Jonny C. Tran

CS Ph.D. in Machine Learning & Bioinformatics

EDUCATION

Ph.D. in Computer Science August 2015 – December 2022
B.S. in Computer Science August 2011 – August 2015
The University of Texas at Arlington, GPA 3.6
Dissertation: *"Graph Representation Learning for Heterogeneous Multimodal Biomedical Data"*

WORK EXPERIENCE

Ph.D. Student / Lab Manager August 2015 – December 2022
BioMeCIS Lab, UT Arlington, TX

- Produced 6 first-author publications to contribute novel **graph ML algorithms** to premier venues such as PSB and BMC Bioinformatics.
- Proposed graph neural network architectures to aggregate heterogeneous relationships for link prediction, node classification, and graph classification tasks.
- Pre-trained an attention-based transformer model on biological **sequence data** (proteins, non-coding RNAs) and optimized its architecture to reduce memory usage for downstream GNNs.
- Deployed automated hyperparameter optimization and model architecture explorations with Weights & Biases on the HPC cluster, achieving nearly 100% utility on A100 GPUs.
- Led efforts to install and manage 6 GPU nodes, 2 data servers, 2 web servers, and deep learning with JupyterHub.
- Created an open-source **Python** framework to aggregate real-world heterogeneous multi-omics bio datasets at the 10's GB scale.

Bioinformatics Intern August 2021 – February 2022
Genentech, San Francisco, CA

- Identified and engineered 100's custom features from TBs of NGS genomics data to identify QC determinants for sequencing quality in a personalized drug pipeline.
- Architected a **data generation pipeline** using dynamic programming to **harmonize datasets** and simulate 3 sequencing parameters; then optimized the DAG workflow, saving 36% HPC runtime and 10's TBs.
- Benchmarked 6 ML baseline models to detect poor-quality samples in an imbalanced multi-task classification and performed interpretability analysis with Random Forests to report key QC features.
- Presented new insights with 20+ interactive data visualizations to cross-functional teams in bioinformatics, statistics & manufacturing.
- Wrote documentation for all code and processes developed, ensuring data and analysis pipeline reproducibility.

Business Intelligence Intern June 2015 – August 2015
USAA, Plano, TX

- Built data pipelines with SQL-based ETL using IBM DataStage and deployed automated jobs on enterprise Linux servers.

CONTACT

- Fort Worth, TX (open to relocate)
- +1 (469) 279-0297
- nhat.c.tran@gmail.com
- linkedin.com/in/nhatctran
- github.com/JonnyTran

SKILLS

Python:
- Pandas / Dask (Advanced)
- NetworkX
- NumPy & SciPy
- PySpark
- Plotly
- SciKit-Learn

Deep Learning:
- **PyTorch**
- PyTorch-Geometric & DGL
- **HuggingFace**
- PyTorch-Lightning
- **Weights and Biases**
- **TensorFlow & Keras**

Distributed Computing:
- Dask
- Docker
- JupyterHub
- Horovod
- SLURM
- Snakemake

Bioinformatics:
- Biopython
- Scanpy
- GATK / Picard
- samtools
- BWA
- **Snakemake**

Software Engineering:
- Java
- R
- JavaScript (AngularJS, D3.js)
- Travis CI
- GitHub Actions
- pytest
- Agile methodologies

Soft skills:
- Technical writing
- Presentation skills
- Collaborative communication
- Data visualization
- Mentoring

- Developed front-end and back-end of an internal production DevOps web app in a mid-size **Agile development** team, from gathering customer requirements to production deployment.

RESEARCH PROJECTS AND SELECT PUBLICATIONS

LATTE2GO

In review @ ISMB.

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

- Developed a method for protein function prediction with knowledge-graph integration of Gene Ontology, protein, and RNA interactions.
- Constructed knowledge graph of 10M nodes from multiple interaction network sources
- Achieved 6% higher AUPR than state-of-the-art using self-attention mechanism 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 to heterogeneous graphs to generate higher-order neighborhood structures automatically.
- Enabled interpretability by identifying salient relationships and outperformed other GNN methods on benchmark datasets by 2-5% AUROC in node classification.

OpenOmics

Tran, Nhat, et al. (2021) Journal of Open Source Software

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

- Created an open-source Python package with modern software engineering, **continuous-integration** and **automated testing** to interface with 20+ online databases.
- Designed an API allowing users to create a pipeline to integrate multimodal data types such as graphs, texts, and tables with scalable out-of-memory data frame operations using Dask.

rna2rna

Tran, Nhat, et al. (2020) Pacific Symposium on Biocomputing

"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 **non-coding RNA** 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-seq expression data.
- Feature groups selected for a Sparse Group Lasso classifier resulted in 10% higher accuracy in predicting cancer stages.

RELATED COURSES

- Optimization on Big Data
- Scalable Learning & Optimization
- Convex Optimization
- Game Theory
- Information Security
- Computer Network
- Compilers

AWARDS

- U-HACK MED @ UTSW: 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.