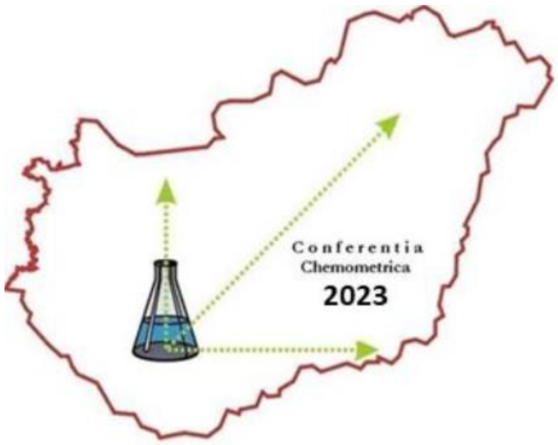
Conferentia Chemometrica 2023

Sopron, Hungary
Hotel Sopron
September 10-13, 2023



Research Centre for Natural Sciences, Institute of Excellence, Hungarian Academy of Sciences
Chemometrics and Chemoinformatics Working Group of the Hungarian Academy of Sciences
Science Port Ltd

International Organizing Committee

K. Héberger (Hungary) chair

R. G. Brereton (UK)

O. M. Kvalheim (Norway)

Kunal Roy (India)

K. Varmuza (Austria)

A. Voelkel (Poland)

I. G. Zenkevich (Russia)

Local Organizers

Head:

Károly Héberger, Research Centre for Natural Sciences, Institute of Excellence, Hungarian Academy of Sciences

Administrative Affairs:

Dávid Bajusz and Anita Rácz, Research Centre for Natural Sciences, Institute of Excellence, Hungarian Academy of Sciences

Financial Affairs:

János Elek SciencePort Ltd.

ISBN 978-963-7067-39-6

Publisher: Research Centre for Natural Sciences, Institute of Excellence, Hungarian Academy of Sciences

Edited by Károly Héberger

Scientific Program of the Conferentia Chemometrica 2023: an overview

Sept. 10, Sunday: 16:00–19:00 Registration 19:00–21:00 Dinner & get-together party;

Sept. 11, Monday: 8:30-9:20 Registration, 9:20-9:30 Opening

September 11 Monday		September 12 Tuesday		September 13 Wednesday	
09:30–10:00	L01 O.M. Kvalheim	09:30–10:00	L12 A. Tropsha	09:30–10:00	L23 K. Khan
10:00–10:30	L02 G. Tóth	10:00–10:30	L13 S. Podlewska	10:00–10:30	L24 M. Putz
10:30–11:00	L03 I.V. Tetko	10:30–11:00	L14 V. Poroikov	10:30–11:00	L25 A L. Pomerantsev
11:00–11:30	Break	11:00–11:30	Break	11:00–11:30	Break
11:30–12:00	L04 T. Andersons	11:30–12:00	L15 D. Kirsanov	11:30–12:00	L26 J. Abonyi
12:00–12:30	L05 S. Kovács	12:00–12:30	L16 B. Grung	12:00–12:30	L27 J. Hageman
12:30–13:00	L06 F. Andrić	12:30–13:00	L17 I. Stanimirova	12:30–13:00	L28 K. Héberger
13:00–14:00	Lunch	13:00–14:00	Lunch	13:00–14:00	Lunch
14:00–14:30	L07 T. Baczek	14:00–14:30	L18 M. Daszykowski	14:00–	Departure
14:30–15:00	L08 A. Farkas	14:30–15:00	L19 M. Csontos		
15:00–15:30	L09 A. Rácz	15:00–15:30	L20 L. Pieszczek		
15:30–16:00	Break	15:30–16:00	Break		
16:00–16:30	L10 R. Brereton	16:00–16:30	L21 J. A. Shimshoni		
16:30–17:00	L11 B. Vajna	16:30–17:00	POSTER SESSION		
17:00– 18:00	POSTER SESSION	17:00– 18:00			
18:00–19:30	Dinner	19:00-21:00	Banquet (Best Poster Award)		
19:30- 21:00	Wine Tasting				

Sunday evening, Sept. 10, 2023

16:00–18:00 **Registration**
18:00–20:00 **Get-together party**

Monday morning, Sept. 11, 2023

09:00–09:20 **Registration**
09:20–09:30 **Opening, technical information**

Latent variable regression, Seriation, Consensus modeling

09:30–10:00 **L01** **Olav M. Kvalheim** *Warren S. Vidar, Tim U.H. Baumeister, Roger G. Linington and Nadja B. Cech*:
Confounding in model interpretation when using latent variable regression methods to make inferences

10:00–10:30 **L02** **Gergely Tóth** *and Rita Lasfar:*
Patch seriation to visualize data and model parameters

10:30–11:00 **L03** **Igor V. Tetko:**
Automatic detection of outlying compounds in the OCHEM platform using consensus modelling

11:00–11:30 Coffee Break

Ranking, multicriteria (multiobject) optimization, sum of ranking differences (SRD)

11:30–12:00 **L04** **Tomass Andersons**, *M. Sawall, K. Neymeyr:*
Multivariate curve resolution with rank deficiency

12:00–12:30 **L05** **Sándor Kovács**, *Attila Gere, Károly Héberger:*
Tailoring MultiCriteria Decision Making methods optimally for individual data sets

12:30–13:00 **L06** **Filip Andrić:**
Multiobjective optimization of effect directed planar chromatography as a tool for fast screening of polypotent natural products

13:00–14:00 Lunch Break

Monday afternoon, Sept. 11, 2023

Data evaluation in chromatography, applications, metabolomics, chemical imaging, QSPR

14:00–14:30 **L07** **Tomasz Bączek:** *Changes of fatty acids' levels in inflammatory bowel diseases in view of the samples collected within the Integrative Human Microbiome Project*

14:30–15:00 **L08** **Attila Farkas,** Á. Kopasz, T. Jámbor, E. Varga, B. Nagy: N*ear-infrared and Raman chemical imaging for prediction of immediate dissolution of acetylsalicylic acid tablets*

15:00–15:30 **L09** *Szilvia Klébert, Dóra Tátraaljai, Krisztina László,* **Anita Rácz:** *Prediction of physical and chemical features of structural materials with machine learning and classical chemometric tools*

15:30–16:00 Coffee Break

16:00–16:30 **L10** **Richard G. Brereton**: *Features of Hotelling's T^2 on simulated and metbolomics data: A tutorial.*

16:30–17:00 **L11** **Balázs Vajna:** *Data science: chemometric techniques generating business value*

17:00–18:00 **POSTER SESSION**

18:00–19:00 Dinner

19:00–21:00 **Wine tasting**

Tuesday morning, Sept. 12, 2023

Drug design, QSAR, consensus modeling

09:30–10:00 **L12** *K. I. Popov, J. Wellnitz, T. Maxfield, and* **Alexander Tropsha:** *Hit Discovery using Docking ENriched by GEnerative Modeling (HIDDEN GEM): A Novel Computational Tool for Accelerated Virtual Screening of Ultra-large Chemical Libraries*

10:00–10:30 **L13** *Agnieszka Wojtuch, Ewelina Jamrozik, Tomasz Danel,* **Sabina Podlewska:** *Explainability approaches in computer-aided drug design*

10:30–11:00 **L14** **Vladimir Poroikov**, *Dmitry Druzhilovski, Nadezhda Biziukova, Oleg Gomazkov, Alexander Veselovsky, Alexander Dmitriev, Sergey Ivanov, Nikita Ionov, Dmitry Karasev, Anastassia Rudik, Polina Savosina, Boris Sobolev, Leonid Stolbov, Vladislav Sukhachev, Olga Tarasova, Dmitry Filimonov: Drug repurposing for Covid-19 therapy: in silico, in vitro, in vivo and in clinics*

11:00–11:30 Coffee Break

Green chemistry, applications, multivariate data analysis

11:30–12:00 **L15** *Mikhail Saveliev, Vitaly Panchuk,* **Dmitry Kirsanov:** *The role of chemometrics in making analytical chemistry green*

12:00–12:30 **L16** **Bjørn Grung:** *Characterizing solvent composition in a CO_2 capture plant using multivariate data analysis of online sensors and spectroscopic data*

12:30–13:00 **L17** **Ivana Stanimirova** *and P.K. Hopke: A strategy for source apportionment analysis*

13:00–14:00 Lunch Break

Tuesday afternoon, Sept. 12, 2023

Chemometric modeling, statistical process control, hyperspectral chemical images, one class classifcation

14:00–14:30 **L18** **Michal Daszykowski**, *L. Pieszczek, I. Stanimirova, S. Krzebietke, H. Czarnik-Matusewicz: Chemometric modelling of vital soil parameters*

14:30–15:00 **L19** **Máté Csontos,** *János Elek*: *Needle in a haystack–NIR method development for quantifying novel foods and additives in rodent diet*

15:00–15:30 **L20** **Lukasz Pieszczek**, *Michal Daszykowski: Combining hyperspectral non-homogeneity measures and a one-class classification concept*

15:30–16:00 Coffee Break

16:00–16:30 **L21** **Jakob Shimshoni:** *Near-Infrared Spectroscopy coupled with multivariate models to predict secondary metabolite and botrytis in cannabis and basil*

17:00–18:00 **POSTER SESSION**

19:00– **Banquet (best poster award)**

Wednesday morning, Sept. 13, 2019

QSAR, QSPR, nanotoxicity, SIMCA

09:30–10:00 **L23** **Kabiruddin Khan**, *Agnieszka Gajewicz-Skretna: The path of computational methods in assessing nano-ecotoxicity: retrospective, current status, and future perspectives*

10:00–10:30 **L24** **Mihai V. Putz:** *Balancing between antifragility and black swans in QSAR*

10:30–11:00 **L25** **Alexey L. Pomerantsev,** *O. Ye. Rodionova: Mutter SIMCA und ihre Kinder [Mother SIMCA and her children]*

11:00–11:30 Coffee Break

11:30–12:00 **L26** *Martin Ferenczi, Ádám Ipkovich, Zsolt Tibor Kosztyán,* **János Abonyi:** *Graph analysis based statistical process control*

12:00–12:30 **L27** **Jos Hageman,** Carla Araya-Cloutier, Sylvia Kalli and Jean-Paul Vincken, *Validating QSAR models: an anti-MRSA case study*

12:30–13:00 **L28** **Károly Héberger:** *The way it was: Personal, idiosyncratic reminiscences of past and present scientific achievements*

13:00–14:00 Lunch

14:00– **Departure**

Poster sessions

Monday and Tuesday afternoon:

17:00–18:00

P01 Mitra R. Alcaraz, M. Antonio, F. Chiappini, J. Zaldarriaga Heredia, S.M. Azcarate, J.M. Camiña, A. Muñoz de la Peña, M.J. Culzoni, H.C. Goicoechea: Higher-order data analysis to leverage the performance of food quality control procedures

P02 Rosa Maria Alonso-Salces, G. E. Viacava, A. Tres, S. Vichi, E. Valli, A. Bendini, T. Gallina Toschi, L. A. Berrueta: Pattern recognition analysis of ^{1}H-NMR fingerprint data for the geographical authentication of virgin olive oils

P03 Katharina Beier, T.-M. Dutschmann, P. M. Puttich, M. Lubienski, T. Beuerle, K. Baumann: Classification of horsetails using machine learning methods on NIR spectra

P04 Bendegúz Borkovits, E. Kontsek, A. Pesti, S. Gergely, I. Csabai, A. Kiss, P. Pollner: Multivariate modelling of mid-infrared spectra of colorectal cancer

P05 Máté Csontos, J. Elek, Z. Vincze: The effect of sample grinding in NIR spectroscopy

P06 Pegah Dehbozorgi, L. Duponchel, V. Motto-Ros, Thomas Bocklitz: Laser-Induced Breakdown Spectroscopy (LIBS) data analysis

P07 Tatjana Djakovic-Sekulić: Quantitative Structure–Retention Relationship study of β-tetralino-spiro-5-hydantoin derivatives

P08 Pawel Dziki, L. Pieszczek, K. Rybicka, M. Daszykowski: Toward more efficient and effective color quality control in the large-scale printing process

P09 J. Slezsák, Z. Gál, A. Salgó, Szilveszter Gergely: Investigation of carbohydrate powder mixtures by near-infrared spectroscopy and multivariate data analysis

P10 Adriano A. Gomes, Ivan Špánik: Multiway one class classification PLS based

P11 Dániel Kovács, Z. Fazekas: Sum of ranking differences (SRD) when differences diminish and reference ranking is ambiguous: the theoretical foundations of weighting schemes

P12 Sándor Kovács, Károly Héberger: Selection of preferable and undesirable distance measures for stochastic optimization by cross entropy

P13 Rita Lasfar and Gergely Tóth: The difference of model robustness assessment using cross-validation and bootstrap methods

P14 Lilla Alexandra Mészáros, Attila Farkas, Zsombor Kristóf Nagy: Machine vision system and multivariate data analysis in the quality assessment of tablets

P15 Nabiollah Mobaraki, K. Baumann: Non-membership probability for assigning a non-member (outlaying) sample in different variants of random forest classification

P16 Márton Mócz, P. P. Hanzelik, J. Slezsák, S. Gergely: Impact of different model transfer algorithms on dilution series and oil samples

P17 Brigitta Nagy, A. Farkas, D. Galata, Zs. K. Nagy: Artificial neural networks in pharmaceutical process development and quality assurance

P18 Anita Rácz, Anna Vincze, György T. Balogh: Extending the limitations in the prediction of permeability with machine learning algorithms based on a diverse PAMPA dataset

P19 Oxana Ye. Rodionova, N. I. Kurysheva, G. A. Sharova, A. L. Pomerantsev: Chemometrics for personalized medicine

P20 N. Tomčić, M. Jankov, P. Ristivojević, J. Trifković, Filip Andrić: High-performance thin-layer chromatography and multivariate image analysis in modelling of adulteration of *Salvia sp*. with olive leaves

P21 Gyöngyi Vastag, S. Apostolov, Š. Ivošević: Chemometrics as a tool to monitoring corrosion degradation of the selected alloys in real conditions

P22 Ekaterina Yuskina, Nikodim Makarov, Maria Khaydukova, Tatiana Filatenkova, Olga Shamova, Valentin Semenov, Vitaly Panchuk, and Dmitry Kirsanov: Contactless chemical analysis with high-frequency inductance coil and chemometrics

P23 Kurt Varmuza, M. Dehmer, P. Filzmoser: Molecular descriptors based on automorphism data

P24 Kurt Varmuza, P. Filzmoser: Adjusted Pareto scaling

P25 N. Vladimirova, E. Puchkova, D. Dar'in, A. Turanov, V. Babain and D. Kirsanov: Application of quantitative structure-property relationship (QSPR) in predicting potentiometric sensor sensitivity to heavy metals

Lectures

Confounding in model interpretation when using latent variable regression methods to make inferences

Olav M. Kvalheim[1], Warren S. Vidar[2], Tim U.H. Baumeister[3],
Roger G. Linington[3] and Nadja B. Cech[2]
[1]Department of Chemistry, University of Bergen, Norway
[2]Department of Chemistry and Biochemistry, University of North Carolina at Greensboro, Greensboro, North Carolina, United States
[3]Department of Chemistry, Simon Fraser University, Burnaby, BC, Canada

Latent variable regression (LVR) methods, such as partial least squares (PLS), are extensively used to model relations between a suite of explanatory variables and one or more outcome variables. Often, there is no assumption of or need for that the explanatory variables should be responsible or causally linked to the outcome. A model with excellent prediction ability can anyway be obtained if there are explanatory variables with good correlation to the outcome. However, if the objective is to use the associations of the explanatory variables to the outcome to generate hypothesis and to make inferences, the situation is more demanding. In this case, we assume that the explanatory variables are responsible for the systematic variation in the outcome and that interpretation of the association pattern between outcome and explanatory variables are meaningful and can be used to achieve the purpose of our investigation.

Investigations of botanical to discover new bioactive compounds represent an application area where the predictive performance of a model is necessary, but not sufficient to be useful for its purpose. Such investigations can be conducted on whole or fractionated botanical extracts. Mass spectrometry is frequently used for profiling samples and the concentrations of the compounds are used as explanatory variables to predict the measured bioactivity. But since the number of measured analytes is usually higher than the number of samples, or, more precisely, higher than the number of underlying latent variables necessary to establish a validated model with good prediction performance, confounding may present a problem. Thus, bioactive candidates obtained by interpreting the association pattern of analytes to the measured bioactivity, may be the result of confounding patterns and not point to true bioactive analytes. The problem is exaggerated by the possibility of strong synergistic or antagonistic behavior between analytes. Thus, models including interactions between analytes should also be examined. This increases the complexity of the model since the number of possible interactions is high. Recently, we published a strategy for reducing the confounding problem in models of botanical extracts with the purpose to reveal bioactive candidates [1]. In this work, we describe the method in a more general and theoretical context and make comparisons with factorial screening designs.

References

[1] Warren S. Vidar, Tim U. H. Baumeister, Lindsay K. Caesar, Joshua J. Kellogg, Daniel A. Todd, Roger G. Linington, Olav M. Kvalheim, and Nadja B. Cech, Interaction Metabolomics to Discover Synergists in Natural Product Mixtures, *J. Nat. Prod.* **86(**4), (2023), 655-671.

Patch seriation to visualize data and model parameters

Gergely Tóth and Rita Lasfar
Institute of Chemistry, Eötvös Loránd University,
1117 Budapest, Pázmány s. 1/a, Hungary, E-mail gergely.janos.toth@ttk.elte.hu

Seriation means how to do row and/or column permutations to enhance visual perception of a table or its representation on a heatmap. Seriation might provide first ideas on hidden clusters, probable classifications, trends, linear or non-linear modelling possibilities or frequency of data values. At least, the confusing stripes of a heatmap can be removed by the rearrangement. Seriation is a non-destructive data analysis, only the row and the column orders are changed [1].

