GlyGen

Automated Publication Analysis

Isaac Kim, MS Bioinformatics, The George Washington University

2026 Spring Volunteership Symposium April 15, 2026 @ 4 PM

Advisors: Rene Ranzinger, Sena Arpinar, Urnisha Bhuiyan

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

About Me

2023

B.S. in Biochemistry

University of Maryland

2026

M.S. in Bioinformatics and Molecular Biochemistry

The George Washington University

Introduction

Aim:

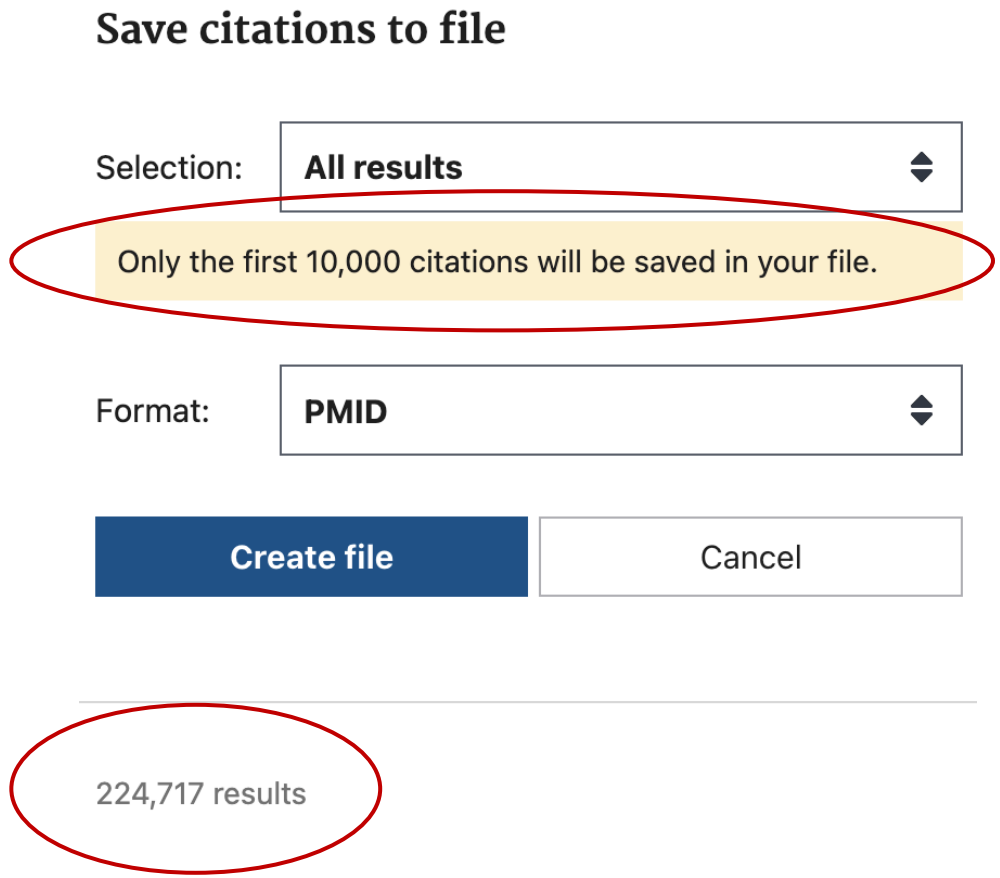
- Calculate the size of a research community (number of people, number of research groups, number of countries)
- For example: glycomics community, proteomics community

Solution:

- Query PubMed to retrieve publications from research community
- Analyze publication abstracts to identify contributing research affiliations, authors, and countries

Compile Various PMID Lists Based on Keywords

- Downloading abstracts from PubMed based on keyword search has limitations
- EDirect bypasses the 10,000 citation limit



Proteomics keyword citation limitation via PubMed API

EDirect Program from PubMed

- NCBI's suite of interconnected databases
- Allows access to Enterez from a Ubuntu terminal window
- Search terms are entered as command line arguments

- URL to Webpage

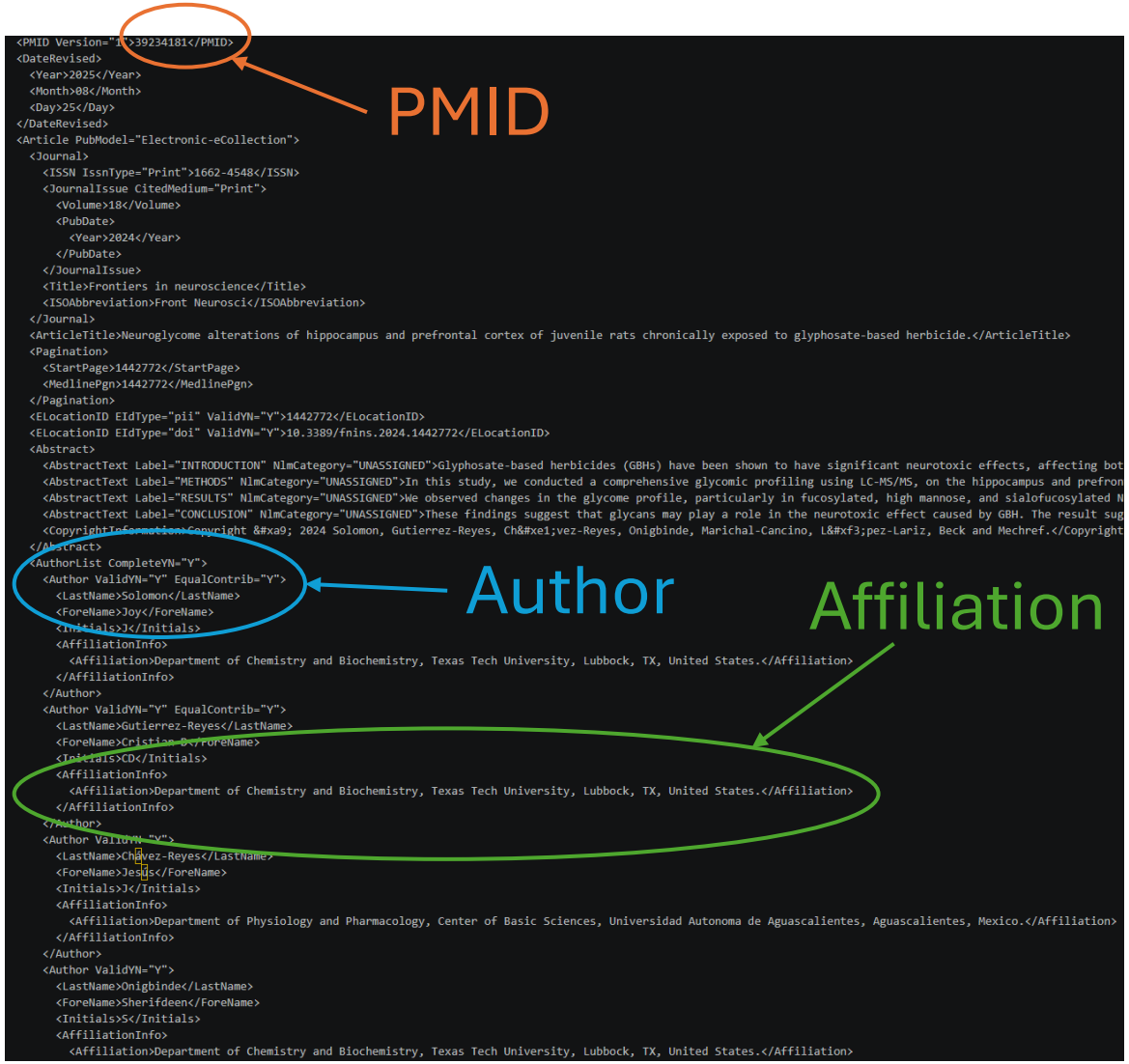
https://www.ncbi.nlm.nih.gov/books/NBK179288/

```
(venv) wildcard@IsaacTower:~$ esearch -db pubmed -query "proteomics" | efetch -format xml > proteomics.xml
```

EDirect Command Line Output

EDirect Filtering

- Command line shows all contents in XML format
- Need only necessary components
  - Author names
  - Country/Affiliation
  - PMID



XML Snippet of “Glycomics” Keyword

Analyze Paper Abstracts for Research Affiliations

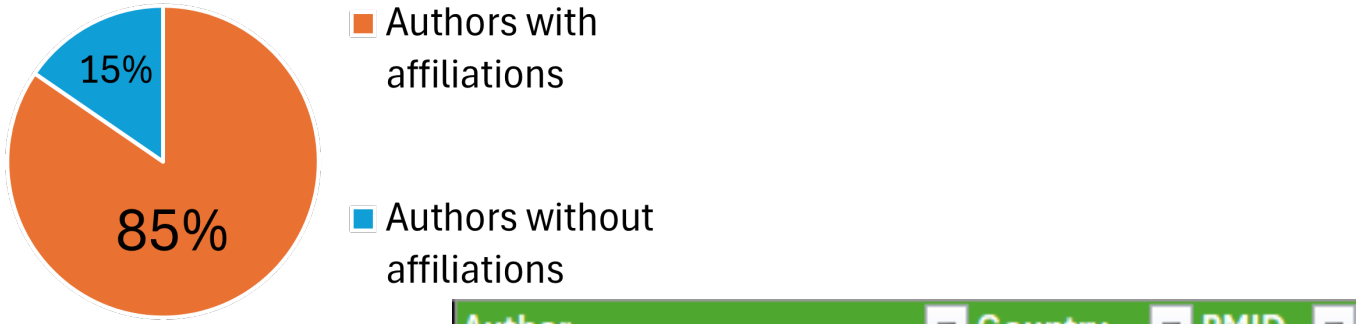
- Tried to use xtract command to parse the authors, affiliations, and country
- Did not work because
  - 1. Issues lining up authors and affiliations
  - 2. Unorganized affiliation list

27573070
KimJWGraduate Program in Functional Genomics, College of Life Sciences and Biotechnology, Yonsei University , Seoul 03722, Korea.Yonsei Proteome Research Center ,
HwangHKorea Basic Science Institute , Ochang 28199, Chungbuk, Korea.
LimJSYonsei Proteome Research Center , Seoul 03722, Korea.
LeeHJYonsei Proteome Research Center , Seoul 03722, Korea.
JeongSKYonsei Proteome Research Center , Seoul 03722, Korea.
YooJSKorea Basic Science Institute , Ochang 28199, Chungbuk, Korea.
PaikYKGraduate Program in Functional Genomics, College of Life Sciences and Biotechnology, Yonsei University , Seoul 03722, Korea.Yonsei Proteome Research Center
36575396
LiangJXDepartment of Oncological Surgery, Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital), No. 1 of Banshan East Road, H
ChenQDepartment of Oncological Surgery, Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital), No. 1 of Banshan East Road, H
GaoWSchool of Medicine, Zhejiang University City College, Hangzhou, People's Republic of China.
ChenDDepartment of Oncological Surgery, Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital), No. 1 of Banshan East Road, H
QianXYSchool of Medicine, Zhejiang University City College, Hangzhou, People's Republic of China.
BiJQSchool of Medicine, Zhejiang University City College, Hangzhou, People's Republic of China.
LinXCSchool of Medicine, Zhejiang University City College, Hangzhou, People's Republic of China.
HanBBSchool of Medicine, Zhejiang University City College, Hangzhou, People's Republic of China.
LiuJSDepartment of Oncological Surgery, Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital), No. 1 of Banshan East Road, Ha

TSV Snippet of Unorganized List

Extraction Script

- Ran a Python script that created a CSV file with strictly 3 columns: author, country, PMID
- Ran into issues with authors not having any affiliation and no country



Author	Country	PMID
Callewaert N	Belgium.	12626389
Schollen E	Unknown	12626389
Vanhecke A	Unknown	12626389
Jaeken J	Unknown	12626389
Matthijs G	Unknown	12626389
Contreras R	Unknown	12626389

```
<AuthorList CompleteYN="Y">
  <Author ValidYN="Y">
    <LastName>Callewaert</LastName>
    <ForeName>Nico</ForeName>
    <Initials>N</Initials>
    <AffiliationInfo>
      <Affiliation>Department of Molecular Biomedical Research, Ghent University and Flanders Interuniversity Institute for Biotechnology, K.l.-ledeganckstraat 35, B-9000 Ghent, Belgium.</Affiliation>
    </AffiliationInfo>
  </Author>
  <Author ValidYN="Y">
    <LastName>Schollen</LastName>
    <ForeName>Els</ForeName>
    <Initials>E</Initials>
  </Author>
  <Author ValidYN="Y">
    <LastName>Vanhecke</LastName>
    <ForeName>Annelies</ForeName>
    <Initials>A</Initials>
```

Conclusion

GitHub Repository Link

https://github.com/glygener/pubmed-analyzer

Challenges

- PubMed does not recognize an author's affiliation in all of their publications
- Affiliation string not systematically formatted
  - **Amster IJ USA.** PMID:41831571
  - **Du W China. Electronic address:weidu@mail.hust.edu.cn** PMID:41833420

Future Directions

- Run statistics of data to create bar chart of top 5-10 countries
- Create line chart showing the number of publications over recent years
- Create a visual world map to highlight countries with most article entries on certain keywords
- Clean up country string
- Extract research group, research institute, and cities from affiliation

Visual world map example

Acknowledgements

- Raja Mazumder
  - The George Washington University, Washington, DC
- Rene Ranzinger
  - University of Georgia, Athens, GA
- Sena Arpinar
  - University of Georgia, Athens, GA
- Urnisha Bhuiyan
  - The George Washington University, Washington, DC

GlyGen

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

QUESTIONS?

Visualizing Glycomics Databases and Their Features

2025 Spring Volunteership Symposium

April 15th 4-6 PM

POC’s: Dr. Rene Ranzinger, Urnisha Bhuiyan, Sujeet Kulkarni

Introduction

Diya Kamalabharathy
Poolesville High School

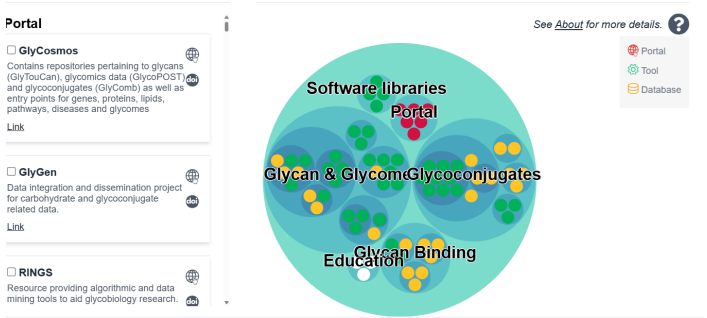
Description of Project

The **Glyspace Alliance**: Collaboration between Glygen, GlycoExpassy, and GlyCosmos. Main focus is on glycan related information through a collection of databases, software tools, data repositories, etc.

Current Issues: Scattered resources and finding resources with specific functionalities

Objective: Creating a tool to help users filter out and find information related to the Glyspace Alliance.

Current - bubble visualization



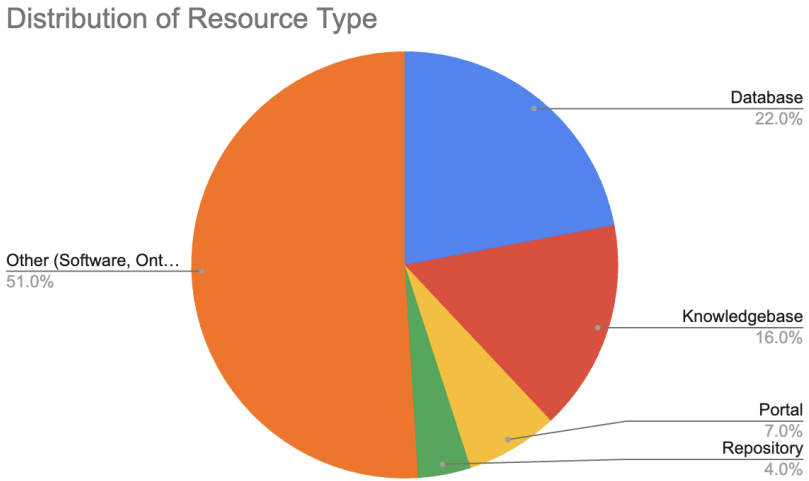
Complementary visualization

Target Features:

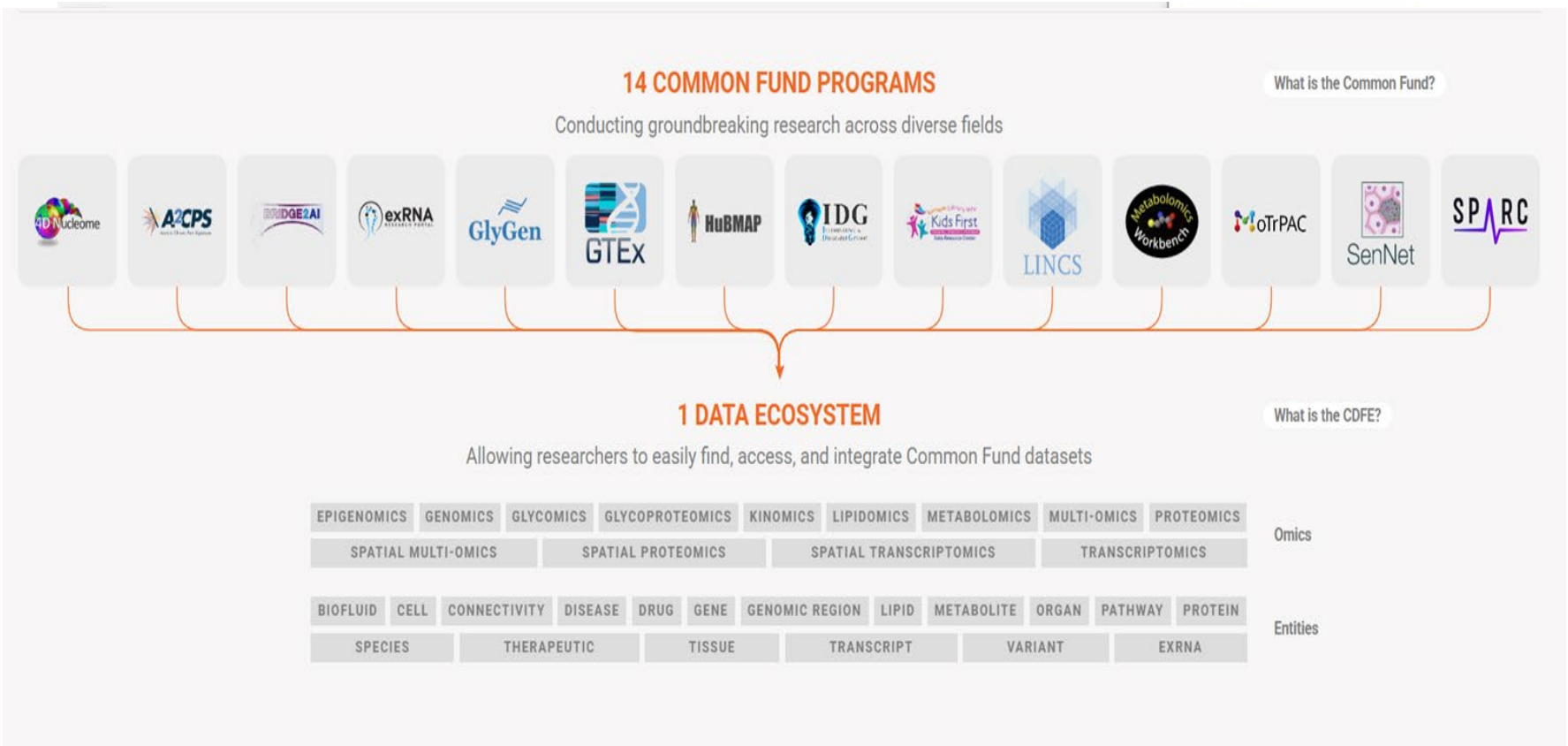
- User friendly
- Flexible
- Filtering options

