Yuzhi Guo

CONTACT INFORMATION

Email: yuzhi.guo@uta.edu **Homepage**: yuzhiguo07.github.io **Phone**: +1 2016681531

RESEARCH INTERESTS

Google Scholar Citations: **893**, h-index: 10

- **Large-Scale AI Model Training**
  - Self-supervised Pre-training Model Development
  - Generative AI models to enhance low-quality data
- **Deep Learning-based Drug Discovery**
- **Multimodal Large Language/Foundation Model Development**
- **Computer Vision/Medical Image Analysis**

EDUCATION

Ph.D. of Science in Computer Science, supervised by Dr. Junzhou Huang August 2022

University of Texas at Arlington, Texas, USA, GPA: 4.0

Master of Science in Computer Science May 2018

Stevens Institute of Technology, Hoboken, NJ, GPA: 3.7

Bachelor of Information and Computing Science June 2016

Beijing University of Technology, Beijing, China

EXPERIENCE

Post-doctoral Associate September 2022 – Present

University of Texas at Arlington, Arlington, TX, USA

- Develop a deep learning-based framework to predict the impact of protein missense mutations on protein functions and protein-protein interactions.
- Spearhead deep learning-based drug discovery projects with a focus on antibody-antigen interactions.
- Develop multimodal models for single-cell and spatial multi-omics data analysis.
- Contribute to grant writing and secure preliminary results through hands-on experiments.
- Mentor 10 Ph.D. students in research methodologies, data analysis, and experimental design.

AWARDS and HONORS

- UTA Rise-100 Innovative Scholars for Excellence (Post-doc) 2024
- Top-10 in Standard Industries Chemical Innovation Challenge (10,000 USD prize) 2024
- First place in the first phase of CoSolve Sprint challenge (3,000 USD prize) 2024

GRANTS

- NIH R01 grant of $3,136,243 for antigen-antibody interaction studies (a main contributor) 2025
- Johnson & Johnson award of $200,000 for LLM based toxicity prediction (a main contributor) 2024
- CPRIT grant of $1,199,997 for TCRs and Neoantigens binding prediction (a main contributor) 2023

COLLABORATORS

- UT Southwestern Medical Center (Collaborate to publish 3 papers, as well as two grants totaling $4,336,240.)
- Johnson & Johnson (Collaborate to publish 2 papers.)

PUBLICATIONS

Postdoctoral Mentorship and Publications

The following publications are the result of my primary mentorship of junior Ph.D. students, in the areas of **foundation models** and **AI for drug discovery**. In these projects, I was responsible for the initial project ideation, experimental design, and extensive manuscript writing and revision.

1. Saiyang Na, **Yuzhi Guo**, Feng Jiang, Hehuan Ma, Jean Gao and Junzhou Huang, "Segment Any Cell: A SAM-based Auto-prompting Fine-tuning Framework for Nuclei Segmentation", IEEE Transactions on Neural Networks and Learning Systems, August 2025. To Appear. (**IF: 14.255**)
2. Haiqing Li, **Yuzhi Guo**, Feng Jiang, Thao Dang, Hehuan Ma, Qifeng Zhou, Jean Gao and Junzhou Huang, "Text-Guided Multi-Instance Learning for Scoliosis Screening via Gait Video Analysis", In Proc. of the 28th Annual International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'25, Daejeon, Korea, September 2025.
3. Thao M. Dang, 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 Proc. of the 28th Annual International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'25, Daejeon, Korea, September 2025.
4. 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 Proc. of the Thirty-Ninth AAAI Conference on Artificial Intelligence, AAAI'25, Philadelphia, Pennsylvania, USA, February 2025.
5. Feng Jiang, **Yuzhi Guo**, Hehuan Ma, Saiyang Na, Weizhi An, Bing Song, Yi Han, Jean Gao, Tao Wang and Junzhou Huang, "AlphaEpi: Enhancing B Cell Epitope Prediction with AlphaFold 3", In Proc. of the 15th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM BCB'24, Shenzhen, China, November 2024.
6. Thao M. Dang, **Yuzhi Guo**, Hehuan Ma, Qifeng Zhou, Saiyang Na, Jean Gao and Junzhou Huang, "MFMF: Multiple Foundation Model Fusion Networks for Whole Slide Image Classification", In Proc. of the 15th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM BCB'24, Shenzhen, China, November 2024.
7. Feng Jiang, **Yuzhi Guo**, Hehuan Ma, Saiyang Na, Wenliang Zhong, Yi Han, Tao Wang and Junzhou Huang, "GTE: A Graph Learning Framework for Prediction of T-Cell Receptors and Epitopes Binding Specificity", Briefings in Bioinformatics, Volume 25, Issue 4, July 2024. (**IF: 8.7**)
8. Weizhi An, **Yuzhi Guo**, Yatao Bian, Hehuan Ma, Jinyu Yang, Chunyuan Li and Junzhou Huang, "Advancing DNA Language Models through Motif-Oriented Pre-training with MoDNA", Biomedinformatics, Volume 4, Number 2, pp.1556-1571, June 2024.

