23

# The method of identification of substances by the position of individual characteristic lines in the raman spectra

© R.A. Gylka, D.R. Anfimov, Ya.R. Chubarkina, P.P. Demkin, I.L. Fufurin

Bauman Moscow State Technical University,
Moscow, Russia

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

*Received December 25, 2024*
*Revised January 14, 2025*
*Accepted February 28, 2025*

The task of substance identification is an actual issue in many areas of research related to quality control and life safety. Rapid identification and analysis of substances by sample-free methods can be used to solve the problems of drug distribution and prevention of terrorist acts. Raman spectroscopy is widely used for these purposes. Substances can be analyzed using mathematical methods for statistical comparison of experimental and standard spectra. For the identification of chemical compounds and the possibility of quantitative analysis of substances, generally, whole Raman spectra are used, that provides the greatest informativeness of the initial data. However, in several cases it is not possible to register the full Raman spectrum, and it is necessary to develop mathematical methods of analysis by the position of peaks of characteristic lines of the Raman spectrum. The method is based on the comparison of peak intensity values in the spectra of the investigated and reference substances. The performance of the proposed methods is investigated by means of correspondence matrices and ROC-analysis. The developed algorithms are compared with the Pearson correlation method. The study was carried out on the basis of a database consisting of 16 powdery substances using a Ventana-785L-Raman diffraction spectrometer equipped with a laser source with an excitation wavelength of 785 nm, maximum laser power up to 120 mW and power instability not more than 1%.

**Keywords:** Raman spectroscopy, spectral analysis, identification, ROC curves.

DOI: 10.61011/EOS.2025.03.61166.8-25

## Introduction

Chemical, electrochemical, physical, chromatographic, and spectroscopic [1,2] methods are used to analyze substances. Molecular spectroscopy methods are the most common methods for the analysis of solids, powders, and liquids in nonlaboratory conditions: diffuse scattering spectroscopy [3–5], Raman spectroscopy [6,7], photoluminescence [8], etc. Diffuse reflectance spectra have low selectivity and Kramers–Kronig relations [9] are generally applied to convert reflectance spectra to absorption spectra.

Raman spectroscopy is one of the methods of rapid analysis without sampling that exhibits the highest probability of correctly detecting substances [10]. The method allows the analysis of substances from a predetermined spectral database in order to identify or classify them by characteristic spectral features [6]. The Raman spectroscopy technique provides the opportunity to obtain an individual spectral fingerprint unique to the molecule in question. It represents the characteristic lines of the Raman spectra. Raman spectroscopy is highly sensitive to small differences in the chemical composition of the substance [11], which is especially relevant in the tasks of quality control of pharmaceuticals.

The application of methods using a database of complete Raman spectra is associated with difficulties in storing large amounts of data, as well as with difficulties in replenishing the database with new spectra. The latter factor is attributable not only to the inaccessibility of the complete Raman spectra, but also to the impossibility of extracting data from open libraries (such as, for example, RRUFF [12], LENS [13]). The database can be populated independently, but it is labor intensive.

## Materials and methods

Fig. 1 shows a schematic of the experimental setup with a Ventana-785L-Raman diffraction spectrometer designed for recording of Raman spectra in the wavelength range of 800–940 nm. The spectrometer has a built-in stabilized laser source with an excitation wavelength of 785 nm, peak laser power up to 120 mW and power instability of no more than 1%.

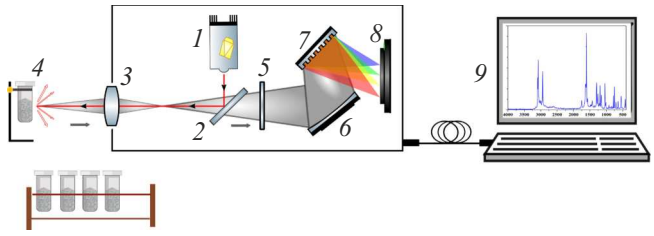


**Figure 1.** Diagram of the experimental setup:*1* — laser source, *2* — light-splitting plate, *3* — focusing lens, *4* — sample, *5* — rejector filter, *6* — reflecting mirror, *7* — diffraction grating, *8* — photodetector, *9* — computer.

