Yuwei Miao

416 Summit Ave, Arlington, TX 76013
Email: yxm9326@mavs.uta.edu
Tel: (+1) 984-215-8487

Research Interests

Gene Function Prediction, Graph Neural Networks, AI for Science, Drug Discovery, Transcriptomics, GWAS, Perturbation Response, Virtual Cell, Large Language Models, Deep Learning

Education

- **The University of Texas at Arlington** — Arlington, TX
  Ph.D. in Computer Science and Engineering, GPA: 4.0/4.0 — *Aug 2022 – Present*
  - Graduate Research Assistant, Advisor: Dr. Junzhou Huang
  - Graduate Teaching Assistant: CSE 5360 Artificial Intelligence; CSE 5301 Data Analysis & Modeling; CSE 2315 Discrete Structures
- **University of North Carolina at Chapel Hill** — Chapel Hill, NC
  B.S. Computer Science; B.A. Biology — *Aug 2018 – May 2022*
  - Relevant Coursework: Algorithms, Machine Learning, Bioalgorithms, Linear Algebra, Molecular Biology, Databases

Research Experience

- **SMILES Lab, University of Texas at Arlington** — Arlington, TX
  Graduate Research Assistant — *Jun 2022 – Present*
- **Orthology-Aligned Foundation Model of Single-Cell Genetic Perturbation Responses**
  Nov 2025 – Present
  - Curated and unified 42 single-cell genetic perturbation datasets from 19 studies across 2 species, resulting in a large-scale corpus of 24.7M cells for perturbation pretraining.
  - Proposed the first self-supervised single-cell genetic perturbation foundation model, capturing perturbation effects as cell state transformations in latent space.
  - A 2-stage self-supervised pretraining with mask and recovery is proposed. It first learns control cell representations, setting the foundation for learning distinguishable genetic perturbation operators in the next stage.
  - Enabled accurate perturbation response prediction and reliable inference of pathway activity changes after perturbations.
  - Manuscript submitted to ICML 2026
- **Optimization Conflict and Correlation Limits in Attention-Based Models**
  Jun 2025 – Present
  - Discovered and formally defined a correlation plateau phenomenon in attention-based regression under joint MSE and correlation optimization, revealing a previously uncharacterized training bottleneck.
  - Developed a theoretical framework analyzing gradient dynamics and convex attention aggregation, and proved an intrinsic upper bound on achievable correlation gains.
  - Designed Extrapolative Correlation Attention (ECA), an aggregation mechanism that provably breaks convex-hull capacity limits while stabilizing training with joint objectives.
  - Co-led the project end-to-end, from problem discovery and theoretical formalization to scalable implementation and large-scale experimental validation.
  - Paper accepted by ICLR 2026.
- **Gene Sequence-to-Phenotype Framework with Structured Multi-Loss Design**
  May 2025 – Present
  - First to formulate gene knockout-induced phenotype abnormality prediction directly from full gene sequences, bridging the sequence–phenotype modality gap.
  - Built a comprehensive benchmarking suite across four datasets with strong baseline coverage and frequency-stratified evaluation.
  - Designed a unified training objective with three complementary losses: contrastive multi-label learning for correlation modeling, exclusivity regularization for logical consistency, and GO bottleneck supervision for mechanism-aware representation.
  - Delivered interpretable predictions with biologically grounded case studies, demonstrating the necessity and synergy of the three-loss design.
  - Paper accepted by AAAI 2026 (Oral)

- **Graph-Structured Representation Learning for Gene Functions**
  Mar 2023 – Jan 2025
  - Proposed the first novel function prediction setting that infers previously unannotated gene functions solely from known GO annotations, reframing gene function discovery as a structured representation learning problem.
  - Formally established this setting as a biologically meaningful and practically valuable task, demonstrating that function co-occurrence and ontology structure alone contain valuable predictive signal.
  - Designed a graph-informed Transformer with multi-view self-supervised objectives to learn ontology-consistent and compositional function representations.
  - Developed a GO-structure-aware masking strategy to prevent hierarchical shortcut recovery and enforce non-trivial representation learning.
  - Paper accepted by AAAI 2025.
- **UniEntrezDB: Large-Scale GO Annotation Benchmark**
  Aug 2023 – Dec 2024
  - Integrated Gene Ontology annotations from 21 databases.
  - Unified heterogeneous gene identifiers to NCBI Entrez IDs.
  - Built standardized benchmarks for scalable structured function prediction.

Publications

- Jingquan Yan*, **Yuwei Miao***, Peiran Yu, and Junzhou Huang, **"Breaking the Correlation Plateau: On the Optimization and Capacity Limits of Attention-Based Regressors"**, In Proc. of the Fourteenth International Conference on Learning Representations, Rio de Janeiro, Brazil, April 2026.
- **Yuwei, Miao**, and Junzhou Huang. **"Deep learning for toxicity prediction."** In Deep Learning in Drug Design, pp. 357-380. Academic Press, 2026.
- Yan, Jingquan*, **Yuwei Miao***, Lei Yu, Yuzhi Guo, Xue Xiao, Lin Xu, and Junzhou Huang. **"GenePheno: Interpretable Gene Knockout-Induced Phenotype Abnormality Prediction from Gene Sequences."** In Proceedings of the AAAI Conference on Artificial Intelligence, 2026 Oral.
- Zheng, Zheng, Yuzhi Guo, Xiao Hu, **Yuwei Miao**, Hehuan Ma, Jean Gao, and Junzhou Huang. **"Heterogeneous Aligned Fusion for Survival Prediction with Missing Modalities."** In Medical Imaging with Deep Learning. 2026.
- Dang, Thao M., Haiqing Li, Yuzhi Guo, Hehuan Ma, Feng Jiang, **Yuwei Miao**, Qifeng Zhou, Jean Gao, and Junzhou Huang. **"HAGE: Hierarchical Alignment Gene-Enhanced Pathology Representation Learning with Spatial Transcriptomics."** In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 228-238. Cham: Springer Nature Switzerland, 2025.
- **Yuwei, Miao**, Yuzhi Guo, Hehuan Ma, Jingquan Yan, Feng Jiang, Rui Liao, and Junzhou Huang. **"GoBERT: Gene Ontology Graph Informed BERT for Universal Gene Function Prediction."** In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 1, pp. 622-630. 2025.
- **Yuwei, Miao**, Yuzhi Guo, Hehuan Ma, Jingquan Yan, Feng Jiang, Weizhi An, Jean Gao, and Junzhou Huang. **"UniEntrezDB: Large-scale Gene Ontology Annotation Dataset and Evaluation Benchmarks with Unified Entrez Gene Identifiers."** arXiv preprint arXiv:2412.12688 (2024).
- **Yuwei, Miao**, Hehuan Ma, and Junzhou Huang. **"Recent advances in toxicity prediction: applications of deep graph learning."** Chemical Research in Toxicology 36, no. 8 (2023): 1206-1226.
- **Yuwei, Miao**, and Wenyi Luo. **"Improve generalization ability of cnn by data augmentation and se block in landmark classification."** In 2022 14th International Conference on Computer Research and Development (ICCRD), pp. 250-255. IEEE, 2022.

Awards & Scholarships

- Graduate Dean's Research Assistantships (2025)
- UNC Dean's list (2019)

Skills

Programming: Python, C++, Java, MATLAB, R
Tools: PyTorch, PyTorch Lightning, PyTorch Geometric (GNNs), Hugging Face Transformers, Linux, CUDA, Distributed Training (DDP), Git, AnnData, Scanpy, FlashAttention, Zarr
Languages: Chinese (Native), English (Fluent)