Fifteenth Winter Symposium on Chemometrics

Modern Methods of Data Analysis

Uzbekistan, Chimgan, 23-27 February, 2026

SCIENTIFIC COMMITTEE

Dr. Federico Marini, University of Rome La Sapienza
Dr. Jose Manuel Amigo, University of the Basque Country
Dr. Alexey Pomerantsev, Semenov Federal Research Center for Chemical Physics RAS
Dr. Oxana Rodionova, Semenov Federal Research Center for Chemical Physics RAS

ORGANIZING COMMITTEE

Organizing Committee

Dr. Bekzod Khakimov, University of Copenhagen
Dr. Dmitry Kirsanov, St Petersburg University
Dr. Sergey Kucheryavskiy, Aalborg University
Dr. Dilbar Dalimova, Center for Advanced Technologies, Tashkent
Dr. Vladimir Tsoy, Center for Advanced Technologies, Tashkent

Secretary

Dr. Anastasiia Surkova, ITMO University, St Petersburg, Russia

http://wsc15.com e-mail: wsc15chemometrics@gmail.com

General Information & Logistics

Venue: Layner Mountain Resort Complex

Chimgan village, Tashkent Region, Uzbekistan

The conference will take place at the Layner Mountain Resort, located in the picturesque Chimgan Mountains (approx. 100km from Tashkent).

Accommodation

Participants will be accommodated in the following buildings (located centrally on the map):

- **Building No. 5** (San Antonio)
- **Building No. 6** (Calypso)
- **Building No. 7** (Victoria)

Conference Sessions

All scientific sessions will be held in the **"Enterprise" Conference Hall**.

- **Location:** Inside the main Administration/Reception building (**Nr. 29** on the map).
- *Route:* From your accommodation (Buildings 5, 6, 7), walk towards the main entrance area to Building 29.

Meals

All meals (Breakfast, Lunch, and Dinner) will be served in the main **Restaurant** (**Nr. 11** on the map).

- **Breakfast time:** 08:00 – 11:00
- **Lunch & Dinner:** Served according to the conference schedule.
- **Banquet**

Leisure & Activities

In your free time, participants are welcome to use the resort's extensive facilities.

- **Indoor/Relaxation:** Spa center (Nr. 17), Billiards, Karaoke (Nr. 31), and Gaming Room (PS5).
- **Outdoor:** Football field (Nr. 18), Tennis court (Nr. 19), Basketball court (Nr. 20), and the All-season slide (Nr. 32).

Scores & Loadings

Traditionally, each conference day concludes with the "Scores & Loadings" gathering to foster informal communication among participants. The exact location (usually a bar or terrace) will be announced daily.

Communication

- **Wi-Fi:** Free Wi-Fi is available throughout the resort territory.
- **Mobile Networks:** The main mobile operators in Uzbekistan are **Ucell, Beeline, Mobiuz, and Uztelecom**. SIM cards can be purchased at the airport upon arrival.

Money

- **Currency:** The local currency is the **Uzbek Sum (UZS)**.
- **Exchange:** You can exchange EUR and USD at the International Airport in Tashkent or at banks in the city.
- **Cards:** Major credit cards (Visa, Mastercard) are generally accepted at the resort reception, but it is highly recommended to carry some cash (UZS) for small expenses or local souvenirs.

Navigation Tip based on the Map:

- **To the Conference:** Walk from the center of the resort (Buildings 5, 6, 7) down towards the main gate area to Building **29**.
- **To the Restaurant:** Walk from your building slightly uphill/centrally to Building **11** (located near the Amphitheater).

Useful Phone Numbers

Anastasiia Surkova, organizing committee
+79277454932 (telegram: @Anastasiia_Melenteva)

Amir-Temur Toshpulatov, *local organizing committee*
+998 90 013 27 04 (telegram: @temuramir)

Dmitry Kirsanov, *organizing committee*
+7921333-1246 (telegram: +7921333-1246)

Layner Venue Map

Building No. 1 "Esperance"
Building No. 2 "Discovery"
Building No. 3 "Santa Maria"
Building No. 4 "Liberty"
Building No. 5 "San Antonio"
Building No. 6 "Calypso"
Building No. 7 "Victoria"
Building No. 8 "Queen Mary"
9) Summer Terrace
10)Covered Terrace
11) Restaurant
12) Conference Hall "Endeavour"
13) Family Zone
14) Service Premises / Staff Only
16) Amphitheater
17) Wellness Complex
18) Football Field
19) Tennis Court
20) Basketball Court
21) Women's Pool
22) Summer Pool
23) Children's Aquapark
24) Changing Rooms / Showers
25) Bar, Fast Food
26) Beach Volleyball
27) Main Parking
28) Additional Parking
29) Reception / Conference Hall "Enterprise"
29.1 Reception
29.2 Conference Hall "Enterprise"
30) Outdoor Parking
31) Karaoke + GameClub
32) All-Season Slide
33) Bus Parking
34) WC
35) Retreat Zone
36) Children's Summer Playground

Monday, February 23, 2026

14:00–15:00	**Registration**
15:00–15:15	**Opening**
15:15–15:45	**Coffee break**
Session 1	**Chair: János Elek**
15:45–16:10	**T1** *Yuri Kalambet* Smoothing method comparison using figures of merits for chromatographic peaks
16:10–16:35	**T2** *Ekaterina Yuskina* Non-invasive screening of kidney and prostate cancer by potentiometric urine analysis and machine learning
16:35–17:00	**T3** *Alexey Shaposhnik* Multivariate Calibration for Selective Analysis
17:00–17:25	**T4** *Dmitriy Matyushin* A Gas chromatography-mass spectrometry for navigating chemical universe
17:25–19:00	**Free time**
19:00–20:00	**Dinner**
20:00–00:00	**Scores & Loadings**

Tuesday, February 24, 2026

08:00–09:30	**Breakfast**
Session 2	**Chair: Yuri Kalambet**
10:00–10:45	**L1** *José Manuel Amigo* A Didactical Unified Framework for Understanding Projection Models
10:45–11:10	**T5** *Ekaterina Boichenko* Behind Blue Dyes: NIR Spectroscopy and Chemometrics for Identification of Pigments in Art
11:10–11:40	**Coffee break**
11:40–12:05	**T6** *Xihui Bian* Advanced signal processing and modeling methods in spectral analysis of complex samples
12:05–12:30	**T7** *Andrey Samokhin* Like a Matryoshka: Nested Interfaces from R to Python to RDKit
12:30–13:00	**Free time**
13:00–14:00	**Lunch**
Session 3	**Chair: Federico Marini**
14:00–14:45	**L2** *Alexey Skvortsov* Combination of multivariate curve resolution with P-spline regression
14:45–15:10	**T8** *Hamid Abdollahi* A Generalized Classical Least Squares Approach for Calibration of Linear and Nonlinear Models with Partial Analyte Information
15:10-15:35	**T9** *Anastasiia Surkova* QSPR-Driven Optimization of Sensitive Elements for Nanophotonic VOC Sensor Systems
15:35–16:00	**T10** *Denis Bikmeev* Quantification of Asphaltenes in Toluene Using Impedance and Infrared Spectroscopy Data
16:00–16:30	**Coffee break**
16:30–19:00	**Poster session**
19:00–20:00	**Dinner**
20:00–00:00	**Scores & Loadings**

Wednesday, February 25, 2026

08:00–09:30	**Breakfast**
Session 4	**Chair: José Manuel Amigo**
10:00–10:45	**L3** *Bekzod Khakimov*. Prediction of Human Blood Lipoprotein Concentrations from ^{1}H NMR Spectra: Model Performances and Cage of Covariance
10:45–11:10	**T11** *Alisa Rudnitskaya* Multidimensional liquid sensing via impedimetric virtual sensor arrays
11:10–11:40	**Coffee break**
11:40–12:05	**T12** *Andrey Stavrianidi* Group-targeted prediction of substituted flavone retention using machine learning
12:05–12:30	**T13** *Dilip Sing* Rapid detection of ethylene glycol (EG) & diethylene glycol (DEG) in propylene glycol using portable NIR Spectroscopy and Chemometrics
12:30–13:30	**Lunch**
13:30–19:00	**Skiing time**
19:00–20:00	**Dinner**
20:00–00:00	**Scores & Loadings**

Thursday, February 26, 2026

08:00–09:30	**Breakfast**
Session 5	**Chair: Bekzod Khakimov**
10:00–10:45	**L4** *Federico Marini* Recent Advances in Chemometrics with Focus on Food Quality Control and Traceability
10:45–11:10	**T14** *A.Z. Hazarika* Estimation of cup characteristics of CTC black tea liquor using NIR spectroscopy and chemometric tools
11:10–11:40	**Coffee break**
11:40–12:05	**T15** *Ivan Krylov* Regularization in PARAFAC: better living through sparsity?
12:05–12:30	**T16** *Mikhail Saveliev* Geochemical production allocation in commingled wells using chromatography and chemometric tools
12:30–13:00	**Free time**
13:00–14:00	**Lunch**
Session 6	**Chair: Dmitry Kirsanov**
14:00–14:25	**T17** *Abhishek* Intensity-Invariant Spectral Analysis via Pairwise Intensity Ratios and Accounting for Feature-Dependencies
14:25–14:50	**T18** *Hadi Parastar* Deep Learning for GC–IMS Data Analysis: A Novel Approach to VOC Characterization
14:50–15: 15	**T19** *Nikolay Sushkov* Laser plasma inhomogeneity as revealed by chemometrics
15: 15–15: 45	**Coffee break**
15:45–16:10	**T20** *Vladislav Deev* Application of generative adversarial networks to improve the accuracy of classification models by generating spectra of underrepresented class of urinary stones
16:10–16: 35	**T21** *Máté Csontos* I found the needle in a haystack – and some other things, too
	Career night for early stage researchers
16: 35–17:30	**José Manuel Amigo** The Role of Chemometrics in the Near Future. Is there even a Future?
17: 30–19:00	**Free time**
19:00–00:00	**Banquet**

Friday, February 27, 2026

08:00–09:30	**Breakfast**
09:30–10:30	**Check out**
10:30–15:00	**Bus to Tashkent and excursion**

Abstracts

L01. A Didactical Unified Framework for Understanding Projection Models

José Manuel Amigo[1,2]
[1] *IKERBASQUE. Basque Foundation for Sciences. Bilbao, Spain*
[2] *Department of Analytical Chemistry. University of Basque Country, Spain,*
josemanuel.amigo@ehu.eus

Projection-based models constitute a central methodological family in Chemometrics. They are used routinely in hyperspectral and multispectral image analysis, metabolomics, bioinformatics, and many other chemometric-related domains. Methods such as Principal Component Analysis (PCA), Classical Least Squares (CLS), Generalized Least Squares (GLS), Multivariate Curve Resolution (MCR), Partial Least Squares Regression (PLS), Partial Least Squares Discriminant Analysis (PLS-DA), Analysis of Variance Simultaneous Component Analysis (ASCA), Independent Component Analysis (ICA), Canonical Correlation Analysis (CCA), Principal Component Regression (PCR), Orthogonal Partial Least Squares (O-PLS), Factor Analysis (FA), Linear Discriminant Analysis (LDA) or Non-negative Matrix Factorization (NMF), among others, are typically regarded as conceptually distinct, each associated with a specific theoretical rationale, optimization criterion, or algorithmic tradition.

In this work, a unified mathematical formulation is introduced showing that all these methods can be expressed within a common bilinear projection framework of the form:
$T=f(X,Y,W,C)$ and $\hat{X}=g(T,P)$
where X denotes the data matrix, Y is an optional reference information (e.g., classes), W and C encode method-specific constraints (orthogonality, non-negativity, statistical independence, correlation maximization, closure, or class separation), and T and P the product of the factorization of X. Through systematic algebraic derivations, we show that the apparent diversity of projection models is largely the result of modifying one or more elements in the same underlying structure: the choice of metric or covariance operator, the introduction of external information through supervised objectives, the imposition of structural constraints on loadings or scores, or the optimization of specific statistical moments.

By combining detailed proofs with geometric visualizations, it is shown that well-known methods emerge as particular solutions to a generic optimization template. This perspective reveals intrinsic connections (e.g., between PCA and FA as eigenvalue problems; between PLS, CCA, and O-PLS through cross-covariance maximization; and between MCR, NMF, and CLS through constrained bilinear decompositions. This analysis not only uncovers the mathematical continuity among these approaches but also provides a principled way to interpret their differences in terms of explicit analytical priorities.

Altogether, this unified framework promotes a more coherent and integrative vision of multivariate analysis. It helps practitioners better understand when two methods are genuinely different and when they are simply different parameterizations of the same underlying model. This conceptual clarity supports more informed methodological choices, encourages the development of hybrid or generalized algorithms, and strengthens the interpretability and robustness of chemometric analyses in both research and applied environments.

L02. Combination of multivariate curve resolution with P-spline regression

A.N. Skvortsov [1,2]

[1] *Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg, Russia*
[2] *Laboratory of the molecular biology of stem cells, Institute of Cytology of the Russian Academy of Science, Saint Petersburg, Russia*

Multivariate curve resolution (MCR) is a chemometric approach, which aims at decomposition of experimental data matrices of mixed systems, the multivariate curves, into meaningful "pure" components basing on physicochemical constraints. Typically, the data matrix is obtained by an experimental analytic method, satisfying generalized Beer law. Modern MCR is a powerful toolbox of various methods; many of them are explicitly or implicitly based on local or global principal component analysis (PCA). The solutions typically possess residual ambiguity, which can vary from negligible to quite significant. Thus, obtaining a single feasible solution is relatively easy, but proper estimation of the whole set of feasible solutions is not. The complexity of the latter task explodes with the number of components, but it also rapidly increases with data matrix size. The compression of data matrix that retains the bilinearity and the constraints would be clearly beneficial for MCR. In optical spectra, the spectral resolution is typically much lower than the data pitch, so the spectral bands are physically smooth, many variables are almost collinear, and compression based on smoothness should be possible.

In the present study we tested the potential benefit of using penalized B-spline regression (P-spline regression) as compression method. The initial 2-way data matrices from various optical spectral experiments were fit to P-splines in both directions, effectively decomposing the dataset into 2-dimensional splines. In the present study we limited ourselves to cubic B-splines, but tested regular and irregular bases. The compressed matrix of spline coefficients was then analyzed by PCA and MCR methods (mostly RFA-like). The ability of the compressed matrix to reproduce the initial dataset was based on comparison of singular values and the angles between principal spaces. The relations between the constraints of the initial data and constraints for the spline coefficients (nonnegativity, unimodality, closure) were formulated. Smoothness is an inherent tunable property P-splines, so smoothness constraint could be set up at P-spline regression step.

The studied combination proved to be quite simple and versatile due to the flexibility of B-splines, their nonnegativity, unimodality, smoothness and minimal overlap. The constraints are readily transferred from initial matrix to spline coefficients. So, most conventional MCR methods and algorithms of estimating residual ambiguity can be used on the matrix of P-spline coefficients with minimal changes. The primary bonus is the significant reduction of matrix size. Additionally, the good interpolating properties of B-splines allow the presence of missing points and gaps in the data matrix, given enough points are left to define the spline basis. The matrix may even have irregular steps in row or column directions. Specifically, the latter property allows splitting the initial data matrix into "independent" calibration and validation sparse submatrices. So, application of conventional calibration-validation approach becomes possible. It can be used for tuning the penalties of P-spline regression. It may also potentially used to test the constraint strength in MCR

It is also worth noting that B-splines are actually piecewise polynomials, so their derivatives and antiderivatives are easily obtained. The benefit of this fact for the analysis and visualization of MCR results is discussed.

