

Методика идентификации веществ по положению характерных линий в спектрах комбинационного рассеяния света

© Р.А. Гылка, Д.Р. Анфимов, Я.Р. Чубаркина, П.П. Дёмкин, И.Л. Фуфурин

МГТУ им Н.Э. Баумана,
Москва, Россия

e-mail: roangy@mail.ru, diman_anfimov@mail.ru

Поступила в редакцию 25.12.2024 г.

В окончательной редакции 14.01.2025 г.

Принята к публикации 28.02.2025 г.

Разработан метод идентификации веществ по положению характерных линий в спектрах комбинационного рассеяния света. Метод основан на сравнении пиковых значений интенсивности в спектрах исследуемого и эталонного веществ. Работоспособность предложенной методики исследована с помощью матриц соответствия и ROC-анализа. Проведено сравнение разработанных алгоритмов с методом корреляции Пирсона. Исследование проведено на основе базы данных, состоящей из 16 порошкообразных веществ, с помощью дифракционного спектрометра Ventana-785L-Raman, оснащенного лазерным источником с длиной волны возбуждения 785 nm, пиковой мощностью лазера до 120 mW и нестабильностью мощности не более 1%.

Ключевые слова: спектроскопия комбинационного рассеяния света, спектральный анализ, идентификация, ROC-кривые.

DOI: 10.61011/OS.2025.03.60249.8-25

Введение

Для анализа веществ применяются методы химические, электрохимические, физические, хроматографические и спектроскопические [1,2]. Наиболее распространенными методами анализа твердых веществ, порошков и жидкостей во внелабораторных условиях являются методы молекулярной спектроскопии: спектроскопия диффузного рассеяния [3–5], спектроскопия комбинационного рассеяния света (КРС) [6,7], фотолюминесценция [8] и пр. Спектры диффузного рассеяния имеют невысокую селективность и, как правило, применяются преобразования Крамерса–Кронига [9] для преобразования спектров рассеяния в спектры поглощения.

Для проведения экспресс-анализа без отбора проб одним из методов, показывающих наибольшую вероятность правильного обнаружения веществ [10], является спектроскопия КРС. Метод позволяет проводить анализ веществ по заранее созданной спектральной базе данных с целью их определения или классификации по характерным спектральным признакам [6]. Метод спектроскопии комбинационного рассеяния дает возможность получить индивидуальный спектральный отпечаток, уникальный по отношению к рассматриваемой молекуле. Он представляет собой характерные линии спектров КРС. Спектроскопия комбинационного рассеяния высокочувствительна к небольшим различиям химического состава вещества [11], что особенно актуально в задачах контроля качества фармацевтических препаратов.

Применение методов, использующих базу данных полных спектров КРС, сопряжено с трудностями хранения данных большого объема, а также со сложностями

пополнения базы новыми спектрами. Последний фактор обусловлен не только труднодоступностью полных спектров КРС, но и с невозможностью извлечения данных из открытых библиотек (таких как, например, RRUFF [12], LENS [13]). Пополнение базы данных можно проводить самостоятельно, но это связано с большими трудозатратами.

Материалы и методы

На рис. 1 изображена схема экспериментальной установки с дифракционным спектрометром Ventana-785L-Raman, предназначенным для регистрации спектров КРС в диапазоне длин волн 800–940 nm. Спектрометр оснащён встроенным стабилизированным лазерным источником с длиной волны возбуждения 785 nm, пиковой мощностью лазера до 120 mW и нестабильностью мощности не более 1%.

