How to Cite: Fomin, V.N., Ainabayev, A.A., Aldabergenova, S.K., Kaikenov, D.A., & Turovets, M.A. (2026). Classification Analysis of Bituminous Coals Using a Combination of GC-MS and Chemometric Methods. *Eurasian Journal of Chemistry.* https://doi.org/10.31489/2959-0663/2-26-6

Article

Received: 15 January 2026 | Revised: 30 April 2026 | Accepted: 5 May 2026 | Published online: 6 May 2026

UDC 543.06:519.23

https://doi.org/10.31489/2959-0663/2-26-6

Vitaliy N. Fomin, Assanali A. Ainabayev, Saule K. Aldabergenova, Dauletkhan A. Kaikenov, Milana A. Turovets*

Karaganda National Research University named after Academician Ye.A. Buketov, Karaganda, Kazakhstan
*(*Corresponding author's e-mail: turovec26.07@mail.ru)*

Classification Analysis of Bituminous Coals Using a Combination of GC-MS and Chemometric Methods

A comprehensive approach to the analysis of organic matter in bituminous coals from Kazakhstani deposits was developed, based on a combination of liquid extraction, gas chromatography with mass spectrometric detection (GC-MS), and chemometric data processing. A triple extraction system of dichloromethane–chloroform–tetrachloromethane (1:1:1) was proposed for sample preparation, providing representative extraction of aliphatic, aromatic and heteroatomic components without additional extract concentration. Chromatographic profiles of 120 coal extracts from four sources in Central and Northern Kazakhstan were analyzed. Optimization of the chromatographic separation conditions was carried out using probabilistic-deterministic design of experiment, which made it possible to establish robust relationships between the instrument parameters (column heating rate, carrier gas pressure) and the chromatogram characteristics. Chemometric data processing using principal component analysis (PCA), hierarchical cluster analysis (HCA), and k-means method revealed reproducible grouping of samples based on chromatographic profile similarity. Compact clusters corresponding to conventionally designated coal sources were formed in the principal component space, which is confirmed by the clustering results. The obtained results demonstrate the applicability of GC-MS and the chemometric approach for coal classification analysis and provide the foundation for the creation of a database of chromatographic fingerprints of the organic phase of coals.

Keywords: chemometrics, gas chromatography, PCA, hierarchical cluster analysis, k-means, coal classification, PDDoE, Kazakhstani deposits

Introduction

Since its inception, gas chromatography has become the primary method for studying volatile organic compounds in coal and its degradation products [1–3]. The wide variety of components in coal extracts and pyrolysis products makes traditional chromatogram interpretation an extremely labor-intensive process, which has driven the active implementation of chemometric methods and machine learning for GC-MS data processing.

In the 1990s, gas chromatographic and combined methods became the primary tools for analyzing pyrolysis products and coal extracts [3–5]. By the late 1990s, three main approaches to coal classification analysis had emerged: extraction analysis of soluble compounds, pyrolysis gas chromatography for thermal degradation products, and direct comparison of chromatograms from different coals. At that time, an analytical approach was commonly used where each peak was considered separately and processing was performed manually.

In the 2000s, gas chromatography evolved from merely a component identification method into a universal tool for molecular analysis of coals and their processing products [6–10]. A study [6] described an elemental gas chromatographic analysis methodology that increased result reproducibility for solid fuels several-fold. The first application of two-dimensional GC×GC-MS to slow pyrolysis tars from lignite showed that multidimensional gas chromatography could separate the most complex product matrices by classes — aromatic, aliphatic, and oxygen-containing [7]. Systematization of the entire spectrum of chromatographic approaches, including GC-MS, LC-MS, and TLC-MS, demonstrated that GC-MS remains the optimal method for the analysis of volatile and semi-volatile organic compounds [8, 9].

During 2010-2020, GC×GC-MS methods actively developed, enabling analysis of thousands of compounds in a single run [10–13]. One research direction in coal science involves studying the solubility of coal organic matter and the fractional composition of extracts using various solvents [12, 14–19]. CS2 and tetra-

© 2026 The Authors. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

hydrofuran extract aromatic and heterocyclic compounds effectively; toluene and pyridine extract alkanes and fatty alcohol fractions [12, 15]. *n*-Hexane predominantly isolates alkane and aromatic fractions, while methanol extracts oxygen-containing compounds. Thus, studies of coal extracts with various solvents have shown that gas chromatography and GC-MS effectively separate dozens and hundreds of volatile and semi-volatile organic compounds in complex matrices. However, with increasing data volumes, the need arose for processing using intelligent tools such as chemometrics and machine learning [20–23].

The transition from classical data processing methods to handling large datasets became a necessity. In mixtures containing numerous components, visual interpretation becomes uninformative [24–27]. The use of chemometric methods such as principal component analysis (PCA), hierarchical cluster analysis (HCA), partial least squares (PLS), and linear discriminant analysis (LDA) enabled a shift from disparate chromatograms to quantitative assessment of relationships between samples and compound classes.

PCA allows combining interrelated variables (peak areas of various compound classes) into several principal components, revealing main trends in complex data [14, 25]. Analysis of extracts from three brown coals using GC-MS and GC×GC-Q-TOF-MS showed that PCA allows of characterization of over 85 % of data variance [28].

Hierarchical clustering methods proved particularly effective in the processing of GC×GC-MS datasets [29, 30]. HCA paired with PCA were used to classify 190 compounds in lignite and subbituminous coal extracts [18]. It was found that extracts from one coal type form stable clusters regardless of solvent, confirming internal consistency of molecular composition.

Partial least squares (PLS) methods found application in quantitative analysis of compound class content based on peak intensities [31]. When applying PLS-R to determine aromatic compounds in coal extracts based on GC-MS signals, the coefficient of determination reached 0.98 [32].

Machine learning (ML) algorithms and neural networks (NN) have recently been actively applied for automatic classification of coals and their processing products [33–35]. Using a combination of HS-GC-IMS (gas chromatography with ion mobility spectrometry) and Random Forest and Support Vector Machine algorithms for identifying coal geographical origin, classification accuracy exceeded 99 %, demonstrating ML advantages [33].

Use of chemometric data processing methods can be more effective with careful physical separation of variables intended for future use as predictors. An effective method for tuning physicochemical analysis instruments involves mathematical experimental design. Probabilistic-deterministic design of experiment was previously successfully applied to optimize gas chromatograph settings for analyzing coal tar hydrogenation products [36], similar in volatile substance composition to extracts. The combination of instrument setting optimization using probabilistic-deterministic experimental design with subsequent chemometric analysis of results has been successfully applied in several LIBS studies, including [37–40].