L03. Prediction of Human Blood Lipoprotein Concentrations from 1H NMR Spectra: Model Performances and Cage of Covariance

Bekzod Khakimov[1,*], Huub C. J. Hoefsloot[2], Age K. Smilde[2], Søren Balling Engelsen[1]

[1] *Department of Food Science, University of Copenhagen, DK-1958 Frederiksberg C, Denmark*

[2] *Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam 1090 GE, The Netherlands*

Lipoprotein subfractions are established biomarkers for early cardiovascular disease risk assessment. However, the reference method for lipoprotein quantification, ultracentrifugation, is labor-intensive and time-consuming, limiting its suitability for large-scale cohort screening. Here, we developed partial least-squares (PLS) regression models that predict ultracentrifugation-derived lipoprotein concentrations from ^{1}H nuclear magnetic resonance (NMR) spectra collected from 316 healthy Danish participants. For the first time, we systematically evaluated multiple ^{1}H NMR spectral regions capturing resonances from lipoprotein-associated molecules and lipid classes to identify parsimonious and robust prediction models. Across 65 lipoprotein main fractions and subfractions, high predictive performance was achieved ($Q^2 > 0.6$) using an optimal region spanning 1.4–0.6 ppm, dominated by lipid methylene and methyl signals. Model generalizability was confirmed in an independent cohort of 290 healthy Swedish participants, with predicted values agreeing with reference measurements by up to 85–95%. Finally, we provide an open access software namely SigMa enabling rapid prediction of blood lipoprotein concentrations from standardized [1] ^{1}H NMR spectra, supporting scalable lipoprotein profiling in epidemiological studies.

Acknowledgements

This study was funded by the Innovation Foundation Denmark through the COUNTERSTRIKE project (4105-00015B), the Counteracting Age-related Loss of Skeletal Muscle (CALM) project (www.calm.ku.dk), and the University of Copenhagen Data+ project funding (Strategy 2013 funds) received for the "Introduction of statistical causality modeling and deep learning to solve the cage of covariance problem in Foodomics/Metabolomics" project.

References

[1] Khakimov, B.; Hoefsloot, H. C. J.; Mobaraki, N.; et al. *Analytical Chemistry* **2022**, *94* (2), 628–636.

L04. Recent Advances in Chemometrics with Focus on Food Quality Control and Traceability

F. Marini [1]

[1]*Department of Chemistry, University of Rome "La Sapienza", 00185 Rome, Italy*

Recent advances in chemometrics have markedly improved the ability to manage and interpret the complex analytical data typically encountered in food quality control, authentication, and traceability. Modern analytical platforms generate highly multivariate and heterogeneous datasets, often affected by strong collinearity and unfavorable signal-to-noise ratios, making advanced multivariate approaches essential. In this framework, chemometrics provides a coherent set of tools that combine statistical modeling, pattern recognition, and variable selection to extract reliable, interpretable, and application-oriented information.

In this communication, particular attention will be devoted to recent developments in variable and interval selection methods, with an emphasis on covariance-based approaches. Methods derived from CovSel and its extensions enable the identification of chemically meaningful variables or contiguous spectral regions, improving model interpretability, robustness, and predictive performance. Window-based and grouping-oriented strategies further extend these concepts by selecting informative spectral windows or clusters of correlated variables, which is especially relevant for spectroscopic and spectrometric data characterized by structured or fragmented signals and for the design of simplified or portable sensing devices [1].

The communication will also focus on recent advances in class modeling for food authentication. In particular, multi-block extensions of SIMCA will be discussed as effective tools for integrating complementary information acquired from multiple analytical techniques, such as MIR, NIR, UV, and Vis spectroscopy. By combining distances from individual block models into a unified decision criterion, these approaches enhance robustness and generalization in real-world applications, including the authentication of artisanal foods and the traceability of products with protected designation of origin.

These findings represent a part of the results obtained within the framework of the National Agritech project (Spoke 9), whose aim is to understand the origin, authenticity, and safety of agricultural productions and agri-food supply chains, in order to promote the alignment of agri-food companies with the 2030 Agenda and the Sustainable Development Goals (SDGs), and to enhance and protect typical and traditional products within agri-food value chains.

Overall, this contribution aims to highlight how these emerging chemometric methodologies can be effectively exploited in food quality control, with specific emphasis on their methodological foundations, practical implementation, and impact on authenticity and traceability studies.

Acknowledgements
This study was carried out within the Agritech National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022).

References
[1] J.M. Roger, A. Biancolillo, B. Favreau, F. Marini, *Chemometr. Intell. Lab. Syst.* **254** (2024) 105223.

T01. Smoothing method comparison using figures of merits for chromatographic peaks

A. Stavrianidi [1], S. Maltsev [2], Y. Kozmin [2], A. Samokhin [1], Y. Kalambet [2]

[1] *Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia*
[2] *Ampersand Ltd., Moscow, Russia*

A multitude of smoothing methods is compared using model system of EMG peak(s) with white noise. Smoothing methods are compared based on effect of smoothing on the most important peak parameters (figures of merit) commonly used for peak evaluation. Smoothing is applied to artificial chromatographic peaks with artificial noise. Raw data: EMG peak modified with additive normal uncorrelated noise are generated by Excel spreadsheet. Peak parameters are selected in a way that excludes oversampling. Generated files are used to

construct a set of calibration chromatograms of the virtual “Peak1” component. Several (five) raw data files with five different implementations of noise are used to construct chromatograms of every concentration level. Generated data are processed by proprietary software Chrom&Spec. The software makes data processing, constructs calibration curves and calculates figures of merit:

- Retention time
- Height
- Area
- Width at half-height
- Asymmetry
- Signal/Noise

for every chromatogram of the set and for each of smoothing methods:

- Median
- Simple moving average
- Gaussian moving average
- Savitzky-Golay moving average
- Exponential recursive moving average
- Eilers’ “A Perfect Smoother”
- Locally Optimal Polynomial Filter (LOPF)
- Improved LOPF (iLOPF)

Calibration curves also provide two figures of merit:

- Relative Standard Deviation
- Estimated SD for unknown concentration analysis

Comparison shows, that LOPF filters have significantly better performance, than other listed filters, including “A Perfect Smoother” by Paul Eilers. Most pronounced improvement occurs in the case of Signal/Noise parameter. Locally optimal filter principles are compared to other families of filters: Fourier transform-based, Matched, Wavelet, Kalman.

T02. Non-invasive screening of kidney and prostate cancer by potentiometric urine analysis and machine learning

E. Yuskina[1], M. Mosoyan[2], I. Jahatspanian[3], A. Vasilev[2], V. Makeev[2], A. Gaponova[1], V. Protoshchak[4], E. Karpushchenko[4], A. Sleptsov[4], D. Kirsanov[1]

[1]Institute of Chemistry, St. Petersburg University, St. Petersburg, Peterhoff, Russian Federation

[2]Almazov National Medical Research Centre of the Ministry of Health, St. Petersburg, Russian Federation

[3] Stock Company Scientific Research and Production Association "Pribor", St. Petersburg, Russian Federation

[4] Department of Urology, S. M. Kirov Medical Academy, St. Petersburg, Russian Federation

Cancer is a large group of diseases that includes over 100 distinct types. While diverse, all cancers share the characteristic of uncontrolled, abnormal cell growth. According to the WHO, cancer is a leading cause of mortality worldwide. This mortality can be reduced through prevention and early detection. Current detection methods include physical examinations, laboratory tests, medical imaging, and biopsies. However, these techniques have drawbacks such as high cost, exposure to radiation, and lengthy imaging procedures. Moreover, the tissue sampling required for biopsies is invasive, painful, and unpleasant for the patient, and is sometimes not feasible. Therefore, the development of non-invasive, portable, and low-cost

diagnostic methods is an urgent task. The chemical composition of biological fluids like blood and urine can differ significantly between patients with diagnosed diseases and healthy individuals. Techniques such as Raman spectroscopy, fluorimetry, impedance spectroscopy, Gas Chromatography - Mass Spectrometry (GC-MS) and electrochemical biosensors are used to analyze these fluids for disease detection. Nevertheless, few studies in the literature focus on the simultaneous identification of multiple cancer types.

The present study [1] aims to develop a rapid screening method for kidney and prostate cancer using a potentiometric multisensor system combined with machine learning. A multisensor system was produced, consisting of 25 cross-sensitive sensors with plasticized, polycrystalline, and chalcogenide glass membranes. The analysis required no specific sample preparation, used a 3 mL sample volume, and had a measurement time of two minutes per sample. 116 urine samples were analyzed: 39 from patients with diagnosed prostate cancer, 38 from patients with diagnosed kidney cancer and 39 from the control group. The sensor responses, which reflect the total ionic composition of the urine, served as the input data for subsequent analysis.

The potentiometric measurement results were used as input for various data visualization (Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP)) and classification methods (Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting Classifier (XGBC), Support Vector Machine (SVM), k-Nearest Neighbors (kNN). The SVM model demonstrated 77% accuracy in differentiating urine samples from kidney cancer patients from the control group, and 79% accuracy for prostate cancer. The RF classifier showed higher accuracy (87%) in distinguishing between the two cancer types. The non-invasive method proposed, upon validation with a larger external dataset, has potential as a promising tool for the simultaneous screening of kidney and prostate cancer. Samples identified as suspicious by this system can subsequently be confirmed using standard clinical diagnostic protocols.

References

[1] E. Yuskina, M. Mosoyan, I. Jahatspanian, et al. *Microchem. J.* **218** (2025) 115589.

T03. Multivariate Calibration for Selective Analysis

A. Shaposhnik [1], P. Moskalev [2], A. Vasiliev[3], K. Oreshkin[1], O. Arefieva[1]
[1] *Department of Chemistry, Voronezh State Agrarian University, 394087 Voronezh, Russia*
[2] *Department of Applied Mathematics, Moscow State University of Technology "STANKIN", 127994 Moscow, Russia*
[3] *Laboratory of Sensor Systems, Dubna State University, 141980 Dubna, Russia*

Multivariate calibration (MC) is widely used for quantitative analysis using multivariate data. Specifically, this method allows for refining the calibration dependence of the response of a semiconductor sensor R operating in thermal modulation mode on the gas concentration φ. In this case, the measurement cycle includes n consecutive values of the electrical resistance R_n, taken at a cyclically changing sensor temperature, which forms a unique stochastically correlated sequence for each gas.

Our proposed method, "Multivariate Calibration for Selective Analysis (MCSA)", differs from the standard MC procedure in that it not only improves the accuracy of quantitative analysis but also enables qualitative analysis, i.e., gas classification.

The first key difference is the determination of the regression relationship between the gas concentration and the sensor's electrical resistance, $\varphi = f_n(R)$, rather than the inverse relationship $R = f(\varphi)$. This type of calibration is uncommon in scientific work, but it is more convenient for writing a gas analyzer algorithm, as it directly converts the measured signal

into a concentration value. The second difference is that it does not find a single average calibration ratio, but a set of $\{f_n\}$ obtained from the training data. Any test experiment also generates a set of sequential electrical resistance values $\{R_n\}$, from which a set of concentration values of the gas being analyzed $\{\varphi_n\}$ can be obtained.

The standard MC procedure involves finding the average concentration value. If this procedure is supplemented by determining the relative standard deviation S_r for the set of $\{\varphi_n\}$, it becomes clear that qualitative analysis can be successfully performed using the Sr value. If the analyzed gas in the test experiment was identical to the gas in the training set, all f_n dependencies operate consistently, and the relative standard deviation is less than a certain critical value: $S_r < S_0$. If the gas was different, the model produces inconsistent predictions, and $S_r > S_0$ [1].

Qualitative and quantitative gas analysis with temperature modulation of the sensor can also be performed using established methods such as principal component analysis (PCA) [2]. However, implementing PCA requires significant computational resources to calculate eigenvectors and projections and can be accomplished using a high-performance processor. We needed to solve this problem using a low-power microcontroller to create an energy-efficient, compact gas analyzer. Using MCSA, based on a series of simple regressions, allows us to solve the problem of selective analysis with minimal computational resources, a key advantage for portable devices.

References

[1] A. Shaposhnik, P. Moskalev, A. Vasiliev et al. *Chemosensors.* **13** (2025) 323.
[2] A. Shaposhnik, P. Moskalev, A. Vasiliev et al. *Sensors & Actuators B*. **334** (2021) 129376.

T04. Gas chromatography-mass spectrometry for navigating chemical universe

D. Matyushin[1], A. Sholokhova [1]

[1]*Frumkin Institute of Physical chemistry and Electrochemistry Russian Academy of Sciences, 31-4, Leninsky prospect, 119071 Moscow, Russia*

The number of small molecules is generally very large, and the number of theoretically possible structures increases incredibly rapidly with the number of atoms in the molecule. For example, the well-known GDB-17 database of "plausible" structures contains 3 billion and 109 billion structures containing 14 and 17 "non-hydrogen" atoms, respectively. This countless number of possible structures forms the "chemical universe".

Only a tiny fraction of all possible structures have ever been synthesized and described—for example, the PubChem database, which includes almost all known structures, lists less than 150 million compounds, not all of which are organic molecules (without metal atoms) with low molecular weights. The number of molecules for which reference gas chromatography-mass spectrometry (GC-MS) data are available is two orders of magnitude lower. Also, the number of known natural products, metabolites, etc., is limited to hundreds of thousands or to several millions.

Gas chromatography-mass spectrometry is one of the main methods for non-target analysis (screening) of volatile small molecules. The space of possible analytes is most often limited by mass spectral databases, but recently, it has become possible to go far beyond this area, exploring the much broader chemical space. There are two main approaches that can be applied: (i) predicting mass spectra for candidate structures and then selecting the one whose mass spectrum best matches the observed one; (ii) predicting molecular fingerprints based on the observed mass spectrum and selecting the candidate whose molecular fingerprint is

closest to the observed one. Two main preliminary "filters" for candidate selection are used: molecular formula (which can be reliably determined using high-resolution mass spectrometry) and retention index (which can be predicted for each candidate).

The aim of this study was to examine what can and cannot be inferred about a molecule's structure given its formula, retention index, and GC-MS mass spectrum. We compared and selected the most accurate algorithms for predicting molecular fingerprints from mass spectra and mass spectra from structure. We also created an accurate algorithm for predicting mass spectra using bond cleavage probability prediction using a graph neural network and considered the latest retention index prediction algorithms developed in 2024-2025.

For various classes of molecules, we assessed the probability of correct identification using each of these two approaches and their combination, considering various chemical spaces for selecting possible candidates. The following spaces were considered: exhaustive isomer enumeration (with the rejection of obviously impossible structures), isomer extraction from PubChem and natural compound databases, and exploration of the chemical universe using generative autoencoders. The impact of the presence of molecules in the training set that are structurally similar to the one being predicted on prediction accuracy was studied. The limitations this imposes on the structure of the analyte that can be identified in this way were demonstrated.