Doctoral Research Publications

My doctoral research focused on two main themes: the development of self-supervised pre-training algorithms and their applications, and deep learning-based drug discovery.

Self-supervised Pre-training Model Development

1. **Yuzhi Guo**, Jiaxiang Wu, Hehuan Ma and Junzhou Huang, "Self-supervised Pretraining for Protein Embeddings Using Tertiary Structures", In Proc. of the Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI'22, Vancouver, Canada, February 2022.
2. Weizhi An, **Yuzhi Guo**, Yatao Bian, Hehuan Ma, Jinyu Yang, Chunyuan Li and Junzhou Huang, "MoDNA: Motif-Oriented Pre-training for DNA Language Model", In Proc. of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, ACM BCB'22, Chicago, Illinois, USA, August 2022.
3. Sheng Wang, **Yuzhi Guo**, Yuhong Wang, Hongmao Sun and Junzhou Huang, "SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction", In Proc. of The 10th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM BCB'19, Niagara Falls, NY, USA, September 2019.

Generative AI models to enhance low-quality data

1. **Yuzhi Guo**, Jiaxiang Wu, Hehuan Ma, Sheng Wang and Junzhou Huang, " Comprehensive Study on Enhancing Low-Quality Position-Specific Scoring Matrix with Deep Learning for Accurate Protein Structure Property Prediction: Using Bagging Multiple Sequence Alignment Learning", Journal of Computational Biology,Volume 28, Number 4, April 2021.
2. **Yuzhi Guo**, Jiaxiang Wu, Hehuan Ma, Sheng Wang and Junzhou Huang, "EPTool: A New Enhancing PSSM Tool for Protein Secondary Structure Prediction", Journal of Computational Biology, Volume 28, Number 4, April 2021.
3. **Yuzhi Guo**, Jiaxiang Wu, Hehuan Ma, Sheng Wang and Junzhou Huang, "Bagging MSA Learning: Enhancing Low-quality PSSM with Deep Learning for Accurate Protein Structure Property Prediction", In Proc. of The 24th International Conference on Research in Computational Molecular Biology, RECOMB'20, Padova, Italy, May 2020.

