МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА

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

АВТОРЕФЕРАТ

диссертации на соискание учёной степени

кандидата химических наук

Москва – 2024

Работа выполнена в Автономной некоммерческой образовательной организации высшего образования «Сколковский институт науки и технологий».

Научный руководитель: **Костюкевич Юрий Иродионович**

доктор химических наук

Официальные оппоненты: **Буряк Алексей Константинович**
доктор химических наук, профессор, член-корреспондент РАН
Федеральное государственное бюджетное учреждение науки Институт физической химии и электрохимии им. А.Н. Фрумкина Российской академии наук, директор
Григорьев Андрей Михайлович
доктор химических наук
Федеральное государственное бюджетное учреждение «27 Научный центр» Министерства обороны Российской Федерации, старший научный сотрудник
Мильман Борис Львович
доктор химических наук
Федеральное государственное бюджетное учреждение «Научно-клинический центр токсикологии имени академика С.Н. Голикова Федерального медико-биологического агентства, ведущий научный сотрудник

Защита диссертации состоится 20 марта 2024 года в 15 часов 00 минут на заседании диссертационного совета МГУ.014.5 Московского государственного университета имени М.В. Ломоносова по адресу: 119991, Москва, ГСП-1, Ленинские горы, д. 1, стр. 3, МГУ имени М.В. Ломоносова, Химический факультет, аудитория 446.

E-mail: dissovet02.00.02@mail.ru

С диссертацией можно ознакомиться в отделе диссертаций научной библиотеки МГУ имени М.В. Ломоносова (Ломоносовский просп., д. 27) и на сайте: https://dissovet.msu.ru/dissertation/014.5/2868

Автореферат разослан « » февраля 2024 года.

Учёный секретарь
диссертационного совета,
кандидат химических наук — И.А. Ананьева

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

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 низкомолекулярных соединений, 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 потребуется 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 (**CCS**) 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 $^{16}O/^{18}O$, а также дейтероводородного 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 хромато-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. Ограниченность библиотек, содержащих масс-спектры, сужает круг потенциально идентифицируемых веществ и снижает эффективность и достоверность идентификации.

Для моделирования хромато-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 $^{16}O/^{18}O$; разработан подход к применению изотопного обмена $^{16}O/^{18}O$ 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 $^{16}O/^{18}O$ 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 $^{16}O/^{18}O$ 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 $^{16}O/^{18}O$ 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 6 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, 27 таблиц и одно приложение.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность тематики работы, сформулирована цель и определены задачи исследования, показана научная новизна и практическая значимость работы.

Первая глава представляет собой обзор опубликованной литературы, систематизирующий вычислительные и экспериментальные подходы, применяемые при идентификации низкомолекулярных соединений в нецелевом хромато-масс-спектрометрическом анализе. Приведена краткая характеристика алгоритмов машинного обучения, применяемых для предсказания хромато-масс-спектрометрических параметров. Рассмотрены существующие вычислительные подходы к моделированию хроматографического удерживания, масс-спектров электронной ионизации и масс-спектров, полученных при диссоциации, индуцированной соударениями. Особое внимание уделено роли методов машинного и глубокого обучения.

Вторая глава содержит описание используемых в работе реагентов, образцов, применяемого оборудования, программного обеспечения, баз данных. Приведены условия и техника проведения экспериментов, и обработки данных.

Хромато-масс-спектрометрический анализ проводили на жидкостных хромато-масс-спектрометрических системах, состоящих из гибридного масс-спектрометрического детектора с квадрупольным масс-анализатором и орбитальной ионной ловушкой (Orbitrap) Thermo Scientific QExactive (Германия), оснащенных источниками электрораспылительной ионизации Thermo Scientific HESI-II Probe, Thermo Scientific Nanospray Flex Ion Source (США), а также системой Spectroglyph MALDI/ESI Injector interface (США), и жидкостных хроматографов Thermo Scientific Ultimate 3000 RSLC Nano (США) и Waters ACQUITY I-Class UPLC system (Великобритания). Регистрацию хроматограмм и обработку данных проводили при помощи программного обеспечения Thermo Scientific™ Xcalibur™ Software (версия 4.0) и Thermo Scientific Compound Discoverer 3.3.

