

**ОТЗЫВ официального оппонента
на диссертацию на соискание ученой степени
кандидата химических наук Хрисанфова Михаила Дмитриевича
на тему: «Поиск ошибок в хроматографических базах данных и
моделирование свойств идентифицируемых молекул
методами машинного обучения»
по специальности 1.4.2. Аналитическая химия**

Работа, представленная Хрисанфовым М.Д. на соискание ученой степени кандидата химических наук, выполнена в области математического обеспечения аналитической химии на примере хроматографии и масс-спектрометрии. Хроматографические и масс-спектральные методы активно применяются в различных сферах для качественного и количественного анализа сложных смесей различных соединений, исследования физико-химических характеристик индивидуальных соединений и смесей. При этом непрерывно растет объем данных об удерживании различных соединений в различных хроматографических режимах, что связано как с расширением ряда анализируемых соединений – от малых молекул до сложных биополимеров и смесей высокомолекулярных соединений, так и с повышением требований к точности идентификации в разных областях, от фармацевтики до экологического мониторинга. Развитие автоматизированных систем и высокопроизводительных методов анализа позволяет сегодня генерировать огромные массивы данных, которые затем могут быть использованы для построения прогностических моделей или оптимизации хроматографических условий. Это особенно важно в нецелевом скрининге сложных смесей (метабомика, экология, токсикология), где нужно идентифицировать неизвестные компоненты: предсказанные данные об удерживании или масс-спектры сужают круг кандидатов и ускоряют интерпретацию полученных экспериментальных данных. При этом неизбежно происходит постепенное накопление ошибок в хроматографических и масс-спектральных базах данных, вопрос о поиске и исключении которых обсуждается не первый год, например, в работах И.Г. Зенкевича и др., однако, проблема до сих пор не

нашла эффективного современного решения. В связи с этим поиск ошибок в базах данных, содержащих информацию о масс-спектрах или о хроматографическом удерживании соединений, критически важен, поскольку неточности в значениях индексов удерживания и масс-спектрах напрямую снижают надёжность идентификации соединений, приводят к ложным совпадениям или пропускам целевых веществ. Ошибочные данные искажают модели, используемые для прогнозирования свойств целевых соединений, мешают оптимизировать условия анализа, что в конечном счёте ведёт к потере времени, ресурсов и снижению достоверности получаемых результатов. В связи с этим работа Хрисанфова М.Д., посвященная усовершенствованию существующих и разработке новых подходов к обработке и предсказанию хроматомасс-спектральных данных, включая поиск ошибок в базах времен и индексов удерживания, а также предсказанию масс-спектров соединений по их структуре и молекулярных отпечатков пальцев по масс-спектрам электронной ионизации с использованием современных методов машинного обучения, является актуальной.

Научная новизна диссертационной работы Хрисанфова М.Д. определяется предложенным автором новым подходом к поиску ошибочных записей в хроматографических базах данных NIST RI и METLIN SMRT с использованием объединения предсказаний нескольких независимых моделей машинного и глубокого обучения путем голосования. Автором предложен новый алгоритм округления значений m/z масс-спектров низкого разрешения до целочисленных, позволяющий минимизировать влияние случайных приборных погрешностей на результаты. В работе Хрисанфова М.Д. улучшены существующие подходы к предсказанию масс-спектров электронной ионизации по структуре молекулы и к предсказанию молекулярных отпечатков пальцев по масс-спектрам электронной ионизации. Доля точно предсказанных масс-спектров впервые увеличена с 5% (NEIMS) до 28% благодаря выполненной автором оптимизации архитектуры нейросетевых моделей.

Практическая значимость работы определяется обеспеченной автором возможностью реализации предложенного им подхода к поиску ошибок в базах хроматографических данных в виде библиотеки для языка программирования Python с открытым исходным кодом, доступным для практической реализации пользователями. Предложенный Хрисанфовым М.Д. подход к обработке данных уже позволил найти 2093 потенциально ошибочные записи в базе данных NIST 17 RI и 1544 в METLIN SMRT, а предложенный подход к предсказанию молекулярных отпечатков пальцев более чем в 30 раз быстрее обучается и работает по сравнению с известными аналогами при более высокой точности, правильности и полноте предсказания.

Достоверность и обоснованность результатов исследований и сделанных научных выводов обеспечивается качеством и репрезентативностью обучающей выборки, грамотным подбором признаков и анализом ошибок модели для выявления и устранения системных проблем, что в комплексе позволяет получать надёжные и воспроизводимые научные выводы, а также применением общепринятых подходов и метрик к сравнению качества предсказания моделей, стандартных в области работы с данными библиотек для языка программирования Python. Сопоставление предлагаемых автором подходов и моделей друг с другом и с известными ранее моделями проведено с применением крупных современных баз данных и является вполне обоснованным и грамотным.

