ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. ХИМИЯ, НАУКИ О МАТЕРИАЛАХ, 2020, том 492–493, с. 116–122

ФИЗИЧЕСКАЯ ХИМИЯ

УДК 542

МАСС-СПЕКТРОМЕТРИЧЕСКАЯ ИДЕНТИФИКАЦИЯ ХИМИЧЕСКИХ СОЕДИНЕНИЙ ДЛЯ РАЗВИТИЯ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

© 2020 г. Ю. В. Бурыкина[1], Д. А. Бойко[1,2], В. В. Ильюшенкова[1], Д. Б. Еремин[1], академик РАН В. П. Анаников[1,2,*]

Поступило 03.04.2020 г.
После доработки 27.05.2020 г.
Принято к публикации 05.06.2020 г.

В данной работе обсуждается число возможных структурных вариантов и затрачиваемого машинного времени на автоматизированную интерпретацию масс-спектрометрических данных в зависимости от точности измерений экспериментальной массы. Проведены экспериментальные измерения для двух типов модельных объектов в диапазоне малых и средних масс в условиях ионизации электрораспылением на приборах с высокой, ультравысокой и сверхвысокой разрешающей способностью. Проанализировано число возможных решений и перспективы использования методов машинного обучения для предсказания новых данных о механизме реакций и поиска скрытых зависимостей в химическом пространстве на базе масс-спектрометрических данных. Показано, что для молекул среднего размера зависимость количества предсказанных решений от точности экспериментальных измерений имеет нелинейный вид, а для молекул большего размера данная кривая демонстрирует линейную зависимость.

Ключевые слова: масс-спектрометрия, масс-спектрометрия ионно-циклотронного резонанса с преобразованием Фурье, масс-спектрометрия с ионизацией электрораспылением, машинное обучение, искусственный интеллект

DOI: 10.31857/S2686953520030048

ВВЕДЕНИЕ

Масс-спектрометрия на сегодняшний день является одним из важнейших аналитических методов для исследования состава и структуры химических соединений [1–4]. Отличительной чертой масс-спектрометрии является высокая чувствительность вплоть до 1×10^{-18} М для рутинных измерений [5]. Благодаря комбинации универсальности метода и высокой чувствительности появилась возможность идентификации сложных смесей. Важнейшим достоинством масс-спектрометрического анализа является накопление большого объема данных об исследуемом объекте с высокой скоростью: от тысяч до десятков тысяч индивидуальных сигналов в спектре регистрируются за короткий промежуток времени [6, 7]. Таким образом, три ключевых фактора: универсальность, высокая чувствительность и скоростное накопление больших объемов данных – делают масс-спектрометрию одним из наиболее востребованных экспериментальных методов для развития алгоритмов машинного обучения и искусственного интеллекта.

Технологии машинного обучения и искусственного интеллекта привлекают все большее внимание исследователей по всему миру в связи с появлением доступных вычислительных возможностей. Оказалось, что для решения сложной задачи можно сконструировать набор из пар исходные данные—корректные решения и потом попытаться подобрать алгоритм, который максимально точно описывает связь между ними. Иногда количество параметров у этих алгоритмов достигает огромных значений: десятки миллионов параметров у нейросетей для классификации изображений, таких как ResNet, Inception, EfficientNet [8]. А так как правила для принятия решения не задаются напрямую, алгоритмы находят в данных скрытые зависимости, необходимые для решения задачи. Сейчас эти технологии все больше и больше проникают в нашу жизнь: распознавание лиц [9], беспилотные автомобили [10], обработка

[1] *Институт органической химии им. Н.Д. Зелинского Российской академии наук, Москва, Россия*
[2] *Химический факультет, Московский государственный университет имени М.В. Ломоносова, Москва, Россия*
**E-mail: val@ioc.ac.ru*

естественного языка (например, перевод текстов) [11].

