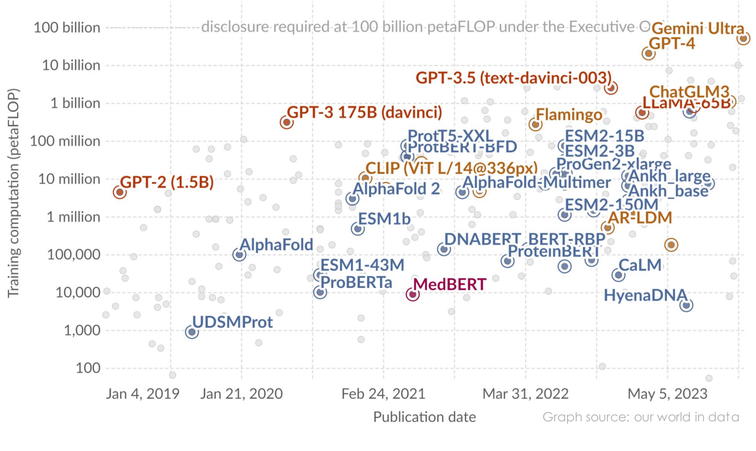
Quality Data and Creative Approaches Drive GenAI's Transformative Journey in Biopharma

Oleg Stroganov, PhD; Dan Rozelle, PhD; Tatiana Khasanova, PhD; Julie Bryant, CEO. *Rancho Biosciences, LLC*

The promise of AI: accelerate discoveries and generate insights

Machine learning techniques have been rapidly evolving since mid 50s, eventually leading to more and more sophisticated algorithms that can play chess, analyze pathology reports, drive cars, and now converse with humans on any topic.



"The McKinsey Global Institute estimates that gen AI could generate $60 billion to $110 billion annually in economic value for pharmaceutical and medtech companies."

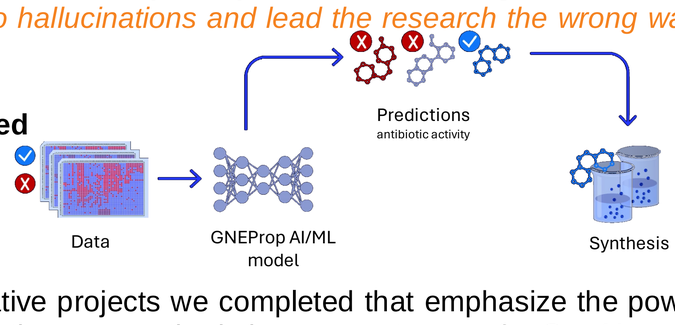
- McKinsey

Life Sciences companies have an opportunity to unlock $5-7 billion (Bn) dollars in value from the use of artificial intelligence (AI). We estimate that nearly 90% of value will be derived from three functional areas: research and development (R&D), manufacturing and supply chain, and commercial. Over the past 18 months, Generative AI (GenAI) has broadened the breadth of value that AI could deliver

- Deloitte

A recent talk by Dr Aviv Regev, of Genentech/Roche, at an NVIDIA forum[1] highlighted the promise of AI in life sciences space. As an example, machine learning was used to rapidly predict activity of candidate small molecules as antibiotics - rate of the successful candidates skyrocketed (50-fold increase) as a result. AI shines when large quantities of data exist, but the underpinnings of the biological context are too complex for humans to decipher. However, a theme is now emerging that is being articulated by many scientists and engineers: *without proper data, insights can quickly turn into hallucinations and lead the research the wrong way*. What does "proper data" mean?

-Data that is **high quality and harmonized**
-Data that is representative
-Data that is deeply annotated
-Data that is relevant to the use case



Here, we are showcasing several innovative projects we completed that emphasize the power of high-quality data and illustrate that creative new methods improve outcomes for GenAI.

