Analytical Methods

TECHNICAL NOTE

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2025, **17**, 9909

Received 6th August 2025
Accepted 25th November 2025

DOI: 10.1039/d5ay01299f

rsc.li/methods

End-to-end workflows for liquid biopsy biotyping analysis using combined MALDI MS and machine learning approach

Lukáš Pečinka,*[ab] Jaromíra Pantůčková,[c] Monika Vlachová,[d] Lukáš Moráň,[ae] Tereza Růžičková,[df] Petra Weselá,[eg] Lubomír Prokeš,[h] Josef Havel,[i] Luděk Pour,[f] Sabina Ševčíková[dj] and Petr Vaňhara*[be]

MALDI MS analysis of liquid biopsy combined with ML enables non-invasive disease screening and monitoring. Here we present an open-source R-based workflow covering all steps from raw data preprocessing to predictive model evaluation. The pipeline is fully customizable and transparent, with validation performed on clinical plasma samples from hemato-oncological patients. This workflow enhances data reproducibility, enables a straightforward end-to-end workflow for liquid biopsy biotyping, and provides a foundation for integrating MALDI MS into routine clinical workflows.

Laser desorption ionization mass spectrometry (LDI MS) was developed over half a century ago. The introduction of matrices based on organic acids that absorb laser wavelengths enabled the expansion of matrix-assisted laser desorption/ionization (MALDI MS) applications to various fields, *e.g.*, identification, structural analysis, and monitoring of biochemical processes *in vivo* and *in vitro*. MALDI MS analytical technique has progressed remarkably, and its clinical applications are expanding rapidly. In clinical practice, MALDI MS applications can be broadly divided into several key areas, including the search for new diagnostic biomarkers at the level of metabolites, peptides, and proteins, and the introduction of new bioanalytical methods into clinical laboratories.[1,2] MALDI MS is currently being used in clinical laboratories on a routine basis as an automated platform for identification and classification of microorganisms (MALDI Biotyper® Library, BRUKER Daltonics), among other applications.[3] MALDI MS profiling is a high-performance technology that compares the specific spectral fingerprint (protein, peptide, and metabolites) obtained from biological materials, such as liquid biopsies or cell cultures. This approach is both time- and cost-effective.[4–7]

The integration of MALDI MS with machine learning (ML) prediction models represents a highly advanced technique, enabling the analysis of complex biological samples including the liquid biopsies (*e.g.*, blood plasma/serum, urine, and cerebrospinal fluid), and generating a comprehensive, patient-specific spectral fingerprint (patient ID).[8–11] Furthermore, it allows for the detection of spectral patterns and their alterations, which may serve as important markers for disease screening, progression monitoring, and treatment response prediction.[12–14] The efficacy of this analytical approach has been substantiated by a number of recent scientific studies.[4,15–17]

A major limitation to the advancement of this methodology is the heterogeneity of data formats generated by MALDI mass spectrometers.[18] The lack of standardized data structures leads to incompatibility between software solutions, forcing users to acquire and manage multiple software platforms. This not only increases analytical complexity and costs but also limits efficiency. Notably, even instruments from the same manufacturer may produce data that are not interoperable across product lines, further complicating comprehensive data integration. This represents a significant barrier to implementing MALDI MS profiling in routine clinical practice. Therefore, efforts are underway to develop and optimize tools that would enable the processing and evaluation of mass spectra from commonly used formats without the constraints of product licenses.

