

Accelerating Translational Research in Immune-Mediated Disorders With 5M Cells From Harmonized Datasets

Introduction

Elucidata partnered with a South San Francisco-based, pioneering translational research company dedicated to developing breakthroughs in the field of inflammatory diseases. The translational research company's unique approach involves targeting the underlying dysregulation that drives diseases.

The company wanted to **perturb gene circuits and manipulate the activity of specific transcription factors associated with regulatory T (Treg)-cell biology**. They wanted to achieve this through the application of network biology, knowledge graphs, and considerations of pre-clinical toxicity. To do so, they needed completely annotated single-cell and bulk RNA-seq data to attain robust predictive efficiency for the targets.

The company partnered with Elucidata to help build a corpus of high-quality public datasets that were relevant to their research.

Challenges

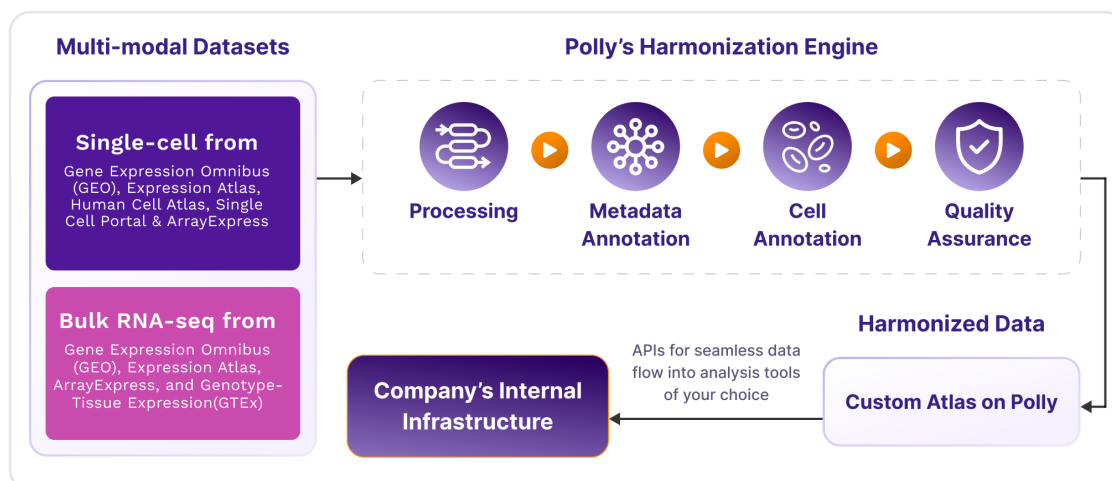
When researchers at the company started exploring public repositories to find and use datasets for their analyses, they hit a dead end because of the following challenges:

- 1. Findability:** Retrieving relevant public datasets was arduous and time-consuming as it required searching across multiple repositories.
- 2. Missing Metadata:** Once found, most public datasets were low quality and lacked well-annotated metadata. Finding and collating metadata from supplementary materials in publications and isolated sections in public repositories was a manual effort.
- 3. Missing Cell Type Annotations:** Most datasets had over **75% missing cell-type annotation information**, making it difficult for researchers to gain insights from these datasets. Manual cell-type annotation was labor-intensive, required specialized knowledge, and relied heavily on the usage of reference markers.
- 4. Integrations:** Getting public datasets in a ready-to-use format, fit for API integration into their in-house infrastructure required significant data preparation.

The Solution: Polly by Elucidata

Elucidata's biomedical data harmonization platform- Polly, helped the customer's bioinformatics team find public datasets of multiple modalities. It identified **~5 million cells and 11,000 multi-modal samples**, most relevant to their research. These datasets were then consistently processed, curated, and QC-ed on Polly and made available through APIs to enable seamless integration of harmonized data into the company's internal infrastructure.

Our solution used a three-step approach (shown below); let's dive right in!



Transforming Uncurated Multi-modal Datasets From Public Repositories Into ML-ready Formats

Retrieving Relevant Datasets

Polly by Elucidata, stores metadata from hundreds of thousands of public datasets spanning various repositories. This capability empowered us to execute advanced queries seamlessly across repositories such as Gene Expression Omnibus (GEO), Expression Atlas, Human Cell Atlas, Single Cell Portal, EMBL Expression Atlas, ArrayExpress, and GTEx—all within a fraction of the usual time. Through programmatic search, we were able to find **~5 million cells and 11,000 samples** relevant to immune-mediated disorders and **250,000 cells** relevant for preclinical toxicity studies.