Обучение, валидацию и оценку эффективности моделей машинного обучения проводили с применением библиотек хроматографического удерживания «METLIN Small molecule retention dataset (SMRT)) (США), центральной химико-аналитической базы данных Организации по запрещению химического оружия версии 21 (OCAD, v.21, 2019), а также масс-спектральной библиотеки NIST 20 (США).

Третья глава содержит оценку применимости методов машинного обучения для предсказания времен хроматографического удерживания в жидкостной хроматографии. Предложены новые подходы к моделированию времени удерживания, основанные на разных

алгоритмах машинного обучения, а именно градиентном бустинге, нейронной сети с архитектурой Трансформер и графовой нейронной сети с распространением сообщений.

При использовании градиентного бустинга сравнивали способы описания молекул с помощью физико-химических дескрипторов из библиотеки Mordred, и круговых фрагментных дескрипторов из библиотеки RDKit. При незначительном выигрыше в точности предсказаний, использование физико-химических дескрипторов требует значительно больших вычислительных ресурсов, в сравнении с фрагментными дескрипторами, поэтому использование последних более предпочтительно. При применении искусственных нейронных сетей с архитектурой Трансформер молекулы представляли в виде текстовых строк SMILES, которые переводились в числовые векторы унитарным кодированием. Использование канонической и неканонических форм записи строк SMILES позволило проводить аугментацию данных из обучающей выборки. Для обучения и применения графовых искусственных нейронных сетей с распространением сообщений молекулы описывали векторами признаков, соответствующих вершинам и ребрам молекулярного графа (т.е. атомам и химическим связям).

Регрессионная задача предсказания времени удерживания была разбита на две подзадачи. Первая сводилась к построению и обучению первичной модели на обучающей выборке большого объема, в качестве которой использовали базу данных METLIN Small molecule retention dataset, содержащую времена удерживания более 80000 химических соединений, измеренных в одних и тех же условиях хроматографического разделения в режиме обращенно-фазовой хроматографии. Однако, данная первичная модель способна моделировать времена удерживания только для идентичных условий разделения. Ввиду того, что в аналитической практике широко распространено применение различных неподвижных фаз, геометрий колонок, разнообразие подвижных и градиентов, данная модель сама по себе имеет ограниченную практическую ценность. Поэтому, на втором этапе решалась задача переноса результатов предсказаний, полученных с помощью первичной модели на другие экспериментальные условия. Для этого в работе предложены два подхода, один основан на построении дополнительной кусочно-линейной модели, описывающей взаимосвязь времен удерживания между двумя хроматографическими системами по обучающей выборке, второй – на «до-обучении» первичной модели в режиме обучения с переносом. В таблице 1 приведены средние значения абсолютной ошибки предсказаний, характеризующие первичные модели, а также перенос предсказаний на отличные условия разделения (набор данных «Eawag_XBridgeC18»). Приведенные результаты показывают, что искусственная графовая нейронная сеть с передачей сообщений превосходит другие предложенные методы по точности предсказаний,

Таблица 1. Оценка точности предложенных подходов к предсказанию.

