

GlyGen: Publication Analysis Project

Points of Contact: Rene Ranzinger and Urnisha Bhuiyan

Background

Bioinformatics resources such as GlyGen depend on active engagement with the scientific communities they serve. To ensure that GlyGen continues to meet the needs of researchers, it is important to understand the size, geographic distribution, institutional landscape, and key contributors within relevant research communities. An initial version of a community analysis platform has already been developed. The software retrieves publication information, extracts author and affiliation metadata, and generates basic visualizations. Source code and a Docker container are available to allow volunteers to rapidly begin extending the platform.

Project Description

The goal of this project is to further develop and expand an existing publication analysis and visualization framework that can be used to characterize and understand scientific research communities. Rather than focusing on a single predefined research domain, the analysis framework is designed to support dynamic community discovery through keyword-based literature searches. Publications identified through these searches will then be included in downstream analyses and graphically represented in form of charts, diagrams or graphs.

The resulting analyses will help answer questions such as:

- How large is a particular research community?
- Where are its researchers geographically located?
- Which institutions and investigators are most active?
- Which organizations appear to be central contributors to the field?
- Which research groups overlap with GlyGen's current user and collaborator communities?
- Which potentially important communities or research groups are currently underrepresented in GlyGen outreach efforts?

Ultimately, the analysis may be integrated with GlyGen usage metrics, such as Google Analytics data, to identify regions and research hotspots where glycobiology research is active, but GlyGen adoption appears limited. Such information can help guide future outreach, training, and community engagement activities.

Existing Software Foundation

An initial version of the analysis platform is already available and includes:

- Retrieval and processing of publication metadata.
- Geographic visualization of author affiliations (separate all authors, first authors, and last authors).
- Word cloud visualizations of MeSH terms associated with publications.
- Dockerized deployment environment.
- Source code repository for continued development.

The primary focus of this project will be the extension of the existing software rather than the development of a new system from scratch.

Volunteer Activities

Potential activities include:

- Reviewing and testing the existing analysis platform.
- Developing new analyses and statistical summaries.
- Creating novel visualizations to better understand community structure and trends.
- Investigating collaboration networks among authors and institutions.
- Identifying highly influential researchers, organizations, and geographic regions.
- Improving software documentation and reproducibility.

Volunteers are encouraged to propose and implement additional analyses that they believe would provide useful insights into scientific communities.

Recommended Skills

- Interest in software development, data analysis and graphical visualization.
- Python programming experience is **recommended but not required**.
- Experience with GitHub, Docker, data visualization libraries, or bibliometric analysis is **beneficial but not required**.
- Willingness to learn new technologies and work independently.

Learning Outcomes

Participants will gain experience in:

- Bibliometric and publication analysis.
- Data visualization and dashboard development.
- Bioinformatics software development.
- Open-source development workflows.
- Scientific communication and presentation.

Deliverables

- New analytical modules added to the existing software platform.
- Additional visualizations and community metrics.
- Documentation of methods and software enhancements.
- Final presentation during the Fall 2026 Volunteer Symposium.
- Contributions to the project's public source code repository.

Participation

- **Maximum of two students** will be selected for this project.
- Suitable for undergraduate students and advanced high school students with an interest in informatics, data analytics and data visualization.

Glycoscience Educational Chatbot

Point of Contact: Rene Ranzinger, Sujeet Kulkarni and Urnisha Bhuiyan

Background

As scientific knowledge continues to expand, researchers, educators, and students increasingly rely on digital tools to access and understand complex information. Glycoscience is a multidisciplinary field with a steep learning curve, and many newcomers struggle to navigate the extensive terminology, concepts, and biological processes described in foundational educational resources such as the *Essentials of Glycobiology* book.

To improve accessibility and support self-directed learning, the GlyGen team has developed an initial alpha version of an educational chatbot capable of answering glycobiology-related questions using content derived from *Essentials of Glycobiology* book. The current prototype is based on a Retrieval-Augmented Generation (RAG) architecture, utilizes an large language model (LLM), and has been implemented using a Python and Docker-based software stack.

The next phase of development focuses on transforming this prototype into a beta system suitable for deployment and evaluation by a selected group of researchers and educators.

Project Description

The goal of this project is to develop a beta version of the Glycoscience Educational Chatbot that can be deployed on a web server and evaluated by a pilot group of users.

The student will analyze the strengths and limitations of the current alpha version and implement improvements that enhance usability, reliability, performance, and safety. Particular emphasis will be placed on improving the system's guardrails to ensure that responses remain focused on educational content, minimize hallucinations, and appropriately handle questions outside the scope of the knowledge base.

The completed beta version will support formal user testing by researchers and educators, providing valuable feedback for future development and broader deployment.