Thus, a combination of existing chemometric methods opens new possibilities for analyzing extracts and pyrolysis products: accelerated analysis, reduced subjectivity, improved reproducibility, and creation of "coal fingerprint" databases. The combination of GC-MS analysis with chemometric processing significantly enhances capabilities, allowing identification of key variables among hundreds of peaks, improving qualitative analysis, and enabling sample classification by rank, geochemical group, and technological application.

It should also be noted that the literature lacks examples of applying such comprehensive methods to coals from Kazakhstan deposits. This opens directions for relevant research contributing to understanding of geochemistry and technologically important properties of domestic coals. Thus, the main goal of this research is a development of an efficient method for classification analysis of coal with combination of GS-MS and chemometrics.

Experimental

Coal samples were purchased from four different commercial sources rather than collected directly from deposits. For convenience, the samples were labeled based on vendor-provided origin information (RA — "Rapid", KA — "Karazhyra", SH — "Shubarkol", EK — "Ekibastuz"). These labels should be considered as nominal identifiers, and not as the confirmed deposit names, since this information could not be independently verified.

Two-stage grinding was performed for material homogenization. At the first stage, a jaw crusher was used, providing coarse grinding and simultaneous mixing of portions weighing approximately 1 kg. Thirty randomly selected coal samples from each source were used. This approach reduces heterogeneity and eliminates systematic differences between samples. After the coarse crushing, the material was quartered and ad-

ditionally ground in a mortar mill in 20-gram batches for 20 minutes until a fine powder with particle size of approximately 20–50 μm was obtained. The resulting powder was thoroughly mixed and dried at room temperature to constant weight.

To clarify the provenance of the coal samples, ash content was determined in accordance with GOST 11022-95, and ash composition by major components was analyzed by LIBS on sodium tetraborate glass fusion discs, following the procedure described in the literature [37]. Ash content measurements were performed three times. Confidence intervals were calculated using the Student's t-coefficient for two degrees of freedom at a 95 % confidence level. Elemental composition was determined on a composite sample, with the confidence interval derived from the calibration curve.

A triple mixture of dichloromethane, chloroform, and tetrachloromethane (all solvents purity is reagent grade, "Komponent-Reaktiv" manufacturer, Russia) in a 1:1:1 molar ratio was selected for extraction. This combination integrates solvent capacity toward aliphatic, aromatic, and heteroatomic components of coal organic matter, ensuring representative extract recovery. The mixture was prepared immediately before use. Extraction was performed by adding 5.00 mL of solvent to 1.00 g of coal and mixing on an orbital thermostated shaker at 120 rpm for 60 minutes at 20 °C. The resulting solution was filtered under atmospheric pressure through paper filter, and the filtrate was used directly for GC-MS.

Probabilistic-deterministic design of experiment was applied for optimization of chromatographic separation conditions based on the previous work with the similar samples [36]. The analysis was performed using the following instrumental parameters: stationary phase — Rxi-5ms capillary column with a length of 30 m, an inner diameter of 0.25 mm, and absorbent thickness of 0.25 μm; injector temperature — 250 °C; temperature program — 60–250 °C. Helium was used as a carrier gas, injection volume — 0.2 μl in split mode (1:1). The heating rate and carrier gas pressure were varied in accordance with a four-factor experimental design incorporating three variation levels, after which the corresponding chromatograms were obtained. Two factor positions within the experimental design matrix were intentionally left vacant. Structural configuration of the PDDoE design, mathematical treatment of the experimental results, and computations based on the derived empirical equations were performed using the software package "PDDoE" [41].

An Agilent 7890A gas chromatograph (USA, manufacturing year — 2008) with an Agilent 5975C mass-selective detector (USA, manufacturing year — 2008) were used for data collection. Mathematical processing of the method application results, as well as calculations based on the obtained empirical equations, were performed using the previously developed "VDPE" program [41] and a more flexible pipeline of scripts. Scripts were developed with "R" programming language in "R" development framework with RStudio IDE [42]. The applicability of the scripts was validated on small data set (singular chromatograms). Developed scripts are applicable to the processing of data from analysis of similar samples (see *Supporting Information*).

Results and Discussion

The previous application of probabilistic-deterministic design of experiment [36] ensures optimization of gas chromatography conditions with a minimum number of experiments and yields robust relationships suitable for extrapolation to other systems. Optimal conditions for the column heating rate and gas pressure levels were calculated and selected based on response surface analysis. The generalized empirical equations for retention time (1) and resolution (2) derived from the PDDoE analysis take the following forms:

$$\overline{t_R} = 81.26 \times \Delta T^{-0.7631} \times \frac{1}{0.03421 + 0.0004116 \times P} / 27.3804 \text{ , } R = 0.9917, t_R = 146.9447 \quad (1)$$

$$R_S = (0.083 + 0.2346 / P) \times (0.1364 + 0.08223 / \Delta T) / 0.1577 \text{ , } R = 0.8003, t_R = 5.4526 \quad (2)$$

Based on the obtained equation, the optimal values of carrier gas pressure and column heating rate were selected to ensure a balance between chromatographic resolution and total analysis duration. Given the substantial number of measurements in the present study — 120 samples in total (four coal types, 30 extracts each) — minimizing individual run time was a practical necessity.

Overall device settings provide a balance between analysis time and separation quality:

– column type — Rtx-100DHA;
– column length — 30 m;
– column diameter — 0.25 mm;
– column adsorbent thickness — 0.5 μm

– injector temperature — 280 °C;
– column temperature — 60–300 °C;
– column heating rate — 8 °C/min;
– ion source temperature — 230 °C;
– quadrupole condenser temperature — 150 °C; EI+ = 70 eV
– carrier gas — helium grade A;
– column gas pressure — 12 psi;
– sample volume — 0.2 μL;
– injection mode — splitless;
– mass spectrum recording mode — scan;
– library — NIST 08;
– analysis time — 32 min.

The chosen parameters (column heating rate — 8 °C/min, column gas pressure — 12 psi) therefore represent a compromise solution: maximizing peak resolution while avoiding unnecessary prolongation of each chromatographic cycle.

Extract components were identified using GS-MSD DataAnalysis software by comparing acquired mass spectra with the NIST 08 library data. The accuracy of the data for subsequent use in the calculations was ensured by 30 individual measurements on each of four samples.

Examples of extract chromatograms for each coal source are presented in Figures 1, 2.

