P-1.190

Pitfalls of Machine Learning in Single Cell Data Analysis at Single Nucleotide Resolution

Fabio Lauria [1], Gabriele Tomè [1], Ilaria Bruno [2], Gabriella Viero [1]

[1] Institute of Biophysics, Trento, Italy

[2] Immagina Biotechnology S.r.l., Trento, Italy

AI has transformed biology and medicine by analyzing complex data, predicting biomarkers, and identifying treatments. AI techniques, like deep learning, are improving single-cell data analysis, revealing gene regulatory networks and rare cell types. However, challenges remain in using AI for single-nucleotide resolution in sequencing data. Factors like data quality, sample size, and biases in training datasets impact AI reliability. This study highlights the ongoing challenges in effectively applying AI in biology. To assess the performance of AI approaches for investigating translation dynamics at single-cell resolution, we tested on the same dataset riboWaltz, a well-established tool for bulk ribosome profiling (RiboSeq), with a machine learning-based random forest (RF) approach, developed to face high noise. RF failed to capture the expected RiboSeq positional features in single-cell data, revealing inaccuracies in ribosome localization. This issue persisted in bulk RiboSeq with ribosome-stalling drugs, where riboWaltz delivered more precise results. In low signal-to-noise conditions, RF further distorted the data, overcorrecting the signal and compromising the interpretation of translation dynamics. AI holds great potential, but requires rigorous validation to ensure reliable and accurate biological interpretations.