[a] *Research Centre for Applied Molecular Oncology (RECAMO), Masaryk Memorial Cancer Institute, Žlutý Kopec 7, Brno 60200, Czech Republic*

[b] *International Clinical Research Center, St. Anne's University Hospital Brno, Pekařská 53, Brno 65691, Czech Republic. E-mail: pvanhara@med.muni.cz*

[c] *RECETOX, Faculty of Science, Masaryk University, Kamenice 753/5, Brno 62500, Czech Republic*

[d] *Babak Myeloma Group, Department of Pathophysiology, Faculty of Medicine, Masaryk University, Kamenice 3, Brno 62500, Czech Republic*

[e] *Department of Histology and Embryology, Faculty of Medicine, Masaryk University, Kamenice 3, Brno 62500, Czech Republic*

[f] *Department of Internal Medicine, Hematology and Oncology, University Hospital Brno, Jihlavská 20, Brno 62500, Czech Republic*

[g] *Biostatistics Department, St. Anne's University Hospital Brno, Pekařská 664/53, Brno 60200, Czech Republic*

[h] *Department of Physics, Chemistry and Vocational Education, Faculty of Education, Masaryk University, Poříčí 538/31, Brno 60300, Czech Republic*

[i] *Department of Chemistry, Faculty of Science, Masaryk University, Kamenice 5, Brno 62500, Czech Republic*

[j] *Department of Clinical Hematology, University Hospital Brno, Jihlavská 20, Brno 62500, Czech Republic*

Although more computationally efficient languages exist, *e.g.*, Python, the R programming language is a popular choice among researchers.[19] R is appreciated for its user-friendly interface and the extensive availability of supporting libraries. These libraries allow researchers to focus on the scripts themselves rather than having to create complex libraries *de novo*.[20] The core functionality is provided by the MALDIquant library. In combination with the MALDIquantForeign, and MALDIrppa libraries, it enables the import of raw mass spectra, their processing, and export for subsequent statistical analysis.[21] These libraries facilitate processing standard MALDI MS formats, including mzML, XML, ASCII, CSV, Bruker daltonics flex files, and the imaging data format (imzML). Many R libraries have been developed for multivariate statistical analysis (ropls, dendextend), data visualization (ggplot2), and ML modelling (caret).[22,23] Therefore, R is preferred over Python in the context of MS. A wide range of open-source software tools *e.g.* Mass++, MZmine 2, mMass 3, OpenMS, and XCMS has been developed to facilitate the visualization, preprocessing, and analysis of MS data.[24–28] While these platforms offer a variety of advanced functionalities, they often lack the flexibility required for fully customizable workflows. In particular, they typically do not support integration of user-defined code or allow fine-tuning of preprocessing parameters and statistical functions. This can limit their applicability in research contexts that demand high adaptability or the incorporation of novel analytical strategies. The creation of a graphical user interface in Shiny (R), Python, or developing new libraries would limit the ability to modify functions and set their parameters. For these reasons, developing an R-based workflow as a script with comprehensive documentation appears to be a viable solution for all types of users. In addition, the user-modifiable workflow, accompanied by detailed descriptions, allows for easy parameter adjustment and the integration of new applications by the user. The objective of this technical note is to present an open-source workflow implemented in R for end-to-end MALDI-MS analysis of complex human plasma samples, to demonstrate its application in a real liquid biopsy dataset (multiple myeloma *versus* healthy controls), and to provide a reproducible guide for the research community.

Experimental section

The systematic workflow presented herein is also suitable for new users of R. It allows incorporation of a text commentary into a script, thus facilitating comprehension of the sequence of steps involved and enabling modification as required. The workflow is divided into several parts for the sake of clarity: (1) mass spectra pre-processing, alignment, feature selection, and descriptive statistics, (2) unsupervised ML methods, primarily Principal Component Analysis (PCA), and (3) supervised ML algorithms, including partial least-squares-discriminant analysis (PLS-DA), Support Vector Machine (SVM), Decision Tree (DT), Random Forests (RF), and Artificial Neural Networks (ANN).[29]