Deep Learning-based Drug Discovery

1. **Yuzhi Guo**, Jiaxiang Wu, Hehuan Ma, Sheng Wang and Junzhou Huang, "Deep Ensemble Learning with Atrous Spatial Pyramid Networks for Protein Secondary Structure Prediction", Biomolecules, 12, 774, June 2022.
2. **Yuzhi Guo**, Jiaxiang Wu, Hehuan Ma, Jinyu Yang, Xinliang Zhu, and Junzhou Huang, "WeightAln: Weighted Homologous Alignment for Protein Structure Property Prediction", IEEE International Conference on Bioinformatics and Biomedicine, BIBM'20, Seout, Korea, December 2020.
3. **Yuzhi Guo**, Jiaxiang Wu, Hehuan Ma, Sheng Wang, and Junzhou Huang, "Protein Ensemble Learning with Atrous Spatial Pyramid Networks for Secondary Structure Prediction", IEEE International Conference on Bioinformatics and Biomedicine, BIBM'20, Seout, Korea, December 2020.
4. Hehuan Ma, Feng Jiang, **Yuzhi Guo** and Junzhou Huang, "Towards Robust Self-training for Molecular Biology Prediction Tasks, Journal of Computational Biology, Volume 31, Issue 3, pp. 213-228, March 2024.
5. Hehuan Ma, Feng Jiang, Yu Rong, **Yuzhi Guo** and Junzhou Huang, "Robust Self-training Strategy for Various Molecular Biology Prediction Tasks", In Proc. of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, ACM BCB'22, Chicago, Illinois, USA, August 2022.
6. Hehuan Ma, Yu Rong, Boyang Liu, **Yuzhi Guo**, Chaochao Yan, and Junzhou Huang, "Gradient-Norm Based Attentive Loss for Molecular Property Prediction", In Proc. of IEEE International Conference on Bioinformatics and Biomedicine, BIBM'21, December 2021
7. Hehuan Ma, Chaochao Yan, **Yuzhi Guo**, Sheng Wang, Yuhong Wang, Hongmao Sun and Junzhou Huang, "Improving Molecular Property Prediction on Limited Data with Deep Multi-Label Learning", IEEE BIBM Workshop on Machine Learning and Artificial Intelligence in Bioinformatics and Medical Informatics, MABM'20, Seout, Korea, December 2020.

Additional Collaborative Works

The following works represent broader scientific collaborations where my contributions included mentoring junior researchers and providing key insights, primarily in the fields of **Computer Vision**, **Medical Image Analysis**, and **Multimodal Large Language Models**.

1. Feng Jiang, Mangal Prakash, Hehuan Ma, Jianyuan Deng, **Yuzhi Guo**, Amina Mollaysa, Tommaso Mansi, Rui Liao and Junzhou Huang, "TRIDENT: Tri-Modal Molecular Representation Learning with Taxonomic Annotations and Local Correspondence", In Proc. of the 39th Annual Conference on Neural Information Processing Systems, NeurIPS'25, San Diego, CA, USA, December 2025.
2. Feng Jiang, Mangal Prakash, Hehuan Ma, Jianyuan Deng, **Yuzhi Guo**, Amina Mollaysa, Tommaso Mansi, Rui Liao, Junzhou Huang, "TRIDENT: Tri-Modal Molecular Representation Learning with Taxonomic Annotations and Local Correspondence", ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada, July 2025

3. Bing Song, Kaiwen Wang, Saiyang Na, Jia Yao, Farjana J. Fattah, Alexandra L Martin, Mitchell S. von Itzstein, Donghan M. Yang, Jialiang Liu, Yaming Xue, Chaoying Liang, **Yuzhi Guo**, Indu Raman, Chengsong Zhu, Jonathan E. Dowell, Jade Homsi, Sawsan Rashdan, Shengjie Yang, Mary E. Gwin, Tuoqi Wu, David Hsiehchen, Yvonne Gloria-McCutchen, Catherine Pei-ju Lu, Prithvi Raj, Xiaochen Bai, Jun Wang, Jose Conejo-Garcia, Yang Xie, Junzhou Huang*, David E. Gerber*, Tao Wang*, "Profiling Antigen-Binding Affinity of B Cell Repertoires in Tumors by Deep Learning Predicts Immune-Checkpoint Inhibitor Treatment Outcomes", Nature Cancer, June 2025. (**IF: 28.5**)
4. Qifeng Zhou, Thao M. Dang, **Yuzhi Guo**, Hehuan Ma, Wenliang Zhong, Saiyang Na, Jean Gao and Junzhou Huang, "Contrastive Pretraining for Computational Pathology With Visual Language Models", In Proc. of IEEE International Symposium on Biomedical Imaging, ISBI'25, Houston, Texas, USA, April 2025.
5. Thao M. Dang, Qifeng Zhou, **Yuzhi Guo**, Hehuan Ma, Saiyang Na, Thao Bich Dang, Jean Gao and Junzhou Huang, "Abnormality-Aware Multimodal Learning for WSI Classification", Frontiers in Medicine, Volume 12, February 2025.
6. Wenliang Zhong, Rob Barton, Weizhi An, Feng Jiang, Hehuan Ma, **Yuzhi Guo**, Abhishek Dan, Shioulin Sam, Karim Bouyarmane and Junzhou Huang, "Zero-Shot Composed Image Retrieval via Dual-Stream Instruction-Aware Distillation", In Proc. of International Conference on Computer Vision, ICCV'25, Honolulu, Hawaii, October 2025.
7. Qifeng Zhou, Wenliang Zhong, **Yuzhi Guo**, Michael Xiao, Hehuan Ma and Junzhou Huang, "PathM3: A Multimodal Multi-Task Multiple Instance Learning Framework for Whole Slide Image Classification and Captioning", In Proc. of the 27th Annual International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'24, Marrakesh, Morocco, October 2024.
8. Wenliang Zhong, Hehuan Ma, Yu Rong, Yatao Bian, Long-Kai Huang, **Yuzhi Guo**, Peilin Zhao and Junzhou Huang, "CoSSL: A Context-based Semi-Supervised Framework for Molecular Property Prediction", ICML Workshop on Computational Biology, Honolulu, Hawaii, USA, July 2023.
9. Jinyu Yang, Chunyuan Li, Weizhi, Hehuan Ma, **Yuzhi Guo**, Yu Rong, Peilin Zhao, Junzhou Huang, "Exploring Robustness of Unsupervised Domain Adaptation in Semantic Segmentation", In Proc. of International Conference on Computer Vision, ICCV'21, October 2021.

