RITWIK GANGULY

+91-629-567-2578 | ritwik24222@iiitd.ac.in

LinkedIn | GitHub | Medium | ORCiD

Kolkata, West Bengal, India

ABOUT

A highly energetic individual, who aspires to learn new things everyday. Having deep foundations of **Computer Science** and **Computational Biology**, I have the experience in **single cell genomics** and done its **synthetic cell generation using graph attention based GAN**. Alongside, done the **transformer fine-tuning** and **RAG-enabled LLMs** over biomedical **knowledge graphs**, demonstrates my strong computational and biological foundations. Currently, I'm advancing **cancer genomics and multi-omics systems** through **heterogeneous GNN architectures** (HGT, GAT, GCN). With these diverse toolkit, I thrive on translating cutting-edge research in **Single Cell, Spatial and Cancer Genomics** with computational technology to make the biological problems more tractable. This interdisciplinary knowledge of **CS fundamentals** and **Modern Biology foundation** make me a **promising candidate** for the **AI centric Biology research and novel exploration**.

EDUCATION

- **Indraprastha Institute of Information Technology (IIIT Delhi)** *July 2024 - June 2026*
  M.Tech in Computational Biology (CB) Delhi, India
  - GPA: 8.75/10 (persuing)
- **Aliah University** *June 2024*
  B.Tech in Computer Science & Engineering (CSE) Kolkata, India
  - GPA: 9.18/10
  - Consolidated Marksheet
- **Kalna Maharajas' High School** *July 2020*
  Higher Secondary Kalna, India
  - Marks (Percentage): 91.4%
  - Marksheet
- **Ichapur S.G. High School** *May 2018*
  Secondary Ichapur, India
  - Marks (Percentage): 92%
  - Marksheet

EXPERIENCE

- **Graduate Researcher - @The Sengupta Lab [🌐]** *July 2024 - present*
  Guide - Dr. Debarka Sengupta Delhi, India
  - Initially worked on **LLM** and **Information Retrieval (IR)** to retrieve the essential biological Information from large **Knowledge Graph (KG)** (e.g. PrimeKG, EnrichR KG), used the **single agentic RAG**, with LLM models.
  - Implemented a **PDF summarizer** named **"Ask Your PDF"**, with the LLM, vector database (chromaDB) and **RAG system**. [GitHub Link] [Demo]
  - With this RAG-enabled LLM knowledge, also build one application, that will **read the medical diagnostic PDF document and tables**, then extract and load the data to excel. This application is **very efficient**.
  - In the initial phase of my M.Tech thesis, worked on sc-rna sequencing, its prepossessing (scanpy) and batch correction and cell type annotation **(scGPT, scANVI)**.
  - Currently I am working in - **Cancer Genomics** and especially in **Triple Negative Breast Cancer (TNBC)** and its SubType Identification with **Multi-omics** data and KG-based GNN approach.
  - In this lab, got the full opportunity to work in **graph-structured data**, especially with heterogeneous graph. Currently work on **KG oriented steiner tree** with GNN architecture in Cancer Genomics.
- **Under Graduate Researcher - CSE Dept @AU [🌐]** *Aug 2023 - May 2024*
  Guide - Dr. Sumanta Ray, Dr. Sk. Md. Mosaddek Hossain Kolkata, India
  - In this lab, initially working on scrna-seq pre-processing and its underlying data transformation.
  - Worked on **Generative Adversarial Network (GAN)** and its diverse architecture, along with f-GAN, w-GAN, VAE.
  - Later worked at a novel Project **"synthetic rna-seq single cell generation using graph attention based GAN"**, where we generated the synthetic svrna-seq cell samplesin conjugation with **used GAT graph attention technique**.
  - From this research lab, got the initial learning and experience of CB and most specific **the single cell genomics**, and got to know, how the computational techniques can detect the heterogeneity in the Biological Data more efficiently.

- **Data Science Intern - Innomatics Research Lab [🌐]** *Feb 2023 - May 2023*
  Guide - Mr. Kanav Bansal Hyderabad, India (remote)
  Perks: [Certificate] [LOR]
  - I primarily worked, how data getting analyzed in large scale, EDA techniques and basic statistical modeling to get the insights from this, make some cool dashboard to get better visualization.
  - Later, enters the **NLP group** and particularly in NLP driven linguistic problems and build - **Laptop Price Recommender**, here used **RandomForest** and **word2vec embedding** and **streamlit** for deployment. [GitHub]
  - Later, worked on two supervised NLP project - **Fake News Detection** and **Spam Email Detection**, used **word2vec** and therefore **TfIdf** technique and **cosine-similarity** for the detection of the Binary Class samples. [GitHub]
  - Appointed as **Team Lead** of the **"Healthcare Chatbot" (RISA)** project using **RASA** framework with my team **@The Skill Prodigies**, after being ranked the top performer of the batch.

