Shania Mitra

Graduate Research Assistant, CSAIL, Massachusetts Institute of Technology

857-500-9365 shania@csail.mit.edu linkedin.com/shania-mitra/ github.com/Shania99

Education

Massachusetts Insititute of Technology (MIT), Cambridge, MA **Sep. 2023 – May 2025**
Master of Science in Computational Science and Engineering *GPA: 5.0/5*

Indian Insititute of Technology (IIT) Madras, Chennai, India **Jul. 2018 – May 2023**
Master of Technology in Data Science, Bachelor of Technology in Chemical Engineering *GPA: 9.36/10, Dept. Rank: 1*

Publications

- **S. Mitra**, L. Huang, M. Kellis *"ProteinRPN: Towards Accurate Protein Function Prediction with Graph-Based Region Proposals" (Under Review)* **[Link]**
- **S. Mitra**, A. K. Tangirala *"Causal Discovery from Natural Language Text using Context and Dependency Information,"* 2022 61st Annual Conference of the Society of Instrument and Control Engineers (SICE), 2022, Japan, pp. 236-241 **[Link]**
- S. K.*, **S. Mitra***, D. Padmanabhan, S. Chakraborti *"Counterfactuals as Explanations for Monotonic Classifiers"* 4th Workshop on CBR for the Explanation of Intelligent Systems (XCBR) held with the ICCBR 2022 Conference, 2022, France **[Link]** *(*Shared First Author)*
- **S. Mitra**, D. Matthew, D. Padmanabhan, S. Chakraborti *"Group Fairness in Case-Based Reasoning"*, International Conference on Case-Based Reasoning (ICCBR 2023) **[Link]**
- S. Chinta, **S. Mitra**, R. Rengaswamy *"Machine Learning-based QSPR Approaches to Predict Solvation Free Energy of Quinone Molecules for Flow Battery Applications" (Submitted)*

Research Experience

Graduate Research Assistant, Compbio Lab, Guide: Prof. Manolis Kellis **Sep. 2023 – Present**
Project: Graph-Based Protein Function Prediction with Region Proposals and Contrastive Learning *CSAIL, MIT*

- Developed novel graph-transformer based region proposal model and contrastive learning techniques to accurately localize functional residues for protein function prediction
- Conducted extensive experiments using PDB and SWISS-MODEL datasets, demonstrating 7% improvement in Fmax over SOTA models; results submitted for publication

Risk Prediction through Causal Discovery in Chemical Plants using NLP **May 2021 – Jul. 2022**
Bachelor's Thesis, Guide: Prof. Arun Tangirala *IIT Madras*

- Developed a novel BiLSTM and Node2Vec-based architecture to extract causes and effects in chemical plant logs
- Built a framework that uses linguistic features, dependency parse trees to disambiguate sentences and detect causality
- Used extracted causal relations to build a causal graph and identify the primary cause of effects observed in the plants

Caching in Computer Vision Architectures for Robustness **Apr. 2020 – Sep. 2020**
Guide: Prof. Pratyush Kumar *IIT Madras*

- Conceptualized a cache-like framework on existing DL models to accelerate inference time, maintaining original accuracy
- Used triplet loss to retrain a segment of existing networks and obtain similar hidden-layer vectors for similar inputs
- Implemented a shallow neural network to emulate KNN on the modified hidden-layer outputs
- Reduced the inference time of DenseNet-121 by 23% maintaining top-1 accuracy at 74.32% on CIFAR100 (original accuracy: 76.58%) while increasing robustness to adversarial attacks

Counterfactuals as Explanations for Monotonic Classifiers **Dec. 2021 – Jul. 2022**
Guides: Prof. Deepak P. (Queen's University, Belfast), Prof. Sutanu C.(IIT Madras) *IIT Madras*

- Developed a novel counterfactual generation algorithm, incorporating domain knowledge through monotone constraints in classifiers
- Proposed the use of skyline corners on the decision surface as counterfactuals to improve proximity to query
- Generated counterfactuals 10% closer to the queries as compared to the baseline, with 70% more coverage

