	1066
173	Мебгидролин	22530	775
174	Мебеверин	4031	832
175	Мевинфос	5355863	615
176	Мезоридазин	4078	747
177	Мезосульфурон-метил	11409499	877
178	Мекарбам	17434	1138
179	Мелатонин	896	593
180	Мелоксикам	54677470	916
181	Мепанипирим	86296	1036
182	Метадон	4095	866
183	Метазахлор	49384	902
184	Метаквалон	6292	826
185	Металаксил	42586	875
186	Метамидофос	4096	41
187	Метамитрон	38854	475
188	Метамфетамин	10836	302
189	Метафлумизон	9827529	1358
190	Метидатион	13709	956
191	Метилендиоксипирролидин валерофенон	20111961	590
192	Метконазол	86210	1118
193	Метобромулон	18290	863
194	Метоклопрамид	4168	507
195	Метоксифенозид	105010	1088
196	Метолахлор	4169	1094
197	Метомил	5360521	268
198	Метопролол	4171	563
199	Метрафенон	6451057	1266
200	Метрибузин	30479	710
201	Метронидазол	4173	108
202	Метсульфурон-метил	52999	825
203	Мефлохин	4046	899

204	Миклобутанил	6336	1040
205	Миртазапин	4205	522
206	Моксифлоксацин	152946	613
207	Моксонидин	4810	112
208	Молинат	16653	997
209	Молсидомин	5353788	259
210	Монокротофос	5371562	386
211	Монтелукаст	5281040	1342
212	Морфин	5288826	65
213	Навелбин	5311497	868
214	Нафазолин	4436	509
215	Нафтидрофурил	4417	887
216	Нафтифин	47641	848
217	Небиволол	71301	859
218	Непафенак	151075	773
219	Никетамид	5497	286
220	Никобоксил	14866	820
221	Никосульфурон	73281	722
222	Нитразепам	4506	787
223	Нитрендипин	4507	1054
224	Нифедипин	4485	926
225	Нифурантоин	5353830	320
226	Нифуроксазид	5390108	610
227	Ницерголин	34040	867
228	Новалурон	93541	1299
229	Нонивамид	2998	1041
230	Нортриптиллин	4543	845
231	Оксадиазон	29732	1367
232	Оксадиксил	53735	727
233	Окрасульфурон	86443	748
234	Оксибупрокаин	4633	729
235	Оксибутинин	4634	901
236	Оксиметазолин	4636	722
237	Оланзапин	135398745	362
238	Олмесартан	158781	668
239	Олмесартан медоксомил	130881	844
240	Олопатодин	5281071	732
241	Омепразол	4594	612
242	Ометоат	14210	118
243	Орлистат	3034010	1612
244	Орнидазол	28061	410
245	Осельтамивир	65028	689

246	Паклобутразол	158076	971
247	Пантопразол	4679	674
248	Папаверин	4680	644
249	Пароксетин	43815	812
250	Педиметалин	38479	1364
251	Пенконазол	91693	1099
252	Пентоксифиллин	4740	575
253	Пенцикурон	91692	1240
254	Перампанел	9924495	929
255	Перициазин	4747	790
256	Перфеназин	4748	881
257	Пиклоксидин	71663	749
258	Пикоксистробин	11285653	1189
259	Пиколинафен	3294375	1315
260	Пилокарпин	5910	92
261	Пиракlostробин	6422843	1220
262	Пирантел	708857	422
263	Пирацетам	4843	52
264	Пиридабен	91754	1465
265	Пиридат	41463	1556
266	Пиридафол	92316	557
267	Пиридостигмин	4991	54
268	Пириметанил	91650	725
269	Пиримикарб	31645	532
270	Пиримифос-метил	34526	1161
271	Пирипроксифен	91753	1339
272	Пиритинол	14190	166
273	Пирлиндол	68802	716
274	Пироксикам	54676228	777
275	Платифиллин	5281742	516
276	Преноксидиазин	120508	867
277	Прометон	4928	686
278	Пропамокарб	32490	142
279	Пропафенон	4932	836
280	Пропахизафоп	16213016	1305
281	Прописамид	32154	1030
282	Пропиконазол	43234	1128
283	Пропоксифен	10100	855
284	Пропоксур	4944	767
285	Пропранолол	4946	703
286	Просульфокарб	62020	1285
287	Профам	24685	878

288	Прохиназид	11057771	1400
289	Прохлораз	73665	1017
290	Процимидон	36242	1095
291	Рабепразол	5029	604
292	Ранитидин	3001055	119
293	Рацекадотрил	107751	1073
294	Ребапимид	5042	740
295	Резерпин	5770	894
296	Рибавирин	37542	42
297	Рилменедин	68712	455
298	Римантадин	5071	640
299	Римсульфурон	91779	825
300	Рисперидон	5073	643
301	Рифаксимин	6436173	1056
302	Рифампин	135550179	1066
303	Ропивакаин	175805	592
304	Ропинирол	5095	481
305	Рутин	5280805	541
306	Себутилазин	23712	1050
307	Секвифенадин	42553	895
308	Секнидазол	71815	224
309	Серотонин	5202	61
310	Силденафил	135398744	767
311	Силодозин	5312125	643
312	Симвастатин	54454	1316
313	Соталол	5253	105
314	Спиродиклофен	177863	1482
315	Спироксамин	86160	961
316	Сульфагуанидин	5324	51
317	Сульфаметоксазол	5329	530
318	Сульфапиридин	5336	222
319	Сульфасалазин	5339	790
320	Сульфацетамид	5320	111
321	Сульфентразон	86369	1391
322	Сульфотеп	19395	1237
323	Суматриптан	5358	199
324	Супракс	6321411	470
325	Таксифолин	439533	518
326	Тамоксифен	2733526	1120
327	Тамсулозин	129211	664
328	Тебуконазол	86102	1081
329	Тебуфенозид	91773	1159