Data Curation

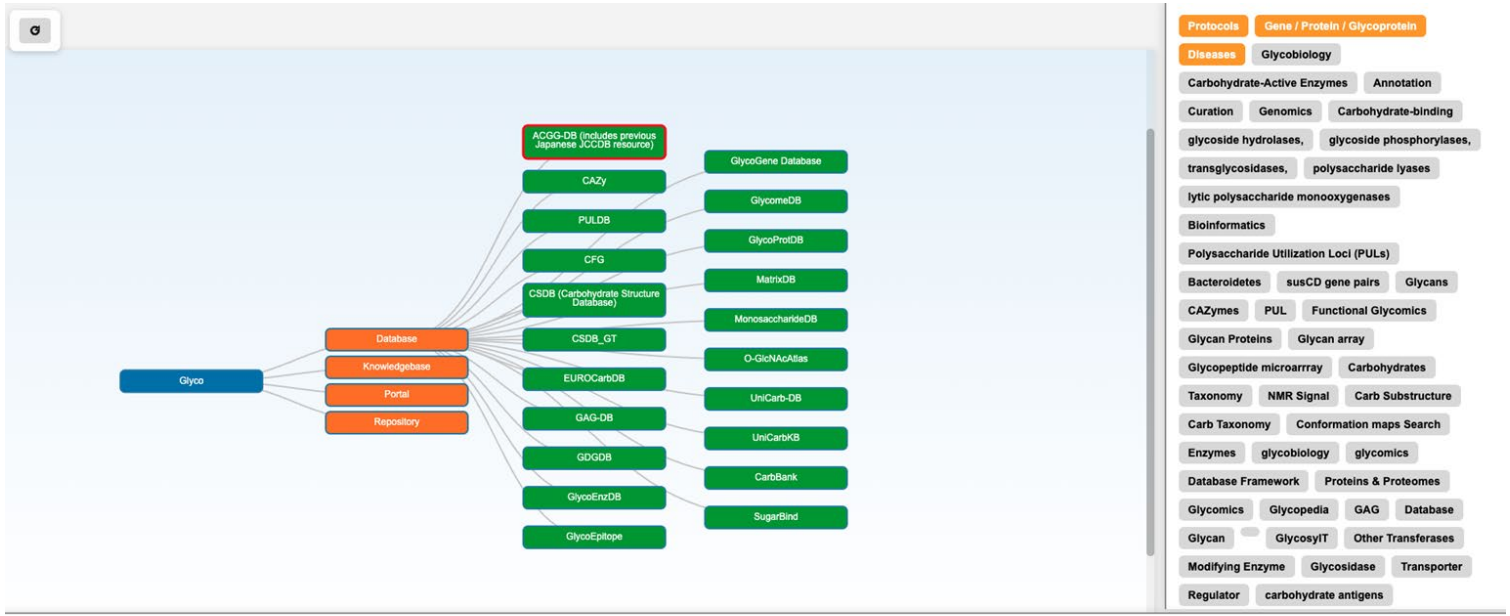
RESOURCE_NAME	RESOURCE_TYPE	DATA_TYPE
ACGG-DB (includes previous Japanese JCCDB	portal	Lectins, Carbohydrate Active Enzyme, Protein, Disease
CAZy	Database	Carbohydrate Active Enzyme
PULDB	Database	Carbohydrate Active Enzyme
CSDB (Carbohydrate Structure Database)	Database	Glycan
CSDB_GT	Database	Carbohydrate Active Enzyme
GAG-DB	Database	Glycosaminoglycan
GDGDB	Database	Disease



First Version – Tree Model



Problems With First Model



1. Scalability Issues

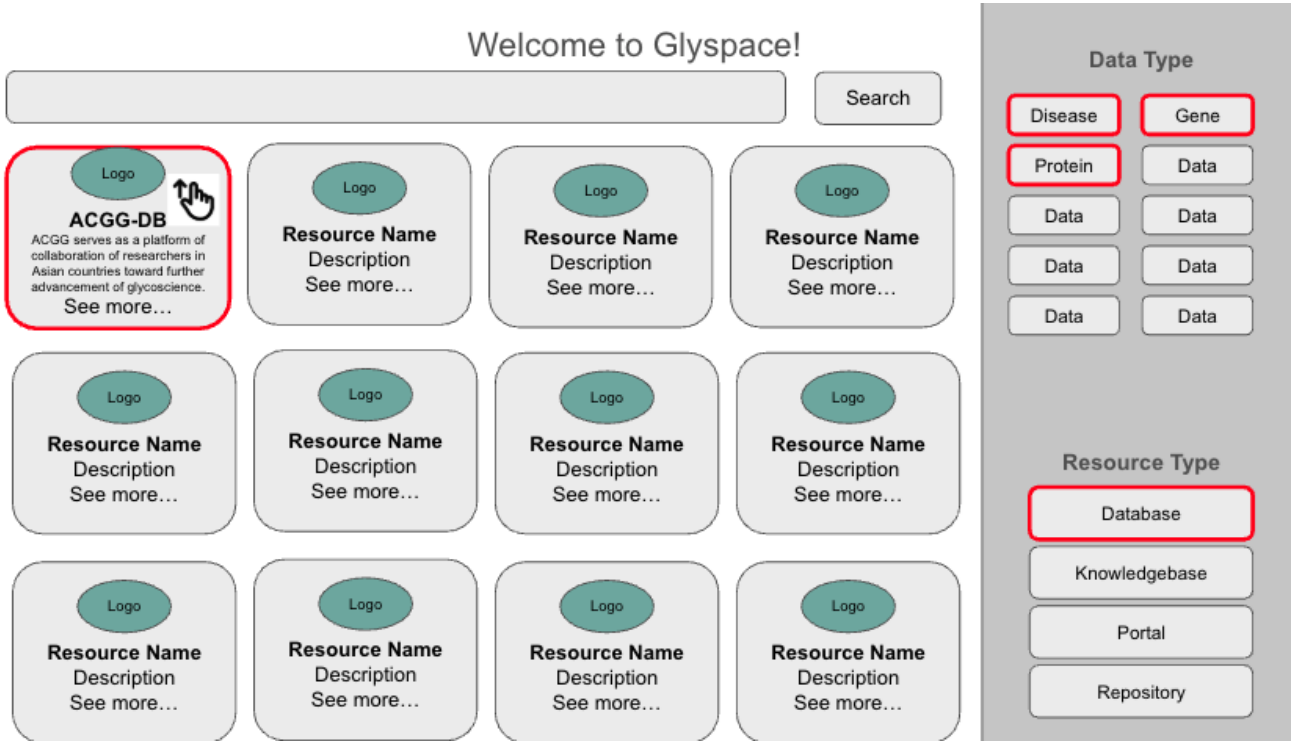
2. Redundant "Root Node" Selection

3. Perspective Switching Complexity

4. Non-Intuitive Layout

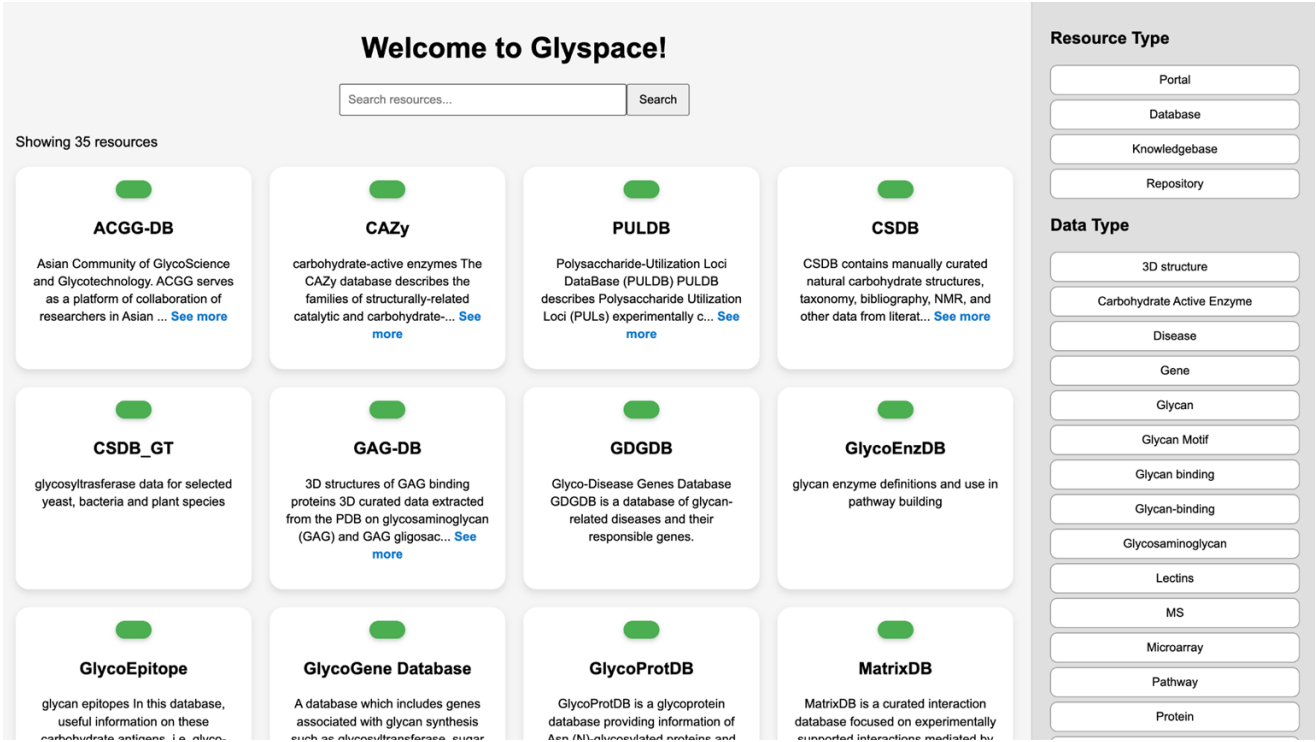
Redesign

- Filtering
- Search bar
- Cards with “summaries”
- Click → redirects to another page with more info

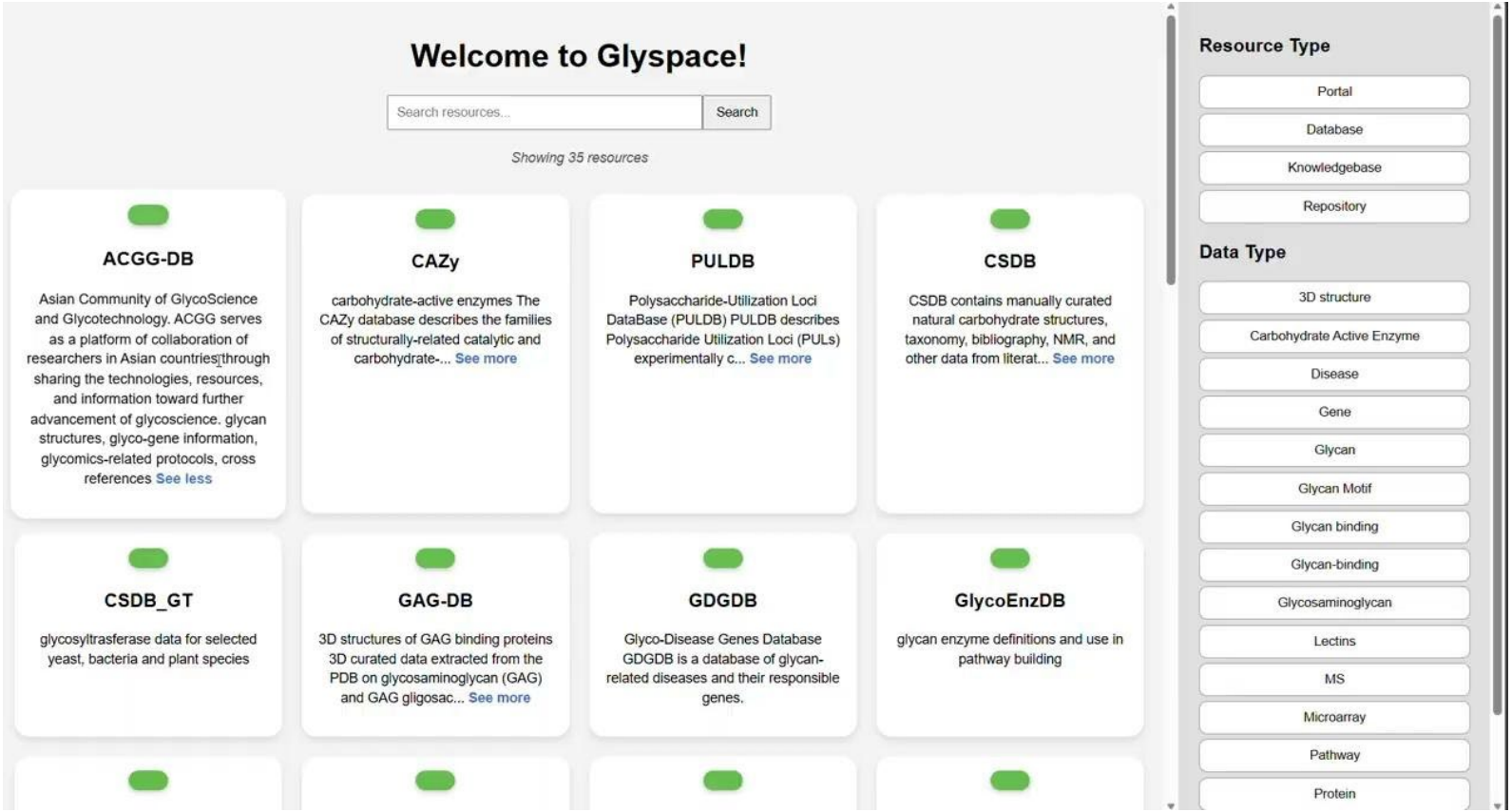


New Prototype

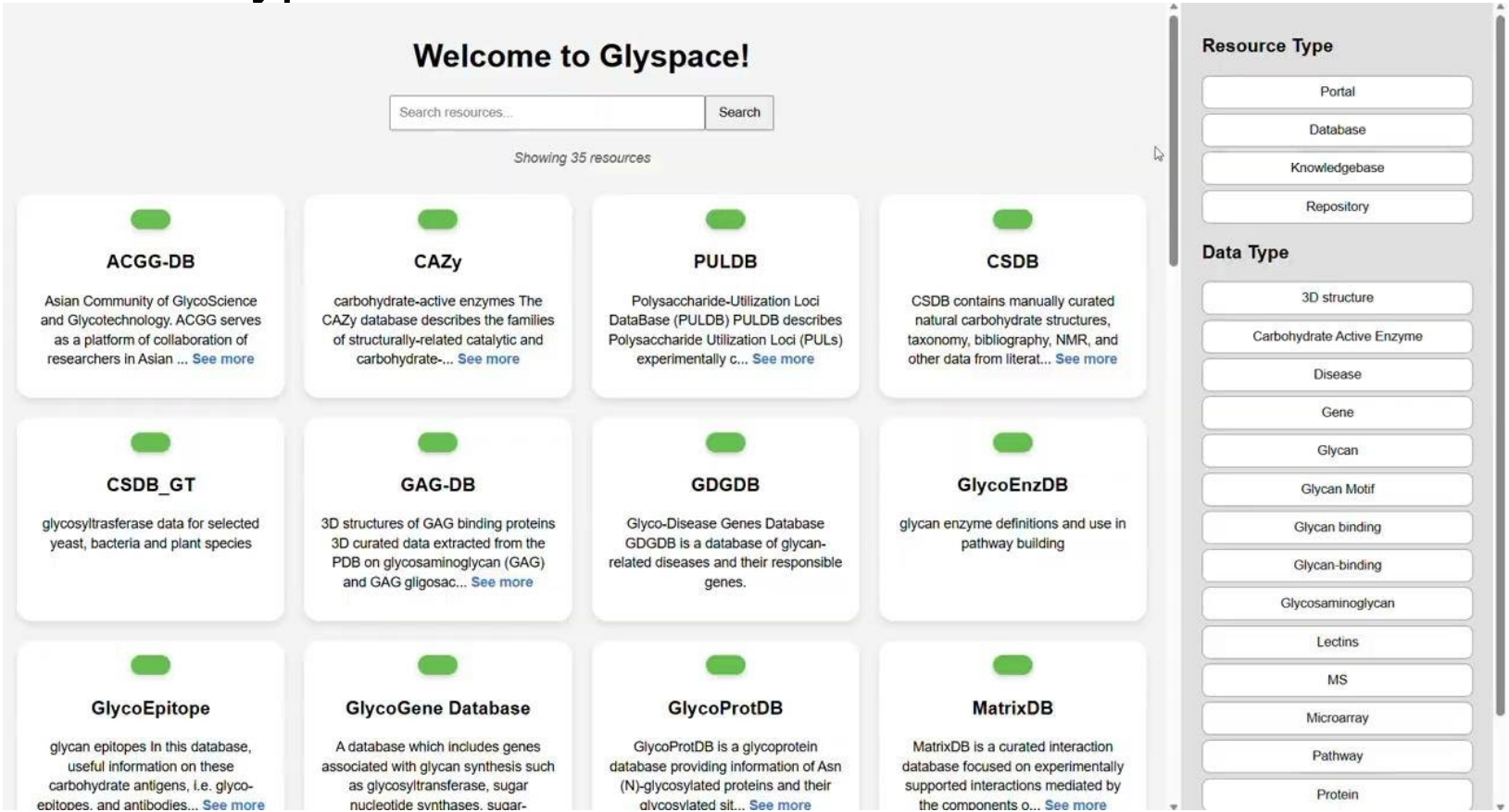
https://glyspace.github.io/glycomics-database-browser/



