elucidata

Making Life Sciences Data AI - Ready

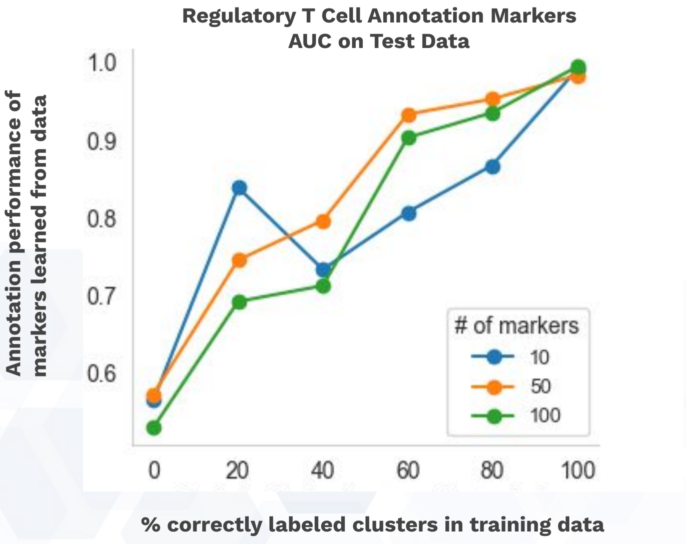
Pistoia Alliance: Collaborate to Innovate

Mya Steadman, Solutions Architect

November 2023

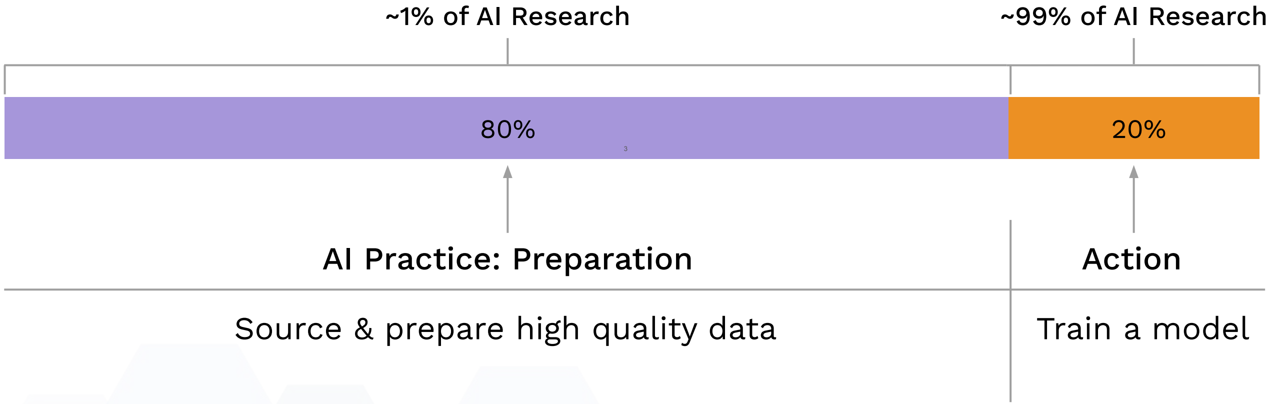
AI/ML Initiatives are Built on High Quality Data

Predicting cell annotations using marker genes derived from a corpus of 200 harmonized single cell samples



- An increase in the percentage of correctly labelled clusters in the training set improves model performance
- Biological signatures learned from harmonized single cell data will be directly impacted by quality of labels

Creating High Quality Datasets is not Trivial



~1% of AI Research

~99% of AI Research

80%

20%

AI Practice: Preparation

Source & prepare high quality data

Action

Train a model

Life sciences R&D teams could spend **~600 hours per quarter,** just on data preparation

Polly’s LLM-powered **Harmonization Engine** transforms messy biomedical data into **‘AI-Ready’ Data** lucidata