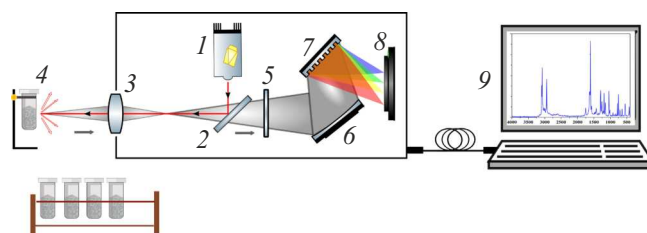


Рис. 1. Схема экспериментальной установки: 1 — лазерный источник, 2 — светоделительная пластинка, 3 — фокусирующий объектив, 4 — образец, 5 — режекторный фильтр, 6 — отражающее зеркало, 7 — дифракционная решетка, 8 — фотоприемное устройство, 9 — компьютер.

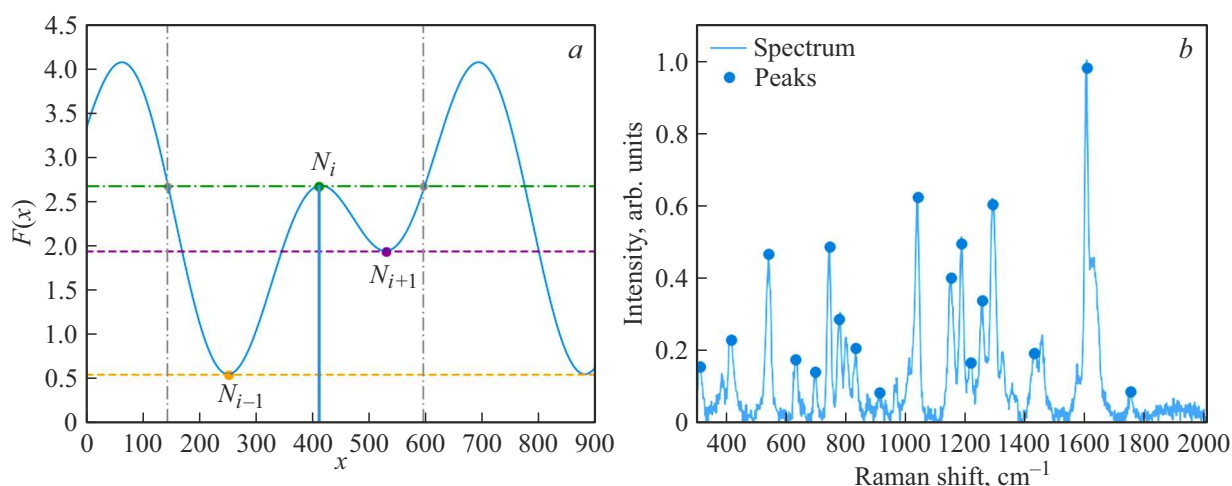


Рис. 2. Принцип поиска пиков (а) в функции $F(x)$ и (б) в спектре аспирина (SNR=16).

Как видно на рис. 1, излучение от лазерного источника 1 проходит светоделительную пластинку 2 и с помощью объектива 3 фокусируется на исследуемом образце 4. Образцы для удобства проведения исследования находятся в специальных пластиковых пробирках, которые прозрачны для возбуждающего излучения. Регистрируемый сигнал от образца 4 через фокусирующий объектив 3 поступает на вход режекторного фильтра 5, который отделяет стоксово рассеяние на входе. После отражения от зеркала 6 излучение падает на дифракционную решетку 7. Регистрируемый на фотоприемном устройстве 8 сигнал передается на персональный компьютер 9 для последующей обработки и анализа.

Для проведения сопоставительного анализа предварительно зарегистрирована база спектров КРС шестнадцати химических соединений: аналгин ($C_{13}H_{16}N_3NaO_4S$), аскорбиновая кислота ($C_6H_8O_6$), аспартам ($C_{14}H_{18}N_2O_5$), аспирин ($C_9H_8O_4$), борная кислота (H_3BO_3), глюконат кальция ($C_{12}H_{22}CaO_{14}$), хлорат калия ($KClO_3$), лимонная кислота ($C_6H_8O_7$), эритритол ($C_4H_{10}O_4$), фруктоза ($C_6H_{12}O_6$), глюкоза ($C_6H_{12}O_6$), мальтитол ($C_{12}H_{24}O_{11}$), парацетамол ($C_8H_9NO_2$), сода ($NaHCO_3$), крахмал ($C_6H_{10}O_5$), ксилитол ($C_5H_{12}O_5$). Одновременно для полной базы референтных спектров КРС тестовых веществ определен набор положений характерных пиков, по которым проводился сравнительный анализ методов идентификации. Число пиков у каждого спектра различно, поэтому данные о пиках для различных веществ отличаются.