В последние годы машинное обучение все чаще используется исследователями для получения новой информации о химических процессах и планирования экспериментов. Развитие структурных решений на базе машинных алгоритмов для масс-спектрометрических исследований имеет несомненное значение. Можно выделить три важных направления: использование масс-спектрометрических данных как векторных представлений без промежуточной идентификации соединений, установление структуры соединений, представленных в масс-спектрах, и использование масс-спектрометрии как просто удобного метода регистрации отдельных соединений. Первое направление задействует ряд различных алгоритмов и включает следующие примеры: классификация масс-спектров биоматериала по наличию определенного заболевания [12, 13] или кластеризация набора масс-спектров [14]. Два других вызывают сложности в настоящее время, особенно для многокомпонентных образцов сложного состава [15]. Решение проблемы идентификации соединений в таких образцах позволит заметно ускорить химические исследования за счет сокращения времени на интерпретацию спектров, а также откроет возможности для выявления сложных и неявных зависимостей в больших наборах данных, собранных для химических процессов, что является целевой задачей для методов машинного обучения и искусственного интеллекта [16–18].

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Для интерпретации масс-спектра необходимо решить три задачи: (1) идентифицировать группы пиков, относящихся к одному соединению; (2) определить брутто-формулы вещества по точной массе с учетом изотопного распределения; и (3), если возможно, на основании спектров фрагментации различными методами активации предложить возможные структуры или структурные фрагменты.

Ключевое значение для однозначного определения молекулярной формулы вещества, особенно в сложных спектрах с перекрыванием сигналов, имеет качество спектральных данных. Так, если в образце присутствует большое количество примесей, то при перекрывании сигналов могут исказиться определяемые массы, а также интенсивности пиков. Это усложнит сравнение с теоретическим изотопным распределением возможного кандидата или сделает невозможным в принципе. В данной работе нами было проведено сравнительное исследование модельных образцов из области малых молекул (хлорид палладия ($PdCl_2$), MW = 177.33, предшественник катализатора в процессах тонкого органического синтеза) и молекул большего размера (лизоцим, MW = = 14.3 кДа, антибактериальный агент, фермент класса гидролаз, разрушающий клеточные стенки бактерий гидролизом пептидогликана) на приборах с высоким, ультравысоким и сверхвысоким разрешением и апробирована возможность получения качественных масс-спектрометрических данных для развития алгоритмов машинного обучения и искусственного интеллекта.

Первая задача решается за счет объединения пиков, массы которых отличаются на заданную величину. На рис. 1 представлен масс-спектр высокого разрешения с ионизацией электрораспылением (ESI-MS), полученный для раствора хлорида палладия с концентрацией 1×10^{-6} моль/л в ацетонитриле в режиме регистрации отрицательных ионов, в котором большинство групп пиков были идентифицированы. Несмотря на простоту исследуемого объекта ($PdCl_2$, CH_3CN), в спектре присутствует более 500 индивидуальных линий [19]. В настоящее время в литературе практически нет опубликованных работ с расшифровкой наблюдаемых сигналов для многих простых солей такого типа.

На рис. 2 приведен фрагмент масс-спектра протонированного белка лизоцима с концентрацией 1×10^{-6} моль/л в диапазоне масс m/z 1000–2200. В ходе работы был проведен сравнительный анализ масс-спектров, зарегистрированных с использованием различных масс-анализаторов: времяпролетных (разрешающая способность до ~23000 для m/z 922 в процессе калибровки) и ячейкой ионно-циклотронного резонанса с преобразованием Фурье (ФП-ИЦР) (разрешающая способность >100000) [20]. Наглядно представлено влияние разрешения на внешний вид спектра анализируемого образца. При высоком разрешении (рис. 2в; разрешающая способность ~10000 для m/z 922 в процессе калибровки) изотопы не разрешены и наблюдается широкий пик, а при ультравысоком значении разрешения (рис. 2б; разрешающая способность до ~26000 для m/z 922 в процессе калибровки) ширина пика на полувысоте (FWHM) составляет 0.0565, что является необходимым, но не достаточным условием однозначной интерпретации масс-спектра и идентификации белка. В то же время результат, полученный на ФП-ИЦР (сверхвысокое разрешение), позволяет надежно идентифицировать изотопную структуру и сделать вывод о составе анализируемого белка (рис. 2а).

Дальнейшее сравнение полученных данных было проведено по разрешению спектральных линий, ширине пика на полувысоте, точности определения массового числа (отношение массы к заряду) и изотопному распределению (табл. 1) [21, 22].

