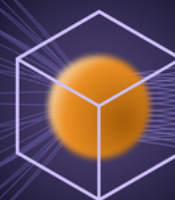


Achieved ~80% Acceleration in Gene Perturbation Target ID and Validation

Find out how an early-stage pharma company leveraged biomolecular data and our in-house expertise in curation, bioinformatics, and machine learning to study the effects of gene perturbation on cell fate and thereby arrive at 2 validated targets.



ABOUT THE COMPANY

An early-stage US-based pharmaceutical company was interested in studying the effects of gene perturbation on cell fate conversion in connection with their research on developing a novel treatment method for widespread health disorders.

THE CHALLENGE

The need was to identify regulatory switches which could facilitate the transformation of relevant cell types for organ repair from abundant, easy-to-obtain patient-derived cells. This required an understanding of the gene regulatory network (GRN) and identifying the transcription factors (TFs) which may play a pivotal role in deciding cell fate.

Given the need for specific scientific expertise to process, normalize and analyze the data at hand, as well as the paucity of resources at their disposal, they were seeking expert help.

REQUIREMENTS

01

To **curate datasets** comprising single cell, transcriptomics, proteomics, and metabolomics data to **process, analyze, and simulate** the effects of gene perturbation.

02

To **identify regulatory switches** (the relevant TFs) which could play a role in cell fate reprogramming from the curated datasets

03

To **analyze the effects** of these TFs in the Gene Regulatory Network (GRN)

04

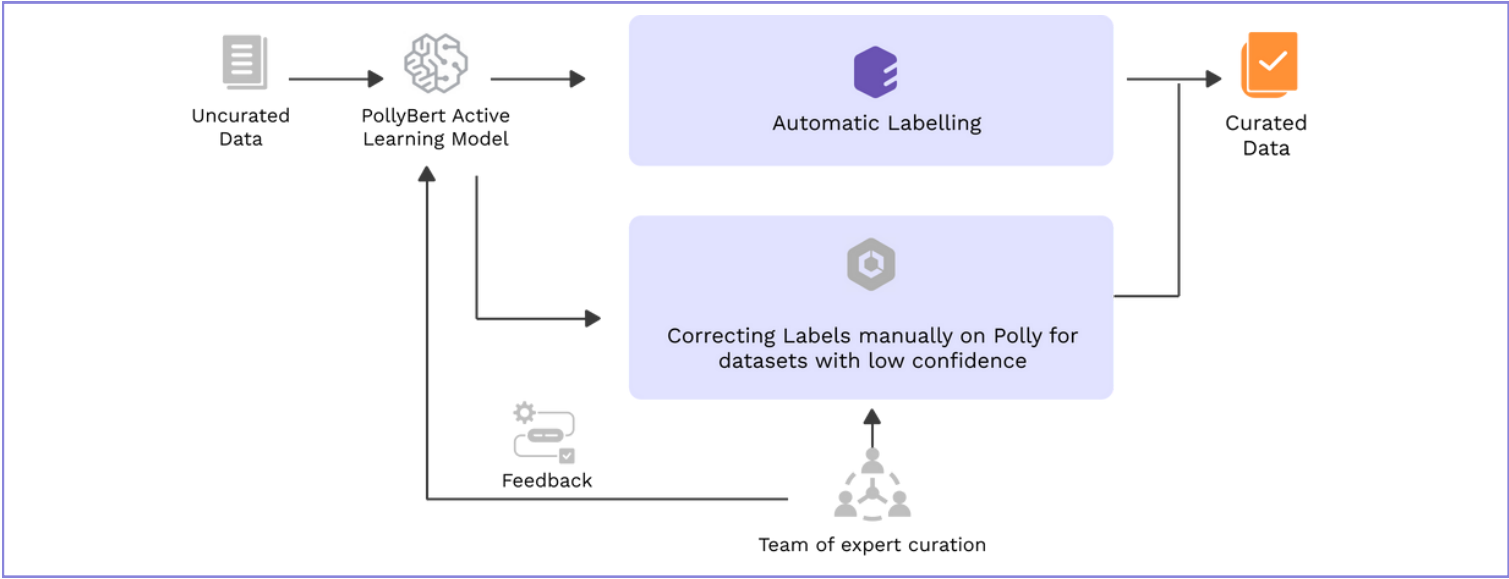
To **validate specific targets** which could decide cell fate reprogramming

OUR APPROACH

1.Data Curation

50+ datasets from the Gene Expression Omnibus (GEO) and various publications were curated into a ‘FAIR-Findable, Accessible, Interoperable, Reusable’ data resource relevant to their research areas of interest.