MALDI MS

Plasma samples from 20 healthy donors and 20 MM (collected at diagnosis) were obtained at the University Hospital Brno. All patients signed informed consent forms approved by the ethics committee of the hospital following the Declaration of Helsinki. All plasma samples were handled as previously described and samples were stored at −80 °C.[30] Extraction of proteins and peptides was performed as described previously.[4] Briefly, 25 μL of plasma sample was precipitated with 50 μL of ACN in a two-step protocol. In the first phase, lipids, salts, and metabolites were removed, while large proteins (*e.g.* Albumin) were removed in the second step (acid hydrolysis).[31] The resulting protein extracts were stored at −20 °C and subsequently subjected to MALDI MS analysis. Extracts were mixed 1 : 1 (v/v) ratio with MALDI matrix (20 mg μL^{-1} sinapic acid in 50% acetonitrile, 2.5% trifluoroacetic acid) and 2 μL spotted in five replicates on a MALDI target plate. In total 200 mass spectra were recorded on MALDI-7090™ TOF–TOF mass spectrometer (Shimadzu, Japan) equipped with a 2 kHz ultra-fast solid-state UV laser (Nd-YAG: 355 nm). Mass spectra were acquired in linear positive mode (2–20 kDa; pulse extraction 12.5 kDa; 1 kHz, 100 μm). Each spectrum was averaged from 5 profiles of 1000 shots and externally calibrated with ProMix1 (2.8–17 kDa). All raw data were exported into a commonly used mzML file format for use in R for further analysis. A detailed description of the sample preparation and instrumentation has been provided in our previous publications.[4,16,32]

Code

The workflow for MALDI MS data analysis employs R open-source libraries and the integration of custom scripts. MALDIquant and MALDIrppa libraries are incorporated into mass spectra processing. Furthermore, ropls, dendextend, and caret libraries are incorporated for multivariate statistical modeling and ML. The pipeline is designed to process the commonly used mzML format natively supported by commercial software. The code can be readily adapted to integrate future experimental setups and the utilization of *e.g.*, TXT and CSV files across the scientific community. The source code and tutorial workflow are available on GitHub (https://github.com/pantuja/Workflow-for-Liquid-Biopsy-Biotyping-Analysis-Using-Combined-MALDI-MS-and-Machine-Learning-Approach-/) and Zenodo (https://doi.org/10.5281/zenodo.14561887).

Data processing workflow

The mass spectra processing involves several steps: smoothing using a Savitzky–Golay filter with a half-window size parameter, baseline correction using the statistics-sensitive non-linear iterative peak-clipping (SNIP) method, intensity calibration ($\sum X_i = 1$, where X_i denotes intensities of corresponding peaks in mass spectra), transformation, spectra alignment to address non-systematic shifts in technical replicates acquired across varying time points, mass range trimming, conversion of technical replicates of each sample into an average mass spectrum, and peak detection utilizing the MAD noise estimation algorithm with a signal-to-noise ratio and a half-window size parameters. To eliminate inter-sample variability within a single group, the feature matrix is constructed only from the signals detected in at least 15% of the total mass spectra. The presence of signals exclusively detected within individual mass spectra or samples may be attributable to the specific treatment undergone, dietary settings, insufficient sample extraction, or contamination of the MALDI target. The feature matrix was employed for subsequent multivariate statistical analysis and the development of selected ML prediction models. The user can easily modify the parameters or use alternative algorithms outlined in the MALDIquant library manual.

Results and discussion

The workflow is illustrated graphically in Fig. 1. The unwanted distortions caused by the event of measurement, namely oscillations in the *m*/*z* values, are eliminated by mass spectra alignment and/or warping to avoid signal readings outside their maxima. To evaluate the effectiveness of alignment and warping, the average shift of each spectrum relative to a reference spectrum was calculated (Fig. 2A). The spectrum with the highest mean correlation was chosen as the reference, ensuring it was the most representative of the overall dataset and reducing the influence of outliers or random variation. Prior to the training of ML models, exploratory multivariate data analysis was performed to assess potential group-wise differences in the data structure. For this purpose, PCA, PLS-DA, and Orthogonal PLS-DA (OPLS-DA) were employed. The 2D or 3D PCA visualization shows the grouping potential of the data without prior knowledge of the diagnostic group (Fig. 2B).

