

Jonny C. Tran

CS Ph.D. in Graph Machine Learning

EDUCATION

Ph.D. in Computer Science

Aug 2015 – Dec 2023

B.S. in Computer Science

Aug 2011 – Aug 2015

The University of Texas at Arlington, GPA 3.6

Dissertation: "Graph Representation Learning for Heterogeneous Multimodal Biomedical Data"

WORK EXPERIENCE

Graduate Research Asst. / Lab Manager

Aug 2015 – Present

BioMeCIS Lab, UT Arlington, TX

- Autonomously carrying out long-term projects that resulted in 6 first-author papers for computational methods in bioinformatics.
- Pre-trained an attention-based **NLP transformer** model on **graph structure** and NLP biological **sequence data**; Optimized its architecture to reduce memory usage for downstream predictions.
- Built a **BERT** model to encode **biological knowledge text** and RNA sequences to find associations within the knowledge-graph.
- Identified a project to create an **open-source** Python package for scalable processing of public multi-omics biological datasets.
- Deployed automated hyperparameter optimization and model architecture explorations with Weights & Biases on HPC cluster.
- Led efforts to install deep learning solutions for lab members with 6 GPU nodes, 2 data servers, 2 web servers, and **JupyterHub/K8s**.

Bioinformatics Intern

Aug 2021 – Feb 2022

Genentech, South San Francisco, CA

- Designed a **pipeline** to **preprocess & harmonize datasets** and **simulate** 3 sequencing parameters with dynamic programming optimizations, saving 36% HPC runtime and 10's TBs.
- Benchmarked 6 ML baselines (e.g. linear, Bayes and tree methods) to detect poor-quality samples in imbalanced multi-task classification.
- Performed **interpretability** analysis with Random Forests to report **statistically** relevant QC features.
- Engineered 100's custom features from TBs of **unstructured** genomics data for QC determinants in a personalized drug pipeline.
- Communicated new findings with 20+ interactive data visualizations to stakeholders in two **interdisciplinary** teams.
- Wrote documentation for all code and processes developed, ensuring data and analysis pipeline reproducibility.

Business Intelligence Intern

Jun 2015 – Aug 2015

USAA, Plano, TX

- Developed front-end and back-end of an internal production DevOps web app from scratch with a mid-size **Agile development** team, from gathering customer requirements to production deployment.
- Built data pipelines with SQL-based ETL using IBM DataStage and deployed automated jobs on enterprise Linux servers.

CONTACT

- Arlington, TX
- +1 (469) 279-0297
- nhat.c.tran@gmail.com
- [linkedin.com/in/nhatctran](https://www.linkedin.com/in/nhatctran)
- github.com/JonnyTran

SKILLS

Python:

- **Pandas** & Dask & PySpark
- **scikit-learn**
- **NetworkX**
- NumPy, SciPy
- **Plotly**, Dash, **seaborn**

Deep Learning:

- **PyTorch & PyTorch-Lightning**
- PyTorch-Geometric, DGL
- **NLP (HuggingFace, SpaCy)**
- **GPT-4**, GH Copilot, gpt-engineer
- Weights and Biases
- TensorFlow & Keras

Distributed Computing:

- **Dask**
- **Docker**
- JupyterHub
- **Hadoop**
- SLURM
- Snakemake

Software Engineering:

- Java
- R
- C++
- JavaScript (AngularJS, D3.js)
- **CI/CD (GitHub Actions, pytest)**
- SQL (**MySQL**)
- **NoSQL** (MongoDB)
- **Agile methodologies**

Soft skills:

- Data exploration
- Technical writing
- Presentation skills
- Collaborative communication
- Mentoring

RELATED COURSES

- Optimization on Big Data
- Scalable Learning & Optimization
- Convex Optimization
- Information Security
- Computer Network
- Artificial Intelligence

RESEARCH PROJECTS AND SELECT PUBLICATIONS

LATTE2GO

In review.

"Protein function prediction by incorporating knowledge graph representation of heterogeneous interactions and gene ontology"

- Developed a GNN method to learn representations in knowledge graphs for end-to-end learning of heterogeneous relationships and taxonomy of classes, while accurately predicting protein function at the same computational complexity as standard node classification.
- Used Dask to collect Gene Ontology, protein, and RNA interaction networks from 13 data sources to construct a heterogeneous graph of 10M nodes spanning multiple species.
- Achieved 6% higher AUPR in blind protein function predictions by learning on higher-order relations across multiple GNN layers.

LATTE

[Tran, Nhat et al. \(2022\) arXiv:2009.08072](#)

"Layer-stacked attention for heterogeneous graph embedding"

- Created a general GNN architecture for heterogeneous graphs to automatically generate higher-order neighborhood structure with a new multi-hop attention mechanism.
- Enabled interpretability by identifying salient relationships and outperformed other GNN methods on benchmark datasets by 2-5% AUROC in node classification.

OpenOmics

[Software](#)

[Tran, Nhat et al. \(2021\) Journal of Open Source](#)

"A bioinformatics API to integrate multi-omics datasets and interface with public databases"

- Open-source Python package to interface with 20+ online databases with scalable out-of-memory dataframe operations using Dask.
- Designed an API allowing users to create a pipeline to integrate multimodal data types such as graphs, texts, and tables into a ML-ready dataset.

rna2rna

[Biocomputing](#)

[Tran, Nhat et al. \(2020\) Pacific Symposium on](#)

"Network representation of large-scale heterogeneous RNA sequences with integration of multi-modal data"

- Created a ConvNet + LSTM architecture for Contrastive Learning with TensorFlow to encode variable-length sequences with an inductive bias capturing both first- and second-order proximities at only first-order computational complexity. Outperforms SoTA graph embedding methods at 90% AUPR in link prediction.

MDSN

[Tran, Nhat et al. \(2018\) BMC Bioinformatics](#)

"Discovering microRNA dysregulatory modules across subtypes in non-small cell lung cancers"

- A graph-based feature selection method using differential analyses with p-value thresholds on multi-omics RNA expression data.
- Feature groups selected for a Sparse Group Lasso classifier resulted in 10% higher accuracy at predicting cancer stages.

AWARDS

- U-HACK MED: won in code sharing and reproducibility category at biomedical hackathon.
- NTx Apps Challenge: Won \$10k with a traffic management system at sustainability hackathon.

RESEARCH CONTRIBUTIONS

Organization:

- Next-Generation Sequencing @ IEEE BIBM '17: As session chair, organized talks, discussions, and papers of researchers in the field.

Paper Reviewing:

- IEEE NNLS '21
- AAAI '19
- IEEE BIBM '20
- KDD '20
- BMC Bioinformatics '18
- IEEE BIBM '18

BIOGRAPHICAL

Citizenship:

- U.S. Citizen

Languages:

- English (native)
- Vietnamese (native)

OTHER INTERESTS

Sports:

- Breakdancing, Lindy Hop, Brazilian jiu-jitsu, rock climbing.

Leisure:

- Data-driven espresso brewing, coffee roasting, hiking & traveling.