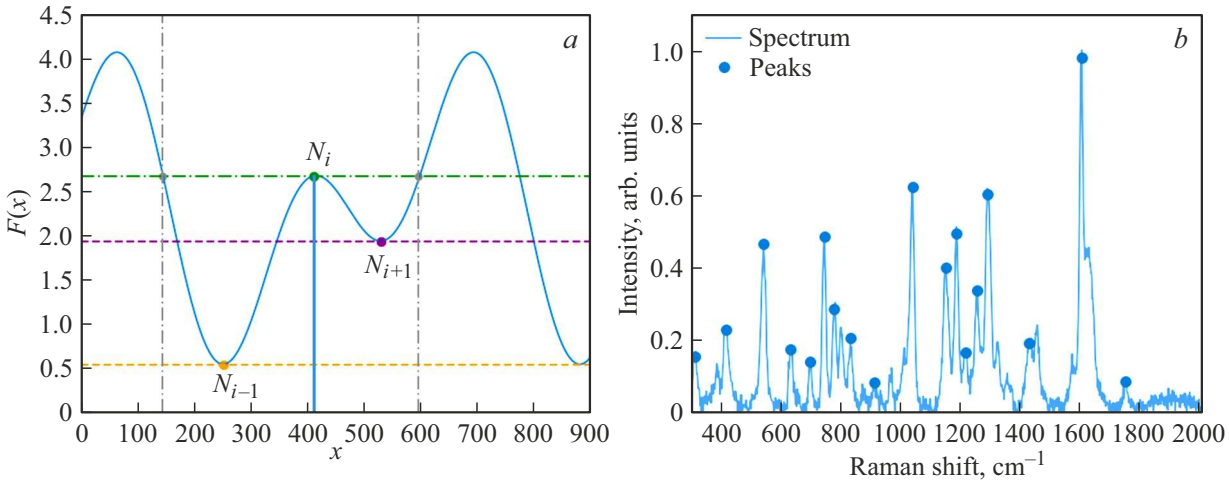


**Figure 2.** Principle of searching peaks ($a$) in the function $F(x)$ and ($b$) in the aspirin spectrum (SNR=16).

As shown in Fig. 1, the radiation from the laser source *1* passes the beam-splitting plate *2* and is focused on the studied sample *4* with the lens *3*. Samples for the convenience of the study are in special plastic tubes, which are transparent to excitation radiation. The recorded signal from the sample *4* is fed through a focusing lens *3* to the inlet of the band filter *5*, which separates the Stokes scattering at the inlet. The radiation falls on the diffraction grating *7* after it is reflected from the mirror *6*. The signal recorded on the photodetector *8* is transmitted to a personal computer *9* for further processing and analysis.

A database of Raman spectra of sixteen chemical compounds was pre-registered for comparative analysis: metamizole ($C_{13}H_{16}N_3NaO_4S$), ascorbic acid ($C_6H_8O_6$), aspartame ($C_{14}H_{18}N_2O_5$), aspirin ($C_9H_8O_4$), boric acid ($H_3BO_3$), calcium gluconate ($C_{12}H_{22}CaO_{14}$), potassium chlorate ($KClO_3$), citric acid ($C_6H_8O_7$), erythritol ($C_4H_{10}O_4$), fructose ($C_6H_{12}O_6$), glucose ($C_6H_{12}O_6$), maltitol ($C_{12}H_{24}O_{11}$), paracetamol ($C_8H_9NO_2$), Sodium bicarbonate ($NaHCO_3$), starch ($C_6H_{10}O_5$), xylitol ($C_5H_{12}O_5$). At the same time, a set of positions of characteristic peaks was determined for the complete base of reference Raman spectra of test substances which were used for a comparative analysis of identification methods. The number of peaks in each spectrum is different, so the peak data for different substances are different.

## Mathematical methods of analysis of Raman spectra

The measured Raman spectra are preprocessed prior to identification. A baseline is determined by the least-squares method, which is subsequently subtracted from the experimental spectrum, after which the Raman spectrum is normalized to unity. The method for determining the baseline is based on the Whittaker smoothing method [14]. The spectral intensity consisting of $n$ elements is defined by the vector $y = \{y_1, y_2, \ldots, y_n\}$, and the smoothed vector $z = \{z_1, z_2, \ldots, z_i\}$ is the modified background for this spectrum.