Алгоритм машинного обучения / переноса результатов предсказаний	Средняя абсолютная ошибка по независимой тестовой выборке, с	
	METLIN SMRT	Eawag_XBridgeC18
Градиентный бустинг /кусочно-линейная функция	45.6±0.4	98.4
Искусственная нейронная сеть с архитектурой Трансформер / обучение с переносом	57±0.6	88.0±8.2
Искусственная графовая нейронная сеть с распространением сообщений / обучение с переносом	31.5±0.1	79.5 ± 10.3

Основной целью моделирования времени удерживания являлась задача фильтрации ложноположительных определений при идентификации химических соединений, зачастую сводящаяся к выбору кандидатов среди изомерных структур. Списки таких изомеров для соединений из тестовых выборок были получены из библиотеки PubChem поиском по элементному составу. При практическом применении подходов, элементный состав компонента может быть установлен по данным масс-спектрометрии высокого разрешения. Проведение фильтрации с использованием предсказанных времен удерживания, и пороговых значений, полученных из ROC-кривых, приводит сокращению пространства поиска в среднем на 23-53% (Рис. 1А). Это способствует снижению затрат при идентификации химических соединений интерпретацией фрагментных масс-спектров и уменьшает количество кандидатов для встречного синтеза. Целесообразность такой фильтрации возрастает при идентификации по обширным общехимическим базам данных (PubChem, ChemSpider). В работе показано, что максимальная эффективность применения данного подхода достигается при идентификации соединений, характеризующихся слабым или сильным удерживанием (Рис. 1Б).

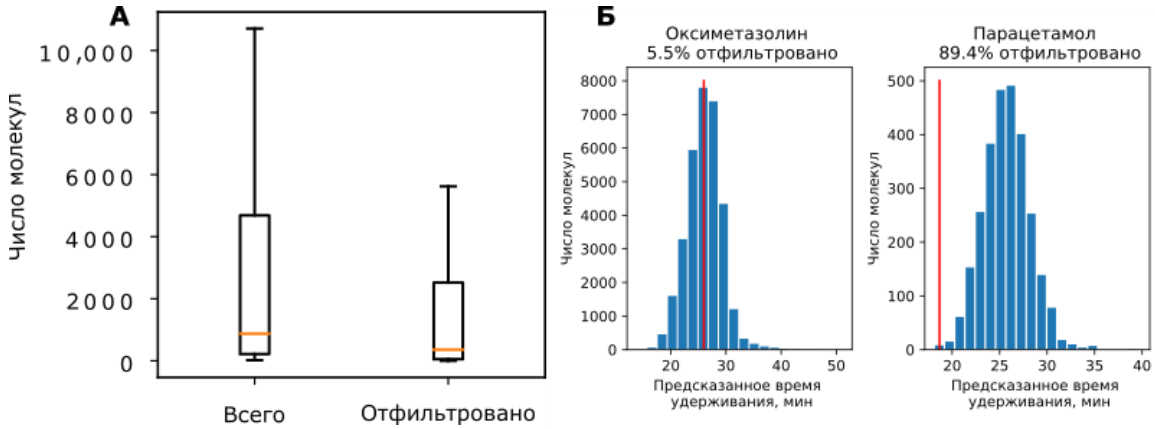

Рис. 1. (А) – Сокращение пространства поиска при фильтрации по предсказанным временам удерживания на примере данных из библиотеки Retip. (Б) – Пример эффективной фильтрации (парацетамол, слабое удерживание) и неэффективной фильтрации (оксиметазолин, среднее удерживание).

Все разработанные модели, алгоритмы находятся в открытом доступе, и доступны по ссылкам:

https://github.com/osv91/RTpredict (дата обращения 14.11.2023)

https://dx.doi.org/10.6084/m9.figshare.13315574 (дата обращения 14.11.2023)

https://github.com/osv91/MPNN-RT (дата обращения 14.11.2023).

Четвертая глава посвящена совместному применению фильтрации по предсказанным временам удерживания и данным изотопного обмена $^{16}O/^{18}O$ полученным хромато-масс-спектрометрией высокого разрешения. На наборе из более чем 100 кислородсодержащих соединений изучена селективность реакции изотопного обмена $^{16}O/^{18}O$, установлены функциональные группы, способные вступать в реакцию обмена, предложены два температурных режима проведения реакции (37°C и 95°C). Разработанный алгоритм фильтрации кандидатов по экспериментально установленному числу обменных атомов кислорода способствовал сокращению пространства поиска в среднем на 62%. На примере варфарина показано, что учет данных по фрагментации молекулы после изотопного обмена позволяет дополнительно отфильтровать 15% ложноположительных определений. Продемонстрировано, что изотопный обмен $^{16}O/^{18}O$ способствует структурной аннотации ионов в фрагментных спектрах, на примере фрагментного масс-спектра мебендазола удалось доказать формирование иона в следствие присоединение остаточных молекул воды в ячейке соударений.

