Sayan Ghosal

Redwood City ✻ sghosal@chanzuckerberg.com ✻ https://sayangsep.github.io ✻ 443-531-5268

Professional Summary

- Computational scientist with a passion for integrating complex statistical models with genomic knowledge to provide insight into disease mechanisms.
- Contributions include novel Bayesian and machine learning models providing mechanistic insights into complex disorders like Alzheimer's, Autism, and Schizophrenia.
- Highly collaborative and motivated to drive new research endeavors in the intersection of methods development and biological discovery.

Academic Background

Johns Hopkins University, Baltimore, USA

Ph.D., Electrical and Computer Engineering 2023

M.S., Applied Mathematics and Statistics 2021

Jadavpur University, Kolkata, INDIA

B.E., Electronics and Telecommunication Engineering 2017

Professional Background

Research Scientist. AI/ML, Chan Zuckerberg Initiative, Redwood City *Present*

Contributions: Foundation Models, Generative Modelling, Multiomics

Computational Scientist, Broad Institute, Cambridge *Oct 2023 - Aug 2024*

Contributions: Graph-Based SV Discovery, Long/Short Read Sequencing, Disease Association

AI Resident, Google X, Mountain View *2022*

Contributions: Genetics, LLM, Mixed Effect Modelling, Time Series Analysis, HPC

ML Intern, Siemens Healthineers, Princeton *2021*

Contributions: Graph Neural Networks, Contrastive Learning, Interpretability

Skills

GL for Genomics Graph neural networks for structural variant calling, graph-attention models for polygenic risk scoring.

ML for Genomics Deep Bayesian models for finemapping, latent-factor models for multimodal imaging-genetics

DL for Genomics Billion-parameter foundation models for gene expression prediction, interpretable autoencoder for multi-omics, generative models for comparative transcriptomics, large-scale distributed training

Statistical Genetics GWAS, structural variant discovery, PRS analysis, finemapping, imaging-genetics

Deep Learning Foundation Models, Hierarchical Transformers, Diffusion Models, Contrastive Learning on Graphs, Autoencoders, GNN

Model Iterpretibility Motif Discovery, Bayesian Feature Selection, Attention Mechanisms, LASSO, Group-LASSO

Relevant experience

Chan Zuckerberg Initiative, Redwood City *Present*
Research Scientist, AI/ML

Large-Scale Generative Genome Modelling

- Building large-scale generative models for genomic sequences to enable *in silico* perturbation and design.
- Scaling training across distributed GPU clusters to model genome-wide regulatory landscapes.

VariantFormer: Personalized Gene Expression Prediction from DNA

- Led development of a 1.2B-parameter hierarchical transformer predicting gene-level RNA abundance across 62 tissues from personalized DNA sequences integrating cis-regulatory elements across >2 Mb context.
- Trained on the largest curated collection of paired WGS and bulk RNA-seq (21K samples, 2.3K donors from GTEx, MAGE, ADNI, ENCODE) at scale on 376 H100 GPUs.
- Achieved state-of-the-art gene correlation (ρ=0.80 protein-coding, 0.54 non-coding), outperforming Enformer, Borzoi, and TWAS baselines; demonstrated variant effect prediction (ρ=0.60) where prior models showed near-zero correlation.
- Demonstrated zero-shot disease risk prediction for Alzheimer's, recapitulating known APOE allele risk architecture through *in silico* mutation of patient genomes.

Deep generative models for comparative transcriptomics

- Developing diffusion models for cross-species single cell genomics data.
- Transfer learning the effects of disease or drugs in single-cell expressions across species.

Broad Institute, Cambridge *Present*
Computational Scientist

Motif Driven Structural Variant Discovery

- Identifying structural variants from the graph representation of sequence alignments.
- Finding novel motifs for complex genomic rearrangements.

Johns Hopkins University, Baltimore *2017-2023*
Research Assistant, Electrical and Computer Engineering

BEATRICE: Bayesian Fine-mapping from Summary Data using Deep Variational Inference

- Developed a deep Bayes variational approach to parse complex heritability resulting in 2.2 fold increase in power and coverage.
- Utilized machine learning with Bayesian inference to handle multiple causal variants and infinitesimal effects from non-causal variants.

A Biologically Interpretable and Non-linear Approach to Generate Polygenic Risk Scores

- Consolidated genetic risk along biological pathways to generate risk scores predictive of disorder.
- Embedded gene ontology in a graph to infer underlying processes and functions linked to disease risk prediction.