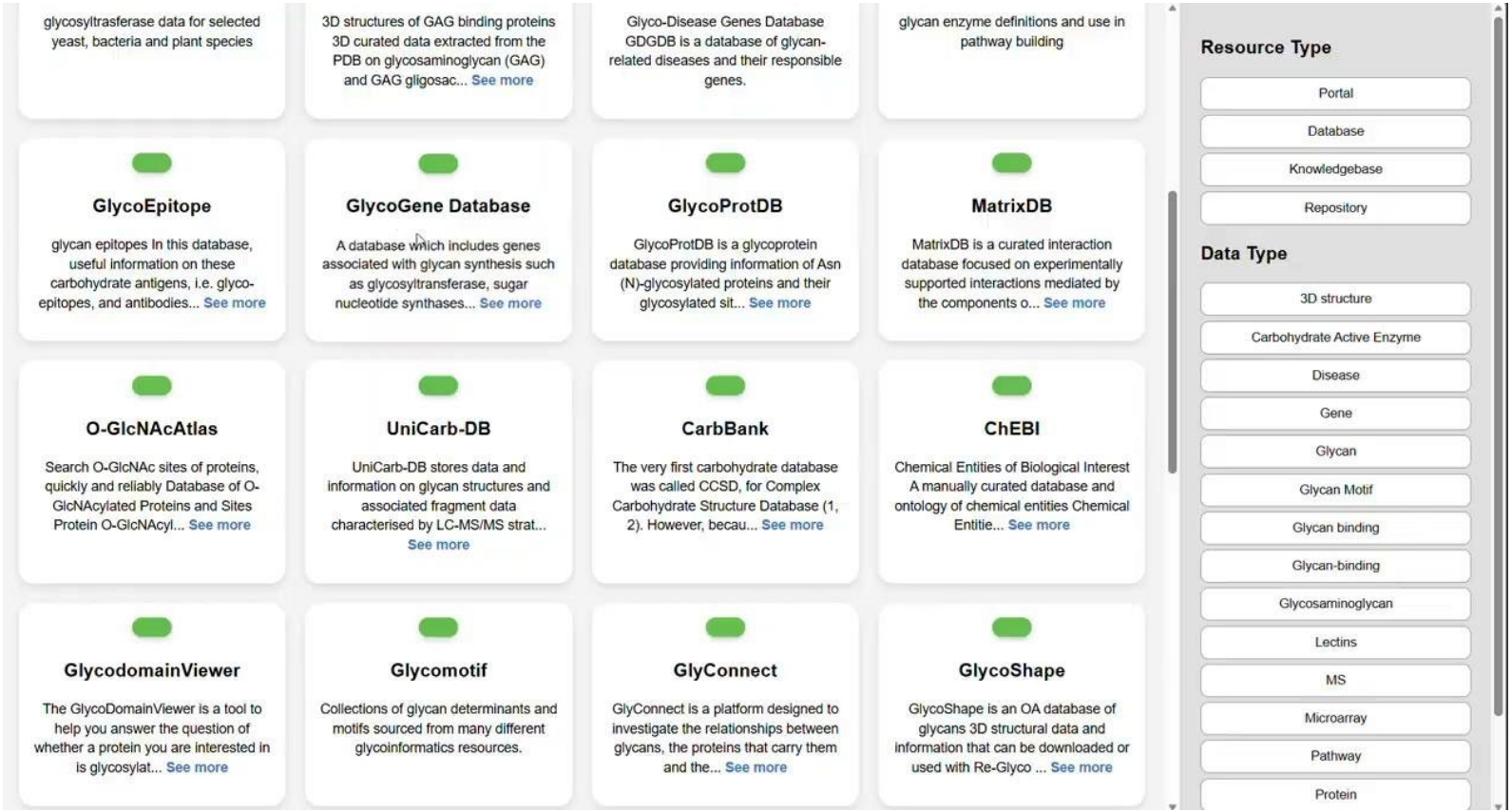
New Prototype Features



New Prototype Features



New Prototype Functionality



Challenges

- Redesigning
  - Initial brainstorming process, choosing to pivot, different ideas to solve problems
- Standardizing data types
  - Reducing number of filters to make prototype more user friendly
- Dependencies to review excel sheet
  - Casing, repeated entries, spacing issues

Next Steps

Short Term:

1. Color coding/icons
2. Using same widget to displaying software tools instead of databases
3. Review other curated columns
4. Receiving feedback (not working, missing resource)

Long Term:

1. Integrating AI to improve search bar to use human language instead of keywords
2. Implementing into webpage
3. Receiving feedback

Acknowledgements

Dr. Raja Mazumder, GW

Dr. Rene Ranzinger, UGA

Urnisha Bhuiyan, GW

Sujeet Kulkarni, UGA

Dr. Frederique Lisacek, SIB

FDA-ARGOS Spring 2026 Volunteership

Venya Gulati

Introduction

- Venya Gulati
- School Without Walls HS
- Interested in biology and public health
- The ARGOS database is a curated collection of pathogen genomes designed to support research and diagnostics

Objectives

1. **Evaluate BioProjects**

Review large datasets of pathogen genome bioprojects

Determine whether they met quality and reliability standards

1. **Identify Emerging Pathogens**

Monitor outbreak reports and journals

Identify pathogens not currently in the ARGOS database

Methods & Materials

Sources Used:

- NCBI BioProject/BioSample/SRA databases
- CDC outbreak reports
- CIDRAP disease monitoring reports

Tools Used:

- Excel/Spreadsheets: sorting + reviewing large datasets
- NCBI databases: verifying assemblies and reads
- Scientific news sources: tracking emerging disease outbreaks

Results: BioProject Review

BioProject	BioSamples	SRA_Reads	Collection_Dates	Platform	Notes	Verdict
PRJNA1027483	586	587	2022-2024	Illumina	Greninger Lab; 2 publications	Use
PRJNA1029161	318	316	2021–2022	Illumina	Greninger Lab; Assemblies are complete genomes	Use
PRJNA1048457	3416	3315	2023–2025	Nanopore GridION	Large dataset;	Use
PRJNA1090887	48	48	2023	Illumina	Peer-reviewed publication (2024)	Use
PRJNA1138732	2	2	2018	Illumina	Small dataset; older collection date	Caution
PRJNA1173516	91	291	2024	Illumina	Volunteer cohort; no publications	Use
PRJNA1223358	54	54	2024	Illumina	2 grants	Use
PRJNA1249236	10	10	2023	Nanopore MinION	No publications	Use
PRJNA1271732	229	229	2024	Nanopore + Illumina prep	1 grant	Use
PRJNA257008	195	148	2014–2015	Illumina	1 publication (2016), Only four complete genomes, Pretty old	Caution
PRJNA507154	11	11	2013–2015	Illumina	Old dataset	Caution
PRJNA671738	58	58	2020	Illumina	Clinical samples	Use
PRJNA701833	5	7	2020	Illumina	2 grants; 5 assemblies (all complete genomes); Rhinovirus	Use
PRJNA917703	73	73	2022–2023	Illumina	51 assembles (complete genomes); ; Rhinovirus	Use
PRJNA939200	30	29	2010–2011	Illumina	Very old samples	Exclude
PRJNA338014	968	965	2016-2018	Illumina	25 publications; 81 assemblies (not complete genomes) 1 reference but scaffold	Caution
PRJNA939712	267	267	2009–2018	Illumina	2 publications (one from 2025); Wide range	Caution
PRJNA1110763	94	94	2023–2024	Illumina	1 publication (2025)	Use
PRJNA1293457	216	150	2025-2026	Illumina	Use samples with reads only	Use
PRJNA1308108	65	60	1961–1990+	Illumina	Extremely old samples	Exclude
PRJNA227457	307	224	2001–2003	Illumina	Very old dataset	Exclude
PRJNA231221	1652	3657	2013	Illumina	6 grants; One publication (2019); Lots of reference genomses; High quality but old	Caution
PRJNA490509	51	49	Not applicable	Illumina	Two publications (2016 & 2017)	Use
PRJNA1066815	1020	1010	2021–2022	Illumina	Large dataset; mostly complete genomes; Rhinovirus	Use
PRJNA1107609	40	53	2023	Illumina	No publications but recent	Use
PRJNA1224818	75	75	2012–2015	Illumina	Older collection dates	Caution
PRJNA927100	16	16	2020	Illumina	1 assembly (complete genome)	Use
PRJNA930027	286	286	2010s	Illumina	255 assemblies (many complete genomes); Older data, could use more recent ones	Caution
PRJNA1017431	62	61	2018–2020	Illumina	2 publications (2024)	Use
PRJNA1142891	8	8	2024	Illumina	No publications	Use
PRJNA723895	1182	1175	2021–2022	Illumina	One publication; Three complete genome assemblies; Rhinovirus	Use
PRJNA613697	1	2	2019	Illumina	Looks good	Use
PRJNA779483	1	37	2018–2019	Illumina	1 complete genome assembly; Slightly old; Lots of reads	Use
PRJNA824010	186	186	2018	Illumina	Slightly old but usable	Use
PRJNA939713	127	127	2010-2015	Illumina	Rhinovirus; Old collection dates	Caution
PRJNA904288	256	181	2022	Illumina	Recent dataset	Use
PRJNA907066	57	70	2021–2022	Illumina	One publication (2023)	Use
PRJNA907865	35	38	2010–2022	Illumina	Use recent samples; One publication (2024); Rhinovirus	Conditional
PRJNA394142	212	253	2017	Illumina	No publications; A little old	Caution
PRJNA737147	23	15	2012	Illumina	Two publications (2023); Two grants; Very old however	Caution
PRJNA1063190	1	1	2023	Illumina	Recent dataset	Use
PRJNA1066385	12	12	2018–2019	Illumina	Reads published 2024	Use
PRJNA1079728	15	30	2020–2022	Illumina	One publication (2024)	Use
PRJNA1098415	18	18	2016–2023	Illumina	Use recent samples only	Conditional
PRJNA1119711	1	1	2023	Illumina	Very small dataset	Use
PRJNA1134214	21	21	2002–2024	Nanopore	Missing metadata for bio samples; filter samples (some have data + host)	Conditional
PRJNA1144058	9	9	2023	Illumina	Recent dataset	Use
PRJNA257197	2	0	2014	missing	No reads present	Exclude
PRJNA344504	115	225	2016	Illumina	One publication (2017)	Use
PRJNA348838	3	3	2015	Illumina	Older dataset	Use
PRJNA360286	36	36	2014	Non-Illumina	Non-human host	Exclude
PRJNA1147890	1	2	2024	Illumina + Nanopore	Use Illumina read; Monkeypox	Use
PRJNA348754	1	1	2011	Illumina	Three publications; Two chromosome assembly only	Exclude
PRJNA849962	714	636	2022	Illumina	467 assemblies; Many complete genomes; Large dataset	Use
PRJNA851991	2	4	2022	Nanopore + ION Torrent	2 assemblies (one complete genome one chromosome)	Caution
PRJNA862948	364	391	2023–2025	Illumina	Large dataset; 358 assemblies (complete genomes)	Use
PRJNA886998	47	47	2022	Illumina	47 assemblies; Two complete genomes	Use
PRJNA931347	2	0	2014–2016	missing	No reads; One publication (2023)	Exclude
PRJNA935252	168	166	2024–2025	Illumina	54 assemblies; Many complete genomes	Use
PRJNA949408	1	1	2022	Illumina	One chromosome assembly only	Caution
PRJNA1077994	140	404	2021	Illumina	Some missing metadata; Rhinovirus	Conditional
PRJNA1170175	24	24	2024	Illumina	Recent publication (2025)	Use
PRJNA396064	126	125	2012-2015	missing	Older dataset; 224 assemblies (mostly complete genomes); One publication (2018)	Caution
PRJNA485481	446	0	missing	missing	No reads; 14,891 assemblies; 6113 publications (all from late 1900's)	Exclude
PRJNA666535	896	1779	2014-2015	Illumina	Metadata missing for many; One publication (2021); Old	Caution
PRJNA679286	128	227	1980	Illumina	Extremely old; 2 assemblies (complete genomes); One publication (2021)	Exclude
PRJNA763062	1	50	2020	Illumina	One complete genome assembly	Use
PRJNA883266	43	43	2016–2017	Illumina	Moderately old; 2 grants; 27 assemblies (all complete genomes)	Use
PRJNA956591	84	84	2024	missing	Non-human hosts	Exclude
PRJNA1034094	38	38	2019–2020	Illumina	Looks good	Use
PRJNA1039588	2	2	2009–2011	Illumina	Very old; One publication	Caution
PRJNA1071434	1	1	2022	Illumina	One publication (2024)	Use
PRJNA1102161	26	26	2009–2018	Illumina	Use recent samples	Conditional
PRJNA1102312	2	2	2020	Illumina	One publication (2020)	Use
PRJNA1111284	19	6	2020	Illumina	Limited reads	Use
PRJNA1158579	120	1	2024	Illumina	Looks fine; Few reads	Use
PRJNA1210160	24	24	2023–2024	Illumina	Recent dataset	Use
PRJNA317709	missing	0	2010	missing	Four assemblies; No reads or complete genomes	Exclude
PRJNA362284	220	35	missing	missing	Missing metadata; difficult to use; One publication; 113 assemblies (12 complete)	Exclude
PRJNA386147	4	4	missing	PacBio	One publication; 1 complete genome assembly; missing collection dates	Exclude
PRJNA436552	565	258	2018	Illumina	Use samples with reads; Four grants; One publication	Use
PRJNA767784	38	38	2016–2017	Illumina	Complete genomes; Two grants; Three assemblies (complete genomes)	Use
PRJNA830667	20	20	2013–2018	Illumina	Use newer samples; 7 assemblies (complete genomes)	Conditional
PRJNA900047	6	2	2019	Nanopore	Very few reads	Caution
PRJNA944717	2	2	2009-2010	Illumina	Too old	Caution
PRJNA952833	3	3	2019–2020	Illumina	Recent enough	Use
PRJNA970478	5	5	2020	Illumina	Looks good	Use
PRJNA9994731	123	123	2021–2022	Illumina	Looks good	Use

Columns: BioProject, BioSamples, SRA_Reads, Collection_Dates, Platform, Notes, Verdict

Criteria

- # of reads
- # of assemblies
- date
- publications

Examples

PRJNA1090887

- 43 biosamples, 43 reads
- 2023
- peer-reviewed publication (2024)
- USE

PRJNA1308108

- 65 biosamples, 60 reads
- 1961-1990 (extremely old)
- EXCLUDE

Results: BioProject Review

Total: 88 BioProjects evaluated ~ 2,314 SRA runs

Sequencing was predominantly Illumina based

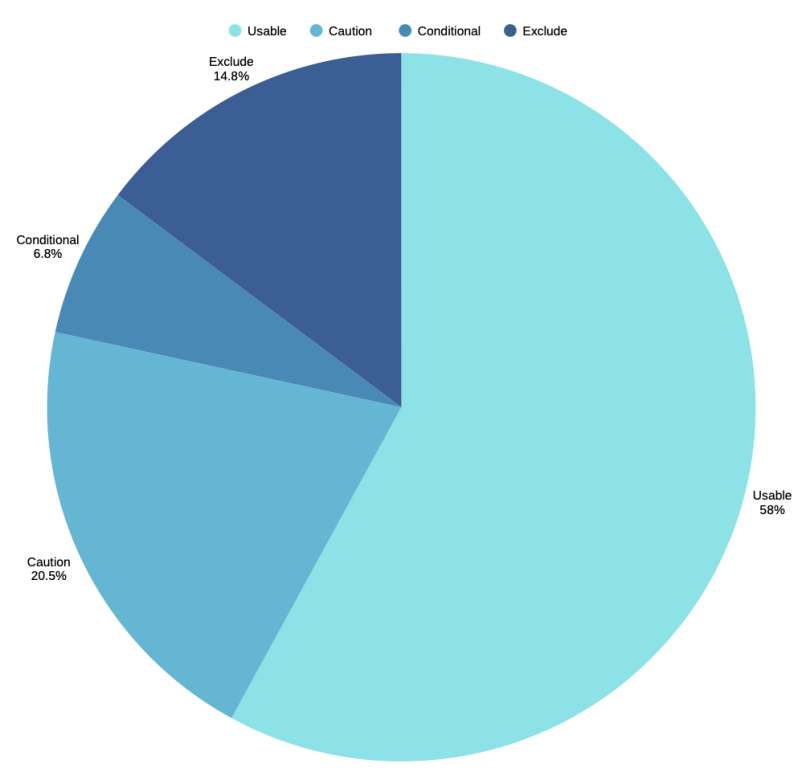
62 published after 2015

Usable (51): Datasets meet criteria

Caution (18): May be usable but require additional validation

Conditional (6): Could be used if certain limitations are addressed

Exclude (13): Do not meet necessary criteria



Results: Emerging Pathogens

Measles morbillivirus (Measles)

- Resurgence of measles cases
- About 588 cases in January 2026
- Biggest outbreak in South Carolina
- Affects mostly unvaccinated people

Henipavirus nipahense (Nipah virus)

- No known cure
- Outbreaks in India (West Bengal's first outbreak since 2007)
- Recurring in South Asia every year (but current cases have drawn attention b/ c of severity)
- Zoonotic

Salmonella Typhimurium & Salmonella Newport

- New outbreak strain
- From contaminated moringa leaf powder
- 65 people infected have been reported from 28 states (as of 1/ 29/ 2026)
- Likely much higher # of infected people

Cochliomyia hominivorax (New World Screwworm)

- Animal infestations reported near the US-Mexico border
- 1,190 cases and 7 deaths in Central America/ Mexico (as of 1/ 20/ 2026(
- No infestations in US yet, however potential for geographic spread
- Transmitted when parasitic flies lay eggs in wounds/ other body cavities
- Previously eradicated in the US by releasing sterile male flies to mate with female NWS fly

Paenibacillus dendritiformis

- Infecting infants with severe neurologic symptoms
- 2 infant cases so far
- Antibiotics may be inadequate
- Gram-positive spore-forming bacterium found in the environment

BioSamples Not Currently in ARGOS

Organism	Taxonomy ID	Assembly Accession	SRA Run(s)	BioSample	Assembly Level	Location	Year	In ARGOS?
Measles morbillivirus	11234	MISSING	SRR37654410	SAMN56533323	MISSING	USA	2026	No
Henipavirus nipahense	3052225	MISSING	SRR35257015	SAMN50880225	MISSING	Malaysia	2025	No
Escherichia coli		MISSING	SRS28576603	SAMN56777044	MISSING	USA	2026	No
Influenza A virus	11320	GCA_054438175.1	SRR36202258	SAMN53405355	Complete genome	USA	2026	No
Influenza A virus	11320	GCA_055463195.1			Complete genome	USA	2026	

- Limited results
- Many already in ARGOS
- Missing reads or assemblies

Skills and Knowledge Gained

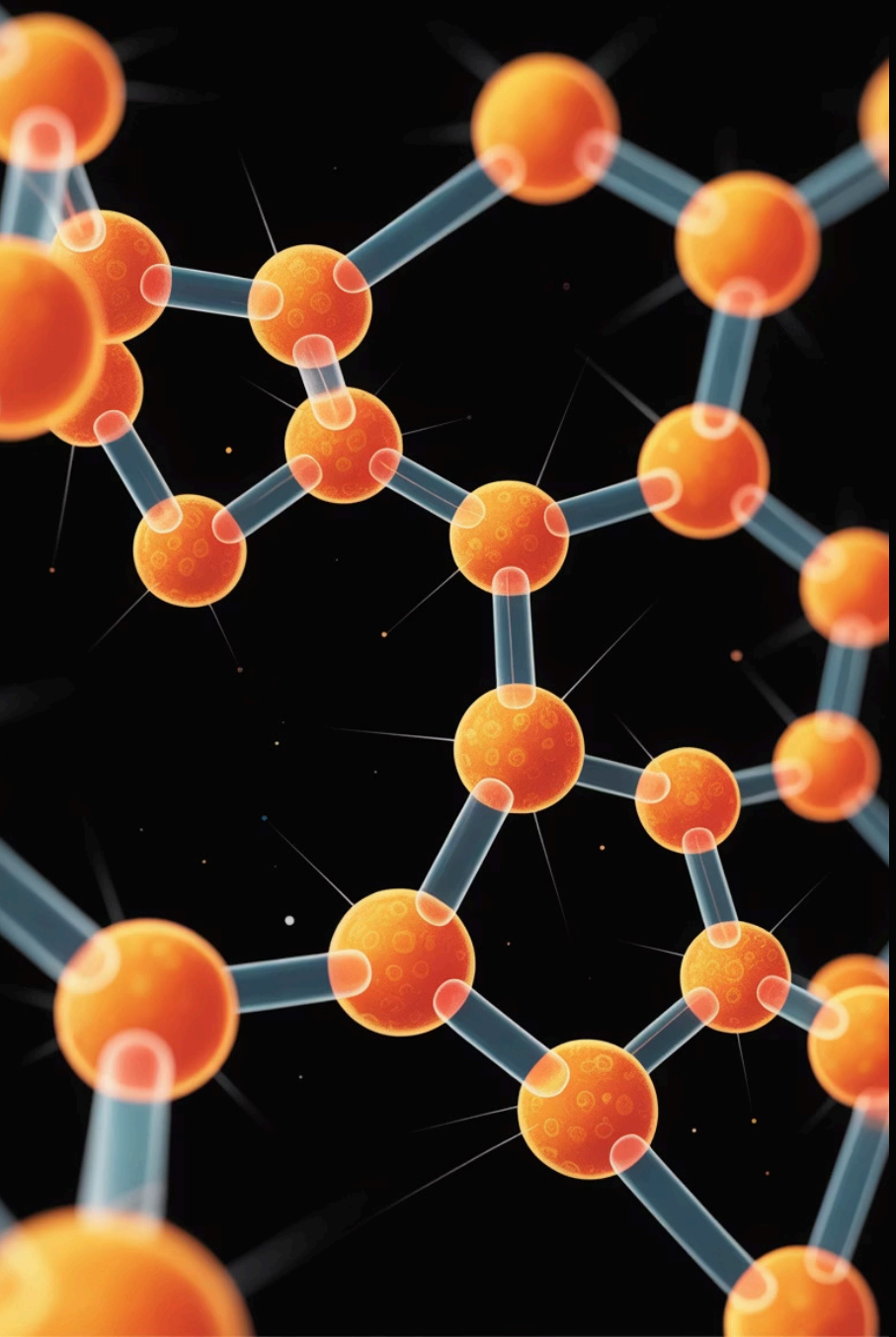
- Technical skills: navigating NCBI databases, reviewing BioProject metadata, organizing large datasets in spreadsheets
- Scientific knowledge: learning about emerging diseases and understand genome sequencing data
- Research skills: monitoring sources for public health information

Acknowledgements

Mentor: Christie Woodside

Lab: Dr. Mazumdar’s Lab

FDA and BARDA



BiomarkerKB: Glycan Biomarker Extraction Pipeline

A Common Fund Data Ecosystem (CFDE) research initiative focused on standardizing and harmonizing glycan biomarker data to build a comprehensive knowledgebase for discovery across cancer, endometriosis, neurodegeneration, and inflammation research, with LLMs used as one tool to support extraction and organization.