Harmonization of Multi-modal Datasets

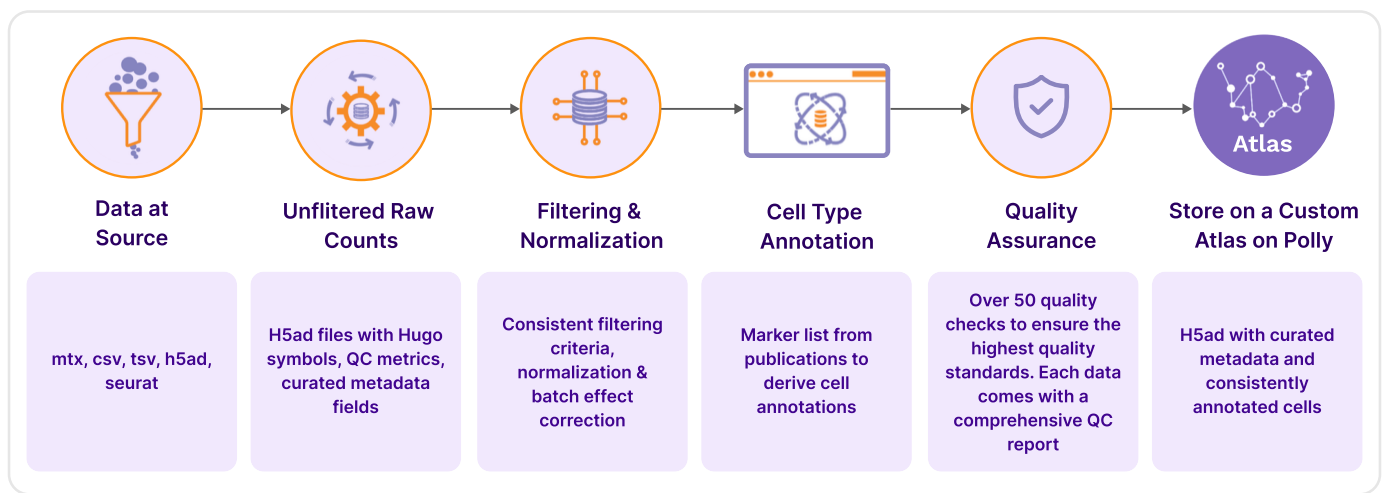
Polly harmonized these datasets, making them ML-ready. This was done by processing the raw data using author-prescribed pipelines, annotating metadata using consistent ontologies to ensure zero missing labels, and rigorous quality checks.

1. Data Processing

A. Single-cell Datasets

Raw, unfiltered single-cell RNA-seq datasets were ingested on Polly. Our single-cell processing pipeline performed QC checks to filter poor-quality cells and genes, data normalization for meaningful comparisons, feature selection, batch effect correction, and dimensionality reduction.

