YUWEI MIAO

416 Summit Ave, Arlington, Texas, 76013 | yxm9326@mvas.uta.edu | Tel: (+1) 9842158487

RESEARCH INTEREST

Gene function prediction, Graph Neural Networks, Drug discovery, Computational Biology, Transcriptomic, Large Language Models, Machine Learning, Deep Learning

EDUCATION

The University of Texas at Arlington (UTA) **Arlington, TX**

Doctor of Philosophy in Computer Science and Engineering **GPA: 4.0/4.0** *Aug. 2022 – Present*

- **Graduate Research Assistant:** Supervised by Professor Junzhou Huang.
- **Graduate Teaching Assistant:** CSE 5360 Artificial Intelligence, CSE 5301 Data Analysis & Modeling Techs

The University of North Carolina at Chapel-Hill *(UNC)* **Chapel Hill, NC**

B.S. in Computer Science, B.A. in Biology *Aug. 2018 – May. 2022*

- **Core Course:** *Data Structures, Computer Organization, Discrete Math, Linear Algebra, MODELS: LANGS/COMPN, Bioalgorithms, Files and Databases, Algorithms & Analysis, Intro to Machine Learning, Computational Photography, Modern Web Programming, Ecology & Evolution, Molec Biol & Genet, Cellular Self Assembly*

RESEARCH & PROJECT EXPERIENCE

GO Annotation with LLM

The University of Texas at Arlington **Arlington, TX**

Graduate Research Assistant, Supervisor: Prof. Junzhou Huang *Nov. 2024 - present*

- Propose an end-to-end pipeline to annotate GO functions from publications.
- Collect and build new datasets and benchmarks for each subtask, such as GO entity recognition, qualifier extraction, etc.
- Preparing a manuscript for submission to ISCB 2025.

Protein-ligand binding affinity prediction

The University of Texas at Arlington **Arlington, TX**

Graduate Research Assistant, Supervisor: Prof. Junzhou Huang *Nov. 2024 - present*

- Responsible for experimental design, parameter tuning, and manuscript proofreading for this project.
- The second author of a manuscript submitted to CVPR 2025.

GoBERT: Gene Ontology Graph Informed BERT for Universal Gene Function Prediction

The University of Texas at Arlington **Arlington, TX**

Graduate Research Assistant, Supervisor: Prof. Junzhou Huang *Mar. 2024 - Dec. 2024*

- Develop a BERT model with two pre-train tasks to capture both explicit and implicit relations among functions.
- Introduce a new task to predict new functions with known functions of gene and gene products.
- Contribute a benchmark dataset for this prediction task.
- Paper Accepted by AAAI 2025.

UniEntrezDB: Large-scale Gene Ontology Annotation Dataset and Evaluation Benchmarks with Unified Entrez Gene Identifiers

The University of Texas at Arlington **Arlington, TX**

Graduate Research Assistant, Supervisor: Prof. Junzhou Huang *Mar. 2024 - present*

- Collect Gene Ontology Annotations from 21 databases.
- Align gene and gene product identifiers from various databases to NCBI Gene Entrez ID.
- Organized dataset published on Zendo https://zenodo.org/records/13335548
- Data processing code and benchmark reproduction code are available at https://github.com/MM-YY-WW/UniEntrezDB
- Preparing a manuscript for submission.

The Influence of Missense Mutation in Protein-Protein Interaction

The University of Texas at Arlington **Arlington, TX**

Graduate Research Assistant, Supervisor: Prof. Junzhou Huang *Jan. 2023 - Jan. 2024*

- Conduct presentations several times about biology knowledge. Investigate, collect, and analyze DNA sequence data from commonly used databases such as NCBI, GeneBank, etc.
- Reproduce baselines and participate in the model design to predict the perturbation in PPI resulting from missense mutation.

Survey of Recent Advances in Toxicity Prediciton

The University of Texas at Arlington **Arlington, TX**

Graduate Research Assistant, Supervisor: Prof. Junzhou Huang *Oct. 2022 - Aug. 2023*

- Comprehensively investigate the method of toxicity prediction from the very beginning of wet-lab experiments to statistical machine learning method and then deep learning and pre-training methods.
- Reproduce the code of representative works throughout the development of toxicity prediction in drug discovery.
- The survey paper was accepted by the Chemical Research in Toxicology.

Digital Image Processing and Segmentation **Remote**

Instructor: Prof. Munib from Harvard University *July. 2021 - May. 2022*

- Performs basic operations on images with Matlab.
- Use data augmentation and incorporate squeeze-and-excitation block to improve the performance of CNN.
- The 2022 14th International Conference on Computer Research and Development (ICCRD) accepted the paper.

PROFESSIONAL EXPERIENCE

Shenzhen Yuntian Lifei Technology Co., Ltd. **Shenzhen, China**

Instructor: Prof. Tianbao Yang from the University of Iowa *July. 2019 - Aug. 2019*

- Learned the basics of machine learning, computer vision, and the design of products like binocular near-infrared face recognition.
- Studied basic algorithms using machine learning; tested and classified data sets under the guidance of Prof. Yang

Silicon Valley Technology Innovation Club, Inc. **Santa Clara, CA**

Instructor: Mrs. Iris Lei *May. 2019 - July. 2019*

- Learned Python, neural networks, Linux systems, and Python libraries (NumPy, Pandas, Matplotlib). Evaluated data annotation quality and trained models.
- Used the knowledge related to computer vision, collected product image information in real-time by installing cameras on the assembly line to detect flaws in products

PUBLICATIONS

- **Yuwei Miao**, Yuzhi Guo, Hehuan Ma, Jingquan Yan, Feng Jiang, Rui Liao, Junzhou Huang. “GoBERT: Gene Ontology Graph Informed BERT for Universal Gene Function Prediction. Accept by *Proceedings of the AAAI Conference on Artificial Intelligence* 2025.
- **Yuwei Miao**, Yuzhi Guo, Hehuan Ma, Jingquan Yan, Feng Jiang, Weizhi An, Jean Gao, Junzhou Huang. “UniEntrezDB: Large-scale Gene Ontology Annotation Dataset and Evaluation Benchmarks with Unified Entrez Gene Identifiers” arXiv:2412.12688.
- **Yuwei Miao**, Hehuan Ma, and Junzhou Huang. "Recent advances in toxicity prediction: Applications of deep graph learning." Chemical Research in Toxicology 36.8 (2023): 1206-1226.
- **Yuwei Miao**, and Wenyi Luo. "Improve Generalization Ability of CNN by Data Augmentation and SE Block in Landmark Classification." 2022 14th International Conference on Computer Research and Development (ICCRD). IEEE, 2022.

SKILLS & INTERESTS

Languages: Chinese (Native); English (Fluent)

Technical Skills: Python, Pytorch, Machine Learning, Deep Learning, Linux System