RECORDED PRESENTATION - VISHAL MUTHUSEKARAN - SPRING 2026 VOLUNTEERSHIP

The Standardization of Glycan Entities

Glycans are biologically important, but unlike genes and proteins, they do not yet have widely adopted standard nomenclatures. Their meaning also changes with context, including the carrier molecule, tissue, and specimen type, making consistent interpretation across studies especially challenging.

The Problem

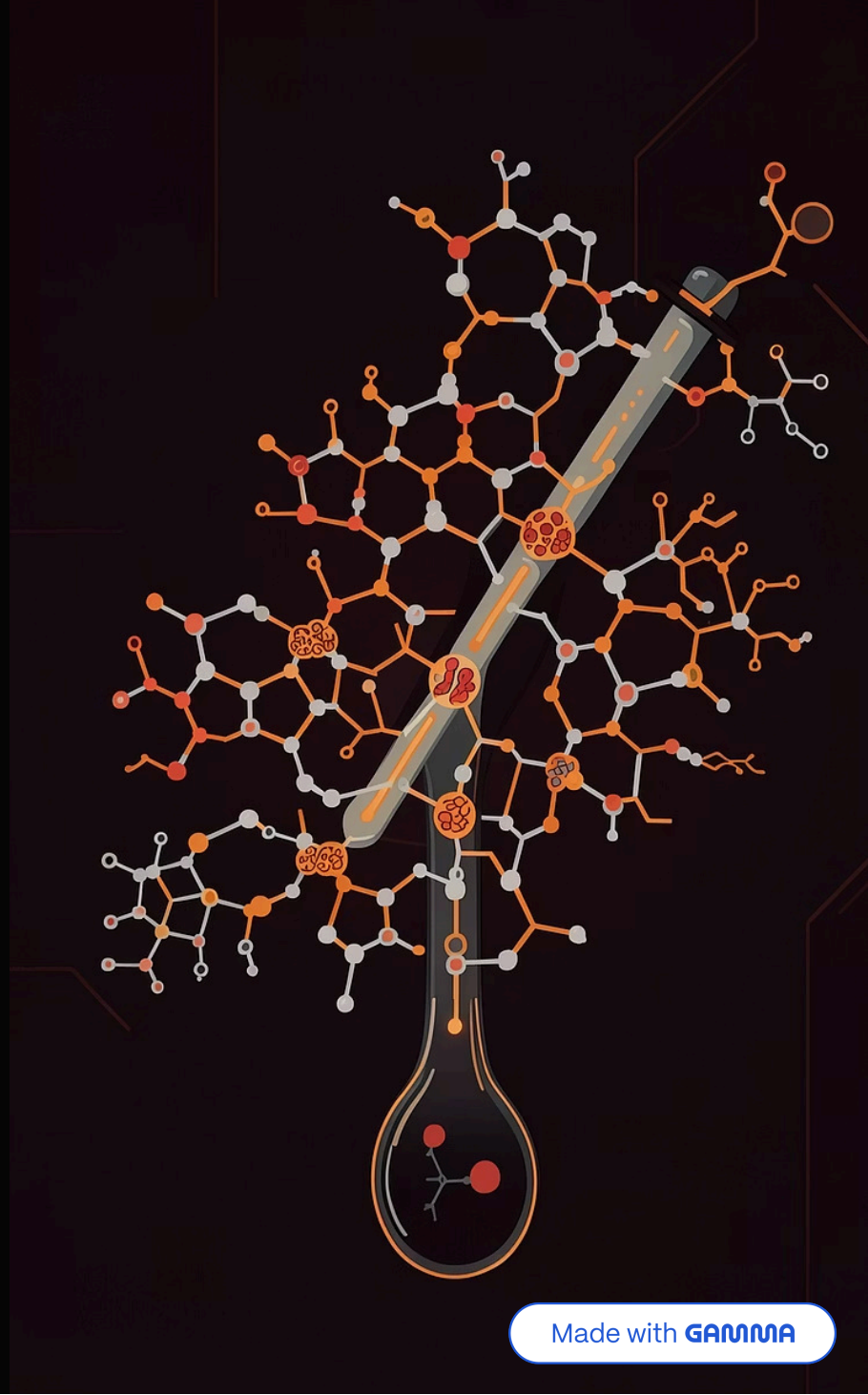
Glycan entities lack a common naming standard, so the same structure may be described differently across resources. Their biological function is also highly context-dependent, varying by protein or lipid carrier, tissue, and specimen type.

The Solution

BiomarkerKB's data model standardizes glycan entities by capturing both the glycan structure and the biological context in which it appears, enabling consistent comparison, curation, and downstream analysis.

The Impact

Researchers can compare glycan biomarkers across datasets with clearer definitions, improved consistency, and better support for cross-study discovery in cancer, endometriosis, neurodegeneration, and inflammation.



CONTRIBUTION 1

The Standard Dataset: 25 Curated Papers

We started with a TSV file that flagged positive and negative glycan biomarker entries. From the positive-flagged papers, we manually curated 25 high-quality peer-reviewed studies into a structured benchmark dataset, creating the definitive Golden Standard ground truth for evaluating LLM extraction performance. Each entry captures the full biological context of a glycan biomarker finding.

1

Evidence & Notes

Concise, paper-sourced evidence text and curator notes for full transparency.

2

Specimen & Carrier

Biological sample type (e.g., serum, tissue) and the molecular carrier (protein/lipid).

3

Direction & Condition

Whether the biomarker is up/down-regulated and the associated disease or condition.

4

Glycan Type & Entity

Specific glycan structure and the biomarker entity identified in the study.

5

PMID & Title

Unique PubMed identifier and paper title for traceability and citation.

25 Curated Papers Dataset

50 Biomarker Set

OBJECTIVES & GOALS

What We Set Out to Build

Primary Objectives

- Curate a high-quality **Golden Standard dataset** of glycan biomarker papers as ground truth
- Develop an **LLM-based extraction pipeline** to automate biomarker identification from literature
- Create a **rigorous evaluation framework** to benchmark and score LLM outputs
- Build a **sustainable, cost-efficient backend** to host models and data without paid APIs

Why These Goals Matter

Manual curation of biomarker literature is an invaluable resource for establishing ground truth. By combining expert-curated data with systematic standardization and automated extraction tools, BiomarkerKB creates a **reproducible, extensible pipeline** that accelerates glycan biomarker discovery and ensures consistency across studies.

The scoring rubric ensures extraction quality is **measurable and comparable** — a critical requirement for any knowledgebase intended for scientific use.

100-Point LLM Scoring Rubric

A glycan biomarker is a measurable glycan structure, glycoform, or glycosylation pattern where the change is associated with a disease state, disease progression, treatment response, or condition.

Total Possible Score: 100 Points

1

CORE CLAIM ACCURACY

40 POINTS

- Entity (25 pts): Must match text exactly or a clear synonym
- Direction (10 pts): Must match reported change
- Condition (5 pts): Must match disease/context

Note: Wrong entity = entry score = 0

2

BIOLOGICAL CONTEXT

30 POINTS

Formula: Context Score = 30 × (correct context fields / total context fields present)

- Correct N/A counts as correct
- Incorrectly filled unsupported fields = 0 for that field
- Includes: specimen type, tissue source, biofluids, molecular carrier, glycan class, assay matrix, organism source

3

EVIDENCE QUALITY

30 POINTS

- Must support the claim (entity + direction + condition)
- Must be concise (<75 words)
- Must be grounded directly in source text
- Factual accuracy (10 pts)
- Text grounding (10 pts)
- Conciseness (10 pts)

4

PENALTY SYSTEM

Applied across the rubric

- Hallucinated entity: Entry score = 0
- Fabricated identifier: −15 pts
- Unsupported filled field: −5 pts each
- Reversed direction: −10 pts

Identifier Field

5-point application rule: Correct ID provided = 5 pts; correctly blank (no ID) = 5 pts; missing but present in source = 3 pts; formatting issues = 2–4 pts; fabricated = −15 pt penalty.

Scoring Implementation: Standard Evaluation

The scoring rubric is important because it lets us objectively evaluate LLM extraction quality, identify specific weaknesses like hallucinations, reversed directions, and poor evidence grounding, and compare models fairly. It also drives targeted improvements and establishes a trustworthy benchmark for automated curation. Without systematic scoring, we cannot trust or improve the LLM extraction pipeline.

The 100-point rubric was applied to each of the 50 curated papers in the Standard dataset. Each entry was systematically scored across the four categories: Core Claim Accuracy, Biological Context, Evidence Quality, and Identifier Field, with penalties applied for hallucinations or errors.

Representative examples from the Standard with their scores:

40843950	Nephropathic Cystinosis (Sialylation altered) (Usable with light review)	70
40890283	Oral squamous cell carcinoma (Disease-specific N-glycopeptides) (Publication-quality)	82
40914662	PDAC (O-GalNAc on MUC1) (Publication-quality)	86
40844495	Salivary duct carcinoma (Tn-MUC1) (Publication-quality)	87
40812640	PDAC (Core-fucosylated glycoproteins) (Publication-quality)	91
40772298	VTE (Plasma N-Glycan) (Publication-quality)	84
40634609	Down syndrome (O-GlcNAcylation) (Usable with light review)	68
40602089	COVID-19 (Serum N-Glycome) (Publication-quality)	94

Scores ranged from 68-94, demonstrating the rubric's ability to differentiate between usable and publication-quality entries.

25 Curated Papers Dataset **50 Biomarker Set**

CONTRIBUTION 3

LLM Development & Extraction Experiments

Models Tested

DistilGPT-2

Lightweight transformer fine-tuned for biomarker entity extraction with custom prompts.

Llama-6

Larger language model tested for complex glycan context understanding and relation extraction.

Experiment Workflow

- **Prompt Engineering:** Iteratively refined prompts to maximize field-level extraction accuracy
- **Local API Testing:** Ran models on local servers to evaluate latency and output quality
- **Cost-Reduction Trials:** Explored self-hosted alternatives to eliminate dependency on paid APIs
- **Benchmark Scoring:** Applied the 100-point rubric to compare model outputs against Golden Standard

CONTRIBUTION 5

Expanded Base Dataset: 50 Additional Papers

Cyrus provided a repository of LLM-curated results from **50 papers**, which I then reviewed and graded using the **100-point scoring rubric** to evaluate model performance. This 50-paper set served as the expanded base dataset for testing and future model iterations, broadening the training and evaluation corpus beyond the original 25-paper Golden Standard.

25

Golden Standard Papers

Fully curated, structured ground truth benchmark for LLM evaluation

50

Base Dataset Papers

LLM-curated results reviewed and graded to expand the training and testing corpus

75

Total Papers Processed

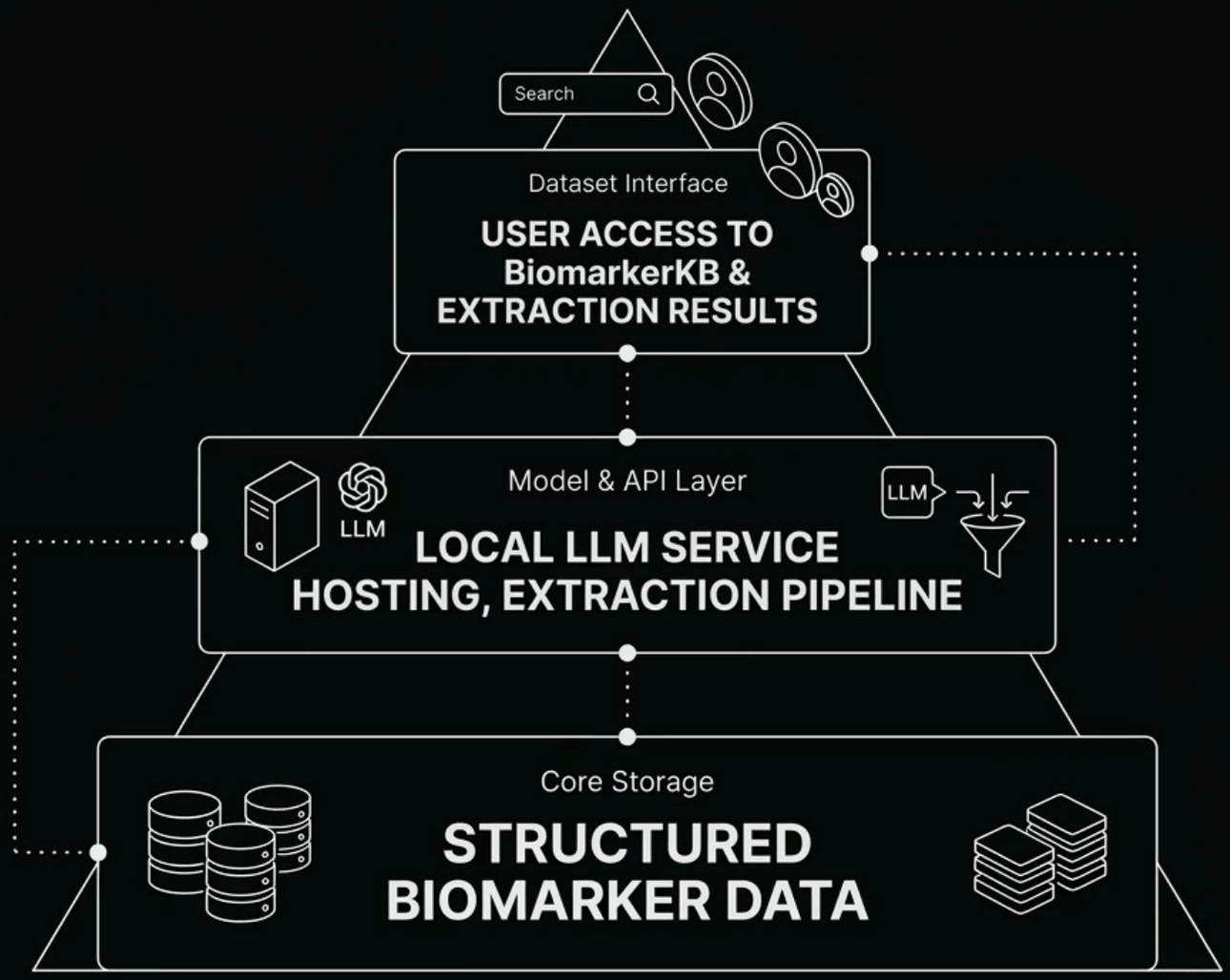
Combined dataset powering the full BiomarkerKB glycan extraction pipeline

100

Scoring Rubric Points

Comprehensive evaluation framework used to assess extraction quality and consistency

Future Directions: System Architecture & Deployment



Note: This represents the system design vision for scaling the project. Implementation of model serving and user access interface is part of future development work.

LOOKING AHEAD

Future Plans & Roadmap

Phase 1: Scale Curation

Expand the Golden Standard dataset beyond 25 papers to cover more disease contexts and glycan types, improving benchmark robustness.

1

Phase 2: Model Refinement

Fine-tune DistilGPT-2 and Llama-6 on the expanded dataset, incorporating rubric scores into a feedback loop for iterative improvement.

2

Phase 3: Backend Deployment

Finalize the Supabase + Viceroy pipeline and deploy a stable, self-hosted LLM API accessible to CFDE collaborators.

3

Phase 4: Knowledgebase Launch

Integrate extraction outputs into the full BiomarkerKB platform, making glycan biomarker data queryable for the broader research community.

4

REFLECTIONS & ACKNOWLEDGMENTS

Skills Gained & Thank You

Skills & Knowledge Gained

- **LLM Fine-Tuning** — Hands-on experience with DistilGPT-2 and Llama models
- **Prompt Engineering** — Designing and iterating prompts for structured extraction
- **Bioinformatics** — Deepened understanding of glycan biology and biomarker curation
- **Backend Development** — Supabase, local server setup, and API design
- **Scientific Communication** — Translating research into structured, evaluable data

Acknowledgments

This project was made possible through the support and collaboration of many individuals:

- **Cyrus** — Close collaborator on glycan biomarker curation and project direction
- **Mentors & Advisors** — For guidance, evaluation frameworks, and scientific rigor
- **Research Team** — For feedback, code reviews, and ongoing collaboration

Questions? Thank you for your attention — happy to discuss any aspect of this work in more detail.

Email: vishal.muthusekaran@gmail.com

Thank You!

BiomarkerKB Biocuration Project

Conner Cognata · Spring 2026 Bioinformatics Volunteership (Symposium)

BIOINFORMATICS | HEALTHCARE DATA | AI / ML

What Is BiomarkerKB?

BiomarkerKB is a structured database that centralizes biomarker information to support biomedical research and enable future AI applications.