It was shown that in some cases (for certain classes of compounds), the exact structure of a molecule can be determined from GC-MS data in 50-80% of cases, even if it is not present in any databases.

Acknowledgments

This research was supported by the Ministry of Science and Higher Education of the Russian Federation (№ 124041900012-4).

T05. Behind Blue Dyes: NIR Spectroscopy and Chemometrics for Identification of Pigments in Art

E.S. Boichenko [1], I.I. Andreev [2], A.A. Reznichenko[3], M.M. Khaydukova [3,4], S.V. Sirro [2], D.O. Kirsanov [1,3]

[1]*ITMO University, Saint Petersburg, Russia*
[2]*The State Russian Museum & ITMO University, Saint Petersburg, Russia*
[3]*St Petersburg University, Saint Petersburg, Russia*
[4]*Institute of Human Hygiene, Occupational Pathology and Ecology, Saint Petersburg, Russia*

Physicochemical methods of analysis in art studies address essential tasks for any museum object – including paintings, sculptures, and other forms – such as dating and attribution. They are also an integral part of the restoration process, as it is necessary to have complete information about the materials used in the creation of the art object, as well as any later additions or restoration interventions. Due to the uniqueness of these objects and the challenges in organizing their transportation between museums and laboratories, the development of nondestructive field analysis methods is especially important; ideally, the object does not need to be moved at all.

This report discusses the suggested in literature nondestructive methods for analyzing art pigments, focusing on the application of near-infrared spectroscopy and chemometrics for this purpose, particularly with regard to ultramarine. Ultramarine is a bright blue pigment used in visual arts since the 6th century. The natural pigment, extracted from minerals, was historically available only to recognized master painters due to its high cost. Industrial

production of synthetic ultramarine began in the 1830s, making its use more widespread. Consequently, determining the origin of ultramarine can help identify late undocumented restorations and, importantly, detect forgeries made with anachronistic dyes. Existing methods for analyzing ultramarine pigments – such as polarization microscopy or mid-infrared spectroscopy – require taking small paint samples from the surface. Energy-dispersive X-ray spectroscopy offers a nondestructive alternative, but its application is limited to determining the elemental composition of the sample, and its sensitivity is insufficient for classifying natural and synthetic dyes, as compositional differences are minor.

This study presents classification results of various ultramarine pigments based on their near-infrared (NIR) spectra, measured from both model samples and real art objects of diverse provenances. The NIR measurements were conducted using a 1.8 mm diameter fiber optic probe (Optofiber, Russia) connected to a portable NIR spectrometer (Avantes, Netherlands) covering the spectral range of 939–1799 nm. Multivariate analysis of the spectral data confirms that this approach enables classification of paints both by the type of binder and by the origin of the ultramarine pigment.

References

[1] M. Bacci, C. Cucci, E. Del Federico, et al. *Vib. Spec.* **49** (2009) 80–83.

T06. Advanced signal processing and modeling methods in spectral analysis of complex samples

X.H. Bian[1], W.B. Yang[2], L.P. Yang[2], Javaria Kousar[2], Y.J. Yan[2]

[1]*School of Pharmaceutical Sciences, Tiangong University, Tianjin, 300387, China*
[2]*School of Chemical Engineering and Technology, Tiangong University, Tianjin, 300387, China*

Spectral analysis combined with chemometrics has been widely used in food, medicine, petrochemical and other fields because of its rapid and non-destructive characteristics [1]. However, the spectral signal is usually affected by noise, background and redundant variables. Our group focus on data grouping [2], spectral preprocessing, variable selection and modeling methods in recent years.

For spectral denoising, VMD is introduced for Raman for Raman spectral denoising [3]. Furthermore, in order to solve information loss problem for the sharp peak after VMD denoising, a based on peak extraction VMD (PE-VMD) is proposed spectral denoising [4].

For variable selection, swarm intelligence (SI) algorithms have attracted more and more attention due to their simple structure, efficiency and powerful ability. Our group introduced grey wolf optimizer (GWO), whale optimization algorithm (WOA), butterfly optimization algorithm (BOA)[5], wild horse optimizer (WHO)[6] and beluga optimization algorithm (BOA)[7] for spectral variable selection of complex samples .

For ensemble modeling, based on the advantages of VMD, VMD unfolded PLS (VMD-UPLS)[8] and VMD unfolded ELM (VMD-UELM)[9] method were proposed. An ensemble modeling method based on Monte Carlo (MC), whale optimization algorithm (WOA) and ELM (MC-WOA-ELM)[10] was proposed for NIR spectral quantitative analysis of binary and ternary adulterated *Angelicae Sinensis Radix* (ASR) samples.

References

[1] Y. Liu, L.W. Zhang, X.Z. Zhang, et al. *Microchem. J.* **213** (2025) 113605.
[2] X.H. Bian, W.B. Yang, K.X. Zhang , et al. *Chemom. Intell. Lab. Syst.* **265** (2025) 105493.
[3] X.H. Bian, Z.T. Shi, Y.J. Shao, et al. *Molecules* **28 (17)** (2023) 6406.
[4] S.M. Lu, Y. Hao, Z.T. Shi, et al. Chin. *J. Anal. Chem.* **52 (9)** (2024)1275-1284.

[5] X.H. Bian, Z.Z. Zhao, J.W. Liu, et al. *Anal. Methods* **15** (2023) 5190-5198.
[6] S.H. Wei, Y.J. Yan, R.X. Wang, et al. *Appl. Sci.* **15** (2025).
[7] Javaria Kousar, Liping Yang, Jiale Xiang, et al. *Foods,* **14** (2025) 4266.
[8] D.Y. Wu, J.B. Johnson, K. Zhang, et al. *Microchem. J.* **196** (2024)109587.
[9] L.L. Shen, J.J. Zhao, D.Y. Wu, et al. *Spectrochim. Acta* (2025)126354.
[10] X.H. Bian, Y.X. Liu, J.Q. Xie, et al. *J. Appl. Res. Med. Aromat. Plants* **46** (2025)100640.

T07. Like a Matryoshka: Nested Interfaces from R to Python to RDKit

A. Samokhin [1,2], M. Khrisanfov [1,2]

[1] *Lomonosov Moscow State University, Moscow, Russia*
[2] *IPCE RAS, Moscow, Russia*

Chemoinformatics frameworks are essential for representing molecular structures in digital form and enabling their use in applications such as QSAR/QSPR modeling, deep-learning based property prediction, and virtual screening. Several open-source chemoinformatics toolkits are widely available, including CDK, Indigo, Open Babel, and RDKit. While all of them supports basic operations such as conversion between common molecular representations (SDF, SMILES, InChI), more advanced capabilities – like maximum common substructure search or chemical reaction handling – are often absent or incomplete in some frameworks.

The Python ecosystem currently dominates modern chemoinformatics, and all the toolkits mentioned above provide Python interfaces. Some of these bindings are officially developed and maintained, while others are community-driven. Among them, RDKit has become the *de facto* standard, serving as the backend of most chemistry-related computational and machine learning workflows.

Within the R ecosystem, chemoinformatics functionality is provided by packages such as rcdk (for CDK) and ChemmineR or ChemmineOB (for Open Babel). However, direct access to RDKit from R remains limited. The available R interface on GitHub is outdated and offers only partial functionality. Alternatively, RDKit can be accessed from R through the reticulate package, which enables calling Python functions from R scripts. While this approach is practical for exploratory or small-scale studies, it is inconvenient for routine use and poorly suited to large-scale, resource-intensive tasks, where an error caused by an unusual molecule can halt long-running computations.

To address this gap, we have begun developing a new R interface to RDKit as part of a private R package. Our R interface uses existing Python API instead of the low-level C++ core. Although this layered architecture may seem less direct, it offers important advantages in terms of simplicity, reliability, and maintainability. Preliminary use of the package in ongoing research has demonstrated its high usability and seamless integration into existing workflows. Benchmark tests have also shown excellent performance, with only minimal overhead compared to the native Python interface. Encouraged by these results, we plan to make the project publicly available to expand access to advanced chemoinformatics tools within the R community.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (No. 124041900012-4).

T08. A Generalized Classical Least Squares Approach for Calibration of Linear and Nonlinear Models with Partial Analyte Information

Hamid Abdollahi [a], Alejandro C. Olivieri [b], Nematollah Omidiki [c], Hamideh Bakhshi [a],

[a] Department of Chemistry, Institute for Advanced Studies in Basic Sciences (IASBS), 444 Prof Yousef Sobouti Blvd, Zanjan 45137-66731, Iran
[b] Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmac´euticas, Universidad Nacional de Rosario, Instituto de Química Rosario (CONICET- UNR), Suipacha 531 (2000) Rosario, Argentina
[c] NIOZ Royal Netherlands Institute for Sea Research, Department of Marine Microbiology & Biogeochemistry, 't Horntje (Texel), the Netherlands.

Multivariate calibration is a powerful analytical strategy that models the relationship between samples and instrumental responses using multiple variables rather than relying on a single measurement. This approach is particularly advantageous for complex systems containing potential interferents. While univariate calibration requires the removal, separation, or masking of interfering components, multivariate calibration inherently compensates for their effects through mathematical modeling, provided that the calibration set properly represents their presence. The Classical Least Squares (CLS) method is the simplest direct multivariate calibration model. However, it is commonly assumed that constructing a valid CLS model requires complete concentration information for all components in the calibration samples. In this work, we demonstrate that the generalized form of this method, Generalized CLS (GCLS), can successfully build a reliable calibration model even when only partial compositional information is available; specifically, when only the analyte concentration is known. The resulting model is capable of accurately predicting analyte concentrations in unknown samples.

Furthermore, previous studies have shown that projecting nonlinear raw data onto a space spanned by nonlinear functions, such as Gaussian basis functions, can lead to transformed data that exhibit a linear relationship with analyte concentrations [1]. Under these conditions, a GCLS model can be constructed using the projected data and subsequently applied to accurately predict analyte concentrations in unknown samples. This concept opens new possibilities for applying direct calibration methods to systems exhibiting nonlinear behavior.

References

[1] Allegrini, F., Olivieri, A. C. *Talanta Open*. 7 (2023) 100235.

T09. QSPR-Driven Optimization of Sensitive Elements for Nanophotonic VOC Sensor Systems

A. Surkova[1], S. Domarev[1], A. Boltenko[1], M. Saveliev[2], E. Boichenko[2], A.O. Orlova[1]

[1] International Laboratory "Hybrid Nanostructures for Biomedicine", PhysNano Department ITMO University, Saint Petersburg, Russia.

[2] Chemical Engineering Center, ITMO University, Saint Petersburg, Russia

Diagnosing various diseases through exhaled breath is becoming a popular direction in modern medicine. Human exhaled breath contains a large number of volatile organic compounds (VOCs), deviations in whose concentration from reference values can serve as an indicator of potential diseases, such as cancer, diabetes, liver and kidney diseases, etc. [1]. One

promising direction in creating diagnostic devices is the development of gas sensors that enable quick, non-invasive measurements without additional sample preparation [2].

For diagnostic purposes, it is insufficient to determine the concentration of just one VOC. It is much more important to determine the concentration pattern of several VOCs, since it is precisely the combination of specific substances that often indicates the presence of a pathology. This task is best addressed by sensor systems such as the e-nose, which consist of arrays of cross-sensitive sensors. In recent years, various nanomaterials, for example, metal oxide semiconductors, carbon nanotubes, and quantum semiconductor nanocrystals, have been actively used as sensing elements in such arrays [3]. The response mechanism of these sensors is associated with a change in the conductivity of the sensing material upon the sorption of molecules from the gas phase onto its surface. Changing the type of nanocrystal, its synthesis methods, and other parameters affects the material's sensitivity. The experimental screening for materials with the required properties is highly laborious and time-consuming.

In this work, we propose to use QSPR modeling to study the functional relationship: material parameters-sensor properties, and to select optimal sensing elements. For this purpose, we selected 20 different nanocrystals (such as CdSe, AIS/ZnS, CdTe) and described them with a set of descriptors representing material characteristics (bandgap width, chemical composition, lattice type, size, etc.). The predicted property was the electrical conductivity of the nanostructures in response to acetonitrile and isopropanol. Regression models were built on this dataset using PLS (Partial Least Squares), enabling the prediction of sensor response (for example, for isopropanol, the variation range was from -6 to 4 Δσ,%, with an RMSEP = 1.8 Δσ,%) to VOCs based on the set of descriptors. An analysis of the regression coefficients for the obtained models was performed, which allowed for the identification of the most significant descriptors.

The obtained results open up new possibilities for the rapid and effective selection of gas sensing elements with the required characteristics, based on constructing QSPR models from small datasets. These sensors can subsequently be utilized in medical applications.

This work was financially supported by the ITMO Fellowship Program.

References

[1] T. Chen, T. Liu, T. Li, H. Zhao, Q. Chen, *Clin. Chim. Acta* **515** (2021) 61–72.
[2] M. Kaloumenou, E. Skotadis, N. Lagopati, et al., *Sensors* **22** (2022) 1238.
[3] X. Zhou, Z. Xue, X. Chen, et al., *J. Mater. Chem. B* **8** (2020) 3231–3248.

T10. Quantification of Asphaltenes in Toluene Using Impedance and Infrared Spectroscopy Data

D.M. Bikmeev [1,2], A.Kh. Bikmeeva [1], D.I. Dubrovsky [2], A.D. Badikova [1]

[1] Ufa State Petroleum Technological University, Department of Physical and Organic Chemistry, 1, Kosmonavtov Str., 450064, Ufa, Russia
[2] RN-BashNIPIneft LLC (Company of Rosneft Group), 86/1, Lenina Str., 450006, Ufa, Russia

Asphaltene–toluene solutions offer a convenient model system for investigating the behavior of heavy oil fractions. Such systems allow for controlled study of composition, aggregation, and physical property changes with increasing concentration. This work demonstrates the potential of chemometric processing of impedance and infrared (IR) spectroscopy data to quantify asphaltene content and detect nonlinear effects.

Impedance spectra were recorded in the 100 Hz to 50 kHz range at 400 mV amplitude under open-circuit conditions. For each concentration (0-10 wt%), three independent series

were acquired. PLS models were built on 30 spectra (five per concentration from the first series), and tested on 50 spectra from the second and third series (2-10 wt%).

Feature vectors were constructed from Re(Z), Im(Z), |Z|, and combined Re+Im data (200 points). PLS models for all input types yielded similar results: RMSEP = 0.79-0.87 wt%, R^2 = 0.961-0.969.

Principal Component Analysis (PCA) visualized composition changes and revealed nonlinearity above 8 wt%. Samples at 10 wt% partially overlapped with lower concentrations, likely due to asphaltene aggregation beyond a critical threshold.

FTIR spectra (range 400-4500 cm^{-1}, resolution 4.0 cm^{-1}) were obtained for the same solutions. PLS models based on cross-validation showed comparable accuracy (RMSECV = 0.74 wt%, R^2 = 0.954) and mirrored the nonlinear trend at 10 wt%.

The results confirm the applicability of both impedance and IR spectroscopy for quantitative analysis of asphaltenes. Impedance provides sensitivity to electrochemical changes, while IR data offer robust spectral validation. Their combined use holds promise for rapid monitoring of stability in complex petroleum systems.