PROJECTS

- **A Graph-Attentive GAN for Rare-Cell–Aware single-cell RNA-seq Data Generation**
  Guide: Dr. Sumanta Ray, Dr. Mosaddek Hossain AUG 2023 - AUG 2025
  - This is the **Thesis Work** in my B.Tech curriculum.
  - Acquiring single-cell RNA-seq (scRNA-seq) data is often limited by high costs and strict patient privacy, restricting its use in downstream analysis. We propose a **Novel Method** to overcome these barriers.
  - In our work, we build one custom **GAN** architecture with **Graph attention Network**(GAT), that makes our model robust. We used a **pytorch** based GAT architecture and at the GAN using **tensorflow & Keras**.
  - Initially we build sample specific graph using **KNN Graph** approach, later used the GAT architecture to put the attention in the nodes based on the cell types.
  - Our intention was to generate synthetic scrna-seq samples. And also, other than generating the cell samples from random noise (that traditional GAN does), we merge some k% of real data with random noise that enhance the data generalization power to our **Model Architecture**.
  - Later after data generation, we calculated the **ARI**, **NMI** and **macro-f1 score** metric and beat almost all **SOTA models for sc-rna seq data generation**.
- **RISA: A Healthcare Chatbot**
  Guide: Mr. Kanav Bansal APR 2023 - JUNE 2023
  - Collected Disease Symptoms Predictor Dataset from **Kaggle**, done data cleaning, required data analysis.
  - We built the **RISA**, using the **RASA NLU** framework, along with at the backend, used **NLG** power with **document wise Vector Embedding** of **disease-symptoms pair** using **BioBert pre-trained model**, that make RISA robust.
  - Also, to make it robust, designed the **RASA own Knowledge Base** with several rules and training examples.
  - At backend, user utterance do **similarity check** using **cosine similarity** and return the most prominent utterance.
  - Also, **RISA** has also 2 important features: it also can answer any disease and symptoms information in 500 words, for that used **wikipidiaAPI** and also it has the **nearest hospital locator** feature with **Folium** map-view.
  - For the **Nearest Hopital Finder**, initially I was scrapped knowledge-base. For a Broad aspect, I am working on the **Geocoding API** of **Google Cloud**, that can capture any location in the world.
  - In future, there is a plan to use the **hugging Face Symptoms2Disease dataset**, to make the training data larger.
- **Bayesian Explainability for Real-Time Anomaly Detection in Medical Diagnostics**
  Guide: Dr. Ranjitha Prasad Oct 2024 - Dec 2024
  - In the modern healthcare landscape, **patient safety and quality of care** have become paramount. **Advanced ML models** have being adopted to **analyze patient data**, **identify anomalies**, and **assist clinicians in decision-making**.
  - But the ML models are **blackbox** as, they have not any interpretability and transparency to their prediction.
  - I used the MIMIC-III diagnostic dataset that consists of diagnostic reports of 40000 patients with abnormality labels.
  - Initially: built one **anomaly detection model** using **LSTM AutoEncoder** that will predict **patients' abnormality**.
  - We train our model on 70 % data and on the validation set we got 98 % accuracy at the **Anomaly Detection**.
  - Next: built one explainable model on the **top of LSTM-AutoEncoder**, used the Explainable AI (XAI) approach. I used the **LIME** and the Bayesian version of this, **BayesLIME** and did a comparative study.
  - For the **'abnormal'** patients, LIME gives the **Point Estimate** or the **Feature Rank**. Whereas, BayesLIME gives the **point estimate + the uncertainty range**. This interpret the uncertainty in Medical Diagnostic Anomaly Detection.
- **Triple-negative Breast Cancer (TNBC) SubType Classification using KG-based GNN Approach**
  Guide: Dr. Debarka Sengupta May 2025 - Present
  - This Work is about my **MTech thesis**, and working on most heterogenous breast cancer **TNBC and its subtyping**.
  - TNBC, lacking hormone receptors **(ER, PR) and HER2 expression**, which makes it unresponsive to targeted hormonal or HER2 therapies. So, subtypes of TNBC will guide to **personalized therapy**, as certain subtypes respond better to specific drugs.

