

Soham Mukherjee

soham.juetce@gmail.com | (614) 270-8809 | website | linkedin | google-scholar | github

Professional Summary

AI Research Scientist with a PhD from Purdue University specializing in Graph Neural Networks, multimodal ML, and generative AI. Track record of designing and scaling ML pipelines on large, heterogeneous datasets — combining graph embeddings, sequential models, and NLP-based feature extraction into unified predictors. Passionate about solving hard, ambiguous problems where data is messy and labels are scarce.

Work Experience

AI Research Scientist, *Cadence Design Systems – San Jose, CA* Feb 2024 - Present

- Built an end-to-end pipeline to convert raw unstructured data into attributed graphs and benchmarked GNN architectures for a regression task on sparse, high-dimensional targets.
- Designed a multimodal predictor combining graph embeddings, log embeddings, and state-space models to capture both structural and sequential signals from heterogeneous data sources.
- Developed a pre-training strategy using Triplet Loss to learn similarity structure from a proxy label before fine-tuning on sparse downstream targets, improving generalization on limited labeled data.
- Built a scalable data augmentation framework that programmatically generates and validates new graph instances, enabling robust dataset enrichment for ML benchmarking and robustness testing.
- Wrote an NLP-based feature extraction framework to parse unstructured logs into structured features for downstream ML modeling.
- Set up distributed MLOps pipelines to train and serve GNNs on graphs with up to 300K nodes using distributed GPUs, with experiment tracking, dataset versioning, and containerized workflows for fast iteration across architectures.

AI Research Intern, *IBM TJ Watson – Yorktown Heights, NY* May 2022 - Aug 2022

- Graph generation with geometrical and topological constraints.

AI Engineering Intern, *Physna Inc. – Columbus, OH* May 2021 - Aug 2021

- Deployed CNNs to predict 3D computer-aided design (CAD) models from 2D images.
- Automated segmentation and registration of point-cloud data obtained from scanning machine parts enabling efficient and accurate inspection.

Education

PhD Computer Science | **Purdue University** (Awarded Scholarship) Aug 2020 - Dec 2023

Relevant Coursework : Machine Learning, Deep Learning, Gen AI

MS Computer Science | **Ohio State University** Aug 2017 - Jul 2020

BE Electronics & Telecommunication | **Jadavpur University** (Gold Medalist) Aug 2013 - Aug 2017

Skills

Languages: Python, C/C++ , Java

Framework: PyTorch, Pytorch-Geometric, Tensorflow, Keras, HuggingFace

Deep Learning & GenAI: GNN, Neural Operators, Contrastive Learning, Computer Vision, NLP, Attention Networks, Transformers, Diffusion Models, Large Language Model (LLM) finetuning & evaluations, Agentic AI, Multimodal Framework.

Selected Patents & Publications

1. Mukherjee, Soham, Karthikeyan Natesan Ramamurthy, and Payel Das. Generative Modeling with Topological Control. U.S. Patent P202203186US01, 2025 (Filed).
2. Mukherjee, Soham, Shreyas N. Samaga, et al. ‘D-GRIL: End-to-End Topological Learning with 2-Parameter Persistence’. Proceedings of the 42nd International Symposium on Computational Geometry (SoCG 2026), 2026.
3. Zhang, Simon, et al. ‘GEFL: Extended Filtration Learning for Graph Classification’. Learning on Graphs Conference, PMLR, 2022, pp. 16–11.
4. Hajij, Mustafa, et al. ‘Topological Deep Learning: Going Beyond Graph Data’. arXiv Preprint arXiv:2206. 00606, 2023.
5. Xin, Cheng, et al. ‘GRIL: A 2-Parameter Persistence Based Vectorization for Machine Learning’. Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML), vol. 221, 333, 2023, p. 313.

Full publication list is available on Google Scholar.