Nima Shoghi

| nimash@gatech.edu | 404-862-0633 | https://nima.sh

Education

(Complete List on Page 2)

Georgia Institute of Technology — ***PhD Machine Learning (ML)*** 2024 - 2028 (expected)
Advisors: Dr. Pan Li and Dr. Victor Fung
Research Interest: Developing ML techniques to solve complex problems in the scientific and engineering domains.

Georgia Institute of Technology — ***MS Computer Science (ML Focus), Summa cum laude*** 2020 - 2021

Georgia Institute of Technology — ***BS Computer Science (ML Focus), Magna cum laude*** 2015 - 2019

Experience

(* indicates first-author; Complete List on Page 2)

Meta Fundamental AI Research (FAIR) — ***AI Resident, FAIR Chemistry Team*** Aug 2021 - Aug 2023

- Developed large foundation models for atomic property prediction, pre-trained on data from diverse chemical domains. Fine-tuned the model to achieve state-of-the-art results across 35 out of 41 downstream tasks. (**ICLR 2024***)
- Contributed to the development of a transfer learning approach using Graph Neural Networks to generalize models across domains in molecular and catalyst discovery, reducing the need for large, domain-specific datasets. (**J Chem Phys 2022**)
- Benchmarked state-of-the-art machine learning interatomic potentials models on the Open Catalyst 2022 dataset, one of the largest datasets for automatic catalyst discovery. (**ACS Catalysis 2023**)

Graph Computation and Machine Learning Lab @ GT — ***Graduate Research Assistant*** Aug 2024 - Present

- Developed parameter-efficient fine-tuning strategies for machine learning interatomic potentials models trained on the Materials Project dataset, achieving near-SOTA performance on the MatBench Discovery benchmark. (**In Submission***)

ProcessMiner — ***Machine Learning Intern*** June 2024 - Aug 2024

- Developed transformer models pre-trained on approximately 500,000 time-series data points from manufacturing processes to predict process outcomes and detect anomalies, achieving accuracy improvements on real-world manufacturing datasets.

HPArch Lab @ GT — ***Research Assistant & Research Staff*** May 2019 - May 2021 & Dec 2023 - May 2024

- Developed an efficient sampling method for Denoising Diffusion Probabilistic Models (DDPMs) which leverages the structure of the latent space to guide sampling, reducing the number of samples needed for high-quality image generation. (**In Submission***)
- Developed novel quantization techniques for deep learning models, achieving >6x memory savings in training and inference. (**MemSys 2020*, IEEE CAL 2021***)

Publications

(Complete List on Page 3)

From Molecules to Materials: Pre-training Large Generalizable Models for Atomic Property Prediction
N Shoghi, *A Kolluru, . . . , CL Zitnick, B Wood* **International Conference on Learning Representations, 2024**

Transfer learning using attentions across atomic systems with graph neural networks (TAAG)
A Kolluru, ***N Shoghi****, M Shuaibi, S Goyal, A Das, CL Zitnick, Z Ulissi* **The Journal of Chemical Physics, 2022**

The Open Catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts
R Tran, J Lan, M Shuaibi, BM Wood, . . . , ***N Shoghi****, . . . , EH Sargent, Z Ulissi, CL Zitnick* **ACS Catalysis, 2023**

SmaQ: Smart Quantization for DNN Training by Exploiting Value Clustering
N Shoghi*, A Bersatti, M Qureshi, H Kim* **IEEE Computer Architecture Letters, 2021**

Neural network weight compression with NNW-BDI
N Shoghi*, A Bersatti, H Kim* **The International Symposium on Memory Systems (MemSys), 2020**

Talks

(Complete List on Page 4)

Unlocking the Potential of Pre-training for Accelerated Discovery in Chemistry
Multiple venues including: AI for Science Institute Beijing (*Sep 2024*, Virtual), Machine Learning for Materials and Molecular Discoveries Symposium (*Aug 2024*, Gothenburg, Sweden), King Abdullah University of Science and Technology (*Jul 2024*, Virtual), SES AI (*Jun 2024*, Virtual), Molecular ML Group (*Apr 2024*, Virtual), ACS Fall (*Aug 2023*, Virtual).