Математические методы анализа спектров КРС

Перед проведением идентификации проводится предобработка измеренных спектров КРС. Методом наименьших квадратов определяется базовая линия, ко-

торая впоследствии вычитается из экспериментального спектра, после чего спектр КРС нормализуется на единицу. Метод определения базовой линии основан на сглаживании методом Уитткера [14]. Спектральная интенсивность, состоящая из n элементов, определяется вектором $y = \{y_1, y_2, \dots, y_n\}$, а сглаженный вектор $z = \{z_1, z_2, \dots, z_i\}$ — это модифицированный фон для данного спектра.

Опорным методом сравнения спектров КРС является метод определения их меры сходства. Одной из возможных мер схожести является статистический коэффициент корреляции Пирсона [15] (здесь и далее будем считать, что спектры представлены в виде числовых векторов конечной длины):

$$r = \frac{(\mathbf{x} - \bar{x})(\mathbf{y} - \bar{y})}{\|\mathbf{x} - \bar{x}\| \|\mathbf{y} - \bar{y}\|}, \quad (1)$$

где \mathbf{x} , \mathbf{y} — векторы сравниваемых спектров, \bar{x} — среднее арифметическое компонент вектора \mathbf{x} , $\|\mathbf{x}\|$ — евклидова норма вектора.

Для поиска пиков используется функция поиска. На рис. 2 представлен результат поиска пиков в функции вида $F(x) = \sin(0.02x) + \cos(0.01x)$ (рис. 2, а) и в спектре КРС аспирина (рис. 2, б). Горизонтальные линии уровня представлены для визуального сравнения значений функции в разных точках. Данная функция выполняет поиск пиков по простому условию: N_i есть пик, если $N_i > N_{i-1}$ и $N_i > N_{i+1}$ при i , для которых $N_i > \bar{x} + k\sigma$. Здесь \bar{x} и σ — среднее арифметическое значение и среднеквадратическое отклонение соответственно для участка спектра без пиков, k — отношение сигнал/шум (signal-to-noise ration, SNR) в измеренном спектре.

Для идентификации веществ по пикам в спектре КРС разработан алгоритм, который учитывает значения интенсивностей, соответствующие пикам эталонного спектра. Для сравнения спектра с известными данными

