Jonny C. Tran

Engineer with a Ph.D. in Machine Learning for Biology

EDUCATION

Ph.D. in Computer Science Aug 2015 – Dec 2022
B.S. in Computer Science Aug 2011 – Aug 2015
The University of Texas at Arlington, GPA 3.6, h-index 4
Dissertation: *"Graph Representation Learning for Heterogeneous Multimodal Biomedical Data"*

WORK EXPERIENCE

Research Scientist, Malaria, Contractor Sep 2023 – Jun 2024
Institute for Disease Modeling, Gates Foundation, Seattle, WA

- Spearheaded an end-to-end scalable **production** data extraction pipeline, leveraging retrieval-augmented generation (RAG) to automate retrieval and extraction of structured data from a corpus of 200 scientific papers at 93.6% precision over 83 domain-specific data fields.
- Deployed a **GPU-optimized** pipeline for PDF parsing service, optimizing for accuracy, latency, and deployment costs for vision transformer architectures with **quantization techniques** and **Pytorch optimizations**.
- Collaborated with interdisciplinary scientists to design a human-in-the-loop workflow for generating high-quality datasets. It reduced manual annotation time by 3x, facilitating downstream modeling efforts.
- Designed **dataset collection** & annotation for RAG finetuning and LLM observability patterns with Langfuse & Weights and Biases integrations.
- Integrated models and microservices into a scalable system using Tilt on **Kubernetes.** Continued contributing to the project as an **open-source** Python & web library (Extralit) providing enhancements post-contract.

Graduate Research Assistant Aug 2015 – Aug 2023
University of Texas at Arlington, TX

- Developed **multimodal deep learning** architecture using **graph neural networks** and **transformers** for biological sequence analysis, for representation learning and semi-supervised recommendation tasks.
- Deployed **infrastructure** for **distributed** PyTorch model training with quantization optimizations using PyTorch Lightning, Ray, Weights & Biases and **Docker** on a self-hosted GPU cluster.

Bioinformatics Intern Aug 2021 – Feb 2022
Genentech, South San Francisco, CA

- Collaborated with biostatistics and manufacturing teams to establish actionable sequencing QC thresholds for large-scale clinical trials, achieving 92% sensitivity in identifying low-quality samples.
- Implemented **production-grade** data workflows with snakemake to process, harmonize, and simulate sequencing on terabytes of whole-exome and whole-genome data, producing a **ML-ready** dataset.
- Optimized the data pipeline with dynamic programming, reducing HPC runtime by 36% and achieving significant storage savings.
- Designed custom interactive **data visualizations** to present & provide interpretation for the proposed QC thresholds with statistical analysis.

CONTACT

- Seattle, Washington
- jonny.bio
- linkedin.com/in/nhatctran
- github.com/JonnyTran

SKILLS

Python:
- Pandas, Dask, PySpark, Pandera
- NumPy, SciPy
- FastAPI, Pydantic
- Pytest
- Snakemake

Machine Learning:
- **PyTorch, Lightning**
- **NLP (Transformers, BERT, SentenceTranformers)**
- **LLM (SFT finetuning)**
- **Graph Attention Networks**
- PyTorch-Geometric (PyG)
- Weights and Biases
- TensorFlow & Keras
- **Huggingface**
- **Llama-Index, LiteLLM, Langfuse**

Big Data / Infrastructure:
- **Docker**, **Kubernetes**, Tilt
- Caddy, Traefik, Nginx
- Weaviate, Elasticsearch
- AWS S3, MinIO
- **Dask, PySpark**
- JupyterHub
- SQL (Postgres, SQLAlchemy, Alembic)
- HPC (SLURM)

Data Visualization:
- **Plotly**, Dash
- ggplot2
- **D3.js**

Software Engineering:
- R
- C++
- Vue.js, Nuxt
- CI/CD (GitHub Actions)
- Agile methodologies

Soft skills:
- Project management
- Data visualization
- Technical writing
- Presentation skills
- Collaboration and Communication

RESEARCH PROJECTS AND SELECT PUBLICATIONS

LATTE2GO

Tran, Nhat et al. (2023) IEEE BIBM

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

- Developed a graph deep learning method using attention mechanisms comparable to transformer architectures to accurately **predict protein functions**, even with limited information, by analyzing a 10M-scale knowledge graphs of protein interactions and gene functions, achieving a 6% accuracy improvement in benchmarks.

LATTE

Tran, Nhat et al. (2022) arXiv:2009.08072

"Layer-stacked attention for heterogeneous graph embedding"

- Created a general graph deep learning model capable of automatically revealing hidden patterns and connections in diverse networks, demonstrating a 2-5% improvement in classification performance over existing graph embedding methods.

OpenOmics

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

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

- Developed an open-source data integration tool for scientists to easily access and integrate diverse biological datasets (up to 20+ public databases) with scalable out-of-memory data workflows 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"

- Built an LSTM-based deep learning model to analyze and classify RNA sequences, accurately predicting their functions and relationships. Achieved a 90% accuracy in predicting interactions for sparsely annotated class of lncRNAs, surpassing existing methods.

MDSN

Tran, Nhat et al. (2018) BMC Bioinformatics

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

- Developed a computational method to identify key RNA molecules involved in different subtypes of lung cancer. Improved accuracy in predicting cancer stages by 10%.

AWARDS

- U-HACK MED '19: Won the code sharing and reproducibility category at biomedical hackathon.
- NTx Apps Challenge '14: Won $10k with a traffic management system at sustainability hackathon.

RESEARCH CONTRIBUTIONS

Organization:

- Next-Generation Sequencing @ IEEE BIBM '17: As session chair, organized talks and facilitated discussions among bioinformatic researchers.

Paper Reviewing:

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