T11. Multidimensional liquid sensing via impedimetric virtual sensor arrays

A. Rudnitskaya
CESAM & Chemistry Dept., University of Aveiro, Aveiro, Portugal

The concept of "virtual sensor arrays" has recently emerged within chemical sensor array research. Initially explored for gas sensing, virtual sensor arrays employ one or a small number of sensing elements whose sensitivity characteristics are modulated by operational parameters such as temperature, illumination, or applied potential. In liquid sensing, multidimensional responses from a single sensor can be generated using dynamic signals in a flow-injection set-up or using electrochemical impedance spectroscopy (EIS). Among transduction techniques, EIS is particularly attractive as a powerful method for probing interactions at the sample/electrode interface. Impedance is highly sensitive because it probes a wide range of timescales, as reflected by the broad span of ac frequencies typically applied. Consequently, the spectrum captures both subtle and pronounced changes at the electrode surface, which is information rich but may also complicate data interpretation. EIS has found numerous applications in biosensing in combination with enzymes, antibodies, DNA, and cells [1]. The use of the whole impedance spectrum for data analysis led to the concept of a one-sensor impedimetric tongue system [2].

Data processing of the impedimetric virtual sensor array may involve modelling of the equivalent electric circuit and using its parameters for model calculation or applying multivariate calibration techniques to the full spectrum. Both approaches will be illustrated using two applications of impedimetric sensor arrays to liquid analysis: detection of marine toxins and detection of acrylamide. In the first case, an enzymatic impedimetric tongue for paralytic shellfish toxins couples EIS with a carbamoylase assay to monitor enzyme conformational changes and resulting adsorption phenomena at the electrode surface [3]. In the second, acrylamide is detected through its polymerization and subsequent adsorption on screen-printed carbon nanotube electrodes.

Acknowledgments
This work was funded by national funds through FCT – Fundação para a Ciência e a Tecnologia I.P., under the project CESAM – Centre for Environmental and Marine Studies, ref. UID/50017/2025 (doi.org/10.54499/UID/50017/2025) and LA/P/0094/2020 (doi.org/10.54499/LA/P/0094/2020); and under the project MISSION (Water4All/0006/2023) within the European Union's Horizon Europe

Programme under the 2023 Joint Transnational Call of the European Partnership Water4All (https://doi.org/10.54499/Water4All/0006/2023).

References

[1] Brett CMA. *Molecules* **27(5)** (2022) 1497.
[2] Rodrigues DR, Fragoso WD, Lemos SG. *Electrochim Acta*. **397** (2021) 139312.
[3] Raposo M, Soreto S, Moreirinha C, et al. *Anal. Bioanal. Chem.* **416** (2024) 1983–1995.

T12. Group-targeted prediction of substituted flavone retention using machine learning

A. Stavrianidi[1,2], I. Rozanov[1,2], A. Buryak[1,2]
[1]*Frumkin Institute of Physical Chemistry and Electrochemistry Russian Academy of Sciences, 31-4, Leninsky prospect, 119071, Moscow, Russia*
[2]*Chemistry Department, Lomonosov Moscow State University, 1/3, Leninskie Gory, GSP-1, 119991 Moscow, Russia*

Predicting the elution order of close structural analogs and isomers is a critical step in plant metabolite dereplication. Machine learning (ML) is an effective approach to automate peak annotation by finding structure-retention relationships (QSRR). In this study, four ML models were trained to predict the elution order of substituted flavone derivatives.

Flavonoids have a core 15-carbon flavone skeleton consisting of two benzene rings (A and B) connected by a three-carbon pyran ring (C). The structure of the ring system, along with the number and position of hydroxyl groups and other substituents, influences the biological activity of flavonoids [1]. Molecular topological fingerprints represent a molecule as a set of structural features and are often used to search for similarities, identify substructures, and model structure-activity relationships [2]. Chromatographic retention can also be predicted using this approach. For similar compounds, such as steroids [3] or flavonoids, an effective way to identify their structural differences is to determine a fingerprint containing information about substituents at different positions. Therefore, a specific molecular fingerprint for flavone-based structures was developed. Neural network-based (NN) models used a simplified (20-bit) version of the fingerprint, while logistic regression models were tested using both a compressed (20-bit) and an extended (92-bit) fingerprint, which included interactions between substituents. Pairwise elution order prediction errors were generally below 10%, demonstrating robust performance under varied reversed-phase LC conditions with an acetonitrile gradient in the mobile phase. In case of isomeric pairs, this error was mostly below 15%. Furthermore, retention times were indirectly estimated from model scores using linear regression and linear interpolation with the mean absolute error close to 1 min.

The study was divided into two parts: an initial database (IDB) containing retention times of over 50 compounds under a linear acetonitrile gradient and independent literature data (ILD) from several sources were used to train the models. In both cases, weights analysis in logistic regression and NN models revealed similar effects of substituents on retention, confirming the validity of the models. The observed retention phenomena were in accordance with those already described in the literature. Although overall performance indicators were comparable, the use of a large experimental dataset or database was preferable compared to fragmented literature data. The proposed approach can be adapted for the use in identifying other groups of plant secondary metabolites with an increased number of substituents.

Acknowledgements
The research was supported by the Russian Science Foundation (Grant № 22-13-00266-P) for Frumkin Institute of Physical Chemistry and Electrochemistry Russian Academy of Sciences.

References
[1] M.C. Dias, D.C.G.A. A.M.S. Pinto, *Molecules*. **26** (2021) Article 5377.
[2] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **50** (2010) 742–754.
[3] G.M. Randazzo, D. Tonoli, S. Hambye, *Anal. Chim. Acta.* **916** (2016) 8–16.

T13. Rapid detection of ethylene glycol (EG) & diethylene glycol (DEG) in propylene glycol using portable NIR Spectroscopy and Chemometrics

Dilip Sing[1, 2]*, Deba Prasad Kanshari[2], Surojyoti Mandal[2], Uday Pratap Sing[2], Soumitra DasMahapatra[2], Ajanto Kumar Hazarika[3], and Rajib Bandyoypadhyay[1,2],
[1] *Department of Instrumentation and Electronics Engineering, Jadavpur University, Salt Lake Campus, Kolkata 700106, India*
[2] *MetaspeQ Division, Ayudyog Private limited, Kolkata-700040, West Bengal, India.*
[3] *Tocklai Tea Research Institute, Tea Research Association, Jorhat 785008, Assam, India.*

Propylene glycol (PG) is a widely used pharmaceutical excipient, yet its safety has been compromised repeatedly by adulteration with ethylene glycol (EG) and diethylene glycol (DEG); both highly toxic industrial solvents implicated in several global, including major Indian, mass-poisoning incidents. [1]. Conventional analytical methods such as GC, GC–MS, and HPLC, though accurate, are labor-intensive, destructive, and unsuitable for rapid, on-site screening of raw materials [2]. This study aims to establish a fast, non-destructive, and cost-effective analytical approach for the simultaneous quantification of EG and DEG in raw PG using portable Near-Infrared (NIR) spectroscopy coupled with chemometric modelling. NIR spectral data (900–1700 nm) were collected using a portable handheld spectrometer, and samples were prepared by spiking pharmaceutical-grade PG with known concentrations of EG and DEG across a wide range of contamination levels. To enhance signal quality, multiple preprocessing techniques—including SNV, unit vector normalization, area and maximum normalization, Savitzky–Golay smoothing, and MSC—were evaluated. Six regression algorithms—PLS, SVR, Random Forest (RF), K-Nearest Neighbor (KNN), Gaussian Process Regression (GPR), and XGBoost—were trained for simultaneous estimation of EG and DEG concentrations. Reference values were obtained using validated chromatographic methods, and model accuracy was evaluated using R^2 and RMSE metrics for both calibration and validation sets. The developed chemometric models demonstrated strong predictive capability, achieving high coefficients of determination and low prediction errors for both analytes. The findings demonstrate that portable NIR spectroscopy, complemented by robust preprocessing and advanced machine-learning regressors, can detect and quantify EG and DEG in PG within seconds. This approach offers a powerful screening tool for pharmaceutical quality control and has strong potential for preventing contamination-related health hazards, ensuring regulatory compliance, and supporting real-time raw material authentication in industrial settings.

References
1. Schier, Joshua G., et al. *Journal of medical toxicology* **7** (2011) 33-38.
2. Sing, Dilip, et al. *Frontiers in Pharmacology* **12** (2021) 629833.
3. Sing, D, et al. *Microchemical Journal*. **199** (2024) 109949

T14. Estimation of cup characteristics of CTC black tea liquor using NIR spectroscopy and chemometric tools

A. K. Hazarika[1], D. Sing[2,3], A. Mahtab[3], S. Banerjee[3], P. Sharma[1], S. S. Selvam[4], R. Bandyopadhyay[2]

[1] *Tocklai Tea Research Institute, Jorhat, Assam, India*

[2] *Department of Instrumentation and Electronics Engineering, Jadavpur University, Kolkata, India*
[3] *MetaspeQ Division, Ayudyog Private Limited, Kolkata, West Bengal, India*
[4] *Tea Academy, Siliguri, West Bengal, Kolkata, India*

Rapid and objective evaluation of cup-quality attributes is becoming increasingly vital for the CTC black tea industry, where traditional sensory assessment of brightness, briskness, strength, colour, and overall liquoring quality relies on expert tasters and is prone to inter-taster variability [1]. This study proposes a comprehensive chemometric framework for predicting these key sensory attributes using near-infrared (NIR) spectral signatures of CTC black tea liquor from Assam, Dooars and Terai regions of India.

The modelling workflow employed spectral preprocessing—Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), and Savitzky–Golay derivative filtering to correct baseline variations, and enhance subtle spectral features [2]. Partial Least Squares Regression (PLSR) yielded acceptable predictive performance; however, nonlinear modelling approaches—particularly Support Vector Regression (SVR) and Random Forest Regression (RFR)—demonstrated markedly superior accuracy for attributes such as brightness and briskness. Their advantage arises from an enhanced ability to capture complex nonlinear interactions between NIR spectral features and the corresponding sensory scores [3].

The developed models exhibited strong quantitative reliability, with high coefficients of determination (R^2), elevated Residual Predictive Deviation (RPD) values, and low Root Mean Square Error (RMSE) across all target attributes. These findings demonstrate that NIR-based chemometric modelling can estimate major cup-quality characteristics of CTC black tea liquor. The approach offers a rapid, objective, and scalable analytical tool capable of augmenting traditional sensory evaluation and enabling real-time quality monitoring in tea manufacturing operations.

References

1. S. S. Turgut, J. A. Entrenas, E. Taşkın, et al. *Food Control.* **142** (2022) 109260.
2. A. K. Hazarika, D. Sing, S. Naskar, et al, *International Conference on NIR Spectroscopy*. Rome, Italy, 8–12 June 2025.
3. M. Z. Zhu, B. Wen, H. Wu, et al, *Journal of Spectroscopy*. **2019** (2019) 9832510.

T15. Regularisation in PARAFAC: better living through sparsity?

I. Krylov [1], T. Labutin [1], A. Drozdova [2]

[1]*Lomonosov Moscow State University, Moscow, Russia*
[2]*Shirshov Institute of Oceanology, Moscow, Russia*

Parallel factor analysis (PARAFAC) is a well-established data decomposition method in fluorescence excitation-emission (FEEM) spectroscopy and other areas where the multi-linearity assumption is applicable. Its second-order advantage makes it possible to work without a separate regressor dataset, obtaining solutions without rotation ambiguity. Since the method is unsupervised, its validation options are limited to fitting subsets of the original dataset and comparing the resulting models: examples range from jack-knifing to bootstrap, with split-half analysis [1] being the typical option.

Occasionally in environmental analysis, a small number of samples contain unique components not encountered in most of the dataset. Those should not be dismissed out of hand as failures of the experiment: they could also be evidence of one-time or very irregular pollution events. On the other hand, the unique nature of these samples makes them outliers from the statistical viewpoint, complicating validation of the PARAFAC models containing

them. Out of all possible ways to split a dataset, many won't contain the unique component in both models being compared, failing the validation test.

It is possible to incorporate all samples into the model and validate it at the same time by formulating the PARAFAC decomposition of a FEEM data cube $\mathcal{X}$ as a constrained regularised inverse problem [2]:

$$\min_{\boldsymbol{\lambda},\,\mathbf{A},\,\mathbf{B},\,\mathbf{C}} \left\| \mathcal{X} - (\mathbf{B} \otimes \mathbf{A})\mathrm{diag}(\boldsymbol{\lambda})\mathbf{C}^{\mathrm{T}} \right\|_2^2 + \kappa \|\boldsymbol{\lambda}\|_1 \text{ s. t. } \|\mathbf{A}\| = \|\mathbf{B}\| = \|\mathbf{C}\| = \text{const}$$

Here, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are the PARAFAC scores and loadings being sought, $\boldsymbol{\lambda}$ is a vector of per-component scales that's being made sparse in order to limit the number of components, $\otimes$ is the Khatri-Rao product, and κ is the regularisation coefficient. The problem can be solved using manifold optimisation methods, and its validation can be performed by means of L-curves.

This method has been applied to fit a 4-component model, incorporating an extra humic-like and a chlorophyll-like factor, to a dataset where split-half only validated a 2-component one.

References

[1] I. Krylov, A. Drozdova, T. Labutin. *Chemom. Intell. Lab. Syst.* **207** (2020) 104176.
[2] I. Krylov, O. Erina, A. Drozdova, et al. *J Appl Spectrosc.* **91** (2024) 1065–1071.

T16. Geochemical production allocation in commingled wells using chromatography and chemometric tools

M. Saveliev, D. Yakovlev, Y. Susanov, V. Panchuk, D. Kirsanov, V. Kurenkov

Advanced Engineering school SPBU, Saint Petersburg, Russia

Currently, geochemical production allocation for commingled wells represents a pressing challenge in the oil industry. The tools and methods conventionally employed for this purpose are notably costly, time-consuming, resource-intensive, and do not permit operational monitoring. A promising solution to this problem lies in the application of reservoir geochemistry techniques, which combine the advantages of instrumental chemical analysis with operational speed and relative low cost. The reservoir geochemistry methodology employed in this work, based on the construction of PLS models, integrated chromatography of the analyzed crude oil with subsequent chemometric data processing.

This study focuses on determining production shares for commingled wells at one of the Russian oil fields. The field contains both reference wells, producing from a single specific reservoir zone, and commingled production wells, which simultaneously extract oil from different reservoir zones in unknown proportions. The first stage involved oil typing for each reservoir zone. Reference wells for each zone within specific field segments were selected, and crude oil samples were collected. Chromatograms for the obtained samples were recorded using gas chromatography with a flame photometric detector. Principal Component Analysis (PCA) demonstrated that reference wells tapping the same reservoir zone exhibit property variations across different field segments. However, within individual segments, it was possible to discriminate between the reservoir zones based on differences in oil composition. Consequently, the field was subdivided into three segments for model development.

