AI x 4Catalyzer Summary Talk + Hackathon Topics

Andrew Kean Gao
Akshita Panigrahi

July 28, 2023

4CATALYZER

The goal of the LLMs for Bio hackathon is to:

Leverage **LLMs and Generative AI** towards a **key need or focus of 4Catalyzer companies.**

Judges are looking for:

- Usefulness and relevance to 4Catalyzer
- Technological complexity/advancement
- Innovativeness
- Uniqueness
- **Integration of an LLM**

4CATALYZER

Judges do **not** want to see:

- Kaggle-style data science/ML projects
- Non-unique projects like Skin Cancer Detection using HAM1000 dataset
- Projects irrelevant to 4C
- Projects only using prompt engineering + the “raw” GPT-X API

4CATALYZER

Broad Project Ideas:

- **Image-to-text:** Given a patient's MRI, retinal, or ultrasound scan image, the LLM will interpret the results, suggest next steps, potential health conditions, and suggestions.
    - This should go *beyond* merely classifying a single disease condition.
- **Sequence-to-text:** Given a patient's DNA sequence for a specific gene/region, the LLM will interpret the sequence. Example: "Based on your DNA sequence, I anticipate that you have blue eyes". "Based on your DNA sequence, I anticipate that you are at higher risk for skin cancer because....."
    - **Integrate** knowledge from databases like GeneCards, NCBI, OMIM, etc.
- Creating a platform or tool that helps researchers working on Directed Evolution for PETase and TdT enzymes