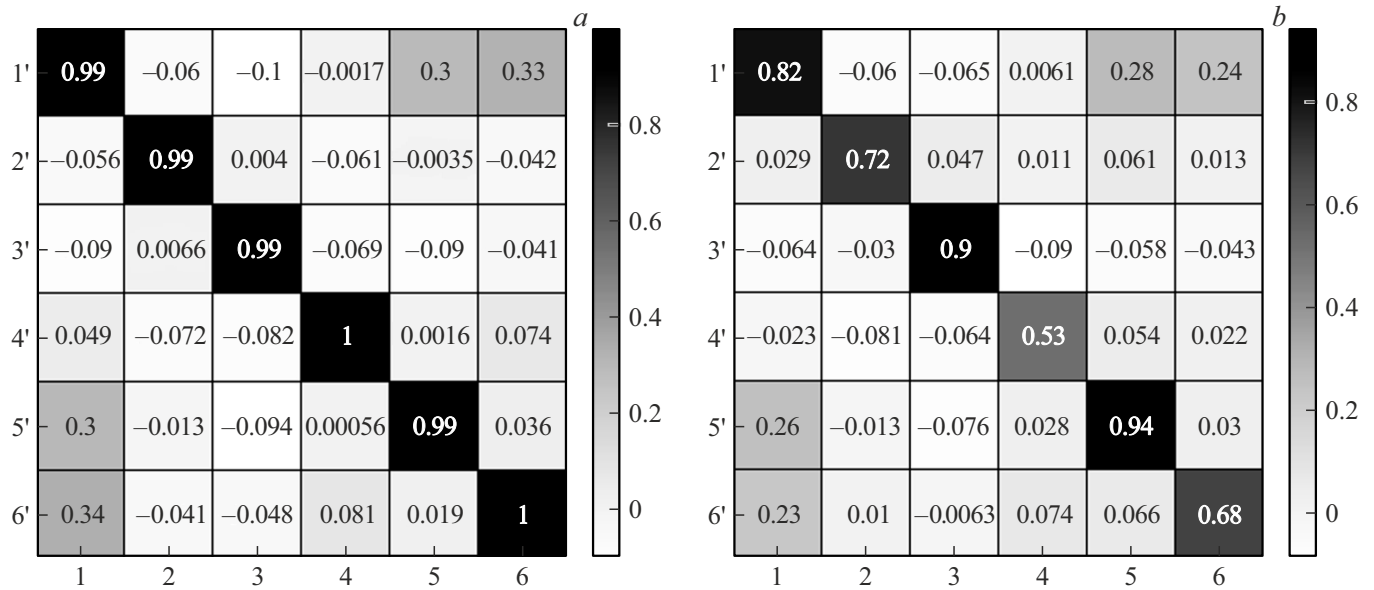


Рис. 3. Матрицы соответствия для опорного метода с (а) высоким и (б) низким SNR. 1 — аспирин, 2 — борная кислота, 3 — хлорат калия, 4 — глюкоза, 5 — парацетамол, 6 — сода.

рассчитывается невязка вида

$$r_p = \|\mathbf{x}_p - \mathbf{y}_p\|, \tag{2}$$

где \mathbf{x}_p и \mathbf{y}_p — векторы интенсивностей пиков эталонного и тестового спектров, положение пиков последнего совпадает с положением пиков эталонного спектра. Референтный набор значений представляет собой координаты характерных пиков при условии, что значения интенсивностей для других длин волн приняты равными нулю.

Метод сравнения заключается в расчете значений невязки для каждого из веществ в базе данных и определении наименьшего из этих значений. Вещество, для которого невязка r_p наименьшая, считается наиболее соответствующим исследуемому веществу и последнее, таким образом, считается идентифицированным в эксперименте. На практике удобно использовать не само значение невязки r_p , а выражение вида $1 - r_p$ (здесь $0 \leq r_p \leq 1$), откуда следует зависимость

$$r_p^* = 1 - r_p \in [0, 1], \tag{3}$$

т.е. чем ближе r_p^* к единице, тем выше схожесть исследуемого и эталонного веществ.

Для сравнения спектров КРС используется порог обнаружения, который выбирается для каждого класса веществ при оптимальных значениях чувствительности и специфичности при ROC(Receiver operating characteristic)-анализе.

Результаты

Для анализа эффективности идентификации веществ разрабатываемого алгоритма были построены матрицы

Показатели чувствительности (Sen), специфичности (Spe) и площади подкривой (AUC)

Вещество	Опорный алгоритм			Разработанный алгоритм		
	Sen	Spe	AUC	Sen	Spe	AUC
Аспирин	1	1	1	1	1	1
Борная кислота	1	1	0.996	1	1	1
Глюконат кальция	0.985	1	0.996	0.987	0.981	0.999
Хлорат калия	1	1	1	1	1	1
Глюкоза	1	1	1	1	1	1
Сода	1	1	1	1	0.942	1
Среднее значение	0.998	1	0.999	0.998	0.997	0.999