Biomarker: A defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions. - NIH/FDA

biomarkerKB

→ The Knowledge Base

This is the "brain" of the operation. It gathers unstructured data from scientific literature, clinical registries, and genomic databases.

→ Biomedical Research & Clinical Trials (Bio)

Once organized, this data becomes a tool for scientists and clinicians for discovery, trials, and validation.

→ Frontend Visualization

The frontend acts as the user interface, translating complex data into actionable insights with dashboards and graphics.

Real-World Applications

- **Treatment Optimization (Cystic Fibrosis):** By monitoring **Sweat Chloride** and **FEV1** levels, doctors can determine within weeks if expensive medications like Ivacaftor are working. If chloride levels don't drop, the therapy can be adjusted immediately, saving the patient time and preventing lung scarring.
- **Therapeutic Selection (Oncology):** Before starting chemotherapy for **NSCLC**, pathologists use **Histology-Based Prediction** to sort patients. Since Pemetrexed is highly effective for "Nonsquamous" cells but potentially harmful for "Squamous" cells, this structured data prevents patients from receiving ineffective, toxic treatments.
- **Proactive Risk Management (Cardiovascular):** Using biomarkers like **hs-CRP** and **HbA1c**, clinicians can identify "asymptomatic" high-risk individuals. This allows for early intervention with statins or lifestyle changes years before a potential heart attack occurs.

Objectives & Goals

Curate High-Quality Data

Build validated biomarker datasets from peer-reviewed sources

Map Relationships

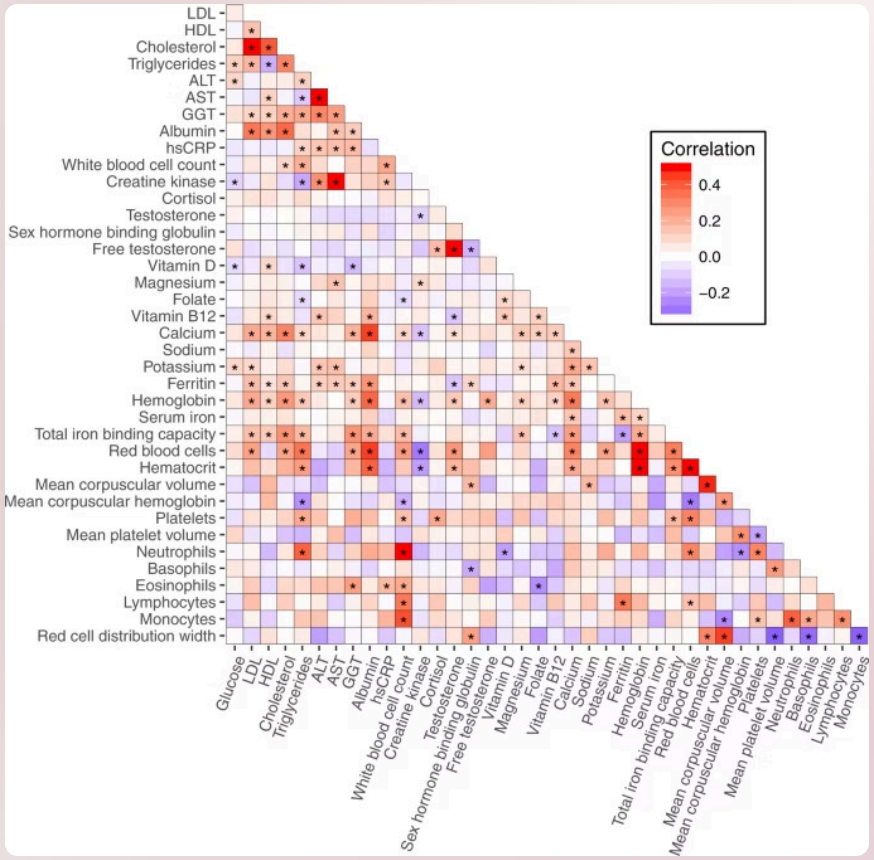
Identify and document biomarker-disease associations

Validate Evidence

Support all entries with PMIDs and DOIs

The Path to AI-Readiness from BMKB to ML

Standardizes and structures complex biomedical data into machine-readable formats.



Methods: How We Curated

The Curation Process (Step-by-Step)

1. **Literature Sourcing:** High-impact clinical articles (provided by Maria) serve as the primary evidence base.
2. **Biomarker Identification:** Systematic scanning of the text for biomarkers that correlate with disease states or drug responses.
3. **Entity Normalization:** Assigning standardized identifiers to each discovery:
   - **UniProt (UPKB)** for proteins.
   - **Disease Ontology (DOID)** for clinical conditions.
   - **LOINC** for laboratory observation codes.
4. **Evidence Linking:** Mapping each biomarker relationship back to its source using **PubMed IDs (PMIDs)** or **DOIs** to ensure scientific traceability.
5. **Role Classification:** Categorizing the biomarker's clinical utility–whether it's for **Risk** (predicting future events), **Monitoring** (tracking drug efficacy), or **Predictive** (selecting the right patient for a drug).

UniProt
Protein identifiers

DOID
Disease ontology

UBERON
Anatomical specimens

Biomarkers Extracted

	Entity	Entity ID	Entity Type	Condition	Condition ID	Exposure Agent	Agent ID	Biomarker Role	Specimen	Specimen ID	LOINC Code	Evidence	Evidence Source
increased CRP	C reactive protein	UPKB:P02741	Protein	Coronary Artery	DOID:3393	-	-	Risk	Serum	UBERON:0001977	30522-7	Rosuvastatin significantly redi	PMID:18997196
increased HbA1c	Hemoglobin A1c	UPKB:P68871	Protein	Coronary Artery	DOID:3393	-	-	Risk	Whole blood	UBERON:0000178	4548-4	Measurement of hemoglobin ,	PMID:18997196
increased urine albumin	Albumin	UPKB:P02768	Protein	Atherosclerosis	DOID:1936	-	-	Risk	Urine	UBERON:0001088	14959-1	In asymptomatic adults with h	PMID:18997196
increased hs-CRP (high-sensitivity C-reactive pr	C-reactive protein	increased hs-CRP (high-sensitivity C-reactive	C-reactive protein	UPKB:P02741	protein	cardiovascular	DOID:1287	risk (predictive)	blood serum	UBERON:0001977		"hs-CRP retains an independe	DOI:10.1161/01.CIR.0000052939.59093.45
increased serum amyloid A (SAA)	Serum amyloid A	increased serum amyloid A (SAA)	Serum amyloid A-1	UPKB:P0DJI8	protein	cardiovascular	DOID:1287	risk (predictive)	blood serum	UBERON:0001977		"newer studies of hs-CRP and	DOI:10.1161/01.CIR.0000052939.59093.45
increased fibrinogen	Fibrinogen	increased fibrinogen	Fibrinogen alpha chain	UPKB:P02671	protein	cardiovascular	DOID:1287	risk (predictive)	blood plasma	UBERON:0001977		"prospective epidemiological s	DOI:10.1161/01.CIR.0000052939.59093.45
increased white blood cell count	White blood cell count	increased white blood cell count	White blood cell count	?	clinical_measur	cardiovascular	DOID:1287	risk (predictive)	blood	UBERON:0001977		"prospective epidemiological s	DOI:10.1161/01.CIR.0000052939.59093.45
increased interleukin-6 (IL-6)	Interleukin-6	increased interleukin-6 (IL-6)	Interleukin-6	UPKB:P05231	protein	cardiovascular	DOID:1287	risk (predictive)	blood serum	UBERON:0001977		"studies of... interleukin-6... st	DOI:10.1161/01.CIR.0000052939.59093.45
decreased sweat chloride concentration	Chloride	CHEBI:17996	ion	cystic fibrosis	DOID:1485	ivacaftor	PubChemCID:16220172	monitoring biomarker	sweat	UBERON:0001988		The mean change from bas	PMID:23590265
increased percent predicted FEV1	Forced expiratory volume in 1 second (FEV1) percent predicted		clinical_measurement	cystic fibrosis	DOID:1485	ivacaftor	PubChemCID:16220172	monitoring biomarker	lung	UBERON:0002048		Patients receiving ivacaftor	PMID:23590265
Nonsquamous histology	Non-small cell lung cancer (nonsquamous subtype)	DOID:3908	histologic subtype	Non-small cell lu	DOID:3908	pemetrexed	CHEBI:46756	predictive biomarker	tumor tissue	UBERON:0001062	N/A	Nonsquamous patients treate	Scagliotti GV et al., The Oncologist, 2009
Squamous histology	Non-small cell lung cancer (squamous subtype)	DOID:3908	histologic subtype	Non-small cell lu	DOID:3908	pemetrexed	CHEBI:46756	predictive biomarker	tumor tissue	UBERON:0001062	N/A	Squamous patients treated wit	Scagliotti GV et al., The Oncologist, 2009

The map organizes complex clinical data into four primary functional categories based on the "Biomarker Role" defined in the sources:

- **Risk Indicators:** These identify the likelihood of developing a condition. For example, HbA1c and Urine Albumin are mapped as risk assessment tools for cardiovascular disease and atherosclerosis in asymptomatic adults
- **Prognostic Indicators:** These predict the likely course or outcome of a disease. C-reactive protein (CRP) and Serum amyloid A (SAA) are categorized here because elevated levels at hospital admission predict poor clinical outcomes (like myocardial infarction or death) in patients with unstable angina
- **Monitoring Indicators:** These track the effectiveness of a treatment over time. The map uses Sweat chloride concentration and FEV1 (lung function) to show how clinicians monitor a patient's response to the drug ivacaftor for cystic fibrosis
- **Predictive Indicators:** These determine if a specific therapy will be effective for a patient. The map highlights Tumor Histology in lung cancer, where "nonsquamous" histology predicts a survival benefit from pemetrexed, while "squamous" histology predicts a lack of benefit

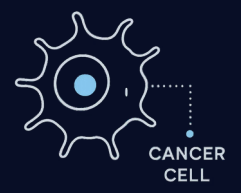
Overview of Personal Discoveries

Cardiovascular Disease

Curated biomarkers linked to atherosclerosis, inflammation, and cardiac risk with validated PMID evidence

Cystic Fibrosis

Structured entries for lung function and treatment monitoring biomarkers



Cancer (NSCLC)

Non-small cell lung cancer biomarkers supporting personalized medicine approaches

Identifying Biomarkers

Step 1: Identification (The "What")

Read the article and look for the Biomarker and the Condition (the disease).

- Action: Highlight the specific protein or lab value the researchers measured.
- Result: You have your "Entity" and your "Condition."

Step 2: Verification (The "Where")

Find the "ID card" for that paper so others can find it.

- Action: Locate the PMID (the 8-digit number on PubMed) or the DOI (the web link).
- Result: This proves your data is backed by real science.

Step 3: Standardization (The "Code")

Give the biomarker and disease a Universal ID so there is no confusion.

- Action: Look up the biomarker in UniProt (e.g., P02741) and the disease in a Disease Ontology (e.g., DOID:3393).
- Result: Even if different doctors use different names, the IDs stay the same.

Step 4: Categorization (The "Why")

Determine how this biomarker is actually used in a hospital.

- Action: Assign a Role. Is it used to find a Risk, to Monitor a patient, or to Predict if a drug will work?
- Result: Your spreadsheet now tells a clinical story.

Achievements & Contributions

1 Extract Biomarker Data (PMID)

I transformed clinical research into a structured database by linking biomarkers to diseases using standardized IDs and verified PubMed evidence.

2 Unstructured Data into Structured

I converted unstructured clinical text into structured data by mapping biomarkers to standardized IDs and verified PubMed evidence.

3 Biological Pattern Discovery

Identified consistent biomarker–disease relationships across studies

4 Integrated Biomarkers into database

I integrated unstructured clinical research into a structured database by mapping biomarkers to standardized IDs and verified PubMed evidence.

I transformed unstructured clinical research into structured, machine-readable data by mapping biomarkers to standardized identifiers and verified PubMed evidence. By identifying recurring patterns across disparate studies, I added significant scientific value and cross-study insights that directly support the growth of the BiomarkerKB database.

Key Findings & Example Biomarkers

What the Data Revealed

Strong, consistent link between **inflammation markers** and cardiovascular disease risk

Biomarkers enable **early disease prediction** before clinical symptoms appear

Patterns were **reproducible across multiple independent studies**

Curated Biomarker Examples

CRP	Cardiovascular risk
HbA1c	Cardiovascular risk indicator
Albumin	Atherosclerosis marker
IL-6, Fibrinogen, WBC	Inflammation markers

Evidence supported by **PMID:18997196** and peer-reviewed DOI studies

Challenges

- Ambiguous Terms Challenge: Papers often use vague names like "blood" or "heart disease."

Solution: Used UBERON and DOID codes to specify the exact sample (e.g., Serum) and the specific condition (e.g., Coronary Artery Disease).

- Missing Identifiers Challenge: Clinical measurements like FEV1 or WBC count don't have protein IDs (UniProt).

Solution: Labeled them as clinical_measurement and used LOINC codes to ensure they remained searchable in the database.

- Non-Molecular Biomarkers Challenge: Some biomarkers are "histology" (cell types) rather than molecules.

Solution: Created a histologic subtype category to capture lung cancer cell types as predictive markers for drug response.

- Fragmented Evidence Challenge: Data is often spread across text, tables, and supplementals.

Solution: Linked every data row to a unique PMID or DOI, creating a "paper trail" that makes the unstructured text fully traceable.

Skills & Knowledge Gained

Technical

- Bioinformatics databases (UniProt, DOID, UBERON)
- Data curation and standardization
- Scientific literature analysis

Analytical

- Pattern recognition across studies
- Evidence validation and cross-referencing
- Structured data quality assessment

Professional

- Critical thinking in scientific contexts
- Scientific communication and documentation
- Collaboration with research teams

Data Modeling & Architecture: I learned how to build a logical framework for biological data. Instead of just making a list, I structured relationships so that every biomarker is correctly linked to its disease and drug, making the data ready for machine learning.

Industry-Standard Curation: I mastered data normalization by translating messy clinical text into universal "scientific codes" (like UniProt and DOID). This ensures the data is "clean" and speaks the same language as global bioinformatics databases.

Scientific Pattern Discovery: I moved from data entry to evidence synthesis. By looking across different studies, I identified recurring patterns–like how various inflammatory markers confirm a single cardiovascular risk–adding real scientific insight to the BiomarkerKB project.

Looking Ahead

1

Expand the Dataset

Grow coverage across additional disease domains and biomarker types

2

Automate with LLMs

Explore large language models to accelerate extraction and reduce manual effort

3

Improve Accuracy

Refine validation pipelines and error-checking protocols

4

Integrate PMID BMKB

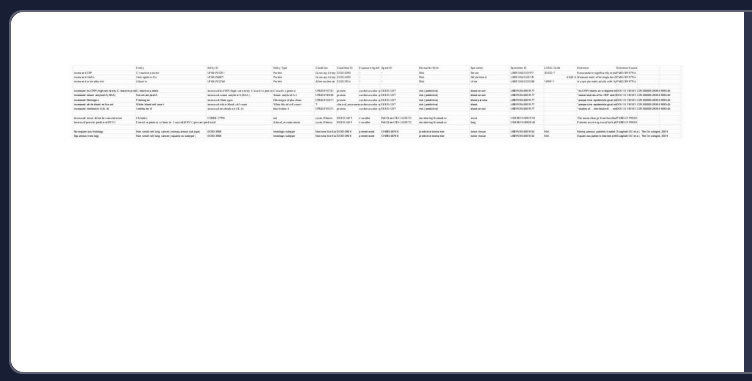
Connect curated data directly to predictive and diagnostic machine learning systems

---

Acknowledgments

Thank you to the **BiomarkerKB Team** – **Maria Kim, Cyrus Yeung, Jeet Vora** – program mentors, and all research contributors who made this work possible.

Questions welcome.



Google Docs

Bio–Marker

PredictMod: PMID Curation for Intervention Outcome Prediction Models

2025 Spring Volunteership Symposium
April 15th 4-6 PM

POC
Lori Krammer

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

HIVE

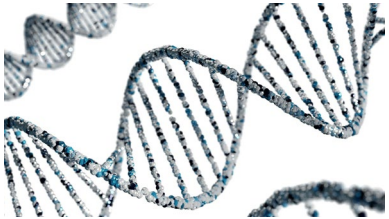
Introduction

Diya Kamalabharathy
Poolesville High School

Ashley Tien
Associate Genomic Analyst
GeneDx

Sampurna Chakravorty
University of Maryland-College Park

The Platform: PredictMod



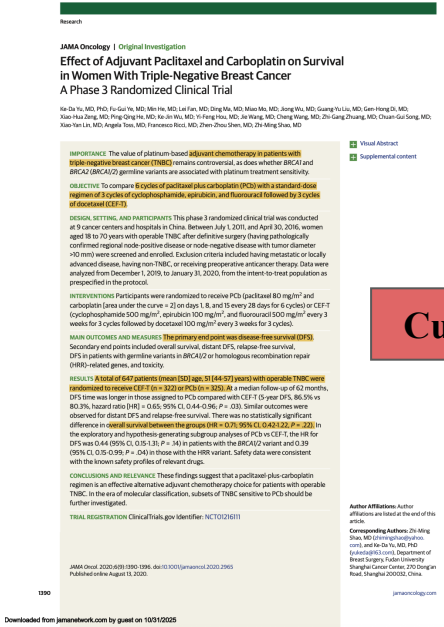
PredictMod is an open-license platform that allows researchers and clinicians to build, use, and share models that predict patient response to treatments.

- **Democratizes AI/ML:** Provides a framework for researchers and clinicians to generate models without needing deep expertise in AI/ML.
- **Optimized for Real-World Research:** Enables pilot studies and early-stage hypothesis generation that can justify larger grants and data collection.
- **Bridges Research and Clinic:** Allows models trained on published data or local datasets to be applied directly to new patient samples for predictive insights.

Our volunteership powers this platform by identifying and validating the publicly available datasets and directly training the ML models that depend on these datasets.

Krammer L, McNeely P, and Bhuiyan U et al. PredictMod: A Platform for Predicting Medical Intervention Outcomes and Sharing Custom ML/AI Models. *NSM*. 2025. Vol. 1(1):57-66. DOI:10.14293/NSM.25.1.0007

Previous Semester



Research

JAMA Oncology | Original Investigation

Effect of Adjuvant Paclitaxel and Carboplatin on Survival in Women With Triple-Negative Breast Cancer
A Phase 3 Randomized Clinical Trial

Ke-Da Yu, MD, PhD; Fu-Gui Ye, MD; Min He, MD; Lei Fan, MD; Ding Ma, MD; Miao Mo, MD; Jiong Wu, MD; Guang-Yu Liu, MD; Gen-Hong Di, MD; Xiao-Hua Zeng, MD; Ping-Qing He, MD; Ke-Jin Wu, MD; Yi-Feng Hou, MD; Jie Wang, MD; Cheng Wang, MD; Zhi-Gang Zhuang, MD; Chuan-Gui Song, MD; Xiao-Yan Lin, MD; Angela Toss, MD; Francesco Ricci, MD; Zhen-Zhou Shen, MD; Zhi-Ming Shao, MD

IMPORTANCE The value of platinum-based adjuvant chemotherapy in patients with triple-negative breast cancer (TNBC) remains controversial, as does whether *BRCA1* and *BRCA2* (*BRCA1/2*) germline variants are associated with platinum treatment sensitivity.

