Leveraging Large Language Models for Scalable Glycan-Disease Relation Extraction

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

GlyGen www.glygen.org

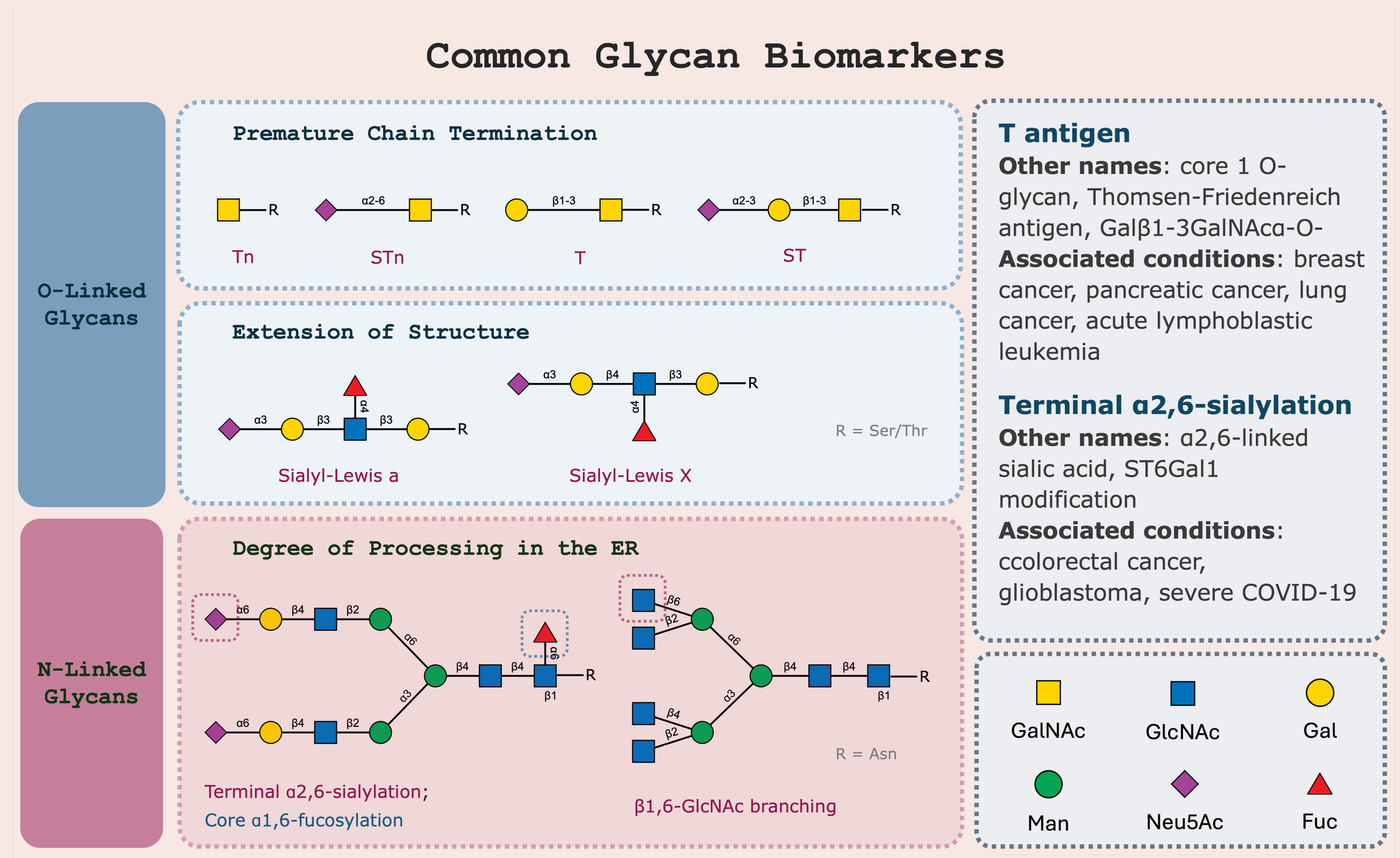
Cyrus Chun Hong Au Yeung, Robel Kahsay

Department of Biochemistry & Molecular Medicine, The George Washington University, Washington, DC 20052

INTRODUCTION

Glycans are complex carbohydrate molecules attached to proteins and lipids, forming essential biological structures known as glycoproteins and glycolipids. Glycosylation is the enzymatic process attaching glycans to biomolecules that plays critical roles in cell-cell recognition, immune response, and disease progression, including **cancers, autoimmune disorders, and infectious diseases**.