соответствия и ROC-кривые для опорного и разработанного методов [16]. На рис. 3 изображены матрицы соответствия для опорного метода сравнения с низким (менее 9) и высоким (более 9) отношением сигнал/шум SNR. Сравниваются эталонные спектры веществ без примесей с SNR более 1000 и тестовые спектры веществ с возможным незначительным содержанием примесей с SNR менее 1000. По вертикали представлены эталонные вещества, по горизонтали — тестовые. На рис. 4 представлена матрица для разработанного метода сравнения по пикам, где используются положения пиков спектра эталонного вещества. На рис. 5 представлены ROC-кривые для опорного (а) и разработанного (б) методов

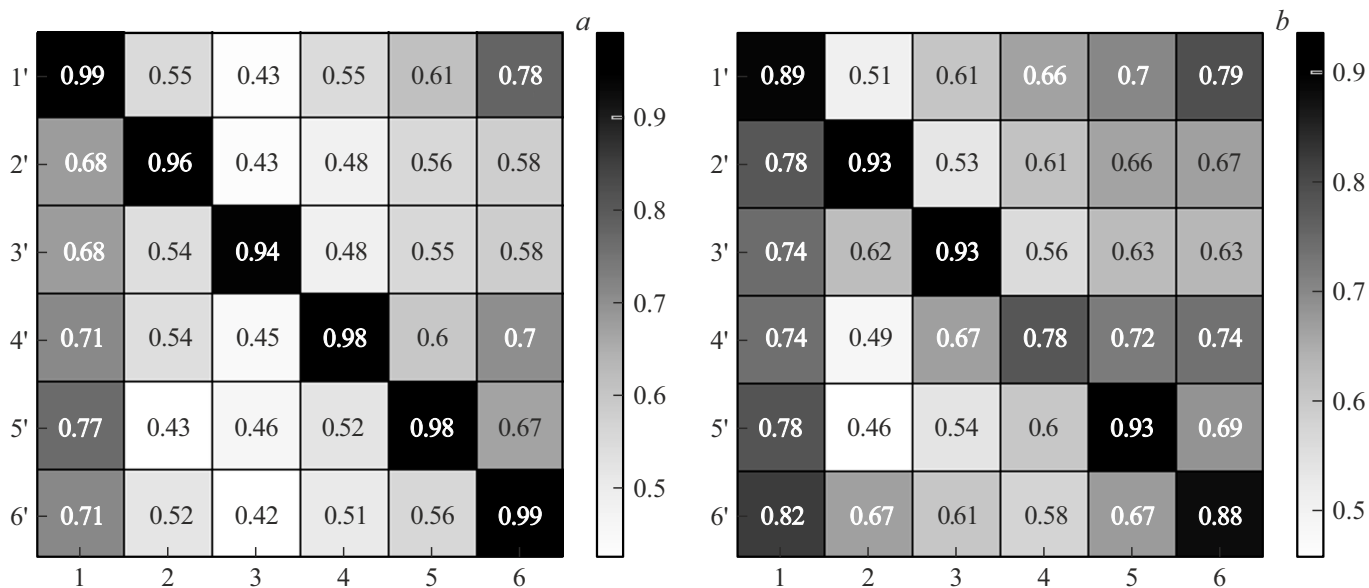


Рис. 4. Матрицы соответствия для разработанного метода с (а) высоким и (б) низким SNR. 1 — аспирин, 2 — борная кислота, 3 — хлорат калия, 4 — глюкоза, 5 — парацетамол, 6 — сода.

