Ali Taghibakhshi

Senior Deep Learning Algorithm Engineer, NVIDIA

Leander, TX | a.t.bakhshi@gmail.com | LinkedIn | Google Scholar | Website

Profile

Senior deep learning algorithm engineer building efficient foundation models across language, reasoning, and biology. Current work spans hybrid Mamba-Transformer architectures, structured pruning, distillation, quantization, elastic multi-budget inference, and distributed training infrastructure in NVIDIA NeMo, Megatron-LM, Megatron Bridge, BioNeMo, and ModelOpt.

Core Skills

Foundation Models	Deep Learning Efficiency, Pruning, Distillation, Quantization, Diffusion Models, Hybrid SSM/Attention Models, Reasoning Models, Model Alignment, Generative Biology Models.
Systems	Distributed Training, Large-Scale Training, Megatron-LM, NVIDIA NeMo Framework, BioNeMo, TensorRT ModelOpt, vLLM.

Professional Experience

NVIDIA Aug. 2023 – Present

Senior Deep Learning Algorithm Engineer Hybrid, Austin, TX

- Core contributor to **NVIDIA Nemotron Models** on the compression side, developing pruning, distillation, quantization, and elastic extraction methods across Nemotron-H, Nemotron Nano 2, **Minitron-SSM**, and **Nemotron Elastic**; reduced model-family training cost by 7x while improving inference efficiency by 2x.
- Contributed to large-scale software stacks in NVIDIA NeMo, Megatron-LM, Megatron Bridge, BioNeMo, and ModelOpt, with focus on distributed training, parallelism design, compression pipelines, model release tooling, and open technical reporting.
- Core contributor to **Evo 2**, a large-scale genomic foundation model published in *Nature*; led parallelization design and BioNeMo integration, and advised architecture and training configuration for **EDEN**, a 28B-parameter metagenomic foundation-model effort for programmable therapeutic design.

NVIDIA (May 2022 – Aug. 2022) — *Deep Learning Algorithms Intern, Remote, Santa Clara, CA*. Hierarchical GNN with cross-attention for billion-edge cross-device user matching; improved SOTA by 5% and produced a patent-submitted internal highlight.

John Deere (May 2020 – May 2022) — *Machine Learning Intern, Champaign, IL*. Computer vision and reinforcement learning for autonomous mower docking, parking assist, planting-condition reconstruction, YOLO perception, DQN control, and hardware validation.

University of Illinois Urbana-Champaign (Sep. 2019 – Aug. 2023) — *Graduate Research Assistant, Urbana, IL*. Graph learning and reinforcement learning for algebraic multigrid and domain-decomposition PDE solvers, with first-author NeurIPS and ICML papers.

Selected Papers

- **Nemotron Elastic: Towards Efficient Many-in-One Reasoning LLMs** (ICML, 2026). ***A. Taghibakhshi***, S. T. Sreenivas, S. Muralidharan, R. Cai, M. Chochowski, A. S. Mahabaleshwarkar, Y. Suhara, O. Olabiyi, D. Korzekwa, M. Patwary, M. Shoeybi, J. Kautz, B. Catanzaro, A. Aithal, N. Tajbakhsh, P. Molchanov.
- **Scaling Laws and Architectural Frontiers in Metagenomic Foundation Models** (ICML, 2026). G. Munsamy, G. Ayres, J. Dona, C. Greco, D. Anderson, S. Sridhar, W. Chow, A. Kollasch, R. Pecoraro, T. Bohnuud, K. Kam, G. Minto-Cowcher, M. Leung, H. Sirelkhatim, J. St. John, ***A. Taghibakhshi***, et al.
- **Minitron-SSM: Efficient Hybrid Language Model Compression through Group-Aware SSM Pruning** (NeurIPS, 2025). ***A. Taghibakhshi***, S. T. Sreenivas, S. Muralidharan, M. Chochowski, Y. Karnati, R. Joshi, A. S. Mahabaleshwarkar, Z. Chen, Y. Suhara, O. Olabiyi, D. Korzekwa, M. Patwary, M. Shoeybi, J. Kautz, B. Catanzaro, A. Aithal, N. Tajbakhsh, P. Molchanov.
- **Elucidating Optimal Reward-Diversity Tradeoffs in Text-to-Image Diffusion Models** (WACV, 2025). R. Jena, ***A. Taghibakhshi***, S. Jain, G. Shen, N. Tajbakhsh, A. Vahdat.
- **Genome Modelling and Design Across All Domains of Life with Evo 2** (Nature, 2026). G. Brixi, M. G. Durrant, J. Ku, M. Naghipourfar, M. Poli, G. Sun, G. Brockman, D. Chang, A. Fanton, G. A. Gonzalez, S. H. King, D. B. Li, A. T. Merchant, E. Nguyen, C. Ricci-Tam, D. W. Romero, J. C. Schmok, ***A. Taghibakhshi***, et al. *Part of the Core Evo 2 team; core contributor.*

- **MG-GNN: Multigrid Graph Neural Networks for Learning Multilevel Domain Decomposition Methods** (ICML, 2023). ***A. Taghibakhshi***, N. Nytko, T. U. Zaman, S. MacLachlan, L. N. Olson, M. West.
- **Learning Interface Conditions in Domain Decomposition Solvers** (NeurIPS, 2022). ***A. Taghibakhshi***, N. Nytko, T. U. Zaman, S. MacLachlan, L. Olson, M. West.
- **Optimization-Based Algebraic Multigrid Coarsening Using Reinforcement Learning** (NeurIPS, 2021). ***A. Taghibakhshi***, S. MacLachlan, L. Olson, M. West.

Technical Reports

- **Nemotron 3 Ultra: Open, Efficient Mixture-of-Experts Hybrid Mamba-Transformer Model for Agentic Reasoning** (NVIDIA, 2026).
- **Nemotron 3 Super: Open, Efficient Mixture-of-Experts Hybrid Mamba-Transformer Model for Agentic Reasoning** (NVIDIA, 2026).
- **NVIDIA Nemotron 3: Efficient and Open Intelligence** (NVIDIA, 2025).
- **Nemotron 3 Nano: Open, Efficient Mixture-of-Experts Hybrid Mamba-Transformer Model for Agentic Reasoning** (NVIDIA, 2025).
- **NVIDIA Nemotron Nano 2: An Accurate and Efficient Hybrid Mamba-Transformer Reasoning Model** (NVIDIA, 2025).
- **Nemotron-H: A Family of Accurate and Efficient Hybrid Mamba-Transformer Models** (NVIDIA, 2025).

Education

University of Illinois Urbana-Champaign Aug. 2019 – Aug. 2023
Ph.D., Mechanical Engineering Advisor: Matthew West

- Thesis focus: efficient scientific machine learning, graph neural networks for numerical solvers, and reinforcement learning for algebraic multigrid.

Sharif University of Technology Sep. 2015 – Jun. 2019
B.Sc., Mechanical Engineering