In our presentation we show a new seriation merit function. At first, a local similarity matrix is calculated, where the average similarity of a neighbouring objects is calculated in a limited variable space. Thereafter, a sum-like function is constructed to maximize the local similarities and cluster them into patches by row and column ordering. Our method identifies data clusters in a powerful way, if the similarity of objects is caused by some variables and these variables differ for the distinct clusters. The method can be used in the presence of missing data and also on more than two-dimensional data arrays. In our talk we show the feasibility of the method on different data sets: on QSAR, chemical, material science, food science, cheminformatics and environmental data in two- and three-dimensional cases.

The method can be used during the development and the interpretation of artificial neural network models by seriating different features of the models. It helps to identify interpretable models by elucidating clusters of objects, variables and hidden layer neurons. An example is shown here on neural network models from our recently submitted article [2].

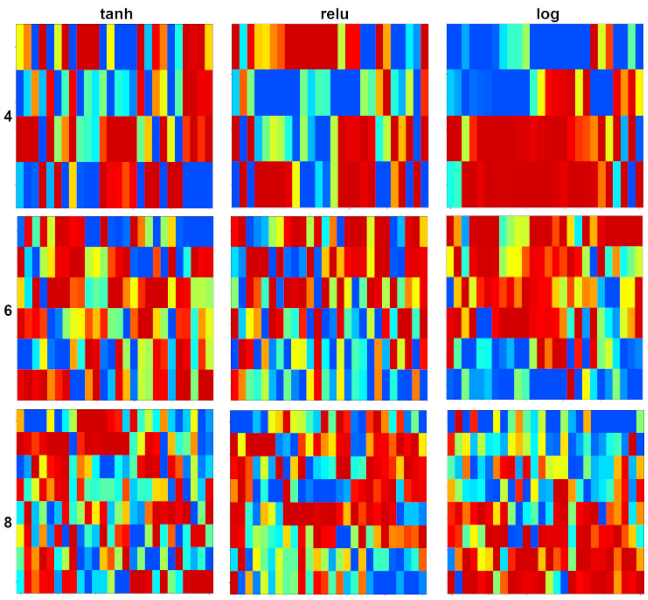


Figure: Seriated matrices. The objects are the neurons and the variables are the scaled weights of the original input variables. Three activation function are used (tangent hyperbolic, relu and logistic) with 4, 6 or 8 neurons in the hidden layer.

References

[1] I. Liiv I *Stat. Anal. Data Min.*, **3** (2010) 70-91.

[2] R. Lasfar and G.Tóth, (2023) https://doi.org/10.21203/rs.3.rs-2780120/v1 submitted

Automatic detection of outlying compounds in the OCHEM platform using consensus modelling

Igor V. Tetko[1]

[1]Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz Munich-Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), DE-85764 Neuherberg, Germany and BIGCHEM GmbH, DE-85716 Unterschleißheim, Germany

The quality of data is crucial for development of quantitative structure activity relationship models. Experimental errors, as well as random errors associated with data collection steps (such as mistakes in molecular structures, incorrect measurement units, or wrong property assignment, etc.) are particularly detrimental to model quality. The presence of such errors markedly decreases the performance of these models. In cases where several measurements are available for the same compound; a common practice is to use a median or mean value for modelling. However, this method cannot be easily applied in cases where the target property is heavily dependent on a specific condition, e.g., temperature for vapor pressure, unless all the measurements were taken under the same conditions.

If references to the original data sources are available, one can manually verify the data and units and in doing so, significantly lower the potential for errors. If original data sources are not available, we can assume that errors in a model follow a normal distribution and simply exclude predictions with large error. This leads to improvements in the accuracy of predictions for some compounds after recalculation of the model, and reduction of errors. Such a procedure was used to detect outlying compounds in the prediction of melting points [1]. The pool of excluded compounds was significantly enriched with compounds that were predicted to decompose before melting. However, we did not explicitly demonstrate if the proposed method could decrease prediction errors for new compounds.

In this study, vapour pressure data available within OCHEM was analysed together with the data collected during the life CONCERT REACH project. Consensus models were developed for each of the two datasets and were used to predict data from the complementary sets, which were used as respective blind test sets. We show that filtering of outliers significantly decreased errors for the blind sets (by 0.3-0.1 log units). Moreover, the models provided even lower errors when applied to the filtered sub-sets. The use of temperature as the parameter of the models allowed us to account for this condition and identify outlying values irrespective of the measurement temperature.

The advantages of using a consensus model as opposed to an individual model to identify outlying molecules will be presented. Several other examples of the consensus modelling which resulted in winning top ranked models for several challenges will be also discussed.

Acknowledgement

This study was partially funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions, grant agreement "Advanced machine learning for Innovative Drug Discovery (AIDD)" No. 956832 as well as by Life CONCERT REACH project LIFE17 GIE/IT/000461.

References

[1] I.V. Tetko, D.M. Lowe and A.J. Williams, *J. Cheminformatics,* **8** (2016), 1-18.

Multivariate curve resolution with rank deficiency

T. Andersons[1], M. Sawall[1], K. Neymeyr[1,2]
[1] Universität Rostock, Ulmenstraße 69, 18057 Rostock, Germany;
corresponding author: tomass.andersons@uni-rostock.de
[2] Leibniz-Institut für Katalyse, Albert-Einstein-Straße 29a, 18059 Rostock, Germany

Multivariate Curve Resolution (MCR) methods serve to decompose a spectral data matrix $\boldsymbol{D}$ into the concentration profiles $\boldsymbol{C}$ of the pure components and the associated spectra $\boldsymbol{S}$. MCR methods work when the data has a bilinear structure, that is, $\boldsymbol{D}=\boldsymbol{C}\boldsymbol{S}^{\mathbf{T}}$. The ambiguity of the pure components can be represented in a low-dimensional way by the Area of Feasible Solutions (AFS). The dimension of the AFS is equal to the rank of the matrix $\boldsymbol{D}$ minus one, and usually the rank of $\boldsymbol{D}$ matches the number of chemical components. However, some chemical reactions, *e.g.*, Michaelis-Menten kinetics, have an inherent rank deficiency. Typically, rank deficiency is caused by linear dependencies in the concentration profiles. This obscures the true chemical structure of the problem. Therefore, classical MCR approaches and AFS cannot be directly applied.

Our approach to rank-deficient MCR problems uses numerical [1] and geometric [2] methods to find a low-dimensional representation of feasible concentration profiles. Thus, a band of feasible concentration profiles for a rank-deficient MCR problem can be computed without the need for additional measurements. Furthermore, it is possible to recover some information about the full-rank factor, that is, spectral profiles [3]. These methods help to overcome the loss of information due to linear dependencies and tackle the inherent difficulties in analyzing rank-deficient problems.

References

[1] M. Sawall and K. Neymeyr, *J. Chemom.*, **35** (2020) e3316.
[2] M. Sawall, T. Andersons and K. Neymeyr, *Chemom. Intell. Lab. Syst.*, **235** (2023) 104782.
[3] M. Sawall, T. Andersons, H. Abdollahi, S. Khodadadi Karimvand, B. Hemmateenejad and K. Neymeyr, *Chemom. Intell. Lab. Syst.*, **226** (2022) 104577.

Tailoring MultiCriteria Decision Making methods optimally for individual data sets

Sándor Kovács[1], Attila Gere[2], Károly Héberger[3],*
[1] Faculty of Economics and Business, University of Debrecen, 4032 Debrecen, Hungary
[2] Institute of Food Science and Technology, Hungarian University of Agriculture and Life Sciences, Budapest, Hungary, E-mail: gereattilaphd@gmail.com
[3] Plasma Chemistry Research Group, Institute of Materials and Environmental Chemistry,
Research Centre for Natural Sciences, Institute of Excellence,
Hungarian Academy of Sciences, Budapest, Hungary
E-mail: heberger.karoly@ttk.hu

Selection of an appropriate Multicriteria Decision Making (MCDM) method is always a challenging task. A plethora of MCDM techniques is available, all of them are planned to select optimal rankings [1-4], still they often produce contradictory (conflicting) results. As the idiosyncrasies of data sets are different and change from case to case, assigning proper (best) technique has not been solved yet: Any solutions have not gained general acceptance to say the least. Three highly different data sets were selected carefully to illustrate the difficulties and visualize the necessary steps: i) a short one with (almost) unambiguous ranking, ii) a medium one with definite bipolar clustering and iii) a larger one with (partly) contradictory ordering.

First, the data structures were studied; then, eleven MCDM techniques were applied to rank the objects of each data set. The direction (optimization) was decided as hypothetical best estimation: maximum for vitamin content of worms (insect), minimum for errors for classifiers (eye) and both for methods to determine pseudorank (chemical rank, Todeschini T5).

Several visualization techniques revealed the inner structure of the ranks: line plots, parallel coordinates, sum of ranking differences (SRD) [4]. The MCDM techniques are grouped and the best ones are selected according to the individual data sets.

Finally, the three date sets were amalgamated into one composite table using Generalized Procrustes analysis (GPA) [5] and overall conclusions are drawn.

Acknowledgement
The research was supported by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund (OTKA, contracts No K 134260 and FK 137577).

References

[1] A. Hafezalkotob, A. Hafezalkotob, H. Liao and F. Herrera, *Inform. Fus.* **51**, (2019) 145-177.
[2] R. Kolde, S. Laur, P. Adler and J. Vilo, *Bioinformatics*, **28** (2012) 573-580.
[3] R. Todeschini, F. Groni and D. Ballabio, *Chemometr. Intell. Lab. Syst*. **191** (2019) 129-137.
[4] B. Roy and D. Vanderpooten, J. Multi-Criteria Dec. Anal., **5** (1996) 22-38.
[5] K. Héberger and K. Kollár-Hunek, *J. Chemometr.*, **25** (2011) 151-158.
[6] J. C. Gower, *Psychometrika* **40** (1975) 33-51.

Multiobjective optimization of effect directed planar chromatography as a tool for fast screening of polypotent natural products

Filip Andrić[1],*
[1] University of Belgrade - Faculty of Chemistry, Studentski trg 12-16,
Belgrade Serbia, *E-mail: andric@chem.bg.ac.rs

Chemical complexity of natural products often results in their pharmacological polypotency. However, selecting a natural product with desirable activity profile is not a straightforward task, especially if optimization of one feature results in deterioration of other facets.

High-performance thin-layer chromatography provides chromatographic fingerprints of natural products in short time, with low costs. Additionally, the separated components on the plate retain their biological activity, which enables HPTLC to be easily coupled with various biological activity assays.

In the present work three different multiobjective optimization algorithms: (a) Derrigner's desirability approach [1], (b) Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) [2], and (d) Sum of Ranking Differences (SRD) [3] have been applied on the HPTLC data collected from the literature [4], and their ranking outcomes of natural products were compared. Furthermore, linear regression and classification algorithms combined with jackknife cross-validation approach were further applied on the preference scores in order to isolate chromatographic features characteristic for the most optimal (polypotent) natural products.

References

[1] G. Derringer, R. Suich, *J. Qual. Technol.* **12(4)** (1980) 214-219.
[2] K. Yoon, *J. Oper. Res. Soc.*, **38** (1987) 277-286.
[3] K. Héberger, K. Kollár-Hunek., *J. Chemom.*, **25 (**2010) 151-58.
[4] G. E. Morlock, J. Heil, A. M. Inarejos-Garcia and J. Maeder, *Antioxidants*, **10** (2021) 117.

Changes of fatty acids' levels in inflammatory bowel diseases in view of the samples collected within the Integrative Human Microbiome Project

Tomasz Bączek*
Department of Pharmaceutical Chemistry, Medical University of Gdańsk,
Hallera 107, 80-416 Gdańsk, Poland,* E-mail: tbaczek@gumed.edu.pl

Data available for 30 fatty acids as part of the Integrative Human Microbiome Project performed for 132 subjects to generate molecular profiles of host and microbial activity during disease were processed. Identification of statistically significant differences found in men and women in the case of both Crohn's disease and ulcerative colitis in comparison to non-inflammatory bowel disease samples was performed. The focus was on statistical analysis of selected data available through the Inflammatory Bowel Disease Multiomics Database at the IBDMDB website (https://ibdmdb.org) [1]. Non-parametric test (Kruskal-Wallis test) was used for the analysis. Post-hoc comparisons focused on by sex or disease and the significance of the differences was done with the FDR correction test. Considering women, statistically significant differences between non-inflammatory bowel disease samples (*n* = 62) and Crohn's disease samples (*n* =116) were found for: adrenate, arachidonate, caproate, docosahexaenoate, docosapentaenoate, eicosadienoate, eicosapenaenoate and eicosatrienoate. In the case of comparisons between non-inflammatory bowel disease samples (*n* = 62) and ulcerative colitis samples (*n* = 93) for women, statistically significant differences were noted exactly for the same metabolites.

Considering men, statistically significant differences between non-inflammatory bowel disease samples (*n* = 73) and Crohn's disease samples (*n* = 149) were found for: 10-nonadecenoate, 17-methylstearate, 3-hydroxyoctanoate, adrenate, arachidonate, caproate, docosahexaenoate, docosapentaenoate, eicosadienoate, eicosapenaenoate, eicosatrienoate, myristoleate, palmitate and palmitoleate. In the case of comparisons between non-inflammatory bowel disease samples (*n* = 73) and ulcerative colitis samples (*n* = 53) for men, statistically significant differences were noted for the same set of metabolites except in-here of the lack of 17-methylstearate, 3-hydroxyoctanoate, myristoleate, palmitate and palmitoleate. For all tested groups the following 8 short-chain fatty acids were always the most statistically significant: adrenate, arachidonate, caproate, docosahexaenoate, docosapentaenoate, eicosadienoate, eicosapentaenoate, eicosatrienoate.

References

[1] J. Lloyd-Price, C. Arze, A.N. Ananthakrishnan, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature,* **569** (2019) 655-662.

Near-infrared and Raman chemical imaging for prediction of immediate dissolution of acetylsalicylic acid tablets

A. Farkas[1], Á. Kopasz[1], T. Jámbor[1], E. Varga[1], B. Nagy[1]
[1] Department of Organic Chemistry and Technology, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary, E-mail: farkas.attila@vbk.bme.hu

In the new era of the pharmaceutical industry, the more complete monitoring of the processes and quality control is particularly emphasized. The new approach of Process Analytical Technology requires developing novel in-, on- and at-line analytical techniques and data collection to better understand the processes, recognize the deviations, and control the product quality. With non-destructive procedures such as NIR and Raman spectroscopy instead of traditional analytical methods, it would enable the examination of entire production batches without random sampling. Furthermore, this approach would prevent any damage to the batches during the analysis. Our previous studies have proven that Raman chemical imaging can effectively predict the dissolution curve of sustained-release tablets considering hydroxypropyl methylcellulose content and particle size [1,2].

In the present study, NIR and Raman chemical images were collected about acetylsalicylic acid tablets. The acquisition time of NIR and Raman chemical images was 1 and 13 minutes, respectively, which proved to be sufficient in terms of representativeness.

During tablet pressing, the particle size of the active ingredient, the applied compression force and the quantity of disintegrant were varied according to an experimental design. The porosity of the tablets was also measured. Then the tablets were dissolved following USP standards and monitoring the API concentration by UV detection for 2 hours. The points of the curves created the outputs of feed-forward neural networks. The chemical images had to be processed for data reduction before using them as inputs for ANNs. The neural networks were optimized by varying the number of neurons in the hidden layer. The trained ANNs could effectively predict the dissolution curve, comparing similarities by f2 values (>50).

References

[1] D.L. Galata, B. Zsiros, L.A. Mészáros, B. Nagy, E. Szabó, A. Farkas and Z.K. Nagy, *J. Pharm. Biomed. Anal.*, **212** (2022) 114661

[2] D.L. Galata, B. Zsiros, G. Knyihár, O. Péterfi, L.A. Mészáros, F. Ronkay, B. Nagy, E. Szabó, Z.K. Nagy and A. Farkas, *Int. J. Pharm.*, **640** (2023) 123001

Prediction of physical and chemical features of structural materials with machine learning and classical chemometric tools, based on IR spectroscopy

Szilvia Klébert[1], Dóra Tátraaljai[2], Krisztina László[2], Anita Rácz[1]
[1] Plasma Chemistry Research Group, Institute of Materials and Environmental Chemistry,
Research Centre for Natural Sciences, Magyar tudósok krt. 2., 1117 Budapest, Hungary
E-Mail: racz.anita@ttk.hu
[2] Department of Physical Chemistry and Materials Science, Faculty of Chemical Technology
and Biotechnology, Budapest University of Technology and Economics,
1521 Budapest, Hungary;

Structural materials nowadays are part of our everyday life, their application areas are endless and they can be found everywhere from our electronical devices to cars, buildings and furniture. Moreover, in the past decades there is a growing need for environmentally friendly, "green" materials (especially plastics) for specific application areas, which can be easily recycled and have no threat to human health and the ecosystem. Therefore, the chemical, physical and mechanical analysis and primary design of these materials play an important role, but are usually a time-consuming and expensive processes. To enhance the determination of the physical and chemical properties of the materials, one can use fast and eco-friendly spectroscopic methods or represent the samples with several computational descriptors for in-silico modeling with machine learning algorithms. Along the evolution of new material science databases with reference measurements and large-scale simulations, and with the highly increased computational power, in-silico modeling with machine learning (ML) became a possibility of huge promise in material science [1].

Our aim was to predict several chemical, physical and mechanical properties of polymers and polymer composites accurately and energy-efficiently, moreover the physical properties of biochar, which is a promising candidate carbon-negative resource for cement substitution in construction materials. For this purpose, we are combining FTIR measurements with "classical" and machine learning algorithms as well. The two different segments of structural materials represent well the importance and usefulness of the chemometric models (classification and regression) in this field, which is still often neglected. The developed models are properly validated and have great performance values; therefore, they are appropriate alternatives for the current exhaustive determinations of the different properties of structural materials.

Reference

[1] Chan, C.H.; Sun, M.; Huang, B. Application of Machine Learning for Advanced Material Prediction and Design. EcoMat 2022, 4, https://doi.org/10.1002/eom2.12194

Features of Hotelling's T^2 on simulated and metbolomics data: A tutorial

Richard G Brereton
School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, U.K.
E-mail: richard.brereton1@gmail.com

Hotellings T^2 is widely used throughout chemometrics, for example to define confidence or control limits or to determine p values. The statistic is just a Mahalanobis distance, and the use of this can only be understood via the probability density and cumulative probability distributions and associated p values. It can be used to identify outliers or whether a sample is likely to be a member of a predefined class. It can be employed in common methods such as QDA or SIMCA.