These datasets were annotated with harmonized metadata labels using a proprietary NLP-based model called PollyBERT. The model assigns a confidence score to each annotation, & datasets with a low score undergo an extensive quality check wherein they’re manually curated by a team of experts. The model learns from these expert annotations and improves its accuracy over time. This semi-automated curation workflow allows [Polly](#) to curate large volumes of datasets with speed and near-human accuracy.



An overview of our curation infrastructure

The datasets encapsulate sample & study level metadata fields, consistent ontologies & samples divided into perturbation & control.

Metadata Fields

Dataset Level Information

Disease

Publication

Tissue

Total no. of Cells

Cell Line

Total no. of Samples

Cell Type

Data Type

Organism

Dataset ID

Drug

Dataset Source

Description

Sample Level Information

Tissue

Cohort Name

Disease

Control

Cell Line

Cell Type

Drug

Type of Genetic Modification

Modification Gene

Age

Source Name

Title

Cohort ID

GEO Accession

Document ID

Additionally, the datasets were processed through a standard pipeline (identifier mapping, normalization & quality check) and made available in consistent tabular formats. Since these are single-cell datasets, the cell type annotations made it feasible to use this data for GRN analysis.

2. Data Processing Workflow

The effect of gene perturbations can be studied using GRN analysis which is carried out using the Cell Oracle. CellOracle is a python library useful for inferring insights by integrating single-cell RNA-Sequencing and epigenomics sequencing data using network analysis and in-silico perturbation simulations.

The workflow for CellOracle is composed of 5 steps:

- Processing ATAC-Seq/Epigenomics Data
- Inferring TF Binding information from ATAC-Seq data.
- Processing scRNA-Seq data
- Network Analysis
- In-silico perturbation identified targets.

3. Solutions Provided

- We collated and curated metadata fields of interest at the dataset level (13 fields) and sample level (15 fields) for all the 50 datasets to make them ready for downstream analyses.
- TF data extracted from the ATAC-Seq/ other epigenomics sequencing data was used to determine TF binding information.
- This processed data was used to create a CellOracle object which was used to study Gene Regulatory Networks of different cell types and further predicted targets of cell fate regulation were perturbed in silico to understand their relevance.

OUTCOMES

- Completed the custom curation process of **50** datasets in a brief period of **2 months**.
- Identified **2 targets** that could regulate cell fate reprogramming.
- Our established **curation infrastructure** and **data processing pipelines** increased the overall speed of analysis which helped them identify and validate the targets in **5-6** months instead of the typical time-period of more than a couple of years!

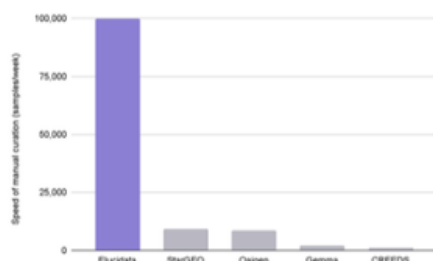
WHY ELUCIDATA

Our **tech-savvy team** of Bioinformaticians, Scientists, ML engineers, and Project managers are highly specialized in biomedical data analysis. Hence, they have the domain knowledge required to understand the problem in-depth, create the relevant data processing workflow and thereby contribute to insight derivation very effectively.

Speed

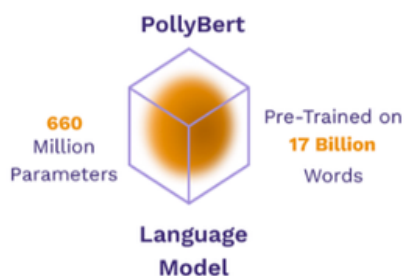
10x Faster than industry standards

ML-readiness throughput



Accuracy

As accurate as human curators on 4 tasks



32x

Growth in datasets on the platform. 5 consecutive quarters of doubling the number of datasets

10x

Datasets (>50,000/month) added to the platform as compared to industry standards

1.5x

Increase in the number of "curated" fields every quarter

ABOUT US

Polly delivers ML-ready biomolecular data that is curated to accelerate drug discovery. Hosting a rich repository of more than **1.5 million** ML-ready multi-omics datasets, Polly is a customizable data-centric platform that assists with the comprehensive analysis of integrated biomedical data. Take a look at our [website](#) to know more about how you can accelerate data-driven research.

For design to deployment support to build **customized, production-ready pipelines** for biomolecular analysis or for **on-demand curation** of public or proprietary data and metadata, get in touch with us [here](#) !