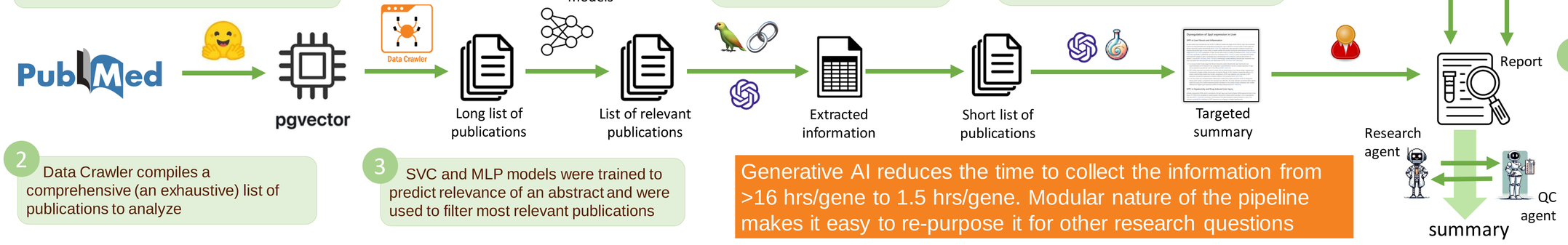
1 Lab in a Loop: AI to Transform Drug Discovery and Development | NVIDIA On-Demand

Large Scale Toxicology Data Extraction and Summarization Benefits from Multi Agent AI pipeline

Scientists want to know **why** the gene is dysregulated in disease and toxicity. Browsing all literature and reading all papers would take weeks – so, we built a pipeline that summarizes the literature and provides a 2-paragraph summary to answer this question, all under 2 hrs per report

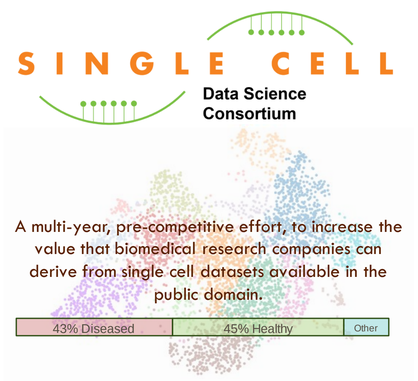
In this pipeline, we introduced a multi-agent workflow to help AI stay focused, don't hallucinate, and perform self-QC

1. PubMed abstracts were vectorized using *gte-large* embedding model and stored in vector DB for extraction
2. Data Crawler compiles a comprehensive (an exhaustive) list of publications to analyze
3. SVC and MLP models were trained to predict relevance of an abstract and were used to filter most relevant publications
4. Specific information was extracted from abstracts (sample type, species, etc) using LLM. This information was used to further filter publications
5. LLM was used to distill relevant abstracts and generate a summary of relevant findings for protein function and association of mRNA expression and disease or toxicity
6. Information from structured sources and manually curated information was collected and assembled into a detailed report
7. GPT-4-based research and QC agents were used to prepare a short summary of gene function in the context of disease and toxicity



Generative AI reduces the time to collect the information from >16 hrs/gene to 1.5 hrs/gene. Modular nature of the pipeline makes it easy to re-purpose it for other research questions

Finding sameness in millions of cells: AI foundational model for single-cell data – driven by high quality data



45 million cells:

>1M cells each from
Blood, lung, liver, heart left ventricle

>500k cells each from
Blood, lung, liver, heart, bone marrow, cortex, colon, pancreas, pleural effusion, ileum, lymph node, retina, brain, pluripotent stem cell, spleen

>250k cells each from
hippocampal formation, interventricular septum, skin of body, adrenal gland, skin epidermis, substantia nigra, mammary gland, heart right ventricle, rectum, sigmoid colon, pleural effusion, anterior cingulate cortex, apical region of left ventricle, lung parenchyma, dorsal plus ventral thalamus, dermis, frontal cortex

6.1 million cells are cancer-related:
lung(1.5M), heme(1.1M), g.i.(817k), breast(219k)

2 million nervous system diseased cells:
HD (262k), PD (223k), AD (411k), MS (141k)

1.8 million cells are immune related
PsA (358k), Ps (291k), UC (397k), dermatitis (198k)

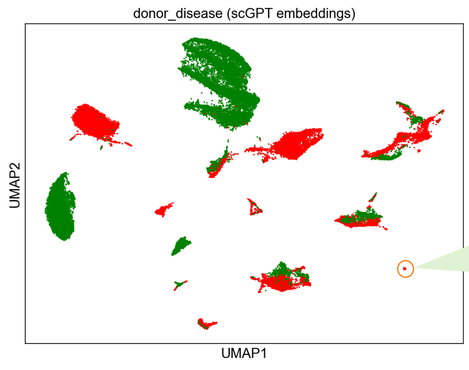
1.5 million cells G.I. dysfunction
Crohn's (532k), IBD (560k), cirrhosis (62k), intestinal cancers (817k)

Routine analysis identified a distinct niche population of diseased cells.
Can these cells be found in other datasets?
- which tissue/perturbation
- what are disease association?

