HARNESSING THE POWER OF AI IN SIALOBIOLOGY RESEARCH & EDUCATION

Arun K. Datta
Department of Computer Science & Cybersecurity, National University, San Diego, CA, USA

Glyco-codes are complex carbohydrate structures consisting of ten monosaccharides that encode important information for various biological processes in human including cell-cell interactions, extracellular signals, and cell differentiation. Glycomics is more difficult than other -omics because it does not follow the linear flow of Central dogma. Rather, it needs meticulous and detailed analysis of glycosyltransferases involved in generating the glyco-codes. However, recent developments in the machine learning technology, particularly ***GenAI***, has created an unprecedented opportunity that can tremendously help in such analysis. In our developmental effort of ***Glycomics Workbench***, this technology has been utilized for aiding glycan research and education. This author finds that a lot of sialic acids related information can be obtained within minutes utilizing this technology that was earlier needed months if not years of manual curation.

For the development of Glycomics Workbench portal that would be supported by the state-of-the-art Cyberinfrastructure (CI), we are utilizing Grid computing, Grid services, and data Grid. Grid technology offers multiple advantages including high scalability and Web accessibility. Grid computing provides accessibility to High-Performance Computing, such as, ACCESS of NSF. Grid services built on Open Grid Collaborating Environments (OGCE) are based on several Web service technologies. Our earlier work on CHOIS was built on OGCE. Data grid is a commodity grid that can host exabytes of data that has become essential for glycome analysis. Our earlier work on C-Grid development was to fulfill such a need. Moreover, a significant number of software tools and databases have been developed over the years for glycome analysis. However, utilization of some of those useful tools is restricted due to the fact that those are not accessible via Web. This also restricts the semantic analysis of a vast amount of experimental data, such as, that were generated under the Consortium of Functional Glycomics (CFG). Our proposed Glycomics Workbench, is designed to integrate such useful computational tools and resources for glycomics that can better serve the Glyco-community.

PROBLEM STATEMENT

- Glycomics is difficult compared to other -omics.
- A significant number of databases and software tools have been developed over time that are useful for Glycome Analysis (1). However, accessibility of some of these tools via Web are restricted due to the technology platform used.
- Gathering the glycan related information is now manual making it tedious and time consuming.