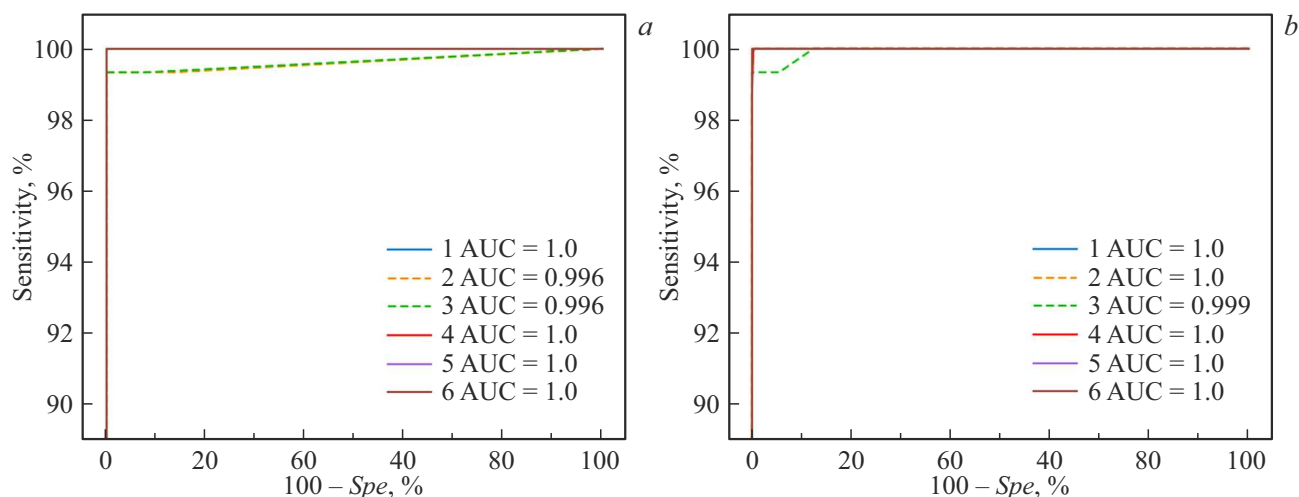


Рис. 5. ROC-кривые для (а) опорного и (б) разработанного методов сравнения. 1 — аспирин, 2 — борная кислота, 3 — хлорат калия, 4 — глюкоза, 5 — парацетамол, 6 — сода.

сравнения. ROC-кривая — график позволяющий оценить качество классификации, показателем которой является площадь под кривой (area under ROC-curve, AUC). По осям отложены показатели специфичность и чувствительность, которые зависят от значений порога обнаружения. Чувствительность (Sen) и специфичность (Spe) определяются следующим образом:

$$Sen = \frac{TP}{TP + FN}, \quad Spe = \frac{TN}{TN + FP}, \quad (4)$$

где TP — истинно положительный результат, FP — ложно положительный результат, TN — истинно отрицательный результат, FN — ложно отрицательный результат. В таблице представлены значения чувстви-

тельности, специфичности и площади под ROC-кривой для некоторых веществ.

Обсуждение полученных результатов

Разработанный метод достаточно хорошо идентифицирует вещества при высоком значении параметра SNR по сравнению с опорным методом. Диагональ соответствия прослеживается как при использовании опорного метода, так и при использовании разработанного. Диагональ соответствия имеет значения невязки векторов, расположенных на главной диагонали матрицы, которые соответствуют сравнению каждого эталонного спектра с соответствующим экспериментальным спектром веще-

ства. При низком SNR диагональ соответствия прослеживается лучше у разработанного метода, но диагональ размывается, что говорит о понижении качества идентификации веществ. Это может быть связано с высоким уровнем шума и появлением из-за этого ложного пика у вещества.

Параметр AUC разработанного метода близок к единице у большинства веществ, что говорит о достаточно высокой дискриминационной способности метода. Показатели чувствительности и специфичности у обоих методов высокие, что также говорит о хорошей способности разделения классов веществ.

Вектор одного спектра вещества содержит 1024 значений интенсивности и их координат. Вектор с пиками эталонного вещества содержит информацию только об интенсивностях характерных пиков и их координатах. База данных пиков веществ занимает в два раза меньше памяти, чем база данных полных спектров. Координатами являются длины волн или волновые числа. Скорость выполнения предложенного алгоритма сравнения приблизительно в три раза выше, чем скорость выполнения алгоритма опорного метода.

