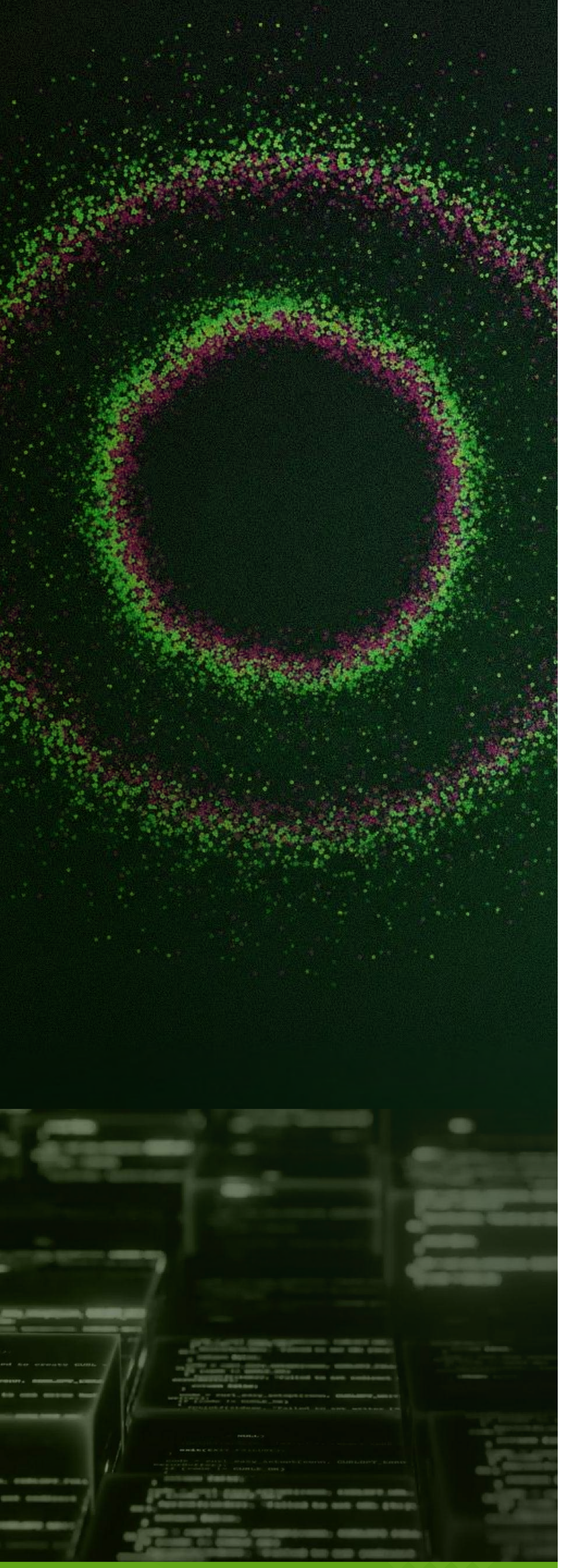
PROTOplast™

ACCELERATE YOUR SCRNA ML TRAINING.



ACCELERATE INSIGHTS WITH US

Contact us today to discuss your project at solutions@dataxight.com.

WWW.DATAXIGHT.COM

Accelerating Scalable ML for Single Cell Data Analysis

PROTOplast is an open-source Python library developed by DataXight, designed to significantly accelerate machine learning model training on large datasets such as **Tahoe-100M.**