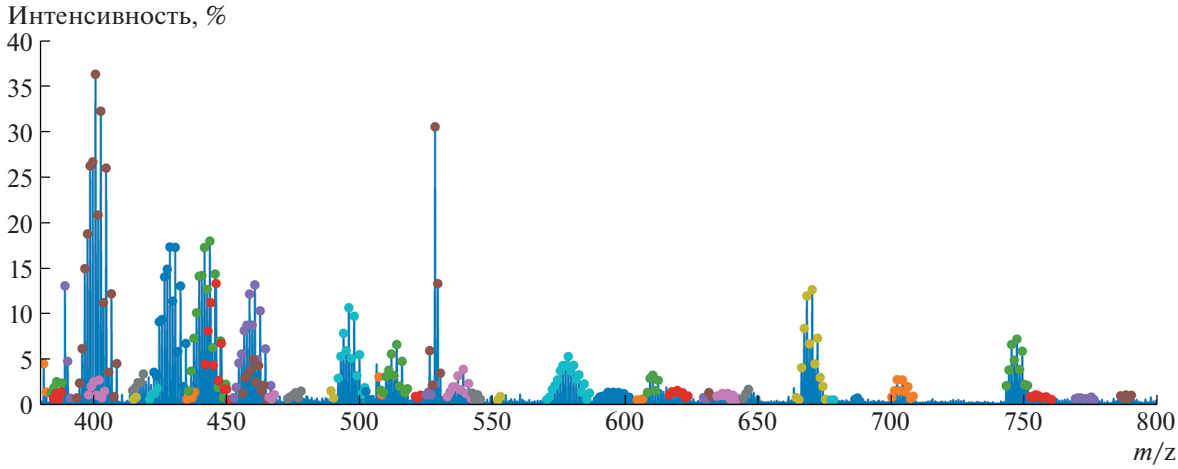

Рис. 1. Общий вид ESI-(–)MS спектра $PdCl_2$ в CH_3CN после поиска групп сигналов индивидуальных компонентов и цветовой маркировки.

Для оценки возможного количества вариантов при заданной точности определения массы нами был осуществлен полный перебор возможных значений. Наглядная демонстрация такого подхода была апробирована на спектре хлорида палладия, зарегистрированного в режиме отрицательных ионов (рис. 3).

На гистограмме количество формул сначала растет, а потом, наоборот, падает (рис. 3а). Это связано с тем, что отличие в массе все больше приближается к массе одного из элементов, который учитывался в переборе, – водорода. Проявление этой зависимости наблюдается на второй кривой: правая ее половина демонстрирует нелинейный рост, а при дальнейшем росте ошибки скорость роста количества формул постепенно снижается (рис. 3б). Из проведенного моделирования видно, что с увеличением точности (уменьшением ошибки Δ, ppm) количество предсказанных молекулярных решений резко снижается, стремясь в пределе к одному решению.

Ситуация для лизоцима выглядит по-другому – масса белка достаточно большая, поэтому для разницы в 1 Да существует огромное количество вариантов построения формулы. При средней скорости проверки в 221 тысячу формул в секунду (скрипт написан на языке программирования Python, запускался в один поток на Intel Xeon W-2145 с частотой 3.7 ГГц) полный перебор для лизоцима занял бы около тысячи лет. Конечно, оптимизация перебора может значительно улучить ситуацию, но при этом перебор все равно займет продолжительное время. Чтобы получить данные для сравнения, симуляция была проведена методом Монте-Карло. Оказалось, что распределение масс в интервале ± 1 Да становится равномерным – на каждом достаточно большом участке этого интервала находится примерно одинаковое количество формул. В результате все столбцы гистограммы имеют примерно одинаковую высоту (рис. 4а), а график для количества формул с точностью, не меньше заданной, – это просто прямая линия (рис. 4б). Из проведенного моделирования видно, что при увеличении точности наблюдается

Таблица 1. Основные характеристики сигналов в масс-спектрометрах*

Тип измерения**	m/z^{exp}	m/z^{calc}	FWHM, $\Delta m_{50\%}$	R	RMS, ppm	Δ, ppm
Высокое разрешение (HRMS, microTOF)	1590.3267	1590.3221	0.3131	5080	3.912	2.892
Ультравысокое разрешение (UHRMS, maXis)	1590.3246	″	0.0565	28143	1.275	1.572
Сверхвысокое разрешение (eXtreme RMS, SolariX)	1590.32178	″	0.01428	111407	0.4524	0.2

* Обозначения: m/z^{exp} – экспериментальное значение m/z; m/z^{calc} – теоретическое значение m/z; FWHM, $\Delta m_{50\%}$ – ширина пика на полувысоте; R – разрешающая способность; RMS (ppm) – среднеквадратичное отклонение; Δ (ppm) – относительная погрешность. ** В скобках приведено название прибора и аббревиатура разрешающей способности.

