Thao Nguyen

thaotn2@illinois.edu

Education

University of Illinois Urbana-Champaign (UIUC)
Ph.D. Candidate in Computer Science — *Aug. 2023 – Present*
Co-advisors: Prof. Heng Ji and Prof. Martin D. Burke

Hanoi University of Science and Technology (HUST) — Hanoi, Vietnam
B.S. in Electronics and Telecommunications Engineering — *Aug. 2016 – Dec. 2020*
Specialization: Biomedical Engineering
Advisor: Prof. Phuc Ngoc Pham

Experience

BLENDER Lab, MMLI, CABBI — University of Illinois Urbana-Champaig
Research Assistant — *2023 – Present*

- **Hypothesis Generation for Natural ProductPprotein Interaction Exploration**
  - Developing a comprehensive framework that integrates biosynthetic gene cluster (BGC) mining to predict novel natural products and a protein-targeting pipeline to identify potential protein partners for each compound.
  - Building a multiple instance learning model that predicts binding potency by focusing on protein pockets, enabling compound–pocket level interaction prediction.
  - Extending the framework to solve the reverse problem: given a protein of interest, identify natural products from large databases that can interact with it.
  - Provides meaningful applications in drug discovery and repurposing, such as identifying bioactive natural products, uncovering novel therapeutic targets, and accelerating the early stages of antibiotic and anticancer drug development.
- **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 VinUniversity

Research Assistant *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 VinBigdata Institute

Intern *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 Hanoi University of Science and Technology

Undergraduate Research Assistant *2017 – 2020*

- Biomedical signal and image processing.

Publications

11. Ziwen Wang, Jiajun Fan, Ruihan Guo, **Thao Nguyen**, Heng Ji, Ge Liu. ProteinZero: Self-Improving Protein Generation via Online Reinforcement Learning. *preprint arXiv:2506.07459, 2025.*

10. 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. *preprint arXiv:2505.12565, 2025.*

9. Ziwen Wang, Jiajun Fan, **Thao Nguyen**, Heng Ji, Ge Liu. Variational Supervised Contrastive Learning. *In Advances in Neural Information Processing Systems (NeurIPS 2025), accepted.*

8. **Thao Nguyen**, Kuan-Hao Huang, Ge Liu, Martin D. Burke, Ying Diao, Heng Ji. FARM: Functional Group-Aware Representations for Small Molecules. *preprint arXiv:2410.02082*, 2024.

7. **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 *ACL 2024 Workshop Language and Molecules*, 2024.

6. Dat T. Ngo, **Thao Nguyen**, Hieu T. Nguyen, Dung B. Nguyen, Ha Q. Nguyen, Hieu H. Pham. Slice-level Detection of Intracranial Hemorrhage on CT Using Deep Descriptors of Adjacent Slices. In *2023 IEEE Statistical Signal Processing Workshop (SSP)*, 2023.

5. **Thao Nguyen**, Hieu H. Pham, Khiem H. Le, Anh Tu Nguyen, Tien Thanh, Cuong Do. Detecting COVID-19 from digitized ECG printouts using 1D convolutional neural networks. In *Plos One*, 2023.

4. **Thao Nguyen***, Anh Tu Nguyen*, Khiem H. Le, Hieu H. Pham, Cuong Do. A novel deep learning-based approach for sleep apnea detection using single-lead ECG signals. In *APSIPA ASC - 2022*, 2022.

3. **Thao Nguyen***, Tam M. Vo*, Thang V. Nguyen, Hieu H. Pham, Ha Q. Nguyen. Learning to diagnose common thorax diseases on chest radiographs from radiology reports in Vietnamese. In *Plos One*, 2022.

2. Khiem H. Le, **Thao Nguyen**, Hieu H. Pham, Tu A. Nguyen, Tien N. Thanh, Cuong D. Do. LightX3ECG: A Lightweight and eXplainable Deep Learning System for 3-lead Electrocardiogram Classification. In *Biomedical Signal Processing and Control.*

1. Khiem H. Le, Hieu H. Pham, **Thao Nguyen**, Tu A. Nguyen, Cuong D. Do. Enhancing deep learning based 3-lead ECG classification with heartbeat counting and demographic data integration. In *IECBES 2022.*

Teaching Assistant Experience

Natural Language Processing at VinUniversity

Algorithms and Data Structures for Data Science at UIUC

Professional Services

Session Chair

- CIKM 2024

Reviewer

- 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@illiois.edu*

Prof. Martin D. Burke
Professor
Department of Chemistry
University of Illinois Urbana-Champaign
Email: *mdburke@illinois.edu*