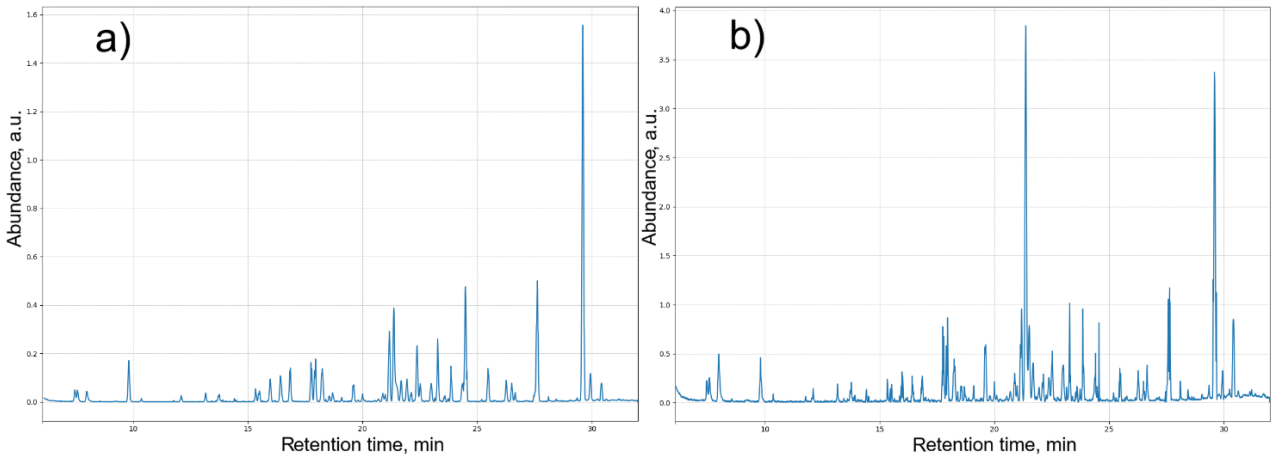


Figure 1. Chromatogram of the coal extract samples: *a* — "KA" sample (KA_1), *b* — "RA" sample (RA_1)

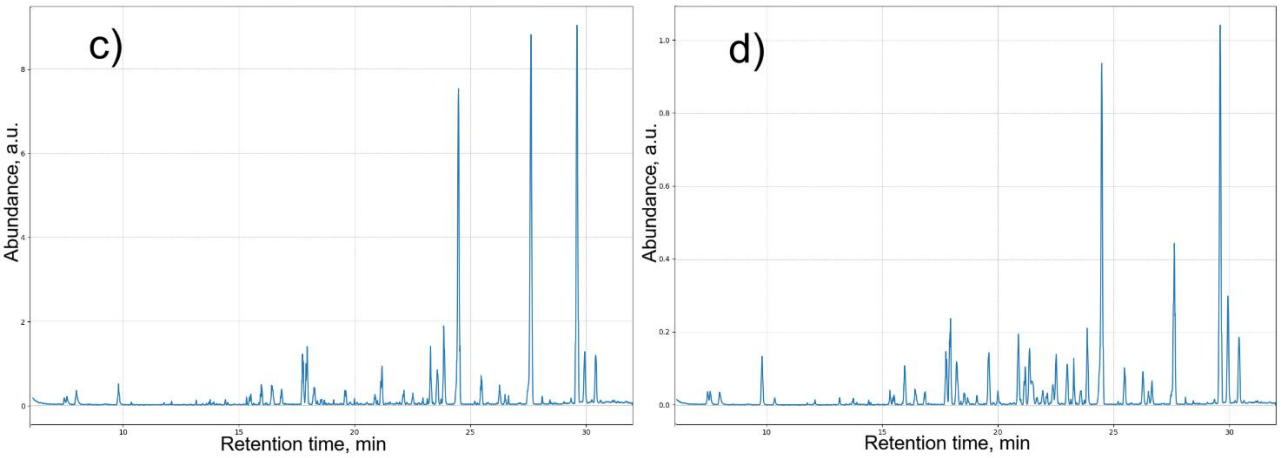


Figure 2. Chromatogram of the coal extract samples: *c* — "SH" sample (SH_1), *d* — "EK" sample (EK_1)

Additional data on the composition and ash content of the coals are presented in Table 1. Overall, the analytical results are consistent with published data on Central Kazakhstan coals [43, 44]; however, they are

not considered sufficient to draw conclusions regarding sample provenance. The use of normalized mineral composition data in combination with chromatographic data through data fusion approaches appears promising, though its applicability is constrained by the labour intensity of data collection.

Table 1

Mineral composition of the studied coals

Sample name	Ash content, %	SiO_2, %	Al_2O_3, %	CaO, %	MgO, %	Fe_2O_3, %	TiO_2, %
KA	22.31±0.12	55.11±3.56	28.48±2.41	2.62±0.025	1.89±0.01	4.51±0.08	2.61±0.05
EK	40.93±0.14	61.1±3.97	23.51±1.99	4.76±0.045	1.01±0.005	4.97±0.09	1.05±0.02
SH	14.82±0.07	56.32±3.64	22.45±1.89	2.11±0.02	1.61±0.008	8.27±0.15	0.87±0.01
RA	24.27±0.17	62.54±4.02	24.77±2.03	3.17±0.036	1.21±0.007	6.08±0.11	0.53±0.01

Scripts were developed in R programming environment for the chromatograms data extraction needed for the calculations. A key feature of used data extraction method is peak alignment of identical components by retention time and assignment of zero areas when a peak is absent in one or more chromatograms. This avoids NA and/or NaN values during further processing.

The script performs line-by-line reading of source files, sample name extraction, recognition of tabular blocks, conversion of text structure to tables, and compilation of a final matrix into a format suitable for further chemometric processing (PCA, clustering, etc.). Chromatograms of 120 coal extract samples obtained from four deposits in central and northern Kazakhstan were used to build the model.

During chemometric processing, all relative peak areas (Area Pct) are treated as a feature vector for each sample. Centering and scaling eliminate the unequal contribution of peaks depending on absolute area.

Principal component analysis transforms the original feature space into new orthogonal axes (PC1, PC2, PC3, etc.), each describing a decreasing proportion of total data variance. This reduces sample data dimensionality from approximately 40–45 original variables (peak areas) to several principal components describing almost all dataset variance.

An interactive scores and loadings plot (See *Supporting Information*) allows rotation of the point cloud to examine spatial clusters and sample relative positions, and to track the directional influence of original variables — chemical components of the mixture — on classification in the space of the first three principal components. Coincident or proximate points correspond to similar chromatographic profiles. Distinguishable point groups indicate differences in extract composition. Adjacent samples (coal samples from one deposit) form compact clusters.

