Jonny C. Tran

CS Ph.D. in Graph and NLP Machine Learning

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

WORK EXPERIENCE

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

- Spearheaded the development of an end-to-end LLM pipeline, leveraging advanced retrieval-augmented generation (RAG) to automate retrieval and extraction of structured data tables at 93.6% precision from 200 scientific papers on complex schemas with up to 82 data fields.
- Worked closely with users, team of malaria scientists and engineers to gather requirements and conduct user testing, ensuring the developed **human-in-the-loop** LLM workflow reduces the annotation time by 3x.
- Established robust evaluations for continuous assessment of LLM performance, utilizing the Argilla data labeling framework to collect user queries, evaluate responses & track key performance indicators.
- Rigorously evaluated the performance of 7 open-source models and OCR services using comprehensive metrics and developed a custom PDF parsing pipeline with optimized accuracy, latency, and deployment costs.
- Architected a custom workflow with configurable deployments and live reloading using **Tilt** on **Kubernetes** to orchestrate vector database, FastAPI servers, cloud storage, ElasticSearch, Postgres, LLM instrumentation, and **Pytorch** model inference servers as microservices.
- Open-sourced "Extralit", a Python data extraction library for the **scientific community** to expand research impact across the organization.

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

- Independently led multiple long-term research projects, culminating in 6 first-author publications in top-tier computational biology conferences and journals.
- Contributed to the open-source scientific community by developing and maintaining OpenOmics, a Python library for scalable & reproducible data workflows.
- Optimized training of Pytorch model with quantization optimizations using Weights & Biases and Docker on multi-GPUs HPC environment.

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

- Collaborated with statisticians, bioinformaticians and genomics experts to apply machine learning to Whole-Genome sequencing and gain new insights in quality control metrics that impacts patient treatment.
- Presented to cross-functional bioinformatics & manufacturing teams to define a set of actionable QC thresholds from clinical trials data.

CONTACT

- Seattle, Washington
- +1 (469) 279-0297
- nhat.c.tran@gmail.com
- linkedin.com/in/nhatctran
- github.com/JonnyTran

SKILLS

Python:
- **Pandas**, Dask, PySpark, Pandera
- NumPy, SciPy
- **Plotly**, Dash
- FastAPI, Pydantic

Machine Learning:
- **PyTorch, Lightning**
- **LlamaIndex**
- **Langfuse, Ragas**
- **Unstructured, Deepdoctection, Nougat, PyMuPDF**
- **NLP (Transformers, BERT, SetFit, GLiNER)**
- PyTorch-Geometric (PyG), DGL
- Weights and Biases
- TensorFlow & Keras

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

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

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

- Architected a pipeline to harmonize 5 datasets, simulate 3 sequencing parameters, extract 100's of custom features from TBs of unstructured genomics data using Python & Spark. Optimized with dynamic programming, saving over 36% HPC runtime and 10's TBs.
- Benchmarked 6 ML baselines in imbalanced multi-task classification to predict poor-quality samples, then performed interpretability analysis on Random Forests to identify statistical thresholds.

RESEARCH PROJECTS AND SELECT PUBLICATIONS

LATTE2GO Tran, Nhat et al. (2023) BIBM

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

- Developed a graph deep learning method to accurately predict protein functions, even with limited information, by analyzing the complex 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 a graph 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: 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 and facilitated discussions among bioinformatic researchers.

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:
- Breaking, Lindy Hop, Brazilian jiu-jitsu, rock climbing.

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