The reference method for comparing Raman spectra is the method for determining their similarity measure. One possible measure of similarity is Pearson's statistical correlation coefficient [15] (hereafter, we will assume that the spectra are represented as numerical vectors of finite length):

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

where $\mathbf{x}, \mathbf{y}$ are vectors of the compared spectra, $\bar{x}$ is the arithmetic mean of the vector components $\mathbf{x}$, $\|x\|$ is the Euclidean norm of the vector.

The search function is used to find peaks. Fig. 2 shows the result of peak search in the function of the form $F(x) = \sin(0.02x) + \cos(0.01x)$ (Figure 2, $a$) and in the Raman spectrum of aspirin (Figure 2, $b$). Horizontal level lines are presented for visual comparison of function values at different points. This function searches for peaks based on a simple condition: $N_i$ is a peak if $N_i > N_{i-1}$ and $N_i > N_{i+1}$ with $i$ for which $N_i > \bar{x} + k\sigma$. Here $\bar{x}$ and $\sigma$ are the arithmetic mean and standard deviation respectively for the peak-free portion of the spectrum, $k$ is the signal-to-noise ration (SNR) in the measured spectrum.

An algorithm has been developed to identify substances based on peaks in the Raman spectrum, which takes into account the intensity values corresponding to the peaks of the reference spectrum. For spectral matching with known data, a discrepancy measure of the form is computed

$$r_p = \|\mathbf{x_p} - \mathbf{y_p}\|, \tag{2}$$

where $\mathbf{x_p}$ and $\mathbf{y_p}$ are the intensity vectors of the peaks of the reference and test spectra, the position of the peaks of the latter coincides with the position of the peaks of the reference spectrum. The reference set of values represents the coordinates of characteristic peaks, provided that the