In the space of the first three principal components, compact clusters are observed corresponding to samples attributed to the same source. Notably, one sample group distinctly separates from the others along the first principal component (PC1), indicating systematic differences in extractable organic fraction composition compared to other samples. The remaining samples form several closely spaced but distinct clusters, the separation of which is primarily evident along PC2 and PC3.

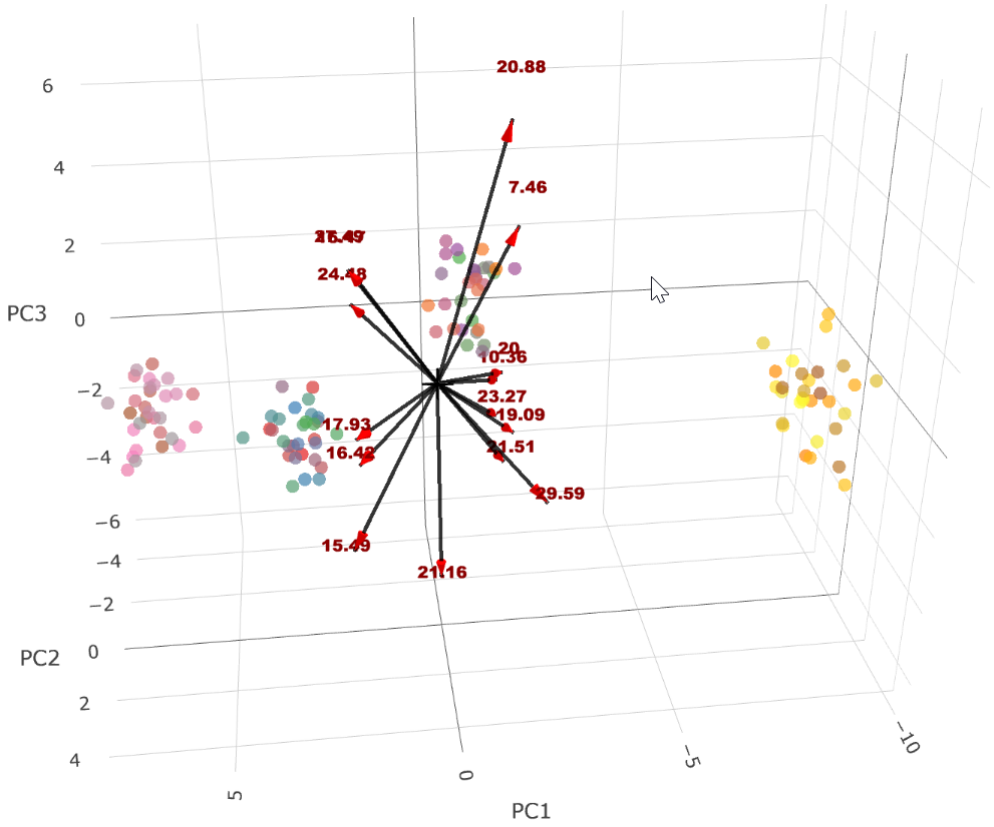
Loading vectors plotted on the graph reflect the contribution of individual chromatographic peaks to principal component formation. Vector direction and length indicate variables most strongly influencing sample separation in PC1-PC3 space. Thus, differences between clusters are determined by the cumulative contribution of several organic phase components rather than single marker compounds, emphasizing the complex nature of coal extract chromatographic “fingerprints.”

The projection of the obtained three-dimensional model onto a plane is presented in Figure 3.

Loading vectors are labeled with the retention time of the corresponding peak:

20.88 — Nonadecane (93 %);
21.16 — 1,2-Benzenedicarboxylic acid, butyl 2-methylpropyl ester (94 %);
10.36 — Dodecane (72 %);
16.42 — Carbonic acid, pentadecyl 2,2,2-trichloroethyl ester (64 %);
15.49 — 1,8-Naphthyridin-2-amine, 5,7-dimethyl- (64 %);
16.47 — Naphthalene, 1,6,7-trimethyl- (94 %);
23.27 — Heneicosane (87 %);
17.93 — Naphthalene, 1,6-dimethyl-4-(1-methylethyl)- (98 %);
7.46 — Benzene, 1,2-diethyl- (90 %);

19.09 — Phenanthrene (80 %);
27.49 — Pentacosane (93 %);
29.59 — Terephthalic acid, di(4-octyl) ester (81 %);
20.00 — Nonadecane (78 %);
24.48 — Phenanthrene, 1-methyl-7-(1-methylethyl)- (99 %).



PC1 axis depicts the coordinates of the 1st principal component,
PC2 axis depicts the coordinates of 2nd principal component,
PC3 axis depicts the coordinates of 3rd principal component

Figure 3. PCA of GC data results: scores and loadings

There are existing clustering methods that work with native variables without dimensionality reduction. One such method is hierarchical clustering of chromatographic profiles. Hierarchical clustering of the obtained chromatographic data enables visual representation of the degree of similarity between samples based on the all registered peaks. In some cases, this approach proves to be even more informative than PCA, as it preserves the metric structure of the original data and allows identification of sample groups with similar profiles without preliminary dimensionality reduction, as shown in Figure 4.

Hierarchical clustering was performed using the Euclidean distance metric and Ward's minimum variance linkage method (Ward.D2). Prior to clustering, the data were autoscaled to prevent domination of variables with larger numerical ranges.

The clustering of coal extract chromatographic profiles revealed a clear multilevel structure of sample similarity. At the top level of the dendrogram, all samples separate into two main groups, indicating fundamental differences in organic extract composition. One of these groups is formed predominantly by samples from one deposit, while the second combines samples from three other sources.

With further clustering depth, each main group subdivides into more compact subgroups, resulting in four stable clusters. These clusters are characterized by high intragroup similarity of chromatographic profiles and are distinctly different from each other based on the total set of registered peaks.

The obtained cluster structure indicates that differences between samples are systematic in nature and determined by peculiarities in the composition of extractable organic matter, rather than random variations in analytical measurements.