The PLS-DA and OPLS-DA (for two clinical groups), a supervised extension of PCA, is designed to improve the interpretability of complex data by separating predictive and non-predictive variations while incorporating clinical groups (class information). This method is optimized according to the R2Y (a measure of goodness-of-fit of the model), Q2Y (model validity, *i.e.*, how well the model predicts new data based on cross-validation), and RMSEE (root mean square error of estimation giving the average estimation error) parameters (Fig. 2C). Parameters are detailed explained in ropls library article.[33] Optimized OPLS-DA with 2 orthogonal components is shown in Fig. 2D, which demonstrates the method's capacity to cluster data according to clinical groups. The user can easily apply the PLS-DA method instead and get the same parameters as for OPLS-DA given in Fig. 2C and D. Consequently, supervised ML algorithms have been incorporated into this workflow to predict clinical groups based on the MS data: PLS-DA, SVM, DT, RF, and ANN. The construction, training, and testing of ML predictive models are contingent upon the number of samples in the study. When the sample size is limited, the data are not divided into a training and test set. Instead, the data are trained as a single batch and tested using 10× repeated 5-fold cross-validation (CV) or leave-one-out CV. If the number of samples is sufficiently high, the data are divided into two distinct sets: a training set comprising 70% of the total data and a test set comprising the remaining 30%. The ML predictive models are trained using the caret library, and their performance is evaluated using standard classification performance metrics, including accuracy, sensitivity, specificity, F1 and F2 score, and area under the curve (AUC).[34] These parameters are compared for training and test data, as well as for CV in the small dataset (Fig. 3A). Evaluation of the models was based on their overall accuracy in predicting the outcome (Fig. 3B).

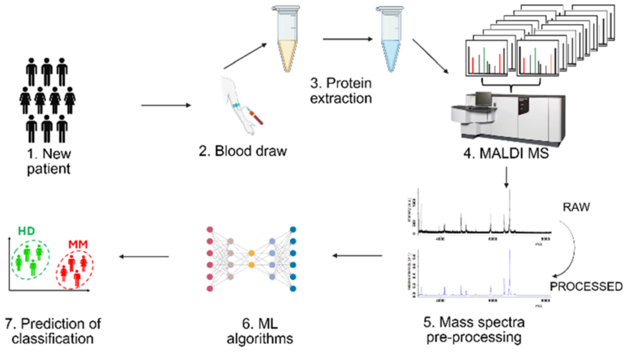


Fig. 1 Illustration of the workflow for liquid biopsy analysis using MALDI MS and machine learning (ML). This approach includes patient material collection (1) and (2), sample preparation (3), mass spectra acquisition (4) and pre-processing (5), training and optimization of ML algorithms (6), outputs from the ML predictive model and the final classification of samples (7).

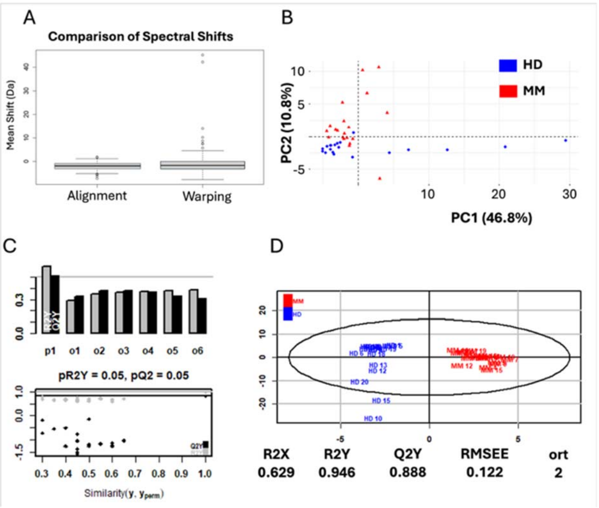
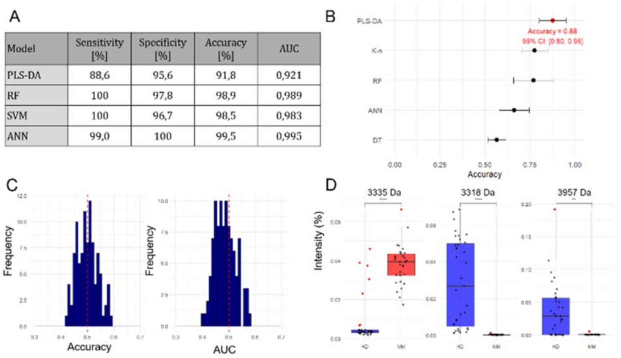