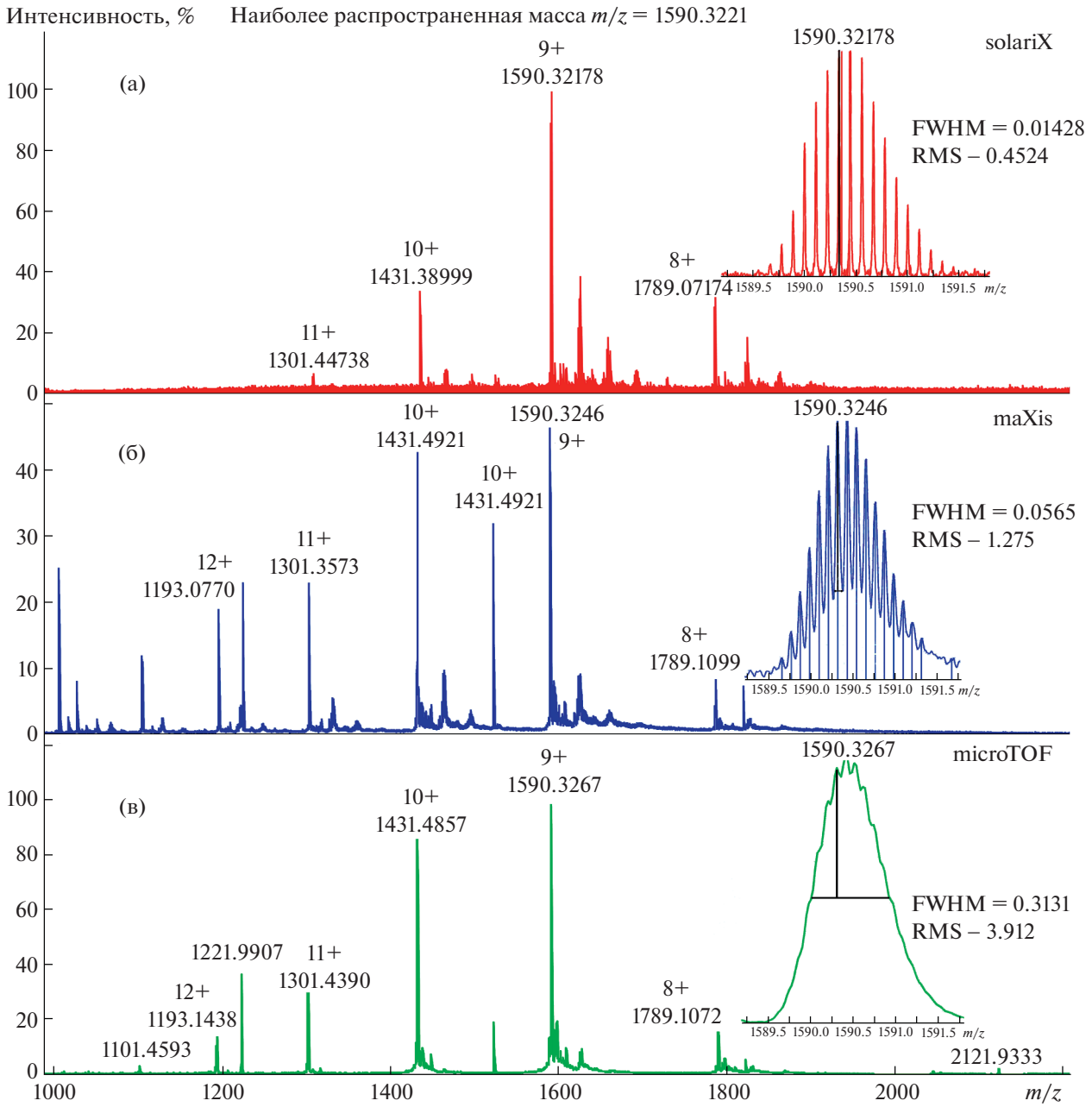

Рис. 2. Общий вид ESI-(+)MS спектров, зарегистрированных в положительном режиме, лизоцима с концентрацией 1×10^{-6} моль/л: (в) – высокое, (б) – ультравысокое, (а) – сверхвысокое разрешение. Расчетные значения m/z 1788.986469 ($z = 8+$), 1590.322115 ($z = 9+$), 1431.390631 ($z = 10+$), 1301.355781 ($z = 11+$), 1192.993405 ($z = 12+$). Значение RMS дано в ppm.