Диссертационная работа Хрисанфова М.Д. состоит из списка основных обозначений и сокращений, введения, обзора литературы, методов и алгоритмов, обсуждения и результатов (в двух главах), заключения, выводов, списка цитируемой литературы из 158 наименований. Диссертация изложена на 110 машинописных страницах и содержит 39 рисунков и 12 таблиц.

Изложению экспериментальной части и обсуждения результатов в работе предшествует обзор литературы, состоящий из двух частей, в котором подробно рассмотрены существующие подходы к предсказанию

хроматографических параметров удерживания (индексов и времен удерживания) и к поиску ошибок в широко применяемых неразмеченных наборах данных, а также подходы, позволяющие решать как прямую, так и обратную задачи предсказания масс-спектра электронной ионизации по структуре молекулы.

Следующая глава диссертации посвящена описанию методов и алгоритмов, используемых автором. В главе подробно описана архитектура предсказательных алгоритмов. Для работы с базой NIST 17 RI для предсказания значений индексов удерживания автор использовал одномерную сверточную нейронную сеть (CNN(1D)), двухмерную сверточную нейронную сеть (CNN(2D)), полносвязную нейронную сеть (MLP), градиентный бустинг в реализации CatBoost и метод регрессии на опорные вектора (SVR). Для обучения моделей в работе использована средняя абсолютная ошибка. Для предсказания времен удерживания и поиск ошибок в базе METLIN SMRT автор использовал полносвязную нейронную сеть с дескрипторными входными данными (FCFP), полносвязную нейронную сеть с входными данными на основе молекулярных отпечатков пальцев (FCFP), одномерную сверточную нейронную сеть (CNN), графовую сверточную нейронную сеть (GNN) и градиентный бустинг в виде CatBoost. Для предсказания масс-спектров-электронной ионизации по молекулярной структуре с NEIMS-PyTorch автор использовал базы экспериментальных масс-спектров электронной ионизации NIST 23 MS и NIST 17 MS. Тестовый набор, используемый автором, содержал 5000 случайно выбранных молекул из числа присутствующих в основной библиотеке (mainlib) NIST 23 MS, но отсутствующих в NIST 17 MS. Для фильтрации данных использовали подход на основе «желтых карточек».

Две следующие главы диссертации посвящены обсуждению полученных автором результатов при обработке хроматографических и масс-спектральных данных. В главе 3 рассматривается подход к фильтрованию данных с использованием метода «желтых карточек», в том числе влияние

ошибок в тренировочном наборе на точность предсказания, проведено соотнесение ошибок предсказания для разных моделей между собой. Автором проведено сравнение эффективности фильтрации «желтых карточек» с другими подходами: отсечением по абсолютной ошибке (все записи с ошибкой предсказания больше заданной считаются ошибочными) или по квантилю/процентиллю (определенная доля наименее точно предсказанных записей считается ошибочными). При поиске ошибок в базе NIST 17 RI автором обнаружен значительный объем данных из одного источника, которые ранее действительно были неверно внесены в базу – ошибка произошла при переносе первичных величин в общую базу данных – и в последующих релизах были скорректированы (в 76% из найденных записей были внесены изменения, а 23% неверно внесенных величин удалены без коррекции), что подтверждает работоспособность предлагаемого автором подхода. Также в главе 3 описаны результаты, полученные автором при поиске ошибок в базе данных METLIN SMRT. Набор потенциально ошибочных времен удерживания для этой базы данных, в соответствии с прогнозом автора, содержит порядка 1200 записей, а размер группы с 5 «желтыми карточками» составляет примерно 1500 записей. К сожалению, в рассматриваемой базе данных отсутствуют очевидные ошибки, которые можно подтвердить внешними методами, однако, два использованных автором независимых способа оценки показали сходные результаты, которые совпадают с пересечением из списков ошибок для 3 повторений, что подтверждает устойчивость модели. Эффективность фильтрации ошибочных записей автор оценивает на синтетических наборах данных с контролируемо добавленными ошибками и сравнивает предложенное решение с существующими альтернативными вариантами в пользу предлагаемого подхода.

В главе 4 автором рассматривается проблема обработки масс-спектральных данных, которые широко используются в современной аналитической химии при нецелевом скрининге, поиске конкретных

соединений в составе сложных смесей, качественном анализе объектов различной природы. Объем данных, генерируемых в ходе одного хромато-масс-спектрального эксперимента, значительно превышает таковой для всех других детектирующих систем в газовой и жидкостной хроматографии, а также получаемые данные являются первостепенными при качественном анализе компонентов смеси и лишь в спорных ситуациях дополняются данными об удерживании. Предлагаемый автором способ уменьшения размерности данных, особенно для масс-спектральных детекторов низкого разрешения, состоит в округлении значений m/z с плавающей запятой до целочисленных: как и любая другая экспериментально полученная величина, значение m/z измеряется с некоторой случайной погрешностью, однако, для дальнейшей обработки дробная часть значений m/z может быть отброшена при условии сохранения постоянства величин: значения m/z , принадлежащие одному фрагменту молекулы, должны округляться к одному и тому же целому. С учетом того, что существующее ПО для обработки масс-спектральных данных использует различные алгоритмы для округления нецелочисленных значений m/z , в работе были сгенерированы два модельных набора данных на основе молекул из базы PubChem и всех комбинаторно возможных фрагментов из NIST 17 MS, что позволило предложить устойчивый алгоритм для округления значений из отрезка $m/z \in [-0,38 : Z + 0,62]$ до целочисленного m/z , находящего на уровне, соответствующем существующим зарубежным коммерческим аналогам $m/z \in [-0,351 : Z + 0,649]$.