Systematically extracting glycan-disease relationships from biomedical literature remains challenging due to the diversity of glycan structures and nomenclatures. Our project addresses this gap by developing a scalable pipeline using **large language models (LLMs)** for automated extraction and curation of glycan biomarkers associated with diseases from PubMed abstracts, enabling the structured representation of glycan entities and their disease contexts.



METHODS

Glycan Biomarker Corpus

PubMed records with relevance to glycan biomarker research was systematically retrieved using a customized query. PubMed corpus consisted of 4,595 articles with **article title**, **abstract**, **keywords**, **MeSH names**, and **chemical names**.

Named Entity Recognition (NER)

System recognizes predefined entities across 5 categories:

- **"assessed_biomarker_entity"**: "glycan entity, glycan entity change, associated glycoprotein, associated glycoenzyme.
- **"associated_disease"**: disease name, medical intervention
- **"sample"**: organism, specimen type
- **"biomarker_role"**
- **"evidence"**

Relation Extraction (RE)

System identifies and links relevant entities such as glycan structures, diseases, and specimen types within text by considering the semantic meaning of sentences

Large Language Model (LLM)

The OpenAI **gpt-4o-2024-08-06 model** with temperature set to 0 was used. All outputs adhered to a predefined JSON schema.

RESULTS

Abstract Text Retrieval

- Custom PubMed query; Entrez E-utilities API fetches XML records

Data Preprocessing

- Remove duplicates (N=2) and records without available abstract (N=96)