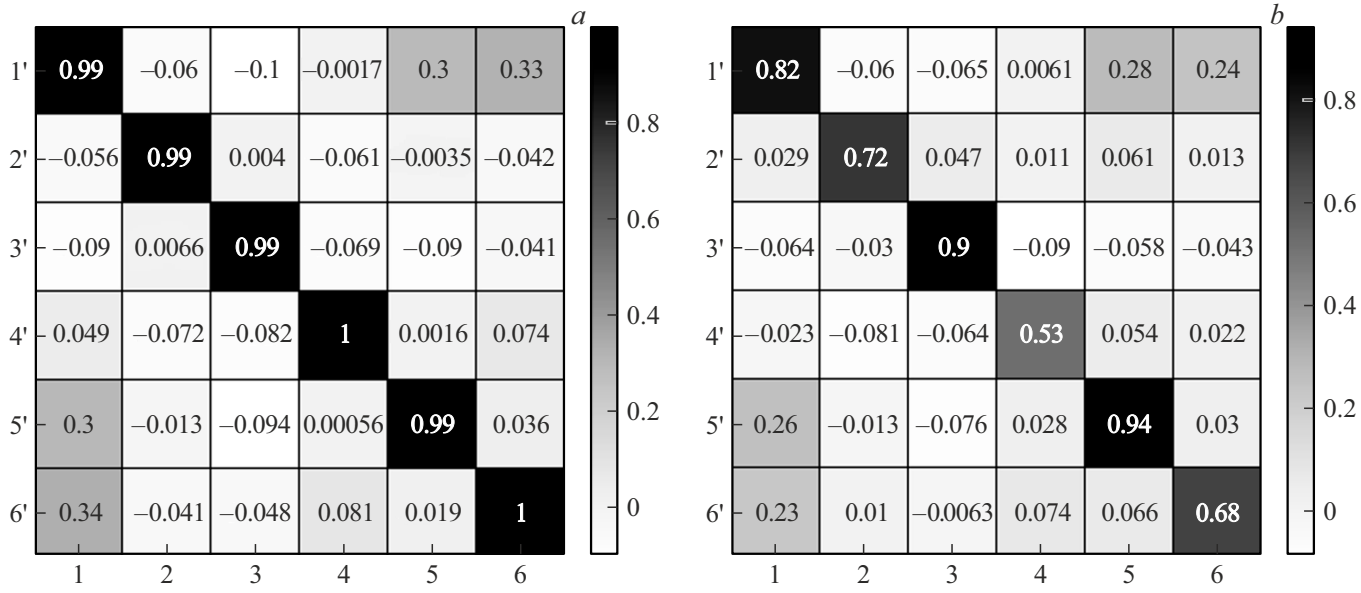


**Figure 3.** Correspondence matrices for the reference method with ($a$) high and ($b$) low SNR. *1* — aspirin, *2* — boric acid, *3* — potassium chlorate, *4* — glucose, *5* — paracetamol, *6* — soda.

intensity values for other wavelengths are assumed to be zero.

The comparison method consists of calculating the measure values for each of the substances in the database and determining the lowest of these values. The substance for which the discrepancy $r_p$ is the smallest is considered to be the most consistent with the substance under investigation and the latter is thus considered to be identified in the experiment. In practice, it is convenient to use not the discrepancy measure value $r_p$ itself, but an expression of the form $1-r_p$ (here $0 \le r_p \le 1$), whence follows the dependence

$$r_p^* = 1 - r_p \in [0, 1], \qquad (3)$$

i.e., the closer $r_p^*$ is to unity, the higher the similarity of the investigated and reference substances.

A detection threshold is used to compare the Raman spectra, which is selected for each class of substances at optimal values of sensitivity and specificity in ROC (Receiver operating characteristic) analysis.

## Results

The identification efficiency of the developed algorithm was evaluated through correspondence matrices and ROC curves, with comparative analysis against the reference method [16]. Fig. 3 shows the correspondence matrices for the reference comparison method with low (less than 9) and high (more than 9) signal-to-noise ratio (SNR). Reference spectra of substances without impurities with SNR greater than 1000 and test spectra of substances with possible minor impurities with SNR less than 1000 are compared. The vertical line represents reference substances, the horizontal line represents test substances. Fig. 4 shows the matrix for the developed peak-to-peak comparison method, where the peak positions of the reference substance spectrum are used. Fig. 5 shows the ROC curves for the reference ($a$) and developed ($b$) comparison methods. The ROC-curve is a graph that allows evaluating the quality of classification, which is measured by the area under ROC-curve (AUC). The specificity and sensitivity values, which depend on the detection threshold values, are plotted along the axes. Sensitivity ($Sen$) and specificity ($Spe$) are determined as follows:

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

where $TP$ is a true positive result, $FP$ is a false positive result, $TN$ is a true negative result, $FN$ is a false negative

Parameters of sensitivity ($Sen$), specificity ($Spe$) and area under the curve (AUC)

| Substance | Reference algorithm | | | Developed algorithm | | |
|---|---|---|---|---|---|---|
| | Sen | Spe | AUC | Sen | Spe | AUC |
| Aspirin | 1 | 1 | 1 | 1 | 1 | 1 |
| Boric acid | 1 | 1 | 0.996 | 1 | 1 | 1 |
| Gluconate calcium | 0.985 | 1 | 0.996 | 0.987 | 0.981 | 0.999 |
| Potassium chlorate | 1 | 1 | 1 | 1 | 1 | 1 |
| Glucose | 1 | 1 | 1 | 1 | 1 | 1 |
| Soda | 1 | 1 | 1 | 1 | 0.942 | 1 |
| Mean value | 0.998 | 1 | 0.999 | 0.998 | 0.997 | 0.999 |