Case Study: Accelerating Data-Driven Target ID with Polly

About the Customer

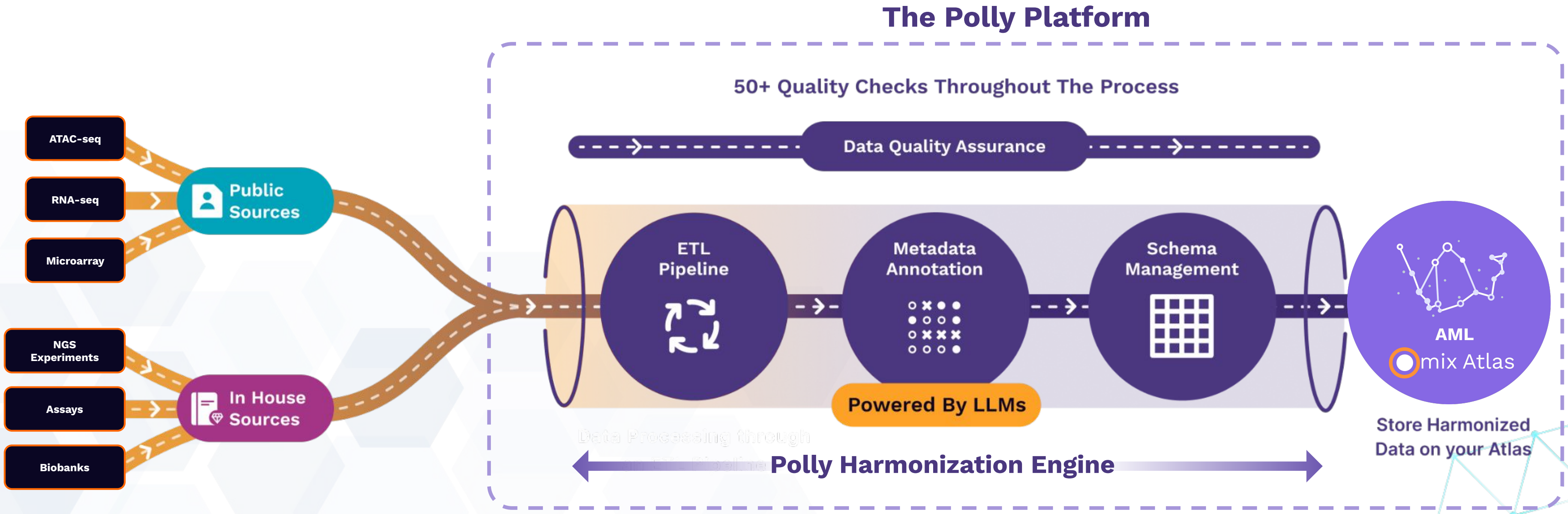
This Boston based pharmaceutical company wants to cure cancer by transforming malignant cells into healthy ones. Their goal is to identify and validate genes that act as potential differentiation based targets in Acute Myeloid Leukemia using multi-omics data.

Needs

- Curate a repository of clean, harmonized multi-omics datasets specific to AML
- Develop and train a patient classifier model to identify the right patient cohorts
- Shortlist target genes and validate using public literature
- Harmonize in-house and public data to common standards, host on the cloud

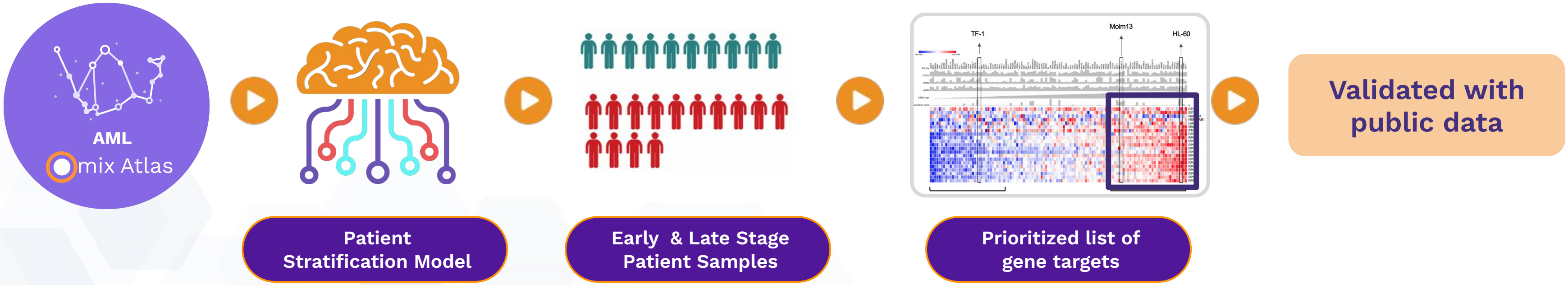
Building a Corpus of Harmonized Data with Polly