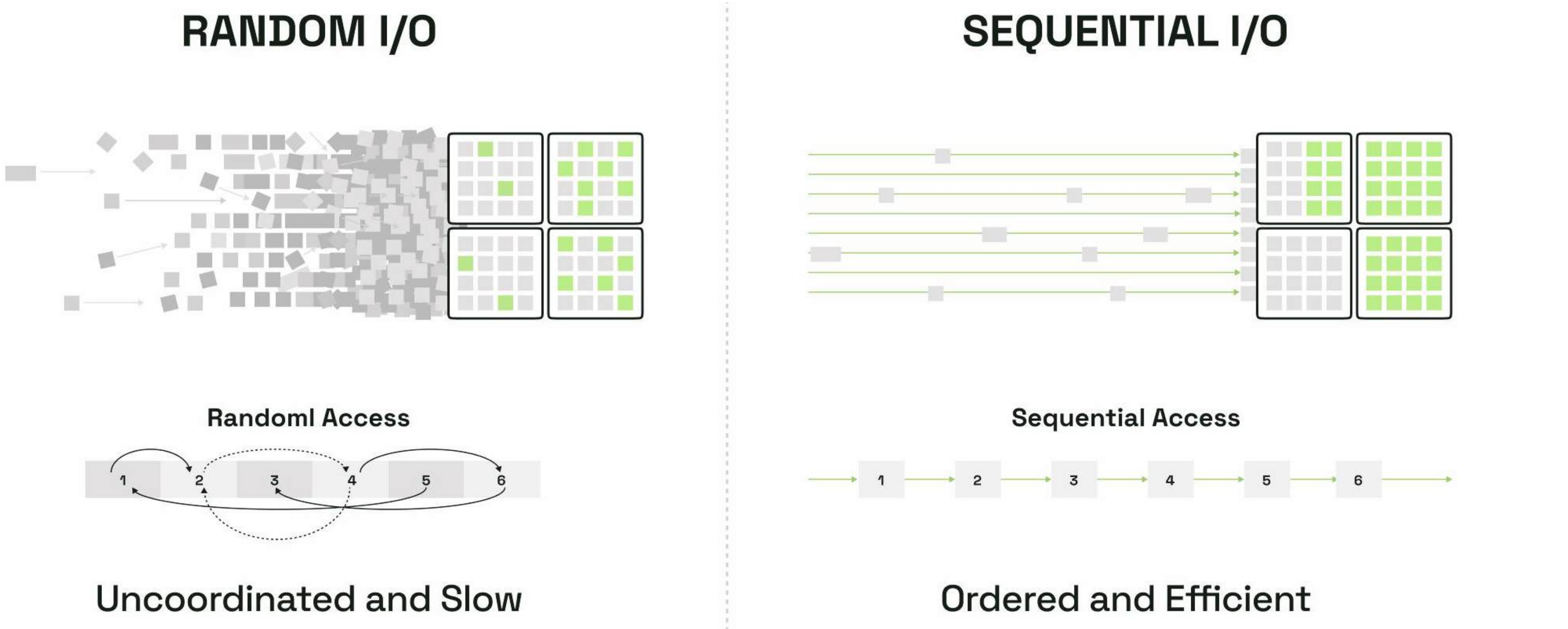


Figure 1: Illustration of data I/O bottlenecks in large single-cell pipelines — a key challenge that limits GPU utilization and scalability.

CHALLENGES

The Problems We're Solving.

{ 1 } DATA MANAGEMENT

Staging data adds overhead

Anndata reads from local file paths, requiring data to be copied to the compute instance prior to analysis

Loading data is time consuming

Large scRNA datasets remain slow to load—often hours to days—even on cloud or HPC systems.

Densification is costly

Sparse matrices, which optimize the amount of storage for scRNA datasets, require densification

{ 2 } SCALABILITY

Memory is constrained

Bottlenecks occur when the size of the data exceeds the amount of physical memory available on a machine

Cluster management is complex

Managing distributed workloads across multiple workers requires specialized expertise

Code environments are fragmented

Rewriting entire analysis pipelines is often necessary when scaling to cluster environments.

HOW DOES PROTOplast HELP?