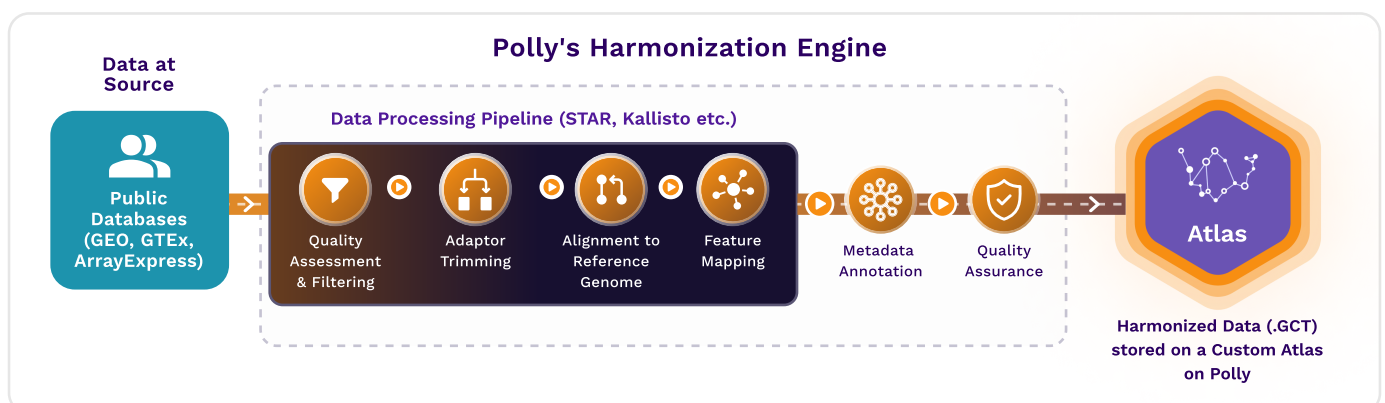
The next step was comprehensively annotating cell types for each dataset. Our single-cell processing pipeline used known marker genes and reference datasets, assigning each cell its respective biological identity to ensure thorough annotation. Special emphasis was placed on annotating Treg cells for pre-clinical toxicity studies (~250K cells) based on reference markers. By leveraging our state-of-the-art single-cell processing pipeline along with an additional layer of human-in-the-loop validation, we were able to guarantee the highest quality annotations in a quick turnaround time. This approach significantly expedited the customer's understanding of Tregs in toxicity studies and helped them identify targets for cells involved in immune-mediated disorders in record time!



Single-cell Processing and Cell-type Annotation

B. Bulk RNA-seq Datasets

Polly's harmonization engine processed over 10,000 bulk RNA-seq samples for the customer using our scientifically validated and optimized STAR/Kallisto pipeline. The pipeline first indexed samples with the reference human genome from the Ensembl for efficient alignment of raw sequencing reads. Subsequently, the sample's raw RNA-seq reads were aligned to the referenced human genome, ensuring accurate mapping of reads to the reference. This ensured gold-standard data quality, which was critical for the customer's research objectives.



Data Processing Pipeline

2. Metadata Annotation

Next, for every dataset, Polly's harmonization engine annotated datasets with metadata by mapping standard fields like disease, tissue, organism, cell line, cell type, and drug to their respective standard ontologies, MeSH, BRENDA Tissue Ontology, NCBI taxonomy, Cellosaurus, Cell Ontology, and PubChem. In addition to these fields, various other metadata fields relevant to the customer's use case were captured from the source publication and custom annotated at three levels:

- **Dataset-level metadata:** General information about the experiment, subject, and transformations. For example, organism, experiment type, PubMed ID, disease, dataset ID, dataset description, total number of samples in a dataset, etc.
- **Sample-level metadata:** Captures information for each sample. For example, drug, tissue, title of the sample, sample ID, type of treatment, etc.
- **Feature-level metadata:** Provides molecular/gene information that is consistent across samples. For example, feature ID, gene ID, ensemble ID, raw counts, total drop count, etc.

Multi-modal datasets with annotated metadata improved the findability of datasets, enabling powerful querying capabilities. This also facilitated the integration of harmonized datasets into the company's internal infrastructure for analysis.

A detailed overview of the dataset, sample, and feature-level metadata fields can be accessed [here](#).

3. Quality Assurance

Finally, each dataset delivered through Polly went through ~50 quality checks to ensure the highest quality standards. The raw counts for each dataset were inspected to ensure raw filtered and processed filtered counts were scientifically valid and distinct.

Our unique, tech-enabled human-in-the-loop approach ensured a balance between technological consistency and scalability and expert precision and quality. Additional quality checks confirmed the presence of QC metrics generated from the ScanPy pipeline, analysis details, UMAP or t-SNE projections (or both), and cluster information. Our QC checks also ensured the absence of batch effects and the annotation of each cell type according to its markers.

All processed multi-modal datasets delivered to the company were accompanied by a "Polly Verified" QC report to ensure complete transparency and visibility of data quality. A sample report can be accessed [here](#).