OBJECTIVE To compare 6 cycles of paclitaxel plus carboplatin (PCb) with a standard-dose regimen of 3 cycles of cyclophosphamide, epirubicin, and fluorouracil followed by 3 cycles of docetaxel (CEF-T).

DESIGN, SETTING, AND PARTICIPANTS This phase 3 randomized clinical trial was conducted at 9 cancer centers and hospitals in China. Between July 1, 2011, and April 30, 2016, women aged 18 to 70 years with operable TNBC after definitive surgery (having pathologically confirmed regional node-positive disease or node-negative disease with tumor diameter >10 mm) were screened and enrolled. Exclusion criteria included having metastatic or locally advanced disease, having non-TNBC, or receiving preoperative anticancer therapy. Data were analyzed from December 1, 2019, to January 31, 2020, from the intent-to-treat population as prespecified in the protocol.

INTERVENTIONS Participants were randomized to receive PCb (paclitaxel 80 mg/m^2 and carboplatin [area under the curve = 2] on days 1, 8, and 15 every 28 days for 6 cycles) or CEF-T (cyclophosphamide 500 mg/m^2, epirubicin 100 mg/m^2, and fluorouracil 500 mg/m^2 every 3 weeks for 3 cycles followed by docetaxel 100 mg/m^2 every 3 weeks for 3 cycles).

MAIN OUTCOMES AND MEASURES The primary end point was disease-free survival (DFS). Secondary end points included overall survival, distant DFS, relapse-free survival, DFS in patients with germline variants in *BRCA1/2* or homologous recombination repair (HRR)–related genes, and toxicity.

RESULTS A total of 647 patients (mean [SD] age, 51 [44-57] years) with operable TNBC were randomized to receive CEF-T (n = 322) or PCb (n = 325). At a median follow-up of 62 months, DFS time was longer in those assigned to PCb compared with CEF-T (5-year DFS, 86.5% vs 80.3%, hazard ratio [HR] = 0.65; 95% CI, 0.44-0.96; *P* = .03). Similar outcomes were observed for distant DFS and relapse-free survival. There was no statistically significant difference in overall survival between the groups (HR = 0.71; 95% CI, 0.42-1.22; *P* = .22). In the exploratory and hypothesis-generating subgroup analyses of PCb vs CEF-T, the HR for DFS was 0.44 (95% CI, 0.15-1.31; *P* = .14) in patients with the *BRCA1/2* variant and 0.39 (95% CI, 0.15-0.99; *P* = .04) in those with the HRR variant. Safety data were consistent with the known safety profiles of relevant drugs.

CONCLUSIONS AND RELEVANCE These findings suggest that a paclitaxel-plus-carboplatin regimen is an effective alternative adjuvant chemotherapy choice for patients with operable TNBC. In the era of molecular classification, subsets of TNBC sensitive to PCb should be further investigated.

TRIAL REGISTRATION ClinicalTrials.gov Identifier: NCT01216111

Visual Abstract
Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Authors: Zhi-Ming Shao, MD (zhimingshao@yahoo.com), and Ke-Da Yu, MD, PhD (yukeda@163.com), Department of Breast Surgery, Fudan University Shanghai Cancer Center, 270 Dongan Road, Shanghai 200032, China.

JAMA Oncol. 2020;6(9):1390-1396. doi:10.1001/jamaoncol.2020.2965
Published online August 13, 2020.

jamaoncology.com

1390

Downloaded from jamanetwork.com by guest on 10/31/2025

Curation

initials	PMID	cancer_type	condition_E	condition_I	N_E	N_I	dataset_E
AS	27613525	lung squamous	1	early-stage lun	1	107 stage I/II tu	1
DK	27613525	Lung	1	Patients compl	1	(the Japan coh	1
AT	27613525	lung	1	"the two-gene (	1	"two independe	1
AT	29682089	breast	1	"DFNA5 methly	1	"The number o	1
FK	29682089	breast	1	we aimed to ar	1	We analyzed Ir	1
AT	36696113	lung	1	"Novel mdeicat	1	"There were ar	1
DK	36696113	Lung	1	We aimed to cl	0	N/A	1
FK	36696113	Non-small cell	1	Medicaid patier	1	2281 person-ye	1
AS	36836779	triple-negative	1	triple-negative	1	1350 patient sa	1
DK	36836779	Breast	1	human sample	1	A total of nine (	1
AS	37313409	esophageal	1	esophageal sq	1	94 ESCC samp	1
DK	37313409	Esophageal	1	After initial scre	1	To start the dat	1
FK	37313409	esophageal sq	1	single-cell RNA	1	GSE53624 dat	1

Resolution

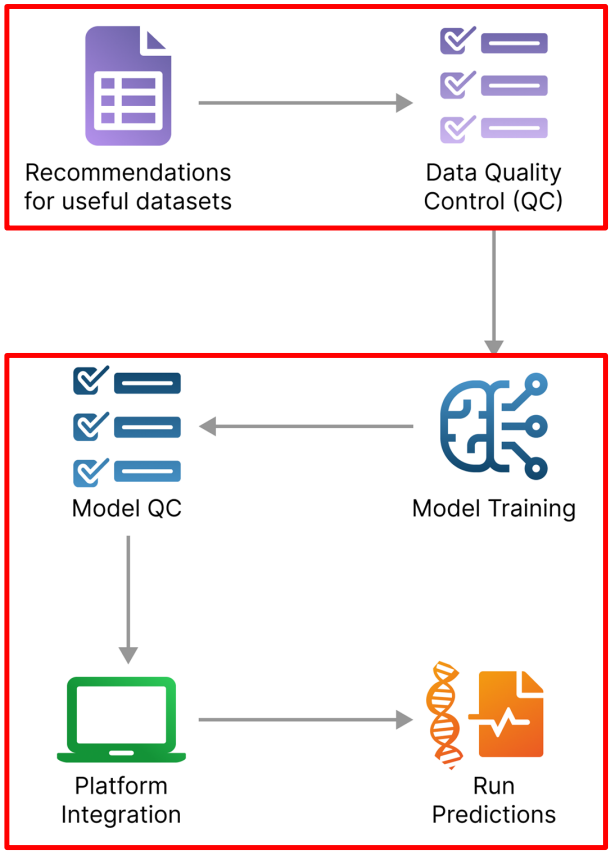
cancer_type	conditio	condition_I	N_E	N_I	dataset	dataset_I	interven	intervent
Breast	1	we utilized comp	1	We used our mu	1	All data can be v	1	Combina
Breast	1	investigated the	1	multiomics TNB(	1	n this study, we (	1	we also p
Breast	1	investigated the	1	multiomics TNB(	1	in this study, we	1	Combina

ntervention_I	pr_endpoir	pr_endpoint_	R_criteria_E	R_criteria_I	NR_criteria	NR_criteria_I	recommend
Combination of (	1		0		0		0
e also propose	1	ferroptosis ac	1	immune resp	0	N/A	1
Combination of (	1	GPX4 inhibit(	1	These data i	1	the increase in G	1

ML Training

This Semester's Objective

Previous Semester



Summer/Fall Semester

Spring Semester

Semester Timeline & Process

STEP 01	Dataset selection	
STEP 02	Data extraction & QC	
STEP 03	Model training & validation	
STEP 04	Documentation & deployment	

Preparation and Data Selection

Environment Set-Up

Read Through Papers

Screen Datasets

Data Type

Sample Size

Resulting Data

Limitations

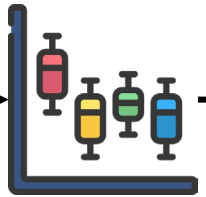
Data Processing

Load Data

```
25-Feb-2026 13:39:07 DEBUG GEOparse - SAMPLE: GSM388076
25-Feb-2026 13:39:07 DEBUG GEOparse - SAMPLE: GSM388077
25-Feb-2026 13:39:08 DEBUG GEOparse - SAMPLE: GSM388078
25-Feb-2026 13:39:08 DEBUG GEOparse - SAMPLE: GSM388079
25-Feb-2026 13:39:08 DEBUG GEOparse - SAMPLE: GSM388080
25-Feb-2026 13:39:08 DEBUG GEOparse - SAMPLE: GSM388081
25-Feb-2026 13:39:09 DEBUG GEOparse - SAMPLE: GSM388082
25-Feb-2026 13:39:09 DEBUG GEOparse - SAMPLE: GSM388083
25-Feb-2026 13:39:09 DEBUG GEOparse - SAMPLE: GSM388084
25-Feb-2026 13:39:09 DEBUG GEOparse - SAMPLE: GSM388085
```

Basic Quality Control

```
Expression matrix shape: (22283, 311)
Missing values total: 0
Duplicate probe IDs: 0
Duplicate samples: 0
```



Read Data Statistics

name	GSM615096
count	22283.000000
mean	7.990332
std	2.257689
min	0.371459
25%	6.570092
50%	8.115227
75%	9.393594
max	17.356685

Format for Training

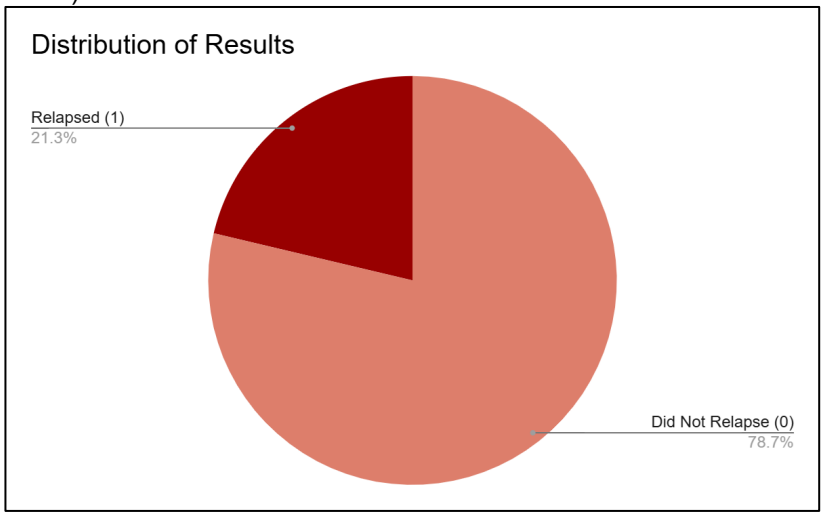
Clean-Up Data

PMID Overview - GSE25066

Original Contribution FREE

A Genomic Predictor of Response and Survival Following Taxane-Anthracycline Chemotherapy for Invasive Breast Cancer

(Hatzis et al., 2011)

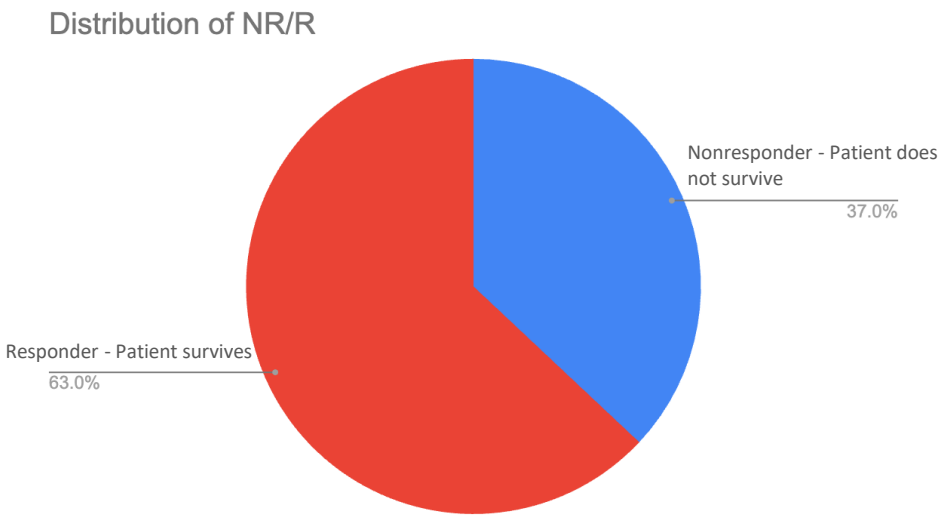


PMID	**2558518**
Disease	Invasive Breast Cancer
Intervention	Taxane-Anthracycline Chemotherapy
Number of Patients	310
Measured Result	Distant Relapse-Free Survival (DRFS)
Dataset Size	6908040 (22284 samples/patient)
Data Type	RNA

PMID Overview - GSE74777

A Two-Gene Prognostic Classifier for Early-Stage Lung Squamous Cell Carcinoma in Multiple Large-Scale and Geographically Diverse Cohorts

Rintaro Noro [a,#], Teruhide Ishigame [a,#], Naomi Walsh [a,b,#], Kouya Shiraishi [c], Ana I Robles [a], Bríd M Ryan [a], Aaron J Schetter [a], Elise D Bowman [a], Judith A Welsh [a], Masahiro Seike [d], Akihiko Gemma [d], Vidar Skaug [e], Steen Mollerup [e], Aage Haugen [e], Jun Yokota [f], Takashi Kohno [c], Curtis C Harris [a,*]



PMID	27613525
Disease	Early-stage lung squamous cell carcinoma
Intervention	Chemotherapy
Number of Patients	Japan (n=121) and NCI-MD (n=73), Total (n=194) Current sample size (n = 107)
Endpoints	Tumor gene expression (primary), survival outcomes (secondary)
Dataset Size	253 candidate genes * 194 patients
Datatype	Boolean data (dead/alive),

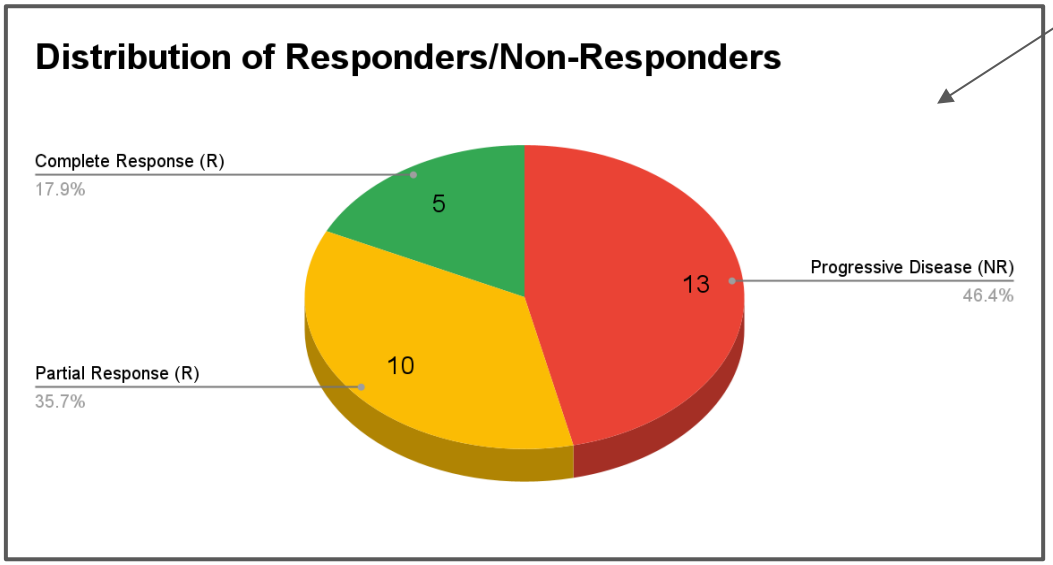
PMID Overview: 37313409; GSE78220

A fibroblast-associated signature predicts prognosis and immunotherapy in esophageal squamous cell cancer

Qianhe Ren [1], Pengpeng Zhang [1], Xiao Zhang [1], Yanlong Feng [1], Long Li [2], Haoran Lin [1], Yue Yu [1]

Affiliations + expand

PMID: 37313409 PMCID: PMC10258351 DOI: 10.3389/fimmu.2023.1199040



PMID	37313409
Disease	Esophageal Squamous Cell Cancer (Melanoma)
Intervention	Immunotherapy: Pembrolizumab
# of Patients	N=28
Endpoints	PD: >= 20% increase in tumor size PR: >= 30% decrease in total diameter CR: = Disappearance of all target tumors
Dataset Size	39,380 gene features 4 metadata features
Data Type	RNA-seq

Documentation: BioCompute Objects (BCOs)

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

In PredictMod, BCOs **fully describe all details of a model**, including:

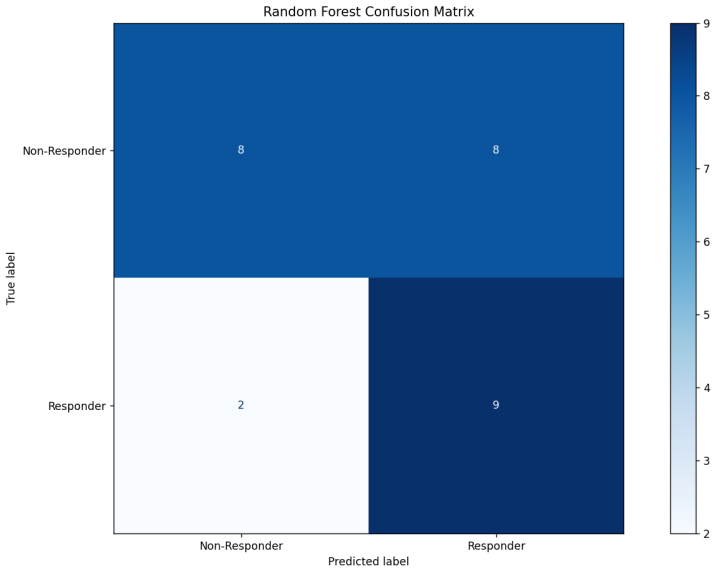
- The data set used for training and associated provenance
- Model type and associated hyperparameters used to create the model
- Software requirements
- Relevant features (variables of the dataset that influenced prediction outcomes)
- Any other information required to fully replicate the model

By adhering to BCO standards, researchers and practitioners can **confidently reproduce results**, **validate findings**, and **collaborate effectively**, thereby advancing scientific rigor and trustworthiness in computational biology and related disciplines.