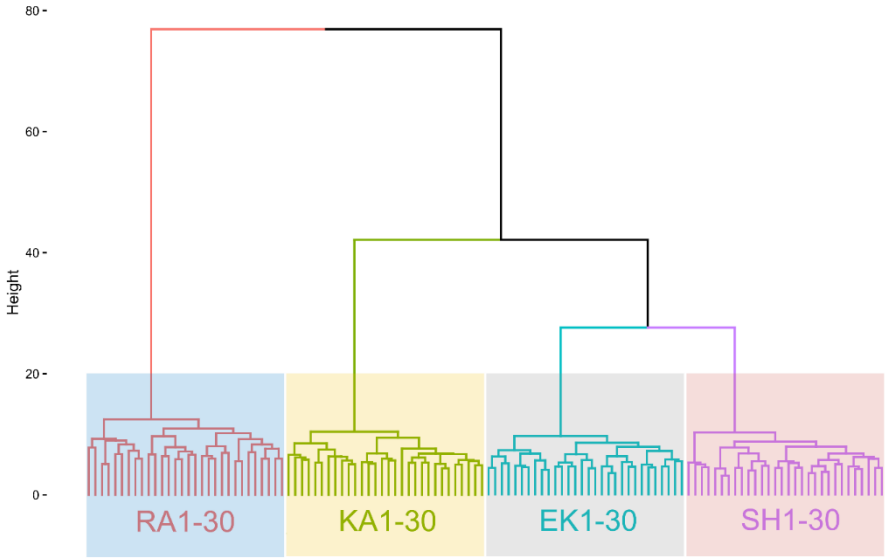


Figure 4. Hierarchical clustering of coal samples (RA1-30 — "Rapid" coal sample, 30 repetitions; KA1-30 — "Karazhyra" coal sample, 30 repetitions; EK1-30 — "Ekibastuz" coal sample, 30 repetitions; SH — "Shubarkol" coal sample, 30 repetitions)

Hierarchical clustering was performed directly on normalized chromatographic data without preliminary dimensionality reduction, which allowed assessment of the natural similarity structure of samples.

To verify the stability of the identified grouping, k-means clustering was additionally applied. The optimal number of clusters for the k-means method was determined using the elbow method, based on analysis of the dependence of within-cluster sum of squares on the number of clusters.

Analysis of the corresponding graph (Fig. 5) shows that the transition from three to four clusters leads to notable improvement in clustering quality, whereas further increase in cluster number yields only marginal gains. Therefore, $k = 4$ was chosen as a compromise between model complexity and its descriptive capacity.

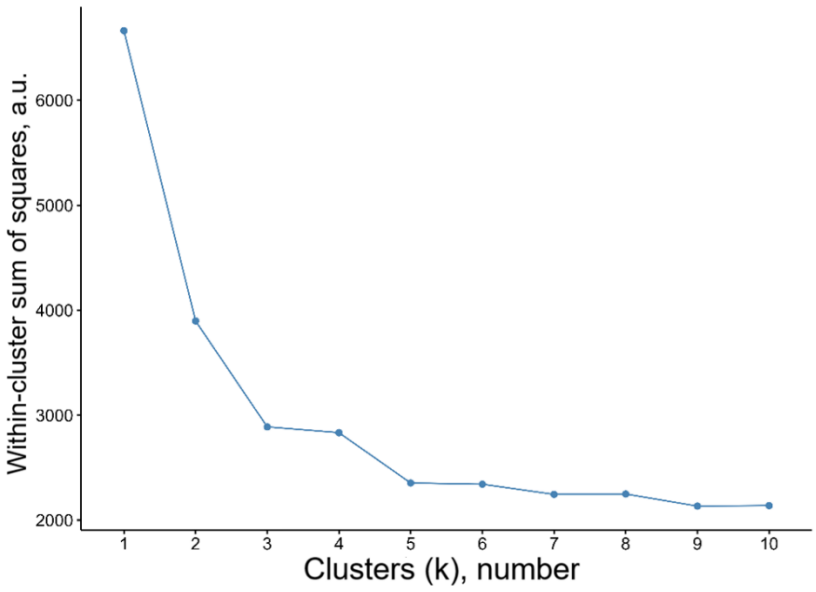
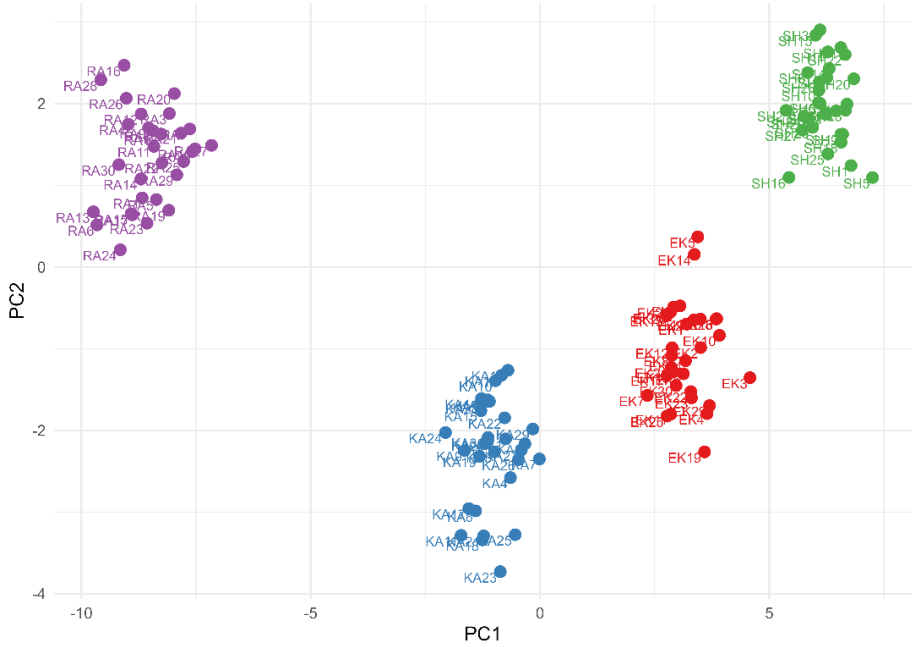


Figure 5. The elbow method — determination of the optimal cluster number

Theoretically, the *k*-means algorithm is based on minimizing the total squared distance between points and cluster centers (so-called centroids). The process iteratively recalculates center positions until the distance between them and the points belonging to the group becomes minimal. Thus, each sample is assigned to the cluster whose centroid is closest to it in multidimensional feature space. In the case of this work, the normalized percentage peak areas from GC-MS identified by retention times were used.

Despite its simplicity, the method reliably identifies main patterns: samples with similar chromatographic profiles group together, while distinct ones form separate clusters. The resulting picture clearly shows the internal data structure: which samples are similar, which differ, and to what extent (Fig. 6). The outcome demonstrates that 4 clusters was defined as “optimal”, meaning the method classified the entire set of chromatograms into 4 groups.

Unlike PCA, where data is projected onto generalized variance axes, *k*-means works directly in the original space of normalized features, making cluster interpretation more straightforward and sometimes more understandable for practical analysis of coal extracts.