Для работы с данными изотопного обмена $^{16}O/^{18}O$ 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 $^{16}O/^{18}O$, 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 14.11.2023 г.). Исходный код доступен по адресу https://github.com/osv91/16O-18O_isotope_exchange (дата обращения 14.11.2023 г.).

Для совместного применения метода изотопного обмена и фильтрации по предсказанным временам удерживания была собрана библиотека времен удерживания 500 соединений, которая была использована для до-обучения графовой искусственной нейронной сети с распространением сообщений, описанной в главе 3. Далее для тестовой выборки из соединений, в которых экспериментально зафиксирован изотопный обмен проведена фильтрация изомерных структур, полученных из базы данных PubChem поиском по элементному составу, по временам удерживания и результатам изотопного обмена. 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%.

Пятая глава посвящена предсказанию индексов удерживания в газовой хроматографии. Существующие универсальные модели предсказания ИУ отличаются высокой точностью, и в работе не удалось предложить более точных подходов с использованием библиотеки NIST Retention index library. Однако, в работе показано, что лучшие из существующих методов, основанных на моделировании ИУ по библиотеке NIST Retention library с применением методов машинного обучения, в частности одномерных сверточных нейронных сетей, не позволяют достичь точности, достаточной для практического применения в определенных задачах. В частности, точность предсказания ИУ для веществ, относящихся к спискам Конвенции по запрещению химического оружия характеризуется средним отклонением 39 единиц. Предложенный в работе подход к обучению модели на выборке соединений с известными экспериментальными значениями индексов удерживания, относящихся к спискам Конвенции, позволил снизить среднее отклонение, определенное в режиме кросс-валидации с 39 до 16 единиц., за счет более высокого структурного сходства молекул. Однако нужно признать, что такой подход ограничивает сферу применения полученной модели. 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.

Дальнейшее повышение точности было достигнуто с применением разновидности инкрементного метода. Соединения в списках Конвенции организованы в гомологические серии, внутри которых молекулы отличаются алкильным радикалом в одной боковой цепи. Считая, что изменение индекса удерживания при изменении радикала не зависит от ядра молекулы, можно вычислить разность между индексами удерживания молекулярной пары в одной гомологической серии по известным индексам удерживания молекулярной пары с аналогичными алкильными радикалами другой серии. Если для одного из соединений в первой серии значение индекса удерживания известно, то можно вычислить значение индекса удерживания второго вещества. Таким образом, задача предсказания времени удерживания может быть сведена к поиску трех молекул с известными индексами удерживания и необходимой структурой (Рис. 2).

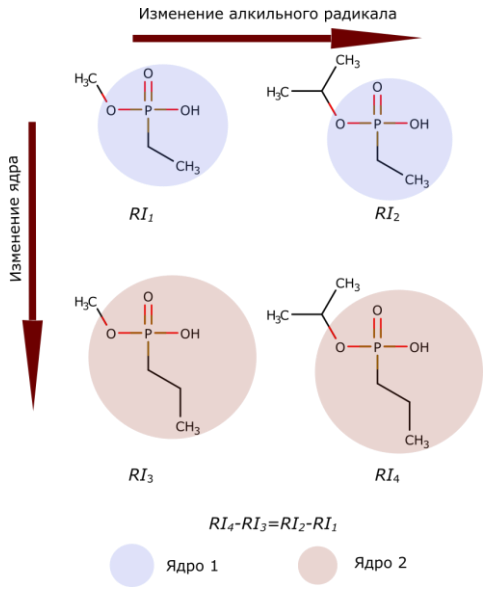

Рис. 2. Принцип предложенного инкрементного подхода к вычислению индекса удерживания.