SOLUTION

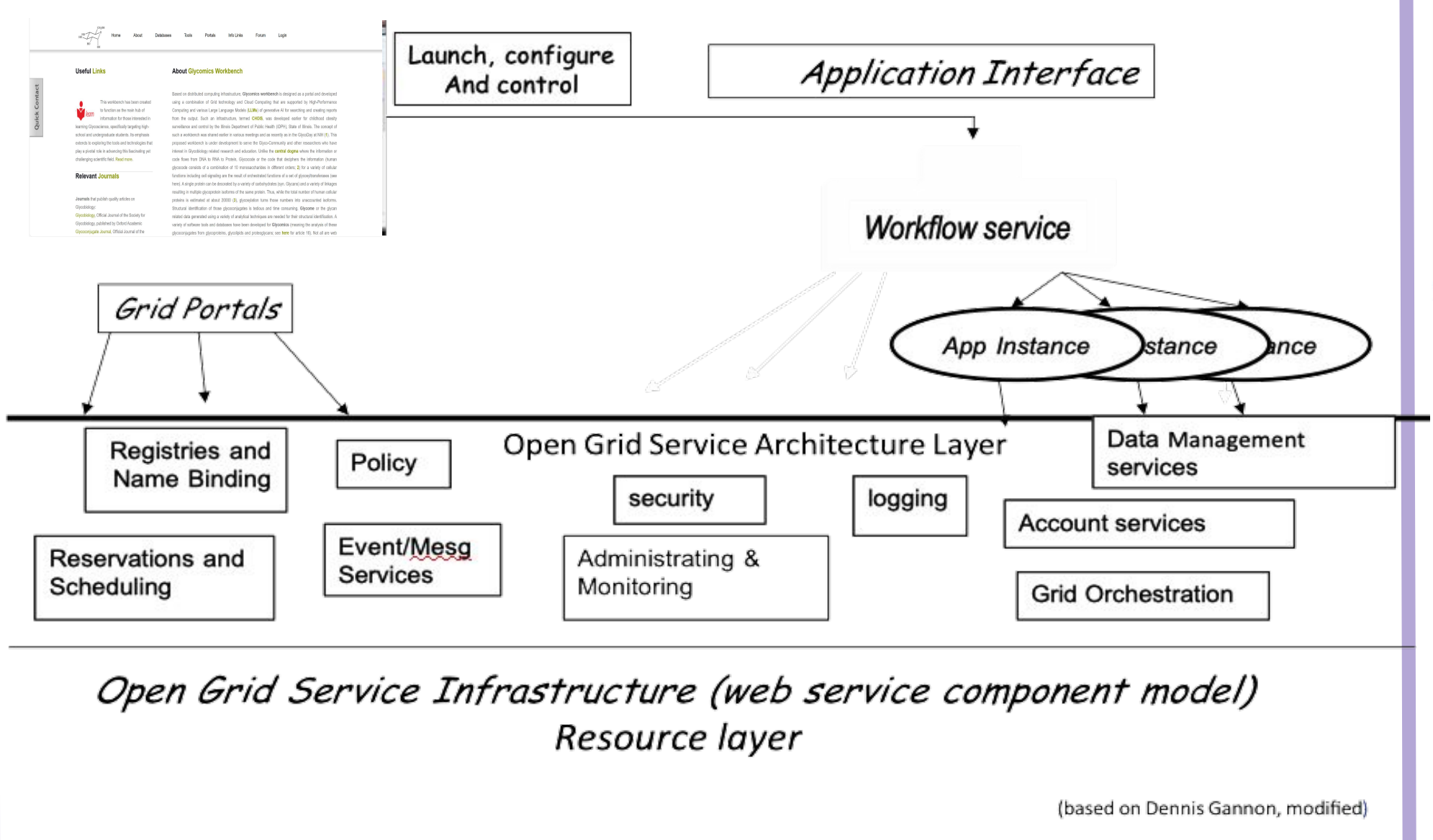
- Developing Glycomics Workbench (Figure 1) for aiding analysis of large amount of data generated on glycans (2).
- As done earlier (3), OGCE will be used for providing scalability and Web accessibility of useful software tools, databases, and other relevant resources.
- C-Grid to store massive amounts of data generated from the analysis (4). APIs will be developed for accessing Grid services and Large Language Models (LLMs) for generative AI (GenAI).
- ACCESS for Grid Computing and related services.

MOTIVATION

- This project is to utilize Grid Technology in our advantage for the development of Glycomics Workbench that will provide Web accessible framework, which can be easily scalable (3, 5).
- Recent development of GenAI (6) has created an opportunity to gather information quickly in unprecedented way. In our effort for developing the proposed Workbench, this powerful tool will now be integrated for gathering relevant data that can be utilized for glycan research and education.

OBJECTIVE

- In collaboration with others, this author is engaged in developing tools and technologies that can benefit glycoscientists, students, and educators.
- Developing Grid technology-based portal to store, manage and share large amounts of glycan related data in a distributed data grid for further analysis by the researchers in a collaborative environment.
- Develop APIs and user-friendly GUIs for accessing the software tools, databases, and relevant information.
- Use of Machine Learning Technology to develop solutions and aiding research and education.


Launch, configure And control
Application Interface
Workflow service
Grid Portals
App Instance
stance
ance
Registries and Name Binding
Policy
Open Grid Service Architecture Layer
Data Management services
security
logging
Account services
Reservations and Scheduling
Event/Mesg Services
Administrating & Monitoring
Grid Orchestration
Open Grid Service Infrastructure (web service component model)
Resource layer

(based on Dennis Gannon, modified)

Figure 1. The proposed CI-supported Grid technology-based Glycomics Workbench. This has three fundamental components: Software-as-a service, C-Grid for distributed storage of data and grid computing supported by HPC/ACCESS. This is also designed to support education and training utilizing GenAI. Data generated will be stored in the C-Grid (figure adapted from 3).