Fig. 2 Box plot showing average shift size of individual spectra compared to the reference spectrum using mass spectra alignment and warping (A). PCA score plot illustrating sample clustering based on unsupervised spectral data analysis (B). Diagnostic evaluation of supervised OPLS-DA model, including CV metrics and clustering performance. pR2Y and pQ2 represent *p*-value for these parameters, ort (number of orthogonal components). (C). Optimized OPLS-DA score plot showing segregation of healthy donors (HD) and multiple myeloma patients (MM) (D).

Overfitting is a common issue in ML, especially when the number of predictors greatly exceeds the number of samples. This is particularly true in data generated in MS.[35,36] In such cases, a model may perform well on training data but fail to



Model	Sensitivity [%]	Specificity [%]	Accuracy [%]	AUC
PLS-DA	88,6	95,6	91,8	0,921
RF	100	97,8	98,9	0,989
SVM	100	96,7	98,5	0,983
ANN	99,0	100	99,5	0,995

Fig. 3 Outputs from ML models: classification performance metrics (A). Comparative accuracy of ML prediction models (B). Accuracy and AUC for randomized class labels (C). Box plots showing the relative intensity of 3 selected signals with the highest discriminatory power between patients with HD and MM. Peaks are labelled according to their *m/z* and ordered according to their importance in the ANN model. Outliers are visualized as red dots (D).

generalize to new, external data, as it learns noise instead of true patterns.[36,37] To detect overfitting and assess model generalization, techniques such as feature reduction, hyperparameter optimization, and permutation testing are often used. To test for overfitting of the ML prediction model, variable reduction was applied using the non-parametric Wilcoxon rank-sum test. Test was applied separately to each *m*/*z* variable, and a ranking score was assigned to quantify the extent of intensity differences between the two groups. Variables were then ordered according to their ranking scores and selected for model retraining. In this case, models retrained with the top 5 and 10 features achieved 94.7% and 98.1% accuracy, respectively, indicating that the original model was not over-fitted. Additionally, permutation testing (100 iterations with randomized class labels) confirmed this, as classification accuracy and AUC values dropped to chance levels (50% and 0.5), supporting the conclusion that the model captured real differences between groups (Fig. 3C). Therefore, eliminating the influence of noise on the predictive power of the model. The number of iterations can be easily modified as the more complex dataset is examined. The variable importance for the ANN model was calculated using the absolute connection weights between the input and hidden layers, reflecting the relative contribution of each predictor to the network's output. Box plot for the most discriminative variables and the *p*-value calculated using Wilcoxon test was presented in Fig. 3D. The presented workflow herein can be easily modified and applied for the analysis of various data generated from MALDI MS analysis. In a recent study published by our research group, we applied a modified workflow for MALDI MS analysis of peripheral blood plasma samples of hemato-oncological diseases. The combination of PLS-DA ML predictive model and MALDI MS profiling enabled discrimination between MM and plasma cell leukaemia, two closely related monoclonal gammopathies, achieving an accuracy of 71.5% under 10× repeated 5-fold CV.[4] A subsequent study comprising 172 patient samples (peripheral blood plasma) successfully differentiated between patients with primary extramedullary disease (EMD) and MM patients. The PLS-DA model demonstrated high sensitivity (86.4%), accuracy (78.4%), and specificity (72.4%) in predicting primary EMD.[16] We also collected a multidimensional dataset combining MALDI MS and clinical data to distinguish between early relapse MM patients (within 6 months) and patients who relapse after 6 months (accuracy for PLS-DA: 74.1%).[32] The transfer of this type of analysis to clinical practice was introduced. Wolrab *et al.* developed a reproducible, robust, and high throughput lipidomic profiling approach combining MS and ML for the detection of pancreatic ductal adenocarcinoma in human serum.[38] It allows screening of at least 2000 samples per month on one MS system. Nowadays, this method has a USA patent, and it is under clinical trials.