PC1 axis depicts the coordinates of the 1st principal component,
PC2 axis depicts the coordinates of 2nd principal component

Figure 6. *k*-means clustering

Thus, unsupervised clustering results confirm that coal extracts possess stable and reproducible features suitable for automatic classification. Methods like *k*-means can be used as a preliminary stage of chemometric analysis for initial sample grouping, selection of representative standards, and verification of models based on PCA, PLS, or neural network approaches. This makes them a valuable tool in constructing coal classification schemes by origin and other potential unifying characteristics.

Conclusions

In the present work, the applicability of a comprehensive analytical approach was evaluated, including sample preparation, gas chromatographic analysis of extracts, and subsequent chemometric data processing, for identifying similarities and differences between bituminous coal samples. Primary attention was given to verifying whether the combination of experimental and computational procedures yields a reproducible and interpretable data structure suitable for further classification and analysis.

Application of principal component analysis showed that the main differences between coal extract samples can be described by a limited number of principal components reflecting the cumulative contribution of multiple chromatographic peaks. In the space of the first three principal components, compact clusters of samples with similar chromatographic profiles form, indicating the presence of stable differences in extractable organic fraction composition.

Hierarchical clustering, performed directly on normalized chromatographic data without preliminary dimensionality reduction, revealed a multilevel structure of sample similarity. At the top level of the dendrogram, samples separate into two main groups, one of which predominantly corresponds to one coal source, while the second combines samples from three other sources. With further clustering depth, four stable clusters form, characterized by high intragroup profile similarity.

Additional *k*-means clustering confirmed the stability of the identified grouping. Analysis of the dependence of within-cluster sum of squares on the number of clusters showed that selecting four clusters results in an optimal compromise between model complexity and its descriptive capacity. The consistency of PCA, HCA, and k-means results, which can be seen by clear identification of four groups by each used method, indicates the systematic nature of differences between samples and the absence of dominant influence from random analytical variations.

The obtained results demonstrate that chromatographic profiles of bituminous coal extracts contain sufficient information for automatic sample grouping without preliminary assignment of deposit affiliation. The advantage of the developed method is that it requires no additional data on the qualitative and quantitative composition of coals, as classification is performed purely based on extract chromatograms. This allows for effective reduction of both analysis time and cost.

The considered approach can be used as a method for primary coal characterization, batch homogeneity control, and formation of "chromatographic fingerprint" databases. In the future, the identified patterns can be compared with technologically significant coal characteristics, such as volatile matter yield, flame length, spontaneous ignition propensity, or thermal degradation features, which opens prospects for expanding the methodology toward predictive models.

Supporting Information

The Supporting Information (Interactive 3D visualization of the results of using PCA; an Excel file with data extracted from chromatograms; a pipeline of R scripts for extracting and processing chromatogram data) is available free at https://github.com/Vtah/SuplimentaryEJoCh2026_2

Funding

This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993009).

*Author Information**

**The authors' names are presented in the following order: First Name, Middle Name and Last Name*

Vitaliy Nikolayevich Fomin — Candidate of Chemical Sciences, Head of LEP "Physical-chemical methods of investigation", Karaganda National Research University named after Academician E.A. Buketov, Universitetskaya street, 28, 100024, Karaganda, Kazakhstan; e-mail: vitfomin@mail.ru; https://orcid.org/0000-0002-2182-2885

Assanali Anuarovich Ainabayev — Candidate of Chemical Sciences, Senior Researcher of the LEP "Physical-chemical methods of investigation", Karaganda National Research University named after Academician E.A. Buketov, Universitetskaya street, 28, 100024, Karaganda, Kazakhstan; e-mail: aliopel_82t@mail.ru; https://orcid.org/0000-0002-3443-446X

Saule Kidirbayevna Aldabergenova — Candidate of Chemical Sciences, Associated Professor, Department of Inorganic and Technical Chemistry, Karaganda National Research University named after Academician E.A. Buketov, Universitetskaya street, 28, 100024, Karaganda, Kazakhstan; e-mail: aldsau@mail.ru; https://orcid.org/0000-0002-4262-911X

Dauletkhan Asanovich Kaykenov — PhD, Lead Researcher of the LEP "Physical-chemical methods of investigation", Karaganda National Research University named after Academician E.A. Buketov, Universitetskaya street, 28, 100024, Karaganda, Kazakhstan; e-mail: krg.daykai@mail.ru; https://orcid.org/0000-0003-4621-7603

Milana Alexandrovna Turovets (*corresponding author*) — Master of Technical Sciences, 2nd year PhD student, Engineer, LEP "Physical-chemical methods of investigation", Karaganda National Research University named after Academician E.A. Buketov, Universitetskaya street, 28, 100024, Karaganda, Kazakhstan; e-mail: turovec26.07@mail.ru; https://orcid.org/0000-0002-0493-6426

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. CRediT: **Vitaliy Nikolayevich Fomin** conceptualization, methodology, vis-

ualization, supervision; **Assanali Anuarovich Ainabayev** data curation, investigation, formal analysis; **Saule Kidirbayevna Aldabergenova** investigation, methodology, validation; **Dauletkhan Asanovich Kaykenov** data curation, formal analysis, visualization; **Milana Alexandrovna Turovets** project administration, methodology, writing-review & editing;

Conflicts of Interest

The authors declare no conflict of interest.

References

1 Alexander, G., & Hazai, I. (1981). Chromatographic fingerprinting of coal extracts. *Journal of Chromatography*, 217, 19–38. https://doi.org/10.1016/S0021-9673(00)88059-0

2 Bartle, K.D., Mills, D.G., Mulligan, M.J., Amaechina, I.O., & Taylor, N. (1984). Molecular mass calibration in size-exclusion chromatography of coal derivatives. *Fuel, 63*(11), 1556–1560. https://doi.org/10.1016/0016-2361(84)90226-6

3 Blanco, C., Prado, J.G., Guillén, M.D., Borrego, A.G., & Iglesias, M.J. (1991). Capillary gas-chromatographic and combined gas-chromatography mass-spectrometric study of the volatile fraction of a coal-tar pitch using OV-1701 stationary phase. *Journal of Chromatography, 539*(1), 157–167. https://doi.org/10.1016/S0021-9673(01)95369-5

4 Assis, L., & Lanças, F. (1999). High-resolution gas chromatography and high-resolution gas chromatography/mass spectrometry study of the volatile fraction obtained from high-inertinite Brazilian coal by supercritical fluid extraction. *Journal of Microcolumn Separations, 11*(7), 501–512. https://doi.org/10.1002/(SICI)1520-667X(1999)11:7<501::AID-MCS2>3.0.CO;2-V

5 Takanohashi, T., Yoshida, T., Iino, M., Katoh, K., & Nishioka, M. (2000). An inverse liquid chromatography study of the interaction of organic compounds with Argonne premium coals. *Energy & Fuels, 14*(3), 720–726. https://doi.org/10.1021/ef990261y