Under laboratory conditions, a calibration set was created by gravimetrically mixing end-member oils to prepare laboratory mixtures ranging from 90%/10% to 10%/90% (with 10% increments) of the two contributing zones. Following chromatographic analysis, specific chromatographic peaks were selected, and their height ratios were calculated. This approach

increased the number of informative variables and reduced their uncertainty. Partial Least Squares (PLS) [1] regression was used to develop the predictive models. Model predictive ability was assessed via cross-validation, with the Root Mean Square Error of Cross-Validation (RMSECV) calculated. For the three field segments, the RMSECV values were 5.1%, 2.3%, and 2.6%, respectively, across the determined fraction range of 0-100%. Additionally, variable importance was assessed by analyzing the relationships between Relative Standard Deviation (RSD) and the R^2 coefficient, and the applicability of models across different field segments was tested. Finally, the developed models were applied to predict the contribution of each reservoir zone in the commingled wells.

References

[1] S, Wold, M. Sjöström, L. Eriksson, *Chem. Intel. Lab*. *Syst*. **58** (2001) 109-130.

T17. Intensity-Invariant Spectral Analysis via Pairwise Intensity Ratios and Accounting for Feature-Dependencies

Abhishek[1], Abdelkhalak Harrak[1], Andreas Seifert[1,2]

[1]*CIC nanoGUNE BRTA, 20018 San Sebastian, Spain*
[2]*Ikerbasque, Basque Foundation for Science, 48009 Bilbao, Spain*

Advanced analysis of spectroscopy data is unlocking new perspectives in biochemical research and holds great promise for medical diagnostics. Current methods involve pre-processing of spectral intensities, such as baseline correction and scaling, followed by dimensionality reduction, feature extraction and classification [1]. In this study, we propose a ratio-based transformation that computes all pairwise ratios of spectral intensities within each spectrum to construct a scale-invariant feature space. This transformation encodes relative relationships between features rather than their absolute magnitudes. Unlike intensity-based methods, our approach puts each spectrum into a relational feature matrix to introduce nonlinear structure into the data. It also identifies dependencies between peaks and provides insights into inter-peak relationships that may carry biological or chemical significance.

In our method, initially the spectra are pre-processed using standard procedures which includes spike removal, background correction and smoothing. The data is then manually tuned for vertical alignment to stabilize pairwise intensity ratios and reduce distortions from small offsets, i.e., noise. Each pre-processed spectrum of n intensity points is then transformed into $n(n-1)/2$ ratio features by dividing each intensity by subsequent values, forming a triangular ratio matrix. A two-stage principal component analysis (PCA) is applied to the transformed high-dimensional data, first to reduce the dimensions and then to extract meaningful components for classification. Both supervised and unsupervised machine learning algorithms were employed to extract informative intensity relations. The performance of multiple classifiers—including supervised methods such as support vector machine, logistic regression, random forest, and boosting algorithm, as well as unsupervised methods such as k-means clustering—was systematically analysed using different cross-validation strategies to assess the discriminative capability. We also mapped the dependencies between different peaks by analysing co-variations among ratio features to identify the spectral regions that change in a particular manner. This ratio-based transformation captures dependencies between spectral peaks and are inherently invariant to scaling, instrumental drifts, and slow varying environmental conditions. Later, the method was validated on two Raman spectroscopy datasets, including 36 cerebrospinal fluid samples (CSF) with 18 Alzheimer's and 18 healthy controls, and 36 plasma samples with 18 lung

cancer and 18 healthy samples. On the CSF dataset, standard methods achieved 67% accuracy while the ratio-based approach reached 75%. On the plasma dataset, standard methods achieved 82–85% accuracy, whereas our method improved it to 91% [2].

This ratio-based transformation opens room for developing new invariant and interpretable chemometric models. Future work may focus on selection of meaningful spectral ratios, combining this ratio-based approach with peak-matching to compute ratios of relative peaks only. The method could also be analysed with deep-learning architectures for nonlinear feature extraction. It has potential for real-time diagnostic systems where invariance to experimental variation and interpretability of peak relationships are critical.

References

[1] Lopez, E., Etxebarria-Elezgarai, J., et al., *Int. J. Mol. Sci.* **25** (2024) 4737.
[2] Hano, H., Lawrie, C.H., Suarez, B., et al., *ACS Omega* **9** (2024) 14084-14091.

T18. Deep Learning for GC–IMS Data Analysis: A Novel Approach to VOC Characterization

Farbod Bayat-Afshary[1], Nima Naderi[1], Philipp Weller[2], Hadi Parastar[1,2]*

[1] *Department of Chemistry, Sharif University of Technology, P.O. Box 11155-9516, Tehran, Iran*
[2] *Institute for Instrumental Analytics and Bioanalytics, Mannheim University of Applied Sciences, 68163, Mannheim, Germany*

In recent years, the analysis of volatile organic compounds (VOCs) has gained increasing attention, especially for complex samples such as biological tissues, food products and environmental materials. This rise is largely due to the advantages of studying compounds in the gas phase, including minimal or no sample preparation [1]. Gas chromatography–ion mobility spectrometry (GC-IMS) has emerged as a powerful tool due to its combination of high-resolution GC and the sensitive, selective capabilities of IMS. However, despite its strengths, GC-IMS also presents data analysis challenges. The two-dimensional heatmap output of GC-IMS introduces analytical challenges that demand advanced data processing. Classical chemometric and machine learning often require significant manual intervention, suffer from preprocessing sensitivity and struggle with high-dimensional data [2, 3]. To overcome these limitations, a novel approach is proposed in this work that integrates computer vision and deep learning (DL) techniques, particularly convolutional neural networks (CNNs), for GC-IMS data analysis. Initially, a simplified CNN model was employed to classify two datasets towards authentication of olive oil (157 samples across 3 classes) [4] and saffron (173 samples across 7 classes) [5]. Only minimal preprocessing (reactant ion peak (RIP) correction and region of interest (ROI) selection) was performed on the data prior to CNN model construction. The simplified model achieved an accuracy of 98.5% for the olive oil dataset; however, it did not reach acceptable classification accuracy for the saffron dataset due to the subtle chemical differences in the chromatographic images and the small dataset size. To address this issue, an image augmentation technique was applied. Using this approach, the enhanced model achieved 98% accuracy on the saffron dataset. Furthermore, saliency maps were utilized to extract features from the images, visualize the model's decision-making process, and identify important regions in the chromatographic images. By superimposing saliency maps onto the original chromatographic heatmaps, one

can visually interpret the spatial importance of different regions within the sample. Furthermore, saliency maps were combined and analyzed using PCA to generate representative maps, highlighting consistent chemical features that contribute to class differentiation across the dataset. This study highlights how advanced DL techniques can revolutionize chromatographic analysis and enhancing feature extraction.

References

[1] H. Parastar, P. Weller, *Trends Anal. Chem.* **170** (2024) 117438.
[2] H. Parastar, P. Weller, *Anal. Chem.* **97 (3)** (2024) 1468-1481.
[3] B. Debus, H. Parastar, P. Harrington, D. Kirsanov, *Trends Anal. Chem.* **145** (2021) 116459.
[4] Gerhardt, N.; Sanders, D.; Rohn, S.; Weller, P. *Anal. Bioanal. Chem.*, **409** (2017) 3933-3942.
[5] Parastar, H.; Yazdanpanah, H.; Weller, P. *Food Chem.* **465** (2025) 142074.

T19. Laser plasma inhomogeneity as revealed by chemometrics

N. Sushkov [1], T. Labutin [1]

[1]*Department of Chemistry, Moscow State University, Moscow, Russia*

Laser-induced breakdown plasma (LIP) is a transient light source which is relatively easy to produce in most materials. The emission of LIP contains characteristic signals of atoms, ions and small molecules present in the plasma. Thus, LIP is broadly used for elemental analysis of materials (which is known as laser-induced breakdown spectroscopy, LIBS) and in non-analytical studies. This practice is complicated by inhomogeneity of the plasma regarding spatial distribution of species and parameters such as temperature and electron number density. The relevant information can be obtained from laborious spatially-resolved observations followed by due mathematical treatment (e.g., reverse Abel transform) [1]. It seems that chemometrics can provide alternative approaches.

We considered a matrix of spectra of diatomic CN radical (violet system, λ = 370–388 nm) obtained in time-resolved LIBS experiments with organic targets, and decomposed it by principal component analysis (PCA) and non-negative matrix factorisation (NMF) algorithms. Among the resulting component loadings, there were quasi-spectra with different shapes of molecular emission bands: one with (0–0) band head dominating the spectrum, and another with more equalised distribution of intensity between heads. This suggested that the components were related to plasma zones with different temperatures: lower (in our case, ~ 4000 K), and higher (~ 7000 K), as far as we could estimate them by fitting the loadings with model spectra [2]. Interestingly, these temperatures were close to electron temperatures obtained by fitting Boltzmann graphs for atomic emission from the same sample, namely, calcium lines with lower excitation energies, and magnesium lines with higher excitation energies, respectively. Thus, matrix decomposition algorithms could be a fast and affordable tool in studying plasma inhomogeneity by spectroscopic methods.

References

[1] A. Rylov, A. Zakuskin, T. Labutin, *Spectrochim. Acta B.* **223** (2025) 107079.
[2] N. Sushkov, N. Lobus, I. Seliverstova, T. Labutin, *Optics and Spectroscopy* **128** (2020) 1343–1349.

T20. Application of generative adversarial networks to improve the accuracy of classification models by generating spectra of underrepresented class of urinary stones

V. Deev [1], E. Boichenko[2], D. Kirsanov [1]
[1]*St Petersburg University, Saint Petersburg, Russia*
[2]*Center of Chemical Engineering, ITMO University, Saint Petersburg, Russia*

Urolithiasis is a common disorder of the genitourinary system. The chemical composition of urinary stones allows to determine the cause and adjust treatment and prevention strategies. The main components are calcium oxalates, phosphates, or uric acid, with their prevalence varying significantly. The most accurate methods for determining chemical composition are X-ray and infrared spectroscopy of stones removed after surgery. Determining the composition during surgery is promising, as this allows for adjustments to stone fragmentation conditions. Near-infrared spectroscopy can be used for this purpose by connecting a waveguide to standard surgical instruments.

Different representation of urinary stone classes leads to decreased accuracy in determining the least represented classes. Obtaining new experimental data requires longer data collection and additional testing. Artificial data augmentation methods may tackle this issue. The literature describes the use of generative adversarial neural networks for generate artificial spectra [1]. The aim of this work is to evaluate the feasibility of using GAN to generate spectra of the least represented class of phosphate urinary stones to improve the accuracy of the classification model.

The set of urinary stones consisted of 157 samples with various chemical compositions (115 oxalates, 28 urates and 14 phosphates). The size of the samples ranged from a few millimeters to centimeters. The chemical composition of urinary stones was determined using the method of X-ray phase analysis with an X-ray diffractometer. NIR spectra were obtained using a portable fiber-optic spectrometer connected to a fiber-optic probe in the spectral range 939-1799 nm with 4 nm resolution 3 times for every sample. The replicated spectra were averaged and min-max normalization was applied.

The parameters and architecture of a generative adversarial neural network for generating NIR spectra were selected. The influence of the number of experimental spectra used for GAN training, number of generated spectra used for SMV training as well as the influence of the method of dividing samples into training and testing sets, were evaluated. 400 artificial spectra were generated based on a subset of experimental phosphate spectra. A support vector machine (SVM) method was used for classification. It was shown that the greatest increase in classification accuracy was observed when generating spectra for a small number of phosphate spectra. As the number of phosphate spectra used for generation increased, classification accuracy increased, but to a lesser extent.

References

[1] Q. Li, Y. Tang, L. Chu, *Expert Syst Appl.* **253** (2024) 124341.

T21. I found the needle in a haystack – and some other things, too

M. Csontos [1,2], J. Elek [1]

[1]*Science Port Ltd. Debrecen, Hungary*
[2]*University of Debrecen, Debrecen, Hungary*

The new generation of food additives and dietary supplements requires animal experiments – typically via the dietary route. Accurate determination of food additive concentrations in rodent diets is critical to ensuring the validity of toxicological research. The exact composition of these biological origin materials is often unknown; moreover, these components show high similarity to rodent diet matrices. Multivariate Near-infrared spectroscopic methods were developed, validated, and used successfully; I found the needle in

the haystack, as I presented at WSC14 in 2024. However, numerous questions were raised. We observed spectral differences between diet batches, which were assumed to be the same; spectral differences of the sample due to pastille pressing were assumed to be the same. An exploratory analysis of the spectroscopic data may help resolve these difficulties and determine their origin.

In this presentation, I show recent results that may answer the questions we faced. Or at least, what else was found in this spectroscopic haystack.

References

[1] S.D. Dimitrov, D.G. Georgieva, T.S. Pavlov, et al. *Environ. Toxicol. Chem.* **34** (2015) 2450-2462.

P01. MOS multisensor based kidney cancer screening via analysis of gas phase above the urine sample using thermocycling mode

M. Saveliev [1,2], M. Grevtsev [2,3], I. Jahatspanian [2]

[1]*ITMO University, Saint Petersburg, Russia*
[2]*Joint-Stock Company, NPO, PRIBOR, Saint Petersburg, Russia*
[3]*Ioffe Institute, Saint Petersburg, Russia*

Timely diagnosis of kidney cancer (KC) is a critical determinant of a patient's full recovery. Currently, the most common and reliable diagnostic methods are biopsy and medical tomography techniques, such as computed tomography (CT) and magnetic resonance imaging (MRI). However, these methods are resource- and time-consuming, making them unsuitable for large-scale early-stage cancer screening. A potential solution to this challenge is the application of multisensor "electronic nose" systems for the analysis of biological fluids. This approach is relatively inexpensive, non-invasive, requires minimal sample preparation, and can be implemented directly in clinical settings. Despite considerable research attention devoted to the use of such systems in clinical diagnostics, significant opportunities remain for the development of advanced analytical methodologies and data processing algorithms.

This study investigates the application of the "ARAMOS-7D" multisensor analyzer, which utilizes an array of seven metal-oxide semiconductor (MOS) sensors. Urine samples were collected from two patient groups: the first group comprised 40 patients with diagnosed kidney cancer, while the second group consisted of 33 healthy patients serving as control group. Each sample was analyzed in triplicate using the multisensor system according to the developed methodology. Sensor responses were recorded in thermal cycling mode, with a linear temperature ramp from 150 to 450 °C. This procedure yielded seven thermal cycles per sample, which were concatenated into a unified signal sequence termed a "thermospectrum." Following preprocessing, the collected data were used to construct binary classification models employing Random Forest (RF) [1], k-Nearest Neighbors (k-NN) [2], Soft Independent Modelling of Class Analogy (SIMCA) [3], Partial Least Squares Discriminant Analysis (PLS-DA) [4], and Support Vector Machines (SVM) [5]. Additionally, an ensemble model was proposed that aggregates predictions from the aforementioned individual models. To ensure statistical robustness, the initial dataset was iteratively split 100 times into training (67%) and testing (33%) subsets. The performance of the developed models was evaluated using accuracy, sensitivity, and specificity metrics, which consistently exceeded 85% on average across all iterations.