The distribution is introduced and its relationship with z (normal), t, χ^2 and *F* distributions described.

The main focus will be on a comparison between T^2 and χ^2 which can be used to calculate *p* values that a sample is a given distance or greater from the centre of a multivariate distribution. It is shown that χ^2 will usually well describe this. The assumption of normality can be checked by a rank graph of *p* values versus rank, which should be roughly linear unless there are outliers or other groups. If an appropriate number of PCs are chosen for the model useful *p* values are obtained and a near linear rank graph is obtained.

Hotelling's T^2 though is found to be very sensitive to the number of degrees of freedom in the model, and if inappropriate can result in highly misleading p values. It is often hard to estimate the number of degrees of freedom accurately, as later dimensions may not be very significant, and the assumption of normality breaks down when later PCs correspond primarily to noise. χ^2 is much more robust as there is only a single degree of freedom involved.

Using both simulated and real data, it is shown that T^2 as normally applied can result in very misleading estimates.

It is recommended not to use T^2 unless the number of samples is very high (in which case it resembles χ^2) or the number of variables is small (for example 2 or 3 PCs). For safe predictions χ^2 is recommended.

Data science: chemometric techniques generating business value

Balázs Vajna[1]
[1] MarketingLens Ltd., E-mail: balazs@marketinglens.com

This presentation aims to provide insight into how certain popular statistical methods that are widely used in chemometrics are also used in the "business world", with the field being labelled as "analytics", "data science", "machine learning", and these days also "artificial intelligence".

The talk will start with a characterization of different sub-fields within data & analytics, and show where chemometric methods (e.g. PCA, clustering, classification, regression) are usually used. This will be followed by listing verticals (e.g. telecommunications, health and public service, retail and e-commerce etc.) and various business areas where "chemometrics" (or data science) is widely adopted, with a number of practical use cases described in more detail such as fraud detection, churn prediction, cross- and up-sell modelling and recommendation engines.

Finally, the key differences will be shown between a "data science project" and a "chemometrics study", especially around the data side. The data preprocessing steps are significantly different: while chemometric evaluation is usually applied on data that is generated by an analytical device, in business use cases generating the data table is the largest and longest part of a data science project, as it has to be created (joined, aggregated and post-processed) from a number of "source" tables with varying granularity and data types.

Chemometrics and "data science" are not vastly different fields; in fact future conversations initiated from this talk may potentially point out areas/approaches that are well established in chemometrics but are under-utilized in business data analytics.

Hit Discovery using Docking ENriched by GEnerative Modeling (HIDDEN GEM): A Novel Computational Tool for Accelerated Virtual Screening of Ultra-large Chemical Libraries

K. I. Popov[1], J. Wellnitz[1], T. Maxfield[1], and A. Tropsha[1]
[1]UNC Eshelman School of Pharmacy, University of North Carolina,
Chapel Hill, NC, USA E-mail: alex_tropsha@unc.edu

The recent growth of make-on-demand, purchasable, chemical libraries comprising dozens of billions if not trillions of molecules has proved beneficial for drug discovery but has also challenged the efficient application of traditional structure-based virtual screening methods. We present a novel, highly computationally efficient approach dubbed HIDDEN GEM (HIt Discovery using Docking ENriched by GEnerative Modeling) for structure-based VS of ultra-large chemical libraries. HIDDEN GEM uniquely integrates and leverages the advantages of all the aforementioned methods including molecular docking, machine learning, and generative modeling while attempting to circumvent their limitations. This novel workflow starts by using traditional molecular docking of a small, chemically diverse chemical library. The docking results are then used to bias a pretrained generative model as well as train a filtering model to create and select novel compounds with better docking scores. The subset of these *de novo* designed molecules confirmed by docking to have high scores are used for massive chemical similarity searching to identify a small set of highly similar, purchasable compounds in any ultra-large library. This small set of molecules selected from the purchasable library can then be docked and scored for nomination as hits or used to iteratively train a new series of generative models. These steps, including the generative modeling, filtering, and similarity searching can be repeated several times if needed. We show that iteration between generation and similarity searching helps HIDDEN GEM rapidly focus on the chemical space with top scoring compounds to deliver both in-library and *de novo* virtual hits.

We evaluated the performance of HIDDEN GEM on sixteen different protein targets from various families using the 37 billion-sized Enamine REAL Space library. We have shown that for all sixteen targets HIDDEN GEM identified sets of hits with docking scores enriched up to 1000-fold over scores from a random search of the chemical library. Surprisingly, these results hold for even a single cycle of generative modeling and similarity searching. The virtual screening for each target was completed in as little as two days leveraging only a single 44 CPU-core machine for docking, an 800 CPU-core computing cluster for similarity searching, and one Nvidia GTX 1080 Ti GPU for generation. We posit that due to high computational efficiency and universal success in identifying the high-scoring hits for diverse targets, HIDDEN GEM offers an attractive alternative to all current VS methods.

Explainability approaches in computer-aided drug design

Agnieszka Wojtuch[1], Ewelina Jamrozik[1], Tomasz Danel[1], Sabina Podlewska[2]
[1] Faculty of Mathematics and Computer Science, Jagiellonian University,
Łojasiewicza 6, 30-348, Kraków, Poland
[2] Maj Institute of Pharmacology, Polish Academy of Sciences,
Smętna 12, 31-343, Kraków, Poland
E-mail: smusz@if-pan.krakow.pl

Great progress in the development of computational strategies for drug design application has revolutionized the process of the search for new drugs [1]. Although the focus of *in silico* strategies is still put on the provision of the desired activity of a compound to the considered target, also the characterization of a compound in terms of its physicochemical and ADMET properties has become an indispensable element of computer-aided drug design protocols [2].

As the initially designed ligands usually undergo optimization to tune its activity and properties, compound evaluation in terms of different properties is not sufficient and there is a desire for development of methods which will support the optimization procedures [3]. Such guidance can be realized e.g., by the provision of the factors which influence particular compound property to the highest extent.

During the presentation, I would like to summarize outcome of our studies oriented at application of explainable predictive modelling to the selected ADMET properties of compounds. The main focus will be put to the Shapley Additive exPlanations (SHAP) revealing how the particular compound substructure influences prediction of compound metabolic stability [4]. Moreover, the application of LIME and counterfactual analysis in the evaluation and optimization of compound cardiotoxicity, genotoxicity, solubility, biological membranes permeability and plasma protein binding will be presented.

All developed solutions are made available to wide community via on-line applications. Their possibilities of their direct service and intuitive graphical user interface made them easily usable by medicinal chemists.

Acknowledgments: The study was supported by the grant OPUS 2018/31/B/NZ2/00165 financed by the National Science Centre, Poland (www.ncn.gov.pl).

References

[1] G. Sliwoski, S. Kothiwale, J. Meiler, and E. W. Lowe, Jr. *Pharmacol Rev.* **66** (2013), 334-395.

[2] T. T. Van Tran, A. S. Wibowo, H. Tayara, and K. T. Chong I. *J. Chem. Inf. Model.* **63** (2023), 2628-2643.

[3] J. Jiménez-Luna, F. Grisoni, and G. Schneider *Nat. Mach. Intell.* **2** (2020), 573-584.

[4] A. Wojtuch, R. Jankowski, S. Podlewska *J. Cheminf.* **13** (2021), 74.

Drug repurposing for Covid-19 therapy: *in silico*, *in vitro*, *in vivo* and *in clinics*

Vladimir Poroikov[1],* Dmitry Druzhilovskiy[1], Nadezhda Biziukova[1], Oleg Gomazkov[1],
Alexander Veselovsky[1], Alexander Dmitriev[1], Sergey Ivanov[1], Nikita Ionov[1],
Dmitry Karasev[1], Anastassia Rudik[1], Polina Savosina[1], Boris Sobolev[1],
Leonid Stolbov[1], Vladislav Sukhachev[1], Olga Tarasova[1], Dmitry Filimonov[1]
[1] Institute of Biomedical Chemistry, Bldg. 8, 10 Pogodinskaya Street, Moscow, 119121,
Russia; *E-mail: vvp1951@yandex.ru

The pandemic of a new coronavirus infection has a significant impact on all aspects of human activity. Over the past three years, scientific knowledge about SARS-CoV-2 and the progression of the infectious process of COVID-19 has expanded significantly. However, the lack of knowledge about the SARS-CoV-2 virus and the mechanisms of development of the pathological process requires further basic and exploratory biomedical research, identification of the main cellular and molecular targets for tissue and organ damage, and the search for new treatments and prevention of coronavirus infection.

Drug repositioning – identifying new indications for approved pharmaceuticals – is the only possible immediate response to the COVID-19 pandemic and future biogenic threats. The search for new pharmacological effects of known drugs is carried out in silico and in vitro. To evaluate in silico characteristics of repositioned drugs and new pharmacological substances, web services have been implemented that provide a search for structural analogues of active compounds among 4000 drugs approved for medical use, predict anticoronavirus effects, side effects and toxicity for molecules planned for synthesis, etc. An analysis of the available experimental data on in vitro testing of the anticoronavirus activity of known drugs allowed to establish priorities for their further research. Using molecular modelling, fifteen drugs (disulfiram, omeprazole, silibinin, saquinavir, etc.) have been selected as potential inhibitors of SARS-CoV-2 protease 3CLpro. Anticoronavirus activity for narlaprevir has been confirmed by the experiment (IC50=2.75 µM, EC50=64 µM, CC50=106 µM). Imatinib inhibited virus replication in cell culture with EC50 = 40.0 µM but was practically inactive against the main protease 3CLpro.

We took part in the international project "JEDI Billion Molecules Against COVID-19 Grand Challenge" (https://www.jedi.foundation/covid19challenge). 130 teams from different countries offered 639,024 hits for synthesis and testing; 820 compounds synthesized and tested; 28 "actives" found (success rate 3.19%). We performed in silico screening among 1.08 billion structures on 4 targets (3CLpro, PLpro, RdRp, TMPRSS2); ten thousand hits were chosen. We were among the 20 teams whose proposals were selected for experimental verification; 36 molecules synthesized; the activity of one molecule (PLpro inhibition) was confirmed in the experiment.

Opportunities and limitations of drug repositioning in the context of the COVID-19 pandemic and ways to reduce the risks of new biogenic threats in the future will be discussed.

Acknowledgement
The study is performed in the framework of the Program for Basic Research in the Russian Federation for a long-term period (2021-2030) (No. 121102900156-6).

The role of chemometrics in making analytical chemistry green

Mikhail Saveliev[1], Vitaly Panchuk[1,2], Dmitry Kirsanov[1]
[1] Institute of Chemistry, St. Petersburg University, St. Petersburg, Russia,
E-mail d.kirsanov@gmail.com
[2] Institute for Analytical Instrumentation RAS, St. Petersburg, Russia

The environmental pollution caused by human activities is one of the major concerns in the modern society and all aspects of this pollution are currently attracting serious attention from researcher. Chemistry, being not very environmentally friendly in its' basis, including analytical chemistry, is also in the spotlight.

Chemometric data processing nowadays is a very important part of analytical chemistry. Application of chemometric modelling in real analytical problems provides for a number of valuable benefits, like reliable qualitative and quantitative analysis in case of non-ideal, overlapped, and poorly selective signals. In this way chemometrics helps in elimination of complex physical manipulations with samples and instruments, thus leading to significant simplification and cost reduction of analytical procedures. And this, in turn, leads to a considerable green impact of chemometrics. Obviously, the motto that chemometrics is green per se is quite popular and various authors have already highlighted this in the recent literature [1,2]. At the same time there are not so many reports where the particular green impact of chemometrics would be characterized numerically. In order to do so one would need corresponding greenness metrics–these were already established in literature with a particular focus on assessing the greenness of analytical methods. These metrics are based on the calculation of certain parameters that allow the numerical or graphical evaluation of the environmental friendliness of the employed analytical procedures. A very popular index is Analytical Eco-Scale method [3], allowing numerical evaluation of the greenness from 0 to 100, where 100 is the so-called "perfectly green analysis". Each component of the analytical procedure (danger and toxicity of reagents, amount of waste, energy consumption of the analysis, etc.) is characterized with certain penalty points, the sum of which is subtracted from 100. The resulting value characterizes the greenness numerically, the higher it is, the more environmentally friendly the technique in question is.

In this study we have evaluated Eco-Scale greenness score for a number of representative studies from recent literature and we have compared chemometric-based analytical procedures with their traditional counter-parts in terms of greenness. A combination of chemometric algorithms with inexpensive spectroscopic, electroanalytical and other techniques eliminates the need in applying dangerous and environmentally unfriendly reagents and procedures. It is shown that Eco-Scale score in addressing particular analytical tasks can be substantially improved through the use of chemometric-based methods.

References

[1] H.-W. Gu, S.-H. Zhang, B.-C. Wu, W. Chen, J.-B. Wang, Y. Liu, *Spectrochim. Acta A* **200** (2018) 93-101.

[2] A. P. Rebellato, E. T. dos S. Caramês, P. P. de Moraes, J. A. L. Pallone, *LWT*, **128** (2020) 109438.

[3] A. Gałuszka, Z. M. Migaszewski, P. Konieczka, J. Namieśnik, *TrAC, Trends Anal. Chem.* **37** (2012) 61-72.

Characterizing solvent composition in a CO_2 capture plant using multivariate data analysis of online sensors and spectroscopic data

Bjørn Grung
Department of Chemistry, University of Bergen, Norway
E-mail: bjorn.grung@uib.no

Discussions on the role of the greenhouse gas CO_2 on the earth's climate have being going on for many years. This presentation will not take part in that debate, but rather focus on *how* chemistry and physics can be used to reduce the amount of CO_2 in the atmosphere. Financial aspects will be presented, as will the historical development (or lack thereof) of carbon capture facilities.

As carbon capture processes are complex and often not understood in detail it should not come as a surprise that chemometrics can contribute to monitoring, modelling, and controlling such processes. The University of Bergen has had a long collaboration with Technology Centre Mongstad, the world's largest testing facility for CO_2 capture. One part of this plant uses amines to bind the CO_2. A large part of the monitoring of the process state depends on purely offline analysis, with results being presented hours or days after sampling. An obvious improvement would be to use multivariate modelling to predict responses online, thus rapidly characterizing the process state. These models have been made from traditional process variables alone, spectroscopic data alone and data fusion models.

A strategy for source apportionment analysis

I. Stanimirova[1,2], P.K. Hopke[2,3]
[1]Institute of Chemistry, University of Silesia in Katowice, Katowice, 40-006, Poland
E-mail: istanimi@us.edu.pl
[2]Department of Public Health Sciences, University of Rochester
Medical Center, Rochester, NY, USA
[3]Institute for Sustainable Environment, Clarkson University, Potsdam, NY, 13699 USA

Information about the pollution sources and evaluating their contributions to the ambient air pollution concentrations is essential for developing or assessing any effective air quality strategy [1]. These insights can be obtained by using emission inventories/source-oriented, or receptor-oriented models. Although emission inventories and source-oriented models, which include Eularian grid with chemistry, artificial neural networks, Gaussian models, and Lagrangian particle dispersion models, among others, are continuously being improved, they still lack the capability to fully describe the relationships between pollution sources and the sites of interest (receptors) that exists in the air [2]. This possibility does arise when analyzing the multivariate measurement data that contains the concentrations of pollutants that are measured at various receptors under the assumption that the total mass of a pollutant in a given time period is the linear sum of the independent contributing sources. The currently most popular receptor-oriented modeling method is positive matrix factorization, PMF.

In this work, we present a general strategy for source apportionment using PMF. Specifically, dispersion-normalized PMF, DN-PMF [3], was used to account for meteorological dilution and thereby to improve the results and to reduce rotational ambiguity. Next, to investigate the long-term seasonal trends, directional dispersion of pollutants depending on the wind speed and wind direction as well as to find the regional locations of pollutant emissions for each of the resolved sources, methods such as the seasonal-trend decomposition procedure combined with locally weighted regression, loess (STL) [4], conditional bivariate probability function (CBPF) and back trajectory analysis [5] were applied.

References

[1] C.A. Belis, B.R. Larsen, F. Amato, I. El Haddad, O. Favez, R. M. Harrison, P.K. Hopke, S. Nava, P. Paatero, A. Prevot, U. Quass, R. Vecchi, M. Viana, European guide on air pollution source apportionment with receptor models, *JRC Reference reports*, 2014.

[2] P.K. Hopke, *J. Air & Waste Manag. Assoc.* **66** (2016) 237-259.

[3] Q. Dai, B. Liu, X. Bi, J. Wu, D. Liang, Y. Zhang, Y. Feng, P.K. Hopke, *Environ. Sci. Technol*. **54** (2020) 9917-9927.

[4] R.B. Cleveland, W.S. Cleveland, I. Terpenning, *J. Off. Stat.* **6** (1990) 3-73.

[5] A.F. Stein, R.R. Draxler, G.D. Rolph, B.J.B. Stunder, M.D. Cohen, F. Ngan, *Bull. Am. Meteorol. Soc.* **96** (2015) 2059-2077.

Chemometric modelling of vital soil parameters

M. Daszykowski[1], L. Pieszczek[1], I. Stanimirova[1], S. Krzebietke[2], H. Czarnik-Matusewicz[3]
[1] Institute of Chemistry, University of Silesia in Katowice,
9 Szkolna Street, Katowice Poland; E-mail: michal.daszykowski@us.edu.pl
[2] Department of Clinical Pharmacology, Faculty of Pharmacy, Wrocław Medical University, 211a Borowska Street, 50-556 Wrocław, Poland
[3] Department of Agricultural and Environmental Chemistry, Faculty of Agriculture and Forestry, University of Warmia and Mazury in Olsztyn,
8 Oczapowskiego Street, 10-719 Olsztyn, Poland

Soil is a very complex natural medium of crucial importance and a precious resource. It is a source of nutrients for plants, assists in food production, stores carbon, plays a significant role in reducing greenhouse gas emissions, and acts as a water filter. Various chemical parameters characterize soil health, overall condition, and fertility. Therefore, efficient monitoring of selected parameters is strongly desired but remains a challenge.

