

ОТЗЫВ официального оппонента
о диссертации на соискание ученой степени
кандидата химических наук Осипенко Сергея Владимировича
на тему: «ПРОГНОЗИРОВАНИЕ ХРОМАТО-МАСС-
СПЕКТРОМЕТРИЧЕСКИХ ХАРАКТЕРИСТИК ХИМИЧЕСКИХ
СОЕДИНЕНИЙ В НЕЦЕЛЕВОМ АНАЛИЗЕ С ПРИМЕНЕНИЕМ
МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ»
по специальности 1.4.2 Аналитическая химия

Диссертация С.В. Осипенко на соискание ученой степени кандидата химических наук по специальности 1.4.2 Аналитическая химия (химические науки) представлена в Диссертационный совет МГУ 014.5. Диссертационная работа подготовлена в Сколковском институте науки и технологий.

В диссертационной работе предложены и реализованы модели оценки/прогнозирования времен и индексов хроматографического удерживания (ИУ) и масс-спектров электронной ионизации (ЭИ) с целью применения этих данных в нецелевом химическом анализе. Рассмотрено применение вычисленных хроматографических параметров в сочетании с данными обмена изотопов кислорода $^{16}\text{O}/^{18}\text{O}$. Определена степень правильности прогнозирования хроматографических характеристик и масс-спектров при различной их валидации и сравнении с достаточно надежными экспериментальными данными.

Диссертационная работа содержит вводную и заключительную части, шесть основных глав, выводы, списки ссылок и сокращений. Работа изложена на 163 страницы, включает 57 рисунков, 27 таблиц и одно приложение.

Введение четко характеризует проблемы, решаемые в диссертационной работе, актуальность ее темы, цель и задачи исследования, несомненную научную новизну и достигнутую практическую значимость. Отчетливо представлены все характеристики работы, включая содержательные положения, выносимые на защиту, соответствие паспорту специальности, степень достоверности результатов, личный вклад автора и указания на апробацию работы на научных форумах.

Глава 1 представляет собой литературный обзор, освещающий вычислительные методы и экспериментальные подходы, применяемые при идентификации низкомолекулярных соединений в рамках нецелевого хроматомасс-спектрометрического анализа. Охарактеризованы алгоритмы машинного обучения и соответствующие способы вычислений, которые применяются для предсказания параметров хроматографического удерживания и масс-спектров. Подробно рассмотрены методы машинного (машинного глубокого) обучения.

В главе 2 указаны используемые химические реагенты, анализируемые образцы, применяемые аналитические приборы, а также релевантные программное обеспечение и базы данных. Детализированы техника проведения экспериментов и способы обработки данных.

В главе 3 рассмотрено применение методов машинного обучения для предсказания времен удерживания в жидкостной хроматографии и приведены результаты оценивания этих характеристик. Предложены три новых подхода к моделированию времени удерживания, которые основаны на разных алгоритмах машинного обучения. Сопоставлены способы описания молекул с помощью различных дескрипторов. Первичная модель предсказания времени удерживания построена на обучающей выборке большого объема (база данных METLIN). Также решена задача переноса оценок, полученных с помощью

первичной модели, на другие условия хроматографирования. Ошибка предсказания составила приблизительно 30 с и 80 с для первой и второй модели соответственно. Разработанные способы предсказания времени удерживания уменьшают число кандидатов на идентификацию, выведенных из данных масс-спектрометрии, вплоть до половины списка кандидатских соединений.

Глава 4 содержит результаты совместной идентификации по предсказанным временам удерживания и данным изотопного обмена $^{16}\text{O}/^{18}\text{O}$, которые получены в условиях применения масс-спектрометрии высокого разрешения. Список кандидатов на идентификацию в результате применения обеих методологий сокращался приблизительно на $\frac{3}{4}$ исходного количества соединений.

В главе 5 обсуждается предсказание газохроматографических индексов удерживания. Предсказание отличается хорошей точностью, если обучающая и тест-выборки включают соединения похожей структуры. Например, в этом случае точность предсказания ИУ для веществ, относящихся к спискам Конвенции по запрещению химического оружия, составляет 16 ед. ИУ. Дальнейшее повышение точности (среднее отклонение 4 ед. ИУ) было достигнуто с применением разработанной диссертантом разновидности известного инкрементного метода. Этот подход включает программный поиск структурных аналогов и компьютерное вычисление изменения ИУ для определенных структурных инкрементов.

В главе 6 рассмотрено предсказание масс-спектров ЭИ методами машинного обучения. Продемонстрировано, что прямое моделирование масс-спектров не приводит к высокой точности их прогнозирования. Для повышения соответствующей точности в работе было предложено применять еще одну независимую модель, в которой масс-спектры были представлены в виде спектров нейтральных потерь. Обучающая выборка сформирована на основе

масс-спектральной библиотеки NIST 20. Приведены показатели сходства (удовлетворительные значения) предсказанных и справочных масс-спектров. Предложенный подход к предсказанию масс-спектров ЭИ с алгоритмом машинного обучения дает лучшие результаты, чем в случае применения квантово-химических расчётов (программный пакет QCEIMS).