Prediction of Blood-brain Barrier Permeability of Drugs using Subset Selection **May 2022 – Jul. 2022**
MITACS Globalink Scholarship, Guide: Prof. Paul Ayers *McMaster University*

- Predicted blood-brain barrier permeability on the PhysProp data set using Stochastic Gradient and Least Angle Regression to show the effectiveness of DiverseSelector
- Profiled the performance of different subset selectors for various selection proportions and data sizes

Professional Experience

Beaver Works Summer Institute, MIT Lincoln Laboratory **Jul. 2024 – Aug. 2024**
Lead Instructor, Medlytics - Machine Learning for Healthcare *Boston, MA*

- Led instruction and managed course operations for Medlytics, overseeing a team of 6 teaching assistants and delivering an ML curriculum to 40 top high school students selected from 1600+ applicants
- Mentored 7 advanced DL projects from ideation to completion, focusing on scalable ML/DL solutions for healthcare challenges, including diagnostic tools and predictive analytics

Goldman Sachs, Inc. - Quantitative Strategy Intern, Model Risk Management **Jun. 2021 – Jul. 2021**
Project: Interpretable Statistical Modeling of Revolver Bias in Loan Portfolios *Bengaluru, India*

- Performed extensive parametric and non-parametric hypothesis testing on loan data from 3 divisions to detect the presence of revolver bias across stress and non-stress periods
- Reformulated revolver bias for improved statistical properties, driving more actionable insights for business deliverables
- Built regression model for new formulation using macroeconomic factors; achieved an adjusted R^2 score of 0.938

Bewgle, Inc. - Natural Language Processing Intern **May 2020 – Jul. 2020**
Project: Product Categorization and Key Phrase Extraction Using NLP and BERT Models *Bengaluru, India*

- Developed a machine learning model for product categorization and key phrase extraction from Amazon Review data, achieving a 0.99 F1-score, demonstrating strong statistical performance and interpretability
- Fine-tuned BERT models for each product category and deployed predictions on customer reviews, translating business requirements into robust NLP solutions

DataHive Labs - Machine Learning Intern **May 2019 – Jul. 2019**
Project: Anomaly Detection in Time Series Sensor Data *Bangalore, India*

- Developed ML models to detect and group breakdowns in mechanical equipment by analyzing large-scale time series sensor data, improving risk detection and equipment reliability
- Predicted sequence anomalies using an epsilon nearest neighbor graph, optimizing for early breakdown detection in real-time sensor data streams
- Applied texture-analysis methods like Gray-Level Co-Occurrence Matrix (GLCM) on spectrogram images to enhance the accuracy of breakdown predictions from sensor data

Technical Skills

Languages: Python, C/C++, HTML, OpenMP, JavaScript, SQL, CypherQL, GraphQL, UNIX/Bash Shell Scripting

Databases: Oracle, MariaDB, MySQL, Neo4j

Libraries/Frameworks & Tools: PyTorch, Tensorflow, Pytorch-Geometric, Higher Performance Computing, OpenCV

Softwares: Git, MATLAB, SIMULINK, Marvin Sketch, ASPEN, VESTA, LaTeX

Leadership / Extracurricular

Teaching Assistant, Machine Learning for Computational Biology, EECS, MIT: Co-instructed, designed exams, and assisted in teaching and evaluating 40+ students applying ML to cutting-edge biological problems at MIT.

Coordinator, Career Development Cell, IIT Madras: Organized 100+ higher studies and placement events, coding events and skill development webinars for 500+ students

Student Mentor, MITr, IIT Madras: Mentored a batch of 5 freshmen through 2 semesters, in all academic and non-academic endeavors

Scholastic Achievements

- Recognized with the Late B Ravichandran Memorial Prize for Best Academic Record in Chemical Engineering, Dual Degree Batch 2018-2023 at IIT Madras
- Awarded the Lakshmi Ravi Prize for the Best Dual Degree Project in the Interdisciplinary Dual Degree program in Data Science at IIT Madras (2023)
- Recipient of the prestigious Kishore Vaigyanik Protsahan Yojana Fellowship (KVPY) by Govt. of India with All India Rank-603 out of 150,000 students *(99.6 percentile)*