Patent and Book Chapter

1. Jiaxiang Wu, **Yuzhi Guo**, and Junzhou Huang. "Protein structure information prediction method and apparatus, device, and storage medium." U.S. Patent 12,288,599, issued April 29, 2025.
2. **Yuzhi Guo** and Junzhou Huang, "Deep learning for protein secondary structure prediction", In Deep Learning in Drug Design: Methods and Applications, Chapter 13, Academic Press, October 2025.

TEACHING AND MENTORING EXPERIENCE

TA Experience

- UTA, CSE5311, Design and Analysis of Algorithms — Spring 2021
- UTA, CSE5311, Design and Analysis of Algorithms — Fall 2020
- UTA, CSE5311, Design and Analysis of Algorithms — Spring 2020
- UTA, CSE5311, Design and Analysis of Algorithms — Fall 2019
- UTA, CSE5311, Design and Analysis of Algorithms — Spring 2019
- UTA, CSE4314, Professional Practices — Fall 2018

Ph.D. Student and High School Student Intern Mentoring

- Mentored 10 junior Ph.D students, leading to the co-authorship of 17 publications
- Supervised a high school intern, resulting in one co-authored paper.

ACADEMIC SERVICES

Academic Symposium/Fair Volunteer

- The volunteer of the North Texas Symposium on Generative AI (GenAI-25) 03/2025
- The judge of the Fort Worth Regional Science and Engineering Fair (FWRSEF-25) 02/2025

Conference Reviewer

- The 40th AAAI Conference on Artificial Intelligence (AAAI-26) 2025
- The 39th Annual Conference on Neural Information Processing Systems (NeurIPS-25) 2025
- The International Conference on Computer Vision 2025 (ICCV-25) 2025
- The 42nd International Conference on Machine Learning (ICML-25) 2025
- The 42nd The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-25) 2024
- The 13th International Conference on Learning Representations (ICLR-25) 2024
- The 39th AAAI Conference on Artificial Intelligence (AAAI-25) 2024
- The 27th International Conference on Pattern Recognition (ICPR) 2024
- The 38th Annual Conference on Neural Information Processing Systems (NeurIPS-24) 2024
- The 33rd ACM International Conference on Information and Knowledge Management (CIKM-24) 2024
- The 41st International Conference on Machine Learning (ICML-24) 2024
- The 33rd International Joint Conference on Artificial Intelligence (IJCAI-24) 2024
- The 38th AAAI Conference on Artificial Intelligence (AAAI-24) 2023
- The 37th Annual Conference on Neural Information Processing Systems (NeurIPS-23) 2023
- The 29th SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-23) 2023
- The 40th International Conference on Machine Learning (ICML-23) 2023
- The 32nd International Joint Conference on Artificial Intelligence (IJCAI-23) 2023
- The 37th AAAI Conference on Artificial Intelligence (AAAI-23) 2022
- The 36th Annual Conference on Neural Information Processing Systems (NeurIPS-22) 2022
- The 1st Learning on Graphs Conference 2022
- The 39th International Conference on Machine Learning (ICML-22) 2022
- The 31st ACM International Conference on Information and Knowledge Management (CIKM-22) 2022