PMID_29988129

Polly verified : V1.0

Summary

Quality Control Metrics

Quality Assurance Checks

Data Exploration

QUALITY ASSURANCE CONTENT

1. Metadata information

2. Data matrix

3. Cell Clusters in umap Embeddings
Colored by Samples: Re-Processed and Polly Datasets

4. Cell Clusters in umap Embeddings
Colored by 'Author Cell Types':
Comparison Between Polly and Re-Processed Datasets

5. Cell Clusters in umap Embeddings
Colored by 'Curated Cell Types':
Comparison Between Polly Dataset and Re-Processed Data

6. Cell Type Frequency Distribution

1. Metadata information

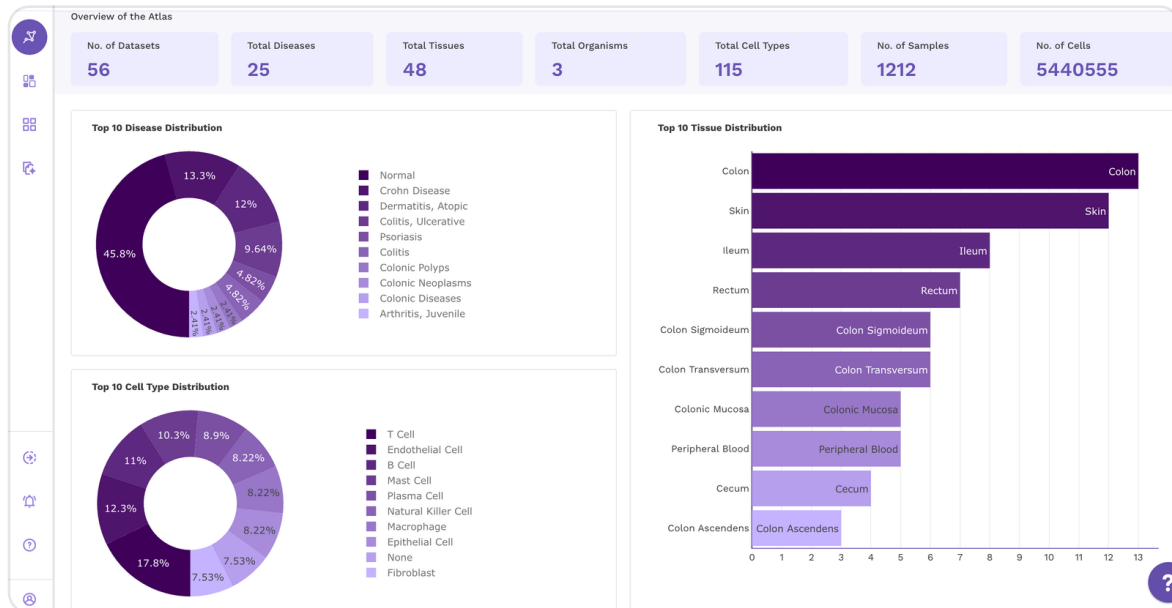
Metadata information	Value
Polly curated metadata fields are present at dataset level i	Pass
Polly curated metadata fields are present at sample level i	Pass
Polly curated metadata fields are present in output file i	Pass
Data Source Link is provided i	Pass
Data Source Link is valid i	Pass
Curated cell Line concordance: Dataset-Level vs. Sample-Level Metadata i	Pass
Curated tissue concordance: Dataset-Level vs. Sample-Level Metadata i	Pass
Curated disease concordance: Dataset-Level vs. Sample-Level Metadata i	Pass
Curated drug concordance: Dataset-Level vs. Sample-Level Metadata i	Pass
Curated cell-type concordance: Dataset-Level vs. Sample-Level Metadata i	Pass
Accuracy of raw counts availability tag i	Pass