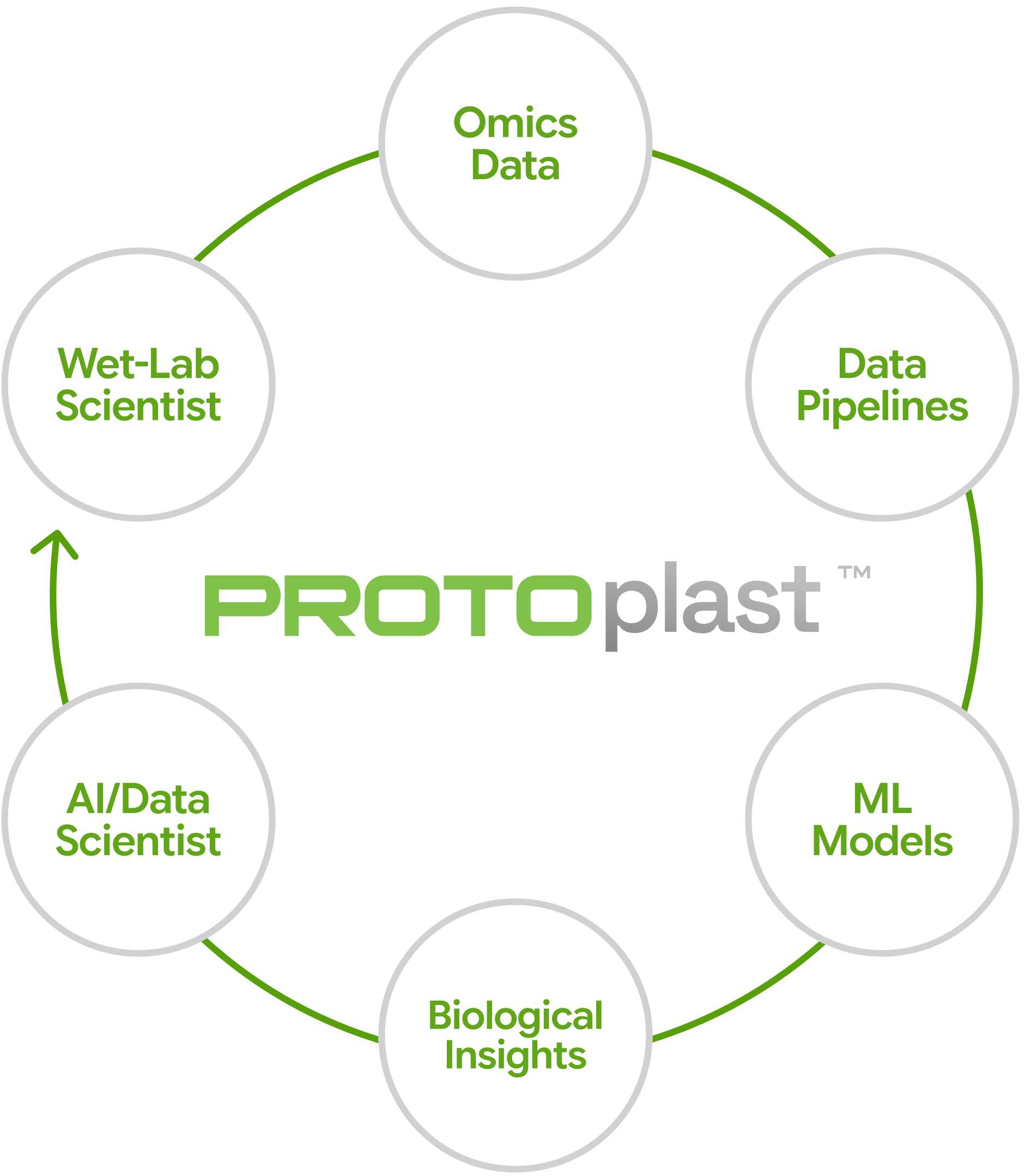


Figure 2: Continuous feedback between wet-lab and data scientists can shorten the time from data generation to actionable insight when supported by robust computational infrastructure and the necessary data-science skills; without those, the loop remains slow and error-prone.

