

Thao Nguyen

thaotn2@illinois.edu · (217) 693-8505 · Personal website

EDUCATION

University of Illinois Urbana-Champaign (UIUC)

Ph.D. Candidate in Computer Science

Aug. 2023 – Present

Advisor: Prof. Heng Ji

Hanoi University of Science and Technology (HUST)

B.S. in Electronics and Telecommunications Engineering

Hanoi, Vietnam

Aug. 2016 – Dec. 2020

Specialization: Biomedical Engineering

Advisor: Prof. Phuc Ngoc Pham

SELECTED PUBLICATIONS

6. **Thao Nguyen**, Heng Ji. MolLingo: Molecule-Native Representations for LLM-Powered Scientific Agents. *preprint arXiv:2605.27853*, 2026.
5. **Thao Nguyen**, Kuan-Hao Huang, Ge Liu, Martin D. Burke, Ying Diao, Heng Ji. FARM: Enhancing Molecular Representations with Functional Group Awareness. *Transactions on Machine Learning Research*, 2026.
4. Ziwen Wang, Jiajun Fan, Ruihan Guo, **Thao Nguyen**, Heng Ji, Ge Liu. ProteinZero: Self-Improving Protein Generation via Online Reinforcement Learning. *Transactions on Machine Learning Research*, 2026.
3. Carl Edwards*, Chi Han*, Gawon Lee, **Thao Nguyen**, Bowen Jin, Chetan Kumar Prasad, Sara Szymkuć, Bartosz A Grzybowski, Ying Diao, Jiawei Han, Ge Liu, Hao Peng, Martin D Burke, Heng Ji. mCLM: A Function-Infused and Synthesis-Friendly Modular Chemical Language Model. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR) (Oral Presentation)*, 2026.
2. Ziwen Wang, Jiajun Fan, **Thao Nguyen**, Heng Ji, Ge Liu. Variational Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
1. **Thao Nguyen**, Tiara Torres-Flores, Changhyun Hwang, Carl Edwards, Ying Diao, Heng Ji. GLaD: Synergizing Molecular Graphs and Language Descriptors for Enhanced Power Conversion Efficiency Prediction in Organic Photovoltaic Devices. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, 2024.

EXPERIENCE

BLENDER Lab, MMLI, CABBI, IIDAI

Research Assistant

University of Illinois Urbana-Champaign

2023 – Present

- **Toward Self-Evolving Chemical Intelligence: A Multi-Agent Framework for Autonomous Closed-Loop Molecular Discovery**
 - * Design a self-evolving, multi-agent chemical intelligence framework that autonomously plans, executes, and refines molecular discovery workflows via closed-loop interaction between generative models, predictive oracles, simulation, and autonomous synthesis and assays.
 - * Case study: Develop and integrate state-of-the-art solubility prediction and optimization pipelines, combining active learning with experimental feedback and task-specialized generative models to iteratively improve aqueous solubility predictions and optimize poorly soluble high-value compounds.
- **Evolution-Guided Agentic Discovery of Natural Product Therapeutics**
 - * Designed an agentic AI framework for autonomous natural product-based drug discovery, integrating evolutionary biology, curated NP-protein datasets, deep learning-based binding affinity prediction, and molecular simulation to prioritize therapeutically relevant NP-protein interactions.
- **mCLM – Modular Chemical Language Model for Drug Discovery**
 - * Contributed to the development of mCLM, a generative model that tokenizes molecules into functional building blocks and learns a bilingual language model of natural language descriptions and molecular building blocks

- * mCLM can significantly improve chemical functions critical to determining drug potentials in FDA-approved drugs and enhance the properties of FDA-rejected drugs (“fallen angels”).

- **ProteinZero – Generative Protein Design with Online Reinforcement Learning**

- * Contributed to the development of a generative model for protein design leveraging online reinforcement learning to continuously improve with new data.
- * Optimized generated proteins for structural accuracy, thermodynamic stability, designability, and diversity, enabling scalable and self-improving protein generation.

- **FARM model – Foundation Model for Small Molecule Representation**

- * Built an algorithm to automatically detect a comprehensive list of 100+ functional groups and annotate each atom with its corresponding functional group.
- * Developed functional group enhanced SMILES, a representation that preserves SMILES sequentialization while embedding functional group information into each atom, expanding the SMILES vocabulary to better resemble natural language and reduce negative transfer.
- * Trained a transformer-based foundation model on a large and diverse set of functional group enhanced SMILES, and used contrastive learning to align sequence and graph representations—leveraging transformer strengths while retaining molecular structural awareness.
- * Achieved state-of-the-art performance on 11 out of 13 MoleculeNet benchmarks, demonstrating strong generalization across molecular property prediction tasks.

- **GLaD model – Graph-Language Aligned Model for Organic Photovoltaics (OPV)**

- * Curated a high-quality dataset of 500 donor–acceptor pairs for organic photovoltaics by systematically mining and standardizing data from literature.
- * Built a PCE prediction model that takes donor–acceptor pairs as input, decomposes them into building blocks, and integrates molecular graph representations with natural language descriptions of OPV properties—mimicking how chemists study building blocks and their functions—to accurately predict power conversion efficiency.

VinUni-Illinois Smart Health Center - VISHC

Research Assistant

VinUniversity

2022 – 2023

- Collected a real-world 3-lead ECG dataset and developed an algorithm to digitize the signals.
- Built high-accuracy classification models for 12 cardiovascular diseases using 3-lead ECG signals and integrated them into a mobile application.
- Performed biomedical signal processing for sleep apnea detection and sleep stage classification.

Department of Medical Image Processing

Intern

VinBigdata Institute

2020 – 2021

- Collect a dataset of 8,000 chest X-ray images and their corresponding medical reports from electronic health records (EHRs) in a provincial general hospital.
- Develop a pipeline to automatically label medical images using data sourced from EHRs.
- Chest X-ray images processing
- Build a robust system capable of accurately classifying Chest X-ray images according to specific regions of pathology identified in the images.

EEG and Rehabilitation Laboratory

Undergraduate Research Assistant

Hanoi University of Science and Technology

2017 – 2020

- Biomedical signal and image processing.

TEACHING ASSISTANT EXPERIENCE

Natural Language Processing @VinUniversity

Algorithms and Data Structures for Data Science @UIUC

PROFESSIONAL SERVICES

Session Chair

- CIKM 2024

Reviewer

- ICLR 2026
- EMNLP 2025
- NeurIPS 2025
- ICML 2025
- ACL Rolling Review 2025
- ICLR 2025
- AI4Research Workshop @ AAAI 2025 - Best Reviewer Award
- Language + Molecules Workshop @ ACL 2024

REFERENCES

Prof. Heng Ji

Professor

Siebel School of Computing and Data Science

University of Illinois Urbana-Champaign

Email: *hengji@illinois.edu*

Prof. Huimin Zhao

Professor

Department of Chemical & Biomolecular Engineering

University of Illinois Urbana-Champaign

Email: *zhao5@illinois.edu*