Заключение

Предложен метод идентификации веществ по характерным линиям спектров КРС. Метод основан на сравнении амплитудных значений выбранных пиков в спектрах комбинационного рассеяния исследуемого и эталонного веществ. Измерения спектров КРС выполнены с помощью дифракционного спектрометра Ventana-785L-Raman, оснащенного лазерным источником с длиной волны возбуждения 785 nm, пиковой мощностью лазера до 120 mW и нестабильностью мощности не более 1%. Разработанный метод апробирован на экспериментальной базе, состоящей из 16 тестовых веществ. Объемы данных, содержащие информацию о положении характерных пиков, в два раза меньше по сравнению с объемом полных спектров КРС. Скорость выполнения алгоритма разработанного метода в три раза выше скорости выполнения алгоритма опорного метода. Эффективность работы алгоритма идентификации для разработанного метода не уступает опорному методу, значение параметра AUC составляет 0.999.

Благодарности

Исследование выполнено за счет гранта „Фонда содействия развитию малых форм предприятий в научно-технической сфере“ в рамках конкурса „Студенческий стартап“ V очередь, договор № 3819ГССС15-L/99612.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

Список литературы

- [1] Ю. Бёккер. *Спектроскопия* (Техносфера, М., 2009).
- [2] А.Н. Морозов, С.И. Светличный. *Основы фурье-спектрорадиометрии* (Наука, М., 2014).
- [3] I.L. Fufurin, A.S. Tabalina, A.N. Morozov, Ig.S. Golyak, S.I. Svetlichnyi, D.R. Anfimov, I. Kochikov. *SPIE Opt. Eng.*, **59** (6), 1 (2020).
- [4] A. Mendizabal, P.G. Loges. *SPIE*, **XXIII**, 34 (2023).
- [5] Д.Р. Анфимов, Иг.С. Голяк, О.А. Небритова, И.Л. Фуфурин. *Химическая физика*, **41** (10), 10 (2022). DOI: 10.31857/S0207401X22100028
- [6] И.Б. Винтайкин, Ил.С. Голяк, Иг.С. Голяк, А.А. Есаков, А.Н. Морозов, С.Е. Табалин. *Химическая физика*, **39** (10), 20 (2020). DOI: 10.31857/S0207401X20100118
- [7] D.W. Shipp, F. Sinjab, I. Notingham. *Adv. Opt. Photon.*, **9** (2), 315 (2017).
- [8] Н.С. Васильев, Иг.С. Голяк, Ил.С. Голяк, А.А. Есаков, А.Н. Морозов, С.Е. Табалин. *ПТЭ*, **1** (1), 181 (2015). DOI: 10.7868/S0032816215010231
- [9] C.W. Peterson, B.W. Knight. *JOSA*, **63** (10), 1238 (1973).
- [10] A.N. Morozov, I.V. Kochikov, A.V. Novgorodskaya, A. Sologua, I.L. Fufurin. *Comp. Opt.*, **39** (4), 614 (2015).
- [11] Р.А. Гылка, А.В. Грицаева. *Политехнический молодежный журнал*, **03** (80), 1 (2023). DOI: 10.18698/2541-8009-2023-03-878.html
- [12] B. Lafuente, R.T. Downs, H. Yang, N. Stone. *Highlights in mineralogical crystallography*, **1**, 1 (2015).
- [13] LENS [Электронный ресурс]. URL: <https://lens.unifi.it/> (дата обращения: 11 ноября 2024 г.)
- [14] S. Oller-Moreno, A. Pardo, J.M. Jimenez-Soto, J. Samitier. *IEEE 11th Int. Multi-Conf. SSD14*, **1**, 1 (2014). DOI: 10.1109/SSD.2014.6808837
- [15] K. Pearson. *Proc. Roy. Soc. London*, **58**, 240 (1895).
- [16] A. Wysoczanski, E. Voigtman. *Spectrochim. Acta B*, **100**, 70 (2014).