Existing Software Foundation

An alpha version of the chatbot has already been developed and includes:

- Retrieval-Augmented Generation (RAG) architecture.

- Knowledge base derived from *Essentials of Glycobiology*.
- Large language model integration.
- Python-based backend services.
- Dockerized deployment environment.
- Initial user interface and query workflow.

The project will focus on enhancing and extending the existing system rather than building a new chatbot from scratch.

Volunteer Activities

Potential activities include:

- Reviewing and testing the existing chatbot architecture.
- Evaluating chatbot response quality and identifying failure modes.
- Improving retrieval strategies and document indexing.
- Enhancing prompt engineering and response generation workflows.
- Developing and testing chatbot guardrails and safety mechanisms.
- Implementing features to improve transparency and citation of source material.
- Improving system configuration, deployment, and monitoring capabilities.
- Analyzing user feedback and identifying areas for future improvement.
- Documenting software components and deployment procedures.

Volunteers are encouraged to propose additional features that improve the educational value, usability, or reliability of the system.

Recommended Skills

- Interest in artificial intelligence, bioinformatics, or educational technologies.
- Interest in software development and scientific communication.
- Experience with Python is **recommended but not required**.
- Experience with Docker, web development, GitHub, machine learning, or LLM applications is **beneficial but not required**.
- Willingness to learn new technologies and work independently.

Learning Outcomes

Participants will gain experience in:

- Large language model applications in science education.
- Retrieval-Augmented Generation (RAG) systems.
- AI safety and guardrail development.
- Software testing and evaluation.
- User-centered design and feedback analysis.
- Open-source software development practices.
- Scientific communication and presentation.

Deliverables

- A deployable beta version of the Glycoscience Educational Chatbot.
- New features and improvements added to the existing software platform.
- Testing and evaluation documentation.
- User feedback summary and recommendations.
- Software documentation and deployment instructions.
- Final presentation during the Fall 2026 Volunteer Symposium.
- Contributions to the project's public source code repository.

Participation

- **Maximum of one student** will be selected for this project.
- Suitable for undergraduate students or highly motivated high school students with programming experience and an interest in artificial intelligence, educational technology, or bioinformatics.

GlyGen AI-Assisted Biocuration Project: Species, Tissue, and Disease Annotation

Points of Contact: Rene Ranzinger and Urnisha Bhuiyan

Background

A major challenge in biological data integration is the harmonization of metadata from diverse sources. Scientific databases, publications, and legacy datasets often describe species, tissues, and diseases using free-text terms that include abbreviations, synonyms, spelling variations, and ambiguous expressions. Before these data can be integrated into resources such as GlyGen, the metadata must be mapped to standardized ontologies and controlled vocabularies.

During the Summer 2026 volunteership, a student developed and evaluated a large language model (LLM)-based workflow for mapping species names to identifiers in the NCBI Taxonomy. The project utilized Python and the ChatGPT API to automate the interpretation and normalization of species-related metadata. Building on these results, GlyGen aims to continue and expand this work to support additional annotation domains, including tissues and diseases.

Project Description

The goal of this project is to improve and expand GlyGen's AI-assisted biocuration workflows for metadata normalization and ontology mapping. The volunteer will review the existing species annotation pipeline, evaluate its performance, and refine the species-specific system prompt used by the LLM. Based on lessons learned from species mapping, the project will then extend the methodology to additional biomedical concepts, particularly tissue and disease annotations.

A major focus of the project will be prompt engineering, performance evaluation, and quality assessment. The student will investigate how well the LLM can identify the correct ontology terms when presented with real-world biomedical metadata containing abbreviations, synonyms, misspellings, and incomplete descriptions.

The resulting workflows will support GlyGen's ongoing efforts to harmonize metadata from publications, databases, and legacy resources, ultimately improving data quality and interoperability across the glycoscience ecosystem.

Existing Software Foundation

A prototype annotation system has already been developed and includes:

- Python-based annotation workflows.

- Integration with the ChatGPT API.
- Species mapping using NCBI Taxonomy identifiers.
- Evaluation datasets for performance assessment.
- Initial validation results generated during Summer 2026.

The project will focus on improving and extending the existing framework rather than building a new system from scratch.

Volunteer Activities

Potential activities include:

- Reviewing and testing the current species annotation workflow.
- Evaluating mapping accuracy and identifying failure cases.
- Refining and optimizing species-specific system prompts.
- Investigating prompt engineering strategies to improve consistency and accuracy.
- Extending the workflow to disease annotation using biomedical disease ontology.
- Extending the workflow to tissue annotation using standardized anatomy ontologies.
- Comparing LLM-generated mappings with manually curated reference datasets.
- Documenting annotation guidelines and best practices.
- Working with GlyGen mentors to analyze challenging annotation cases and edge conditions.