сходимость к одному решению, но сам график на рис. 4б теряет нелинейность, а количество вариантов возрастает примерно на шесть порядков.

Сравнение полученных масс-спектров лизоцима и хлорида палладия для различных значений разрешающей способности приборов и оценка использования этих данных для машинного обучения показала, что корректная интерпретация результатов может быть проведена на основе масс-спектров с высоким разрешением по точно определенному отношению массы к заряду путем перебора для масс, меньших 1 кДа. Учет изотопного распределения анализируемого вещества позволяет дополнительно улучшить качество ранжирования возможных формул, решая проблему систематических погрешностей. Минимальное значение разрешения, необходимое для полного разделения двух соседних изотопных пиков, эквивалентно массе зарегистрированного иона. Например, для разрешения сигнала иона

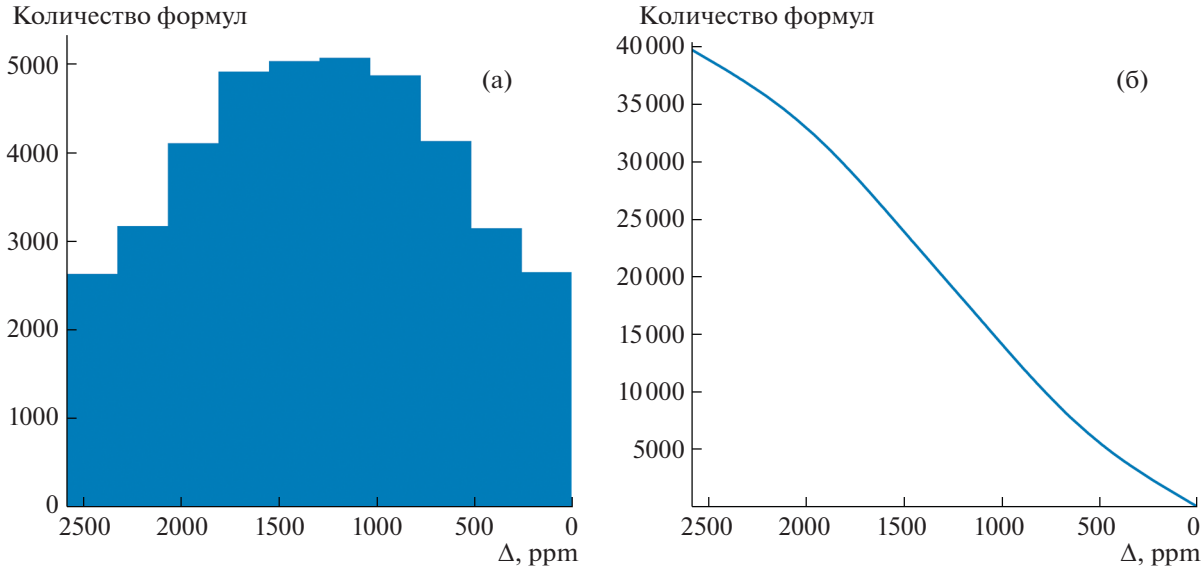

Рис. 3. (а) – гистограмма относительных ошибок Δ для сгенерированных формул в случае Pd_2Cl_5 (заряд = –1, моноизотопная масса – 386.65230); гистограмма отражает количество формул в интервале ошибок, не включая формулы с большей или меньшей ошибкой относительно верхней и нижней границы интервала соответственно; (б) – зависимость количества формул от максимальной относительной ошибки (для каждой ошибки указано количество формул, ошибка для которых не превышает заданную).

