Md Saidul Hoque Anik

Research Interest: ML Systems Optimization, Sparse Computation

anik@tamu.edu | in/in/onixhoque/ | +1 470-929-2551 | College Station, TX, USA

Education

Texas A&M University TX, USA
Ph.D. Candidate in Computer Science & Engineering Spring 2025–Fall 2027 (Expected)
Specialization: Scalable and Differentiable Sparse Kernels for Graph Learning

Indiana University Bloomington IN, USA
Master's in Intelligent Systems Engineering (CGPA: 4.00/4.00) Fall 2022–Summer 2024
Specialization: High Performance Computing and Graph Neural Networks

Bangladesh University of Engineering and Technology (BUET) Dhaka, Bangladesh
Master's in Computer Science & Engineering (Part-time) April 2017–Aug 2022
Bachelor's in Computer Science & Engineering May 2012–Feb 2017

Selected Publications & Posters

Hoque Anik, M. S. & Azad, A. (2025, May). *SparseTransX: Efficient Training of Translation-Based Knowledge Graph Embeddings Using Sparse Matrix Operations*. In Eighth Conference on Machine Learning and Systems (MLSys) 2025. [Paper] [Talk] [GitHub]

Anik, Md Saidul Hoque and Azad, Ariful, *A Sparse Approach for Translation-based Training of Knowledge Graph Embeddings*, Poster presented at SC24 – ACM/IEEE Supercomputing Conference, Atlanta, GA. *Finalist (top-6) for Best Research Poster Track*. [Summary]

Hoque Anik, M. S., Badhe, P., Gampa, R., & Azad, A. (2024, May). *iSpLib: A Library for Accelerating Graph Neural Networks using Auto-tuned Sparse Operations*. In Companion Proceedings of the ACM on Web Conference 2024 (pp. 778-781). [Paper] [GitHub]

Agrawal, Abhigya, **Md Saidul Hoque Anik**, and Ariful Azad. *Predicting Interactions in the Weapons of Mass Destruction Knowledge Graphs*. International Conference on Complex Networks and Their Applications. Cham: Springer Nature Switzerland, 2024. [Paper]

Research Experience

Amazon AWS with Sagemaker Inference Team Santa Clara, CA
Applied Scientist Intern May 2025–Aug 2025

Developed a robust, end-to-end **differentiable GPU kernel autotuner** for **vLLM** that works well with as low as **1% ground truth** of the total search space. The solution also performs **transfer learning** and can leverage cheaper kernel tuning data to reduce new kernel tuning time **from days to hours**.

- Across six datasets, our auto-tuner obtained an improvement in cross-validation accuracy of 0.85× − 1.60× compared to **16 other ML and tabular transformer models** (commonly used for performance modeling) on **1% ground truth**. Accuracy continued to improve and outperformed all baselines as the training data increased.
- In transfer learning, up to 12.7% accuracy improvement was observed when an expensive **CUDA** kernel was tuned using only 100 config-perf pairs and leveraged with similar and cheaper **Triton** kernel's tuning data.

HipGraph Lab with Dr. Ariful Azad College Station, TX and Bloomington, IN
Graduate Research Assistant Aug 2022–Present

Currently working on: (1) a **distributed SpMM** PyTorch operator for massive matrices and (2) an **LLM-based code generator** for SpMM kernels. Developed several high-performance PyTorch and LibTorch-based **linear algebra** kernels for **Graph Neural Networks** and **Knowledge Graph** Embeddings training. Added parallel, distributed, and streaming functionalities in the GNN and KG libraries. Mentored three graduate students.

- *[WIP] Data-Driven LLM-Based Code Generator for SpMM Kernels*: Developing a framework for **data-driven** and efficient SpMM kernel generation using LLM. Developed a dataset curation pipeline utilizing NVIDIA Nsight Profiler feedback, enabling automated **model distillation** with **Chain-of-Thought** (CoT) reasoning. Incorporated contrastive learning using **direct preference optimization** (DPO). Using **Quantization** and **LoRa** to perform efficient fine-tuning.
- *[WIP] Scalable Multi-GPU SpMM Operator for Graph Learning in PyTorch*: Developing a high-performance, distributed Sparse Matrix–Dense Matrix Multiplication (SpMM) implementation in PyTorch, tailored for **efficient backpropagation** and large-scale graph learning. The approach leverages **partitioning**, **CPU offloading**, and **communication overlap** to improve memory usage and training speed.