330	Телмисартан	65999	871
331	Темазепам	5391	897
332	Тенонитрозол	19646	861
333	Теобромин	5429	90
334	Теофиллине	2153	130
335	Тепралоксидим	135594055	1050
336	Теразозин	5401	554
337	Тербинафин	1549008	944
338	Тербуфос	25670	539
339	Тетраконазол	80277	1012
340	Тиабендазол	5430	269
341	Тиаклоприд	115224	649
342	Тиаметоксам	5485188	413
343	Тизанидин	5487	149
344	Тилозин	5280440	836
345	Тилорон	5475	536
346	Тимолол	33624	544
347	Тинидазол	5479	271
348	Тиоридазин	5452	958
349	Тирозин	6057	50
350	Тифенсульфурон-метил	73674	765
351	Толбутаид	5505	872
352	Толифлуанид	12898	1230
353	Толклофос-метил	91664	1233
354	Толтеродин	443879	829
355	Топирамат	5284627	685
356	Торсемид	41781	716
357	Тофизопам	5502	832
358	Тразодон	5533	669
359	Тралкоксидим	135741327	1353
360	Трамазолин	5524	580
361	Трандолаприл	5484727	881
362	Триадименол	41368	980
363	Триадимефон	39385	1042
364	Триазофос	32184	1098
365	Тримебутин	5573	796
366	Тримепразин	5574	836
367	Триметоприм	5578	424
368	Триптофан	1148	144
369	Тритиконазол	6537961	1007
370	Трифлуксистеробин	53627428	1298
371	Трифлумизол	91699	1111

372	Трифлумурон	47445	1183
373	Трихлорфон	5853	485
374	Трицилазол	39040	602
375	Троспий	5284632	734
376	Уденафил	135413547	820
377	Униконазол	6436604	999
378	Фабомотизол	9862937	554
379	Фамотидин	3325	94
380	Фамцикловир	3324	452
381	Фелодипин	3333	1168
382	Феназахин	86356	1342
383	Феназепам	40113	932
384	Фенамидон	10403199	1036
385	Фенамифос	31070	1045
386	Фенаримол	43226	1021
387	Фенацетин	4754	611
388	Фенбуконазол	86138	1102
389	Фенгексамид	213031	1059
390	Фенилбутазон	4781	1114
391	Фенилэфрин	6041	51
392	Фенмедифам	24744	972
393	Фенобукарб	19588	969
394	Феноксикарб	51605	1116
395	Фенпироксимат	6002001	1396
396	Фенпропидин	91694	938
397	Фенспирид	3344	253
398	Фентиконазол	51755	1150
399	Фентоат	17435	1196
400	Финастерид	57363	974
401	Флазасульфурон	93539	933
402	Флоникамид	9834513	347
403	Флувоксамин	5361192	850
404	Флуконазол	3365	533
405	Флуметсулам	91759	590
406	Флумиоксазин	92425	992
407	Флуоксетин	3386	890
408	Флуоксостробин	9804219	1132
409	Флуометурон	16562	821
410	Флуопиколид	11159021	1069
411	Флурбипрофен	3394	1041
412	Флусилазол	73675	1091
413	Флутриафол	91727	852

414	Флуфенацет	86429	1119
415	Флуфеноксурон	91766	1334
416	Флухинконазол	86417	1050
417	Фозалон	4793	1249
418	Фозиноприл	55891	1319
419	Форамсульфурон	11419598	763
420	Фосмет	12901	997
421	Фостиазат	91758	842
422	Фосфамидон	3032604	722
423	Фуберидазол	19756	359
424	Фуразолидон	5353636	383
425	Хиналфос	26124	1151
426	Хинаприл	54892	866
427	Хинокламин	17748	691
428	Хиноксифен	3391107	1249
429	Хлорамфеникол	5959	621
430	Хлоридазон	15546	518
431	Хлорпирифос	2730	1369
432	Хлорпротиксен	667467	899
433	Хлорсульфурон	47491	824
434	Хлорфенвинфос	5377791	1160
435	Холин	305	35
436	Целекоксиб	2662	1130
437	Цетиризин	2678	857
438	Цефтриаксон	13216808	492
439	Цефуроксим	5361202	514
440	Циазофамид	9862076	1162
441	Цилостазол	2754	921
442	Цимиазол	43714	608
443	Цимоксамил	5361250	597
444	Циннаризин	1547484	1008
445	Цинхофен	8593	1171
446	Ципродинил	86367	893
447	Ципроконазол	86132	1040
448	Элперенон	443872	794
449	Эпоксиконазол	3317081	1045
450	Эрдостеин	65632	137
451	Эритромицин	12560	796
452	Эрукамид	5365371	1627
453	Эсциталопрам	146570	757
454	Этион	3286	1391
455	Этиохоланолон	5880	904

456	Этиримол	135424354	552
457	Этифоксин	135413553	841
458	Этоксихин	3293	806
459	Этопрофос	3289	1048
460	Этофенпрокс	71245	1159
461	Этофумезат	33360	1056