C-GRID

- C-Grid, the Community Grid web portal developed earlier (4, Figure 2), would serve as a gateway for the collaborative institutions and organizations to utilize CI supported resources for data analysis and helps the users to create and manage "virtual data collection" that can be stored in heterogeneous data resources across distributed network.
- Remote management of this data grid is performed using iRODS, the Integrated Rule-Oriented Data System, which is a middleware developed by the Data Intensive Cyber Environments (DICE) research group, and collaborators.
- GUI component, termed ez-PRODS created earlier (Figure 3) as a component of C-Grid to interact with iRODS located in the data grid using PRODS API (7).
- Video and other big data can be stored in C-Grid (4, Figure 4).

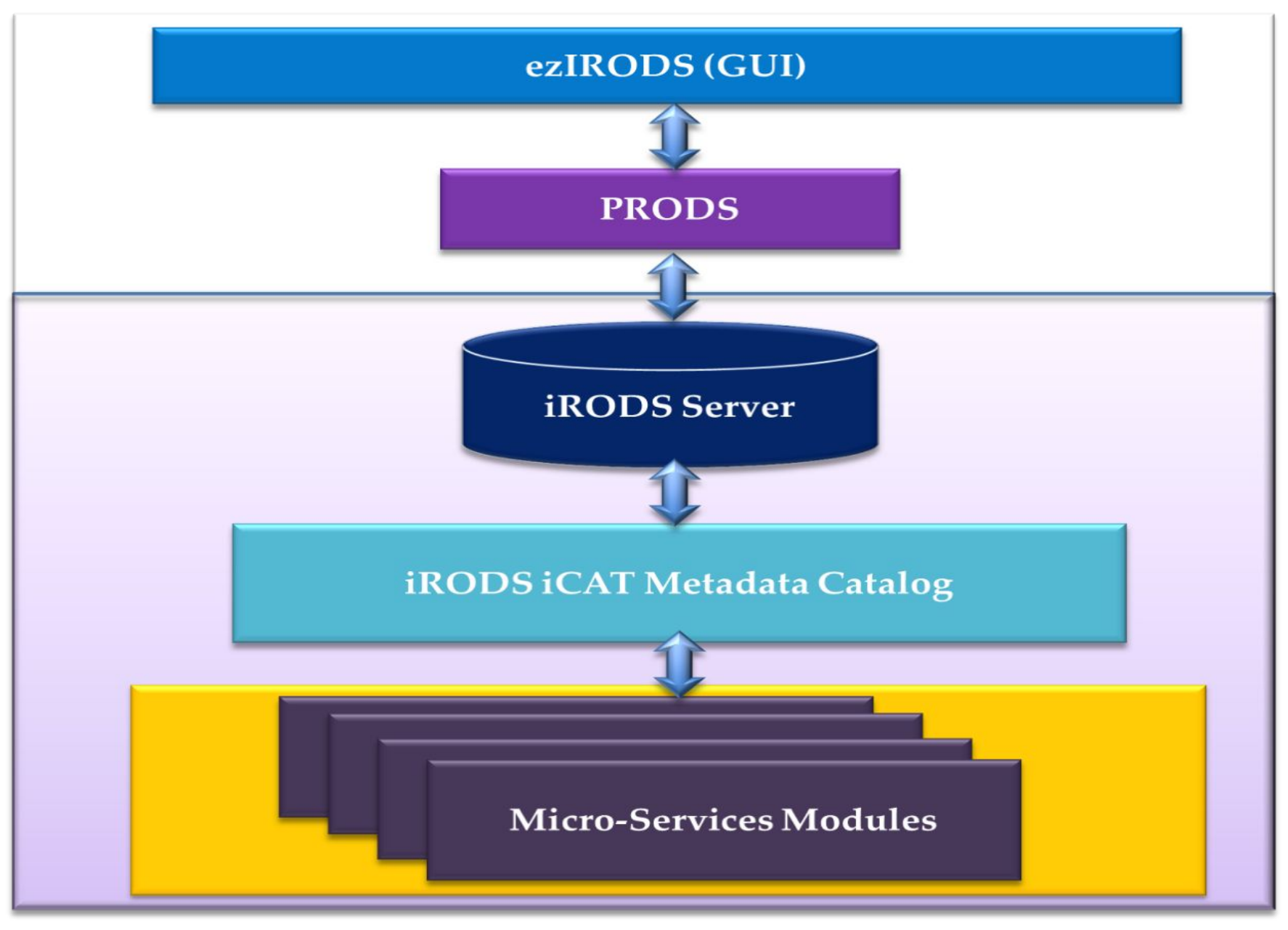
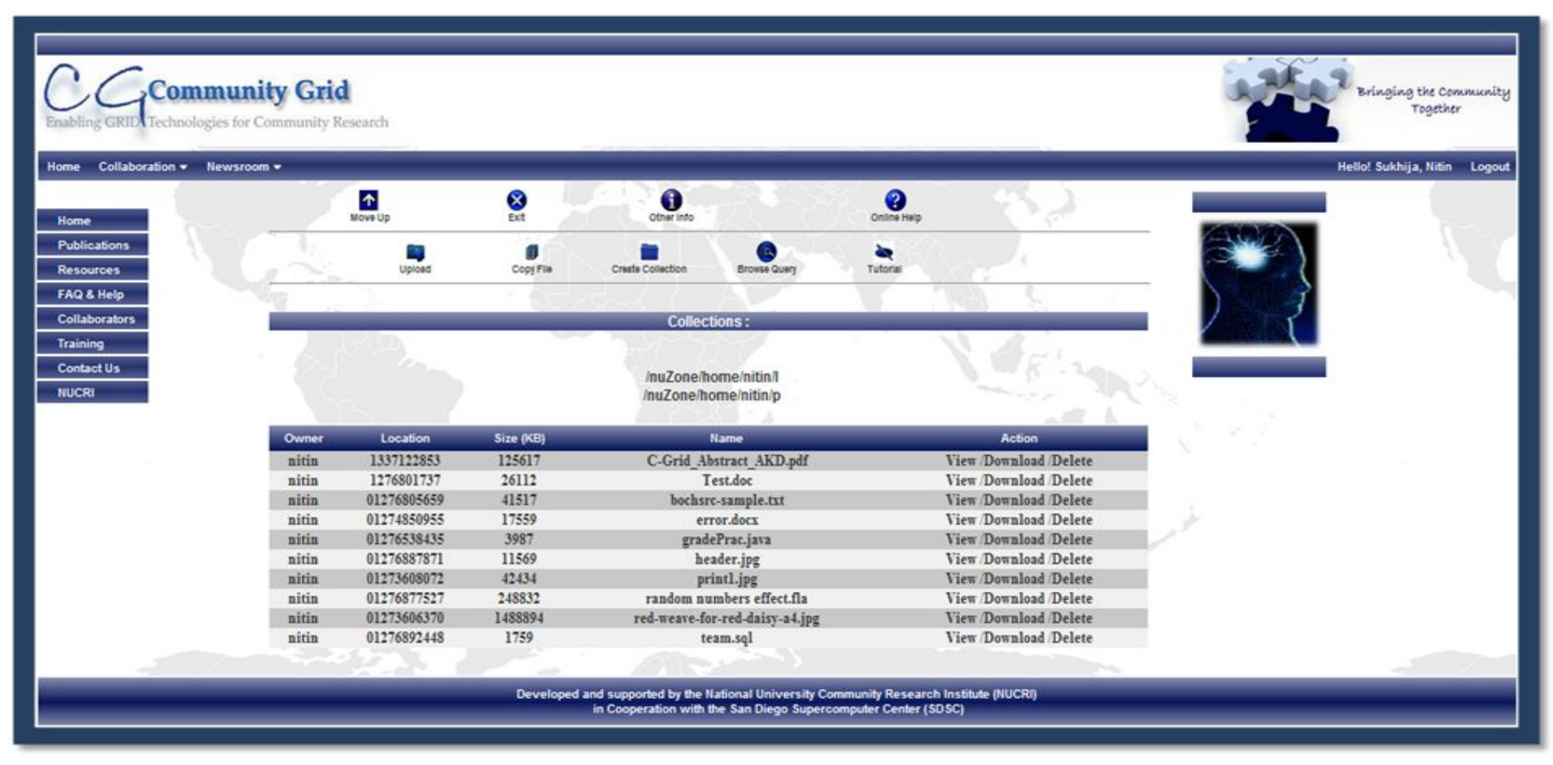


Figure 2. C-Grid or Community Grid Portal. Portal framework.



