Silesian University of Technology
Faculty of Automatic Control, Electronics
and Computer Science

Skipping batch effect correction: clustering-based methods for analyzing confounded single-cell RNA-sequencing data

PhD Thesis

Author: **Tomasz Kujawa**
Supervisor: **prof. dr hab. inż. Joanna Polańska**
Co-supervisor: **dr inż. Michał Marczyk**

Gliwice, June 2023

ABSTRACT

Single-cell RNAseq experiments are often conducted on a large scale, involving multiple laboratories or measurements taken at different times. Perfectly balanced experimental designs for such large projects may be infeasible, resulting in the need to conduct experiments in batches. Consequently, batch effects inevitably arise. Batch effects introduce variation that is unrelated to the biological variability under investigation, thereby obscuring it. If left unaddressed, batch effects can result in misleading conclusions drawn from the analysis. Therefore, batch effects have to be computationally corrected or removed.

Although batch effects have a detrimental impact on the data, the process of correction for them can also be harmful, particularly at the gene-level. There are several downsides to batch correction, including the lack of a measure to quantify the uncertainty associated with the correction process, requiring caution in the application of correction tools. Existing algorithms often prioritize achieving complete mixing of cells between batches rather than preserving the underlying population structure.

This work aims to provide a pipeline that utilizes iterative subspace clustering, combined with functional analysis of gene sets, to mitigate the negative impact of the batch effect on scRNAseq data. The crucial aspect of the functional analysis involves identifying cluster-specific pathways and establishing their linkage between batches. Therefore, the proposed workflow eliminates the need for applying batch-effect correction and enables consolidated analysis of batches that were generated separately.