Recommended wet chemistry methods for soil analysis require extensive sample preparation, are time-consuming, and return a single measurement per analysis. On the other hand, spectroscopic methods, combined with advanced chemometric modeling of various spectra, offer an attractive alternative for rapid and high-throughput sampling and comprehensive soil characterization using spectroscopic fingerprints. Spectroscopic fingerprints reflect soils' chemical composition, condition, and ongoing processes. Most applications concern the use of reflectance spectra in visible (Vis), near (NIR), short (SWIR), or the medium infra-red (MIR) range, i.e., from about 350 nm to about 25,000 nm [1]. Spectra of soil samples are collected in the laboratory or in situ. However, increasing attention is paid to remote sampling using devices installed on a drone, airplane, balloon, or satellite.

This study focuses on modeling 19 different physio-chemical parameters describing the status of cultivated Haplic Luvisol soils readily from the near-infrared reflectance spectra. These parameters are organic carbon, total nitrogen, available soil nutrients (P, K, and Mg), exchangeable cations (Ca, Mg, and K), pH (H_2O), pH (KCl), hydrolytic acidity and eight elements (Cd, Cu, Pb, Ni, Cr, Zn, Mn, and Fe).

Predictions offered by partial least squares models indicate that near-infrared spectroscopy can effectively monitor many vital soil parameters and trace changes in concentrations of potentially toxic elements in regularly cultivated Haplic Luvisol soil [2].

References

[1] F. da Silva Terra, R. Rizzo, E. Ben-Dor, J.A.M. Dematte, Soil sensing by visible and infrared radiation, in: *Handbook of Near-Infrared Analysis*, 4th ed., CRC Press, 2021.

[2] S. Krzebietke, M. Daszykowski, H. Czarnik-Matusewicz, I. Stanimirova, L. Pieszczek, S. Sienkiewicz and J. Wierzbowska, *Talanta*, **251** (2023) 123749.

Graph analysis-based statistical process control

Martin Ferenczi[1], Ádám Ipkovich[1], Zsolt Tibor Kosztyán[2], János Abonyi[1]
[1] ELKH-PE Complex Systems Monitoring Research Group, University of Pannonia,
Egyetem u. 10, H-8200 Veszprém, Hungary
E-mail: janos@abonyilab.com
[2] Department of Quantitative Methods, University of Pannonia, Egyetem u. 10,
H-8200 Veszprém, Hungary

This work enriches the statistical process control (SPC) toolbox by incorporating visibility-graph-based network analysis. The role of SPC is to identify and classify faults when the process is not in control, using patterns represented by the Western Electric or Nelson rules[1]. The key idea of this work is that we convert the time series of SPC charts into visibility graphs to improve the ability to detect and visualise systematic patterns.

The visibility graph is a network interpretation of a time series, where each data point is treated as a node and edges are formed between the nodes if they are 'visible' or 'reachable' to each other based on a specific criterion[2]. This graph-based representation offers a fresh angle to interpret intricate patterns in time series data. Beyond visualisation, the network can preserve specific patterns revealed using network analysis methods and graph metrics. A decision tree is trained as an interpretable classifier using machine learning to map graph metrics from the visibility graphs and extract their importance.

The applicability of visibility graphs within SPC is demonstrated by analysing real-world industrial and synthetically generated data. In a broader context, beyond SPC, the proposed method can also be applied to other single-variable fault detection and time series analysis-based decision support tasks:

i) Statistical process control and usability in visibility graphs.

ii) Typical error patterns – *e.g.*, Western electric rules.

iii) Network analysis and Decision Tree for determining patterns.

References

[1] L. S. Nelson, "The Shewhart Control Chart—Tests for Special Causes", *Journal of quality technology*., **16** (1984) 238-239.

[2] L. Lacasa, B. Luque, F. Ballesteros, J. Luque & J.C. Nuno (2008). From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences*, **105**(13), 4972-4975.

Combining hyperspectral non-homogeneity measures and a one-class classification concept

Lukasz Pieszczek[1], Michal Daszykowski[1]
[1] Institute of Chemistry, University of Silesia in Katowice, Poland,
E-mail: lukasz.pieszczek@us.edu.pl

The near-infrared hyperspectral imaging (NIR-HSI) has enabled studying chemical distribution on the surface of a sample. Considering many wavelengths at a time, NIR-HSI provides detailed information about sample composition and reveals non-homogeneities on its surface. Moreover, information about chemical variability can be captured and modeled using machine learning algorithms and selected statistical descriptors (including hyperspectrograms [1]–a simplified representation of joint sample features represented by principal component analysis scores). Such measures extract meaningful features from the hyperspectral data and reveal the intrinsic properties of samples.

Combining sample chemical non-homogeneity measures with a one-class classifier [2,3] can enhance the detection of anomalous or non-homogeneous samples. The non-homogeneity measures describe the overall sample's chemical variability, while the one-class classifier models the expected behavior of groups of samples with an a priori known set of spectral distributions. Therefore, samples exhibiting substantial deviations from the expected behavior can be identified as anomalies.

In this study, we focus on the potential of hyperspectral imaging, one-class classifier, and hyperspectrogram representation to verify pharmaceutical tablets' quality and authenticity. The Specim FX17e camera, operating in the 900-1700 nm spectral range, was used to characterize samples and detect changes that may occur during pharmaceutical production of tablets. Different groups of three-component tablets (with cellulose, magnesium stearate, and ascorbic acid), with respect to selected production factors, were prepared and examined. The results demonstrate the effectiveness of the proposed data analysis approach in accurately identifying the different types of tablet production variations, surpassing an alternative solution relying on the analysis of individual spectra obtained from the hyperspectral image of a sample.

Acknowledgement

This research was supported by the National Science Centre, Poland (research grant no. 2018/29/N/ST4/01547).

References

[1] S. Kucheryavskiy, A new approach for discrimination of objects on hyperspectral images, *Chemometr. Intell Lab. Syst*., **120** (2013) 126–135.

[2] O.Ye. Rodionova, P. Oliveri, A.L. Pomerantsev, Rigorous and compliant approaches to one-class classification, *Chemometr. Intell Lab. Syst*., **159** (2016) 89–96.

[3] L. Pieszczek, H. Czarnik-Matusewicz, M. Daszykowski, Identification of ground meat species using near-infrared spectroscopy and class modeling techniques–aspects of optimization and validation using a one-class classification model, *Meat Sci.*, **139** (2018) 15–24.

Near-Infrared Spectroscopy coupled with multivariate models to predict secondary metabolite and *Botrytis* in cannabis and basil

Matan Birenboim[1], David Kengisbuch[1], Tarin Paz-Kagan[2], and Jakob A. Shimshoni[1*]
[1] Institute for Postharvest and Food Sciences, Agricultural Research Organization,
P.O. Box 15159, Rishon LeZion, 7505101 Israel.
[2]Ben-Gurion University of the Negev, Sede Boqer Campus, 8499000, Israel.
*E-mail: jakobs@volcani.agri.gov.il

Currently, the quantitative chemical analysis of secondary metabolites in edible and medicinal plants is achieved through the use of laborious, expensive, and time-consuming technologies, such as high-pressure liquid-chromatography and/or gas chromatography-mass spectroscopy. We aimed to develop a simple, accurate, fast, and cheap technique for the quantification of major cannabinoids and terpenes as well as latent *Botrytis* infestation in cannabis inflorescence and terpene, eugenol and polyphenolic concentration in basil leaves, using Fourier transform near infra-red spectroscopy (FT-NIRS) [1,2]. For authenticity purposes, we have developed accurate cultivar classification models for cannabis inflorescence and basil cultivars. For that purpose, FT-NIRS was coupled with state-of-the art multivariate classification and regression models, namely partial least square-discriminant analysis (PLS-DA) and partial least square regression (PLS-R) models. The PLS-DA model yielded an absolute major cannabis and basil class separation and perfect class prediction. The prediction of cannabinoid and terpene concentrations in medicinal cannabis inflorescence and rosmarinic acid, eugenol and terpenes in basil leaves by PLS-R, yielded robust models with high predictive capabilities R^2_{CV} and $R^2_{\mathrm{pred}} > 0.85$, RPD > 2.5, RMSECV/RMSEC ratio < 1.2). The prediction of latent *Botrytis* provided reasonable prediction of the ΔΔCt values, which were accurately measured using real-time PCR (R^2_{CV} and $R^2_{\mathrm{cv}} > 0.87$ and 0.70). Our results confirm that there is sufficient information in the FT-NIRS to develop excellent prediction models of secondary metabolites and *Botrytis* load in cannabis and basil and major-cultivar classification models.

References

[1] Birenboim M et al. Use of near-infrared spectroscopy for the classification of medicinal cannabis cultivars and the prediction of their cannabinoid and terpene contents. *Phytochemistry*. 2022:113445.

[2] Birenboim M et al. Optimization of sweet basil harvest time and cultivar characterization using near-infrared spectroscopy, liquid and gas chromatography, and chemometric statistical methods. *J Sci Food Agric*. 2021.

New applications of recurrences. detection of analytes' hydration during HPLC analysis

Igor G. Zenkevich*, Daria A. Nikitina, Abdennour Derouiche
St. Petersburg State University, Institute for Chemistry,
*Universitetskii prosp., 26; St. Petersburg 198504; *E-mail: izenkevich@yandex.ru*

The applications of recurrent relations (1, 2) in chemistry and chromatography seem to be highly diverse. For chromatographic retention times (t_R) and any quantities proportional to them (retention volumes, capacity factors, etc.) they are:

$$\mathbf{t_R(n + 1) = at_R(n) + b} \qquad (1)$$

$$\mathbf{t_R(x + \Delta x) = at_R(x) + b,\ \ \Delta x = const} \qquad (2)$$

The equation (1) is applicable to the physicochemical properties of homologs (the number of carbon atoms in a molecule is integer by definition). Another relation (2) allows extending the use of recurrences to continuous variables (temperature, pressure, or concentration), but it requires selection and fixing the increment Δx = const.

In reversed phase high performance liquid chromatography (RP HPLC) there are no simple single equation for describing the dependence of retention parameters *vs*. content of organic solvent in a water-containing eluents. Their set of relations includes the Scott-Kuchera approach ($1/t'_R = aC + b$), the Soczewinski-Wachtmeister equation ($\log t'_R = aC + b$), the Snyder-Soczewinski equation ($\log t'_R = a\log C + b$), Schoenmakers' polynomial approximation ($\log t'_R = aC^2 + bC + c$), and others. However, using the all of them can be replaced by the application of the single recurrent relation (2, $x = C$).

If under separation conditions (RP HPLC) the target analyte (X) does not demonstrate any specific interactions with an eluent (e.g., prototropic equilibria, or hydration), the plot of recurrent dependence (2, $x = C$, ΔC = 5% CH_3CN) is close to be linear (R > 0.999), as it is illustrated by **Fig. 1** for synthetic drug with trivial name "Gefitinib"). However, if the reversible formation of hydrates (3) takes place in the presence of water in an eluent, the similar plots demonstrate the deviations from linearity in area corresponding to just high waterc ontents:

$$\mathbf{X + H_2O \rightleftharpoons X{\times}H_2O} \qquad (3)$$

It is illustrated by **Fig. 2** for another synthetic drug with trivial name "Pazopanib":

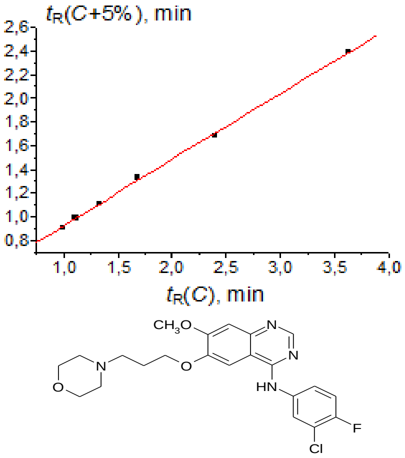


Fig. 1. Typical linear recurrent dependence (1) for analyte having no specific interactions with HPLC eluent.

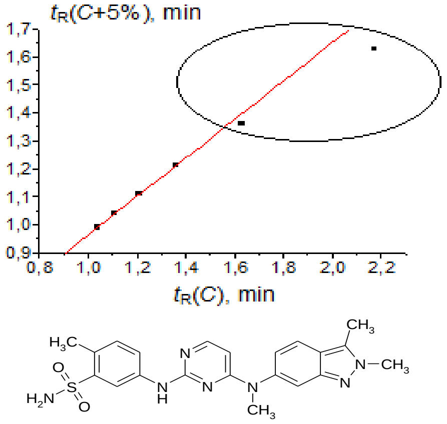


Fig. 2. Deviations of recurrent dependence (1) from linearity for analyte that forms a hydrate at high content of water in eluent.

It means the deviations of recurrences from linearity maybe not less informative than the absence of any anomalies.

Path of computational methods in assessing nano-ecotoxicity: retrospective, current status, and future perspectives

Kabiruddin Khan[1], Agnieszka Gajewicz-Skretna[1]
[1]Laboratory of Environmental Chemoinformatics, Faculty of Chemistry, University of Gdansk, Gdansk, Wita Stwosza 63, 80-308 Gdansk Poland
E_mails: Kabiruddin.khan@ug.edu.pl and agnieszka.gajewicz@ug.edu.pl

The production and widespread use of nanomaterials have raised concerns regarding their potential impact on the environment and human health. Assessing the environmental risks associated with nanotechnology requires a thorough understanding of nano-ecotoxicity, which investigates the toxic effects of nanomaterials on living organisms and ecosystems. In recent years, computational modeling has emerged as a powerful tool for predicting and comprehending the ecotoxicological effects of nanomaterials [1-2]. Here we provide a comprehensive survey of the advancements made in the predictive modeling of nano-ecotoxicity.

The initial focus of our study is to present a comprehensive overview of the fundamental concepts and principles that form the basis of nano-ecotoxicity modeling. This includes examining the physicochemical properties of nanomaterials, their interactions with biological systems, and the underlying mechanisms responsible for their toxic effects. Various computational approaches employed in modeling and predicting nano-ecotoxicity are then explored, encompassing quantitative structure-activity relationship (QSAR) models, molecular docking simulations, and molecular dynamics simulations. The present analysis critically discusses these computational methods' strengths, limitations, and applicability domains. Furthermore, it highlights recent progress in developing predictive models for nano-ecotoxicity [3]. These advancements include the integration of high-throughput screening data, the utilization of machine learning algorithms, and the incorporation of multi-scale modeling techniques. The potential of these models to predict the toxicity of diverse nanomaterials across different biological levels, spanning from molecular interactions to population dynamics, is thoroughly examined. Moreover, it addresses the challenges and future directions in the field of predictive modeling of nano-ecotoxicity. It discusses the necessity for standardized datasets, improved model validation strategies, and the consideration of uncertainties in modeling predictions. Emphasizing the significance of interdisciplinary collaborations among toxicologists, computational scientists, and experimentalists, the analysis underscores their pivotal role in advancing the field further.

In conclusion, our work offers valuable insights into the advancements and challenges in the predictive modeling of nano-ecotoxicity. This talk provides researchers with an overview of the current state-of-the-art in this field. It also guides future research endeavors to develop reliable computational models for assessing the environmental risks associated with nanomaterials.

Acknowledgement
This research is part of project No. 2022/45/P/NZ7/03391 co-funded by the National Science Centre and the European Union Framework Programme for Research and Innovation Horizon 2020 under Marie Sklodowska-Curie grant agreement no. 945339.

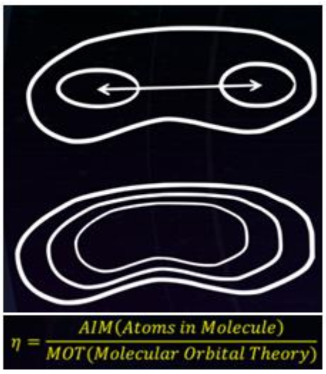
References

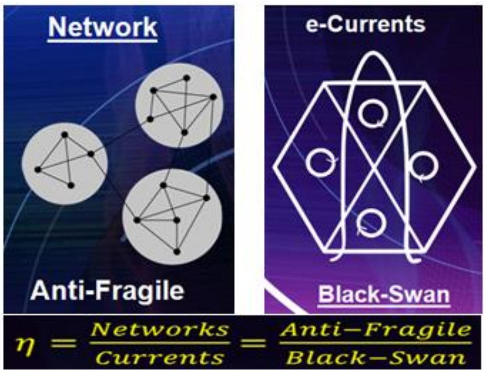
[1] Puzyn, Tomasz, *et al*. *Nature nanotechnology* **6.3** (2011): 175-178.
[2] Fourches, Denis, *et al*. *ACS nano* **4.10** (2010): 5703-5712.
[3] Chatterjee, Mainak, *et al*. *Environmental Science: Nano* **9.1** (2022): 189-203.

Balancing between Antifragility and Black Swans in QSAR

Mihai V. Putz[1,2]
[1] Laboratory of Computational and Structural Physical-Chemistry for Nanosciences and QSAR, Biology-Chemistry Department, Faculty of Chemistry, Biology, Geography, West University of Timișoara, Pestalozzi Str. No. 16A,RO-300115 Timișoara, Romania; E-mail: mihai.putz@e-uvt.ro;
[2] Scientific Laboratory of Renewable Energies-Photovoltaics, R&D National Institute for Electrochemistry and Condensed Matter (INCEMC-Timisoara), Dr. Aurel Podeanu Str. No. 144, RO-300569 Timișoara, Romania

In post-modern statistical theory and applications two major concepts that depart from the fashioned Gaussian normal distribution have appeared: a) *the Black Swan (BS) concept* that regards the asymmetry in distribution for the samples equally spanning small and large cardinals (so differently behaving than the large number theorem prescribes for the normal distribution in small error conditions); b) *the antifragility (AF) concept* that is merely related with non-linear correlation, customarily connected with the polynomial generalization of the multi-linear correlation analysis. Bothe of these concepts (BS & AF) highly impacted on the quantitative structure-activity [property] relationships (QSA[P]R) since both affecting the sample conditions and domain (i.e. non-normal distribution of entry structural chemical data) as well as the effect and the mechanism of action (i.e. the non-multi-linear correlations with the recorded biological activity or chemical properties). The present lecture and investigation is therefore original in addressing for the first time the balance between these two non-classical statistical concepts to chemical space, so advancing a new chemometry in QSARs studies. Essentially, two levels of analysis are performed: i) *the molecular space analysis* in itself, in which case the new BS vs. AF is advanced as the (Bader's basins of) Atoms-in-Molecules towards Molecular Orbital Theory indices respectively, e.g. chemical hardness or electronegativity; ii) *the chemical graph space analysis* by which the electronic currents superposition among the fragments vs. the clustering or meta-networking of them stays for such non-normal non-linear chemometric operator, $\hat{\eta} = AF / BS$. Further discussion of $\hat{\eta}$ impact on the QSAR studies is exposed in the context of maximum hardness principle [1] and of Thom catastrophes [2], respectively.