/nuZone/home/nitin/l
/nuZone/home/nitin/p

Owner	Location	Size (KB)	Name	Action
nitin	1337122853	125617	C-Grid_Abstract_AKD.pdf	View /Download /Delete
nitin	1276801737	26112	Test.doc	View /Download /Delete
nitin	01276805659	41517	bochsrc-sample.txt	View /Download /Delete
nitin	01274850955	17559	error.docx	View /Download /Delete
nitin	01276538435	3987	gradePrac.java	View /Download /Delete
nitin	01276887871	11569	header.jpg	View /Download /Delete
nitin	01273608072	42434	printl.jpg	View /Download /Delete
nitin	01276877527	248832	random numbers effect.fla	View /Download /Delete
nitin	01273606370	1488894	red-weave-for-red-daisy-a4.jpg	View /Download /Delete
nitin	01276892448	1759	team.sql	View /Download /Delete

Developed and supported by the National University Community Research Institute (NUCRI) in Cooperation with the San Diego Supercomputer Center (SDSC)

Figure 3. C-Grid or Community Grid Portal. Portal interface for viewing data collections in C-Grid after role-based access with authentication (4).

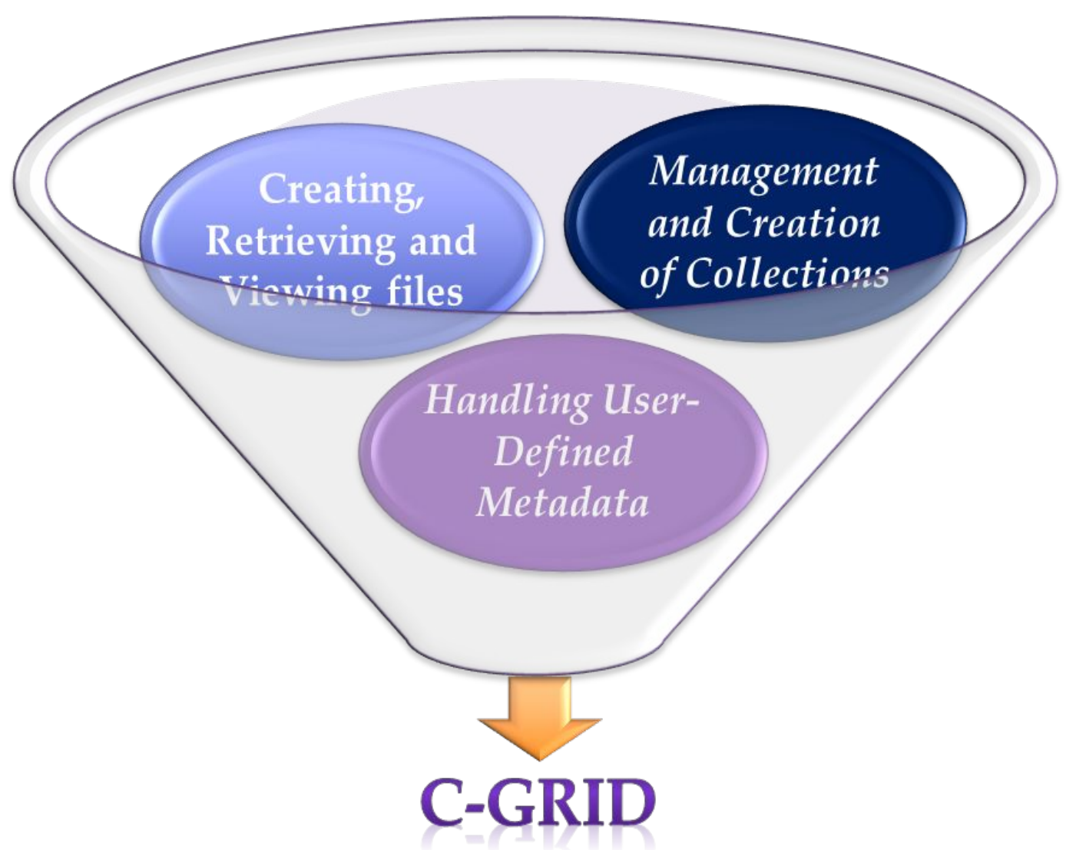

Creating, Retrieving and Viewing files
Management and Creation of Collections
Handling User-Defined Metadata
C-GRID