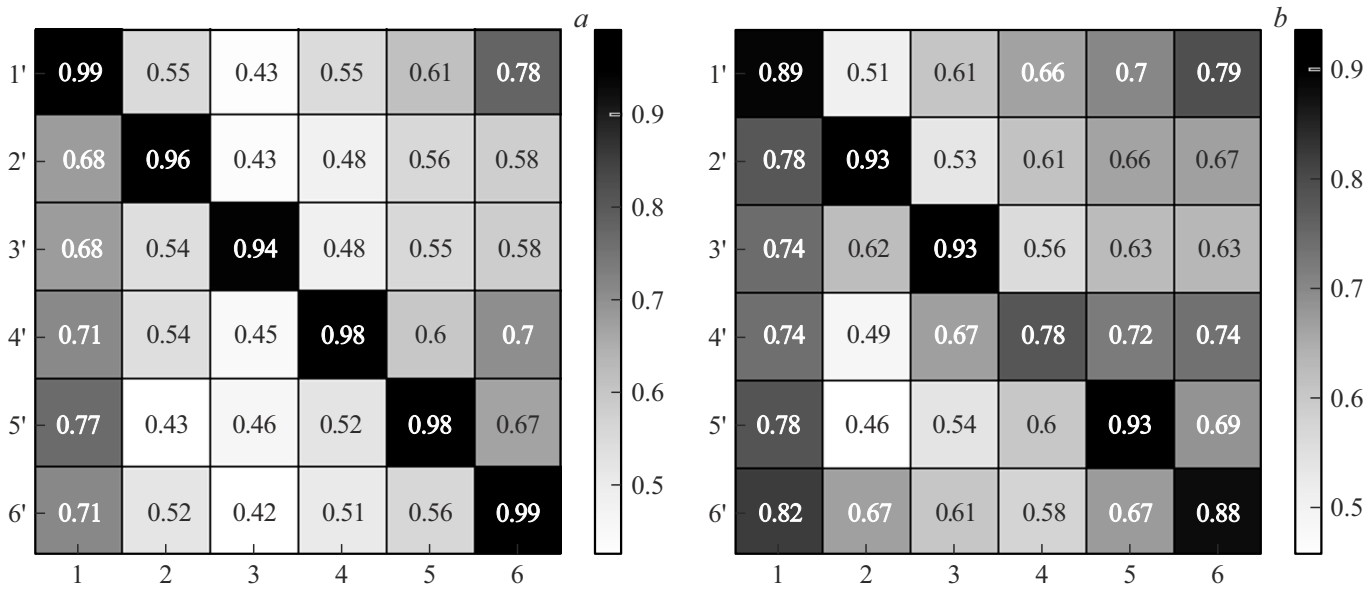


**Figure 4.** Correspondence matrices for the developed method with (*a*) high and (*b*) low SNR. *1* — aspirin, *2* — boric acid, *3* — potassium chlorate, *4* — glucose, *5* — paracetamol, *6* — soda.

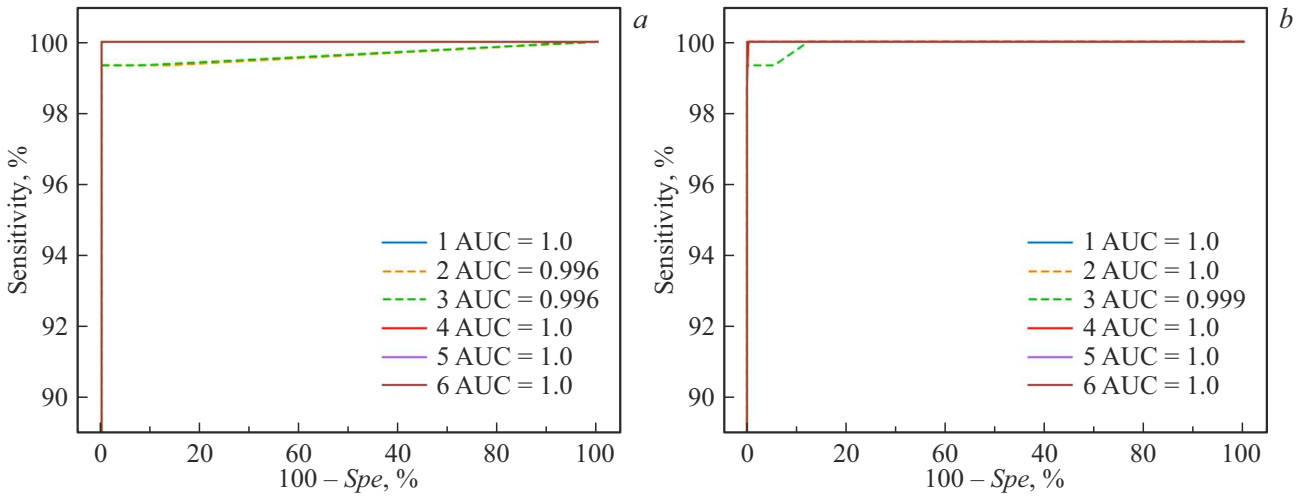


**Figure 5.** ROC curves for (*a*) reference and (*b*) developed comparison methods. *1* — aspirin, *2* — boric acid, *3* — potassium chlorate, *4* — glucose, *5* — paracetamol, *6* — soda.

result. The table shows the sensitivity, specificity, and area under the ROC curve for some substances.