Prompt Engineering

- Provide instructions for automated biocuration

LLM-based NER-RE

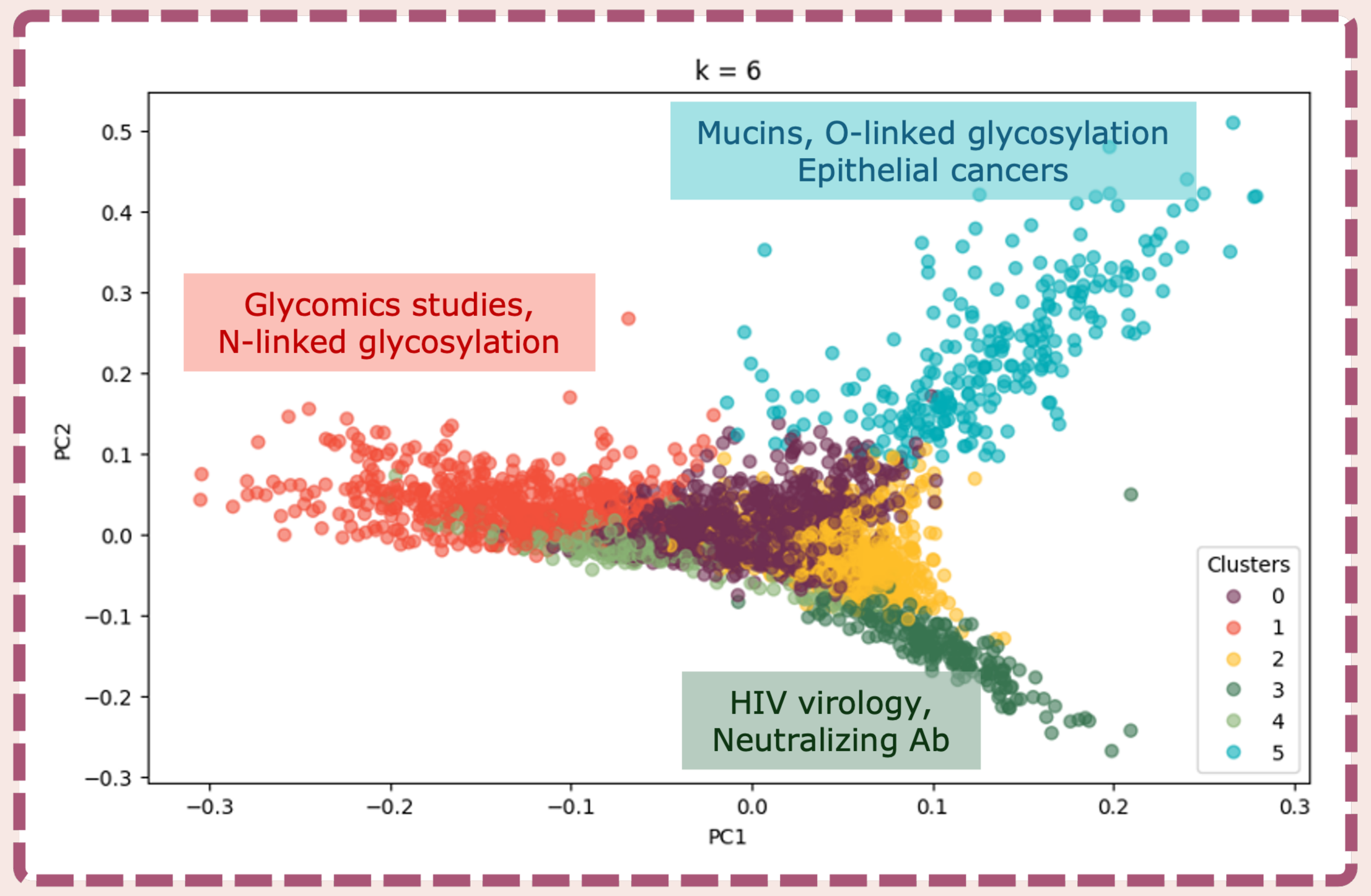
- Parallelized OpenAI API workflow

JSON Parsing

- Conversion of JSON output to a TSV dataset
- 6,782 glycan biomarker entries

Term Standardization

- Map glycan terms to the Glycan Structure Dictionary[1]
- GlyTouCan accession, Uberon, Disease Ontology



Retrieval Summary

Targeted PubMed query retrieved **4,595** citations on glycan biomarker research.