Figure 4. Core functionality offered by C-grid web portal.

LLMs for GenAI

- APIs for GenAI will be developed utilizing various LLMs that will serve to create educational materials, which will also serve to aid the research.
- Particular emphasis will be on Gemini of Google (Figure 5), which currently provides more than 1.5 million token capabilities.
- APIs for other machine learning models, such as, DALL-E 3, Grammarly, etc. will also be developed as needed.

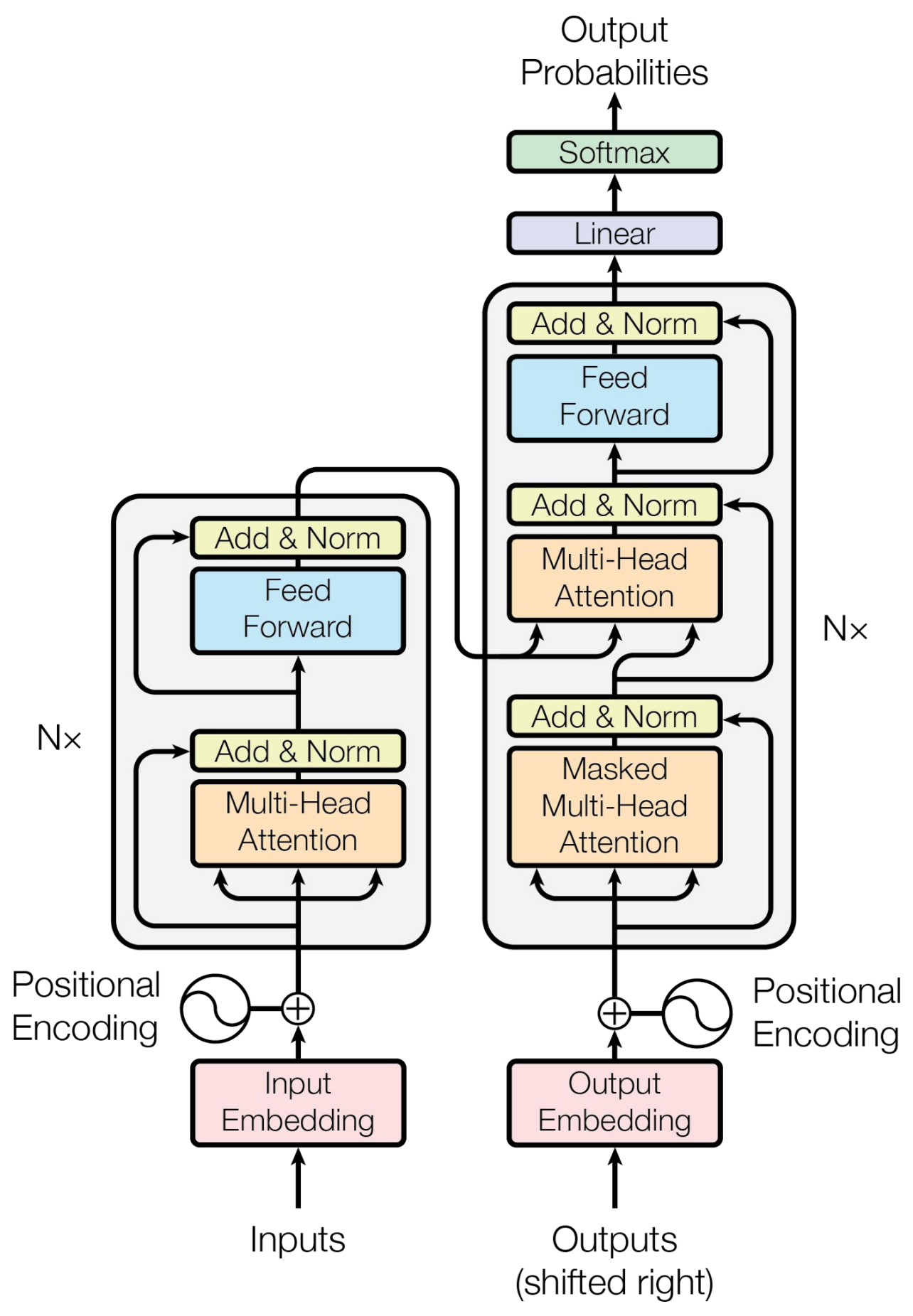


Figure 5: The Transformer - model architecture (6) serves as the 'brain' for GenAI.

CONCLUSION

- Recent developments of various LLMs for Machine Learning has provided an unprecedented opportunity to quickly generate a lot of information on sialic acids (8, 9) that can be utilized for research and education.
- Glycomics Workbench is under development that will utilize these LLMs for Glycoscience research and education.
- C-GRID of Glycomics Workbench would provide collaborative data sharing and maintenance of distributed storage resource collections. It would also address the challenge of dealing with the problem of scalability of data and the data visualizations.

FUTURE WORK

Along with the LLMs, this proposed framework will implement CI for supporting Glycan research and education.

ACKNOWLEDGEMENT

This author acknowledges the contribution of N. Sukhija of Slippery Rock University (PA, USA) for the development of C-Grid under collaboration.

REFERENCES

1. Datta, AK., and Sukhija, N. (2021). Glycobioinformatics in deciphering the Mammalian Glycocode: Recent Advances In "Glycome: The Hidden Code in Biology" (D. Banerjee, ed.) Nova Science Publishers, article 16. ISBN: 978-1-53619-377-0.