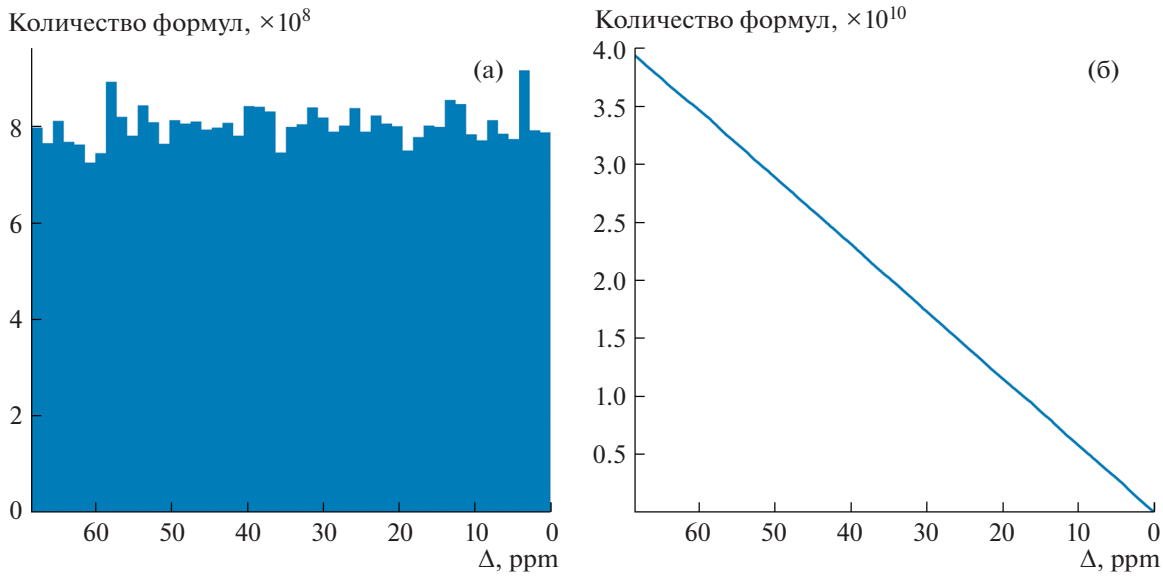

Рис. 4. (а) — гистограмма относительных ошибок Δ для сгенерированных формул в случае лизоцима (заряд = +9, моноизотопная масса – 1589.431144); гистограмма отражает количество формул в интервале ошибок, не включая формулы с большей или меньшей ошибкой относительно верхней и нижней границы интервала соответственно; (б) — зависимость количества формул от максимальной относительной ошибки (для каждой ошибки указано количество формул, ошибка для которых не превышает заданную).

лизоцима (заряд = +9, наиболее распространенная масса – 1590.322115) минимальное значение разрешающей способности прибора составляет >15900 [4]. Дальнейшее подтверждение этого подхода будет проводиться с использованием широкого набора масс-спектров для исследуемого класса соединений.

Использование методов машинного обучения для предсказания новых данных о механизме реакций и поиска скрытых зависимостей в химическом пространстве является актуальной задачей для современного ученого. Повышение точности результата интерпретации предложено проводить полным перебором, и, в случае больших масс,

снижать количество вариантов, заменяя атомы на целые группы (например, аминокислоты для белков). Алгоритмы машинного обучения следует использовать для сложных предсказательных задач, которые невозможно решить при помощи правил или простых, вручную построенных моделей. Проведенные в данной работе экспериментальные оценки и компьютерное моделирование показали, что в зависимости от массы иона может наблюдаться различная зависимость между точностью определения m/z и количеством возможных формул: нелинейная S-кривая для средних масс и линейные зависимости для больших масс. При этом количество возможных формул начинает расти очень быстро в обоих случаях, делая точность измерения отношения массы к заряду ключевым параметром, определяющим дальнейшее использование спектра.

ОБОРУДОВАНИЕ И МЕТОДИКА ЭКСПЕРИМЕНТА

Лизоцим энзим (EC 3.2.1.17, 51100 ед./мг, 14295.8 г/моль) производства Sigma (Merck-Sigma). Хлорид палладия(II) (99.9%, ca. 60% Pd) производства ABCR (ABCR GmbH & Co. KG). Ацетонитрил (LC-MS Grade, J.T. Baker), метанол (LC/MS, LiChrosolv, Merck, Praha), муравьиная кислота (LC-MS, Sigma-Aldrich, Germany). Деионизированная вода (milliQ, Merck).