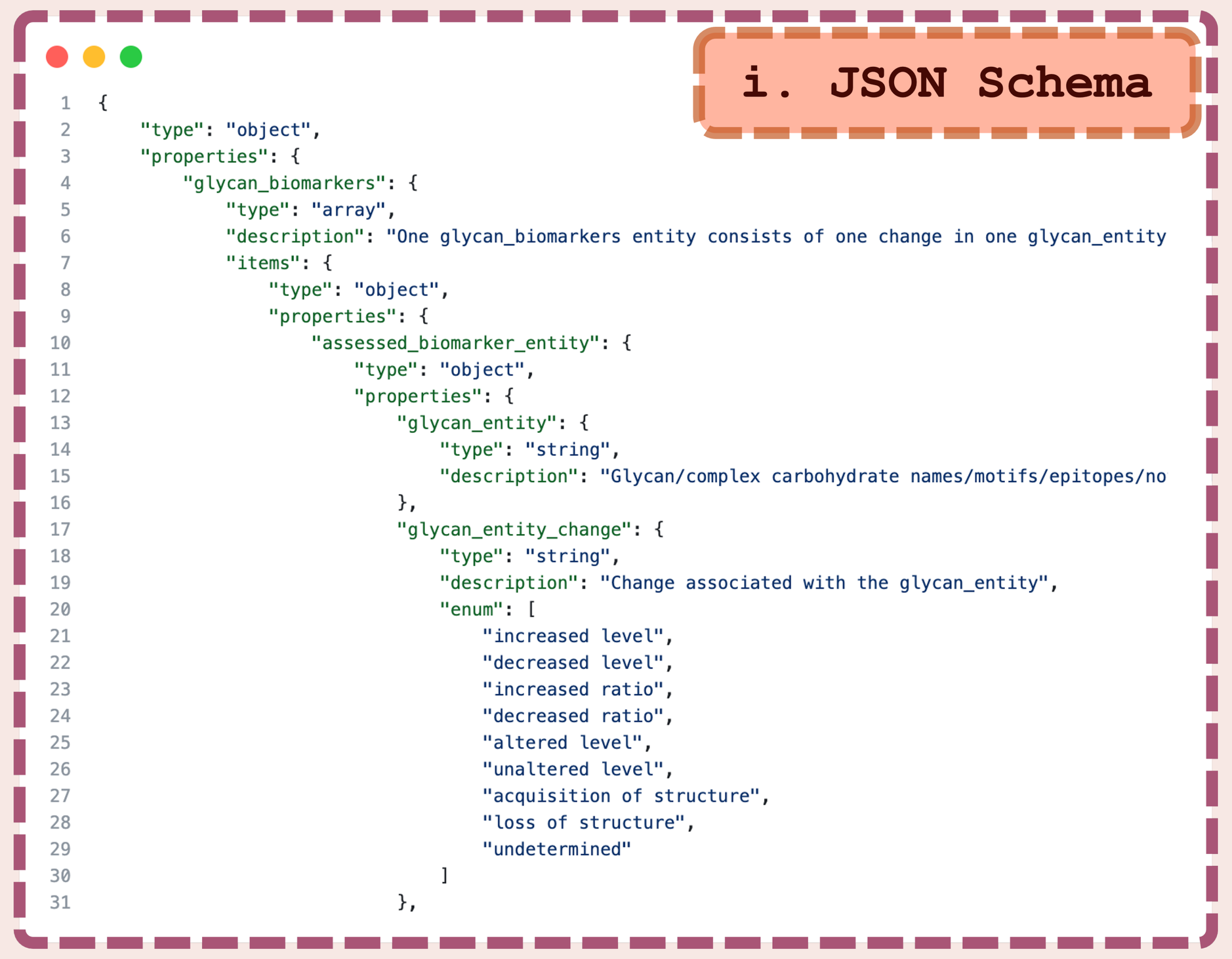
- 97.9% contain full abstract
- 98.5% are journal articles