## Discussion

The developed method identifies substances quite well with high SNR parameter value compared to the reference method. The diagonal of correspondence can be traced for both the reference method and the developed method. The correspondence diagonal contains the residual errors of vectors located on the main matrix diagonal, which correspond to the comparison of each reference spectrum with its matching experimental spectrum of the substance. The diagonal correspondence is better traced for the developed method in case of a low SNR, but the diagonal is blurred, indicating a lower quality of substance identification. This may be due to the high noise level and the appearance of a false peak in the substance because of this.

The AUC parameter of the developed method is close to unity for the majority of substances, which indicates a sufficiently high discriminatory ability of the method. The sensitivity and specificity values of both methods are high, which also indicates a good ability to separate classes of substances.

The vector of one spectrum of a substance contains 1024 intensity values and their coordinates. The vector with peaks of the reference substance contains information only about intensities of characteristic peaks and their coordinates. The

database of substance peaks occupies half as much memory as the database of full spectra. Coordinates are wavelengths or wave numbers. The execution speed of the proposed comparison algorithm is approximately three times faster than that of the reference method algorithm.

## Conclusion

A method for identification of substances by characteristic lines of Raman spectra is proposed. The method is based on comparison of amplitude values of selected peaks in Raman spectra of the studied and reference substances. The Raman spectra were measured using Ventana-785L-Raman diffraction spectrometer with a laser source with an excitation wavelength of 785 nm, peak laser power up to 120 mW, and power instability of no more than 1%. The developed method was tested using 16 test substances. The datasets containing information on the position of characteristic peaks are two fold smaller compared to the dataset of full Raman spectra. The execution speed of the developed method algorithm is three times faster than the execution speed of the reference method algorithm. The efficiency of the identification algorithm for the developed method is not inferior to the reference method, the value of the AUC parameter is 0.999.

### Acknowledgments

The study was supported by the grant of „Foundation for Assistance to Small Innovative Enterprises in the Scientific and Technical Sphere“ within the framework of the competition „Student Startup“ stage V, contract № 3819GSSS15-L/99612.

### Conflict of interest

The authors declare that there is no conflict of interest.

## References

[1] Y. Bekker. *Spectroskpiya* (Technosphera, M., 2009) (in Russian).
[2] A.N. Morozov, S.I. Svetlichnyi. *Osnovy fur'e-spektroradiometrii* (Nauka, M., 2014) (in Russian).
[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] D.R. Anfimov, Ig.S. Golyak, O.A. Nebritova, I.L. Fufurin. Khimicheskaya fizika, **41** (10), 10 (2022) (in Russian). DOI: 10.31857/S0207401X22100028
[6] I.B. Vintajkin, Il.S. Golyak, Ig.S. Golyak, A.A. Esakov, A.N. Morozov, S.E. Tabalin. Khimicheskaya fizika, **39** (10), 20 (2020) (in Russian). DOI: 10.31857/S0207401X20100118
[7] D.W. Shipp, F. Sinjab, I. Notingher. Adv. Opt. Photon., **9** (2), 315 (2017).
[8] N.S. Vasiliev, I.S. Golyak, Il.S. Golyak, A.A. Esakov, A.N. Morozov, S.E. Tabalin. PTE, **1** (1), 181 (2015) (in Russian). 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] R.A. Gylka, A.V. Gritsayeva. Politekhnicheskij molodezhnyj zhurnal, **03** (80), 1 (2023) (in Russian). 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 [Electronic source]. URL: https://lens.unifi.it/ (date of access: November 11, 2024)
[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).

*Translated by A.Akhtyamov*