Показано, что данный подход превосходит методы машинного обучения по точности, характеризуясь средним отклонением в 4 единицы. В рамках работы разработано программное обеспечение, реализующее соответствующий поиск по библиотеке Организации по запрещению химического оружия, а в случае отсутствия необходимых данных, предсказывает значение моделью машинного обучения. Данное программное обеспечение доступно по ссылке https://rt-cwc.anvil.app (Дата обращения 14.11.2023 г.)

Шестая глава посвящена моделированию масс-спектров электронного удара при помощи машинного обучения. Задача предсказания масс-спектров была сформулирована как задача многоцелевой регрессии следующим образом: в качестве независимых переменных использовали вектор фрагментных дескрипторов молекул, входящих в обучающую выборку, в качестве зависимых переменных использовали вектор интенсивностей сигналов в масс-спектре, где координата x_i вектора соответствует относительной интенсивности иона с $m/z = i$. Данное представление подразумевает использование целочисленных значений m/z, что в случае спектров электронной ионизации не влечет потери информации, т.к. большинство библиотек содержат масс-спектры с целочисленными представлениями m/z. Решение многоцелевой регрессионной задачи проводили с применением алгоритма градиентного бустинга реализованного в библиотеке XGBoost v.1.6.2 для языка программирования Python. Этот алгоритм хорошо зарекомендовал себя в задачах моделирования химических свойств молекул, благодаря точности предсказаний, а также высокой скорости обучения моделей.

Выявлено, что прямое моделирование масс-спектров имеет ограниченную точность, особенно при предсказании интенсивностей ионов в области высоких значений m/z. Наиболее вероятным объяснением является тот факт, что данные «тяжелые» ионы, образующиеся в

результате нейтральных потерь, характеризуются значительным разбросом по молекулярным массам, в отличие от «легких» ионов, которые являются общими структурными фрагментами для различных молекул. В результате возникает недостаток обучающих примеров для моделирования масс-спектра в области высоких масс. Для повышения точности конечных предсказаний в работе было предложено применять дополнительную независимую модель, при обучении которой спектры представлялись в виде спектров нейтральных потерь. Ввиду ограниченности списка возможных нейтральных потерь, это позволяет увеличить количество обучающих примеров, соответствующих каждой нейтральной потере. Для получения конечного результата предсказанные спектры (после обратного преобразования спектра нейтральных потерь в канонический вид) усредняются. Для каждой модели был подобран набор гипепараметров, в качестве функции потерь при обучении обеих моделей была использована средняя квадратичная ошибка. Для обучения модели использовали обучающую выборку на основе масс-спектральной библиотеки NIST 20 mass spectral library. Разработанное программное обеспечение доступно по ссылке https://figshare.com/articles/software/GBEIMS/21538965 (Последнее обращение 07 ноября 2023 г.).

Оценку эффективности модели проводили в режиме кросс-валидации с разбиением исходной обучающей выборки на 5 частей, а также независимой тестовой выборкой, которая не использовалась в процессе обучения. Выбор метрики для оценки результатов был сделан в пользу косинусной меры сходства библиотечного и предсказанного спектра, в т.ч. «взвешенной», вычисляемой по формуле:

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

где I_q, I_l –векторное представление искомого и библиотечного масс-спектров, m_k – значение m/z соответствующее координате k вектора I.

Данная метрика используется как мера сходства масс-спектров электронного удара в подавляющем большинстве программных продуктов, позволяющих проводить библиотечный поиск. Результаты кросс-валидации приведены в таблице 2. Необходимо отметить повышение точности предсказаний при использовании двух моделей, по сравнению с индивидуальными моделями предсказания масс-спектра и спектра нейтральных потерь. На рисунке 3А представлено распределение взвешенной косинусной меры сходства библиотечных спектров и спектров, предсказанных предложенным подходом GBEIMS. Можно видеть, что для подавляющего большинства молекул тестовой выборки значение данной меры превышает 0.6, что является пороговым значением при идентификации химических соединений в нецелевом газохроматографическом анализе, ниже которого сходство спектров считается неудовлетворительным.