10k+ AML specific multi-omics datasets from public and in-house sources were processed and metadata-annotated with Polly's Harmonization Engine



Training Patient Classifiers with Harmonized AML Datasets

Harmonized datasets were used to train a patient stratification model, extract gene signatures from the right patient segments, and identify a list of possible gene targets



AML Omix Atlas

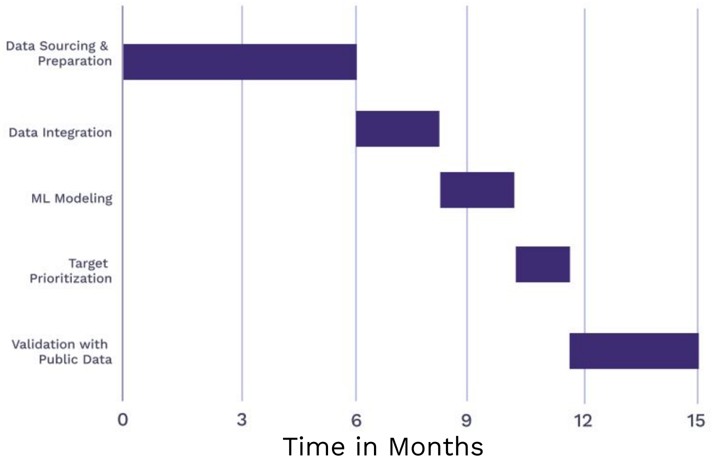
Patient Stratification Model

Early & Late Stage Patient Samples

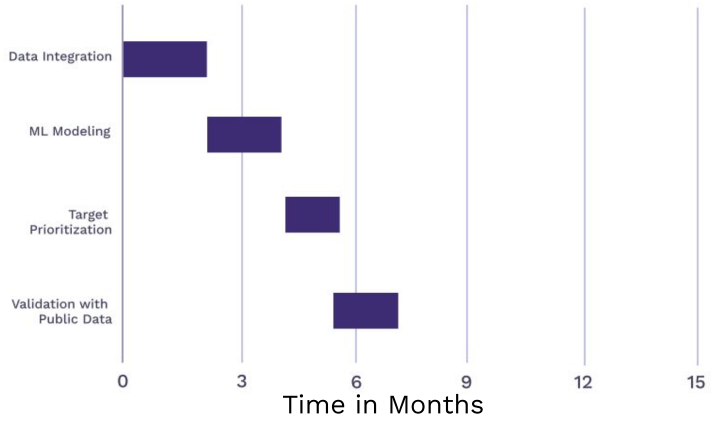
Prioritized list of gene targets

Validated with public data

Without Polly



With Polly



Impact

- **2+ Differentiation based targets** in AML identified using an integrative multi-omics approach

- **6 Months** to identify & validate targets vs the average **1-2 year** time frame