Conclusions

The development and implementation of programming tools have significantly facilitated the handling and interpretation of MALDI MS data. We present an R-based, fully scriptable workflow for MALDI MS data processing and ML analysis tailored to liquid biopsy profiling. The pipeline enables robust spectral preprocessing, feature selection, and classification modelling with demonstrated applicability to haematological diseases. Its open-source nature ensures transparency, adaptability, and reproducibility, supporting broader use in translational research. This tool represents a step toward standardizing MALDI MS profiling for potential clinical integration.

Author contributions

Conceptualization: L. Pe., J. P., M. V., S. Š., P. V., data curation: L. Pe., L. Pr., formal analysis: L. Pe., P. V; funding acquisition: S. Š., P. V., L. M.; methodology: L. Pe., M. V.; validation: P. V., L. Pr., P. W.; writing – original draft: L. Pe., M. V.; writing – review and editing: L. Pe., J. P., M. V., L. M., T. R., P. W., L. Pr., S. Š., J. H., L. Po, P. V. All authors read and agreed to the published version of the manuscript.

Conflicts of interest

There are no conflicts to declare. This study was conducted in accordance with the current version of the Helsinki Declaration. This research has been approved by the Ethics committee of the University Hospital Brno. Informed consent was obtained from all subjects involved in the study.

Data availability

Experimental data, source code and a tutorial workflow are available on GitHub (https://github.com/pantuja/Workflow-for-Liquid-Biopsy-Biotyping-Analysis-Using-Combined-MALDI-MS-and-Machine-Learning-Approach-/) and Zenodo (https://doi.org/10.5281/zenodo.14561887).

Acknowledgements

The authors wish to thank all MM patients and their caregivers for participating in this study. This study was supported by the internal grants of Masaryk University MUNI/A/1738/2024, MUNI/A/1587/2023, MUNI/11/PU/1/2025, MUNI/LF-ACC/1421/2024; internal grant of St. Anne's University Hospital in Brno; by grants of the Ministry of Health of the Czech Republic, grants no. NU23J-08-00006, NU23-08-00241 and NW24J-03-00038; and by European Union projects SALVAGE (OP JAC; reg. no. CZ.02.01.01/00/22_008/0004644, co-funded by the European Union and by the State Budget of the Czech Republic), and CREATIC (grant agreement no. 101059788).

References

1 D. Li, J. Yi, G. Han and L. Qiao, *ACS Meas. Sci. Au*, 2022, **2**, 385–404.

2 C. Zambonin and A. Aresta, *arXiv*, MDPI, 2022, preprint, arXiv:27061925, DOI: 10.3390/molecules27061925.

3 A. Haider, M. Ringer, Z. Kotroczó, C. Mohácsi-Farkas and T. Kocsis, *arXiv*, MDPI, 2023, preprint, arXiv:14010008, DOI: 10.3390/microbiolres14010008.

