Mahbuba Tasmin

PhD Candidate — Machine Learning for Biological Foundation Models

University of Massachusetts Amherst

413-479-9565 | mtasmin@umass.edu | linkedin.com/in/mahbuba-tasmin | github.com/Tasmin153 | Google Scholar

Research Profile

PhD candidate studying machine learning methods for biological sequence modeling and phenotype prediction. My research focuses on protein foundation model representations, mutation-sensitive prediction tasks, and evaluation frameworks for reliable AI in genomics. I build GPU-accelerated training pipelines, design controlled ablation studies, and analyze representation trade-offs (token-level vs pooled vs compressed embeddings) to understand how biological foundation models generalize across sparse and evolving genomic datasets.

Education

University of Massachusetts Amherst Amherst, MA
Ph.D. Candidate in Computer Science (Advisor: Prof. Anna G. Green) *Expected May 2027*

- Research focus: Foundation models for biological sequences, phenotype modeling, robustness and interpretability in mutation-sensitive prediction tasks.
- Award: Sudha and Rajesh Jha Scholarship (2023).

University of Massachusetts Amherst Amherst, MA
M.S. in Computer Science - GPA: 3.9 / 4.0 *Sep. 2022 – May 2025*

- Thesis aligned with PhD research on predictive and interpretable models of antibiotic resistance.

North South University Dhaka, Bangladesh
B.S. in Computer Science and Engineering, Summa Cum Laude *Jan. 2016 – Dec. 2019*

- Concentration: Artificial Intelligence and Algorithms; GPA: 3.89 / 4.0

Research Experience and Projects

Graduate Research Assistant Sep. 2022 – Present
SAGE Lab, University of Massachusetts Amherst *Amherst, MA*

- **Foundation Model Representations for Mutation-Sensitive Prediction**
  - Studied representation properties of protein foundation models (ESM) for mutation-sensitive phenotype prediction, building GPU-accelerated PyTorch pipelines for resistance modeling in *Mycobacterium tuberculosis*.
  - Analyzed representation trade-offs (token-level, PCA-compressed, sequence-mean) to balance expressivity and computational efficiency.
  - Investigated how different representation granularities influence biological signal capture in mutation-level prediction tasks.
  - Conducted controlled ablations on evolutionary augmentation to diagnose performance gains and instability under distribution shift.
  - Evaluated models via stratified cross-validation, ROC-AUC, sensitivity/specificity, and mutation-category analysis.
- **BIG-TB: Benchmarking Clinical Genomic Models**
  - Co-developed a reproducible benchmarking framework spanning linear models, CNNs, Transformers, and protein language models for genomic phenotype prediction.
  - Standardized heterogeneous genomic, protein, structural, and curated resistance datasets into a unified evaluation ecosystem.
  - Introduced interpretability metrics (SHAP, causal variant recall@k) to assess biological validity beyond predictive accuracy.
  - Characterized model failure modes in sparse-variant and low-data regimes.
- **Structure-Aware Modeling and Resistance Forecasting**
  - Developed structure-informed fused ridge objectives incorporating spatial proximity constraints into convex optimization.
  - Integrated sequence embeddings, structural perturbation metrics, and evolutionary features for resistance forecasting.
  - Demonstrated improved stability and interpretability in sparse, low-data regimes.

Applied Machine Learning Experience

AI Engineer Mar. 2022 – Jul. 2022
NITEX Solutions Ltd. *Dhaka, Bangladesh*

- Implemented Detectron2-based instance segmentation and OCR pipelines for automated product identification.
- Built NLP- and CV-driven fashion trend analysis tools for business workflow automation.

Software Engineer (AI & ML) Jul. 2020 – Feb. 2022
M2SYS Technology *Dhaka, Bangladesh*

- Developed image spoofing detection systems and NLP-based contextual recommendation engines.
- Deployed production ML systems and automated workflows using Camunda across distributed environments.

Technical Skills

Programming: Python (advanced), C/C++, R, Bash, LaTeX
Deep Learning: PyTorch, CNNs, Transformers, protein language models (ESM), sequence representation learning
Foundation Models: Token-level embeddings, representation compression (PCA), fine-tuning, robustness evaluation, ablation design
Experimentation: Cross-validation, hyperparameter optimization, performance diagnostics, interpretability (SHAP)
Bioinformatics: UniProt, InterPro, Rosetta, AAIndex, protein structure mapping
Systems & Infrastructure: GPU-based training, SLURM clusters, memory-mapped datasets, Docker, Linux

Selected Publications

Tasmin, M., Mohanty, S., Kulkarni, S., Farhat, M. R., **Green, A. G.**†
BIG-TB: A benchmark for evaluating prediction and interpretability of sequence-based machine learning using Mycobacterium tuberculosis genomes. PLOS Computational Biology (under review), 2026.

Green, A. G., **Tasmin, M.**, Vargas, R., Farhat, M. R.
The structural context of mutations in proteins predicts their effect on antibiotic resistance. eLife, 2025.

Tasmin, M., Green, A.
Beyond Sequence-only Models: Leveraging Structural Constraints for Antibiotic Resistance Prediction in Sparse Genomic Datasets. ICLR 2025 MLGenX Workshop.

Yang, Z., Yao, Z., **Tasmin, M.** et al.
Unveiling GPT-4V's hidden challenges behind high accuracy on USMLE questions. Journal of Medical Internet Research, 2025.

Software and Data Resources

BIG-TB: Scalable Benchmark for Biological Sequence Models 2024–Present
Open-source framework *github.com/SAGE-Lab-UMass/Big-TB-benchmark*

- Reproducible end-to-end pipeline for large-scale training and evaluation of biological sequence models.
- Supports linear models, deep neural networks, and protein foundation model embeddings under unified evaluation protocols.
- Implements structured cross-validation, robustness diagnostics, and interpretability benchmarking.

Structure-Aware Variant Analysis Toolkit 2023–Present
Research codebase *github.com/aggreen/MTB_Mut_Clust*

- Pipelines for mapping mutations to protein structures and quantifying spatial clustering effects.
- Supports structure-informed modeling and mechanistic interpretation of mutation effects.

Selected Talks and Presentations

BIG-TB: A Benchmark Dataset for Genomic Resistance Prediction and Interpretability.. MLCB Workshop, 2025 (Spotlight).

Protein Structure-Informed Regularized Models for Antibiotic Resistance.. Harvard PQG Conference, 2024.

Teaching and Professional Services

Head Teaching Assistant Spring 2023-Current
COMPSCI 520: Software Engineering *UMass Amherst*

Graduate Representative 2025-2026
Faculty Senate, College of Information & Computer Sciences, UMass Amherst