Приборы и оборудование. Масс-спектры высокого разрешения зарегистрированы на приборах Bruker micrOTOF (времяпролетный масс-анализатор), Bruker maXis QTOF (тандемный квадрупольный/времяпролетный масс-анализатор) и Bruker solariX XR (ИЦР масс-анализатор, сверхпроводящий магнит с напряженностью поля 15 Тл) (Германия), оборудованных источником ионизации электрораспылением (ИЭР). Диапазон сканирования m/z 50–3000 для лизоцима и m/z 100–1400 для хлорида палладия(II). Время регистрации масс-спектра на приборах microTOF, maXis и solariX XR составляет 45 с, в случае solariX XR количество сканов 32, количество точек 4M. Внешнюю калибровку шкалы масс для microTOF и maXis осуществляли с помощью низкоконцентрированного калибровочного раствора “Tuning mix” (Agilent Technologies), для solariX XR – с помощью раствора трифторацетата натрия (0.1 мг/мл в смеси ацетонитрил : вода в отношении 1 : 1). Ввод образцов осуществляли с помощью шприца Hamilton RN 1750 (Швейцария) на 500 мкл. Измерения проводились в режиме регистрации положительных ионов (+) для лизоцима (заземленная игла распыления, высоковольтный капилляр – 4500 В; разность потенциалов с защитным экраном спрея –500 В) и отрицательных ионов (–) для хлорида палладия(II) (заземленная игла распыления, высоковольтный капилляр +4000 В; разность потенциалов с защитным экраном спрея –500 В). Скорость потока при вводе контролировали шприцевым насосом (3 мкл/мин). Азот использовали в качестве газа-распылителя (1.0 бар) и газа осушителя (4.0 л/мин, 200°C). Для детектирования низкоконцентрированного образца лизоцима (1×10^{-8} M) на масс-спектрометре SolariX XR использовали накопление в ИЦР ячейке количество сканов 1024, количество точек 8M. Данные обрабатывали с использованием программного пакета Bruker Data Analysis 5.0.

Пробоподготовку образцов лизоцима осуществляли взятием навески (1 мг) с последующим растворением в смеси ацетонитрил : вода = 1 : 1 с добавлением 1 мкл муравьиной кислоты. Диапазон концентраций от 1×10^{-8} до 1×10^{-4} M. Масс-спектры хлорида палладия(II) готовили разбавлением навески 1 мг в 1 мл ацетонитрила с последующим разбавлением до 1×10^{-6} M.

Автоматический анализ масс-спектров. Все программное обеспечение было написано на языке Python с использованием библиотек для анализа данных NumPy и Pandas. Графики были построены с использованием библиотеки Matplotlib. Спектры конвертировались в mzXML-формат, а затем обрабатывались при помощи библиотеки pyOpenMS.

Определение количества формул для масс, меньших, чем 1000 Да, было проведено с использованием полного перебора возможных структур. Для больших масс количество возможных вариантов оценено с использованием метода Монте-Карло ввиду большого количества вычислений и невозможности полного перебора за разумное время. В расчетах по методу Монте-Карло случайно выбирались 5×10^9 формул, для которых проводилось сравнение с экспериментальной массой и соответствующее масштабирование.

Для соли металла было допущено наличие следующих элементов (компоненты самой соли, возможные лиганды и молекулы растворителя): палладий, хлор, водород, углерод, азот и кислород. В элементном составе лизоцима учитывались водород, углерод, кислород, азот и сера.

СПИСОК ЛИТЕРАТУРЫ

1. *Dass C.* Fundamentals of Contemporary Mass Spectrometry. New Jersey: Wiley; 2007. 608 p.
2. *Kanawati B., Schmitt-Kopplin P.* Fundamentals and Applications of Fourier Transform Mass Spectrometry. Amsterdam: Elsevier; 2019. 778 p.
3. *Лебедев А.Т.* Масс-спектрометрия в органической химии. Москва: Техносфера; 2015. 704 с.
4. *de Hoffmann E., Stroobant V.* Mass Spectrometry: Principles and Applications. 3nd ed. Chichester: Wiley; 2007. 502 p.

5. *Tsedilin A.M., Fakhrutdinov A.N., Eremin D.B., et al.* How Sensitive and Accurate are Routine NMR and MS Measurements? // Mendeleev Commun. 2015. V. 25. № 6. P. 454–456.

6. *Marshall A.G., Rodgers R.P.* Petroleomics: The Next Grand Challenge for Chemical Analysis // Acc. Chem. Res. 2004. V. 37. № 1. P. 53–59.

7. *Rodgers R.P., McKenna A.M.* Petroleum Analysis // Anal. Chem. 2011. V. 83. № 12. P. 4665–4687.

8. *Tan M., Le Q.V.* EfficientNet: Rethinking model scaling for convolutional neural networks // 36th Int. Conf. Mach. Learn. ICML 2019; 10–15 June 2019; Long Beach, California; P. 10691–10700.