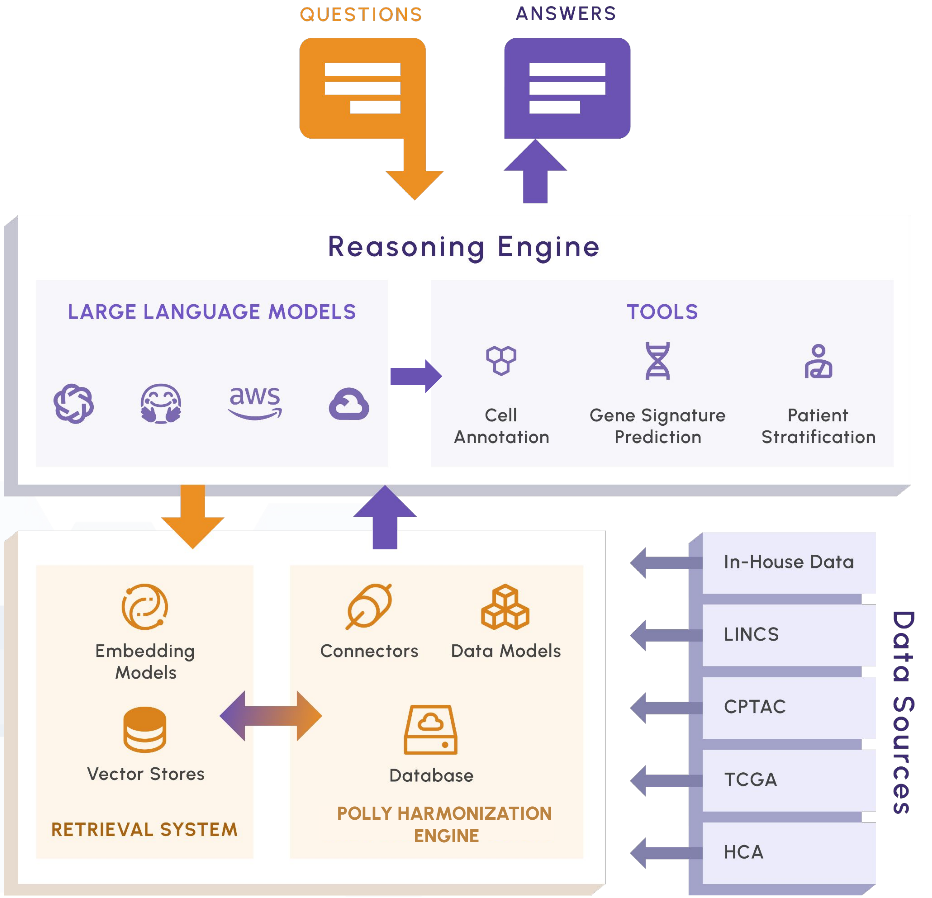
- **75% decrease** in time spent on data acquisition & preparation

We Predict..

A shift to LLMs trained on biological data to perform specific tasks using Instructions in Natural Language

Video

The Polly Ecosystem: Empowering R&D to become AI-Ready



Appendix

Benchmarking the Polly Harmonization Engine

Experimental design

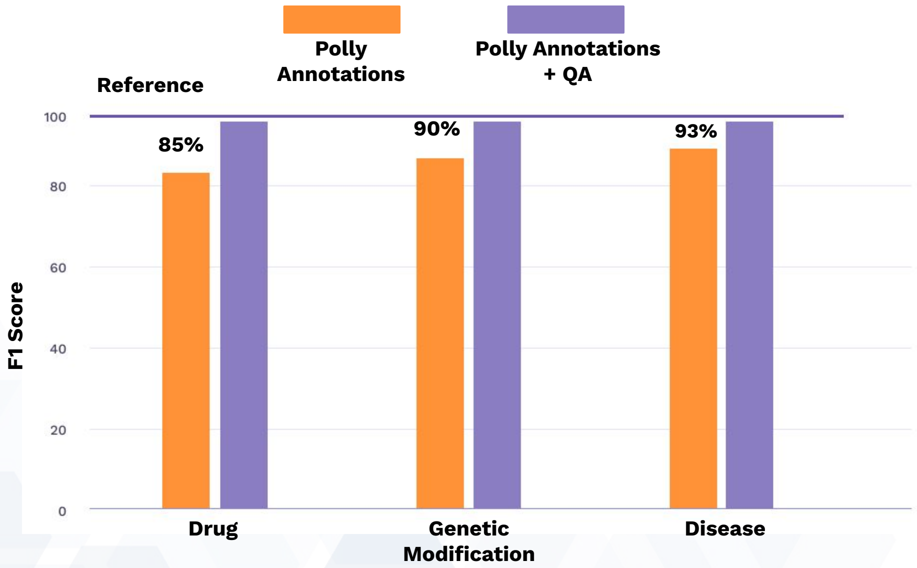
- **Reference Datasets:** We took 1500 datasets corresponding to ~30,000 samples across 300 diseases, 271 drugs and 871 genes from the *CREEDS corpus (Source)