$\eta = \frac{AIM(Atoms\ in\ Molecule)}{MOT(Molecular\ Orbital\ Theory)}$


Network
Anti-Fragile
e-Currents
Black-Swan
$\eta = \frac{Networks}{Currents} = \frac{Anti-Fragile}{Black-Swan}$

References

[1] M.V. Putz *Int. J. Mol. Sci.* 11(4) (**2010**) 1269-1310; DOI: 10.3390/ijms11041269.

[2] M.V.Putz, M. Lazea, A.M. Putz, C. Seiman-Duda *Int. J. Mol. Sci.* 12(12) (**2011**) 9533-9569; DOI: 10.3390/ijms12129533;

Mutter SIMCA und ihre Kinder

Aleksey L. Pomerantsev, O. Ye. Rodionova
Federal Research Center for Chemical Physics RAS,
E-mail: forecast@chph.ras.ru

SIMCA is a well-known method that uses two characteristics: score and orthogonal distances (SD and OD) to develop a decision rule in the frame of class modelling, also known as one-class classification [1]. Now it is in great demand, mainly for solving authentication problems in different areas. In this lecture, we will not discuss SIMCA itself, but, on the contrary, we will turn to several new methods that directly follow from this concept, and which, therefore, can be called the children of SIMCA.

The eldest son of SIMCA is the concept of Cumulative Analytical Signal (CAS), which develops the idea of characteristic distances and introduces new characteristics as the sum of old ones [2]. CAS gives birth to a series of new methods (SIMCA grandchildren) such as outlier detection [3], representative subset selection [4], limit of detection estimation [2], multi-block data fusion [5], etc.

The beloved daughter of SIMCA is the precise and personalized medicine (PPM) that solves the problems of similarity/novelty, evaluation of curing results, selection of individual treatment, etc [6]. Details of the PPM approach are provided in our poster.

The youngest SIMCA son is Procrustes Cross-Validation (PCV) [7], which gives rise to a new concept of validation based on a pseudo-test set generated from the results of conventional cross-validation and the CAS rules. The main advantage is that the PCV set exists in reality as a set of numbers, and not as a virtual procedure that provides us with the final results of modelling only [8].

In the lecture, we discuss all these methods, focusing on ideas and concepts, and not on the mathematical apparatus, which can be found in the references.

References

[1] A.L. Pomerantsev, O.Ye. Rodionova, *J. Chemom.*, **34**, (2020) e3250
[2] A.L. Pomerantsev, O.Ye Rodionova, *Trends. Anal. Chem*, **143**, (2021). 116372.
[3] O.Ye. Rodionova, A.L. Pomerantsev, *Anal. Chem.*, 92, (2020) 2656−2664.
[4] A.L. Pomerantsev, O.Ye. Rodionova, *Microchem.J.*, **190**, (2023) 108654.
[5] O. Rodionova, A. Pomerantsev, *Anal. Chim. Acta*, **1265**, (2023) 341328.
[6] O. Rodionova, N. Kurysheva, et al, *Anal. Chim. Acta*, **1250**, (2023) 340958.
[7] S. Kucheryavskiy, S. Zhilin, et al, *Anal. Chem.* **92**, (2020) 11842-11850.
[8] S. Kucheryavskiy, O. Rodionova, et al, *Anal. Chim. Acta*, **1255**, (2023) 341096.

Needle in a haystack–NIR method development for quantifying novel foods and additives in rodent diet

Máté Csontos[1,2], János Elek[1]
[1] Science Port Kft. 4031 Debrecen, Köntösgát sor 1, Hungary,
E-mail: info@scienceport.hu
[2] University of Debrecen, 4025 Debrecen, Egyetem tér 1.

The registration and toxicological testing of the next generation of the food additives and dietary supplements – so called novel foods - bring new challenges from the analytical point of view as well. A large number of these materials belong to the group of UVCB materials: Unknown or Variable Composition, Complex Reaction product or Biological Origin [1]. These multi-component test items may contain a variety of proteins, fibres, fat compounds, and many other types of constituents. Not only the separation of these components can be strenuous with the classical analytical methods, but also the selection and targeted analysis of the main components. In this presentation I will demonstrate the development and validation of a near-infrared method combined with multivariate data analysis. The method was validated for the quantitation of a multicomponent biological novel food in a quite complex matrix: semi-synthetic rodent diet.

References

[1] Dimitrov SD, Georgieva DG, Pavlov TS, Karakolev YH, Karamertzanis PG, Rasenberg M, Mekenyan OG., *Environ. Toxicol. Chem.*, **34**(11) (2015) 2450-2462.

Validating QSAR models: an anti-MRSA case study

Jos Hageman[1], Carla Araya-Cloutier[2], Sylvia Kalli[3] & Jean-Paul Vincken[4]
[1] Biometris - Mathematical and Statistical Methods, Wageningen University & Research,
Wageningen, The Netherlands, E-mail: jos.hageman@wur.nl
[2] Laboratory of Food Chemistry, Wageningen University & Research,
Wageningen, The Netherlands
[3] Laboratory of Food Chemistry, Wageningen University & Research,
Wageningen, The Netherlands
[5] Laboratory of Food Chemistry, Wageningen University & Research,
Wageningen, The Netherlands

Quantitative Structure-Activity Relationship (QSAR) modelling plays a crucial role in predicting the activity of chemical compounds based on their molecular descriptors. This poster demonstrates a meticulous workflow: a variable selection methodology based on genetic algorithms (GA), coupled with careful model validation and applicability domain (AD) analysis. The methodology is illustrated by the QSAR modelling of antibacterial prenylated (iso)flavonoids with anti-MRSA activity.

By utilizing a deterministic division of data into training and validation sets via the Kennard-Stone algorithm, an overoptimistic division in training and test sets, commonly encountered in random splits, is mitigated. Molecular activity is modelled using multiple linear regression (MLR) with the most influential descriptors being selected by a GA. Model performance is evaluated using Leave-one-out cross-validation (LOOCV). A William's plot is used to define the applicability domain (AD) of the regression models. Outliers with high leverage or high residual values were carefully identified and excluded from the dataset to refine the AD. By assessing the AD, predictions are made only for compounds falling within the reliable scope of the developed models, enhancing the accuracy and validity of the QSAR predictions. After deciding on a final model, an external validation set was used for the final assessment of model accuracy.

This lecture highlights the significance of the GA-MLR approach in QSAR modelling, emphasizing the robustness approach and reliability of predictions through careful validation. The integration of GA allows for the selection of optimal predictors, enhancing the model's predictive power. The utilization of LOOCV ensures rigorous evaluation, while the applicability domain assessment guarantees the reliability of predictions beyond the training set. The developed GA-MLR models can help predict the activity of novel compounds with enhanced antibacterial activity.

The way it was: Personal, idiosyncratic reminiscences of past and present scientific achievements

Károly Héberger
Plasma Chemistry Research Group, Institute of Materials and Environmental Chemistry,
Research Centre for Natural Sciences, Centre of Excellence, Hungarian
Academy of Sciences, Budapest, Hungary
E-mail: heberger.karoly@ttk.hu

It is highly difficult to summarize more than 46 years' experiences in science. I started in Student's Scholarly Circles, finished my Diploma work (MSC) in Tallin, Estonia, and started to work at the Central Research Institute for Chemistry of the Hungarian Academy of Sciences. First supervisor Dezső Gál introduced me, how to work, and how to study radical oxidation processes in liquid phase. The first independent achievement was to observe a phenomenon that antioxidants may cause additional initiation [1].

Some non-linear regularities have been observed between gas chromatographic retention data of alkylbenzenes in Göttingen under the supervision of H. Gg Wagner [2]. The non-linear fits helped me to give the uncertainties of Arrhenius parameters (with full covariance matrix), reveal the effect of wrong weighting, and catapulted me from Hungary to the University of Zürich (Prof. Hans Fischer), where I made kinetic measurements (addition reactions of substituted benzyl, isopropylol and cyano isopropyl radicals [3]. I am especially proud of the latter, despite of repeated efforts nobody could measure the rate constant of addition reaction for this radical on various olefins before.

In the meanwhile, I have found that some definition of (multiple) correlation coefficient may produce higher correlation coefficients than one using certain definitions, if one of the vectors came from nonlinear relation. It shocked the scientific community, but the truth remained. There was another big throw: the pair correlation method and its generalization [4]. The last idea was the algorithm of sum or ranking differences (SRD) and it took ten years to be elaborated to the full (and it is under development till).

Acknowledgement
Herewith I express my sincere gratitude to all, who helped me during these years: Teresa Kowalska, Adam Voelker, Miklós Görgényi, John H. Kalivas, Caroline West, Klára Kollár-Hunek, Dávid Bajusz, Anita Rácz, Attila Gere, to all my coworkers and students.

References

[1] K. Héberger: On the Interaction of Hydroperoxide and Inhibitor molecules, *International Journal of Chemical Kinetetics*, **17** (1985) 271-275.

[2] K. Héberger: Discrimination between Linear and Non-Linear Models Describing Retention Data of Alkylbenzenes in Gas-Chromatography, *Chromatographia*, **29** (1990) 375-384.

[3] K. Héberger and H. Fischer: Rate Constants of the 2-Cyano-2-Propyl Radical to Alkenes in Solution. *International Journal of Chemical Kinetics*, **25** (1993) 249-263.

[4] Károly Héberger* and Róbert Rajkó, Generalization of Pair-Correlation Method (PCM) for Nonparametric Variable Selection, *Journal of Chemometrics*, 16 (2002) 436-443.

[5] Klára Kollár-Hunek and Károly Héberger*, Method and Model Comparison by Sum of Ranking differences in Cases of Repeated Observations (Ties) *Chemometrics and Intelligent Laboratory Systems*, **127** (2013) 139-146.

Posters

Higher-order data analysis to leverage the performance of food quality control procedures

M.R. Alcaraz[1,2], M. Antonio[1,2], F. Chiappini[1,2], J. Zaldarriaga Heredia[2,3], S.M. Azcarate[2,3], J.M. Camiña[2,3], A. Muñoz de la Peña[4], M.J. Culzoni[1,2], H.C. Goicoechea[1,2]
E-mail: malcaraz@fbcb.unl.edu.ar
[1] LADAQ-FBCB, Littoral, Santa Fe, Argentina
[2] CONICET, CABA, Argentina
[3] INCITAP, FCEyN, National University of La Pampa, La Pampa, Argentina
[4] Department of Analytical Chemistry, University of Extremadura, Badajoz, Spain

The demonstrated potential of using chemometrics in analytical chemistry has been escorted by a tireless pursuit of new advantages and benefits of multidimensional data analysis. Recent investigations in multi-way data analysis have revealed that second- and higher-order models can profitably exploit the second-order advantage, albeit additional benefits in third- and higher-order models are still under discussion. However, more theoretical and practical investigations should be conducted to clarify the basic theory of multi-way methods to explore the essence of higher-order methodologies [1,2].

Fluorescence excitation-emission matrix (EEM) spectroscopy coupled with multi-way analysis is a powerful tool for the analysis of fluorophore mixtures or complex systems because of its straightforwardness, selectivity, and high sensitivity. When combined with an extra instrumental or experimental mode, it can render appealing outcomes in terms of analytical performance. This work is devoted to demonstrating that third-order data analysis can be conveniently utilized in pursuing leveraging the analytical performance of the methods, in the field of food quality control.

First, to demonstrate the enhancement of the analytical performance of a method with quantitative aims, a four-way multivariate calibration method was developed for the simultaneous determination of 5 pesticides (thiabendazole, carbendazim, pirimiphos-methyl, imidacloprid and clothianidin) in citrus. Third-order data were acquired by registering the EEM photo-induced fluorescence at different times of UV irradiation in organized media. On the other hand, the effect of increasing the number of instrumental modes for classification analysis was also evaluated. Here, aiming at discriminating virgin olive oils from extra virgin olive oils, third-order data analysis was carried out on data generated by the EEM monitoring of the thermal degradation of olive oil.

In both cases, different data arrays (first- (in classification), second- and third-order data and data fusion) were built and subjected to several chemometric models to evaluate the properties and advantages of each data structure. Accordingly, PARAFAC, MCR-ALS and PLS-based modelling were used in each case. Quantitative results were evaluated through the predictive performance and detection capabilities, whereas classification results were evaluated through global indices, such as average sensitivity, non-error rate, and average precision. The results revealed different degrees of improvement by the inclusion of an additional mode to the data structure. The obtained results shed light on the fact that the use of higher-order data is an attractive approach to be explored in the classification field, particularly, in the study of samples with very similar spectral profiles, for which no evident classification patterns are observed.

References

[1] Alcaraz MR, Monago-Maraña O, Goicoechea HC, Muñoz de la Peña A. *Anal. Chim. Acta* **41** (2019); 1083.

[2] Azcarate SM, de Araújo Gomes A, Muñoz de la Peña A, Goicoechea HC, *Trends Anal. Chem*. **107** (2018) 151-168.

Pattern recognition analysis of ^{1}H-NMR fingerprint data for the geographical authentication of virgin olive oils

R. M. Alonso-Salces[1], G. E. Viacava[2], A. Tres[3], S. Vichi[3], E. Valli[4],
A. Bendini[4], T. Gallina Toschi[4], L. A. Berrueta[5]
[1] CONICET, CIAS-IIPROSAM, Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina.
E-mail: rosamaria.alonsosalces@gmail.com
[2] CONICET, GIIA, Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina
[3] Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Facultat de Farmàcia i Ciències de l'Alimentació, INSA-UB, Universitat de Barcelona, Santa Coloma de Gramenet, Spain
[4] Dipartimento di Scienze e Tecnologie Agro-alimentari, Alma Mater Studiorum - Università di Bologna, Cesena, Italy
[5] Departamento de Química Analítica, Facultad de Ciencia y Tecnología, Universidad del País Vasco/Euskal Herriko Unibertsitatea, Leioa, Spain

Food authenticity and food traceability are of great concern to the consumer, producer, food processor, retailer and regulatory bodies. An increasingly important issue of authenticity is the geographical origin, which is tackled in several EU regulations. Regarding EU Regulation No. 29/2012, the geographical origin of extra virgin and virgin olive oils must be declared on the label, referring to the EU, the EU member state or to the third country of origin, accordingly. In the case of blends of olive oils produced in more than one EU or non-EU country or in both EU and non-EU countries, the corresponding blend must be specified on the label. In accordance with EU Regulation No. 2081/92, quality agricultural food products, such as certain extra virgin and virgin olive oils, are permitted to be marketed using a Protected Designation of Origin (PDO) or a Protected Geographical Indication (PGI) label on the basis of their production area. Owing to the fact that geographical origin is among the main factors influencing consumer choice, and given the financial benefits associated with PDO/PGI labels, economic fraud is highly likely to occur, leading to a high rate of non-compliance related to false declaration of origin, labelling a non-PDO/PGI product as PDO/PGI, adulterating a PDO/PGI product with olive oils that do not fulfil the PDO/PGI requirements, among others. The lack of official analytical methods to verify the declared geographical origin and the authenticity and traceability of PDO/PGI olive oils triggers counterfeiting. Therefore, validated methods are needed to address these issues to protect both the consumer and the producer from illicit practices in this sector.

In the present study, a statistical significant number of extra virgin and virgin olive oils from EU and non EU countries (434 samples) collected during two harvest seasons were fingerprinted by ^{1}H-NMR. Multivariate data analysis of the ^{1}H-NMR fingerprint of these olive oils allowed to achieve classification models to verify the geographical origin of EU and non-EU olive oils at the national and PDO/PGI levels.

Acknowledgements: OLEUM "Advanced solutions for assuring the authenticity and quality of olive oil at a global scale" was funded by the EC within the H2020 Programme (2014–2020), GA no. 635690. The information expressed in this abstract reflects the authors' views; the EC is not liable for the information contained therein.

Classification of horsetails using machine learning methods on NIR spectra

K. Beier[1], T.-M. Dutschmann[1†], P. M. Puttich[2], M. Lubienski[2], T. Beuerle[2], K. Baumann[1]
[1]Institute of Medicinal and Pharmaceutical Chemistry, TU Braunschweig,
Beethovenstraße 55, 38106 Braunschweig, Germany, E-mail: katbeier@tu-bs.de
[2]Institute of Pharmaceutical Biology, TU Braunschweig, Mendelssohnstraße 1,
38106 Braunschweig, Germany
[†]Current address: Information Research, AbbVie Deutschland GmbH & Co. KG,
Knollstrasse, Ludwigshafen, 67061 Germany

Common horsetail (*Equisetum arvense L.*, syn.: field horsetail) holds a long tradition in the supportive treatment of numerous diseases [1,2]. A frequently observed problem is the risk of confusing *Equisetum arvense* plants with the closely related species *Equisetum palustre* (syn.: marsh horsetail) due to its morphological similarities. Distinguishing between the two species when harvesting is further complicated by the fact that both species share similar habitats [3]. This, however, is of particular importance because *E. palustre* contains the toxic alkaloids palustrine and palustridiene while this is not the case for *E. arvense* used for medicinal purposes (Equiseti herba) [4,5].

The aim of this study was the classification of horsetails using NIR spectroscopy in combination with machine learning techniques. Therefore, over 370 *E. arvense* and *E. palustre* samples originating from all over Germany, consisting of two years of harvest, were analysed using two different devices: a miniature (handheld) NIR device and a benchtop NIR device.