Skills

- Extensive experience in training and deploying large-scale deep learning models for scientific and engineering applications.
- GPU programming (CUDA, Triton) and developing custom kernels for efficient model training and inference.
- Experience with generative modeling techniques, including autoregressive language models, VAEs, GANs, and DDPMs.
- Strong programming skills (Python, C++) and software engineering practices (Git, Docker, CI/CD, testing).
- Proficient in PyTorch and JAX for large-scale model development and distributed training on HPC systems.
- Strong background in graph neural networks, geometric learning, and equivariant neural networks.
- Skilled in optimizing ML models for computational efficiency, enabling larger models, faster training, and faster inference

Education

Georgia Institute of Technology — ***PhD Machine Learning (ML)*** 2024 - 2028 (expected)
Advisors: Dr. Pan Li and Dr. Victor Fung
- NSF Graduate Research Fellowship Honorable Mention (2024)

Georgia Institute of Technology — ***MS Computer Science (ML Focus), Summa cum laude*** 2020 - 2021
Advisor: Dr. Hyesoon Kim
- "Thank a Teacher" Award, Georgia Tech Center for Teaching and Learning (2020, 2021)

Georgia Institute of Technology — ***BS Computer Science (ML Focus), Magna cum laude*** 2015 - 2019
- ACM SIGBED Student Research Competition Bronze Medal (2019)
- Zell Miller Scholarship Recipient (2015 - 2019)

Druid Hills High School — ***International Baccalaureate Diploma, 4.0 GPA*** 2011 - 2015
- IB Diploma with Higher Level Mathematics & Physics
- Yale Book Award (2014)
- AP Scholar Award (2014)
- UGA Merit Award (2014)
- QuestBridge Scholarship Finalist (2014)

Experience

(Reverse Chronological Order; * indicates first-author publication)

Graph Computation and Machine Learning Lab @ GT — ***Graduate Research Assistant*** Aug 2024 - Present
- Working with Dr. Pan Li and Dr. Victor Fung on robust fine-tuning strategies for large-scale pre-trained GNN models.
- Developed parameter-efficient fine-tuning strategies for machine learning interatomic potentials models trained on the Materials Project dataset, achieving near-SOTA performance on the MatBench Discovery benchmark. (**In Submission***)

ProcessMiner — ***Machine Learning Intern*** June 2024 - Aug 2024
- Worked with Dr. Kamran Paynabar to develop novel pre-trained transformer models for manufacturing process data.
- Developed transformer models pre-trained on approximately 500,000 time-series data points from manufacturing processes to predict process outcomes and detect anomalies.
- Fine-tuned models to achieve accuracy improvements (relative to previous production models) on real-world manufacturing datasets.

High Performance Computer Architecture Lab @ GT — ***Temporary Research Staff*** Dec 2023 - May 2024
- Worked with Dr. Hyesoon Kim and Dr. Stefano Petrangeli on efficient inference strategies for pre-trained image diffusion models, with a focus on generating diverse, high-quality images.
- Developed an efficient sampling method for Denoising Diffusion Probabilistic Models (DDPMs) which leverages the structure of the latent space to guide sampling, reducing the number of samples needed for high-quality image generation. (**In Submission***)

Meta Fundamental AI Research (FAIR) — ***AI Resident, FAIR Chemistry Team*** Aug 2021 - Aug 2023
- Worked with Dr. Larry Zitnick, Dr. Abhishek Das, and Dr. Brandon Wood on the Open Catalyst Project, focusing on atomic property prediction and catalyst discovery using large-scale pre-trained models.
- Developed large foundation models for atomic property prediction, pre-trained on data from diverse chemical domains. Fine-tuned the model to achieve state-of-the-art results across 35 out of 41 downstream tasks. (**ICLR 2024***)
- Benchmarked state-of-the-art machine learning interatomic potentials models on the Open Catalyst 2022 dataset, one of the largest datasets for automatic catalyst discovery. (**ACS Catalysis 2023**)
- Co-authored a paper discussing the challenges and potential of developing generalizable machine learning models for catalyst discovery, highlighting the importance of large-scale datasets like the Open Catalyst 2020 dataset (OC20). (**ACS Catalysis 2022**)
- Contributed to the development of a transfer learning approach using Graph Neural Networks to generalize models across domains in molecular and catalyst discovery, reducing the need for large, domain-specific datasets. (**J Chem Phys 2022**)

High Performance Computer Architecture Lab @ GT — ***Research Assistant*** May 2019 - May 2021
- Developed software-level and hardware-level techniques for accelerating deep learning training and inference under the guidance of advisors Dr. Hyesoon Kim and Dr. Moinuddin Qureshi.
- Introduced SmaQ, a quantization scheme that leverages the normal distribution of neural network data structures to efficiently quantize them, addressing the memory bottleneck in single-machine training of deep networks. (**IEEE CAL 2021***)
- Developed NNW-BDI, a neural network weight compression scheme that reduces memory usage by up to 85% without sacrificing inference accuracy on an MNIST classification task. (**MemSys 2020***)
- Demonstrated the feasibility of running ORB-SLAM2 in real-time on the Raspberry Pi 3B+ for embedded robots through optimizations that achieved a 5x speedup with minor impact on accuracy. (**SRC ESWEEK 2019, 3rd Place***)
- Co-authored a paper on a context-aware task handling technique for resource-constrained mobile robots, enabling concurrent execution of critical tasks with improved real-time performance. (**IEEE Edge 2023**)
- Contributed to a study that formalized the subsystems of autonomous drones and quantified the complex tradeoffs in their design space to enable optimized solutions for diverse applications. (**ASPLOS 2021**)
- Collaborated on the development of Pisces, a power-aware SLAM implementation that consumes 2.5× less power and executes 7.4× faster than the state of the art by customizing efficient sparse algebra on FPGAs. (**DAC 2020**)