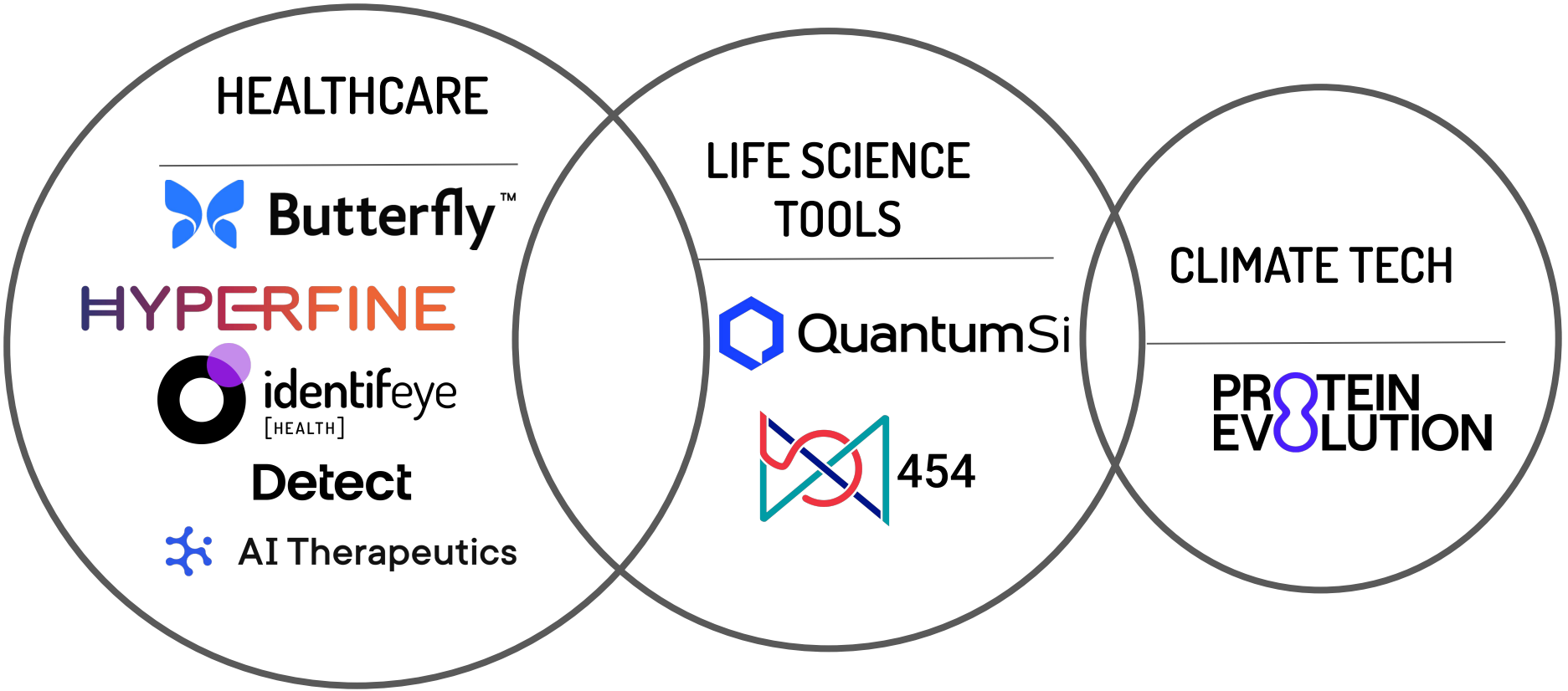
4C Overview - Abridged Version

4Catalyzer

4Catalyzer is a technology incubator founded by Dr. Jonathan Rothberg with the **mission to save lives and maximize societal impact**.

We work at the intersection of engineering, machine learning and natural sciences to support innovation in healthcare, life sciences, and climate tech.

4Catalyzer



CURRENT STATE - part I

Health information is often inaccessible to many patients

Confined to specialists' offices

Imaging and testing is primarily confined to urban settings, large hospitals and specialists' offices, making access to care inconvenient and difficult for patients.

Large and complex devices, complex workflows

Existing devices are immoble or complex and don't allow providers to meet patients where they are

Prohibitively expensive

The price point of existing devices puts them out of reach for most non-specialized facilities and use cases

CURRENT STATE - part II

Life science tools are often inaccessible to many researchers

Confined to specialized labs

testing is primarily confined to specialized labs and

Large and complex devices

Existing devices are complex, have large physical overhead

Prohibitively expensive

The price point of life sciences tools such as sequencing platforms and mass spectrometers (including reagents and test kits and data analytics) puts them out of reach for most non-specialized facilities and use cases

CURRENT STATE - part III

Drug Development is costly and Drugs can be inaccessible to patients

Lengthy, complex, and costly - high degree of uncertainty that a drug will actually succeed.

Target Identification is challenging - unknown pathophysiology for many disorders

Heterogeneity of the patient population -requires processing of large amounts of data (clinical phenotyping and subtyping)

Lack of tools for improved target identification and validation

Hackathon Topics

4CATALYZER

Butterfly™

Butterfly is commercializing the world's first ultrasound on a semiconductor chip. "BFLY" went public in February of 2021 on the NYSE raising half a billion dollars to continue to increase hospital connectivity as well as to further develop AI applications, ultimately enabling home use.

NYSE: BFLY

Democratizing access to Ultrasound Imaging

- World's first full body ultrasound scanner on a semiconductor chip, fused with AI and cloud technology
- The quality of a $60k device for a fraction of the price
- Two thirds of the world does not have access to medical imaging, and two thirds of diagnostic issues can be resolved with simple imaging.
- Saving lives on all 7 continents, partnership with the Gates Foundation for large scale deployments in Limited Resource Settings

Butterfly Project Ideas

Ultrasound Physician Assistant

An **LLM-based assistant that will use information immediately available during a doctor visit (e.g. primary care visit or checkup) to suggest an ultrasound imaging procedure appropriate for improving the overall assessment.** The LLM can be trained/fine tuned on published literature about the utility of ultrasound for assessing various conditions coupled with general medical knowledge linking information available during a doctor visit to target possible conditions. The resulting solution can take baseline information on the patient, augmented with additional information obtained during the visit (and possibly gleaned from a transcription of dialog between patient and doctor) and provide a real time suggestion on possible ultrasound scans that should be performed during the visit. The solution can also suggest training resources for obtaining the scan (from the public domain and from Butterfly Academy resources). Information utilized can include all or a subset of:

- A description (written or transcribed) of observations obtained in the course of the visit, including
  - Patient supplied descriptions of symptoms
  - Doctor's immediate observations
  - Live transcription of dialog between patient and doctor
- Basic information immediately available in the patient's EHR, such as age, sex, height, and ethnicity
- Metrics obtained from the current visit
  - Blood pressure
  - Heart rate
  - Weight/BMI
  - Blood oxygenation
- Historical health information including
  - Historical values of basic metrics, including weight, blood pressure, etc.
  - Lab test results
  - Past diagnoses
  - Current and past prescriptions
  - Write-ups from previous visits

Follow on: Utilizing the above information *and* an ultrasound image scan, suggest a diagnosis.