A Biologically Interpretable Graph Convolutional Network to Link Genetic Risk Pathways and Neuroimaging Markers of Disease

- Developed a novel geometric deep learning tool for whole-brain whole-genome analysis of schizophrenia.
- Collaborated with cross-functional teams of biologists, data scientists, and clinicians, which led to a future million-dollar grant, scholarships[2], awards[1], and two publications.

Multimodal Imaging Genetics Models for Biomarker Identification and Schizophrenia Risk Prediction

- Developed novel latent factor models utilizing autoencoder and dictionary learning to identify correlated brain and genetic networks from brain imaging and genetics study of schizophrenia.
- Received *special mention* in the Hopkins magazine and a best paper award[3] at SPIE.

Supervising Activity

Johns Hopkins University, Baltimore *2021- 2023*
Supervisor

- Advising a computer science graduate student on deep learning projects aimed to learn the longitudinal effect of genetic variations on morphological changes in brain regions of 1K Alzheimer's patients.
- Authored a senior author paper at the International Conference of the IEEE Engineering in Medicine and Biology Society.

Honors And Awards

[1]Organization for Human Brain Mapping awarded $700 for noteworthy abstracts.	2023
[2]MINDS fellowship awarded $\$30K$ for spring tuition.	2022
[3]Best Paper Award, SPIE Medical Imaging (Image Processing Conference)	2021
[4]MICCAI travel award of $500.	2020
[5]Dept. of Electrical and Computer Engineering, JHU, PhD fellowship	2017-2018
[6]Mitacs Globalink Research Fellowship Award	2016

Patents

Ghosal, S., Jacob, A. J., Sharma, P., & Gulsun, M. A. (2023). Subpopulation Based Patient Risk Prediction Using Graph Attention Networks. US Patent App. 17/647,613.

Publications

S. Ghosal, *et al.*, *VariantFormer: A Hierarchical Transformer Integrating DNA Sequences with Genetic Variation and Regulatory Landscapes for Personalized Gene Expression Prediction.* bioRxiv 2025.10.31.685862 (*Preprint*).

S. Ghosal, *et al.*, *GUIDE-PRS:A Biologically Interpretable and Non-linear Approach to Generate Polygenic Risk Scores.*(*In Prep*).

S. Ghosal, *et al.*, *BEATRICE: Bayesian Fine-mapping from Summary Data using Deep Variational Inference.*(*Submitted in* ***Oxford Bioinformatics***).

S. Wu, A. Venkataraman, **S. Ghosal**. *GIRUS-net: A Multimodal Deep Learning Model Identifying Imaging and Genetic Biomarkers Linked to Alzheimer's Disease Severity.*

S. Ghosal, *et al.* *A Biologically Interpretable Graph Convolutional Network to Link Genetic Risk Pathways and Neuroimaging Markers of Disease.* **ICLR: International Conference on Learning Representations**, 2022 (**Accepted**).

S. Ghosal, *et al.* *A Generative Discriminative Framework that Integrates Imaging, Genetic, and Diagnosis into Coupled Low Dimensional Space.*NeuroImage: 238:118200, 2021

S. Ghosal, *et al.* *G-MIND: An End-to-End Multimodal Imaging-Genetics Framework for Biomarker Identification and Disease Classification.* Proc. SPIE, Medical Imaging 2021: Image Processing.
Selected for Special Oral Presentation (<15% of Papers), and received best student paper award

S. Ghosal, *et al. Bridging Imaging, Genetics, and Diagnosis in a Coupled Low-dimensional Framework.*MICCAI: Medical Image Computing and Computer Assisted Intervention, 2019. **Selected for Early Acceptance (Top 18% of Submissions)**

S. Ghosal, *et al. A generative-predictive framework to capture altered brain activity in fMRI and its association with genetic risk: application to Schizophrenia.* Proc. **SPIE** 10949, Medical Imaging 2019: Image Processing.

S. Ghosal, Nilanjan Ray. *Deep deformable registration: Enhancing accuracy by fully convolutional neural net.* **Pattern Recognition Letters**.

S. Ghosal, *et al. A novel non-rigid registration algorithm for zebrafish larval images.* 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (**EMBC**), 2017.

Invited Seminars and Talks

Title: Benefits of Deep Learning to Parse Complex Genetic Architectures to Provide Mechanistic Insights

MIT (Host: Manolis Kellis) *2023*

Title: Deep Imaging Genetics to Parse Neuropsychiatric Disorders

Regeneron (Host: Yu Bai) *2023*

Google-Genomics, Google Health (Host: Farhad Hormozdiari) *2022*

Title: Biologically Inspired Regularization Models Integrating Multimodal Data to Parse Neuropsychiatric Disorders.

ECE Seminar Series (Host: Archana Venkataraman) *2022*