В главе 4 также описаны разработанные автором два варианта полносвязных нейросетевых предсказательных моделей: EI2FP:Lite и EI2FP:Full, способные предсказывать несколько типов молекулярных отпечатков пальцев одновременно по масс-спектрам электронной ионизации. Обе модели превосходят представленный ранее подход DeepEI по удобству использования, точности предсказания и скорости. Полученные в работе модели могут быть использованы в составе комплексного пакета для обработки ГХ/МС данных, а предсказанные по масс-спектрам электронной

ионизации молекулярные отпечатки пальцев позволяют проводить поиск неизвестных соединений по общехимическим базам данных и могут применяться в алгоритмах ранжирования для сокращения числа возможных кандидатов. Подтверждением этого служат приведенные в заключительной части результаты улучшения известных подходов к предсказанию масс-спектров электронной ионизации по структуре молекулы и к предсказанию молекулярных отпечатков пальцев по масс-спектрам электронной ионизации. Доля точно предсказанных масс-спектров увеличена с использованием результатов автора с 5% (NEIMS) до 28% благодаря оптимизации архитектуры используемых нейросетей.

В целом прочтение диссертации позволяет сделать вывод, что полученные диссертантом результаты вносят существенный вклад в расширение возможностей современного математического обеспечения методов аналитической химии. По теме диссертации опубликовано 4 статьи в рецензируемых научных журналах, рекомендованных для защиты в диссертационном совете МГУ по специальности и отрасли наук. Результаты работы представлены в 14 тезисах докладов на отечественных и международных конференциях.

В то же время, как и любое большое исследование работа не лишена спорных моментов и вопросов, основные из которых отмечены ниже:

- 1) Приведенные погрешности (см., например, Табл. 5 или 6) должны быть округлены по первой значащей цифре, результат – до того же десятичного разряда, что и округлённая погрешность. Это особенно важно в предсказательных моделях, когда автором прогнозируются конкретные числовые значения индексов или времен удерживания.

2) Что автор подразумевает по «приемлемым качеством аппроксимации» (стр. 77)? Необходимы четкие критерии.

3) Прогнозирование абсолютных значений времен удерживания является достаточно ограниченным подходом, поскольку времена удерживания зависят не только от используемой неподвижной фазы, но и от конкретного приборного обеспечения, используемых хроматографических условий. Насколько предлагаемая авторами модель может быть распространена на вариативные условия хроматографирования?

4) Проводился ли авторами подтверждающий хроматографический эксперимент или моделирование проводили только на основании информации, представленной в базах данных?

5) В тексте диссертации и автореферате отсутствует ссылка на библиотеку с открытым исходным кодом, доступным для практической реализации пользователями.

Вместе с тем, указанные замечания не умаляют значимости диссертационного исследования. Диссертация Хрисанфова М.Д. 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 пп. 2.1-2.5 Положения о присуждении ученых степеней в Московском государственном университете имени М.В.Ломоносова. Диссертационное исследование оформлено согласно требованиям Положения о совете по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук Московского государственного университета имени М.В.Ломоносова.

Таким образом, соискатель Хрисанфов Михаил Дмитриевич заслуживает присуждения ученой степени кандидата химических наук по специальности 1.4.2. Аналитическая химия.

Официальный оппонент:

кандидат химических наук
ведущий научный сотрудник

Канатьева
Анастасия Юрьевна

« 4 » августа 2026

Федеральное государственное бюджетное учреждение науки Ордена
Трудового Красного Знамени Институт нефтехимического синтеза им.
А.В.Топчиева Российской академии наук (ИНХС РАН)
Лаборатория спектральных и хроматографических исследований

Контактные данные:

Специальность, по которой официальным оппонентом защищена
диссертация: 02.00.04 – Физическая химия

Адрес места работы:

ГСП-1, 119991, г. Москва, Ленинский проспект, д. 29

Федеральное государственное бюджетное учреждение науки Ордена
Трудового Красного Знамени Институт нефтехимического синтеза им.
А.В.Топчиева Российской академии наук (ИНХС РАН)

Подпись Канатьевой А.Ю. удостоверяю
Ученый секретарь ИНХС РАН
Д.х.н.

О.В.

уста 2026