scSSL-Bench: Benchmarking Self-Supervised Learning for Single-Cell Data

Olga Ovcharenko Florian Barkmann[ETH] Philip Toma[ETH]

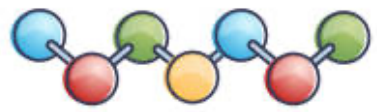
Imant Daunhawer[ETH] Julia Vogt[ETH] Sebastian Schelter Valentina Boeva[ETH]

BIFOLD

berlin

ETH zürich

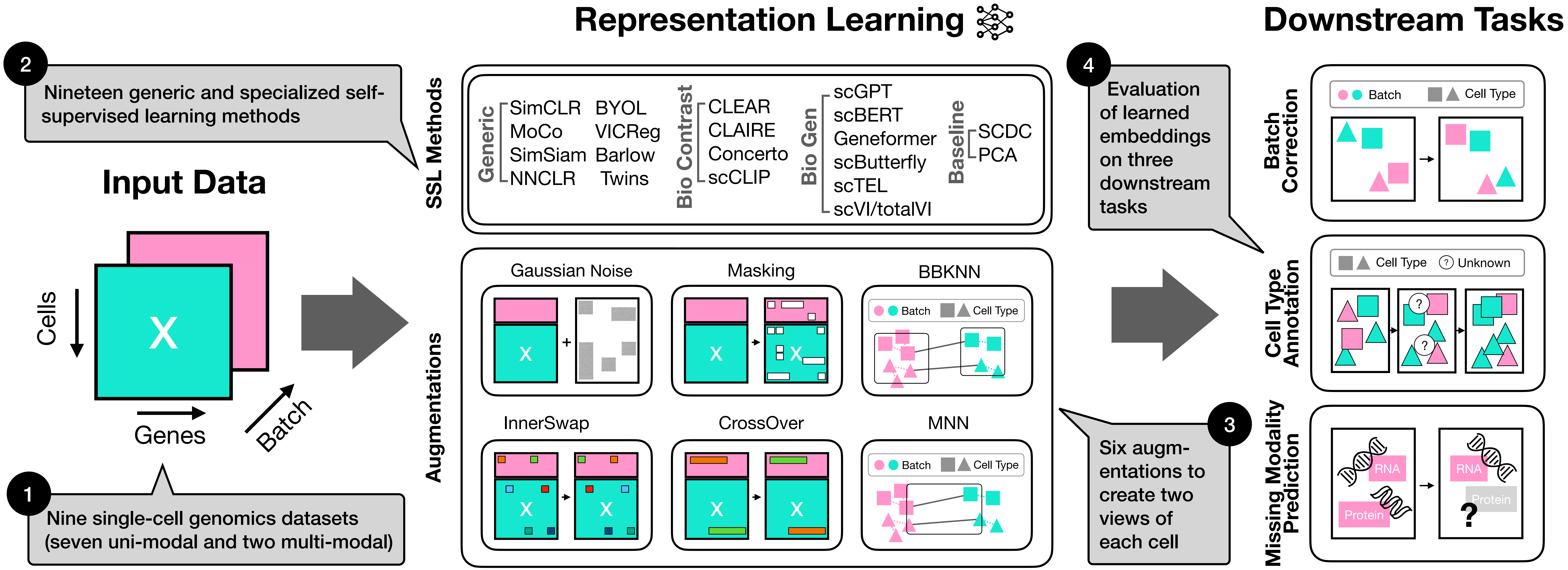
Motivation

- **Why** is genomics analysis an important research direction?
  - Profile cells at different resolutions and modalities
  - Understand diseases, develop personalized treatments, trace origins of conditions like cancer and autoimmune disorders
- **What** is single-cell data?
  - High-dimensional gene expression (**GEX**) levels in individual cells
  - Additional molecular features measured together with **GEX**
    - **Protein** levels + 
    - Open **chromatin** accessibility +

Why is a Benchmark Needed?

- Self-supervised learning (SSL) is a powerful approach for the representation extraction from single-cell data
- **Research questions**:
  - Benchmark if **specialized** single-cell SSL methods outperform **generic** methods
  - Assess **hyperparameters** and **augmentation** techniques of generic SSL approaches
  - Evaluate if genomics benefit from techniques proposed for images