4 L. Pečinka, M. Vlachová, L. Moráň, J. Gregorová, V. Porokh, P. Kovačovicová, M. Almáši, L. Pour, M. Štork, J. Havel, S. Ševčíková and P. Vaňhara, *J*, 2023, **34**, 2646–2653.

5 L. Pečinka, L. Moráň, P. Kovačovicová, F. Meloni, J. Havel, T. Pivetta and P. Vaňhara, *Heliyon*, 2024, **10**(9), 29936–29946.

6 M. Deulofeu, E. M. Peña-Méndez, P. Vaňhara, J. Havel, L. Moráň, L. Pečinka, A. Bagó-Mas, E. Verdú, V. Salvadó and P. Boadas-Vaello, *ACS Chem. Neurosci.*, 2023, **14**, 300–311.

7 P. Vaňhara, L. Kučera, L. Prokeš, L. Jurečková, E. M. Peña-Méndez, J. Havel and A. Hampl, *Stem Cells Transl. Med.*, 2018, **7**, 109–114.

8 P. Vaňhara, L. Moráň, L. Pečinka, V. Porokh, T. Pivetta, S. Masuri, E. Maria Peña-Méndez, J. Elías Conde González, A. Hampl and J. Havel, in *Mass Spectrometry in Life Sciences and Clinical Laboratory*, IntechOpen, 2021.

9 Y. Wang, K. Zhang, T. Tian, W. Shan, L. Qiao and B. Liu, *ACS Appl. Mater. Interfaces*, 2021, **13**, 4886–4893.

10 S. Long, Q. Qin, Y. Wang, Y. Yang, Y. Wang, A. Deng, L. Qiao and B. Liu, *Talanta*, 2019, **200**, 288–292.

11 M. Buszewska-Forajta, P. Pomastowski, F. Monedeiro, A. Król-Górniak, P. Adamczyk, M. J. Markuszewski and B. Buszewski, *Talanta*, 2022, **236**, 122843.

12 J. Bai, Y. Yang, J. Wang, L. Zhang, F. Wang and A. He, *Clin. Proteomics*, 2019, **16**, 17.

13 M. A. Koc, T. A. Wiles, D. C. Weinhold, S. Rightmyer, A. L. Weaver, C. T. McDowell, J. Roder, S. Asmellash, G. A. Pestano, H. Roder and R. W. Georgantas III, *J. Mass Spectrom. Adv. Clin. Lab*, 2023, **30**, 51–60.

14 J. S. Weber, M. Sznol, R. J. Sullivan, S. Blackmon, G. Boland, H. M. Kluger, R. Halaban, A. Bacchiocchi, P. A. Ascierto, M. Capone, C. Oliveira, K. Meyer, J. Grigorieva, S. G. Asmellash, J. Roder and H. Roder, *Cancer Immunol. Res.*, 2018, **6**, 79–86.

15 F. Barceló, R. Gomila, I. de Paul, X. Gili, J. Segura, A. Pérez-Montaña, T. Jimenez-Marco, A. Sampol and J. Portugal, *PLoS One*, 2018, **13**, e0201793.

16 M. Vlachová, L. Pečinka, J. Gregorová, L. Moráň, T. Růžičková, P. Kovačovicová, M. Almáši, L. Pour, M. Štork, R. Hájek, T. Jelínek, T. Popková, M. Večeřa, J. Havel, P. Vaňhara and S. Ševčíková, *Sci. Rep.*, 2024, **14**, 18777.