Таблица 2. Результаты кросс-валидации (n=5) предложенного подхода по предсказанию масс-спектров электронного удара

	Взвешенная косинусная мера			Косинусная мера		
	Прямая модель	Модель нейтральных потерь	Усредненная модель	Прямая модель	Модель нейтральных потерь	Усредненная модель
Обучающая выборка	0.798±0.003	0.833±0.003	0.815±0.001	0.875±0.001	0.836±0.001	0.915±0.001
Валидационная выборка	0.798±0.003	0.833±0.003	0.851±0.001	0.875±0.001	0.836±0.002	0.915±0.005
Тестовая выборка	0.609±0.001	0.752±0.001	0.759±0.001	0.633±0.001	0.616±0.001	0.783±0.001

В диссертационной работе проведено сравнение возможностей предложенного подхода, основанного на алгоритме машинного обучения и открытого программного пакета Quantum-chemical electron ionization mass spectra (QCEIMS) для моделирования масс-спектров электронного удара. На рисунке 3Б представлено распределение взвешенной косинусной меры сходства спектров, предсказанных предложенным подходом GBEIMS и спектров, рассчитанных квантово-химическим программным пакетом QCEIMS и опубликованных в литературе. Можно видеть, что предложенный подход характеризуется существенно более высокой точностью, в тоже время обладая подавляющим преимуществом в скорости вычислений.