scSSL-Bench Design



Input Data

1 Nine single-cell genomics datasets (seven uni-modal and two multi-modal)

2 Nineteen generic and specialized self-supervised learning methods

Representation Learning

SSL Methods

Generic: SimCLR, MoCo, SimSiam, NNCLR, BYOL, VICReg, Barlow Twins

Bio Contrast: CLEAR, CLAIRE, Concerto, scCLIP

Bio Gen: scGPT, scBERT, Geneformer, scButterfly, scTEL, scVI/totalVI

Baseline: SCDC, PCA

Augmentations

Gaussian Noise, Masking, BBKNN, InnerSwap, CrossOver, MNN

3 Six augmentations to create two views of each cell

Downstream Tasks

4 Evaluation of learned embeddings on three downstream tasks

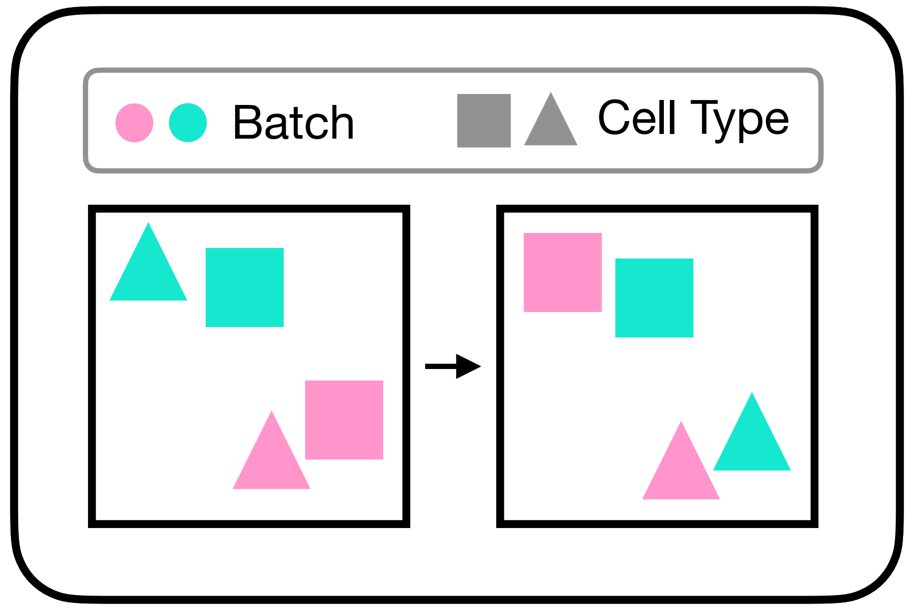
Batch Correction

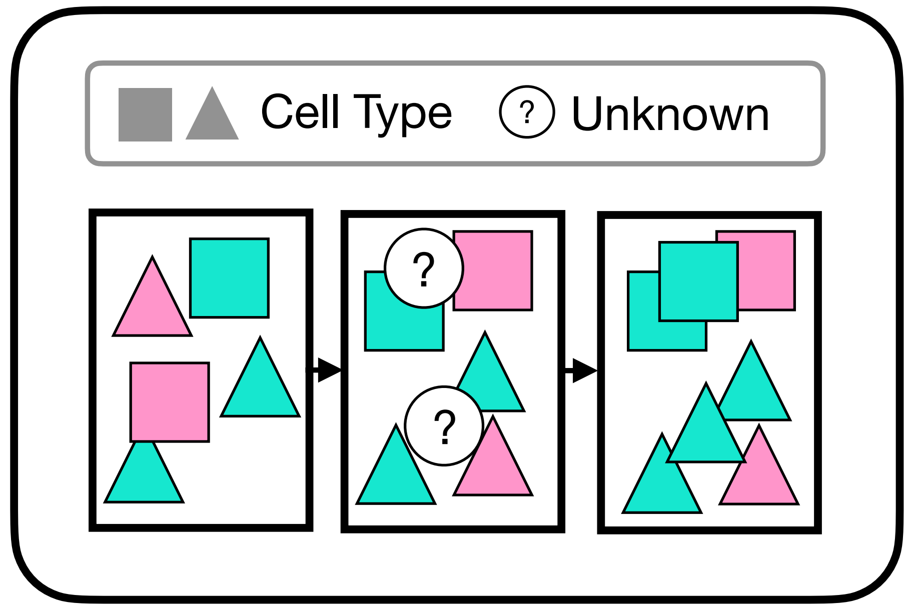
Cell Type Annotation

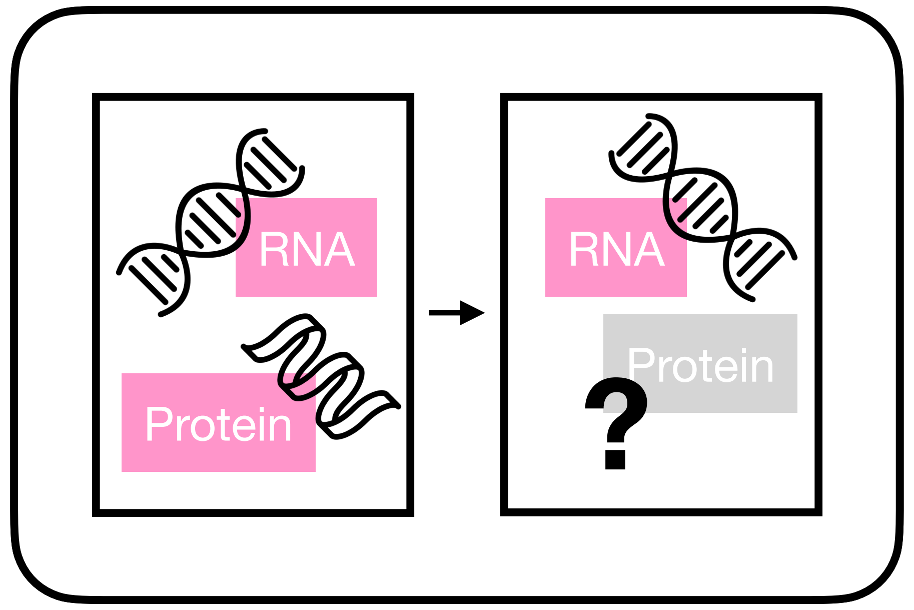
Missing Modality Prediction

Downstream Tasks on Learned Cell Representations

- **Batch effect correction**
  - Reduce technical biases introduced while sequencing
  - Preserve true biological signal
- **Cell typing**
  - Annotate cells of hold-out dataset by mapping them to seen cells
- **Missing modality prediction**
  - Infer unmeasured modalities for the hold-out dataset







Conclusions

- **Specialized** single-cell SSL methods work better for **uni-modal** data
  - scVI, CLAIRE, fine-tuned scGPT
- **Generic** SSL methods succeed in **multi-modal** single-cell data integration
  - SimCLR, VICReg
- **Masking** is the best augmentation technique
- **Moderately-sized embeddings** lead to better results

scSSL-Bench: Benchmarking Self-Supervised Learning for Single-Cell Data

Code

Paper

Contact Me

BIFOLD

TU berlin

ETH zürich