Naiyu Yin

Postdoctoral Research Associate, Department of Mathematics, Lehigh University

Google Scholar | naiyu.yin@gmail.com | LinkedIn | GitHub Personal Page

Summary

Machine learning researcher specializing in *causal discovery and inference*, *causal representation learning*, *Bayesian inference*, *uncertainty quantification*, *foundation models for scientific applications*, and *mechanistic interpretability of large language models (LLMs)*. First-author publications at **CVPR**, **AAAI**, **ECCV**, and ICPR, with additional work at **ICML**, **IJCAI**, NeurIPS and ICML workshops. Proven ability to translate research into deployable methods through industry collaborations with IBM Research, spanning LLM interpretability, causal discovery and causal representation learning for computer vision tasks.

Research Interests

- Foundation model pre-training and fine-tuning for scientific applications.
- Mechanistic interpretability (circuit learning and sparse autoencoders) of large language models.
- Causality-related research: causal discovery, causal inference, causal representation learning.
- Bayesian-related research: Bayesian inference, Bayesian neural networks, uncertainty quantification.

Education

Rensselaer Polytechnic Institute — Troy, USA
Ph.D. in Electrical, Computer, and Systems Engineering GPA: 3.77/4.00 — *May 2019 – Dec 2024*
M.S. in Applied Mathematics — *May 2019 – Dec 2023*

Duke University — Durham, USA
M.S. in Electrical and Computer Engineering GPA: 3.85/4.00 — *Aug 2017 – May 2019*

Beihang University — Beijing, China
B.Eng. in Electronic Information Engineering — *Sep 2013 – Jun 2017*

Selected Publications

- **Yin, N.**, Wang, H., Yu, Y., Gao, T., Dhurandhar, A., Lee, C.-H., & Ji, Q. (2026). CGU-Bayes: Causal graph uncertainty-guided Bayesian inference for domain generalization. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*
- **Yin, N.**, Wang, H., Yu, Y., Gao, T., Dhurandhar, A., & Ji, Q. (2024). Integrating Markov blanket discovery into causal representation learning for domain generalization. *European Conference on Computer Vision (ECCV).*
- **Yin, N.**, Gao, T., Yu, Y., & Ji, Q. (2024). Effective causal discovery under identifiable heteroscedastic noise model. *AAAI Conference on Artificial Intelligence.*
- **Yin, N.**, Yu, Y., Gao, T., & Ji, Q. (2024). Efficient nonlinear DAG learning under projection framework. *International Conference on Pattern Recognition (ICPR).* (Oral)
- **Yin, N.**, Wei, D., Gao, T., Dhurandhar, A., Ramamurthy, K., & Yu, Y. (2026). Scalable circuit learning for interpreting large language models. *Mechanistic Interpretability Workshop, ICML.*
- **Yin, N.**, & Yu, Y. (2026). Rational sparse autoencoder. *Mechanistic Interpretability Workshop, ICML.*
- **Yin, N.**, Liu, N., Lattimer, B. Y., Lua, J., & Yu, Y. (2025). Fake it till you make it: Multi-physics synthesis breaks the data barrier in chemical language models. *Machine Learning and the Physical Sciences Workshop, NeurIPS.*
- Cui, Z., **Yin, N.**, Wang, Y., & Ji, Q. (2022). Bayesian approaches for robust constraint-based causal discovery under insufficient data. *International Joint Conference on Artificial Intelligence (IJCAI).*
- Yu, Y., Gao, T., **Yin, N.**, & Ji, Q. (2021). DAGs with no curl: An efficient DAG structure learning approach. *International Conference on Machine Learning (ICML).*

Invited Talks

- **Yin, N.**, & Yu, Y. "Attention-Mechanism-Based Foundation Model for Causal Graph Learning." Invited talk, *18th U.S. National Congress on Computational Mechanics (USNCCM18)* and *SIAM Annual Meeting*, 2025.

Skills

Programming Languages: Python (proficient), MATLAB (proficient), R, C/C++, Java.