Butterfly Project Ideas

Ultrasound Pathology Simulator

Problem: A challenge in advancing the adoption of point of care ultrasound (POCUS) is rooted in the **lack of training or knowledge in how to use ultrasound** during the course of practice. One benefit of handheld ultrasound is the ability to practice on one's self or on colleagues to gain proficiency in image acquisition. However, a key part of training is learning how to identify anomalies and pathologies in the images acquired by the operator and one barrier to training is the lack of access to patients with said pathologies.

Idea: Since it is quite easy to practice POCUS with a handheld, the ability to **simulate and introduce pathologies into a scan** would greatly increase the operator's confidence in their ability to leverage ultrasound to formulate a diagnosis. Recent capabilities in generative AI can potentially bridge this gap by **taking a live ultrasound image, a textual description of the desired pathology (i.e. "gallstones", "malignant breast tumor", "abdominal aortic aneurysm") and produce a new image** substantially similar to the given image, but with the requested pathology present

Butterfly Project Ideas

Ultrasound Radiologist in a Box

Build upon recent advances in semantic image AI to implement an anatomical feature and pathology detector for ultrasound images. Given an ultrasound image, **highlight key anatomical landmarks** (e.g. Iliac Vein, Right Kidney, gallbladder, etc.) and also **highlight abnormalities and potential pathologies** (e.g. gallstone, enlarged cardiac septum, etc.). Leverage published **literature** to extract images, image callouts, and captions for training. In addition to a textual **description** of observations, spatially indicate the **location** of various semantic components in the image.

4CATALYZER

HYPERFINE

"HYPR" went public in 2021 on NASDAQ raising quarter billion dollars to advance the product pipeline,continue to increase hospital connectivity as well as to further develop AI to give clear images

NASDAQ: HYPR

Democratizing access to MRI

- MRI is one of the safest forms of medical imaging - but is unavailable in 90% of the world - Hyperfine is changing that
- World's first Portable MRI
    - Low Cost, easy to transport to the patient's bedside, plugs into a wall outlet
- Powered by AI to give clear images and insights
- In use at over 100 sites worldwide

Hyperfine Project Ideas

Learn to **generate the appearance of stroke lesions in and then insert them into another data.** This would help to train and improve stroke detection / segmentation projects. We will not have large-enough dataset for this yet (have to wait for ACTION-PMR), but there are some publicly available datasets this could be done with.

1. Use labeled dataset for **lesion simulation** and generation: http://www.isles-challenge.org/
2. **Insert** the **lesion** into another dataset, e.g.: https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release

Extra points:

1. Retrain segmentation algorithm on generated lesion and submit to the challenge (this should improve ranking).
2. Downgrade the data by adding noise and reducing resolution as a simple transformation to low-field-like appearance.

Hyperfine Project Ideas

1. Use LLM to train (or finetune opensource LLM model from Meta) on large scale radiology report dataset (e.g.: https://physionet.org/content/mimic-cxr/2.0.0/). I wish we had one for MRI.
2. Develop system to retrieve reports with specific queries about the findings. This will be much more powerful than query search.

