ETH zürich

medical data science

BOEVA LAB

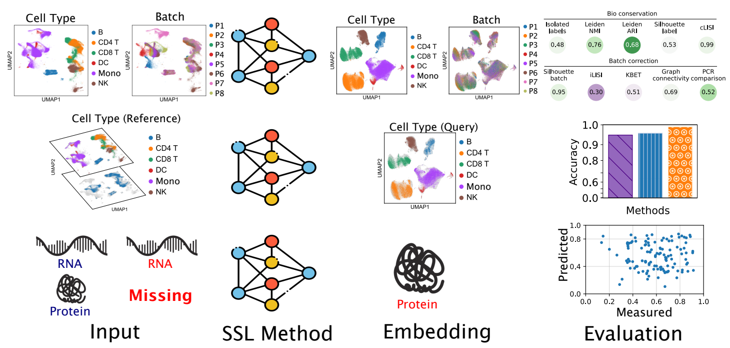
Benchmarking Self-Supervised Learning for Single-Cell Data

Philip Toma[1,†], Olga Ovcharenko[1,†], Imant Daunhawer[1],
Julia Vogt[1], Florian Barkmann[1,‡], Valentina Boeva[1,2,3,4,‡]

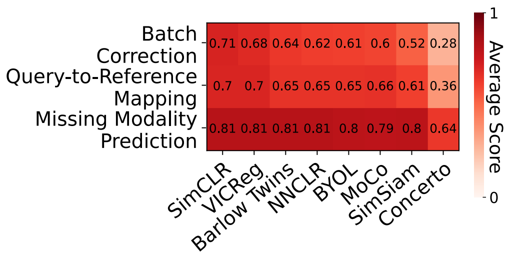
1: Institute for Machine Learning, Department of Computer Science, ETH Zürich, Zürich, Switzerland; 2: ETH AI Center, ETH Zürich, Zürich, Switzerland;
3: Swiss Institute for Bioinformatics (SIB), Lausanne, Switzerland; 4: Cochin Institute, INSERM U1016, CNRS UMR 8104, Paris Descartes University, Paris, France;
†‡: equal contribution

Self-supervised learning (SSL) has emerged as a powerful approach for learning **biologically meaningful** representations of single-cell data. To establish best practices in this domain, we present a comprehensive benchmark evaluating eight SSL methods across **three downstream tasks and eight datasets**, with various data augmentation strategies. Our results demonstrate that **SimCLR and VICReg consistently outperform other methods** across different tasks. Furthermore, we identify random masking as the most effective augmentation technique. This benchmark provides valuable insights into the **application of SSL to single-cell data analysis**, bridging the gap between SSL and single-cell biology.

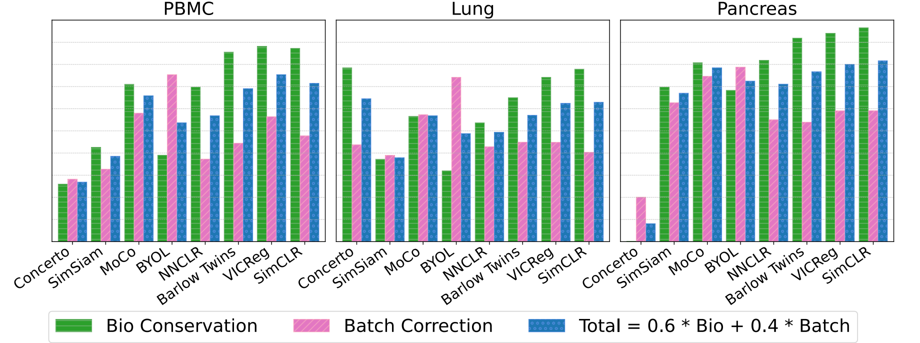
Evaluation on three downstream tasks.



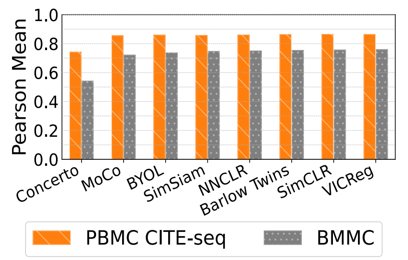
SimCLR and VICReg are the best-performing methods across all downstream tasks.



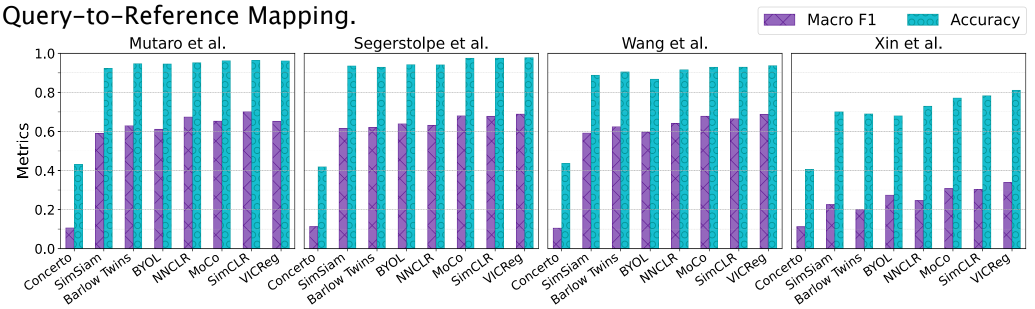
SimCLR and VICReg are the best methods for batch integration.



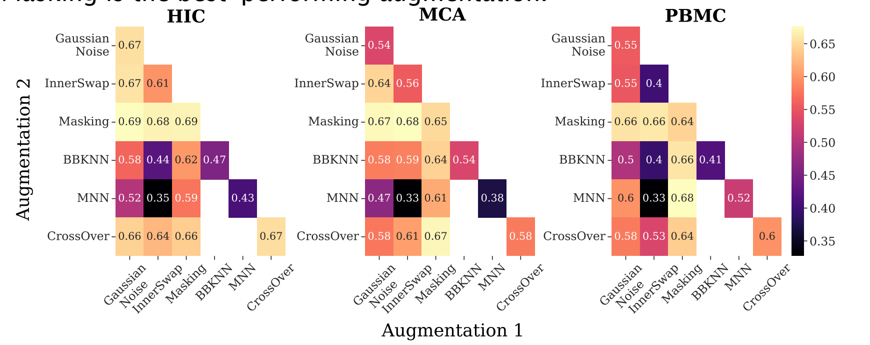
Missing Modality Prediction.



Query-to-Reference Mapping.



Masking is the best-performing augmentation.



In conclusion, SimCLR and VICReg emerge as the **top performing methods.** Masking augmentation proves to be the **most impactful augmentation.** We provide a benchmark to compare SSL methods on a new modality, enabling systematic **evaluation and advancement** of self-supervised learning methods for single-cell data.

Preprint

Poster

Code

NEURAL INFORMATION PROCESSING SYSTEMS

D INFK