Initial dimension reduction and clustering techniques (PCA, t-SNE) provided insightful visualizations for the distribution of both species within the data space. After applying feature screening to the spectral data, a variety of supervised machine learning models based on different algorithms (SVM, RF, *k*NN) were trained to predict the species from an individual spectrum. In a cross-validation (CV) approach, it could be shown that the spectra from both spectrometers are sufficient to achieve high classification accuracies around 90 %. The validity of the complete workflow is further highlighted by assessing its reliability through posterior probabilities, which were high for the predicted class labels, implying a satisfying model certainty.

References

[1] A.-E. Al-Snafi, *IOSR Journal of Pharmacy*, **7** (2017) 31-42.

[2] T. Boeing, K. G. T. Moreno, A. J. Gasparotto, L. M. da Silva, P. de Souza, *Evid-based Complement Altern Med.* **2021**, ID 6658434.

[3] M. Nowak, I. Tipke, L. Bücker, K. Franke, M. Lubienski, T. Beuerle, *Planta Med,* **88** (2022) 447-454.

[4] L. Cramer, L. Ernst, M. Lubienski, U. Papke, H. M. Schiebel, G. Jerz, *Phytochemistry*, **116** (2015) 269-282.

[5] J. Müller, P. M. Puttich, T. Beuerle, *toxins*, **12** (2020) 710-724.

P03

Multivariate modelling of mid-infrared spectra of colorectal cancer

B. Borkovits[1], E. Kontsek[2], A. Pesti[2], S. Gergely[3], I. Csabai[1], A. Kiss [2], P. Pollner[4]
[1] Department of Physics of Complex System, Eötvös Lorand University, Budapest,
E-mail: borbende@phys-gs.elte.hu
[2] Department of Pathology, Forensic and Insurance Medicine, Semmelweis University Budapest, kontsekendre@gmail.com
[3] Department of Applied Biotechnology and Food Science, Budapest University of Technology and Economics, Budapest
[4] MTA-ELTE Statistical and Biological Physics Research Group of the Hungarian Academy of Sciences, Budapest

Keywords: infrared, FTIR, colorectal, spectroscopy

Introduction: The applicability of techniques based on spectroscopy outside the visible light range is increasingly being investigated. The mid-infrared technique is non-invasive and non-destructive, which is one of its main advantages over the use of ionizing radiation. In addition, it can provide sufficient chemical information to effectively predict whether a patient has cancer using appropriate machine learning methods.

Materials and Methods: Sections of tissue microarrays of formalin-fixed tissue samples embedded in paraffin-embedded tissue were selected and analysed. Spectra were acquired using a Fourier transformation mid-infrared Perkin Elmer Spotlight microscope. Conventional H&E stained sections were digitized using a 3DHistech P1000 scanner. The 32 cores are containing intact colon mucosas (NC) and primary colorectal carcinomas (CRC). A digital database was created to organize and assemble data from different modalities. Both unsupervised (PCA) and supervised methods (Random Forest, Linear Discriminant Analysis, Support Vector Machine, XGBoost, U-Net) were used to process the data.

Results: 7744 spectra were collected from each core. Point clouds visualized for the first two principal components of PCA show some mixing of tumor versus normal spectra. The Random Forest algorithm resulted in a model accuracy of 0.575, Linear Discriminant Analysis 0.644, Support Vector Machine 0.593, XGBoost 0.592 and the U-Net 0.532. Linear Discriminant Analysis produced the highest sensitivity of 0.885.

Conclusions: The accuracies were poor in the discrimination of normal colon mucosa from colorectal cancer on the non-filtered spectra. Therefore, data preprocessing is suggested to be performed on larger cohort after spectral background filtration.

The effect of sample grinding in NIR spectroscopy

M. Csontos[1,2], J. Elek[1], Z. Vincze[2]
[1] Science Port Kft. 4031 Debrecen, Köntösgát sor 1, Hungary
E-mail: info@scienceport.hu
[2] University of Debrecen, 4025 Debrecen, Egyetem tér 1.

The spectroscopic analysis of solid samples often requires physical sample preparation steps: one of these is grinding. Several demands shall be met for the successful outcome of the process, the particle size of the ground sample should be small enough and homogenously distributed, the grinding should be reproducible even between laboratories. These properties are crucial in order to minimize the undesired physical thus spectral differences between samples caused by the light scattering. Moreover the particle size also affects the homogeneity obtained in the various sample pre-treatment steps; for instance mixing and drying.

A common type of these instruments is the blade grinder, where the only controlled parameter is the process length: the longer grinding - in theory - the smaller the particle size. In this study the effect of the grinding time was investigated of two commercial blade grinders in order to develop a standardized grinding method for pelleted diets. The NIR spectra of the same type but differently ground diet samples were collected and analysed using PCA in order to investigate how the grinding time affects the particle size and homogeneity.

Reference

[1] M. Otsuka, *Powder Technology*, **141(3)** (2004) 244-250.

Laser-Induced Breakdown Spectroscopy (LIBS) data analysis

Pegah Dehbozorgi [1,2], Ludovic. Duponchel [3], Vincent. Motto-Ros[4], Thomas. Bocklitz[1,2,5]
[1] Leibniz Institute of Photonics Technology, Member of Leibniz Health Technologies, Member of the Leibniz Centre for Photonics in Infection Research (LPI), Albert-Einstein-Strasse 9, 07745 Jena, Germany. E-mail: pegah.dehbozorgo@uni-jena.de
[2] Institute of Physical Chemistry (IPC) and Abbe Centre of Photonics (ACP), Friedrich Schiller University Jena, Member of the Leibniz Centre for Photonics (LPI), Helmholtzweg4, 07743 Jena, Germany.
[3] Univ. Lille, CNRS, UMR 8516 – LASIRE – Laboratoire de Spectroscopie pour Les Interactions, La R´eactivit´e et L'Environnement, Lille, F-59000, France.
[4] Institut Lumière Matière, UMR5306 Université Lyon 1-CNRS, Université de Lyon 69622 Villeurbanne cedex, France.
[5] Institute of Computer Science, Faculty of Mathematics, Physics & Computer Science, University of Bayreuth Universitaetsstraße 30, 95447 Bayreuth, Germany.

The Laser-Induced Breakdown Spectroscopy (LIBS) technique is widely used to measure the concentration of elements in different types of samples [1,2]. This study was established to investigate and compare the performance of two approaches, classical regression using Partial Least Square (PLS) and Deep Learning (DL), in predicting the concentration of 24 elements from LIBS spectra. The main challenge for developing predictive models was the variation of electron density and temperature of the plasma, which can completely modify the spectra. Therefore, besides PLS, we tried implementing more advanced tools such as Convolutional Neural Networks (CNNs). The study used the training set of 20000 simulated LIBS spectra and 5000 simulated LIBS spectra as the test set. To develop the models, a pre-processing step was conducted to normalize the data to the (0,1) range. However, the models were also trained and tested with the original data (without normalization) to make the study more comprehensive. For DL, a simple CNN with six convolutional layers was designed. The performance of the models was evaluated based on their stability and accuracy in predicting the concentration of the 24 elements within the test set. Our findings suggest that DL outperformed classical regression in predicting the concentration of presented elements within the simulated test LIBS spectra. The DL model showed greater stability and higher accuracy in predicting concentrations of elements. Overall, this study provides important insight into the application of DL in LIBS analysis as a powerful and stable tool for accurate and reliable elemental analysis.

Acknowledgement

This work is supported by the Federal Ministry of Education and Research of Germany (BMBF), funding program Photonics Research Germany (FKZ: 13N15706, 13N15466, 13N15710) and is integrated into the Leibniz Center for Photonics in Infection Research (LPI).

References

[1] O. Nicolini, LIBS multivariate analysis with machine learning, 2020.
[2] C. Pasquini, J. Cortez, L. Silva, F.B. Gonzaga, Laser induced breakdown spectroscopy, *Journal of the Brazilian Chemical Society* **18** (2007) 463-512.

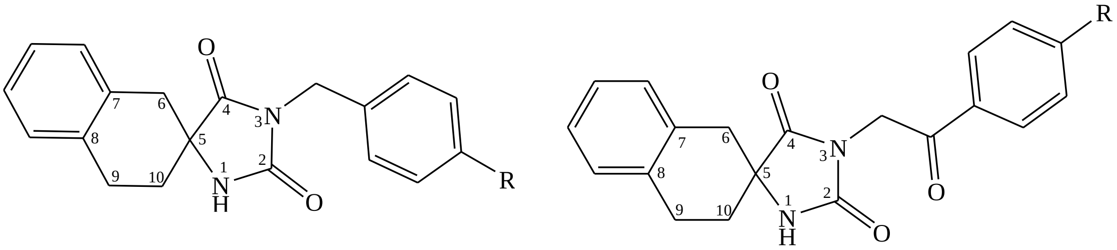
P06

QSRR study of β-tetralino-spiro-5-hydantoin derivatives

Tatjana Lj. Djaković Sekulić[1], Anamarija Mandić[2], Anita Lazić[3]
[1] University of Novi Sad, Faculty of Sciences, Department of Chemistry, Biochemistry and Environmental Protection, Novi Sad, Republic of Serbia,
E-mail tatjana.djakovic-sekulic@dh.uns.ac.rs
[2] University of Novi Sad, Institute of Food Technology, Novi Sad, Republic of Serbia
[3] University of Belgrade, Faculty of Technology and Metallurgy, Belgrade, Republic of Serbia

In recent years spiro compounds have attracted significant interest due to their unique conformational features and their structural implications on biological systems. Since spiro compounds contain two rings with only one shared atom have a good balance between conformational restriction and flexibility, makes them adaptable to many biological targets. Therefore, knowledge of their interactions in aqueous systems is of crucial importance for the understanding of biological response and hence the initial step for the selection of the compound for the drug candidate.

In this poster presentation the retention data for two series of spiro compounds derivatives of β-tetralino-spiro-5-hydantoins with a 4-substituted benzyl group or a 2-(4-substituted phenyl)-2-oxoethyl group in the position 3 (structures presented in Figure 1) were investigated.



Series 1 Series 2

Figure 1. General formula of investigated β-tetralono-spiro-5-hydantoins. Substituent R in Series 1: H (**1a**), CH_3 (**1b**), OCH_3 (**1c**), Cl (**1d**), Br (**1e**), CN (**1f**), NO_2 (**1f**); substituent R in Series 2: OCH_3 (**2a**), F (**2b**), Cl (**2c**), Br (**2d**), CN (**2e**), and NO_2 (**2f).**

The attention of this work will be focused on the evaluation of chromatographic lipophilicity parameters determined by means of reversed-phase thin layer chromatography (RP-TLC) and reversed phase high-performance liquid chromatography (RP-HPLC). Principal Component Analysis (PCA) will be used in order to find similarities and differences among the investigated compounds, their retention data and structural parameters relevant for the activity.

Acknowledgment: The authors gratefully acknowledge the financial support of the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Grant No. 451-03-47/2023-01/200125)

Toward more efficient and effective color quality control in the large-scale printing process

P. Dziki[1,2], L. Pieszczek[2], K. Rybicka[1], M. Daszykowski[2]
[1] Walstead Kraków Sp. z o.o., 11 Obroncow Modlina Street, 30-733 Krakow, Poland;
E-mail: pawel.d.dziki@walstead-ce.com
[2] Institute of Chemistry, University of Silesia in Katowice,
9 Szkolna Street, Katowice Poland

In the printing industry, basic quality control focuses, among others, on monitoring color appearance and its potential variability. For this purpose, control templates with multiple color fields are screened manually, field by field, using a simple handheld spectrophotometer. Instead of considering collected spectral profiles, they are replaced by the corresponding L*, a*, and b* values (the CIELAB system) and then compared with expected values to determine the color difference (ΔE) [1].

Despite measurement simplicity, some drawbacks of a compact handheld spectrophotometer are revealed when there is a need to carry out quick and multiple measurements over the entire process or at its selected time points. Multiple and precise measurements require much attention. With increasing operator fatigue, it is likely to measure incorrect color fields. Moreover, compact handheld spectrophotometers have difficulty dissipating heat, translating into a systematic increase in measurement error over time.

Hyperspectral cameras offer an attractive alternative to handheld spectrophotometers and open new possibilities for rapid and advanced quality control, including color. With little effort, the so-called 'push-broom' hyperspectral camera can be installed over the production line and monitor the progress of a given process. It can characterize a sample's surface by a dozen hundred spectra. In the context of printing and quality control, their spatial and spectral resolution can provide very detailed information concerning the stability of the printing process and reveal its fluctuations over time.

On the other hand, the use of advanced hyperspectral imaging systems requires efficient data processing and modeling. Chemometrics can greatly support color quality control, its density, and offers an excellent framework for designing efficient expert systems based on exploring information from hyperspectral images. The main challenge concerns efficient characterization and comparing large spectra distributions representing different production stages.

In this study, we illustrate how chemometrics and a hyperspectral imaging system, with an FX10 camera (Specim, Oulu, Finland) collecting spectra in the visible range, can uncover potential changes in the large-scale printing process.

References

[1] S. Westland, C. Ripamonti, V. Cheung, Computational colour science using MATLAB, the 2nd edition, Wiley, Hoboken, NJ, 2012.

Investigation of carbohydrate powder mixtures by near-infrared spectroscopy and multivariate data analysis

J. Slezsák, Z. Gál, A. Salgó, S. Gergely
Dept. of Applied Biotechnology and Food Science, Budapest University of Technology and Economics, Budapest, Hungary,
E-mail: gergely.szilveszter@edu.bme.hu

In the pharmaceutical and food industry, there is often a need for fast, high-throughput testing of not only pure components, but also various powder mixtures. In these cases, not only the chemical composition, but also physical characteristics such as particle size affect the product quality. Among the vibration spectroscopy methods, the non-destructive near-infrared (NIR) technique is able to provide information about both the chemical and physical variability of powder mixtures through the spectra obtained as a result of complex light-matter interactions.

The applicability of the NIR technique was investigated by measuring different carbohydrate model systems with different physical and chemical characteristics. Through the spectra and their outputs obtained by multivariate data analysis, three parameters were examined: a) the effect of sample preparation (grinding), b) the effect of spectrophotometers with different optical and measurement systems, [1] c) the effect of mathematical pretreatments [2].

The primary goal was to find a combination of the used spectrometers and evaluation methods, which would later enable the accurate classification of powder mixtures similar to the tested systems based on either qualitative or quantitative properties. The elaborated model system can help in deciding whether the delivered pure components to be processed can be blended directly on the basis of their particle size, or whether they require other technological operations involving time and cost before blending.

References

[1] A. Kazeminy, S. Hashemi, R. L. Williams, G. E. Ritchie, R. Rubinovitz and S. Sen, *J. Near Infrared Spectrosc.*, **17** (2009) 233-244.

[2] Å. Rinnan, F. v. d. Berg and S. B. Engelsen, *Trends Anal. Chem.*, **28** (2009) 1201-1222.

PLS based multiway one class classification

Adriano A Gomes[1,2], Ivan Špánik[2]
[1] Federal University of Rio Grande do Sul - UFRGS;
E-mail: ivan.spanik@stuba.sk
[2]Slovak University of Technology -STU

Authentication models are often recommended to be used when inspecting frauds in commodities like food, drugs, fuel, among others. Considering that the fraud process is very complex and dynamic, measuring chemical information as much as possible may help achieve more reliable models. In this context, multiway data can be a useful approach rather than using just traditional first-order data. In this work, we propose to perform one-class classification based on unfolded and multidimensional partial least square [1-2]. The performance of those new algorithms, so-called OC-U PLS and OC-N PLS, was evaluated in a simulated case study. The simulated data were generated starting from Gaussian vectors (1 × 100), which were multiplied to generate unitary X matrices (100 × 100). In total, three individual matrices were generated: $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, which were combined by a weighted sum to generate the signal of each simulated sample (**S**). Gaussian noise has been added to the data. In both cases, the models were rigorously adjusted, using only the target samples. Subsequently, they were evaluated on a test set containing samples belonging to the target class and samples not belonging to it. The appropriate number of latent variables (LVs) was selected based on minimizing the standard deviation of prediction errors via LOOCV. The results found are summarized in **Table 1**.

Table 1: Statistical summary of model fit and testing.

Model	**LVs**	***a***	SEN_{train}	SEN_{test}	SPC_{test}	Eff_{test}
OC U-PLS	3		1.00	1.00	1.00	1.00
OC N-PLS	4	0.01	1.00	1.00	1.00	1.00
OC U-PLS	3		0.95	0.96	1.00	0.98
OC N-PLS	4	0.05	0.97	0.98	1.00	0.99
OC U-PLS	3		0.9.0	0.94	1.00	0.97
OC N-PLS	4	0.1	0.91	0.94	1.00	0.97

For all cases, the sensitivity in the training (*SENtrain*) stage is in agreement with the adopted alpha, suggesting the absence of overfitting. In addition, high sensitivity (*SENtest*), specificity (*SPCtest*), and efficiency (*Efftest*) were found in the test step. These findings indicate that the use of one-class classifiers applied to multiway data can be a useful tool in authenticity studies.

Acknowledgement
This work was supported by the Slovak Research and Development Agency under contract number APVV-21-0211.

References

[1] A. A. Gomes, S. M. Azcarate, I. Špánik, L. Khvalbota, H. C. Goicoechea. Pattern recognition techniques in food quality and authenticity: A guide on how to process multivariate data in food analysis. *Trends Anal. Chem.* 164 (2023) 117105.

[2] S. M. Azcarate, A. A. Gomes, A. M. Peña, Héctor C. Goicoechea. Modeling second-order data for classification issues: Data characteristics, algorithms, processing procedures and applications. *Trends Anal. Chem.* 107 (2018) 151-168.