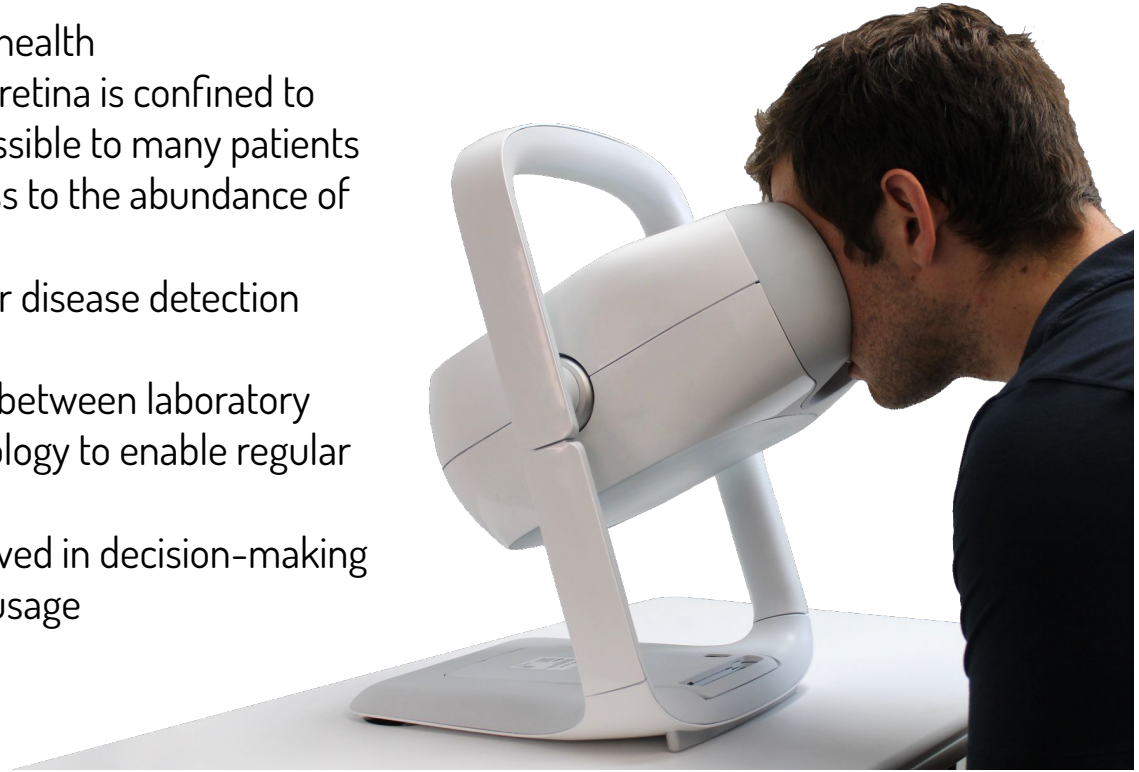
4CATALYZER

identifeye
[HEALTH]

identifeye HEALTH (fka Tesseract Health) raised $80M through Series B to develop and launch its first product. We are building an intuitive, consumer friendly, medical device to capture health information from the eye.

Create a New Branch in Diagnostics – Making Retinal Imaging as simple as an eye selfie

- Eyes are a window to the body and health
- Access to health information in the retina is confined to specialists' offices and often inaccessible to many patients
- Our mission is to democratize access to the abundance of health information in the eye
- Prevent vision loss by making ocular disease detection easier and more accessible
- Create a new branch in diagnostics between laboratory medicine (e.g. bloodwork) and radiology to enable regular non-invasive health monitoring
- Empower patients to be more involved in decision-making by reducing barriers to access and usage



Identifeye Project Ideas (open-ended)

We are currently focusing on fundus images - and our first application is Diabetic Retinopathy and in the pipeline we have a big focus on cardiovascular.

Some of the most straight forward from the public databases:

- MESSIDOR/MESSIDOR-2/IDRiD and Kaggle DR set for DR
- DRIVE and STARE for a variety but they're often used for vessel segmentation
- AIROGS and REFUGE for glaucoma
- AREDS for AMD
- UKB is great but access is not trivial to get access to

If you want something easy and straightforward to obtain, start with the **DR** and maybe **DRIVE** and **STARE**

Liminal Sciences

Liminal Sciences is developing a non- invasive brain monitor to sense, understand, and ultimately heal the brain.

Liminal is building the world's first non-invasive brain monitor

- The Brain is as important to monitor as the heart - but today there is no way to do this without drilling a hole in the patients skull
- Liminal is building the first wearable brain monitor for acute and chronic conditions including epilepsy, stroke, and traumatic brain injury.
- Enabling a brain monitor as ubiquitous as the heart monitor
- Partnered with AI for quicker and more valuable health insights.

4CATALYZER

Detect

In response to the global pandemic, Detect built and brought-to-market a PCR-quality rapid molecular home Covid-19 test authorized for EUA by the FDA. It is now focusing on a platform for POC and home use.

Access to rapid PCR quality testing

- Detect's proprietary technology serves as the platform for its future home tests, including the next-generation Detect Covid+Flu Test, as well as rapid molecular home tests for respiratory health, Strep and STIs.
- Super easy to use, low cost, and fast
- Enables stakeholders to test on site, without sending to a lab