- Participated in an in-depth analysis of the hardware and software components of autonomous drones, characterizing the performance of the ArduCopter flight stack and providing insights to optimize flight controllers and increase drone range. (**ISPASS 2020**)

Georgia Institute of Technology — ***Graduate Teaching Assistant*** Aug 2020 - May 2021
- Led weekly recitations, graded assignments, and held office hours to help students understand course material for CS 4510: Automata and Complexity, a senior-level undergraduate course on the theory of computation. Taught the course in Fall 2020 with Dr. Merrick Furst and in Spring 2021 with Dr. Zvi Galil.
- Received two "Thank a Teacher" awards from the Georgia Tech Center for Teaching and Learning in recognition of outstanding contributions and positive impact as a teaching assistant. (2020, 2021)

Cyber Forensics Innovation Lab at Georgia Tech — ***Research Assistant*** Jan 2020 - Aug 2020
- Developed Graph Neural Network (GNN) based machine learning models to analyze social media data for detecting incoming cyber attacks, under the guidance of advisor Dr. Maria Konte.
- Utilized GNNs to effectively capture the complex relationships and patterns within social media networks, enabling early detection and prevention of potential cyber threats.

Ciena Corporation — ***Software Engineering Intern*** May 2017 - Aug 2018
- Developed software to interface with network devices and maintained CI/CD pipelines for build processes.
- Collaborated with cross-functional teams to ensure smooth integration of software components and timely delivery of projects.
- Gained valuable experience in software development best practices, version control, and agile methodologies.

Publications (Ordered by Date)

From Molecules to Materials: Pre-training Large Generalizable Models for Atomic Property Prediction
N Shoghi, *A Kolluru, JR Kitchin, ZW Ulissi, CL Zitnick, B Wood*, **International Conference on Learning Representations (ICLR), 2024**
- Introduces Joint Multi-domain Pre-training (JMP), a supervised pre-training strategy that leverages diverse data to advance atomic property prediction across chemical domains, achieving state-of-the-art performance on 34 out of 40 downstream tasks.

Distribution Learning for Molecular Regression
N Shoghi, *P Shoghi, A Sriram, A Das*, **arXiv preprint arXiv:2407.20475, 2024**
- Introduces Distributional Mixture of Experts (DMoE), a robust method for molecular property regression that outperforms baselines on multiple datasets and architectures.

The Open Catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts
R Tran, J Lan, M Shuaibi, BM Wood, S Goyal, A Das, J Heras-Domingo, A Kolluru, A Rizvi, ***N Shoghi****, A Sriram, F Therrien, J Abed, O Voznyy, EH Sargent, Z Ulissi, CL Zitnick*, **ACS Catalysis 13 (5), 2023**
- Introduces the Open Catalyst 2022 (OC22) dataset, consisting of 62,331 DFT relaxations, to accelerate machine learning for oxide electrocatalysts and establish benchmarks for the field.

Transfer learning using attentions across atomic systems with graph neural networks (TAAG)
A Kolluru, ***N Shoghi****, M Shuaibi, S Goyal, A Das, CL Zitnick, Z Ulissi*, **The Journal of Chemical Physics 156 (18), 2022**
- Introduces a transfer learning approach using Graph Neural Networks to generalize models across domains in molecular and catalyst discovery, reducing the need for large, domain-specific datasets.

Context-Aware Task Handling in Resource-Constrained Robots with Virtualization
R Hadidi, ***N Shoghi****, B Asgari, H Kim*, **2023 IEEE International Conference on Edge Computing and Communications (EDGE), 2023**
- Introduces a context-aware task handling technique for resource-constrained mobile robots, enabling concurrent execution of critical tasks with improved real-time performance.

Open challenges in developing generalizable large-scale machine-learning models for catalyst discovery
A Kolluru, M Shuaibi, A Palizhati, ***N Shoghi****, A Das, B Wood, CL Zitnick, JR Kitchin, ZW Ulissi*, **ACS Catalysis 12 (14), 2022**
- Discusses the challenges and potential of developing generalizable machine learning models for catalyst discovery, highlighting the importance of large-scale datasets like the Open Catalyst 2020 dataset (OC20).

