“Machine Learning in 21st Century Analytical Science”

Olatomiwa Bifarin, Facundo M. Fernández[1]

[1]School of Chemistry and Biochemistry and Petit Institute of Bioengineering and Bioscience, Georgia Institute of Technology.

Advances in analytical chemistry instrumentation and computational tools have now made it possible to examine biological processes in complex systems with an exquisite level of detail. This includes collections of analytes (or “omes”) with a given specific nature such as the transcriptome (nucleic acids), the proteome (proteins and peptides) and the metabolome (metabolites and lipids). Of these, the metabolome is the most sensitive to perturbations and interventions, yielding a molecular profile that is closest to the physiological phenotype. Metabolomic profiles are therefore sensitive to reprogramming observed in early disease stages and disease progression, which are more difficult to detect at the proteome or transcriptome levels.

The metabolome of a given organism is the total collection of biologically-active small molecules with molecular weights lower than about ~1.5 kDa. This includes endogenous molecules that are biosynthesized by metabolic networks in primary metabolism, specialized secondary metabolite signaling or defense molecules, molecules derived from diet or environmental exposures (the exposome), and molecules derived from the biosynthetic interactions with associated microbes (the microbiome). Metabolomics can either be “targeted” to a set of known compounds, for example certain lipids, or “non-targeted”, which attempts to detect and relatively quantify as many metabolites or lipids as possible.

While mapping the metabolome with high resolution separation techniques such as liquid chromatography, nuclear magnetic resonance spectroscopy and mass spectrometry, generation of highly dimensional and complex datasets is the norm, rather than the exception. Machine learning (ML) is therefore used in metabolomics to analyze and interpret these vast amounts of complex data, with tasks including classification, regression, clustering, dimensionality reduction and anomaly detection, among others. Despite its power, ML is not without limitations. These include Data Dependency (ML models heavily rely on the quality and quantity of data available for training. Biased or incomplete data can lead to inaccurate or unfair predictions), Interpretability (ome machine learning models, like deep neural networks, can be complex and challenging to interpret. This lack of transparency can be a drawback in fields where explainability is crucial, such as healthcare), and Overfitting (models that are overly complex or trained on limited data may memorize the training examples instead of learning generalizable patterns. This can lead to poor performance on unseen data).

In this talk, we will discuss the analytical tools and workflows involved in metabolomics, provide examples of using ML-enabled metabolomics to diagnose and estimate the prognosis of kidney and ovarian cancer, and discuss the pitfalls associated with using ML as a “black box”. We will also showcase recent results on how explainable ML is helping solve the interpretability issue in the context of metabolomics, and how natural language processing (NLP) is enabling computers to understand and interpret the vast knowledge in the field of metabolomics in a way that it is both meaningful and useful.