Faculty of Chemistry
University of Vienna
Guest Seminar

Name:	Michael D. Sorochan Armstrong
Institution:	Department of Signal Theory, Telematics and Communications, University of Granada
Title:	**Knowledge generation from heterogeneous Omics datasets**
Date:	5th June, 2025
Time:	10:00 CET
Location:	Lecture Hall 2

Modern analytical instruments are extremely sensitive, and can capture the chemical complexity of interesting biological samples in incredible detail. However these instruments are still relatively low-throughput, generating a situation in which datasets present a much greater number of features than observations. While there are a multitude of different mathematical techniques for dimensionality reduction that address this specific problem, the limited number of replicate samples available in most "Omics" studies may limit the statistical power of the findings. Traditional meta-analyses rely on previously proposed hypotheses for a limited number of features, which may not be consistent across multiple studies, and engaging with the peak table summaries themselves is difficult when no common modality (i.e. samples, or features) exists between the datasets.

General Linear Models (GLMs) and their multivariate extensions via ANOVA-Simultaneous Component Analysis + (ASCA+) are a mathematical generalization of ANOVA to unbalanced, and multivariate data respectively [1]. Through ASCA+, nominally orthogonal factorizations of individual datasets can be tested for significance as a function of the experimental characteristics of interest and their interactions. Common experimental characteristics also span a common subspace along the feature modalities, which can be exploited to analyze multiple datasets, although the underlying multivariate structure can still vary across different experiments.

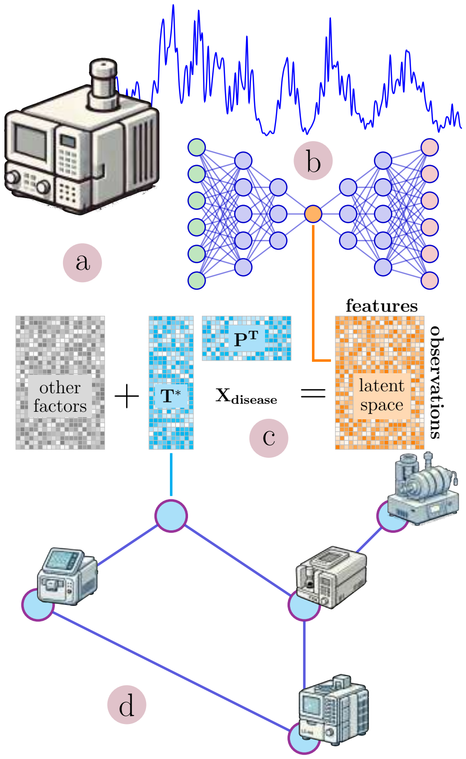

Figure 1: *Outline of an idealized workflow - data with experimental characteristics are collected in a), and are transformed through a "pre-processing" algorithm in b) which represents the raw data as a latent space of observations and chemical features. This latent space is factorized according to its experimental characteristics in c), before this representation of the data is analysed in the context of similar problems spanning different instruments in d).*

Inconsistent and often poorly transparent methods for feature extraction and integration across multiple samples are extremely common for hyphenated chromatographic-mass spectrometric datasets. This can be addressed by imputing missing values, but our knowledge of the underlying biological mechanisms for disease can nonetheless be improved through a better understanding of the instrumental data structure used as an intermediary in its representation. This talk will outline the steps necessary to create a workflow based on modern methods for statistical and machine learning that integrate domain-specific, state-of-the-art research in chemometrics (Figure 1).

[1] Thiel, Michel, Baptiste Féraud, and Bernadette Govaerts. "ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs." Journal of Chemometrics 31.6 (2017): e2895.