SmaQ: Smart Quantization for DNN Training by Exploiting Value Clustering
N Shoghi*, A Bersatti, M Qureshi, H Kim*, **IEEE Computer Architecture Letters 20 (2), 2021**
- Introduces SmaQ, a quantization scheme that leverages the normal distribution of neural network data structures to efficiently quantize them, addressing the memory bottleneck in single-machine training of deep networks.

Quantifying the design-space tradeoffs in autonomous drones
R Hadidi, B Asgari, S Jijina, A Amyette, ***N Shoghi****, H Kim*, **Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2021**

- Formalizes the subsystems of autonomous drones and quantifies the complex tradeoffs in their design space to enable optimized solutions for diverse applications.

Neural network weight compression with NNW-BDI
***N Shoghi**, A Bersatti, H Kim*, **Proceedings of the International Symposium on Memory Systems (MemSys), 2020**
- Introduces NNW-BDI, a neural network weight compression scheme that reduces memory usage by up to 85% without sacrificing inference accuracy on an MNIST classification task.

Pisces: power-aware implementation of SLAM by customizing efficient sparse algebra
*B Asgari, R Hadidi, **N Shoghi**, H Kim*, **2020 57th ACM/IEEE Design Automation Conference (DAC), 2020**
- Introduces Pisces, a power-aware SLAM implementation that consumes 2.5x less power and executes 7.4x faster than the state of the art by customizing efficient sparse algebra on FPGAs.

Understanding the software and hardware stacks of a general-purpose cognitive drone
*S Jijina, A Amyette, **N Shoghi**, R Hadidi, H Kim*, **2020 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), 2020**
- Conducts an in-depth analysis of the hardware and software components of autonomous drones, characterizing the performance of the ArduCopter flight stack and providing insights to optimize flight controllers and increase drone range.

Secure Location-Aware Authentication and Communication for Intelligent Transportation Systems
***N Shoghi**, R Hadidi, L Jaewon, J Chen, A Siqueria, R Rajan, S Dhawan, P Shoghi, H Kim*, **arXiv preprint, 2020**
- Introduces a scalable, infrastructure-independent, location-aware authentication protocol for intelligent transportation systems, providing trustworthy communication and efficient sender localization using visual authentication beacons.

SLAM performance on embedded robots
***N Shoghi**, R Hadidi, H Kim*, **Student Research Competition at Embedded System Week (SRC ESWEEK), 2019**
- Demonstrates the feasibility of running ORB-SLAM2 in real-time on the Raspberry Pi 3B+ for embedded robots through optimizations that achieved a 5x speedup with minor impact on accuracy.

Talks (Ordered by Date)

Unlocking the Potential of Pre-training for Accelerated Discovery in Chemistry
Multiple venues: AI for Science Institute Beijing (*Sep 2024*, Virtual), Machine Learning for Materials and Molecular Discoveries Symposium (*Aug 2024*, Gothenburg, Sweden), King Abdullah University of Science and Technology (*Jul 2024*, Virtual), SES AI (*Jun 2024*, Virtual), Molecular ML Group (*Apr 2024*, Virtual), ACS Fall (*Aug 2023*, Virtual).
- Presented on unlocking the potential of large-scale pre-training methods to accelerate discovery in chemistry, highlighting key challenges and opportunities in this rapidly evolving field.

SmaQ: Smart Quantization for DNN Training by Exploiting Value Clustering
Georgia Institute of Technology, Atlanta, GA, Apr 2021
- Introduced Smart Quantization (SmaQ) technique for DNN training, which exploits value clustering in DNNs to reduce memory usage during training by up to 6.7x with no loss in accuracy.

Legal Text Summarization Using Transformer Models
Georgia Institute of Technology, Atlanta, GA, Nov 2020
- Presented work on a new transformer-based encoder-decoder architecture for abstractive legal text summarization, achieving state-of-the-art performance on the BIGPATENT dataset.

Attention is All You Need: The Transformer Architecture
Georgia Institute of Technology, Atlanta, GA, Sep 2020
- Presented the seminal Transformer paper by Vaswani et al. (2017) and discussed its impact on the field of natural language processing.

Projects (Ordered by Date)

Legal Text Summarization Using Transformer Models
Class project for CS 7643: Deep Learning & CS: 8803-DLT
- Developed a novel transformer-based encoder-decoder architecture for abstractive legal text summarization, achieving state-of-the-art performance on the BIGPATENT dataset.
- Converted the pretrained PEGASUS model to a PEGASUS-Longformer model to handle longer input sequences. Implemented training pipeline with support for distributed training, learning rate finding, batch size scaling, and gradient accumulation.
- Evaluated model performance against PEGASUS baselines and provided detailed analysis of results. Open-sourced codebase with instructions for reproducing results and extending the work.