ML / AI Frameworks and Tools: PyTorch, TensorFlow, JAX, Hugging Face, Scikit-learn, Git, Docker.

Specialized AI / ML Techniques: Causal discovery and inference, causal representation learning, Bayesian inference, Bayesian neural networks, uncertainty quantification, pre-training and fine-tuning transformer-based foundation models, mechanistic interpretability of language models, and nonlocal neural operators.

Research Experience

Lehigh University Bethlehem, USA

Postdoctoral Research Associate *Feb 2025 – Present, Full-time*

- Built an attention-based foundation model for causal discovery that captures the mapping from observational data to causal graphs via an attention kernel, enabling efficient zero-shot causal graph inference; achieved substantial accuracy and efficiency gains over SOTA baselines on multiple benchmark datasets. *(Paper under review.)*
- Developed the *Rational Sparse Autoencoder (RSAE)*, replacing fixed SAE encoder activations with a trainable rational function to improve reconstruction and downstream fidelity for LLM interpretability. Published at the *Mechanistic Interpretability Workshop, ICML 2026*.
- Developed the *Poly4mer*, a physics-augmented chemical language model for polymer property prediction and targeted discovery. Published at the *Machine Learning and the Physical Sciences Workshop, NeurIPS 2025*.

IBM-Rensselaer Future of Computing Research Collaboration Yorktown Heights, USA

Research Scientist Intern *May 2024 – Aug 2024*

- Designed the *CircuitLasso*, a scalable circuit learning method for mechanistic interpretability of LLMs, using sparse linear regression to efficiently uncover relationships among high-dimensional sparse-autoencoder features, revealing how interpretable semantic features propagate through the model, and improving downstream task performance. Published at the *Mechanistic Interpretability Workshop, ICML 2026*.

IBM-Rensselaer AI Horizons Research Collaboration Yorktown Heights, USA

Research Scientist Intern *May – Aug 2020, 2021, 2022*

- Proposed an uncertainty-guided Bayesian inference framework to improve domain generalization performance in computer vision (CV) tasks. Introduced multi-level uncertainty quantification to evaluate model reliability. Published at *CVPR 2026*.
- Developed an invariant causal Markov blanket feature learning method to improve out-of-distribution generalization in CV tasks. Published at *ECCV 2024* and the *Causal Representation Learning Workshop, NeurIPS 2023*.
- Proposed an efficient causal discovery approach by proving the equivalence between the set of causal graphs and the set of weighted gradients of graph potential functions, hence enabling acyclicity-constraint-free causal graph learning over the equivalence space. Published at *ICPR 2024* (Oral) and *ICML 2021*.

Rensselaer Polytechnic Institute Troy, USA

Graduate Research Assistant *May 2019 – Dec 2024, Full-time Ph.D. Student*

- Built a unified framework combining causal discovery and representation learning for computer vision, designing structural-causal-model-based generative algorithms to capture and manipulate the image generation process; demonstrated counterfactual and controllable image generation via causal inference. *(Paper under review.)*
- Investigated causal discovery under heteroscedastic noise in structural equation models (SEMs); derived relaxed, implementable sufficient conditions proving identifiability for a general class of SEMs and proposed a two-phase iterative algorithm handling varying noise variance across variables and observations. Published at *AAAI 2024*.

Awards & Honors

Travel grant, *Society for Industrial and Applied Mathematics – New York – New Jersey – Pennsylvania Section (SIAM-NNP)*, 2025 Annual Meeting.

Best Industrial Paper Award nomination, *ICPR* 2024.

Professional Activities

Peer Review: Reviewer for NeurIPS, ICML, ICLR, CVPR, ICCV, AAAI, UAI, ECCV, AISTATS, and ARR, and for journals including the *Journal of Machine Learning Research (JMLR)*, *Journal of the American Statistical Association*, *IEEE Transactions on Information Theory*, *Computers & Mathematics with Applications* and *Communications Engineering*.

Conference Organization: Organized and hosted the minisymposium "Foundation Models: From Theory to Practice" at the *SIAM-NNP Annual Meeting*.