- Collected 4 omics data including, microarray, rna-seq, methylation and CNV with their TNBC subtype (BL1/BL2/LAR/M). Integrate these multi-omics and use in downstreaming is our one of the keys.
- As, some papers have different nomenclature of subtypes, did the **TNBC Subtype mapping** to 4 subtypes.
- Next, for all samples, performed the **GSEA Analysis** to validate the subtype-specific pathway enrichment and **Refernce Component Analysis (RCA)** to validate & correlate the biological cell type specific TNBC Subtype.
- Currently working in **Knowledge Graph based dual-view GNN approach** on heterogeneous graph topology, to get the **subtype classification** as well as **graph explainability** and therefore build a **robust TNBC Subtype Classifier**.
- Also, working in another approach of **Conditional VAE (CVAE) based classification** and Bayesian Explainability using BayesLIME, to build our approach more robust towards explainability.

Publications

- **A Graph-Attentive GAN for Rare-Cell–Aware single-cell RNA-seq Data Generation**
  First Author 2025
  - Submission Link - [BioRxiv Link] [Genome Biology Link]
- **Mechanism-aware inference of response to targeted cancer therapies**
  Co-Authors 2025
  - Pre-print Link - [BioRxiv Link]
- **AI for Computational Biology: Highlights from the first BioAI Hackathon at University of Warsaw**
  Co-Authors 2025
  - Pre-print Link - [BioHackrXiv Link]

Skills

- **Programming Languages:** Python, R, SQL, git
- **Data Analysis & Visualization:** Pandas, NumPy, Matplotlib, Seaborn, Plotly, Folium
- **Database Systems:** MYSQL, MongoDB, Neo4j, ChromaDB
- **Data Science & Machine Learning:** Scikit-learn, Pytorch, Tensorflow, Streamlit
- **Graph Neural Network (GNN):** Pytorch Geometric, GCN, GAT, GraphMAE, HeteroGNN, Hetero Data
- **Natural Language Processing (NLP):** RNN, LSTM, Autoencoder, Vector Embedding, RASA
- **Large Language Model (LLM):** Langchain, RAG, Hugging Face, LangGraph
- **Genomics:** Deconvolution (BayesPrism), GSEA analysis (pygsea, gsva), single-cell preprocessing (scanpy), Multi-omics, scGPT, Omics Batch Correction (pycombat), RCA Projection
- **Specialized Area:** Cancer Genomics, Single Cell Genomics, ML, DL, GNN, NLP, RAG-enabled LLM, Finetuning LLM, Database Management, Knowledge Graph, Variational AutoEncoder (VAE)
- **Research Skills:** Linux, LateX, VScode, Excel, Overleaf, Inkscape

Achievements

- **GATE CS/IT & DS-AI 2024 Qualified**
  Organised by – IISc Bangalore 2024
  - Secured AIR 2856 in GATE DS-AI 2024.
- **B.Tech University 2nd Rank Holder**
  Aliah University 2020 – 2024
  - Ranked 2nd in the university among B.Tech graduates.
- **Participated at BioAI Hackathon**
  Organised by – University of Warsaw (CeNT) May, 2025
  - Our problem statement was - **"Toxicological Profiling Of Mol Compound Using GNN And Pre-Trained Molecular Embedding"** - [Perticipation Certificate] [Paper Link]

Additional Information

Languages:

- English (Professional working proficiency)
- Hindi (Limited working proficiency)
- Bengali (Native or bilingual working proficiency)

Interests: Cricket, Problem Solving, Scientific Reading

REFERENCES

1. **Dr. Debarka Sengupta**
   Professor, Department of Computational Biology & CS-AI
   Indraprastha Institute of Information Technology, New Delhi-110020, India
   Email: debarka@iiitd.ac.in
   Phone: +91-9831307912
   Relationship: M.Tech Thesis Supervisor in IIIT Delhi

2. **Dr. Sumanta Ray**
   Associate Professor, Department of Data Science
   The West Bengal National University of Juridical Sciences (NUJS), Kolkata-700098, India
   Email: sumantaray@nujs.edu
   Phone: +91-9231879956
   Relationship: B.Tech Thesis Supervisor in Aliah University

3. **Dr. Jaspreet Kaur Dhanjal**
   Assistant Professor, Department of Computational Biology
   Indraprastha Institute of Information Technology, New Delhi-110020, India
   Email: jaspreet@iiitd.ac.in
   Phone: +91 99712 66508
   Relationship: M.Tech Mentor & Guide in IIIT Delhi