В целом, перед нами объемное законченное исследование, выполненное на высоком научном уровне. Избранная тема, несомненно, актуальна, поскольку реализация моделей предсказания хроматомасс-спектрометрических данных способствует расширению массивов справочной информации, необходимой для нецелевого химического анализа. Многие модели машинного обучения являются новыми. Научные положения, выводы и рекомендации, сформулированные в диссертации, полностью достоверны и обоснованы, поскольку базируются на применении современной экспериментальной техники и улучшенных моделей предсказания обсуждаемых характеристик, при перекрестном сравнении полученных результатов между собой и сопоставлении с результатами работ других авторов. Сформулированные цели исследования достигнуты, а задачи – решены. Подходы к предсказанию справочных хроматомасс-спектрометрических данных и массивы этих данных применимы практически для решения многих задач нецелевого анализа. Обширность апробации полученных результатов и личный вклад автора не вызывают сомнений. Опубликовано достаточное количество статей (6 публикаций) в высокорейтинговых научных журналах.

Отметим также, что в тексте диссертации обнаруживаются некоторые неточности, затрагиваются дискуссионные вопросы.

1. Эффективность предсказания времен удерживания проверялась для набора кандидатских соединений с определенной брутто-формулой (с. 59 диссертации и другие страницы). При сокращении числа кандидатов

учитывались все соответствующие изомеры из базы данных PubChem. В этом – определенная избыточность: в практическом анализе чаще всего обнаруживаются популярные/распространенные соединения (см. *Milman B.L., Zhurkovich I.K. Statistics of the Popularity of Chemical Compounds in Relation to the Non-Target Analysis // Molecules. 2021. Vol. 26, No. 8. P. 2394*), редкие соединения нечасто становятся идентифицированными аналитами, их необязательно учитывать. Таким образом, реальная степень сокращения списка кандидатов на идентификацию нуждается в переоценке.

2. Утверждение типа «... для проведения однозначной идентификации, которая требует совпадения масс-спектров со схожестью выше 800...» (с. 129 диссертации) не является полностью корректным. Критерии однозначной идентификации определяются при разработке методик целевого анализа, одного значения показателя сходства масс-спектров может быть недостаточно, чаще всего выбирают сходство масс нескольких характеристичных ионов и интенсивностей их пиков плюс сходство параметров хроматографического удерживания.

3. В работе диссертанта масс-спектры электронной ионизации предсказаны с хорошей точностью: доля правильных результатов в тестовой идентификации составляет 73 % (с. 125). Это вполне «приличный» результат: эталонная правильность идентификации по библиотекам масс-спектров экспериментального происхождения – приблизительно 80 % (классические работы Стайна и МакЛафферти). Сравнение же полученных прогнозов с результатами применения других алгоритмов, на наш взгляд, оказалось неполным. Сопоставление с другим алгоритмом машинного обучения (NEIMS, с. 126) оказалось в пользу диссертанта, но было бы логично также провести сравнение с весьма популярной программой CFM-ID, версии 2.0 и 3.0 (<https://cfmid.wishartlab.com/versions>).

Вместе с тем, указанные замечания не умаляют значимости диссертационного исследования. Диссертация С.В.Осипенко отвечает требованиям, установленным Московским государственным университетом имени М.В.Ломоносова к работам подобного рода. Содержание диссертации соответствует специальности 1.4.2 Аналитическая химия (по химическим наукам), а также критериям, определенным Положением о присуждении ученых степеней в Московском государственном университете имени М.В.Ломоносова.

Таким образом, соискатель Сергей Владимирович Осипенко заслуживает присуждения ученой степени кандидата химических наук по специальности 1.4.2 Аналитическая химия.

Официальный оппонент:

доктор химических наук,
ведущий научный сотрудник лаборатории токсикологической химии
органических соединений химико-аналитического отдела
Федерального государственного бюджетного учреждения
«Научно-клинический центр токсикологии имени академика
С.Н. Голикова Федерального медико-биологического агентства»
(ФГБУ НКЦТ им. С.Н. Голикова ФМБА России)

МИЛЬМАН Борис Львович

1 марта 2024 г.

Контактные данные:

тел.: 7 (921) 766 52 96, e-mail: bormilman@yandex.ru

Специальность, по которой официальным оппонентом
защищена диссертация:

02.00.02 – Аналитическая химия

Адрес места работы:

192019, г. Санкт-Петербург, ул. Бехтерева, д. 1,

ФГБУ НКЦТ им. С.Н. Голикова ФМБА России,

лаборатория токсикологической химии органических соединений

Тел.: +7 (812) 365-06-80; e-mail: institute@toxicology.ru

Подпись сотрудника

ФГБУ НКЦТ им. С.Н. Голикова ФМБА России

Мильмана Б.Л. удостоверяю:

Подпись Мильмана Б.Л.