Journal Reviewer

- Scientific Reports
- Journal of Cheminformatics
- Cognitive Computation
- BMC Bioinformatics
- InderScience
- Journal of Visual Communication and Image Representation

INVITED TALKS

- Guest lecture about “Pre-training Models in Drug Discovery” (UTA CSE 6363) 2025
- Guest lecture about “Foundation Model in Medical Image Processing” (UTA CSE 6363) 2025
- Invited talk about “The Impact of Missense Mutations on Protein-Protein Interaction Prediction” at UT Austin 2024
- Invited talk about “Deep Self-supervised Pre-training for Hierarchical Protein Embeddings” at UT Southwestern 2023

PROFESSIONAL REFERENCES

Dr. Junzhou Huang
Jenkins Garrett Professor, FAIMBE
Department of Computer Science and Engineering,
The University of Texas at Arlington,
Arlington, TX 76019, USA.
☎ +1-817-272-9596
✉ send.Huang.1D0EBCA0A1@interfoliodossier.com

Dr. Hong Jiang
Wendell H. Nedderman Endowed Professor and Chair
Department of Computer Science and Engineering,
The University of Texas at Arlington,
Arlington, TX 76019, USA.
☎ +1-817-272-3605
✉ send.Jiang.A8C698566A@interfoliodossier.com

Dr. Tao Wang
Associate Professor
Department of Bioinformatics and Computational Biology,
The University of Texas MD Anderson Cancer Center,
Houston, TX 77030, USA.
☎ +1-713-689-4310
✉ send.Wang.4B0DB2C4C7@interfoliodossier.com

Engineering Research Building.
500 UTA Blvd., Arlington, TX 76010
Homepage: https://yuzhiguo07.github.io/
yuzhi.guo@uta.edu

Yuzhi Guo
Post-Doctoral Associate
Computer Science and Engineering Department
University of Texas at Arlington

Dear Faculty Search Committee,

I am writing to apply for the tenure-track Assistant Professor position in Computer Science at Texas A&M University-Corpus Christi, with a focus on **Artificial Intelligence and Machine Learning**. I received my Ph.D. in Computer Science from the University of Texas at Arlington in 2022 under the supervision of Dr. Junzhou Huang, and I am currently a Postdoctoral Associate in his group. My research centers on machine learning, large-scale model development, and interdisciplinary applications in biomedical science, and I am eager to contribute to the department's growth in AI and data-driven research.

My research program advances computer science in two complementary directions. First, I develop **large-scale pretraining** and **generative models**, including novel self-supervised models that leverage molecular, genomic, and imaging data, and generative AI models to enhance low-quality data. Second, I pursue **AI for Science**, where I apply these methods to impactful problems such as drug discovery, protein-antibody interactions, multi-omics integration, and medical image analysis. This work has resulted in publications in leading venues including *NeurIPS, AAAI, RECOMB, MICCAI, ICCV, TNNLS,* and *Nature Cancer*. My research accomplishments have been recognized with the **UTA RISE-100 Innovative Scholars for Excellence Award**, which highlights postdoctoral researchers with exceptional potential for academic leadership. These efforts establish a coherent research trajectory that combines methodological innovation with transformative scientific applications.

I have also demonstrated strong potential in securing competitive research funding. As a key contributor, I contributed to an **NIH R01 ($3.1M)**, a **CPRIT award ($1.2M)**, and a **Johnson & Johnson award ($200K)**, totaling over $4.4M. These experiences highlight my ability to drive high-impact interdisciplinary projects and build a strong foundation for an independent funding program.