Sum of ranking differences (SRD) when differences diminish and reference ranking is ambiguous: the theoretical foundations of weighting schemes

D. Kovács[1], Z. Fazekas[1]
[1] Hevesy György PhD School of Chemistry, ELTE, Eötvös Loránd University, Budapest, Hungary, E-mail: kovacs.daniel@ttk.elte.hu

The sum of ranking differences (SRD) method is a versatile vector comparison tool with diverse scientific applications [1,2]. The theoretical basis of SRD as a nonparametric statistical test has been discussed in detail upon its introduction [3-5]. SRD analysis is based on comparing the ranking vectors of objects by different measurement techniques/ experts/ predictors to the rankings yielded by a reference vector. However, this means that SRD in the presently known form is only applicable if the ranking of the objects by the reference is unambiguous. If the reference values are subject to experimental or intrinsic uncertainty (as most values usually are), then consequently, the reference ranking might become ambiguous making the application of SRD unfeasible. We explore the possibility of solving this problem by introducing weights in SRD, *i.e.* augmenting Eq. 1 with the $w_{i,j}$ weights in equation Eq. 2 and proposing mathematical definitions for these weights.

$$SRD_j = \sum_{i=1}^{n} \left| r_{i,j} - r_{i,ref} \right| \tag{1}$$

$$SRD_{j,weighted} = \sum_{i=1}^{n} w_{i,j} \times \left| r_{i,j} - r_{i,ref} \right| \tag{2}$$

In Eq. 1 and 2, *i* and *j* are indices for the rows (objects) and columns (techniques/ experts/ predictors) of the data matrix, respectively. We denote rank values by r_{ij}.
The authors are aware of only a single publication in which column-wise weighted SRD was applied [6]. The present work recommends a more sophisticated row-wise approach with a detailed discussion of the theoretical properties of weighted SRD values.

We explain, how experimental or intrinsic uncertainty of reference values may lead to an ambiguous reference ranking of the studied objects and how weighting could resolve this ambiguity. We also show, how weighting facilitates reintroducing some of the information into SRD that is lost during rank transformation of the reference vector.

Furthermore, we describe the heuristics behind a proposed weighting scheme and the most important mathematical properties of the defined weights. The particular feasibility of weighting in the case of using row average as reference in the absence of a real golden standard is highlighted.

Computer simulation results demonstrating the effects of weighting are also provided.

References

[1] D. Bajusz, A. Rácz, K. Héberger, *Molecules*, **24** (2019), 2690
[2] D. Kovács, A. Bodor, *RSC Adv*., **13** (2023) 10182-10203
[3] K. Héberger, *Trends Analyt. Chem*., **29** (2010) 101-109
[4] K. Héberger, K. Kollár-Hunek, *J. Chemom*., **25** (2011), 151-158
[5] K. Kollár-Hunek, K. Héberger, *Chemometr. Intell. Lab. Syst*., **127** (2013) 139-146.
[6] A. Gere, D. Szakál, K. Héberger, *Appl. Sci*., **12** (2022) 6303

Selection of preferable and undesirable distance measures for stochastic optimization by cross entropy

Sándor Kovács[1], Károly Héberger[2],*
[1.]University of Debrecen, Faculty of Economics and Business, Institute of Statistics and Research Methodology, Department of Economical and Financial Mathematics, Debrecen, Hungary E-mail: kovacs.sandor@econ.unideb.hu
[2] Research Centre for Natural Sciences, Institute of Excellence, Hungarian Academy of Sciences, H-1117 Budapest, Magyar Tudósok krt. 2, Hungary
E-mail: heberger.karoly@ttk.hu

Rank aggregation is one of the most difficult and unresolvable tasks for large data sets (*np*-hard). Stochastic optimization criterion for aggregating top-*k* lists conforms to the generalized Kemény guidelines. Shortly: it is the weighted sum of distances between the aggregate rankings and the input lists [1].

A fair way to compare distance measures was developed, based on sum of ranking differences (SRD) [2,3] and compared with classical rank aggregations: mean, median, geomean, L_2-norm, Markov Chain (three versions).

The orderings were compared to the benchmark ranking (SRD): row average of all ranks by stochastic optimization methods based on cross entropy.

Three data sets:

1) well defined (short) ordering (Insect data)
2) Clustering into two distinct rankings (eye data)
3) Zigzag pattern, contradictory rankings (Todeschini data)

The following novel distance measures using cross entropy optimization were compared: Kendall (CET), Ulam (CEU), Spearman (CES), Cosine (CECS), Cayley (CEC), Soergel (CESO), Euclidean (CEE), Hamming (CEH), Correl (CEPR), Dice (CED), Cayley&SRD [4] (CECSRD).

A quasi-continuous ordering of rank aggregation methods can be observed.

Distance measures are worth to be selected according to the data structure (inherent characteristic of data sets).

From among the classical rank aggregation methods the median should be preferred. Kendall- (CET) Spearman- (CES) Cayley- (CEC), *etc*. distances should be selected differently (optimally) from data set to data sets.

Geometric mean (geomean) and especially Ulam (CEU), analogous to Levenshtein, distance cannot be recommended: they cannot pass the randomization tests.

References

[1] S. Lin, Rank aggregation methods, *WIREs Computational Statistics*, **2** (2010) 555-570.

[2] K. Héberger, Sum of ranking differences compares methods or models fairly. *TRAC - Trends Anal. Chem.*, **29** (2010) 101-109.

[3] K. Kollár-Hunek and K. Héberger, Method and Model Comparison by Sum of Ranking differences in Cases of Repeated Observations (Ties). *Chemometr. Intell. Lab. Syst.*, **127** (2013) 139-146.

[4] L. Sipos, A. Gere, A., J. Popp, S. Kovács, A novel ranking distance measure combining Cayley and Spearman footrule metrics, *J. Chemometr*. **32(4)**, e3011 (2018)

The difference of model robustness assessment using cross-validation and bootstrap methods

Rita Lasfar and Gergely Tóth
Institute of Chemistry, Eötvös Loránd University, 1117 Budapest, Pázmány s. 1/a, Hungary, E-mail: rita.lasfar@gmail.com

The OECD Guidance [1] describes three necessary aims of model validation: the assessment of goodness of fit, robustness and predictivity. The first and the second ones are suggested to be performed as internal validation and predictivity should be checked on external data set. In the case of robustness, cross-validation methods are the most frequently used ones. The choice of the leave-one-out and/or the leave-many-out variants might depend on the selected modelling method [2]. The other method can be bootstrap to assess robustness.

In our poster we would like to show the similarities and the differences of the cross-validation and bootstrap methods. There are a rather large number of bootstrap variants, where the so-called out-of-badge [3] is conforming to the robustness defined in the OECD guidance. We investigate different features by showing their trends with respect to the sample size. As a preliminary result, we concluded that their information content relates similarly comparing to the goodness-of-fit and predictivity metrics. There is also a significant rank correlation between the two ways. The bootstrap method provides slightly less fluctuating individual results due to the repetition scheme applied there. On contrary, the known underestimation feature of bootstrap [3] might lead to difficulties during the interpretation of the actual values of this validation metric.

References

[1] "OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models" 2004, "Guidance document on the validation of (Quantitative) Structure-Activity relationships[(Q)SAR models" 2007 Organisation for Economic Cooperation and Development (OECD), https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm.

[2] P. Király, R. Kiss, D. Kovács, A. Ballaj, G. Tóth. *Mol. Inf.* **41** (2022) 2200072

[3] T. Hastie, R. Tibshirani, J Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer (2009) NewYork

Machine vision system and multivariate data analysis in the quality assessment of tablets

Lilla Alexandra Mészáros[1], Attila Farkas[1], Zsombor Kristóf Nagy[1]

[1] Department of Organic Chemistry and Technology, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics,
E-mail: meszaros.lilla.alexandra@vbk.bme.hu

The pharmaceutical industry is focused on the paradigm shift from traditional batch to continuous manufacturing. Thus, modern, innovative manufacturing lines and analytical tools need to be developed and introduced in the future. The published Process Analytical Technology guideline and the Quality-by-Design approach also support the shift toward continuous manufacturing. Machine vision coupled with multivariate data analysis can lead to innovative solutions in the mentioned field.

The presented work outlines a non-destructive, rapid, digital UV/VIS, imaging-based tool for predicting several critical quality attributes of various tablets. The quantitative prediction of the compression force, crushing strength and the content of active pharmaceutical ingredients (API) was executed with the partial least squares method and feed-forward artificial neural networks. The input dataset for the compression force and crushing strength prediction was obtained by processing VIS images with discrete wavelet transformation. The applied data analysis techniques accurately predicted the target level with less than 10% relative error. The input dataset used for the prediction of API content was based on the processing of the colour components of VIS and UV images. With these quantitative predictions, the relative error was obtained around 5% for the target level.
For the qualitative prediction, classification tasks were executed using pattern recognition neural networks. These were used to classify various tablets in connection with the particle size distribution and the content of the active pharmaceutical ingredients. Based on the obtained goodness parameters, the developed system would be capable of the qualitative examination and the comparison of tablets. Finally, the predicted critical quality attributes were applied for *in vitro* dissolution profile predictions using feed-forward artificial neural network.

The presented non-destructive, rapid machine vision system with multivariate data analysis can provide a new, simpler, cheaper alternative to the current measurement technology tools and techniques. It can support the quality assessment at continuous manufacturing lines individually or complement spectroscopic methods.

Non-membership probability for assigning a non-member (outlaying) sample in different variants of random forest classification

N. Mobaraki[1], K. Baumann[1]

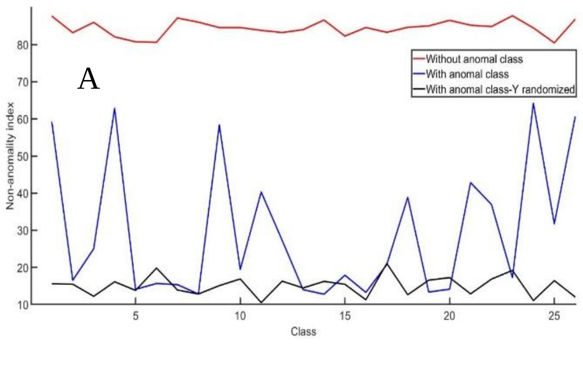
[1] First Institute for Medicinal and Pharmaceutical Chemistry, University of Technology Braunschweig, Beethovenstraße 55, 38106 Braunschweig, Germany;

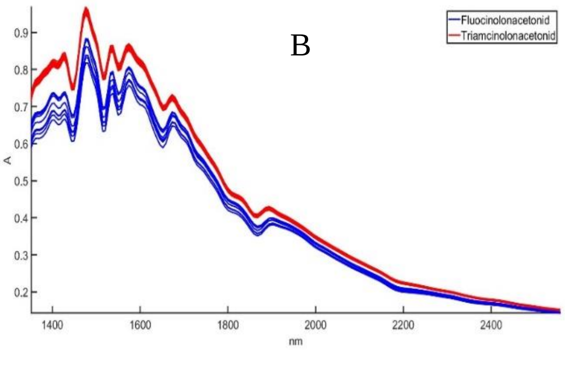
E-mail: k.baumann@tu-braunschweig.de

In most classification methods, the goal is assigning unknown samples to one of the predefined classes. In these methods, it is assumed that the unknown sample belongs to one of the predefined classes in the training phase. However, if the unknown sample does not belong to any predefined classes, most methods fail to distinguish it and hence classify it in one of the predefined classes. Random forest is a popular method for classification as well as regression tasks [1]. Until now, a lot of efforts have been made to improve the classification results of RF [2]. In all mentioned studies, the focus has been on improving the classification accuracy by using some innovative hyphenated methods. But one important point is missing here. Besides the classification accuracy, it is very important to have an estimation about non-membership probability to identify those samples that are not the member of any classes considered in the training phase. This will be referred to as "non-membership probability" here. This parameter is very crucial in cases of classifying non membership samples in practical problems. Therefore, the main aim of this research is defining class non-membership probability for predicted samples in the random forest classification.

Spectra of 26 different steroids were collected. These spectra were used to find the appropriate threshold for classification of non-member samples. Four hand-held vibrational spectroscopy instruments were used (NIR instruments: MicroNIR, SCIO, Neo Spectra, and Raman instrument: Metrohm MIRA-M1 Advanced).

After applying the RF method, ECVA method was used on each of the RF nodes. Based on the number of misclassifications obtained from this method, a factor was obtained to find non-membership classes (Figure A). As can be seen, our method is able to detect unknown members very well. Figure B shows the spectrum of two classes. As can be seen, the spectra of these two classes are very similar, and this shows that the proposed method has a high efficiency in detecting anomalies.





References

[1] L. Breiman, *Mach. Learn.*, **45** (2001) 5-32.

[2] P.S. Akash, M.E. Kadir, A.A. Ali. M.N. Ahad Tawhid and M. Shoyaib, in: 2019 Int. Conf. Robot. Signal Process. Tech., IEEE, (2019) 611-616.

Impact of different model transfer algorithms on dilution series and oil samples

M. Mócz[1,2], P. P. Hanzelik[1], J. Slezsák[2], S. Gergely[2]
[1] Group Enterprise Data MOL Plc., Budapest, Hungary; pphanzelik@mol.hu
[2] Dept. of Applied Biotechnology and Food Science, Budapest University of Technology and Economics, Budapest, Hungary; E-Mail: mmocz@edu.bme.hu

The model transfer allows for using existing well-performing models on new instruments, saving time and resources required for model development. It is essential for maintaining consistency and comparability of analytical measurements across different devices or laboratories. Model transfer ensures the model performs accurately and reliably when applied to new instruments or laboratories. Various techniques are used for model transfer [1], including complete recalibration, prediction correction, and standardization of spectral responses. These methods involve collecting reference data from source and target instruments [2] and using mathematical algorithms to align the calibration models.

The concept of model transfer was examined by investigating a dilution series with three compounds. The dilution series allowed us to explore how well the model transfer performed across different concentration levels in a draft model system. Samples were measured at four laboratories, and predictive result errors consistently decreased following transfer of the model completion. After understanding behaviour of the models with the dilution series, we applied it to 'real' samples, which are oil samples from the industry. These samples consisted of essentially different components than the dilution series, presenting a practical model transfer evaluation scenario.

In conclusion, the success of model transfer relies on the availability of representative samples and a thorough understanding of the sources of variation. Proper validation and verification procedures are essential to ensure the accuracy and reliability of the transferred calibration models [3].

References

[1] L. Li, W. Huang, Z. Wang, S. Li, X. He and S. Fan, *Postharvest Biol. Technol.,* **183** (2022) 111720.

[2] M. F. Abdelkader, J. B. Cooper and C. M. Larkin, *Chemom. Intell. Lab. Syst.,* **110** (2012) 64-73.

[3] H. Leion, S. Folestad, M. Josefson and A. Sparén, *J. Pharm. Biomed. Anal.* **37** (2005) 47–55.

Artificial neural networks in pharmaceutical process development and quality assurance

B. Nagy[1], A. Farkas[1], D. Galata[1], Zs. K. Nagy[1]
[1]Department of Organic Chemistry and Technology, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111, Budapest, Hungary, E-mail: nagy.brigitta@vbk.bme.hu

Artificial intelligence (AI) is getting tremendous attention nowadays, both in everyday life and in the manufacturing industries. The fourth industrial revolution (Industry 4.0 or Pharma 4.0 in the pharmaceutical industry) embraces the opportunities brought by AI to facilitate big data processing, digitalization, and automation while striving for interconnected manufacturing systems and smart factories capable of autonomous decisions [1]. However, the stringent quality control expectations of the pharmaceutical industry hinder their introduction. Consequently, very few available studies evaluate how artificial intelligence methods, such as artificial neural networks (ANNs), can be utilized with pharmaceutical process and analytical data to aid process and product development and routine quality assurance [2].

In this presentation, multiple case studies will be introduced, where the applicability of ANNs on pharmaceutical analytical and process data was assessed to achieve real-time, non-destructive quality assessment. First, the quality of pharmaceutical tablets (*e.g.*, *in vitro* dissolution, tensile strength) was predicted using Raman and NIR spectroscopy, raw material properties (*e.g.*, particle size distribution), and various process conditions. The capability of multilayer perceptron (MLP) neural networks was compared to the traditional modeling approach, *i.e.*, partial least square regression (PLS). Results showed that ANNs could significantly improve the quantification and serve as an excellent and straightforward tool for data fusion applications. Furthermore, the problem of black-box modeling was addressed by studying the interpretability of ANNs using various tools, such as sensitivity analysis [3]. The applicability of automatically registered time-series process data was also evaluated to predict the outcome of the manufacturing process. For this, MLP models were compared with deep neural networks applicable for time-series analysis.

The results of this work could contribute to the more efficient analysis of routinely collected pharmaceutical manufacturing datasets, help to achieve consistent product quality, more effective product development, and eventually reach the vision of smart factories.

References:

[1] Arden, N. S., Fisher, A. C., Tyner, K., Lawrence, X. Y., Lee, S. L., Kopcha, M., *Int. J. Pharm.*, **602** (2021), 120554.

[2] Nagy, B., Galata, D. L., Farkas, A., Nagy, Z. K., *The AAPS Journal*, **24(4)**, (2022) 74.

[3] Nagy, B., Szabados-Nacsa, Á., Fülöp, G., Nagyné, A. T., Galata, D. L., Farkas, A., Mészáros, L. A., Nagy, Z. K., Marosi, G. (2023), *Int. J. Pharm.*, **633 (**2023), 122620.

Extending the limitations in the prediction of permeability with machine learning algorithms based on a diverse PAMPA dataset

Anita Rácz[1], Anna Vincze[2], György T. Balogh[2,3]
[1] Plasma Chemistry Research Group, Institute of Materials and Environmental Chemistry,
Research Centre for Natural Sciences, Magyar tudósok krt. 2., 1117 Budapest, Hungary
E-mail: racz.anita@ttk.hu
[2] Department of Chemical and Environmental Process Engineering, Budapest University
of Technology and Economics, Műegyetem rakpart 3., 1111 Budapest, Hungary
[3] Department of Pharmacodynamics and Biopharmacy,
University of Szeged, 6720 Szeged, Hungary

Gastrointestinal absorption is a key factor amongst the ADME-related (absorption, distribution, metabolism and excretion) pharmacokinetic properties; therefore, it has a major role in drug discovery and drug safety determinations. The Parallel Artificial Membrane Permeability Assay (PAMPA) can be considered as the most popular and well-known screening assay for the measurement of gastrointestinal absorption. Our study provides quantitative structure property relationship (QSPR) models based on experimental PAMPA permeability data for almost four hundred diverse molecules, which is a major extension of the applicability range of the models regarding chemical space. Two- and three-dimensional molecular descriptors were applied for the model building in every case. We have compared the performance of a classical partial least squares regression (PLS) model with two major machine learning algorithms: artificial neural networks (ANN) and support vector machine (SVM). Due to the applied gradient pH in the experiments, we have calculated the descriptors for the model building at pH values of 7.4 and 6.5, and compared the effect of pH on the performance of the models. After a complex validation protocol, the best model based on the ANN algorithm has an $R^2 = 0.91$ for the training set, and $R^2 = 0.84$ for the external test set. We have extended the applicable chemical space with a larger open-source dataset to provide a robust and more precise QSPR model for further applications. The developed models are capable of quickly predicting new compounds with an excellent accuracy compared to the previous PAMPA related QSPR models.