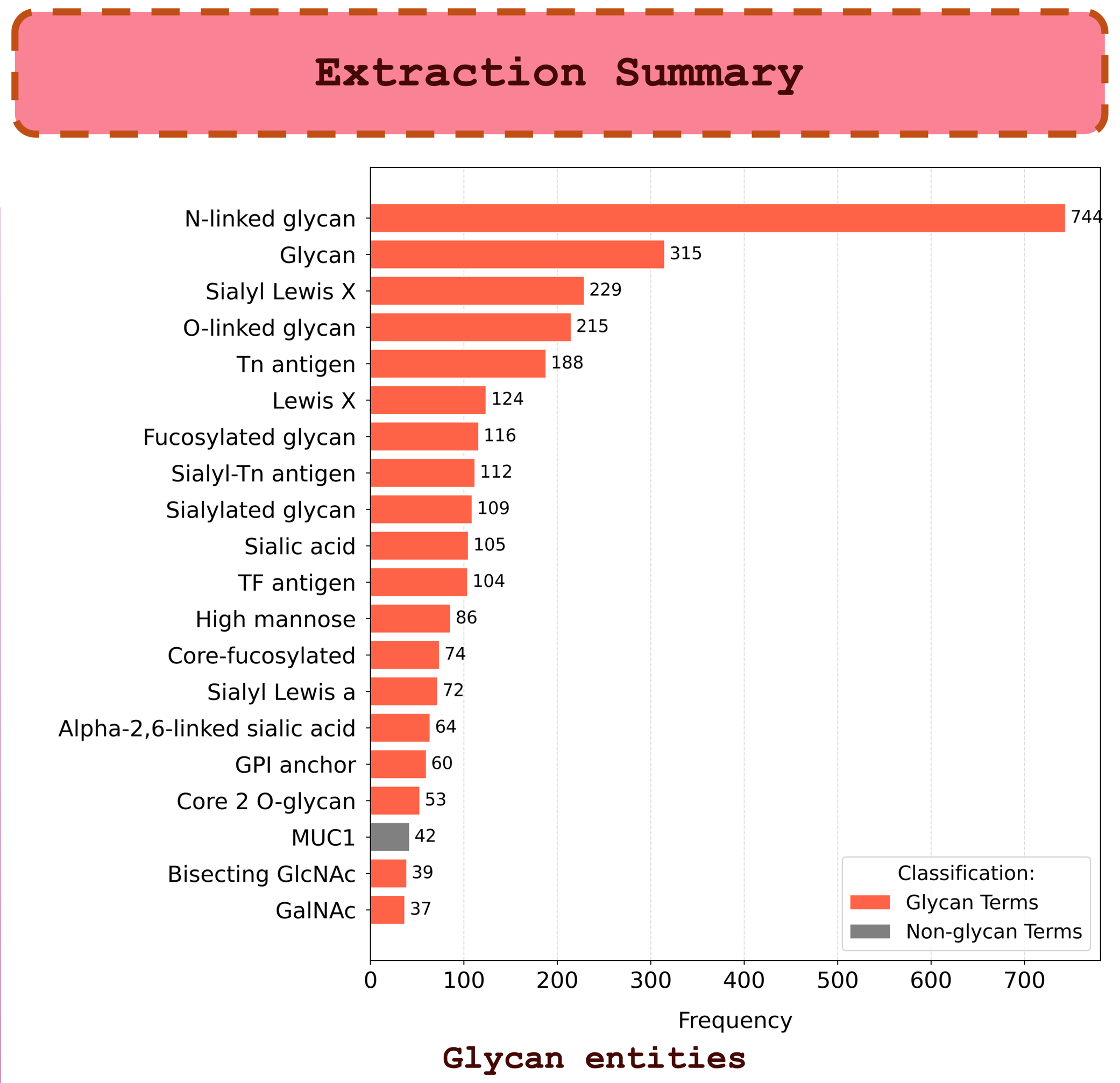
Top Substance Headings

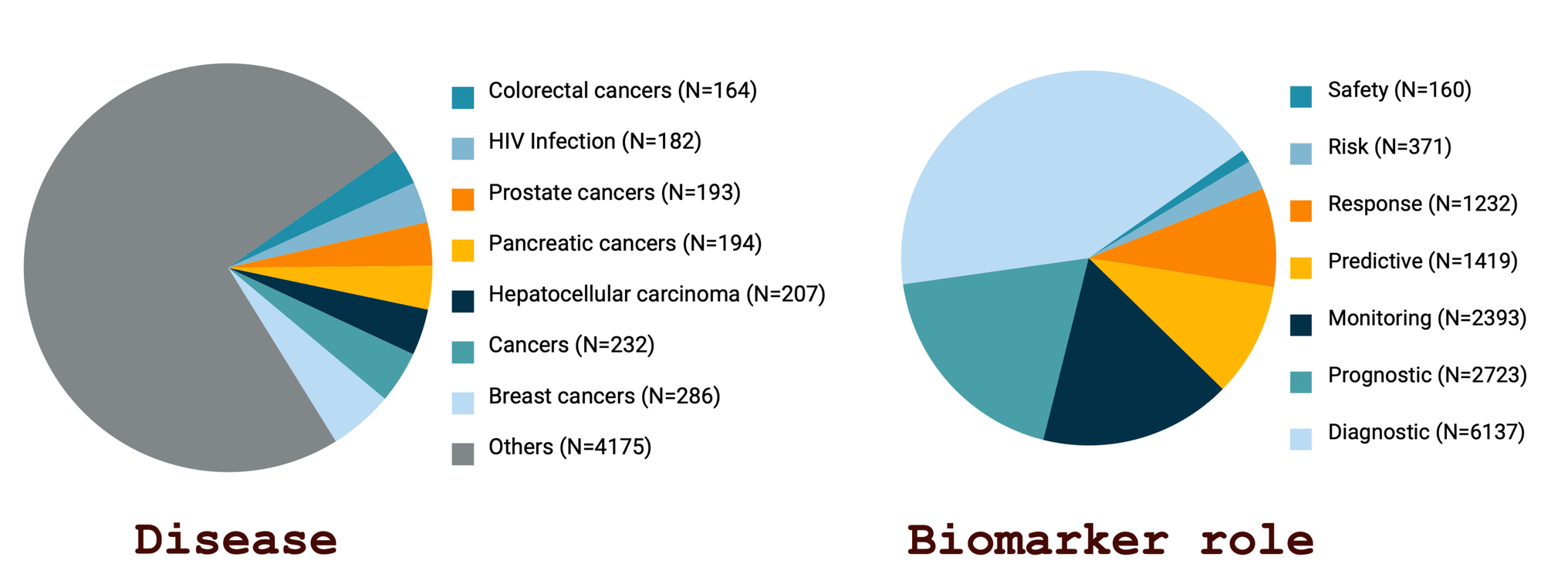
- Polysaccharides (26.2%)
- Biomarkers (22.3%)
- Biomarkers, Tumor (16.1%)