PROTOplast transforms large-scale Machine Learning *(ML)* training, eliminating what was once a significant bottleneck and making it a routine, manageable part of the workflow.

This breakthrough is unprecedented, allowing researchers to bypass a typical three-week waiting time. Those who adopt **PROTOplast** can now start their research in mere minutes, gaining the luxury to focus on their core work.

It was built to remove these bottlenecks.

{ 1 } 1300X FASTER I/O

1300X faster I/O than standard AnnData: Training one epoch on the full **Tahoe-100M dataset**, which previously took 22.5 days using AnnLoader (AnnData), now takes only 14.5 minutes with **PROTOplast** using a 4-L40S instance.

WORKFLOW	ELAPSED	# OF WORKERS
AnnData	**22.5 days**	12
PROTOplast	**14.5 minutes**	12

*The benchmark was timed on 1 epoch, 2 MLP classifier, 4 NVIDIA L40S GPUs.[1]

{ 2 } SEAMLESS INTEGRATION

Simply subclassing PyTorch Lightning's LightningModule maintains complete compatibility with the existing PyTorch ecosystem, providing you with the necessary flexibility to develop specialized models for your molecular and single-cell data.

```
from state.tx.models.embed_sum import EmbedSumPerturbationModel
from protoplast import RayTrainRunner
trainer = RayTrainRunner(
   EmbedSumPerturbationModel,
   ...
)
```

{ 3 } NATIVE CLOUD INTEGRATION

Eliminates the need for intermediate downloads altogether. Now you can stream data directly from remote storage (S3, GCS, Azure).

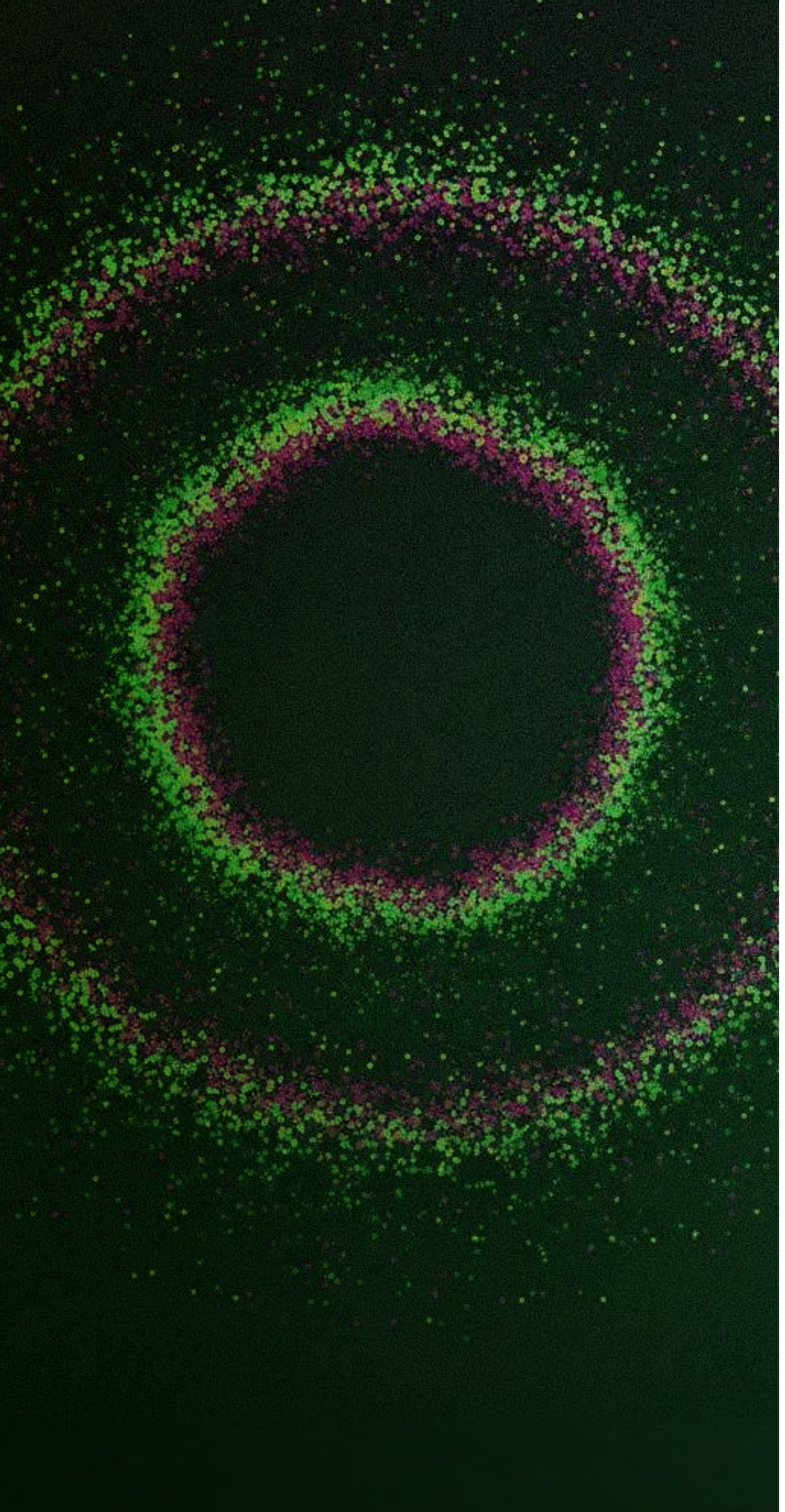
```
trainer.train([
   "s3://collaborator-1/cohort_1.h5ad",
   "gcs://collaborator-2/cohort_2.h5ad",
   "adl://collaborator-3/cohort_3.h5ad",
   "dnanexus://project-xxx:/cohort_4.h5ad",
], ...)
```