References

[1] L. Breiman, *Mach. Learn.* **45** (2001) 5-32
[2] T. Cover, P. Hart, *IEEE Trans. Inf. Th.* **13** (1967) 21-27
[3] S. Wold, M. Sjöström, C*hem.: Th. Appl.* **52** (1977) 243-282

[4] M. Barker, W. Rayens, *J. Chem.* **17** (2003) 166-173
[5] C. Cortes, V. Vapnik, *Mach. Learn*. **20** (1995) 273-279

P02. Voltammetry multisensory system for quality control of mineral water

Ch. Mukhametdinov[1], R. Zilberg[1], A. Saveleva[1]

[1]*Ufa university of science and technology, Ufa, Russia*

Quality control of bottled mineral water, both at the production stage and at the product sales stage, is an urgent and essential part of product safety. Nowadays, the chromatographic and spectral methods used need a wide set of expensive instrumentation and multidisciplinary highly qualified specialists. The proposed voltammetric multisensory system provides an affordable, simple solution with the ability to perform rapid, integral assessment of the quality of the analyzed samples [1-4]. Sensors based on glass-carbon electrodes modified with composites of a polyelectrolyte complex of chitosan with single-walled carbon nanotubes, reduced graphene oxide, and gold nanoparticles were used to impart cross-sensitivity to the sensor system. 11 samples of mineral waters from different manufacturers, presented in retail trade, different in total mineralization and ionic composition, were taken as analytes. All measurements were registered in cyclic voltammetry mode. Chemometric processing of voltammetric data using PCA and SIMCA-classification methods was used to identify and classify mineral waters. The PCA modeling score plots shows a clear division into separate, non-overlapping clusters of mineral water samples by manufacturer, with the placement of samples along the PC1 correlating with total mineralization and along the PC2 with the qualitative composition of the water. The SIMCA classification shows that type I and type II errors don't exceed 12.5%, which shows that the multisensory system can identify mineral waters by brand and classify samples based on their nature.

The work was supported by the Russian Science Foundation (grant no. 23-73-00119)
https://rscf.ru/project/23-73-00119/

References

[1] A. V. Sidelnikov, R. A. Zilberg, F. Kh. Kudasheva, et al. *J. Anal. Chem.* **63** (2008) 1072–1078
[2] Y. A. Yarkaeva, D. I. Dubrovskii, R. A. Zil'berg, et al. *Russ. J. Electrochem.* **56** (2020) 544-555.
[3] R. Zilberg, E. Bulysheva, Y. Teres, et al. *Chim. Techno Acta* **12** (2025) 12204.
[4] R. A. Zilberg, J. B. Teres, E. O. Bulysheva, et al. *Electrochimica Acta* **492** (2024) 144334.

P03. Chemometric Analysis of Drill Cuttings Composition: From Spectra to Prediction of Trace Elements

A.A. Nikolaev [1], D.M. Bikmeev [1,2], A.A. Ishkov [3], A.V. Levin [3]

[1] *RN-BashNIPIneft LLC (Company of Rosneft Group), 86/1, Lenina Str., 450006, Ufa, Russia*
[2] *Ufa State Petroleum Technological University, Department of Physical and Organic Chemistry, 1, Kosmonavtov Str., 450064, Ufa, Russia, e-mail: bikmeev@gmail.com*
[3] *RN-KrasnoyarskNIPIneft LLC (Company of Rosneft Group), 65d, 9 Maya St., Krasnoyarsk, 660098, Russia*

A comprehensive methodology for the quantitative analysis of major rock-forming oxides and trace components in drill cuttings using WDXRF spectroscopy has been developed

and implemented. Special emphasis was placed on interpreting inter-element relationships via chemometric techniques.

Correlation analysis revealed stable dependencies between major oxides and trace elements. Principal component analysis enabled the visualization of key geochemical trends and synchronous variations in elemental content throughout the vertical profile. Cosine similarity was employed to evaluate relationships between loading and score vectors, facilitating interpretation of element clustering and sample grouping. Distinct groups were identified: vanadium, iron, manganese with silicate components; cerium with calcium, magnesium, and strontium; rubidium and niobium with halite.

PLS regression was used to predict trace element concentrations from the contents of rock-forming oxides. The resulting models demonstrated high accuracy (R^2 up to 0.931), with RMSECV ranging from 0.74 to 5.95 ppm and relative errors from 1.8 to 35.2%, as assessed via cross-validation. These results indicate the potential for application in rapid, routine analysis.

The methodology was tested under production laboratory conditions and showed high reproducibility of spectral data (mean deviation <3%), confirmed by multi-level quality control including instrument stability checks and inter-operator testing. The proposed approach enables further automation and integration into geochemical monitoring systems during drilling operations.

P04. Two Types of "Electronic Nose" — Two Approaches to Multivariate Data Processing

A. Shaposhnik [1], P. Moskalev [2], A. Vasiliev[3], M. Kulikov[4], S. Ryabtsev[4]
[1] *Department of Chemistry, Voronezh State Agrarian University, 394087 Voronezh, Russia*
[2] *Department of Applied Mathematics, Moscow State University of Technology "STANKIN", 127994 Moscow, Russia*
[3] *Laboratory of Sensor Systems, Dubna State University, 141980 Dubna, Russia*
[4] *Department of Physics, Voronezh State University, 394018 Voronezh, Russia*

The concept of an "electronic nose" is based on generating unique "fingerprints" of gas mixtures for their subsequent identification. The traditional approach utilizes an array of *n* chemiresistive sensors. In this case, the multivariate system response is a set of electrical resistance values {R_1, R_2, ..., R_n} measured in a quasi-stationary mode. Standard multivariate methods such as Principal Component Analysis (PCA), Partial Least Squares (PLS) or Artificial Neural Networks (ANN) are successfully applied for the selective analysis of such data. However, this approach contains an inherent contradiction: while each individual sensor is a simple and low-cost miniature device, the need for complex multivariate data processing requires computer resources, which offsets the advantages of the sensors and increases the cost of the entire gas analyzer.

An alternative paradigm involves the creation of an "electronic nose" based on a single sensor operating in a dynamic, non-stationary mode, for instance, under programmed temperature modulation. Here, the multivariate signal is generated as a sequence of n resistance measurements over a single measurement cycle, also forming a set {R_n}. The key difference lies in the nature of these data: the resistance values constitute a correlated time series, rather than a set of disparate points from different sensors. This opens prospects for developing specialized methods for processing correlated data that do not require significant computational resources.

Three promising directions for such processing can be identified. The first is based on an investigation of the kinetics of chemisorption and surface reactions governing the sensor's response dynamics. The second option is applicable when the time-dependent resistance

curve exhibits pronounced extrema characteristic of a specific gas. Determining the coordinates of these extrema allows for solving both qualitative and quantitative analysis tasks. The third, most universal option, is the application of our specially developed algorithm, "Multivariate Calibration for Selective Analysis (MCSA)," which effectively combines the accuracy of multivariate calibration with computational efficiency.

References

[1] A. Shaposhnik, P. Moskalev, A. Vasiliev et al. *Sensors*. **19** (2019) 1135
[2] A. Vasiliev, A. Shaposhnik, P. Moskalev et al. *Sensors.* **23** (2023) 3730.
[3] A. Shaposhnik, P. Moskalev, A. Vasiliev et al. *Int. J. Hydrog. Energy*. **82** (2024) 523.
[4] A. Shaposhnik, P. Moskalev, A. Vasiliev et al. *Chemosensors.* **13** (2025) 323.

P05. Intrinsic radioactivity in spent nuclear fuel – the way to X-ray fluorescence measurement for element quantification

V. Panchuk[1], Y. Petrov[2], E. Lisovskaya[2], V. Semenov[3], D. Kirsanov[3]

[1]*Institute for Analytical Instrumentation RAS, St. Petersburg, Russia*
[2]*Khlopin Radium Institute, St. Petersburg, Russia*
[3]*Institute of Chemistry, St. Petersburg University, St. Petersburg, Russia*

Quantification of chemical elements in spent nuclear fuel (SNF) during the reprocessing is an important analytical task. The knowledge on the chemical composition is required to ensure a safe and effective process run. As such, SNF reprocessing solutions are very challenging objects for chemical analysis. This is due to their very high radioactivity and related personnel safety requirements, very strong acidity and simultaneous presence of numerous elements with very diverse chemical properties. There is an urgent need for the methods that would provide for remote on-line quantification of elements in spent nuclear fuel and its reprocessing technological solutions.

Here we demonstrated that the intrinsic radioactivity of spent nuclear fuel samples can be employed as the source of X–ray fluorescence excitation for analytical purposes. This idea was implemented in a simple device consisting of Si PIN detector for quantification of elemental composition in complex radioactive samples. In this case the X-ray excitation conditions will obviously vary from sample to sample; moreover, the resulting spectra will be a complex superposition of numerous signals from soft gamma emitters and X-ray fluorescence of various nature. These complex spectra can be effectively treated with chemometric data processing for quantification of particular elements.

The accuracy of the proposed technique was assessed in quantification of Zr, Mo, La, Ce, Nd and U in the mixtures based on real spent nuclear fuel samples and was found to be sufficient for technological needs in rapid element monitoring in highly radioactive SNF reprocessing media. In our opinion this approach can be easily extended to other elements having appropriate XRF lines. The accuracy will depend on line intensities and the extent of overlapping with other signals. Taking into account the simplicity of the device and the complex nature of the analytical task these results can be considered as a good promise for the development of novel approach for SNF analysis. The important feature of the proposed method is that it does not require the direct contact of radioactive samples with the instrument and can serve to minimize personnel dose burdens.

P06. Complex changes in a simple system: metrology through the glasses of chemometrics

J. Elek, D. Csilla

Hydrogen Revolution Hungary Ltd, Hungary

A system for hydrogen storage in aqueous alkali bicarbonate–formate media is patented by our company. The chemical equilibrium appears very simple, with only hydrogen entering and leaving the system during the reaction cycles.
An automated system is proposed to regulate the pressure and temperature zones; therefore, precise reaction monitoring is a key feature of the technology. Raman spectra of the reaction partners were collected, and a concentration-independent PLS model was developed to predict the reaction progression (transformation %). The model performed well, but when the bicarbonate solutions were exposed to air for some time, we observed a strange phenomenon: a shoulder in the carbonate signal, leading to a significant difference from the HPLC results. The signal is too large for CO_2 absorption; ^{13}C NMR does not support it either: no additional carbon signal appeared. It seems to be a “mystic” unknown component. The shoulder overlaps with the formate; however, the HPLC also did not show evidence for the presence of any new chemical entity. To overcome this interference, a dataset of 509 analyses with both HPLC results and Raman spectra available was split into training and test sets (85–15%). The samples were taken during or research operations at different initial concentrations and for different durations in the open air to cover the maximum possible variations. The HPLC results served as a reference, and the Raman spectra were used as predictors in an XGB regression model. In this work, we show how successful this new approach solved the prediction issues.

P07. Quantification of the minimal inhibition concentration for antimicrobial peptides by PLS

M.M. Khaydukova [1, 2], O.V. Shamova [1, 2]

[1] *Laboratory of design and synthesis of biologically active peptides, Department of Pathology and Pathophysiology, FSBSI Institute of Experimental Medicine, Saint Petersburg, Russia*
[2]*Faculty of Biology, St Petersburg University, Saint Petersburg, Russia*

The second half of the 21st century has been hailed as the "golden age" of antibiotics. The discovery of penicillin stripped many lethal diseases of their once-feared status. However, this breakthrough led to the widespread and often unnecessary global application of antimicrobial therapy. Over time, microorganisms evolved resistance mechanisms, resulting in the emergence of antibiotic-resistant strains. Methicillin-resistant Staphylococcus aureus (MRSA) was among the first documented strains resistant to β-lactam antibiotics. Today, the prevalence of resistant strains continues to rise. In 2017, the World Health Organization identified six highly virulent, antibiotic-resistant bacterial pathogens as priority threats [1]. Effective drugs against these pathogens are urgently needed, a situation further exacerbated by the COVID-19 pandemic.

Antimicrobial peptides (AMPs) are natural molecules that form an essential component of the innate immune system, protecting living organisms from bacterial, viral, and fungal infections. These molecules are also interesting because resistance mechanisms against them are anticipated to develop more slowly than those against conventional antibiotics. Numerous studies have employed machine learning techniques to classify peptides as antimicrobial based on their amino acid sequences and physicochemical properties. Such models have even been integrated into databases. Class boundaries are typically defined by the minimum inhibitory concentration (MIC) – the lowest concentration of an antimicrobial agent that inhibits visible microbial growth – with lower values indicating

greater potency. However, few studies have developed mathematical models capable of quantitatively predicting MIC solely from amino acid sequences.

In this study, we constructed three partial least squares (PLS) models. The independent variable matrices comprised numerical representations of amino acids from peptide sequences. Dependent variables were MIC values against bacteria from ESKAPE group obtained from the DRAMP database (http://dramp.cpu-bioinfor.org/). Then MICs were predicted for randomly generated peptide sequences. Six novel peptides with predicted MICs below 5 μM were synthesized via solid-phase peptide synthesis, purified, and characterized. Their antimicrobial activity was evaluated against both standard and resistant bacterial strains. The results will be presented in a poster.

References

[1] Organization W. H. WHO publishes list of bacteria for which new antibiotics are urgently needed, Book WHO publishes list of bacteria for which new antibiotics are urgently needed, Editor, 2017.

P08. Graph neural networks vs molecular descriptors and fingerprints for predicting molecular properties: who wins?

A. Sholokhova[1], D. Matyushin[1]

[1]Frumkin Institute of Physical chemistry and Electrochemistry Russian Academy of Sciences, 31-4, Leninsky prospect, 119071 Moscow, Russia

In the field of chemoinformatics, predicting various molecular properties (such as chromatographic retention, toxicity, and boiling point) based on molecule structure using statistical methods is of great importance. For several decades, molecular descriptors (MD) and molecular fingerprints (MF) have been used to address this issue. MD are values that describe a molecule and can be easily calculated based on its structure. MD include very simple quantities, such as molecular mass, as well as rather complex topological indices whose physical meaning is not always clear. MF are long, sparse vectors of integers (or bits) that indicate the presence or number of certain structural features in a molecule. After calculating the MD and MF, optional pre-selection is performed, and these values are used as input features in statistical and machine learning methods, such as linear regression, support vector regression (SVR), decision tree–based methods, and neural networks. The resulting model links the MF and MD to the predicted molecular property.

Currently, the literature tends to consider this approach obsolete. Graph neural networks are increasingly considered the "default" approach for constructing models that link molecular structure to predicted properties. This method demonstrates outstanding performance in many cases and is quite versatile. The generation of MD and MF is unnecessary because the molecular graph itself serves as the input feature for this method.

In recent years, many new machine learning methods have been developed to work with tabular data, in which each sample is a vector of numbers. These methods primarily rely on deep neural networks (e.g., TabR, TabM, and TabPFN). The novel methods for tabular data have the potential to revitalize the use of MD and MF in predicting molecular properties. However, there are still very few publications.

In this study, we systematically compared several variants of graph neural networks with several traditional machine learning methods, such as random forest, gradient boosting, linear regression, SVR, and multilayer perceptron, as well as several cutting-edge deep learning methods for tabular data. The considered tasks included predicting retention times and indices (gas and liquid chromatography), collision cross-sections, boiling points, LD50, and other molecular properties.

It was shown that MD and MF-based methods perform well for most tasks, in many cases exceeding the accuracy achieved using graph neural networks. The TabPFN method, which does not require pre-training, demonstrates good accuracy for most tasks. Overall, the so-called "*no free lunch theorem*" is confirmed: different methods are more effective in different cases.