- *[MLSys25, SC24] SpTransX*: A sparse kernel-based high-performance **Knowledge Graph** embedding training library using **PyTorch**. The current codebase supports **four** translational models and exhibits up to 5.3× speedup on the CPU and a 4.2× speedup on the GPU. The implementation scales across multiple nodes and GPUs using **PyTorch DDP** and **FSDP**. Achieved up to an 11.1× improvement in CUDA memory efficiency, along with a 3.9× speedup in training performance across 64 GPUs at the NERSC Supercomputer.
- *[WebConf24] iSpLib*: A **PyTorch**-based auto-tuned sparse kernel library for **GNN** that can speed up PyTorch 2.1 implementation of various GNNs up to 93×. Achieved up to 54× speedup for **GCN**, 32× for **GraphSAGE-SUM**, 23× for **GraphSAGE-MEAN**, and 51× for **GIN** compared to equivalent **PyTorch Geometric** 2.4 implementations. The project is an extension of a **C**-based Fused BLAS library with a code generator supporting SIMD instruction tuning across Intel, AMD, and ARM processors. Enabled easy integration into compatible PyG codes with **only two lines** of Python code.
- *Interfacing:* [Complex Network 2024] Designed a **Neo4j**-based training pipeline for Knowledge Graph training using **TorchKGE**. Also developed a **PyBind11** interface for FastGraph, an **OpenMP**-based **C++** parallel sparse matrix library similar to **NetworkX**.

BUET Next-gen Computing Lab with Dr. A. B. M. Alim Al Islam — Dhaka, Bangladesh
Graduate Student — April 2017–Aug 2022
Worked on several **NLP** and **HCI**-related ML projects. Published two conference papers, two journals, and a poster (IETE Technical Review 2022, Neural Computing and Applications 2021, NSysS 2021, 2018, & 2017).

Graduate Course Projects

CSCI 565P - Data Mining with Dr. Dongruo Zhou — IUB, Fall 2023
Developed a **spatio-temporal link prediction** pipeline using **GC-LSTM** for dynamic graphs in **PyTorch**. Achieved over 75% Hits@100 accuracy for the protein-protein interaction **graph sequences** of DDPIN dataset. This pipeline can also be used in recommendation systems or knowledge graph completion.

ENG 503 - Intro to Intelligent Systems with Dr. Ariful Azad — IUB, Fall 2023
Performed CPU **time-profiling** on several translational, bilinear, and deep **KG embedding models** using **py-spy**. Found the gradient computation for embedding to be the hotspot for translational models.

CSE 6305 - Programming Languages and Systems with Dr. Rifat Shahriyar — BUET, Fall 2017
Performed **time-profiling** on **multi-threaded** programs written in three implementations of Python: **IronPython (.NET)**, **Jython (Java)**, and **CPython**. Found Jython to outperform the other two.

Teaching Experience

United International University (UIU), Dhaka, Bangladesh — Jan 2020–Present
(On-leave) Tenure Faculty at Dept. of CSE
UG Courses Conducted: Computer Architecture, Software Engineering, Web Programming, and 7 others.

Military Institute of Science & Technology (MIST), Dhaka, Bangladesh — Feb 2017–Jan 2020
Adjunct → Tenure Faculty at Dept. of CSE
UG Courses Conducted: Structured and Object-Oriented Programming Language (C, C++, Java), Data Structures and Algorithms, and 4 others.

Software & Frameworks

Programming Language Python, C, C++, Java, JavaScript. Familiar: VB6, Assembly (MIPS)
Machine Learning PyTorch, LibTorch, NumPy & Scikit-learn, PyTorch Geometric, TorchKGE, NetworkX
HPC DDP and FSDP, SLRUM, PyBind11, Neo4j, CUDA, NVIDIA NCU, PySpy. Familiar: OpenMP
Web Development Flask, Django, PHP, VueJS, PWA, JavaScript, Whoosh, MySQL, SQLite

- Aayatun (aayatun.com), a **Flask**-based Quran encyclopedia with 150K+ monthly visitors [GitHub]

Android App Development Java Android SDK, Volley Framework, ORACLE, MySQL, Java Socket

- Conducted a **5-day workshop** on Android app development at BUET's System Design community

Cross-platform Desktop App Development ElectronJS, VueJS, Java Swing
Automation PyQT5, Selenium

Awards & Recognitions

Finalist of Best Research Poster Award in **Supercomputing'24** (Top 6)	Nov 2024
Google **Foobar** Challenge	2023
Luddy Summer Fellowship, $8,400	Aug 2022
BRAC Hackathon 2015 Android App Development Champion, $2,500	Dec 2015
BUET Undergraduate Scholarship	2013

Other Experience

TAMU CSCE-654: Supercomputing, lecture on PyTorch DDP [GitHub], **Guest Lecturer**	Fall 2025
Asian CHI Symposium 2020, **Paper Reviewer**	2020
UIU Innobotics, **Assistant Technical Coordinator**	Feb 2020
MIST **Postgraduate Coordinator**	Jan 2019–June 2019
BUET System Analysis, Design and Development community, **Coordinator**	2015