Contribution ID: **56** Type: **Oral contribution**

Toward Reliable Synthetic Omics: Statistical Distances for Generative Models Evaluation

Tuesday 8 July 2025 11:05 (30 minutes)

Synthetic data generation is emerging as an approach to overcome the limitations of real-world data scarcity in omics studies, especially in precision medicine and oncology. Omics datasets, with their high dimensionality and relatively small sample sizes, often lead to overfitting, especially in deep learning models. Generative models offer a promising way to generate realistic synthetic data preserving the original data distribution. However, there is still no objective consensus on how to evaluate their performance. In this talk, we set out to validate generative networks for transcriptomics data generation by using statistical distances as robust evaluation metrics. In particular, we observe that statistical distances enable simultaneous evaluation of global and local data fidelity of generated synthetic data. Because these distances satisfy the properties of true metrics, they also enable formal hypothesis testing to assess whether generative models have in fact converged or are merely approaching the reference distribution. Crucially, optimizing for these distances was found to implicitly select models maximizing other widely used metrics of generative performance, providing evidence of their broad applicability. Overall, our findings indicate that the adoption of these metrics can play a key role in guiding the development of generative models across a wide range of domains.

Author: JURMAN, Giuseppe (Fondazione Bruno Kessler & Humanitas University)

Presenter: JURMAN, Giuseppe (Fondazione Bruno Kessler & Humanitas University)