6 Platonov, V.V., Proskuryakov, V.A., Korotaeva, K.V., & Yur'ev, E.M. (2002). Development and optimization of the procedure of gas-chromatographic elemental analysis of high-carbon solid fossil fuels. *Russian Journal of Applied Chemistry, 75*(5), 834–839. https://doi.org/10.1023/A:1020387318686

7 Rathsack, P., & Otto, M. (2014). Classification of chemical compound classes in slow pyrolysis liquids from brown coal using comprehensive gas-chromatography mass-spectrometry. *Fuel, 116*, 841–849. https://doi.org/10.1016/j.fuel.2013.05.100

8 Zubkova, V., & Witkiewicz, Z. (2016). Chromatographic analysis of chemical compositions of coals and changes in them during technological processing. *Critical Reviews in Environmental Science and Technology, 46*(7), 701–755. https://doi.org/10.1080/10643389.2016.1154779

9 Zuber, J., Ecker, D., Otto, M., & Wüst, E. (2016). Gas chromatography/atmospheric pressure chemical ionization-Fourier transform ion cyclotron resonance mass spectrometry of pyrolysis oil from German brown coal. *International Journal of Analytical Chemistry*, 2016, Article 5960916. https://doi.org/10.1155/2016/5960916

10 Li, G., Li, Z., Ma, C., Zhang, L., Cheng, J., & Hou, Y. (2019). Molecular characteristics of the soluble components from three low-rank coals based on the analyses using GC/MS and GC/Q-TOF MS. *Fuel*, 254, Article 115671. https://doi.org/10.1016/j.fuel.2019.06.010

11 Li, W., Zhang, D., Hou, Y., Qiao, E., & Cui, S. (2019). Analysis of light weight fractions of coal-based crude oil by gas chromatography combined with mass spectroscopy and flame ionization detection. *Fuel*, 241, 392–401. https://doi.org/10.1016/j.fuel.2019.04.108

12 Wang, X., Zhu, Z., & Li, X. (2024). Analysis of the Organic Chemical Fractions of Three Coal Extracts. *Applied Sciences, 14*(19), 8933. https://doi.org/10.3390/app14198933

13 Wang, X., & He, X. (2022). Cluster Analysis of Soluble Organic Fractions in Two Low-Rank Coals. *Applied Sciences, 12*(22), 11562. https://doi.org/10.3390/app122211562

14 McGregor, L., Gauchotte, E., Habets, F., & Goovaerts, P. (2012). Multivariate statistical methods for the environmental forensic classification of coal tars from former manufactured gas plants. *Environmental Science & Technology, 46*(7), 3744–3752. https://doi.org/10.1021/es203708w

15 Zhang, X., Wei, X., Li, G., Wang, Y., Zhang, L., & Zong, Z. (2020). Structural characteristics of soluble organic matter in four low-rank coals. *Fuel*, 267, Article 117230. https://doi.org/10.1016/j.fuel.2020.117230

16 Fan, X., Yu, H., Liu, Z., Yu, J., Zhou, Q., & Wei, X. (2018). Molecular characteristics of Shenfu coal characterized by mass spectrometers with three ion sources. *ChemistrySelect, 3*(37), 10383–10387. https://doi.org/10.1002/slct.201802238

17 Xu, H., Zhang, L., Wei, X., Fan, X., & Zong, Z. (2024). Exploring the molecular characteristics of organic matter in low-rank coals using GCxGC/TOF-MS plus data mining. *Journal of Analytical and Applied Pyrolysis*, 181, Article 106605. https://doi.org/10.1016/j.jaap.2024.106605

18 Hamilton, J.F., Webb, P.J., Lewis, A.C., & Reviejo, M.M. (2007). Comprehensive two-dimensional gas chromatography coupled to time-of-flight mass spectrometry of coal liquids produced during a coal liquefaction process. *Energy & Fuels, 21*(1), 286–294. https://doi.org/10.1021/ef060366i

19 Zhang, X., Li, G., Wei, X., Zhang, L., & Zong, Z. (2019). Cluster analysis of molecular characteristics for soluble organic matter in coals. *Chinese Journal of Analytical Chemistry, 47*(1), 99–105. https://doi.org/10.19756/j.issn.0253-3820.181585

20 Zeng, Z., Hugel, H.M., Marriott, P.J., & Schoenmakers, P.J. (2014). Interpretation of comprehensive two-dimensional gas chromatography data using advanced chemometrics. *TrAC Trends in Analytical Chemistry*, 53, 150–166. https://doi.org/10.1016/j.trac.2013.08.009

21 Li, G., Li, Z., Jiang, H., Mo, W., Hu, H., & Zhang, L. (2019). Insight into molecular information of Huolinguole lignite obtained by Fourier transform ion cyclotron resonance mass spectrometry and statistical methods. *Rapid Communications in Mass Spectrometry, 33*(13), 1107–1113. https://doi.org/10.1002/rcm.8448

22 Huang, J., Li, G., Xu, H., Li, Z., Qin, Z., Fan, X., & Wei, X. (2024). Molecular characteristics of eight lignites based on the big data obtained from Orbitrap mass spectrometry. *Journal of the Energy Institute*, 114, Article 101569. https://doi.org/10.1016/j.joei.2024.101569

23 Li, G., Qin, Z., Li, Z., Xu, H., & Wei, X. (2024). Combination of chemometrics and mass spectrometric methods for the data mining of molecular structure information of coal and biomass. *Fuel*, 361, Article 130714. https://doi.org/10.1016/j.fuel.2023.130714

24 Roy, A., Varma, A. K., Sar, T. K., Biswas, S., & Gupta, S. (2021). Insights from principal component analysis applied to Py-GCMS study of Indian coals and their solvent extracted clean coal products. *International Journal of Coal Science & Technology, 8*(6), 1504–1514. https://doi.org/10.21203/rs.3.rs-127356/v1

25 Li, Y., Li, G., Xu, H., Li, Z., Fan, X., Qin, Z., & Wei, X. (2024). Accurate classification of the molecular characteristics of soluble portions from various lignites: Joint analysis of thermal dissolution experiments and data mining methods. *Journal of Analytical and Applied Pyrolysis*, 180, Article 106536. https://doi.org/10.1016/j.jaap.2024.106536

26 Fan, H.-H., Li, J., Song, L., & Cui, Y. (2025). Comprehensive quantitative analysis of coal-based liquids by Mask R-CNN-assisted two-dimensional gas chromatography. *Separations, 12*(2), Article 22. https://doi.org/10.3390/separations12020022