We used chromatographic retention datasets created by our group along with datasets from the literature. This work presents several datasets in the field of gas chromatography-mass spectrometry, which can be further used by researchers in the fields of cheminformatics and chemometrics.

Acknowledgments

This research was supported by the Ministry of Science and Higher Education of the Russian Federation (№ 124041900012-4).

P09. Spectrophotometric determination of synthetic food colors E102 and E110 in soft drinks using magnetic nanoparticles and chemometric algorithms

K. Andreeva, N. Burmistrova, T. Rusanova

Saratov State University, Saratov, Russia

A new method for the determination of synthetic food dyes (SFDs) in their combined presence is proposed. This method is based on their extraction with magnetic nanoparticles, recording the absorption spectra of the reextract, and chemometric data processing. The SFDs of tartrazine (E102) and sunset yellow (E110) were determined in the soft drinks "YES! Fruit" and "Fresh Orange." Solid-phase extraction with magnetic Fe_3O_4 nanoparticles modified with polyethyleneimine was used to extract the dyes from the samples. After reextraction of the dyes with an alkali solution, they were spectrophotometrically determined using absorption spectra in the visible region. Individual determination of dyes with overlapping absorption bands was achieved using the chemometric algorithm of projection onto latent structures (PLS). For this purpose, a model of the absorption spectra's dependence on the dye concentrations in their mixtures was preliminarily constructed using the PLS method, the optimal number of latent variables was selected, and the model's performance was assessed. The determination error in the validation set was no more than 6% and 4% for tartrazine and sunset yellow, respectively. In the "YES!Fruit" drink sample, the tartrazine content was 2.19 ± 0.14 mg/L; in sunset yellow – 5.2 ± 0.6 mg/L, respectively. The dye content in the "Fresh Orange" drink was 2.4 ± 0.2 and 8.5 ± 1.2 mg/L for tartrazine and sunset yellow, respectively. The accuracy of the determination was confirmed using the "added-found" method and comparison with the results of another method—HPLC using a gradient elution mode (acetonitrile:ammonium acetate, 0.1 mol/L). Thus, it was demonstrated that the combination of solid-phase extraction with magnetic nanoparticles, spectrophotometry, and chemometrics enables the separate determination of OPCs with overlapping absorption bands in soft drinks.

P10. Chemometric analysis of geochemical parameters of hydromineral raw materials of the chechen republic

A. Chernyshova[1], M. Baskhanova[2], A. Shaipov[2], T. Rusanova[1]
[1]Saratov State University, Saratov, Russia
[2]Grozny State Petroleum Technological University, Grozny, Chechen Republic, Russia

The physicochemical properties and chemical composition of 22 samples of hydromineral raw materials obtained from flooded oil wells in the Chechen Republic were determined. The acidity, density, and dry residue values were found, as well as the concentrations of lithium, sodium, potassium, calcium, magnesium, and iron cations; hydrocarbonate, chloride, sulfate, iodine, and bromide ions; the boron and petroleum product contents were determined; and the water type was determined according to V.A. Sulin. The data were processed using the principal component analysis (PCA) method using the Python software environment. Outliers were preliminarily removed from the data, interval values were transformed into mean values, and deposits were encoded with numbers. The first several principal components explain a significant proportion of the total data variance (~87% is explained by the first 4 components). PCA successfully identified the main factors of variability of the geochemical composition of water: mineralization and temperature (PC1), acidity (PC2), as well as factors associated with the content of sulfate ions and total Fe (PC3). The nature of the dependence of physicochemical parameters on the sampling depth was studied. Many parameters, especially those closely related to PC1 (mineralization, temperature), have a strong positive correlation with depth. pH demonstrates a negative correlation with depth. A complete correlation matrix between the physicochemical parameters was constructed. The group of mineralization parameters (Na^+, K^+, Li^+, dry residue, chloride ions, иromide ions, density, temperature) strongly correlates with each other. Boron and iodide ions are closely related to each other. Iron and sulfate ions have more complex and less pronounced correlations with the main group. Cluster analysis successfully identified groups of objects with similar geochemical characteristics, which can be used to classify waters and determine their potential origin or formation processes.

P11. Voltammetric electronic tongue for identification of timolol pharmaceuticals by manufacturer

R. Zilberg, Ch. Mukhametdinov, E. Bulysheva, Yu. Teres

Ufa university of science and technology, Ufa, Russia

According to the World Health Organization, the share of counterfeit drugs on the pharmaceutical market in some countries can reach 30%, while the share of counterfeit drugs on the domestic pharmaceutical market is estimated at 12%. The development of rapid, simple, and inexpensive methods for identifying and monitoring the quality of pharmaceuticals is a pressing issue. Multisensor systems such as the "electronic tongue" [1-5] meet these requirements; they can quickly and easily determine several components of the analyzed solution simultaneously. The sensors exhibit cross-sensitivity to the components and excipients being determined. Using chemometric processing of voltammetric data, it becomes possible to identify drugs by manufacturer. Aluminosilicate and aluminophosphate zeolites, differing in structure and pore size, were used as modifiers for glassy carbon electrodes (GCE). The electrochemical behavior of the proposed composite sensors was studied using cyclic voltammetry and electrochemical impedance spectroscopy. Based on the linear dependence of peak currents on the square root of the potential scan rate and Semerano's criterion values close to 1, it was established that the electrode oxidation process of timolol preparations on the GUE/PEC@AEL, GUE/PEC@AFI, GUE/PEC@FAU, GUE/PEC@MFI, and GUE/PEC@BEA sensors is controlled by the adsorption of the electroactive substance to the electrode surface. Differential pulse voltammetry was used to analyze timolol pharmaceutical preparations from seven manufacturers, followed by chemometric data processing using principal component analysis (PCA) and Soft Independent Modeling of Class Analogy (SIMCA). The best separation of preparations on the score plots of

PCA models was observed using composite sensors modified with the following zeolites: FAU, AFI, and CHA. Based on these sensors, a three-sensor system was developed consisting of the SUE/PEC@AFI, SUE/PEC@FAU, and SUE/PEC@CHA. According to the SIMCA classification results, the identification of timolol pharmaceuticals is free of Type I errors, and Type II errors do not exceed 20%.

The work was supported by the Russian Science Foundation (grant no. 23-73-00119) https://rscf.ru/project/23- 73-00119/

References

[1] R. Zilberg, E. Bulysheva, Y. Teres, A. Volkova, G. Ishmakaeva, G. Mukhametdinov, I. Vakulin, Chim. Techno Acta 12 (2025) 12204
[2] R. A. Zilberg, J. B. Teres, E. O. Bulysheva [et al.], Electrochim. Acta 492 (2024) 144334.
[3] Y. A. Yarkaeva, D. I. Dubrovskii, R. A. Zil'berg, V. N. Maistrenko Russian Journal of Electrochemistry 56 (2020) 544-555;
[4] A. V. Sidelnikov, R. A. Zilberg, F. Kh. Kudasheva, V. N. Maistrenko, G. F. Yunusov, S. V. Sapelnikova, J. Anal. Chem. 63 (2008) 1072–1078;
[5] R. A. Zil'berg, V. N. Maistrenko, Y. A. Yarkaeva, D. I. Dubrovskii, J. Anal. Chem. 74 (2019) 1245-1255.

P12. Transferring multivariate calibrations between different types of spectrometers: NIR and FTIR in e-cigarette refill fluids analysis

N. Iurgenson[1], Y. Monakhova[2,3], D. Kirsanov[4,5]

[1]*Institute of Chemistry, University of Debrecen, Debrecen, Hungary*
[2]*FH Aachen University of Applied Sciences, Department of Chemistry and Biotechnology, Jülich, Germany*
[3]*Saratov State University, Institute of Chemistry, Saratov, Russia*
[4]*Institute of Chemistry, St Petersburg University, St. Petersburg, Russia*
[5]*ITMO University, St. Petersburg, Russia*

This work assesses the potential for transferring multivariate calibration models between Near Infrared (NIR) and Attenuated Total Reflection Infrared (ATR-IR) spectroscopic platforms. Using the analysis of e-cigarette refill fluids as a case study, we demonstrate that the Direct Standardization (DS) method enables the effective transfer of Partial Least Squares (PLS) models between these instruments.

The target analytes were nicotine (0 – 36.5 mg/mL), glycerol (7.5 – 61.9 v/v %) and propylene glycol (30.8 – 85.8 v/v %). Both NIR and ATR-FTIR data yielded reasonably accurate PLS models for quantification of all three components, nicotine was quantified with somewhat bigger relative errors due to its' spectral features. Calibration transfer was performed in both directions: the NIR data were converted into ATR-FTIR format and vice versa.

While transferred models show a modest increase in prediction error (RMSEP values typically 1.0 to 1.5 times higher than native models), their performance is highly dependent on the transfer set selection. In optimal cases, transferred models can achieve accuracy comparable to models built directly on the target instrument's data, indicating a viable path toward robust, platform-independent calibration models for diverse analytical applications [1].

The study was supported by the Russian Science Foundation (grant no. 25-43-20018)

References

[1]. Iurgenson, N., Monakhova, Y., Kirsanov, D. (2025) Microchemical Journal, 212, art. no. 113375, DOI: 10.1016/j.microc.2025.113375

P13. Aquaphotomics and NIR spectroscopy for monitoring of protein microbubble structural integrity for drug delivery systems

Y. Ladanova [1], T.M. Estifeeva [2], P.G. Rudakovskaya [2], J. Muncan [3], A.O. Orlova[1], A. Surkova [1],

[1] International Laboratory "Hybrid Nanostructures for Biomedicine", PhysNano Department ITMO University, Saint Petersburg, Russia
[2] Center for Photonic Science and Engineering, Skolkovo Institute of Science and Technology, Moscow, Russia
[3]Department of Biochemistry, Molecular biology, Entomology & Plant Pathology, Mississippi State University, USA

Proteins are key functional components of biological and biotechnological systems. Preserving the native spatial structure of a protein is essential for its biological activity, stability, and predictable behavior in solution. Protein denaturation — the disruption of secondary, tertiary, and quaternary structures under physical or chemical factors — can lead to aggregation, loss of functionality, and altered interactions with the environment. Therefore, developing methods for rapid process control is a crucial task in biophysical research and the production of protein-based formulations [1].

Various spectroscopic techniques, such as circular dichroism (CD), infrared (IR), and Raman spectroscopy, are commonly employed for protein structural analysis. CD spectroscopy is widely used to determine secondary structure content in protein solutions and to monitor conformational transitions such as denaturation [2]. Despite its high information content, practical use for *in situ* control is often limited by stringent optical transparency requirements, high cost, and equipment complexity.

Near-infrared (NIR) spectroscopy is a powerful, non-destructive analytical tool enabling rapid and precise *in situ* analysis, widely used for studying aqueous and biological systems. NIR spectroscopy detects overtones and combination bands of hydrogen-bonded group vibrations (O–H, N–H, C–H), making it particularly sensitive to the composition and structure of complex matrices. In protein studies, NIR spectroscopy can determine total protein content and monitor thermal effects, aggregation, and changes in solution properties [3]. However, the direct spectral signals of proteins in this region are weak. Simultaneously, the method exhibits high sensitivity to changes in the state of the aqueous matrix surrounding the macromolecule. The analysis of these changes, fundamental to aquaphotomics [4], allows for the indirect yet highly sensitive detection of protein conformational transitions (e.g., denaturation and folding) and the study of molecular interactions by monitoring rearrangements in the hydration shell.

The aim of this work was to evaluate the applicability of NIR spectroscopy combined with chemometrics and aquaphotomics for monitoring structural changes of bovine serum albumin (BSA) under thermal stress. BSA was investigated in an aqueous solution, in a microbubble-based system, and in complexes with stabilizing additives (spermine and polyvinylpyrrolidone). NIR spectra were analyzed using principal component analysis (PCA) to identify key factors of spectral variability and reduce data dimensionality, complemented by an aquaphotomics approach to interpret changes in the water matrix. The results demonstrate that this integrated methodology serves as an effective tool for the non-destructive monitoring of protein denaturation and analysis of the stabilizing effect of the additives in complex aqueous systems.

This work was supported by the Russian Science Foundation (grant no 24-73-10191). J. Ladanova acknowledges funding from the ITMO University Academic Mobility Program.

References
[1] N. Gooran, K. Kopra, *Int. J. Mol. Sci.* **25(3)** (2024) 1764.
[2] C. Jones, *J Pharm Biomed Anal.* 219 (2022) 114945.
[3] K. Beć, *Molecules* **25(12)** (2020) 2948.
[4] Tsenkova, R., et al., *Front. Chem.* **6** (2018) 636.

P14. Investigation of Water Structure in Allogeneic Bioimplants Using NIR Spectroscopy and Aquaphotomics

M. Yaroslavova [1], D. Sinitsyn [2], N. Ryabov [3], A. Volov [3], L. Volova [3], J. Muncan [4], A. Orlova [1], A. Surkova [1]

[1]*International Laboratory "Hybrid Nanostructures for Biomedicine", ITMO University, Saint Petersburg, Russia*
[2]*Department of Analytical and Physical Chemistry, Samara State Technical University, Samara, Russia*
[3]*Research Institute of Biotechnology "BioTech", Samara State Medical University, Samara, Russia*
[4]*Aquaphotomics Research Field, Graduate School of Agricultural Sciences, Kobe University, Kobe, Japan*

In regenerative medicine allogeneic bioimplants are widely used to restore connective and supporting tissues, which are taken from donors of the same biological species [1]. For successful and safe integration, these bioimplants must undergo several stages of purification and testing for various parameters, both during manufacturing stage, storage and use [2]. One of the most important parameters for evaluating these materials is residual moisture, because it is directly related to physicochemical properties, mechanical strength and stability. Moisture control at an optimal level allows bioimplants to be stored at room temperature and facilitates their transportation.

Nowadays, the most common method of moisture control is thermogravimetric analysis, but it has a few disadvantages, including the removal of the sample from the sterile package and destroying the analyzed material [3]. For these reasons, it is necessary to develop a fast non-destructive method of moisture control through protective packaging. A possible alternative is near-infrared spectroscopy (NIR) combined with chemometrics [4]. The method is based on the analysis of absorption spectra, which provides information about the water content in the samples.

In this study, we evaluated the applicability of fiber-optic NIR spectroscopy for non-destructive moisture determination of bioimplants from human allogeneic bone tissue. The study confirmed that plastic packaging has a minimal effect on spectral measurements. At first, the NIR spectra of dry and wet samples were analyzed. The processing of spectral data by principal component analysis (PCA) revealed distinct qualitative differences between the states of the samples, which confirms the sensitivity of the method to changes in moisture. NIR spectra were then recorded during sequential drying of the materials. A regression model based on the partial least squares regression (PLSR) method was developed to quantify the moisture content. The PLSR model for the dataset consisted of 60 measurements during the drying or wetting processes yielded an RMSEP of 5.43% in 0.62–64.36% range of moisture content. Thus, it can be concluded that NIR spectroscopy can be used for moisture quantification in allogeneic bioimplants.