[1] dataxight.com/blogs/introducing-protoplast-scalable-machine-learning-for-molecular-data-analysis

DATAXIGHT®

PROTOplast™

ACCELERATE YOUR SCRNA ML TRAINING.



ACCELERATE INSIGHTS WITH US

Contact us today to discuss your project at solutions@dataxight.com.

WWW.DATAXIGHT.COM

Why Partnering With **DataXight** is The Strategic, Intelligent Decision.

Data-To-Insight Journey

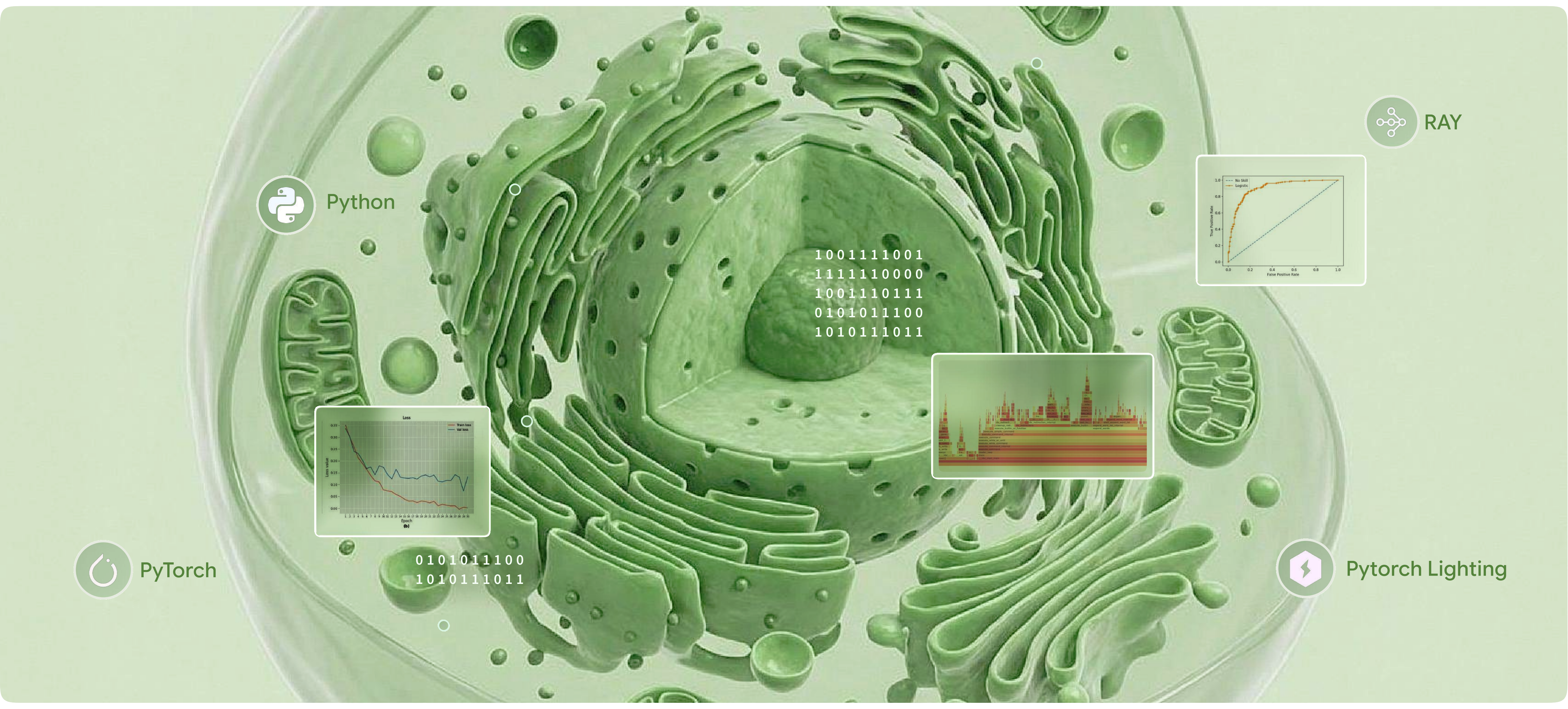
In solving the toughest challenges along the data-to-insight journey, DataXight offers you end-to-end, high-quality software services and solutions.

Comprehensive Solutions for Science & Healthcare

We cover the entire data journey with expertise in software engineering, AI/ML, and data science, complemented by deep industry knowledge. Our technical capabilities allow us to engage at any stage, from integrating complex systems and capturing experimental data to developing predictive models and delivering actionable insights.

Fit For Purpose

When we optimize development for your specific use, we emphasize quality and align with regulatory requirements. Deploying pipelines must be carefully and thoroughly scrutinized in order for us to continually perfect every step of each data journey, making each project an opportunity to ensure innovation, speed, and ease of use.



QUICK START

(IT'S SIMPLE)

Installation guide:

```
pip install protoplast
```

A minimal code that showcase end-to-end:

```
from protoplast import RayTrainRunner, DistributedCellLineAnnDataset, LinearClassifier
import glob

trainer = RayTrainRunner(
   LinearClassifier,  # replace with your own model
   DistributedCellLineAnnDataset,  # replace with your own Dataset
   ["num_genes", "num_classes"],  # change according to what you need for your model
)
trainer.train(
   file_paths=glob.glob("/data/tahoe100/*.h5ad"),
   batch_size=1024,
   test_size=0.0,
   val_size=0.0,
)
```

That's it — no extra code, no tuning. **PROTOplast** automatically scales across GPUs, nodes, or clusters.

RESOURCES

{ 1 } EXAMPLES

Training perturbation prediction models on scRNA-seq data.
Advancing precision in drug and gene response modeling

Handling datasets at the 100M+ cell scale.
Seamless integration with external and custom models

Create a submission to the Virtual Cell Challenge
Step-by-step guide to packaging and submitting your model for evaluation

{ 2 } GET STARTED

Documentation
Tutorials & Examples
Installation

Join our community

Github