Volunteers will be encouraged to propose additional evaluation approaches and improvements that strengthen the reliability of the annotation process.

Recommended Skills

- Interest in biomedical research, bioinformatics, biocuration, or medical informatics.
- Educational background in biology, biomedical sciences, health sciences, medicine, or a related discipline.
- Strong attention to detail and interest in data quality.
- Python programming experience is **beneficial but not required**.

- Familiarity with ontologies, controlled vocabularies, or AI tools such as ChatGPT are **beneficial but not required**.

Learning Outcomes

Participants will gain experience in:

- Biomedical data curation.
- Ontology and terminology mapping.
- Prompt engineering for large language models.
- Evaluation of AI systems in scientific workflows.
- Metadata harmonization and quality control.
- Scientific communication and presentation.

Deliverables

- Improved species annotation prompts and workflows.
- Prototype workflows for tissue and disease annotation.
- Evaluation reports comparing automated and manual annotations.
- Documentation of prompt designs, validation methods, and best practices.
- Final presentation during the Fall 2026 Volunteer Symposium.
- Contributions to the project's public source code repository.

Participation

- **Maximum of one student** will be selected for this project.
- Students with a background in biology, biomedical sciences, medicine, public health, or related disciplines are especially encouraged to apply.

Glycoscience Resource Discovery and Search Platform

Point of Contact: Rene Ranzinger and Urnisha Bhuiyan

Background

The glycoscience research community relies on a diverse ecosystem of databases, software tools, analysis platforms, and educational resources. One widely used catalog of glycoscience resources is maintained through the Glyco@Expasy portal, which organizes databases and tools using a visualization based on intersecting circles and predefined resource categories.

While this approach provides a valuable overview of the available resources, it can become difficult for users to identify the most relevant databases and tools for a specific research question, particularly as the number of available resources continues to grow.

During Spring 2026, a volunteer developed a prototype alternative discovery platform inspired by modern e-commerce and product discovery websites. The prototype allows users to browse resources using search and filtering functions, view resources as information cards, and access additional details through expandable displays. The system currently uses a JSON-based data store and a lightweight web technology stack.

The next phase of development focuses on transforming this prototype into a production-ready application and exploring the use of large language models (LLMs) to support natural language resource discovery.

Project Description

The goal of this project is to develop a production-ready search and discovery platform for glycoscience databases and software tools. The volunteer will enhance an existing prototype that provides a modern, user-friendly alternative to traditional resource catalogs. The platform will enable researchers to efficiently discover relevant databases, software tools, and analytical resources based on their scientific needs.

A major component of the project will be expanding the underlying resource catalog to include not only databases but also software tools and analysis platforms. The volunteer will improve the platform's filtering and search capabilities and investigate how LLMs can be incorporated to support natural language queries.

Existing Software Foundation

A functional prototype has already been developed and includes:

- Card-based resource browsing interface.

- Search functionality.
- Category-based filtering.
- Expandable resource detail views.
- JSON-based resource repository.
- Web-based user interface implemented using standard web technologies.

The primary objective is to extend and mature the existing prototype rather than creating a new system from scratch.

Volunteer Activities

Potential activities include:

- Reviewing and testing the existing prototype.
- Expanding the resource collection to include software tools and analysis platforms.
- Improving resource classification and metadata organization.
- Enhancing search functionality and filtering options.
- Designing new user interface components to improve usability.
- Integrating and evaluating LLM-based natural language search capabilities.
- Documenting the software architecture and deployment procedures.

Volunteers are encouraged to propose additional features that improve resource discovery, accessibility, or user experience.

Recommended Skills

- Interest in web technologies, artificial intelligence, or software development.
- Web development experience is **beneficial but not required**.
- Python programming experience is **beneficial but not required**.
- Experience with JavaScript, JSON, GitHub, UI/UX design, search technologies, or LLMs are **helpful but not required**.

Learning Outcomes

Participants will gain experience in:

- Resource discovery systems.

- Web application development.
- Search and information retrieval methodologies.
- Large language model applications.
- Open-source software development.
- Scientific communication and presentation.

Deliverables

- A production-ready version of the glycoscience resource discovery platform.
- Expanded database and software tool catalog.
- Enhanced search and filtering capabilities.
- Prototype implementation of LLM-assisted natural language search.
- Software documentation and deployment instructions.
- Final presentation during the Fall 2026 Volunteer Symposium.
- Contributions to the project's public source code repository.

Participation

- **Maximum of one student** will be selected for this project.
- Suitable for undergraduate students or advanced high school students interested in web development, information retrieval, or artificial intelligence.