Aquaphotomics provided insight into changes in the state of water within the biomaterials [5]. Aquagrams served as a tool to track water-material interaction dynamics by visualizing changes in water's molecular structure. The analysis revealed that the drying process primarily involved the loss of free and weakly bound water, while the absorption band at 1461 nm remained stable. This spectral stability indicates the preservation of collagen's structural integrity during production. Thus, this study demonstrates the effectiveness of integrating NIR spectroscopy with aquaphotomics, enabling not only

quantitative moisture control of allogeneic bioimplants but also a qualitative assessment of their structural integrity.

References

[1] A. Aratikatla, N. Maffulli, H.C. Rodriguez, et al. *J. Orthop. Surg.* **17** (2022) 307.
[2] W. Liang, C. Zhou, J. Bai, et al. *Heliyon* **10** (2024) e36152.
[3] P. Matejtschuk, C. Duru, K. Malik, et al. *Am. J. Anal. Chem.* **07** (2016) 260–265.
[4] K.B. Beć, J. Grabska, C.W. Huck, *Adv. Food Nutr. Res.* (2025) S1043452625000026.
[5] J. Muncan, V. Matovic, S. Nikolic, J. Askovic, R. Tsenkova, *Talanta* **206** (2020) 120253.

P15. Machine Learning Prediction of Key Supramolecular Descriptors for Calixarene Host-Guest Complexes

Arbukhanova G. A. [1], Kalashnikov A. M. [1], Ignatiuk E. S.[1], Boichenko E. S. [1], Muraviev A.A.[1], Kirsanov D. O.[1,2]

[1]*ITMO University, Saint Petersburg, Russia*
[2] *Saint Petersburg State University, Russia*

Calix[n]arenes are macrocyclic compounds with a cavity suitable for host-guest interactions with various analytes. Their tunable cavity size and functional groups make them promising for developing sensitive and selective chemical sensors for hazardous gases, pollutants, and biologically active molecules [1,2]. It is especially applicable in the field of massive screening of cancer by analyzing a patient's sample of exhaled air, which is a rich source of volatile organic compounds associated with presence of specific tumors. However, designing effective calixarene-based sensors empirically requires time-consuming stages of synthesis of calixarenes (which is rarely a straightforward process) and analysis of model samples. Replacing the experimental work by theoretical assessment of analytical characteristics of calixarenes and compiling a list of top candidates with highest sensitivity for further tests is a key advancement allowing for efficient design of new types of sensors. This approach is known as a quantitative structure-property relationship (QSPR). QSPR modeling predicts chemical properties based on molecular structure and is widely used in drug design, where large datasets of known compounds with described properties are available. Nevertheless, recent studies show QSPR also works with small samples, for example to predict the analytical performance of potentiometric sensors from ionophore structure [3]. To address this challenge, we trained a PLS model correlating the molecular structure of calixarenes with the DFT-calculated Gibbs free energy of complexation with target analytes – a primary indicator of binding strength and potential sensor sensitivity. This avoids collecting extensive experimental data. Our model predicts both Gibbs free energy of complexation with various guest molecules and HOMO-LUMO energy levels. The PLS model was trained on a set of calixarenes with diverse structures, described by a comprehensive set of molecular descriptors. The optimized model accuracy was measured by 5-fold cross-validation test. The obtained R^2 values ranging from 0.7 to 0.92 confirms the model's reliability in prediction of the studied properties. This work provides a practical approach for the rapid in silico screening of calixarenes, speeding up the initial design of supramolecular receptors for sensing applications.

References

[1] R. Nag, C. P. Rao, *Chem. Commun.* **58** (2022) 6938-6951.
[2] D. Quaglio, F. Polli, C. Del Plato, et al. *Supramol. Chem.* **33** (2021) 345-369.
[3] N. Vladimirova, V. Polukeev, J. Ashina, et al. *Chemosensors* **10** (2022) 43.

P16. Application of chemical sensors for detection of coliform bacteria in milk

A. Shuba[1], E. Anokhina [2]
[1]*Deparment Physical and Analytical Chemistry, Voronezh State University of Engineering Technologies, Voronezh, Russia*
[2]*Laboratory of metagenomic and food biotechnology, Voronezh State University of Engineering Technologies, Voronezh, Russia*

The use of sensors in biomedical research is often based on the assessment of microbial metabolites in various environments. However, to improve the accuracy and reliability of the analysis, it is necessary to predict the presence of specific metabolites in the sample. Various omics technologies are used to study metabolic pathways; however, the data obtained in this way may be insufficient for specific analytical conditions. Therefore, it is important to identify potential metabolites in the sample to optimize sensor data processing.

The aim of this study was to develop a method for identifying coliform bacteria in milk based on projection methods for analyzing sensor data, taking into account the products of amino acid metabolism in milk.

The study subjects were pasteurized milk samples of varying fat content, pre-inoculated with *E. coli* (a strain from the university collection). The milk samples were heated for 4 hours, and the number of colony-forming units (CFUs) in the samples was determined hourly according to GOST standards. The gas phase above the milk samples was simultaneously analyzed using an array of 8 piezoelectric sensors with composite coatings. Liquid chromatography was used to analyze the content of 28 amino acids in samples collected at gas phase sampling points. Projection regression methods were used to determine the *E. coli* content in milk samples. It was shown that including sensor parameters responsible for changes in amino acid content in milk samples increased the accuracy of the regression model. It has also been shown that taking into account the calculated parameters of sensors for the products of amino acid metabolism by bacteria under anaerobic conditions allows for increasing the accuracy of identification of *E. coli* and *Staphylococcus aureus* in milk samples. This work was supported by the Russian Science Foundation, Grant No. 22-76-10048.

P17. On effective strategies for applying chemometric methods to process chemical sensor data

A. Shuba[1], T. Kuchmenko[1], N. Ramya[2], R. Sowmyalakshmi[3], H. Karami[4]

[1]*FSBEI HE "Voronezh State University of Engineering Technologies", Voronezh, Russia*
[2]*Saranathan College of Engineering, Trichy, Tamil Nadu, India*
[3]*University College of Engineering, BIT- Campus, Tiruchirappalli, Tamil Nadu, India*
[4]*Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology, Baldiri Reixac 10-12, 08028 Barcelona, Spain*

The use of sensor systems and technologies to solve a wide range of problems is gaining momentum in modern analytical methods. To extract meaningful information from the acquired data, a data processing method is often selected experimentally, comparing the metrological characteristics of the resulting models to address the research objectives. Based on personal research experience and scientific publications, the authors attempted to

systematize and apply various chemometric methods and other mathematical data processing techniques to solve a wide range of problems using chemical gas sensors.

This study examines the application of chemometric methods to process data from various sensor types (semiconductor, quartz crystal microbalance, and electrochemical) to determine chemical and biologically active substances, predict complex integral indicators, assess membership in established classes, and identify groups and subsets within a series of objects. The resulting effective mathematical models were analyzed taking into account the sample matrix, validation method, and complexity of the algorithms used. The authors also analyzed the application of various projection methods, artificial neural networks, discriminant and cluster analysis, support vector machines, and random forests, as well as methods proposed by the authors. A scheme for selecting methods for processing sensor data is proposed for solving various types of analytical problems based on a priori information about the complexity of the composition and origin of the sample, the number of analytes to be determined, and possible limitations in the experiment.

P18. NIR Spectroscopic Analysis of Quinine in Cinchona Bark through Optimised Wavelength Selection and Multivariate Calibration

Dilip Sing[1], Md Banaz Alam[2], Arbab Mahtab[1], Ajanto Kumar Hazarika[3], Samuel Rai[2], Rajib Bandyopadhyay[1]

[1]Department of Instrumentation and Electronics Engineering, Jadavpur University, Salt Lake Campus, Kolkata 700106, India
[2]Directorate of Cinchona and Other Medicinal Plants, Mungpoo, Darjeeling – 734313, India
[3]Tocklai Tea Research Institute, Tea Research Association, Jorhat 785008, Assam, India

Quinine, the primary alkaloid in cinchona bark, remains crucial for malaria treatment and has diverse applications in pharmaceuticals, beverages, and chemical synthesis. India hosts one of the largest cinchona plantations, covering approximately 2,800 hectares in the Darjeeling hills. However, extensive cross-pollination and seed propagation over the past century have led to declining quinine content, reducing market value. Large-scale plantation rejuvenation through vegetative propagation requires rapid, cost-effective methods for assessing quinine content in millions of plants, making conventional analytical techniques impractical.

This study developed a field-deployable method for predicting quinine content in cinchona using Near-Infrared (NIR) spectroscopy combined with advanced chemometric techniques. A total of 220 bark samples were collected from eight plantation divisions across Darjeeling (1,200–6,000 ft altitude). NIR spectra were acquired using a portable spectrometer with an InGaAs detector array covering 900–1700 nm, utilising a tungsten halogen lamp (12V/20W) as the light source [1]. Standard Normal Variate (SNV) and Savitzky-Golay (SG) preprocessing were applied, and two wavelength selection algorithms — CARS and IRIV — were used to optimise inputs for Partial Least Squares Regression (PLSR).

The PLSR models developed using characteristic wavelengths selected by CARS and IRIV algorithms demonstrated excellent predictive performance, achieving coefficient of determination (R^2) values exceeding 0.90, Root Mean Square Error of Prediction (RMSEP) below 0.85%, and Ratio of Performance to Deviation (RPD) greater than 3.2. The best-performing model, obtained using SNV preprocessing combined with CARS wavelength selection and PLSR, achieved R^2 = 0.93, RMSEP = 0.78%, and RPD = 3.68, indicating excellent prediction capability.

The integration of CARS and IRIV wavelength selection algorithms with PLSR significantly improved model performance by reducing spectral dimensionality while retaining critical chemical information. The portable NIR-based method offers a rapid, non-destructive, and

field-applicable solution for quality assessment in cinchona plantation management, with potential extension to biomarker profiling and quality trait prediction in other medicinal plants, supporting precision agriculture and breeding programmes.

Acknowledgements: Indian Council of Agricultural Research (ICAR) for the programme National Agricultural Science Fund vide Sanction Order No. NASF/PA-10009/2023-24 dated 26.02.2024.

References

[1] D. Sing, S. Banerjee, R. Mallik, et al. *Microchem. J.* **199** (2024) 109949.

P19 Feature Selection in Spectral Classification Models for Biological Samples: Enhancing Urolithiasis Treatment

E.S. Boichenko [1,2], K.N. Lastochkina [2], D.O. Kirsanov [1,2]

[1]ITMO University, Saint Petersburg, Russia
[2]St Petersburg University, Saint Petersburg, Russia

Urolithiasis is a prevalent condition characterized by stone formation in the kidneys, bladder, and upper urinary tract. Accurate determination of urinary stone composition is crucial for selecting treatment strategies and preventing recurrences. Developing in vivo technologies for stone analysis is thus highly relevant, enabling efficient lithotripsy and extraction during surgery, along with tailored prophylaxis recommendations. Portable near-infrared (NIR) spectrometers equipped with fiber-optic probes, combined with chemometric data processing, represent a promising approach [1].

NIR spectral data exhibit high dimensionality, with many redundant and noisy variables that can cause overfitting or poor performance of machine learning models. In this study, four feature selection methods were compared – interval selection, ReliefF, minimum redundancy maximum relevance (mRMR), and neighborhood component analysis (NCA) – in ECOC-based multiclass classification of urinary stones, evaluating their impact on precision and recall of the models.

We analyzed 260 urinary stones (73% oxalates, 18% urates, 9% phosphates) using an AvaSpec-NIR256-1.7-USB2 spectrometer and AvaLight-HAL-S-Mini source (Avantes, the Netherlands), and flexible fiber optic probe (Endoprobe 7NIR + 1NIR, Optofiber, Russia). Diffuse reflectance spectra (939-1799 nm, 4 nm resolution) were measured under standard conditions (air) and in saline medium.

Under standard conditions, all methods matched full-spectrum performance for oxalates (G-mean 92-95%) and urates (87-90%). Phosphate classification was different: interval selection (1381-1517 nm) and ReliefF demonstrated G-mean 72% and 70% vs. 70% for the full spectra, improving precision (82% vs. 77%) while maintaining recall (61-64%). In saline, performance declined predictably, but ReliefF proved resilient for oxalates (92%) and urates (83%), with G-mean 59% for phosphates; interval selection (1521-1651 nm) improved the G-mean for phosphates (61% vs. 52% full).

These results highlight that interval selection and ReliefF were most suitable feature selection algorithms, particularly for challenging phosphate stones, reducing the number of spectral variables from 225 to 36-40. This approach supports real-time intraoperative classification, advancing personalized urolithiasis management.

References

[1] E. Boichenko, M. Paronnikov, A. Reznichenko, et al., *Anal. Chim. Acta* **1354** (2025) 344007.

P20 Exploring the Potential of NIR Spectroscopy Combined with Aquaphotomics for Urine-Based Cancer Screening

A. Latysheva [1], E. Boichenko [2], J. Muncan [3], A.O. Orlova [1], A. Surkova [1]
[1] International Laboratory "Hybrid Nanostructures for Biomedicine", PhysNano Department ITMO University, Saint Petersburg, Russia.
[2] Chemical Engineering Center, ITMO University, Saint Petersburg, Russia
[3]Department of Biochemistry, Molecular biology, Entomology & Plant Pathology, Mississippi State University, USA

The presence of cancer cells induces systemic metabolic alterations, which are reflected in the chemical profile of extracellular biofluids (e.g., patient's urine, blood, plasma, or cerebrospinal fluid) [1]. The analysis of biological fluids is one of the most promising noninvasive approaches for rapid cancer screening. A multitude of methods are used for fluid analysis, including a variety of chromatographic and spectroscopic techniques. These methods, however, are often costly and involve lengthy, tedious procedures. There is a pressing need to develop efficient, rapid, and simple alternatives for cancer screening.

Near-infrared (NIR) spectroscopy holds high potential for rapid screening and dynamic patient monitoring through the analysis of biological fluids. NIR spectroscopy enables the assessment of key biomolecules such as hemoglobin and water in biological fluids [2]. However, due to the broad and heavily overlapping absorption bands characteristic of the NIR range, the information content of raw spectra is often limited. Specialized data preprocessing methods and chemometrics are applied to extract relevant information. NIR spectra are particularly valuable for studying water structure, especially in the region of the first overtone of O–H stretching vibrations (1300–1600 nm), which contains numerous characteristic water absorption bands. This region forms the basis for a new scientific discipline – aquaphotomics [3] – which investigates the molecular organization of water through the analysis of NIR spectra.

This study explores the potential of aquaphotomics approach in discrimination of urine samples from patients with diagnosed prostate cancer, kidney cancer, and control group. 124 samples were studied with Shimadzu UV-3600 spectrophotometer in the 900-1600 nm range. Various spectral preprocessing methods (normalization, SNV, EMSC) were explored and the yielded data were employed as inputs for machine learning classification algorithms. The results indicate a certain potential of the proposed approach, however, further dedicated studies are required to clarify the real world applicability.

References

[1] M. Mayhew, O. Megram, S. Roshan, et al., *Metabolomics* **22(1)** (2025)13.
[2] C. W. Huck, in *Near-Infrared Spectroscopy* (Eds.: Y. Ozaki, C. Huck, S. Tsuchikawa, S. B. Engelsen), Springer, Singapore (2021) 413–435.
[3] Tsenkova, R., et al., *Front. Chem.* **6** (2018) 636.