9. *Schroff F., Kalenichenko D., Philbin J.* FaceNet: A unified embedding for face recognition and clustering // Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.; 07–12 June 2015; Boston. P. 815–823.

10. *Bojarski M., Del Testa D., Dworakowski D., et al.* End to End Learning for Self-Driving Cars // arXiv:1604.07316. 2016. P. 1–9.

11. *Lample G., Conneau A.* Cross-lingual Language Model Pretraining // arXiv:1901.07291. 2019.

12. *Behrmann J., Etmann C., Boskamp T., et al.* Deep learning for tumor classification in imaging mass spectrometry // Bioinformatics. 2018. V. 34. № 7. P. 1215–1223.

13. *Huang Y.-C., Chung H.-H., Dutkiewicz E.P., et al.* Predicting Breast Cancer by Paper Spray Ion Mobility Spectrometry Mass Spectrometry and Machine Learning // Anal. Chem. 2020. V. 92. № 2. P. 1653–1657.

14. *Palmer A.D., Bunch J., Styles I.B.* The Use of Random Projections for the Analysis of Mass Spectrometry Imaging Data // J. Am. Soc. Mass Spectrom. 2015. V. 26. № 2. P. 315–322.

15. *Marshall A.G., Blakney G.T., Chen T., et al.* Mass Resolution and Mass Accuracy: How Much Is Enough? // Mass Spectrom. 2013. V. 2 (Special Issue): S0009. P. 1–5.

16. *Saeys Y., Inza I., Larrañaga P.* A review of feature selection techniques in bioinformatics. // Bioinformatics. 2007. V. 23. № 19. P. 2507–2517.

17. *Granda J.M., Donina L., Dragone V., et al.* Controlling an organic synthesis robot with machine learning to search for new reactivity // Nature. 2018. V. 559. № 7714. P. 377–381.

18. *Ahneman D.T., Estrada J.G., Lin S., Dreher S.D., Doyle A.G.* Predicting reaction performance in C-N cross-coupling using machine learning // Science. 2018. V. 360. № 6385. P. 186–190.

19. *Šádek V., Schröder D., Tsierkezos N.G.* Clustering of palladium(II) chloride in acetonitrile solution investigated by electrospray mass spectrometry // Int. J. Mass Spectrom. 2011. V. 304. № 1. P. 9–14.

20. *Marshall A.G., Hendrickson C.L., Jackson G.S.* Fourier transform ion cyclotron resonance mass spectrometry: A primer // Mass Spectrom. Rev. 1998. V. 17. № 1. P. 1–35.

21. *Brenton A.G., Godfrey A.R.* Accurate mass measurement: Terminology and treatment of data // J. Am. Soc. Mass Spectrom. 2010. V. 21. № 11. P. 1821–1835.

22. *Wells G., Prest H., Russ IV C.W.* Why Use Signal-To-Noise As a Measure of MS Performance When It Is Often Meaningless? // Curr. Top. Mass Spectrom. 2011. P. 28–33.

COMPREHENSIVE MASS-SPECTROMETRIC MAPPING OF CHEMICAL COMPOUNDS FOR THE DEVELOPMENT OF ALGORITHMS FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

J. V. Burykina[a], D. A. Boiko[a,b], V. V. Ilyushenkova[a], D. B. Eremin[a], and Academician of the RAS V. P. Ananikov[a,b,#]

[a] *Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, Moscow, Russian Federation*

[b] *Lomonosov Moscow State University, Moscow, Russian Federation*

[#] *E-mail: val@ioc.ac.ru*

Herein, the influence of the accuracy of mass measurements on the number of possible structural compositions and the computation time of computer-aided interpretation of mass-spectrometric data is evaluated. The analysis was performed for two model objects in the range of small and medium masses using high, ultra-high and extreme-high resolution electrospray ionization mass-spectrometers. Based on that, we describe prospects of using machine learning in conjunction with mass-spectrometric data for inference of new data about reaction mechanisms and search for hidden relationships in the chemical space. It was also shown, that two types of relationships between the molecular formula and mass determination error may be observed, depending on the mass of the ion. Nonlinear curve was predicted for small molecules and linear relationship was predicted for large molecules.

Keywords: mass-spectrometry, FT-ICR-MS, ESI-MS, machine learning, artificial intelligence