From data to insights: evaluating preprocessing and modelling approaches in (un)targeted metabolomics

Caterina Durante[1] & Eugenio Alladio[2]

[1] *Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, via Campi 103, Modena, 41125, Italy.* caterina.durante@unimore.it

[2] *Department of Chemistry, University of Turin, Via P. Giuria 5, 10125 Turin, Italy.* eugenio.alladio@unito.it

---

In recent years, the advancement of high-resolution mass spectrometry technologies has enabled the development of untargeted approaches for metabolomic analysis, which do not require prior knowledge of the compounds being investigated, leading to the discovery of unexpected or unidentified compounds. However, in the use of liquid chromatography-mass spectrometry (LC-MS) or hyphenate techniques in general, the signals produced are extremely rich in information but can also reach sizes up to gigabytes. For their analysis, it is necessary to reduce the file size and compress the information to avoid memory issues on computers with limited capacity. Compression must be carefully designed to prevent the loss of useful information, eliminating possible sources of error or spurious variability, and providing a set of data that can be effectively used for classification, calibration and other downstream applications. Furthermore, it is important to consider that the desired information is often hidden within the complexity of the data, and the detection and selection of relevant features can be particularly challenging.

In this seminar, based on targeted and untargeted analyses, LC-MS signals and results, obtained from the analysis of food and environmental samples, were preprocessed using different machine learning approaches, highlighting the strengths and weaknesses of the different procedures. Furthermore, despite their importance, the effective application of machine learning techniques often relies on programming skills that are rarely addressed in traditional chemistry education. To bridge this gap, we showcase SpectrApp, an open-source application built in Python and R Shiny, aimed at facilitating machine learning analysis for students, educators, and researchers without the need for coding expertise.

Short bio

Caterina Durante is an Associate Professor of Analytical Chemistry at the Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia. During her Ph.D., she began to interface with the development of analytical methods for food authentication, collaborating with different national and international research groups. Over time expanded her expertise toward the integration of machine learning techniques in analytical chemistry. Currently, her research focuses on the development of innovative untargeted and fingerprinting approaches, combining chromatographic techniques with mass spectrometry, as well as solvent-free spectroscopic methods that require minimal sample amounts, in accordance with the principles of green chemistry.

She is currently principal investigator of two research projects: '***Exwaster Project: Exploitation of targeted and untargeted analytical strategies for wastewater monitoring: toward a sustainable water management according to the principles of circular economy*'** PRIN 2022 PROJECTS- FUNDED BY THE EUROPEAN UNION – NEXT GENERATION EU and ***GreenTrace project : Integration of Eco-Sustainable Analytical Methods, Machine Learning, and Data Management for the Authenticity and Quality of Food*** supported by University of Modena and Reggio Emilia - Modena Foundation through FAR Mission Oriented 2024 funds.

Eugenio Alladio is an Associate Professor at the Department of Chemistry, University of Turin. He also serves as a collaborator at the "A. Bertinaria" Anti-Doping Center in Orbassano, Turin, where he specializes in developing applications of chemometrics and machine learning, validating analytical methodologies, and performing data analysis in forensic chemistry and biology. He earned his PhD in Chemistry from the University of Turin in 2017 and subsequently served as a Sub-Lieutenant at the Carabinieri Scientific Investigation Department (RIS) in Rome. In addition to his academic role, Alladio is the CEO of DataBloom Srl, a spin-off company focused on data analysis for forensic, clinical, industrial, and sports applications. He is currently principal investigator of the research project: '***BORDER - Bipolar disORDER: beyond psychology and towards multi-omics analysis****'* PRIN PNRR 2022 PROJECTS-FUNDED BY THE EUROPEAN UNION – NEXT GENERATION EU.