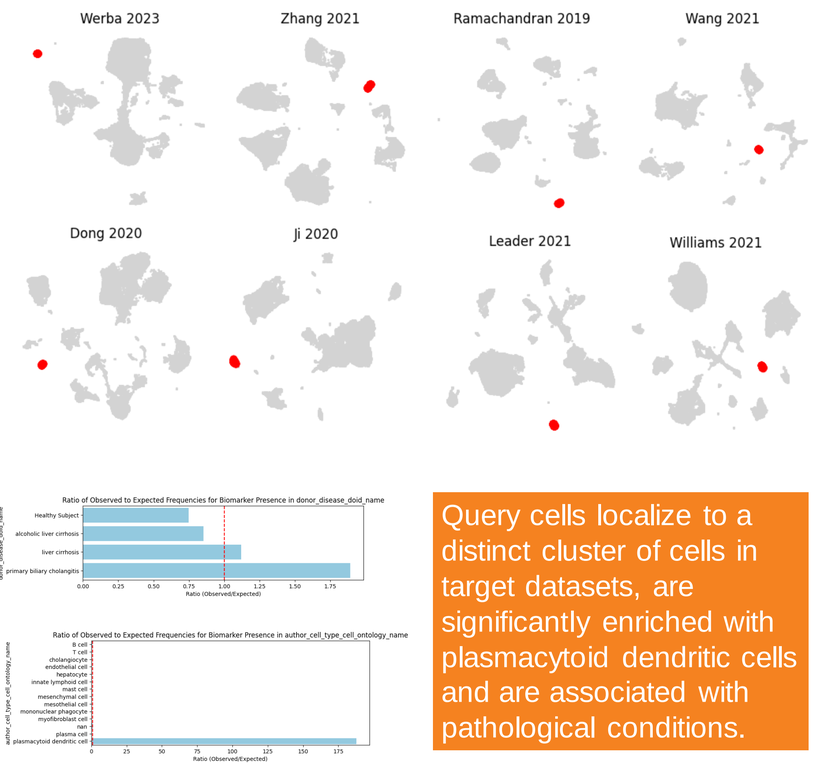
With typical methods source datasets need to be integrated with query datasets and determine if cell populations are co-localized. This is a complicated and time-consuming process.

Single cell transcriptomic foundational models such as scGPT makes this integration possible, and relatively easy to implement with harmonized data.

scGPT embeddings were calculated for ***70 AI-ready datasets from Rancho's SCDS*** (11.2 million cells) and loaded into a vector database. Query cells were compared to reference cell types.



This rare cell subtype appears only in Crigler Najjar samples.
Goals:
- identify other datasets with similar cells.
- determine associated factors



Query cells localize to a distinct cluster of cells in target datasets, are significantly enriched with plasmacytoid dendritic cells and are associated with pathological conditions.

What's next in AI for IT, R&D and clinical operations

There is a 2-3 order of magnitude difference between big tech and big pharma in investment into compute.

However, pharma operates with the same scale of data: e.g. 200M cells is ~8 trillion tokens, which is more than went into GPT-3 training.

The compute gap offers great opportunities for pharma: applications of text-based models, and development of new foundational models

What's next for R&D?

- Retrieval-Augmented Generation (RAG) applications to increase data accessibility is a quick win
- Content summarization powered by AI agents to reduce hallucinations
- Multi-modal applications in omics and images for advanced analytics

- Data-powered autonomous AI agents will lead to R&D advances
- New foundational models based on data that only pharma has – multi-omics, time series, clinical outcomes

What's next for clinical operations?

- Enhanced information sharing and access through RAG applications
- AI-agents empowered on-demand data analysis and dashboarding
- Biomarker discovery through existing multi-modal AI applications in single cell transcriptomics

- Extended multi-modal applications and dataset integrations
- Foundational models integrating clinical outcomes, biomarkers and existing knowledge graphs

What's next for IT?

- Compute-intensive GenAI support for day-to-day inference tasks
- Data needs to be ready for AI consumption
- Vectorization layer for data infrastructure
- AI knowledge management to prevent prompt injection and other attacks

- High compute and data requirements to train breakthrough AI models mean collaboration might be the key to unlock GenAI potential