17 X. Han, Y. Yang, J. Lu, Y. Lin, D. Zhang, L. Lin and L. Qiao, *Chin. Chem. Lett.*, 2025, **36**, 110183.

18 A. Halder, A. Verma, D. Biswas and S. Srivastava, *Drug Discovery Today:Technol.*, 2021, **39**, 69–79.

19 J. Lai, C. J. Lortie, R. A. Muenchen, J. Yang and K. Ma, *Ecosphere*, 2019, **10**, e02567.

20 F. M. Giorgi, C. Ceraolo and D. Mercatelli, *arXiv*, MDPI, 2022, preprint, arXiv:1205064, DOI: 10.3390/life12050648.

21 S. Gibb and K. Strimmer, *Bioinformatics*, 2012, **28**, 2270–2271.

22 M. Kuhn, *J. Stat. Softw.*, 2008, **28**, 1–26.

23 H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo and H. Yutani, *J. Open Source Softw.*, 2019, **4**, 1686.

24 S. Tanaka, Y. Fujita, H. E. Parry, A. C. Yoshizawa, K. Morimoto, M. Murase, Y. Yamada, J. Yao, S. Utsunomiya, S. Kajihara, M. Fukuda, M. Ikawa, T. Tabata, K. Takahashi, K. Aoshima, Y. Nihei, T. Nishioka, Y. Oda and K. Tanaka, *J. Proteome Res.*, 2014, **13**, 3846–3853.

25 H. P. Benton, D. M. Wong, S. A. Trauger and G. Siuzdak, *Anal. Chem.*, 2008, **80**, 6382–6389.

26 M. Strohalm, D. Kavan, P. Novák, M. Volný and V. Havlíček, *Anal. Chem.*, 2010, **82**, 4648–4651.

27 T. Pluskal, S. Castillo, A. Villar-Briones and M. Orešič, *BMC Bioinf.*, 2010, **11**, 395.

28 M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert and O. Kohlbacher, *BMC Bioinf.*, 2008, **9**, 163.

29 U. W. Liebal, A. N. T. Phan, M. Sudhakar, K. Raman and L. M. Blank, *arXiv*, MDPI AG, 2020, preprint, arXiv:10060243, DOI: 10.3390/metabo10060243.

30 J. Gregorova, P. Vychytilova-Faltejskova, T. Kramarova, Z. Knechtova, M. Almasi, M. Stork, L. Pour, J. Kohoutek and S. Sevcikova, *Neoplasma*, 2022, **69**, 412–424.

31 C. H. Lin, H. Su, C. C. Hung, H. Y. Lane and J. Shiea, *Molecules*, 2021, **26**(15), 4457–4472.

32 T. Růžičková, M. Vlachová, L. Pečinka, M. Brychtová, M. Večeřa, L. Radová, S. Ševčíková, M. Jarošová, J. Havel, L. Pour and S. Ševčíková, *Cell Div.*, 2025, **20**, 4.

33 E. A. Thévenot, A. Roux, Y. Xu, E. Ezan and C. Junot, *J. Proteome Res.*, 2015, **14**, 3322–3335.

34 S. Wang, H. Zhu, H. Zhou, J. Cheng and H. Yang, *BMC Bioinf.*, 2020, **21**, 439.

35 E. S. Lee and T. J. S. Durant, *J. Mass Spectrom. Adv. Clin. Lab*, 2022, **23**, 1–6.

36 J. H. Harrison Jr, J. R. Gilbertson, M. G. Hanna, N. H. Olson, J. N. Seheult, J. M. Sorace and M. N. Stram, *Arch. Pathol. Lab. Med.*, 2021, **145**, 1228–1254.

37 O. A. Montesinos López, A. Montesinos López and J. Crossa, in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Springer International Publishing, 2022, pp. 109–139.

38 D. Wolrab, R. Jirásko, E. Cífková, M. Höring, D. Mei, M. Chocholoušková, O. Peterka, J. Idkowiak, T. Hrnčiarová, L. Kuchař, R. Ahrends, R. Brumarová, D. Friedecký, G. Vivo-Truyols, P. Škrha, J. Škrha, R. Kučera, B. Melichar, G. Liebisch, R. Burkhardt, M. R. Wenk, A. Cazenave-Gassiot, P. Karásek, I. Novotný, K. Greplová, R. Hrstka and M. Holčapek, *Nat. Commun.*, 2022, **13**, 124.