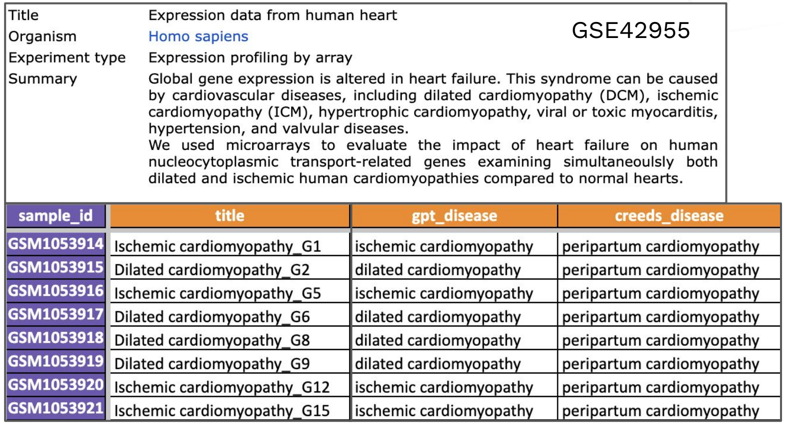
- **Polly Harmonized Datasets:** All datasets from source were processed through the 3 steps of Polly Harmonization Engine

- **Quality and TAT Performance Comparison:** We compared the quality and TAT of the performance of Polly Harmonization Engine against manually curated CREEDS datasets, for 3 fields: Disease, Drug and Genetic Modification

**CREEDS: Crowd Extracted Expression of Differential Signatures; The Maya'an Lab*

Polly Annotates Data 25X Faster, with Precision





Title	Expression data from human heart
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Global gene expression is altered in heart failure. This syndrome can be caused by cardiovascular diseases, including dilated cardiomyopathy (DCM), ischemic cardiomyopathy (ICM), hypertrophic cardiomyopathy, viral or toxic myocarditis, hypertension, and valvular diseases. We used microarrays to evaluate the impact of heart failure on human nucleocytoplasmic transport-related genes examining simultaneoulsly both dilated and ischemic human cardiomyopathies compared to normal hearts.

GSE42955

sample_id	title	gpt_disease	creeds_disease
GSM1053914	Ischemic cardiomyopathy_G1	ischemic cardiomyopathy	peripartum cardiomyopathy
GSM1053915	Dilated cardiomyopathy_G2	dilated cardiomyopathy	peripartum cardiomyopathy
GSM1053916	Ischemic cardiomyopathy_G5	ischemic cardiomyopathy	peripartum cardiomyopathy
GSM1053917	Dilated cardiomyopathy_G6	dilated cardiomyopathy	peripartum cardiomyopathy
GSM1053918	Dilated cardiomyopathy_G8	dilated cardiomyopathy	peripartum cardiomyopathy
GSM1053919	Dilated cardiomyopathy_G9	dilated cardiomyopathy	peripartum cardiomyopathy
GSM1053920	Ischemic cardiomyopathy_G12	ischemic cardiomyopathy	peripartum cardiomyopathy
GSM1053921	Ischemic cardiomyopathy_G15	ischemic cardiomyopathy	peripartum cardiomyopathy

***Ontologies** help Polly annotate samples with more precise labels than CREEDS*

Polly can annotate **25 datasets** with **30+ fields** within **1 hour.**
This is **25X faster** than manual curators (1 dataset and limited fields within 1 Hour)

elucidata

Elucidata at a Glance

Founded in 2015

8 Years in the business. **150+** team of bioinformatics scientists, ML engineers & data scientists

Traction

Discovery programs at **Eli Lilly, Janssen, Pfizer** & **30** other Biopharma partners

Funding

Capital Raised: **$23+ M**
Backed by: **F Prime Capital**, **Eight-Roads Ventures** & others

Platform

Our SaaS platform Polly curates biomedical data with human level accuracy. Scales to **10+** R&D data types

Curation

Used by leading life sciences companies on ~ **2 Million Biomedical** Datasets

Impact

Enabled the discovery of **5 drug targets** with 3 Biopharma companies, using Polly