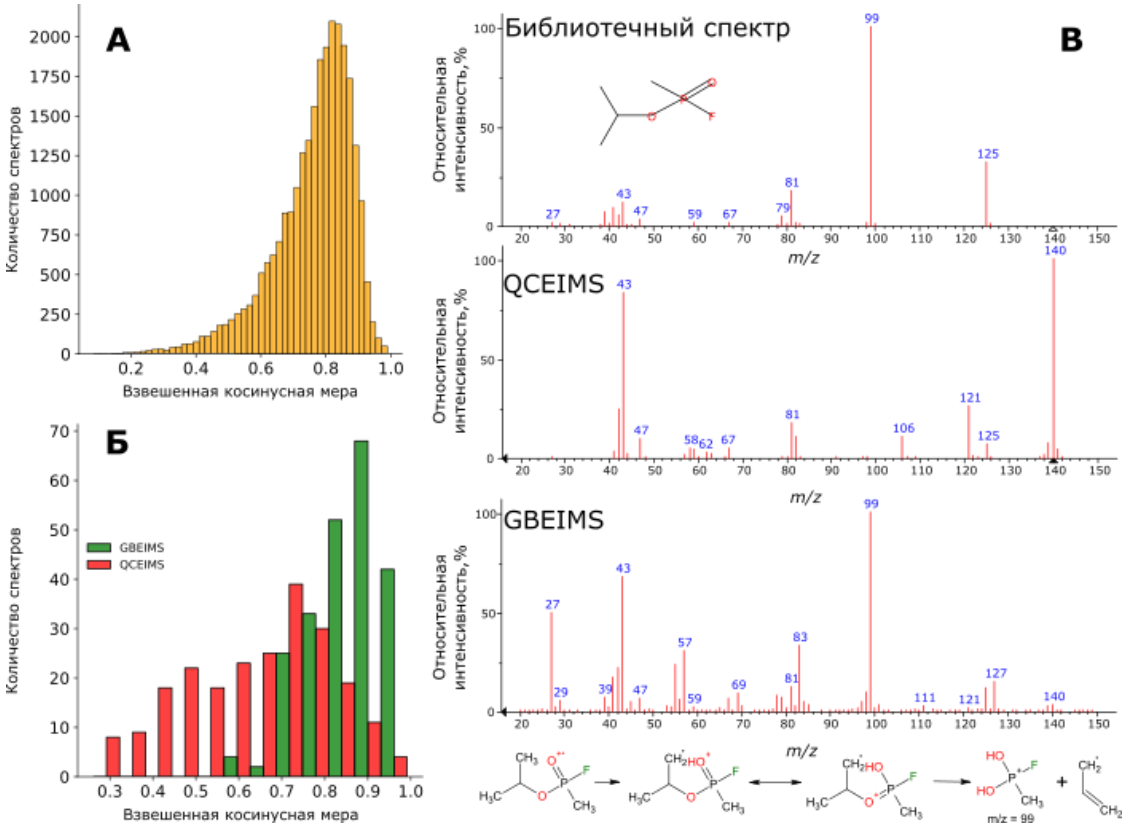

Рис. 3. (А) – Распределение взвешенной косинусной меры сходства библиотечных спектров и спектров, предсказанных предложенным подходом GBEIMS. (Б) – Сравнительное распределение взвешенной косинусной меры библиотечных спектров и спектров, предсказанных

предложенным подходом GBEIMS и квантово-химическим программным пакетом QCEIMS. (B) – сравнение библиотечного спектра Зарина и спектров, предсказанных предложенным подходом GBEIMS и квантово-химическим программным пакетом QCEIMS.

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.

ЗАКЛЮЧЕНИЕ

В результате проведенных исследований предложены новые подходы к моделированию аналитических характеристик химических 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 $^{16}O/^{18}O$. 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 химических 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 средне 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 химических соединений, используемых при идентификации химических соединений, и показано, что предсказанные величины могут быть использованы там, где экспериментальные 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 гомологичными соединениями 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 $^{16}O/^{18}O$ 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.

Основные результаты работы изложены в следующих публикациях:

<u>Научные статьи, опубликованные в рецензируемых научных журналах</u>, индексируемых в базах данных Web of Science, Scopus, RSCI и рекомендованных для защиты в диссертационном совете МГУ по специальности 1.4.2 Аналитическая химия:

1. **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. ИФ (Web of Science) – 4.478. 50%. 1.155 п.л.

2. **Osipenko S.**, Botashev K., Nikolaev E., Kostyukevich Y. Transfer learning for small molecule retention predictions // Journal of Chromatography A. 2021. Т. 1644. С. 462119. ИФ (Web of Science) – 4.601. 50%. 1.0395 п.л.

3. 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. ИФ (Web of Science) – 3.344. 50%. 1.155 п.л.

4. **Osipenko S.**, Nikolaev E., Kostyukevich Y. Retention Time Prediction with Message-Passing Neural Networks // Separations. 2022. Т. 9, № 10. С. 291. ИФ (Web of Science) – 3.344. 75%. 1.0395 п.л.

5. **Osipenko S.**, Zherebker 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. ИФ (Web of Science) – 3.262. 50%. 1.0395 п.л.

6. **Osipenko S.**, Nikolaev E., Kostyukevich Y. Amine additives for improved in-ESI H/D exchange // Analyst. 2022. Т. 147, № 14. С. 3180-3185. ИФ (Web of Science) – 5.227. 50%. 0.693 п.л.

Благодарности

Автор выражает искреннюю благодарность и признательность научному руководителю д.х.н. Костюкевичу Ю.И. за помощь в постановке задач и обсуждении результатов исследования; д.х.н., проф. чл.-корр. РАН Николаеву Е.Н. за ценные советы и наставления; д.х.н. Кирееву А.Ф., к.ф.-м.н. Кононихину А.С. за поддержку в научно-исследовательской работе и ценные замечания и комментарии при подготовке диссертации, к.х.н. Жеребкеру А.Я., PhD Соснину С.Б., к.ф.-м.н. Ковалевой О.А. за плодотворную совместную работу, а также всем сотрудникам лаборатории масс-спектрометрии Сколтеха.