Chemometrics for personalized medicine

O. Ye. Rodionova[1], N. I. Kurysheva[2], G. A. Sharova[3], A. L. Pomerantsev[1]
[1]Federal Research Center for Chemical Physics RAS,
E-mail: oksana@chph.ras.ru
[2]The Ophthalmological Center of the Federal Medical and Biological Agency of the RF;
[3]Ophthalmology Clinic of Dr. Belikova, Moscow, Russia

Personalized medicine is a modern trend, which is an extension of traditional expert medicine and currently widespread population medicine. Personalized medicine involves the use of genetic or other biomarker information, and includes patient anatomical, environmental, and lifestyle factors. We believe that this approach is a typical problem of multivariate data analysis, which can be investigated using well-known chemometric methods.

Three issues among a wide variety are considered.

The first issue is the concept of novelty and similarity. This is complex concept that has numerous applications in medicine, especially in image analysis, in healthcare and clinical trials. Novelty is a popular concept, and there are many publications about it. Similarity topic in multivariate case has not been sufficiently studied yet. Two important principles for the detection of novelty and similarity are presented. The first one concerns the proper complexity of the model. The second point is interpretability, the important issue that is often overlooked. DD- SIMCA is applied for estimation of datasets similarities in clinical trials [1].

The second issue is a generalized characteristic of the success of a treatment. In order to evaluate the results of medical intervention, presented as multivariate data, we propose to assess the proximity of subjects to the Healthy group, which is chosen as a target class. Full distances calculated in the DD-SIMCA model is used as the univariate numerical characteristic to measure and compare the results of medical intervention [2].

The third issue is the development of the methodology for personalized choice of an effective method of medical intervention. Using principal component regression and variable selection method we have developed a selection criterion for the choice of the method of treatment of glaucoma. The special model with the reduced number of variables can be applied in routine clinical practice [3].

All issues are illustrated with the help of the real example of of treatment of an ophthalmic disease, primary angle closure of the anterior chamber of the eye.

References

[1] N.I. Kurysheva, A.L. Pomerantsev, O.Ye. Rodionova, G.A. Sharova, *J. Glaucoma*, **32** (2023), e43–e55

[2] O. Rodionova, N. Kurysheva, G. Sharova, A. Pomerantsev, *Anal. Chim. Acta*, **1250** (2023) 340958.

[3] N.I. Kurysheva, O.Ye. Rodionova, A.L. Pomerantsev, G.A. Sharova, *Biomed. Signal Process.Control,* **85** (2023) 104884.

High-performance thin-layer chromatography and multivariate image analysis in modelling of adulteration of *Salvia sp*. with olive leaves

N. Tomčić[1,2], M. Jankov[3], P. Ristivojević[1], J. Trifković[1], F. Andrić[1],*
[1] University of Belgrade - Faculty of Chemistry, Studentski trg 12-16, Belgrade, Serbia,
*E-mail: andric@chem.bg.ac.rs
[2] Profilab – Laboratory for testing, quality control and certification,
Vele Nigrinove 1, Belgrade, Serbia
[3] Innovative Centre of the Faculty of Chemistry Ltd., Belgrade, Serbia

Herbs and spices are important ingredients in various foods, medicines, and cosmetics. In order to increase profit margins, economically motivated adulteration of herb products is heavily present in food industry. According to the study carried out at the University of Bristol, 60% of oregano spices present on the EU market are heavily adulterated with olive, myrtle, sumac, cistus, and hazelnut leaves [1]. According to the same authors the sage products are adulterated by similar bulking agents [2]. However, no analytical method for detection of sage adulteration by these adulterants has not yet been reported to the best of our knowledge.

The aim of this study was to develop an analytical method for detection of sage adulteration by olive leaves using thin-layer chromatography coupled with digital image analysis and multivariate linear regression/classification (PLS and PLS-DA).

Twenty-four samples (4 – pure sage, 4 – pure olive leaves, 16 – mixtures of olive and sage leaves with content of added olive leaves varying in 5, 10, 20 and 70%) have been prepared, extracted, and analysed under typical normal-phase conditions. Chromatographic derivatization was done by Neu reagent (detection of polyphenolics), 2% $AlCl_3$ solution (detection of flavonoids), anisaldehyde-sulphuric acid reagent (detection of terpenes and terpenoids), DPPH reagent (detection of antioxidants) and 3% $FeCl_3$ solution (detection of phenolic compounds). Derivatized plates were inspected under visible (VIS) or UV-light (FLD) and digital images of chromatograms were recorded. After acquisition of chromatographic signals, the PLS-DA and PLS models of various complexities were built.

PLS and PLS-DA models based on chromatographic signals obtained after derivatization by $FeCl_3$, anisaldehyde-sulphuric acid, and DPPH demonstrated good statistical performances with R^2 ranging 0.894 – 0.998, and relative prediction error of 4-12%. Misclassification error < 4% was obtained in the case of DPPH and anisaldehyde-sulphuric acid derivatization. Contrary, derivatization by Neu reagent and $AlCl_3$ did not result in statistically satisfactory PLS and PLS-DA models.

Therefore, the high-performance thin-layer chromatography combined with multivariate image analysis could be used as fast, relatively simple, low cost and green analytical tool for assessment of sage adulteration by olive leaves.

References

[1] C. Black, S. A. Haughey, O. P. Chevallier, P. Galvin-King and C. T. Elliott, *Food Chem.*, **210** (2016) 551-557.

[2] https://rb.gy/uq1nl (Last time accessed on 11 July 2023)

Chemometrics as a tool to monitoring corrosion degradation of selected alloys in real conditions

Gy. Vastag[1], S. Apostolov[1], Š. Ivošević[2]
[1]University Novi Sad, Faculty of Sciences, Novi Sad, Serbia,
E-mail: djendji.vastag@dh.uns.ac.rs
[2]University of Montenegro, Faculty of Maritime Studies Kotor, Montenegro

Monitoring of the corrosion process of alloys in the real conditions often results in extensive data, characterized by complex interdependence, but in the other hand also by the large degree of mutual deviation. As first, large scattering of the obtained results, make it very difficult to form the correct conclusions about the real impact of tested parameters on the corrosion behaviour of alloys. On the other hand, in many cases the high interdependence between the corrosion factors can also greatly burden the analysed system and therefore significantly complicate the recognition of the main impact. Multivariate analysis, especially the Principal component analysis (PCA), become increasingly popular in the processing of this type of data, due to its ability to recognize and eliminate redundant data. The goal of this study was to examine the possibility of using multivariate analysis methods in the processing of the corrosion test results obtained under real conditions.

In this work, with the aim of monitoring the corrosive degradation, alloy samples after different exposure times, in various marine environments, were analysed by using the Energy Dispersive Spectrometer (EDS). The obtained results were, as expected, very voluminous, complex and with lot of unexpected deviations.

In order to identify the main corrosion factors in the given conditions, regardless of deviating data the obtained EDS results were processed by using the selected multivariate analysis. The results of Cluster analysis and Principal component analysis indicate that used multivariate methods in combination with EDS analysis can be used successfully for identification the most important corrosion factors, as well as their influence on the degradation of these alloys under the given conditions.

Contactless chemical analysis with high-frequency inductance coil and chemometrics

Ekaterina Yuskina[1], Nikodim Makarov[1], Maria Khaydukova[2,3], Tatiana Filatenkova[2], Olga Shamova[2], Valentin Semenov[1,4], Vitaly Panchuk[1,4], and Dmitry Kirsanov[1]
[1] Institute of Chemistry, St. Petersburg University, St. Petersburg, Russia
[2] Laboratory of Alternative Antimicrobial Biopreparations, World-Class Research Center "Center for Personalized Medicine", FSBSI Institute of Experimental Medicine, St. Petersburg, Russia
[3] Laboratory of Peptide Chemistry, Institute of Human Hygiene, Occupational Pathology and Ecology, Saint Petersburg, Russia
[4] Institute for Analytical Instrumentation RAS, St. Petersburg, Russia

An urgent task of modern analytical chemistry is the analysis of real objects in field conditions and the preference is given to contactless methods that allow analysis without sampling and sample preparation. In this work, a simple and inexpensive detector with wide analytical application is proposed. The device is based on the ideas of high-frequency contactless conductometry, which was actively studied in the middle of the twentieth century; however, instead of recording the signal at a certain single AC frequency, as it was done in those years, it allows recording a whole spectrum at different frequencies. The sensor device is based on an inductance coil connected to a high-frequency electric field generator (4-113 MHz). The sample in a plastic tube is placed inside the coil and becomes the core of the inductor, where it changes the properties of the electrical signal passing through the coil. A receiver connected to the coil records the response spectrum, which depends in a complex way on the conductivity, dielectric constant, polarizability of the sample, as well as its magnetic and capacitive properties. The obtained spectra are processed by chemometric algorithms to obtain qualitative and quantitative information about the samples.

To date, it has already been shown that this variant of electrochemical spectroscopy allows to distinguish between the substances with different physical and chemical properties; allows to carry out quantitative analysis of aqueous solutions of inorganic salts; to determine integral characteristics of complex multicomponent samples (e.g., fat content of dairy products, water content in EtOH-H_2O mixtures); to recognize media with cultures of different cells and bacteria [1]. The presentation will provide the details on the construction and performance of this new contactless sensor device.

Acknowledgement
This research was funded by Russian Science Foundation, grant number RSF 23-23-00114.

References

[1] E. Yuskina, N. Makarov, M. Khaydukova, T. Filatenkova, O. Shamova, V. Panchuk and D. Kirsanov, *Anal.l Chem*, **94(35)**, (2022) 11978-11982.

Molecular descriptors based on automorphism data

K. Varmuza[1,2], M. Dehmer[3,4], P. Filzmoser[1]
[1] Vienna University of Technology, Institute of Statistics and Mathematical Methods in Economics, Computational Statistics, Vienna, Austria
Email: kurt.varmuza@tuwien.ac.at
[2] Vienna University of Technology, Institute of Chemical, Environmental and Bioscience Engineering, Vienna, Austria
E-mail: kurt.varmuza@tuwien.ac.at
[3] UMIT TIROL - Private University For Health Sciences and Health Technology, Eduard Wallnöfer Zentrum, Hall in Tyrol, Austria
[4] Swiss Distance University of Applied Sciences, Department of Computer Science, Brig, Switzerland

Chemical structures are represented by colored graphs with vertices (atoms) given here by one of three elements (C, N, O), and edges (bonds) given here by one of four bond types (single, double, triple, aromatic). The complete automorphism group of such chemical structure graphs is determined [1,2] and evaluated by functions in the programming environment R [3].

Molecular descriptors based on automorphism data comprise symmetry and entropy measures, as well as roots of graph polynomials. For instance, the orbit polynomial [4] has been defined by orbit data for vertices and edges obtained from the automorphism group of a graph. A previous work using this concept for exhaustive sets of isomeric alkanes [5] is extended to coloured graphs (CHNO molecules). Such descriptors can be used together with others [6,7] for QSPR (quantitative structure property relationships) models for the prediction of molecular properties and class memberships. QSPR models have been created by linear PLS regression optimized and evaluated by the strategy repeated double cross validation [8].

References

[1] K. Varmuza, et al.: *Croatica Chemica Acta*, **78** (2005) 141-149.

[2] H. Scsibrany, K. Varmuza: Software SubMat, 2004, http://www.lcm.tuwien.ac.at

[3] R, A language and environment for statistical computing, R Development Core Team Foundation for Statistical Computing, http://www.r-project.org , Vienna, Austria, 2023.

[4] M. Dehmer, et al.: *IEEE Access*, **8** (2020) 36100-36112.

[5] K. Varmuza, et al.: *Croatica Chemica Acta*, **94** (2021) 47-58.

[6] R. Todeschini, V. Consonni: Molecular descriptors for chemoinformatics, Wiley-VCH, Weinheim, Germany (2009).

[7] CORINA-Symphony, Chemoinformatics program package, Molecular Networks GmbH & Altamira LLC, Nuremberg, Germany (2023).

[8] P. Filzmoser, B. Liebmann, K. Varmuza, *J. Chemometr.*, **23** (2009) 160-171.

Adjusted Pareto scaling

K. Varmuza, P. Filzmoser
Vienna University of Technology, Institute of Statistics and Mathematical Methods in Economics, Computational Statistics, Vienna, Austria
Email: kurt.varmuza@tuwien.ac.at

The performance of multivariate models in chemometrics often depends on the applied method of scaling the x-variables. Autoscaling and Pareto scaling are widely used; here an *adjusted Pareto scaling* is suggested–covering the range from no scaling via classical Pareto scaling to autoscaling, and is compared with range scaling and vast scaling [1].

(1) *Autoscaling* of a variable is performed by x_c /s with x_c for the centred original variable, and s the standard deviation of the variable. (2) *Pareto scaling* is performed by $x_c /s^{0.5}$. The scaling effect is weaker than with autoscaling, noise is less amplified, and variables with a high original variance retain part of their importance for the model. (3) *Adjusted Pareto scaling* is performed by x_c /s^{P} with P varying between 0 (no scaling) and 1 (autoscaling), typically in steps of 0.1, thus including classical Pareto scaling with $P = 0.5$. (4) *Range scaling* is defined by $x_c/(x_{HIGH} - x_{LOW})$ with the variable spread given for instance by the quantiles 0.98 and 0.02. (5) *Vast scaling* considers the relative standard deviation $s_{RSD} = s/c$ (c is the mean or median of the variable) possessing rather small values for stable variables and scaling by $x_c/(s.s_{RSD})$.

These scaling methods are compared for PLS regression models and applying the strategy repeated double cross validation [2]. In one example GC retention indices are modelled by molecular descriptors, in the other the glucose content of fermentation samples is modelled by NIR absorbances. Parameters for scaling are varied and robust versions are considered.

Results show that appropriate scaling of x-variables by methods based on the spread may improve the performance of multivariate regression models. However, the effect has to be tested and optimized. Standard Pareto scaling with $P = 0.5$ may be not optimal and varying P between 0 and 1 is recommended. For some data sets, (robust) scaling methods based on the variable spread, like range scaling or vast scaling, may outperform (adjusted) Pareto scaling. No general rules, related to data properties, seem to be evident.

References

[1] J. Walach, *et al.*: In J. Jaumot, *et al.*: Comprehensive analytical chemistry. Data analysis for omics sciences: Methods and applications, Elsevier, Amsterdam, 165-196 (2018).
[2] P. Filzmoser, B. Liebmann, K. Varmuza: Repeated double cross-validation *J. Chemometr.*, **23** (2009) 160-171.

Author index

P.K. Hopke	L17
N. Ionov	L14
Á. Ipkovich	L19
S. Ivanov	L14
Š. Ivošević	P21
M. Jankov	P20
T. Jámbor	L08
E. Jamrozik	L13
K. Khan	L23
S. Kalli	L27
D. Karasev	L14
M. Khaydukova	P22
D. Kirsanov	L15, P22, P25
A. Kiss	P04
Sz. Klébert	L09
E. Kontsek	P04
Á. Kopasz	L08
Z. T. Kosztyán	L19
D. Kovács	P11
S. Kovács	L05, P12
S. Krzebietke	L18
N. I. Kurysheva	P19
O. M. Kvalheim	L01
R. Lasfar	L02, P13
K. László	L09
R.G. Linington	L01
M. Lubienski	P03
N. Makarov	P22
T. Maxfield	L12
L.A. Mészáros	P14
N. Mobaraki	P15
M. Mócz	P16
V. Motto-Ros	P06
B. Nagy	P17, L08
Z.K. Nagy	P14, P17

K. Neymeyr	L04
D. A. Nikitina	L22
A. Muñoz de la Peña	P01
L. Pieszczek	L18, L20, P08
V. Panchuk	P22, L15
A. Pesti	P04
S. Podlewska	L13
P. Pollner	P04
A.L. Pomerantsev	L25, P19
V. Poroikov	L14
E. Puchkova	P25
K.I. Popov	L12
P.M. Puttich	P03
M.V. Putz	L24
A. Rácz	L09, P18
P. Ristivojević	P20
O.Ye. Rodionova	L25, P19
A. Rudik	L14
K. Rybicka	P08
A. Salgó	P09
M. Saveliev	L15
P. Savosina	L14
M. Sawall	L04
V. Semenov	P22
O. Shamova	P22
G. A. Sharova	P19
J. Shimshoni	L21
J. Slezsák	P09, P16
B. Sobolev	L14
I. Špánik	P10
I. Stanimirova	L17, L18
L. Stolbov	L14
V. Sukhachev	L14
O. Tarasova	L14
D. Tátraaljai	L09

I.G. Tetko	L03
N. Tomčić	P20
G. Tóth	L02, P13
A. Tres	P02
J. Trifković	P20
A. Tropsha	L12
A. Turanov	P25
B. Vajna	L11
E. Valli	P02
E. Varga	L08
K. Varmuza	P24, P25
Dj. Vastag	P21
A. Veselovsky	L14
G. E. Viacava	P02

S. Vichi	P02
W.S. Vidar	L01
A. Vincze	P18
Z. Vincze	P05
J-P. Vincken	L27
J. Wellnitz	L12
N. Vladimirova	P25
A. Wojtuch	L13
E. Yuskina	P22
J. Zaldarriaga Heredia	P01
I.G. Zenkevich	L22