2. Datta, AK., and Sukhija, N. (2020). Glycomics Workbench, a grid technology-based workbench for Glycome Analysis. Proceedings in the 13th annual NIH & FDA Glycoscience Research Day, Bethesda (Maryland), May 15, 2020.
3. Datta, AK., Jackson, V., Nandkumar, R., Sproat, J., Zhu, W., Krahlin, H. (2010). CHOIS: Enabling grid technologies for obesity surveillance and control. In 'Healthgrid Applications and Core Technologies (Eds. , T. Solomonides, I. Blanquer, V. Breton, T. Glatard, and Y. Legre), vol 159, p191-202, IOS Press, Washington D.C. ISBN 978-1-60750-582-2; Presented at the HealthGrid 2010, Paris, France.
4. Sukhija, N., and Datta, AK (2013). C-Grid: Enabling iRODS-based Grid Technology for Community Health Research. Information Technology in Bio- and Medical Informatics, Lecture Notes in Computer Science (M. Bursa, S. Khuri, and M. E. Renda, Eds.), LNCS 8060, pp 17-31, Springer-Verlag, Berlin, Heidelberg, ISBN 978-3-642-40092-6).
5. Foster, I., Kesselman, C (2003). The grid 2: Blueprint for a new computing infrastructure. Morgan Kaufmann Publishers. San Francisco (CA), USA.
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, AN., Kaiser, L., Polosukhin, I. (2017). Attention Is All You Need. arXiv:1706.03762.
7. Sukhija, N., Datta, A., Sevin, S., Coulter, JE. (2018). Grid Technology for Supporting Health Education and Measuring the Health Outcome ACM ISBN 978-1-4503-6446-1/18/07; presented at the PEARC'18, July 22–26, 2018, Pittsburgh, PA, USA.
8. Schauer R, Kamerling JP. (2018). Exploration of the Sialic Acid World. Adv Carbohydr Chem Biochem. 2018;75:1-213.
9. Varki A, Schauer R. (2009). Sialic Acids. In: Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Bertozzi CR, Hart GW, Etzler ME, editors. Essentials of Glycobiology. 2nd edition. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2009. Chapter 14.

SCAN ME