Polly Verified Report for Single-cell Datasets

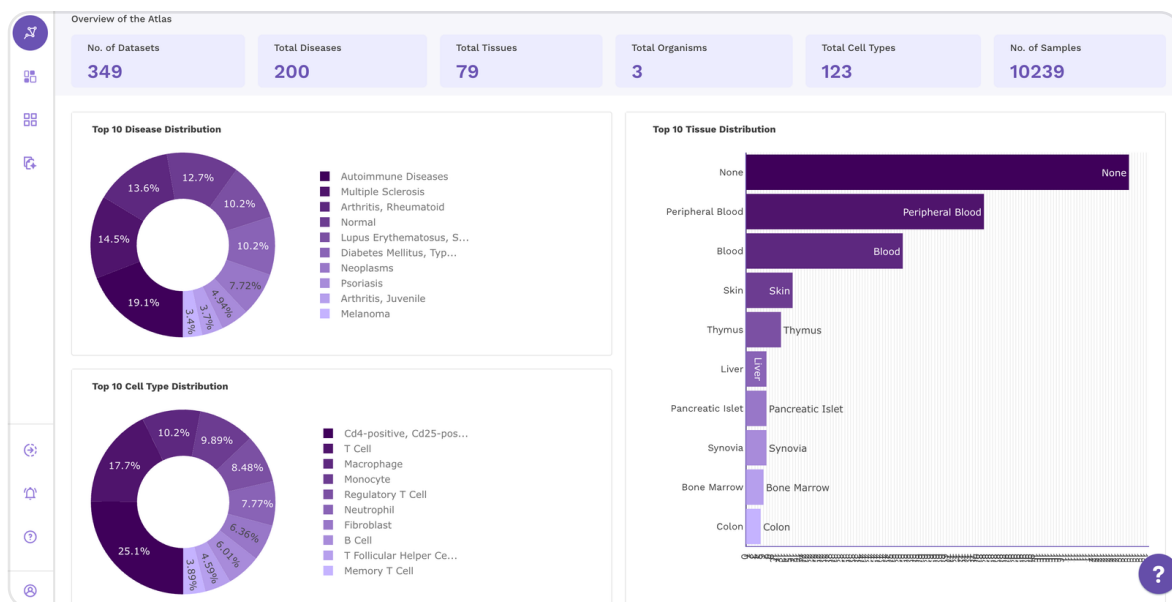
Programmatic Access To Harmonized Data

The customer's research team wanted programmatic access to facilitate the integration of harmonized datasets directly into their infrastructure, eliminating the need for downloading and uploading datasets. This method reduces the risk of data integrity loss and interoperability issues by reducing intermediate and manual steps.

Using Polly's programmatic access, the customer could efficiently ingest **100 harmonized datasets in 5 seconds, the time it typically takes to ingest a single dataset without Polly!** This automation not only expedited the entire ingestion process but also accelerated downstream analyses. Our APIs allowed the customer's bioinformatics teams the flexibility to download datasets without any restrictions and stream them to apps, dashboards, and analysis tools of their choice. All this while seamlessly integrating Polly's harmonized public datasets into their existing workflows!



Single-cell Datasets Delivered to the Company



Bulk RNA-seq Datasets Delivered to the Company

Outcome

This successful partnership over 1 year has helped the company:

1. Save ~1600 hours of data wrangling efforts:

- Finding **5 million cells** and **~11,000 relevant multi-modal samples** from public datasets and annotating metadata from the source would have taken **~1000 hours** without Polly.
- Curation and annotation of **over 5 million cells** with reference markers would take **~600 hours** without Polly.

2. Reduce internal resource requirements:

- Without Elucidata, the company would have hired **1 additional full-time bioinformatician and 1 full-time data engineer** to set up the infrastructure to process and automatically ingest public datasets into their internal infrastructure.

3. Expedite translation research milestones:

- Polly's high-quality datasets and comprehensive cell-type annotation based on reference markers helped accelerate the generation of knowledge graphs for transcription factors.
- Additionally, the customer was able to accelerate their target identification program with the help of Polly's curated pre-clinical toxicity studies.

About Elucidata

Polly by Elucidata optimizes data quality for pre-clinical drug discovery by harmonizing multi-omics and assay data into ML-ready formats. Polly's powerful harmonization engine is utilized by trained experts to curate diverse data types, annotate metadata, and ensure consistent processing at affordable costs. These ML-ready data are stored on an Atlas on the platform, ideal for analysis and management.

Polly's advanced technology accommodates 25+ data types, catering to teams in pre-clinical drug discovery and diagnostics R&D. Polly is trusted by over 25 research organizations, including 4 of the top 10 pharma companies, to accelerate their discovery programs.

[Book a Demo](#)

For more information, visit Elucidata | Home or reach out to us at info@elucidata.io.



San Francisco | 114 Sansome St, 250, San Francisco, CA 94104, USA

Cambridge | 625 Massachusetts Ave 2nd Floor, Cambridge, MA 02139, USA

Delhi | Avanta Business Center, Saket, New Delhi, Delhi 110017, India

Bangalore | UrbanVault HSR Layout 761, Bengaluru, Karnataka 560102, India