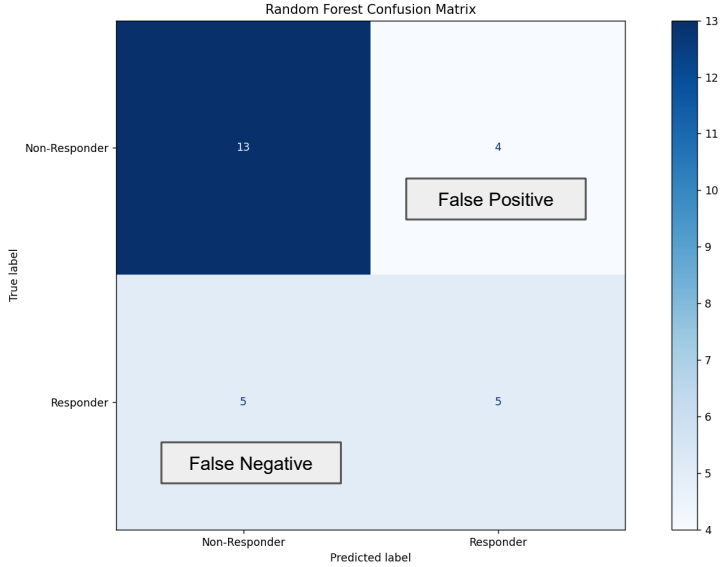
8 Top Level Domains	Required	Optional
1	**Provenance Domain**: Metadata describing the BCO	
2	**Usability Domain**: Free text field for researcher to explain the analysis and relevant details	
3	**Extension Domain**: User-defined fields	
4	**Description Domain**: Steps of the analysis, external resources needed for the steps, and the relationship of I/O objects	
5	**Execution Domain**: Information about the environment in which the analysis was run	
6	**Parametric Domain**: Records any parameters that were changed from default values	
7	**Input and Output Domain**: A list of global input and output files	
8	**Error Domain**: Used for describing errors	

Results - Diya

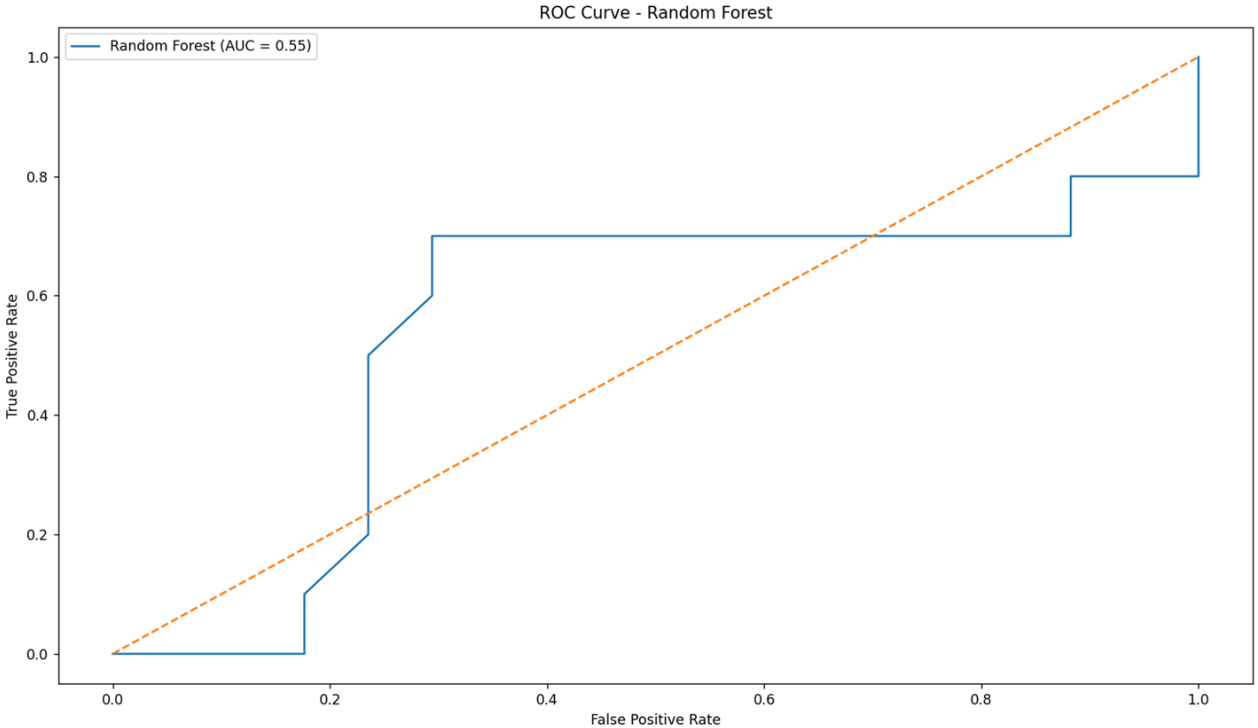
Confusion Matrix With Overfitting



Confusion Matrix Without Overfitting

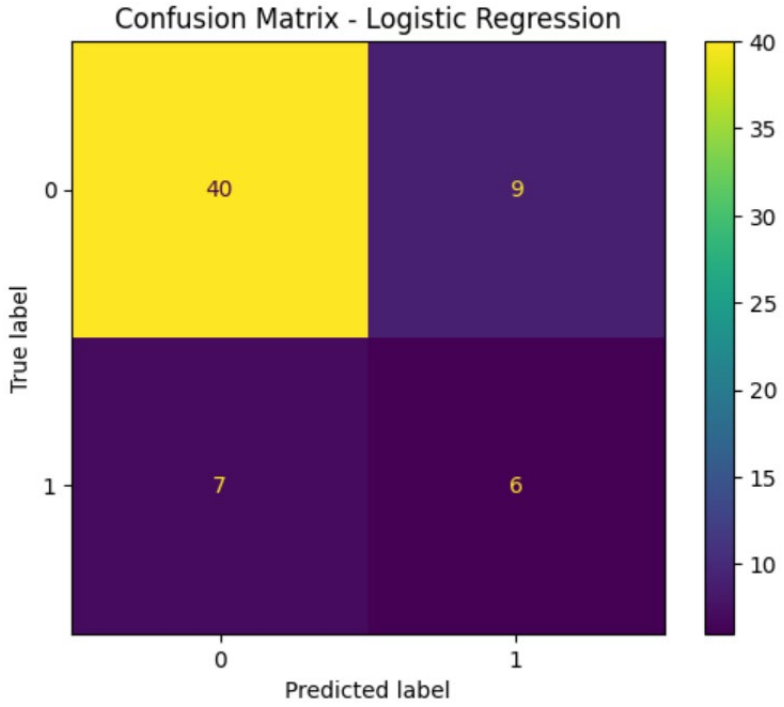


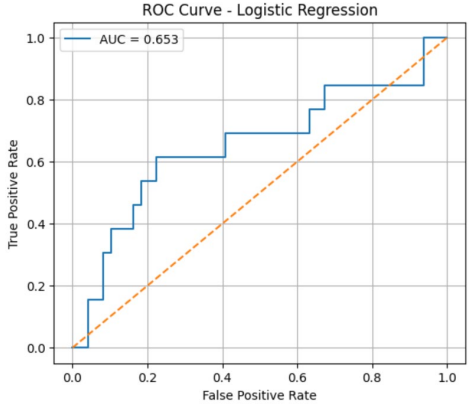
Results - Diya

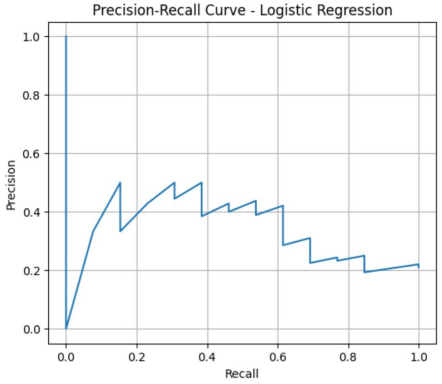


Results - Sampurna

```
drfs_1_event_0_censored
0    244
1     66
Name: count, dtype: int64

Reduced feature count: 1000

Training set shape: (248, 1000)
Test set shape: (62, 1000)

Model Performance:
accuracy: 0.7419
precision: 0.4000
recall: 0.4615
f1: 0.4286
auc: 0.6531
```







Results - Sampurna

Gene	Coeff.	Biological Relevance
EE1A2	1.518858	Gene that gets overexpressed in breast tumors when its is usually undetectable in normal breasts (Tomlinson et al., 2005)
PTN	-1.2778893	Growth factor that contributes to immunosuppression and metastasis (Ganguly et al., 2023)
GALNT12	-1.155978	Gene codes for modifying protein and has been correlated to increased risk of colorectal cancer (NCBI, 2026)
C14orf105	-1.058550	Expresses in a statistically significant way when comparing normal and cancerous breast tissues (Chen & Yang, 2015)
DLX2	1.039317	Its protein shuts down cell signalling and promotes tumor growth (Yilmaz et al., 2011)

Magnitude of the coefficient indicates whether the genes is expressed more or less
Sign of the coefficient indicates whether it is associated with a positive or negative outcome

Results - Ashley (GSE78220) Preprocessing Summary

```
[STEP 1] Loading datasets ...
  Expression matrix : 28 samples x 39376 genes
  Metadata          : 28 samples x 5 columns
```

```
Gene columns (first 5)     : [100287102, 653635, 102466751, 107985730, 100302278]
Metadata columns           : ['response', 'gender', 'age', 'disease_status', 'prev_mapki']
Expression NaN count       : 0
Metadata NaN count         : 0
Duplicate expr samples     : 0
Duplicate meta samples     : 0
Duplicate gene columns     : 0
```

```
Merged dataset shape (X_combined)    : 28 samples x 39380 features
  Gene features                      : 39376
  Clinical features                  : 4
Target vector shape                  : 28 labels
Responders (1): 15   Non-responders (0): 13
```

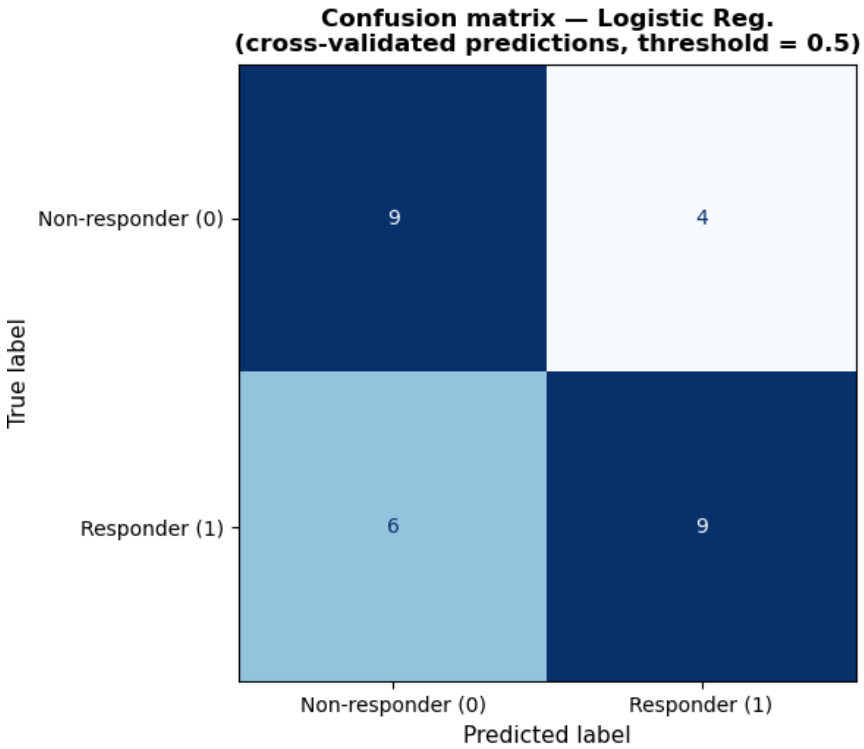
```
[STEP 4] Building CV-safe pipelines ...
  Total feature columns in X_combined  : 39380
  Gene columns identified              : 39376
  Clinical columns identified          : 4 ['Gender', 'Age', 'Disease_Stage', 'MAPK_History']
  N_GENES selected per fold (kbest)    : 25
  N_CLIN  selected per fold (kbest)    : 4
  Total features entering classifier   : 29
```

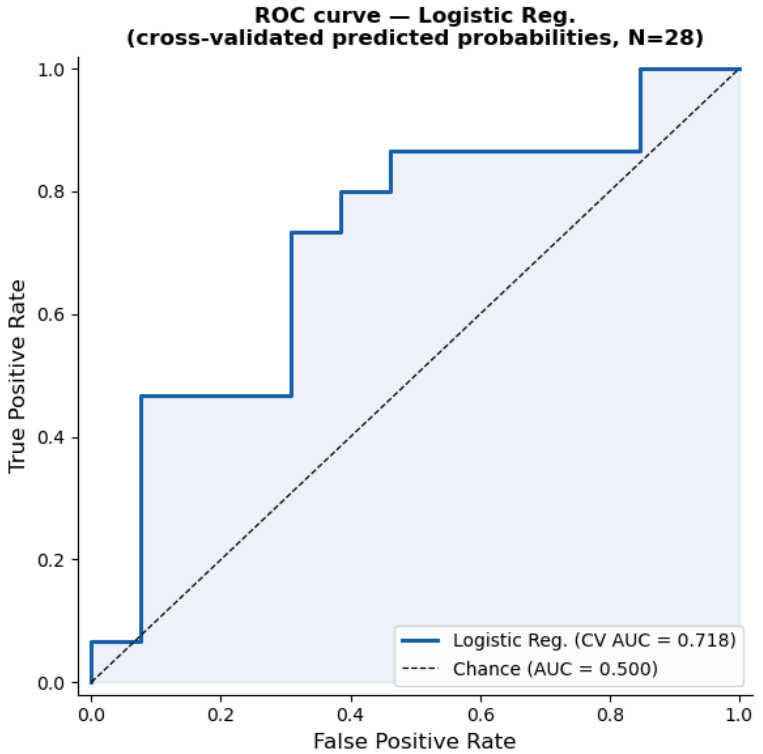
Results- Ashley

```
-- Full Cross-Validated Results --------------------------------
               AUROC     F1  Recall  Precision  Accuracy  Train AUROC  CV Gap Overfit?
Model
Random Forest  0.648  0.670   0.711      0.692     0.649          1.0   0.352      YES
SVM            0.667  0.602   0.600      0.697     0.636          1.0   0.333      YES
Logistic Reg.  0.767  0.648   0.644      0.696     0.642          1.0   0.233      YES
Decision Tree  0.472  0.484   0.511      0.473     0.478          1.0   0.528      YES
XGBoost        0.656  0.626   0.644      0.646     0.600          1.0   0.344      YES
```

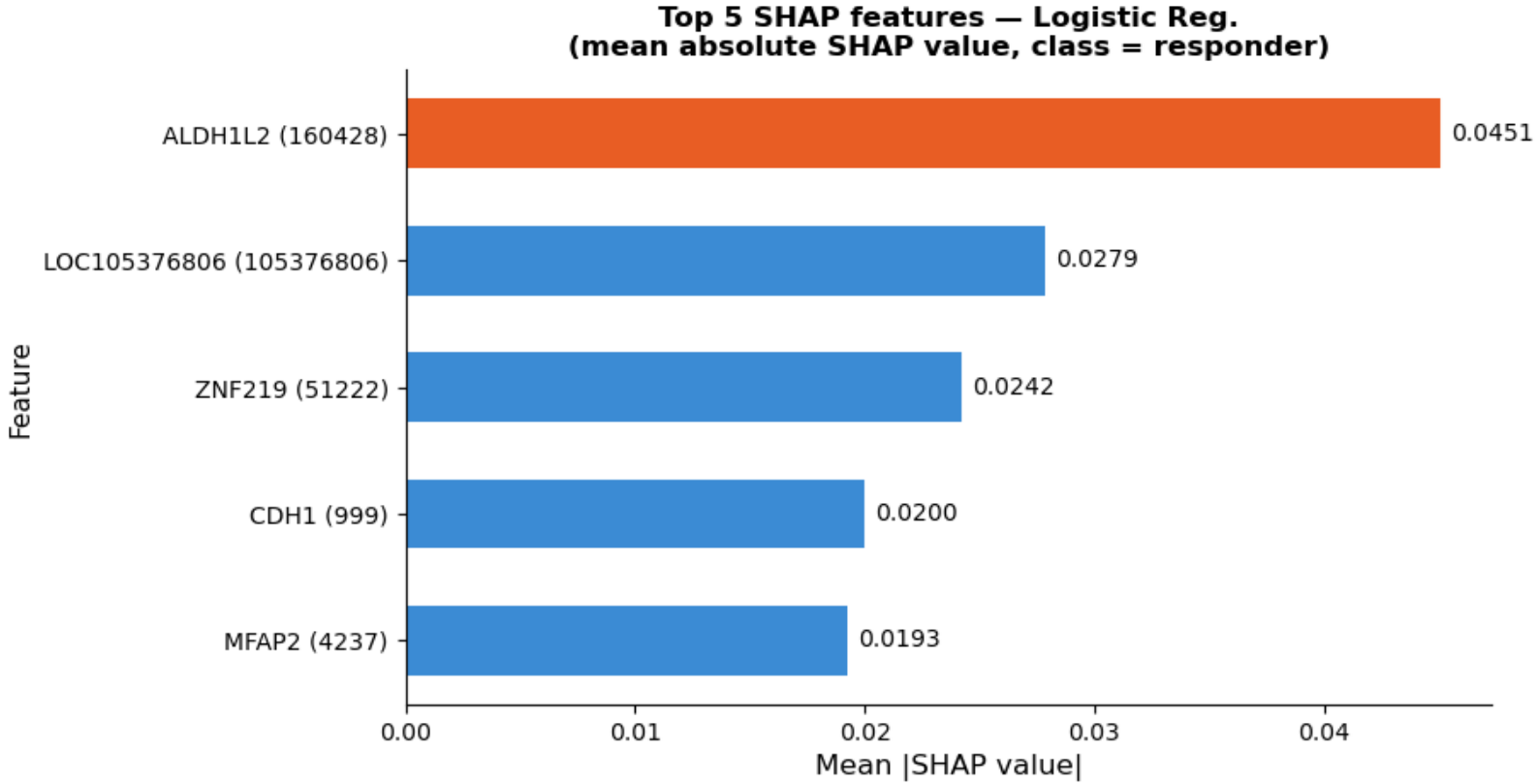
```
[STEP 5b] Leave-One-Out CV on best model (Logistic Reg.) ...

  LOO CV results for Logistic Reg. (N-1=27 training samples per iteration):
  Metric                   LOO   k-fold CV    Difference
  ----------------------------------------------------
  AUROC                  0.836       0.767       +0.069
  F1                     0.774       0.648       +0.126
  Recall                 0.800       0.644       +0.156
  Precision              0.750       0.696       +0.054
  Accuracy               0.750       0.642       +0.108
  ----------------------------------------------------
  Positive difference = LOO is more optimistic (less training-data bias)
  LOO AUROC (0.836) vs k-fold AUROC (0.767): LOO higher -- k-fold was pessimistic due to smaller training folds
```

Results - Ashley





Results - Ashley



Challenges and Solutions

Step 1: Set up

- Learning curve: Setting up environment and learning how to code

Step 2: Data Preprocessing

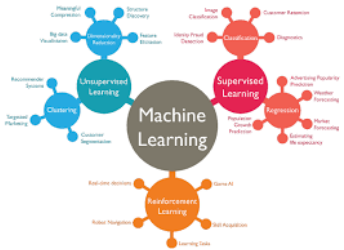
- Data extraction complexity
- Extracting resulting data from the data file
- Ensuring clean, usable datasets

Step 3: Model Training

- Ensuring no data leakage
- Finding ways to optimize our model performance

Solution:

- Peer Reviews and Check-Ins
- Utilizing ML Tutorials on the PredictMod Wikipage



Future Considerations and Next Steps

Recommendations for improving the models:

- Find Support for results by incorporating samples from other publicly available datasets
- Identify biological pathways from the top features to look for correlation
- Conduct verifications of findings
- Replicate/reproduce findings with more samples in new dataset/experiment

Next Steps:

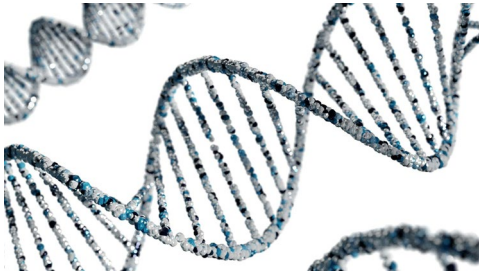
- Dockerization
- Finalization and delivery: Uploading to Github

GitHub

Acknowledgements

Thank you to:

- Dr. Mazumder
- Lori Krammer
- Patrick McNeely



Research Paper Recommendation For Machine Learning Modeling Using Large Language Model

Presenter: Vishal Bakshi
Mentor: Dr. Raja Mazumder

04/15/2026
Bioinformatics Symposium

BACKGROUND AND MOTIVATION

- **The screening bottleneck**: A researcher spends 15-20 minutes per paper evaluating if it has usable data for ML modeling. Across 100 candidates, that's 25+ hours just to find 1-2 usable papers.
- **The needle-in-a-haystack problem**: The information needed to make this decision - cohort definitions, dataset access, interventions, endpoints is scattered across abstract, methods, results, and supplementary sections. No single section gives the answer.
- **The real cost:** Every hour spent screening papers is an hour not spent building models. As published research grows exponentially, manual review becomes the bottleneck, not the modeling itself.