27 Li, B., Li, G., Jiang, H., Mo, W., Hu, H., & Zhang, L. (2019). Insight into molecular information of Huolinguole lignite obtained by Fourier transform ion cyclotron resonance mass spectrometry and statistical methods. *Rapid Communications in Mass Spectrometry, 33*(13), 1107–1113. https://doi.org/10.1002/rcm.8448

28 Khare, P., Baruah, B.P., & Rao, P.G. (2011). Application of chemometrics to study the kinetics of coal pyrolysis: A novel approach. *Fuel*, 90(11), 3299–3305. https://doi.org/10.1016/j.fuel.2011.05.017

29 Roy, A., Varma, A.K., Rao, K.S., Prasad, M., & Reddy, B.S. (2019). Py-GCMS studies of Indian coals and their solvent extracted products. *Fuel*, 256, Article 115981. https://doi.org/10.1016/j.fuel.2019.115981

30 Yin, H., Yang, Y., Liu, X., Li, J., & Zhang, X. (2021). Application of chemometrics for coal pyrolysis products by online py-GCxGC-MS. *ACS Omega, 6*(5), 3763–3770. https://doi.org/10.1021/acsomega.0c05359

31 Lu, W., Li, H., Wang, X., Zhang, Y., & Chen, J. (2024). Discrimination of coal geographical origins through HS-GC-IMS assisted with machine learning algorithms in larceny case. *Journal of Chromatography A*, 1735, Article 465330. https://doi.org/10.1016/j.chroma.2024.465330

32 Zhang, L., Li, G., Wei, X., Fan, X., & Zong, Z. (2022). Characterization of nitrogen-containing compounds in coal tar and its subfractions by comprehensive two-dimensional GC x GC-TOF and ESI FT-ICR mass spectrometry based on new separation method. *Fuel Processing Technology*, 227, Article 107213. https://doi.org/10.1016/j.fuproc.2022.107213

33 Fan, H.-H., Li, J., Song, L., & Cui, Y. (2025). Comprehensive quantitative analysis of coal-based liquids by Mask R-CNN-assisted two-dimensional gas chromatography. *Separations, 12*(2), Article 22. https://doi.org/10.3390/separations12020022

34 Wu, Z., Rodgers, R.P., & Marshall, A.G. (2003). Resolution of 10,000 compositionally distinct components in polar coal extracts by negative-ion electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Energy & Fuels, 17*(4), 946–953. https://doi.org/10.1021/ef030026m

35 Li, Z., Li, G., Xu, H., Qin, Z., Fan, X., & Wei, X. (2025). Classification of soluble proportions derived from coals and their correlation with coal type: Conjoint analyses of extraction, thermal dissolution and machine learning. *Journal of the Energy Institute*, 119, Article 102010. https://doi.org/10.1016/j.joei.2025.102010

36 Fomin, V.N., Aynabaev, A.A., Kaykenov, D.A., Sadyrbekov, D.T., Bakhytkyzy, I., & Aldabergenova, S.K. (2021). Optimization of coal tar gas chromatography conditions using probabilistic-deterministic design of experiment. *Bulletin of the Karaganda University. Chemistry Series*, 4(104), 39–46. https://doi.org/10.31489/2021Ch4/39-46

37 Turovets, M.A., Fomin, V.N., Kelesbek, N.K., Ainabayev, A.A., & Sadyrbekov, D.T. (2024). Chemometric approach for the determination of vanadium by the LIBS method. *Eurasian Journal of Chemistry, 29*(4), 61–70. https://doi.org/10.31489/2959-0663/4-24-10

38 Fomin, V.N., Aldabergenova, S.K., Kelesbek, N.K., Ainabayev, A.A., Sadyrbekov, D.T., Kaykenov, D.A., Borsynbayev, A.S., Azhibay, N.T., & Turovets, M.A. (2024). Method of classification and quantitative analysis of vein quartz using LIBS and chemometric techniques. *Bulletin of the L.N. Gumilyov Eurasian National University. Chemistry. Geography. Ecology Series, 2*(147), 48–60. https://doi.org/10.32523/2616-6771-2024-147-2-48-60

39 Fomin, V.N., Aldabergenova, S.K., Rustembekov, K.T., Omarov, K.B., Rozhkovoy, I.E., Dik, A.V., & Saulebekov, D.M. (2021). Optimization of the parameters of a laser induced breakdown spectrometer (LIBS) using probabilistic-deterministic design of experiment. Industrial Laboratory. *Diagnostics of Materials, 87*(5), 14–19. https://doi.org/10.26896/1028-6861-2021-87-5-14-19

40 Fomin, V., Turovets, M., Kelesbek, N., Ainabayev, A., Sadyrbekov, D., Kaykenov, D., Borsynbayev, A., Azhibay, N., & Aldabergenova, S. (2025). LIBS of low-alloyed lead systems: Chemometric data processing and quantitative analysis. *Analytica, 6*(4), Article 55. https://doi.org/10.3390/analytica6040055

41 Fomin, V.N. (2018). *Veroyatnostno-determinirovannoe planirovanie eksperimenta (VDPE)* [Probabilistically-deterministic design of experiments (PDDE)] (Certificate of Authorship of the Republic of Kazakhstan No. 26 dated 01.10.2018) [Computer software]. Kazpatent. https://copyright.kazpatent.kz/?!.iD=kux

42 Fomin, V. (2025). Software implementation of probabilistic-deterministic design of a chemical experiment on R. *Bulletin of the L.N. Gumilyov Eurasian National University. Chemistry. Geography. Ecology Series, 2*(151), 130–142. https://doi.org/10.32523/2616-6771-2025-151-2-130-142

43 Buldakov, Y.M., Egizekov, M.G., Kulenova, N.A., Reymer, Y.A., & Skorikov, S.P. (2018). Tovarnyy ugol' i produkty yego szhiganiya — perspektivy razvitiya novykh proizvodstv [Commercial coal and its combustion products: prospects for new industry development]. *Novosti nauki Kazahstana — News of Kazakhstan Science*, *1*(135), 99–116. https://vestnik.nauka.kz/storage/docs/2018/03/9-%D0 %91 %D1 %83 %D0 %BB%D0 %B4 %D0 %B0 %D0 %BA%D0 %BE%D0 %B2.pdf [in Russian]

44 Ermagambet, B.T., Kasenov, B.K., Nurgaliyev, N.U., Kazankapova, M.K., Kasenova, Zh.M., & Kuanyshbekov, E.E. (2020). Chemical Composition and Electrophysical Characteristics of the Ash of Bogatyr Coal. *Solid Fuel Chemistry, 54*(2), 99–104. https://doi.org/10.3103/s0361521920020020