ai4bmr-datasets: A unified interface for spatial omics data access for computer vision and machine learning

2025-05-28

Summary

ai4bmr-datasets is an open-source Python package that provides a harmonized and standardized interface for accessing spatial omics datasets, including imaging mass cytometry (IMC) and multiplexed ion beam imaging (MIBI) data. The package enables researchers to load raw spatially-resolved omics data from multiple studies in a unified format, apply and retrieve data structures ready for downstream analysis or model training. By focusing on open-source raw data processing and enforcing common data schemas (e.g., standardized image and single-cell data formats), `ai4bmr-datasets` promotes reproducible and efficient research in computational and spatial biology. The library is currently used internally within our group, but is designed to serve the broader community working on spatial omics by easing data access and integration into machine learning workflows.

Statement of Need

Spatially-resolved omics technologies generate high-dimensional datasets with complex formats that often vary greatly between studies. In emerging fields like spatial proteomics (e.g., IMC) and spatial transcriptomics, researchers face a lack of unified frameworks to access and process heterogeneous datasets, hindering reproducibility and cross-study analyses. Each dataset typically requires custom scripts to handle its specific data structure, which creates barriers to entry, complicates usage and hinders robust benchmarking. `ai4bmr-datasets` addresses this gap by providing a single interface to multiple well-known spatial omics datasets, abstracting away dataset-specific idiosyncrasies. This unified approach allows scientists to focus on biological and analytical questions rather than data wrangling, and it supports consistent preprocessing pipelines across different studies that people can easily adapt and extend by providing predefined interfaces for common processing steps.

The importance of data standardization in spatial omics has been highlighted by recent efforts in the community (e.g., developing standards for sharing spatial transcriptomics data (K. C. Jackson and Pachter 2023)), underscoring the need for tools like `ai4bmr-datasets` that facilitate data sharing, reproducibility, and comparative analysis. By harmonizing data access, our package enables more straightforward integration of spatial omics data into machine learning and statistical modeling frameworks, ultimately accelerating biomedical discovery.

Supported Datasets

The package supports the following public spatial omics datasets: # TODO: add stats

- **Keren et al. 2018** – IMC of triple-negative breast cancer (Keren et al. 2018)
- **Jackson et al. 2023** – IMC of breast cancer (H. W. Jackson et al. 2020)
- **Danenberg et al. 2022** – IMC of breast cancer (Danenberg et al. 2022)
- **Cords et al. 2024** – IMC of NSCLC (Cords et al. 2024)

Additionally, dummy datasets are provided to mimic real data structure for development and testing purposes.

Each dataset is accessible through a standardized class interface that mimics the lightning philosophy and includes methods for downloading, preparing, and accessing processed components (images, masks, metadata, and spatial coordinates). These datasets follow consistent naming conventions and data schemas, making them immediately usable for downstream tasks.

Example Usage

```python
from ai4bmr_datasets import Keren2018
from pathlib import Path

dataset = Keren2018(base_dir=Path("<PATH>"))
dataset.prepare_data()
dataset.setup(image_version="v1", mask_version="v1")

print(dataset.images)  # list of images
print(dataset.masks)  # list of masks
print(dataset.intensity.shape)  # cell x marker matrix
print(dataset.metadata.shape)  # cell x annotation matrix
```

Conclusion

`ai4bmr-datasets` lowers the technical barrier to working with spatial omics data by providing unified, open access to several published datasets and processing routines. Its modular design and standardized outputs make it a practical tool for researchers developing computational methods in spatial biology. We welcome contributions and extensions from the community and envision this package as a foundation for reproducible spatial omics analysis.

References

Cords, Lena, Stefanie Engler, Martina Haberecker, Jan Hendrik Rüschoff, Holger Moch, Natalie De Souza, and Bernd Bodenmiller. 2024. "Cancer-Associated Fibroblast Phenotypes Are Associated with Patient Outcome in Non-Small Cell Lung Cancer." *Cancer Cell* 42 (3): 396–412.e5. https://doi.org/10.1016/j.ccell.2023.12.021.

Danenberg, Esther, Helen Bardwell, Vito R. T. Zanotelli, Elena Provenzano, Suet Feung Chin, Oscar M. Rueda, Andrew Green, et al. 2022. "Breast Tumor Microenvironment Structures Are Associated with Genomic Features and Clinical Outcome." *Nature Genetics* 54 (5): 660–69. https://doi.org/10.1038/s41588-022-01041-y.

Jackson, Hartland W., Jana R. Fischer, Vito R. T. Zanotelli, H. Raza Ali, Robert Mechera, Savas D. Soysal, Holger Moch, et al. 2020. "The Single-Cell Pathology Landscape of Breast Cancer." *Nature* 578 (7796): 615–20. https://doi.org/10.1038/s41586-019-1876-x.

Jackson, Kayla C., and Lior Pachter. 2023. "A Standard for Sharing Spatial Transcriptomics Data." *Cell Genomics* 3 (8): 100374. https://doi.org/10.1016/j.xgen.2023.100374.

Keren, Leeat, Diana Marquez, Marc Bosse, Roshan Angoshtari, Samir Jain, Sushama Varma, Soo Ryum Yang, et al. 2018. "A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging." *Cell* 174 (6): 1373–1387.e19. https://doi.org/10.1016/j.cell.2018.08.039.