PROBLEM STATEMENT

- There is currently no reliable automated mechanism to determine whether a biomedical research paper is suitable for machine-learning modeling. Existing approaches rely heavily on manual expert review or keyword-based searches, which fail to capture methodological completeness and consistency.
- As a result, identifying ML-compatible papers is slow, subjective, and difficult to scale across large volumes of research literature.

OBJECTIVE

Build an end-to-end, self-hosted pipeline that turns raw PubMed papers into ML-ready dataset recommendations.

- Scheduled extraction of cancer research papers from PubMed's API, parsed from XML into structured text, running weekly on in-house servers with no cloud dependency or API costs.
- Batch-process hundreds of papers through a locally deployed LLM (Ollama) that scores each paper against 6 ML-readiness criteria, running overnight on existing infrastructure, not expensive GPU clusters.
- Validate every LLM output against a human-curated golden dataset, retry on malformed responses, and store all results in MongoDB, building an ever-growing, searchable knowledge base of evaluated papers that eliminates redundant screening.

DATA SOURCE (PubMed - Entrez API)

- PubMed is the global resource library for biomedical literature, providing access to millions of full-text research papers essential for our recommendations.
- **Entrez API:** Utilized the Entrez API to automate the search and download of full-text XML data, replacing manual review.
- **Query:** (TCGA OR GEO OR SEER OR publicly available data OR open access data OR public dataset OR data repository) AND (cancer OR neoplasm OR carcinoma OR tumor OR malignancy) AND (treatment OR therapy OR drug OR chemotherapy OR radiotherapy OR immunotherapy OR clinical trial OR intervention OR targeted therapy OR pharmacotherapy)

A PEEK INTO DATA

XML

```xml
<?xml version="1.0"  ?>
<!DOCTYPE pmc-articleset PUBLIC "-//NLM//DTD ARTICLE SET 2.0//EN" "https://dtd.nlm.nih.gov/ncbi/pmc/a
<pmc-articleset>
    <article article-type="research-article" xml:lang="en" dtd-version="1.4">
        <front>
            <journal-meta>
                <journal-id journal-id-type="nlm-ta">Cell Transplant</journal-id>
                <journal-id journal-id-type="iso-abbrev">Cell Transplant</journal-id>
                <journal-id journal-id-type="pmc-domain-id">3327</journal-id>
                <journal-id journal-id-type="pmc-domain">cll</journal-id>
                <journal-id journal-id-type="publisher-id">CLL</journal-id>
                <journal-title-group>
                    <journal-title>Cell Transplantation</journal-title>
                </journal-title-group>
                <issn pub-type="ppub">0963-6897</issn>
                <issn pub-type="epub">1555-3892</issn>
                <publisher>
                    <publisher-name>SAGE Publications</publisher-name>
                </publisher>
            </journal-meta>
            <article-meta>
                <article-id pub-id-type="pmcid">PMC7873767</article-id>
                <article-id pub-id-type="pmcid-ver">PMC7873767.1</article-id>
                <article-id pub-id-type="pmcaid">7873767</article-id>
                <article-id pub-id-type="pmcaiid">7873767</article-id>
                <article-id pub-id-type="pmid">33327771</article-id>
                <article-id pub-id-type="doi">10.1177/0963689720978722</article-id>
                <article-id pub-id-type="publisher-id">10.1177_0963689720978722</article-id>
                <article-version article-version-type="pmc-version">1</article-version>
                <article-categories>
                    <subj-group subj-group-type="heading">
                        <subject>Original Article</subject>
                    </subj-group>
                </article-categories>
                <title-group>
```

JSON


```json
{
  "metadata": {
    "title": "Molecular Classification and Emerging Targeted Therapy in Endometrial Cancer",
    "pmcid": "PMC6685771",
    "pmid": "30741844",
    "doi": "10.1097/PGP.0000000000000585",
    "authors": [
      "Ting-Tai Yen",
      "Tian-Li Wang",
      "Amanda N. Fader",
      "Ie-Ming Shih",
      "Stephanie Gaillard"
    ],
    "year": "2020",
    "keywords": [
      "endometrial cancer",
      "molecular classification",
      "targeted therapy"
    ]
  },
  "abstract": "Recent advances in molecular studies, especially genome-wide analyses, have revealed
  "introduction": "Introduction\nUterine cancer is the most common cancer of the female reproducti
  "methods": "",
  "results": "",
  "discussion": "",
  "conclusion": "Conclusions\nSince publication of the TCGA data in 2013 [ 35 ], there have been e
  "data_availability": "",
  "supplementary": "",
  "tables": "Table 1.\nCharacteristics of endometrial carcinoma by histology.\nLG-Endometrioid | H
  "figures": "Fig. 1.\nAssociations of histological and TCGA classifications in endometrial cancer
}
```

KEYWORD OVERVIEW

Entity	Definition
Condition_E (Experimental Cohort Definition)	The experimental or responder group is defined.
N_E (Sample Size)	The sample size or data volume for the experimental cohort is explicitly stated.
Dataset_E (Dataset Availability)	The data source for the experimental group is clearly specified and publicly accessible.
Intervention_E (Intervention Description)	The experimental treatment, intervention, or biological condition is clearly described.
Pr_endpoint_E (Primary Endpoint)	A primary outcome/response endpoint is defined.
R_criteria_E (Responder Criteria)	Criteria used to define responders are stated.

EVALUATION CRITERIA

- **Zero-Inference Policy:** Disallow assumptions or free-form explanations; only explicitly stated data is accepted.
- **Predictive vs. Prognostic Distinction:** Studies focusing only on survival correlations without a recorded treatment intervention are automatically flagged as invalid for response modeling.
- **Human-Only Data Requirement:** Research design is ignored if the data exists only in cell lines or animal models; human clinical data must be present.
- **Deterministic JSON Output:** All evaluations must be returned in a strict JSON schema to ensure they pass the Pydantic Validation layer.

PROMPT DESIGN

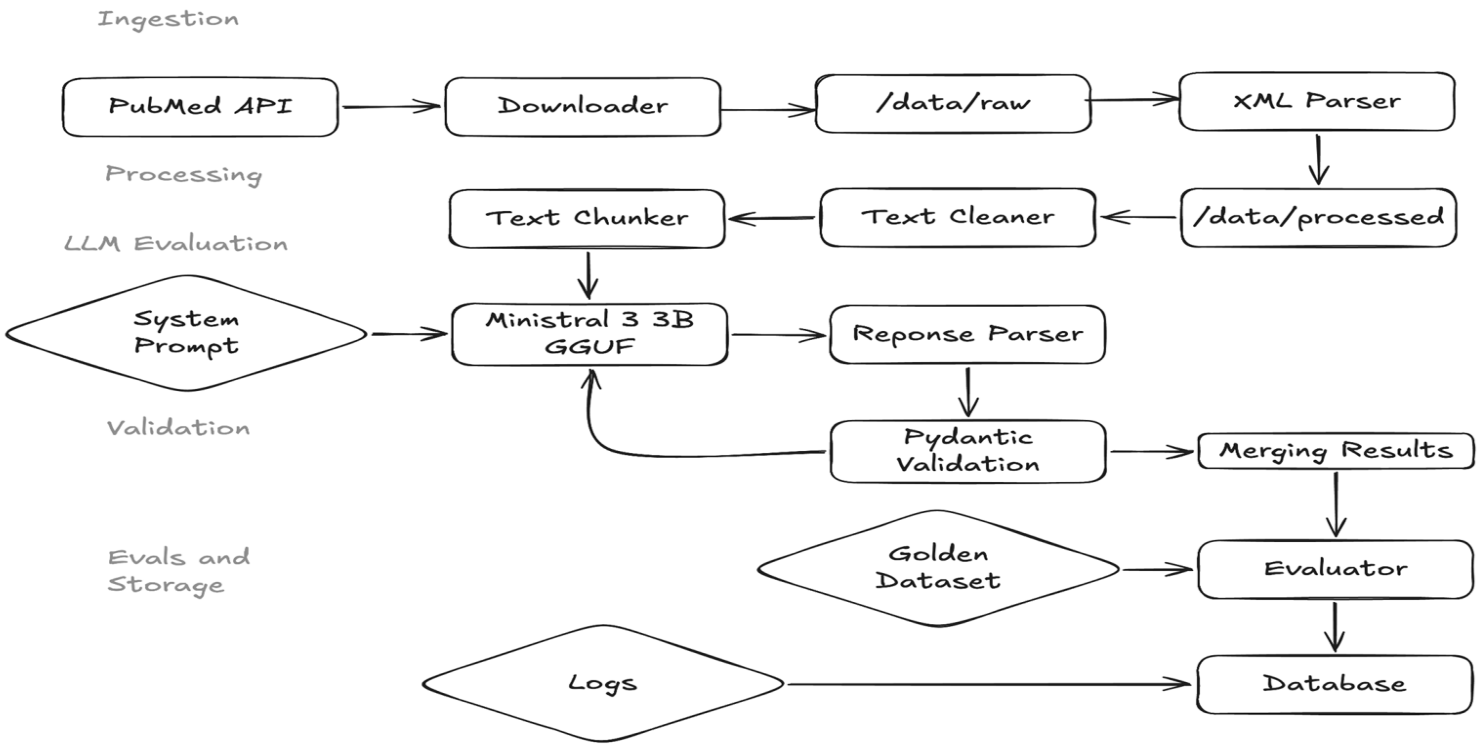
1. Shifted From Research Design to Data Availability Perspective

A paper could have brilliant research with interventions and outcomes, but if those elements only exist in cell lines or mouse models and NOT in the human dataset, it's useless for ML modeling. The old prompt was approving papers that looked good on paper but had no usable data.

2. Added Predictive vs Prognostic Distinction With Logic Chain

The old prompt was incorrectly flagging prognostic studies as valid. For example the Ferroptosis paper - "GPX4 correlates with poor survival" was being counted as a valid endpoint. But survival without a recorded treatment is useless for building a treatment response prediction model.

ARCHITECTURE



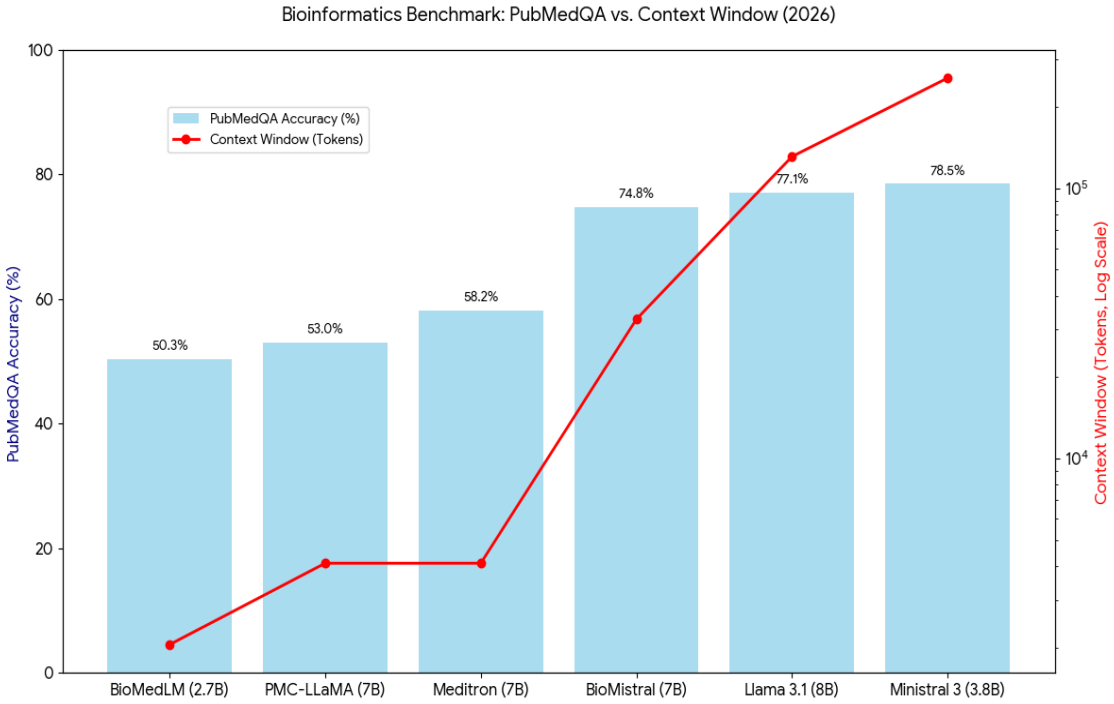
ARCHITECTURE KEY COMPONENTS

1. **Automated Scalable Ingestion**
   We've moved away from manual literature review by implementing a fully automated ingestion pipeline. The system autonomously fetches full-text XMLs via the PubMed API, standardizing raw data into a local /data/processed lake.

1. **Schema-Strict Extraction via Pydantic**
   To bridge the gap between unstructured research papers and structured databases, we use Pydantic-enforced extraction. By integrating Ministral 3 3B within a self-healing loop, the system validates every extracted entity against our strict bio-schema

1. **Systematic Quality Control & Regression Testing**
   Every new batch is automatically benchmarked against a manually curated Golden Dataset. This serves as continuous regression testing; it allows us to quantify precision and recall for our model's extractions, ensuring that any logic updates don't compromise the reliability of our stored results.

COMPARATIVE ANALYSIS



Unmatched Context Utility: While traditional Bio-LLMs (BioMedLM) fail at 2k tokens, Ministral 3 handles 256k tokens with **78.5% PubMedQA accuracy**, enabling the analysis of full-length research papers and complex genomic data that smaller-window models cannot process.

Edge-Optimized Performance: At only 3.8B parameters, it delivers "large-model" reasoning on local hardware via GGUF quantization, ensuring total data privacy for sensitive clinical records while maintaining high-speed inference without expensive server infrastructure.

MODEL DETAILS

Name of the LLM	Ministral 3 3B Instruct 2512
Tensor Type	Q4_K_M: Optimized 4-bit format that retains ~99% of model intelligence while reducing memory footprint by 70%
Context Window	256K Tokens
Inference Speed	~8-12 Tokens/second on CPU
Architecture	Mistral V3

INPUT - OUTPUT STRUCTURE

INPUT

```
{
"model":"ministral-3:3b-instruct",
 "messages": [
{ "role": "system",                    "content":
"system_prompt"},
{"role": "user", "content":
"parsed_paper_text"}],

  "stream": false,
  "options": {
    "temperature": 0.1,
    "num_ctx": 32254
  }
}
```

OUTPUT

```
{
  "paper_title": "string",
  "condition_E": 0,
  "condition_E_reason": "string",
  "N_E": 0,
  "N_E_reason": "string",
  "dataset_E": 0,
  "dataset_E_reason": "string",
  "intervention_E": 0,
  "intervention_E_reason": "string",
  "pr_endpoint_E": 0,
  "pr_endpoint_E_reason": "string",
  "R_criteria_E": 0,
  "R_criteria_E_reason": "string"
}
```

CONFUSION MATRIX FOR GROUND TRUTH

	Predicted: Recommended	**Predicted: Not Recommended**
Actual: Recommended	7 (TP)	0 (FN)
Actual: Not Recommended	1 (FP)*	6 (TN)

RESULT FROM LLM

```
 LLM Output:

{'paper_title': 'Afibroblast-associated signature predicts prognosis and immunotherapy in esophageal squamous cell cancer', 'co
ndition_E': 1, 'condition_E_reason': 'Six CAF clusters were identified in ESCC based on scRNA-seq data, three of which had pro
gnostic associations.', 'N_E': 1, 'N_E_reason': 'A total of 642 genes were found to be significantly correlated with CAF clust
ers from a pool of 17080 DEGs, and 9 genes were selected to generate a risk signature.', 'dataset_E': 1, 'dataset_E_reason': '
The GEO database provided the single-cell RNA sequencing (scRNA-seq) data. The GEO and TCGA databases were used to obtain bulk
 RNA-seq data and microarray data of ESCC, respectively.', 'intervention_E': 1, 'intervention_E_reason': 'A risk signature bas
ed on CAF-related prognostic genes was constructed using Lasso regression.', 'pr_endpoint_E': 1, 'pr_endpoint_E_reason': 'The
risk signature was significantly correlated with stroma and immune scores, as well as some immune cells.', 'R_criteria_E': 1,
'R_criteria_E_reason': 'Multivariate analysis demonstrated that the risk signature was an independent prognostic factor for ES
CC, and its potential in predicting immunotherapeutic outcomes was confirmed.'}

 Recommedation:

✅ Paper IS RECOMMENDED
```

RECOMMENDATIONS

Based on the recent automated pull from the PubMed database. These are the recommendations from the system.

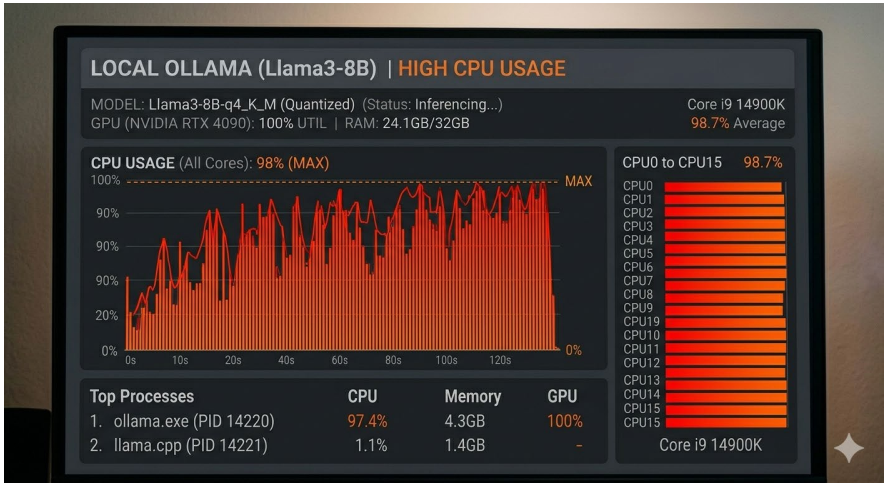
Partially Recommended:

1. PMC12872018 - Link
2. PMC7658512 - Link
3. PMC7570778 - Link

Fully Recommended:

1. PMC12872174 - Link

LIMITATIONS



Inference latency is currently constrained by CPU-bound processing, resulting in a lower tokens-per-second (TPS) throughput.

FUTURE DIRECTIONS

1. GUI-Driven Workflow
2. DevOps Integration

ACKNOWLEDGEMENTS

1. Dr. Raja Mazumder
2. Lori Krammer
3. Patrick McNeely