Top MeSH Terms

- Humans (80.6%)
- Glycosylation (55.2%)







DISCUSSIONS

Strengths of the LLM-Based Pipeline

- GPT-4o provides deep **paragraph-level semantic understanding**, resolving synonymy and context better than classic rule-based or BioBERT systems, which boosts relation-extraction accuracy.
- Through asynchronous API calls, the NER-RE module can parse **~50 PubMed abstracts in a minute**, reducing week-long manual tasks to a few hours. Curation speed is currently limited by the API rate limit.
- Prompt-driven control means new glycan or biological entity types does not require model retraining. Adjustments are done through prompt/schema edits.
- This pipeline can be expanded to any other biomedical fields, extracting user-defined terms and generalizing knowledge from discrete research.

Key Applications

Knowledgebase Integration

- JSON-standardized outputs import seamlessly into resources such as BiomarkerKB[2], enriching glycan–disease coverage for downstream users.

Pattern Discovery & Target Finding

- Aggregate analytics highlight recurrent aberrant glycan structures and expose understudied glycoenzymes as potential drug targets.

Semantic Text-Clustering

- Embedding vectors let us (i) query novel sentences reporting glycan biomarkers from larger corpora through embeddings search, and (ii) auto-detect term variants via cosine-similarity grouping, helping with the discovery of new glycan structures and terms.

ACKNOWLEDGEMENTS

We would like to thank the GlyGen project[3] for providing comprehensive glycan structure resources and ontology mappings. Such resources facilitated the standardization and annotation of the extracted glycan biomarker data, enhancing data interoperability for integration and downstream analysis.

REFERENCE

1. Vora J, Navelkar R, Vijay-Shanker K, et al. The glycan structure dictionary—a dictionary describing commonly used glycan structure terms. *Glycobiology (Oxford)*. 2023;33(5):354–357. https://www.ncbi.nlm.nih.gov/pubmed/36799723. doi: 10.1093/glycob/cwad014.
2. Biomarker Knowledgebase (BiomarkerKB). 2025. BiomarkerKB: A knowledgebase that integrates and standardizes biomarkers and related data from the Common Fund Data Ecosystem (CFDE) and other public sources. https://www.biomarkerkb.org.
3. York WS, Mazumder R, Ranzinger R, et al. GlyGen: Computational and informatics resources for glycoscience. *Glycobiology*. 2020;30(2):72–73. https://www.ncbi.nlm.nih.gov/pubmed/31616925. doi: 10.1093/glycob/cwz080.