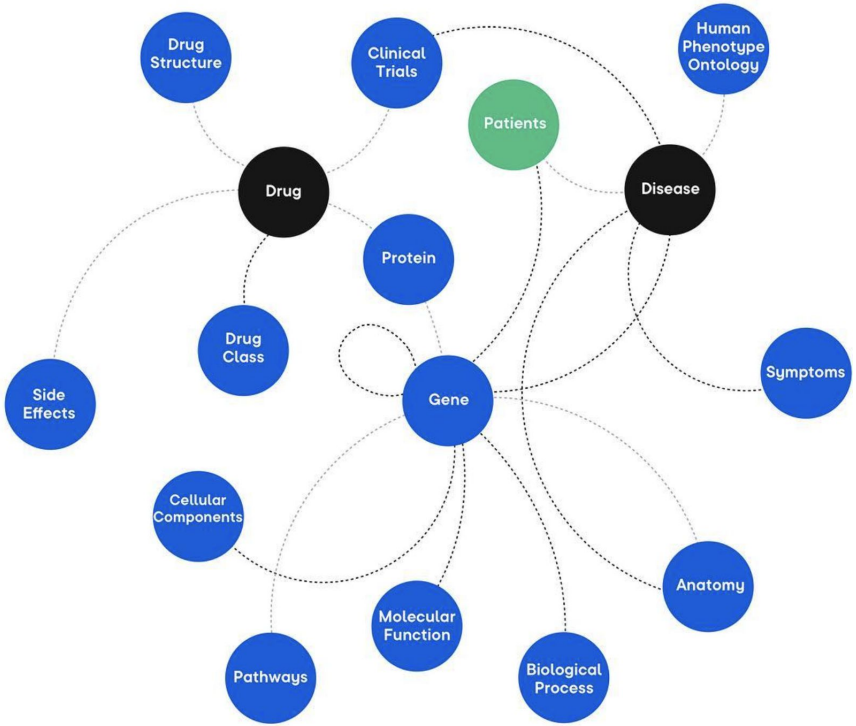
4CATALYZER

AI Therapeutics

AI therapeutics is identifying and developing treatments for rare orphan conditions.

Access to rapid PCR quality testing

- Uses AI Algorithms to identify promising assets to add to their pipeline
- AI analyzes what these safe drugs can be used for
- Helping find treat for orphan/rare diseases
- Current clinical stage drug candidates target several orphan disorders including amyotrophic lateral sclerosis, pulmonary arterial hypertension, bronchiolitis obliterans, pulmonary sarcoidosis, Ewing sarcoma, rhabdoid tumor, and SWI/SNF mutated or dysregulated cancers.



4CATALYZER

QuantumSi

Quantum-Si has built the world's first single molecule protein sequencer on a semiconductor chip. QSI went public in 2021 raising over $500 million to revolutionize the future of diagnostics.

NASDAQ: QSI

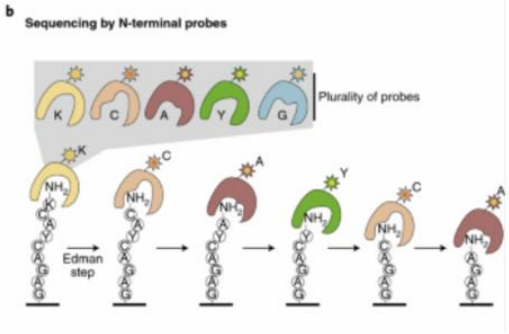
Access to rapid PCR quality testing

- First ever single molecule protein sequencing device
- DNA tells you what may happen, proteins tell you what is happening/about to happen
    - Could give us the ability to predict a heart attack before it happens with high accuracy
- Deeper proteomic insights will advance science and human health

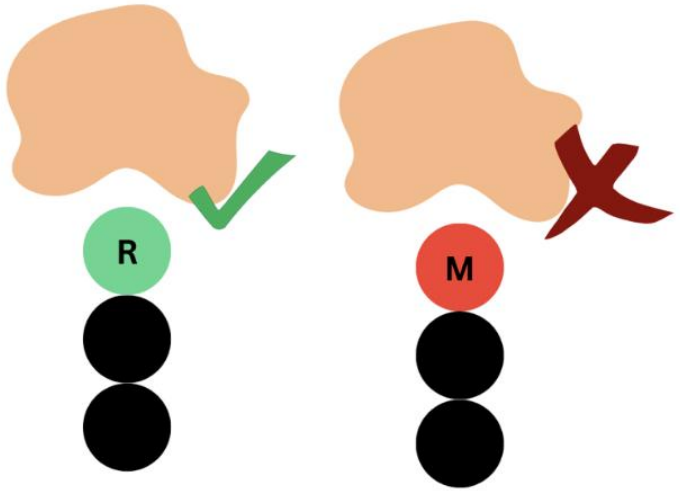
Quantum-Si Project Ideas

Amino acids to target: G, P, or M

Designing proteins that selectively bind to a specific amino acid from the side (N-terminus) and two more amino acids, of any residue.



Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device (Reed et. al)
https://www.science.org/doi/10.1126/science.abo7651



Example: this binder selectively binds to arginine, but not other amino acids

QSI De Novo Residue Calling: Single molecule protein sequencing

Utilize machine learning methods to assign binding events (ROIs) to amino acids without the use of a reference peptide sequence. Dataset will be shared.

	Input	Method	
Current Method	ROIs w/Binder Assignment and Pulse Width Peptide Reference Sequences	Custom Aligner w/Scoring Based on Matching Expected Pulse Width, Gaps, and Deletions	
Proposed Method	ROIs w/Binder Assignment, Pulse Width, Inter-pulse duration, ROI duration, and inter-ROI duration	Assign residue based on prediction from trained model	Amino acid sequence of read

4CATALYZER

454 Bio

Founded in 2022, 454 Bio is building the first Next Generation DNA sequencer (NGS) fit for at-home use

Decentralized On-site Sequencing

Today, personal genomics is limited by expensive and cumbersome equipment - turnaround time is slow, and sequencing requires expertise and resources

454 Bio's mission is to bring genomics out of the lab and enable universal access to affordable and fast DNA sequencing.

Low Cost device, easy to use kits, seamless data transfer for results without expertise

454 Bio is supporting a wide range of use cases in public health, research, and consumer genomics.

454 Bio Project Ideas

Variant Evolution

Link SARS variant evolution with geography and/or clinical severity, and then try to predict which novel variants may be "next". For example, it could be any respiratory virus..

Find a link between sets of variants across a human genome to disease (any) with onset.

SARS datasets: https://www.ncbi.nlm.nih.gov/sars-cov-2/

Cancer Genome Stability

Identify regions in the human genome which are stable in cancer or the opposite.

Manually curated somatic cancer mutation data: https://cancer.sanger.ac.uk/cosmic/download

Cancer study data: https://www.cbioportal.org/datasets

Genomic Data Commons Data Portal: https://portal.gdc.cancer.gov/

Raw SNP data: https://www.ncbi.nlm.nih.gov/snp/

4CATALYZER

PROTEIN EVOLUTION

Protein Evolution is leveraging recent breakthroughs in natural science and artificial intelligence to design enzymes to break down end-of-life textile and plastic waste into the building blocks that make up new textile and plastic products. Protein Evolution aims to help the chemicals industry transition to a lower-carbon, circular economy.

Infinitely recyclable High quality plastic

- Identify the waste - water bottle, car tire, or piece of clothing.
- Engineer enzymes - Uniting natural science and artificial intelligence, develop enzymes that break down the waste source so we can recycle it in an economical, sustainable way.
- Break down the plastic into its “building blocks” in a low-emission, eco-friendly process.
- Reproduce materials by using these “building blocks” to create good-as-new plastic bottles, textiles, and other infinitely renewable plastic products.

PEI Project Ideas

AI Lab Trainer

Using public resources, train an AI model that can teach lab scientists how to perform a new experiment. The model should generate lab protocols and explain the caveats of the experiment and provide information about expected results.

We envision it to be a conversational AI model but please feel free to think outside the box.

AI Safety Inspector

Given lab protocols, download all relevant MSDS sheets and generate a safety report. For example, "this experiment requires BSL-3" or "this experiment will accidentally create mustard gas as an intermediate". It would be great if the AI Safety Inspector could also suggest solutions to mitigate the safety issues. As an additional aim, AI Safety Inspector could highlight where relevant information is missing in the protocol and propose to add it. Relevant AutoGPT paper: https://arxiv.org/ftp/arxiv/papers/2304/2304.05332.pdf. We envision this to be a Generative AI model but please feel free to think outside the box.

PEI Project Ideas

A protein-specific, scalable vector search platform
Given a protein sequence, find the nearest neighbors in function space. For example, given an enzyme, find other enzymes in the database that have similar optimum pH.

ProteinLLM to predict optimum pH
Given an enzyme, predict the optimum pH at which the enzyme would perform.
This is a difficult problem. See this relevant paper: https://www.biorxiv.org/content/10.1101/2023.06.22.544776v1. The RSME can be as high as 2-3 pH units, which is too high an error to be useful. It would be interesting to **even** learn more about why the error is so high and how we can improve it. We envision this to be a discriminative AI model but please feel free to think outside the box.