Equally important is my commitment to teaching and mentoring. At UT Arlington, I served as a graduate teaching assistant (GTA) for multiple semesters in **both graduate and undergraduate courses** such as Design and Analysis of Algorithms and Professional Practices. I have also been invited several times as a **guest lecturer for the Machine Learning course**, where I delivered advanced topics in deep learning and foundation model design. My strong background in algorithms, AI, and machine learning prepares me to **teach a wide range of undergraduate and graduate courses in computer science**. In addition, I have **mentored ten Ph.D. students and a high school intern**, guiding them toward successful publications in top venues and instilling in them rigorous research skills. These experiences reflect both my dedication to student success and my ability to integrate research and teaching.

I am drawn to TAMU–Corpus Christi's mission as an emerging research institution with a strong emphasis on student-centered learning and interdisciplinary collaboration. I envision establishing a research program that not only **advances machine learning methodology** but also **engages students directly in research, broadening participation in AI** and **empowering them with skills that are relevant to both academia and industry.**

Thank you for your time to read my application and your kind consideration. I am looking forward to hearing from you.

Sincerely yours,

Yuzhi Guo

Yuzhi Guo

My commitment to a career in academia is driven by a profound passion for teaching and mentorship. I view education as a powerful catalyst, empowering students with the knowledge to pursue their aspirations and bridging gaps created by diverse backgrounds. This passion has been a cornerstone of my academic journey, from my time as a teaching assistant to my role as a postdoctoral mentor, and it is a privilege I am eager to carry forward.

Teaching Experience

During my Ph.D. and postdoctoral journey, I honed my teaching skills through various roles, including serving as a teaching assistant, delivering guest lectures on advanced topics, and giving invited talks at other institutions. Over the course of several years, I served as a Graduate Teaching Assistant (GTA) at the University of Texas at Arlington, where I contributed to both graduate and undergraduate courses. These included the graduate-level core course "Design and Analysis of Algorithms" and the undergraduate course "Professional Practices." These experiences have helped me cultivate a teaching philosophy centered on three core principles:

Connecting Theory to Real-World Impact. I firmly believe that student engagement soars when abstract concepts are grounded in tangible, real-world applications. In my guest lectures on foundation models, for example, I begin not with the complex mathematics and deep learning algorithms, but with a practical demonstration of how models like "Segment Any Cell" are revolutionizing how we analyze medical images. By framing complex topics around solving real problems—from accelerating drug discovery to improving disease diagnosis—I aim to ignite students' curiosity and motivate them to master the underlying principles. This approach helps demystify what can be intimidating subject matter, making it more accessible to students from all backgrounds. By showcasing the direct line from a theoretical algorithm to a tangible scientific breakthrough, I also help students envision their own potential career paths and contributions to the field.

Fostering Critical and Independent Thinking. My goal as an educator is not simply to transmit information, but to cultivate the next generation of independent researchers and problem-solvers. In my role as a TA, I designed assignments that encouraged students to go beyond rote memorization and apply algorithmic concepts to novel problems. When mentoring students on research projects, I make it a practice to challenge their assumptions and ask probing questions about their design choices. This approach encourages them to think critically about the "why" behind their methods, leading to deeper understanding and more innovative solutions.

Providing Individualized and Supportive Mentorship. Recognizing that students come from diverse academic and personal backgrounds, I am committed to creating an inclusive and supportive learning environment. As a postdoctoral associate, I have had the privilege of mentoring 14 Ph.D. students. This experience has taught me the importance of tailoring my mentorship style to each individual's needs, whether it's providing detailed guidance on experimental design, suggesting relevant literature, or helping to scope a project's contributions. I strive to be an approachable and patient resource, offering comprehensive feedback that not only corrects errors but also empowers students to find the right answers themselves.

Mentoring and Diversity

I have substantial experience in mentoring junior Ph.D. students by identifying their individual strengths and weaknesses to help them succeed. For example, when I mentored a junior PhD student on an antigen-antibody interaction project, I recognized his strength in literature review but also his weakness in hands-on implementation. I guided him step-by-step through the data processing pipeline, explained the practical implementation of several classic AI papers on proteins and antibodies, and recommended further literature for him to reproduce. Subsequently, I helped him establish clear project expectations, contributions, and scope. Drawing from my own broad collaborations, I helped him identify potential

collaborators who could significantly enhance his work. Through his dedicated efforts, this project culminated in three papers accepted by a prestigious journal, a top-tier bioinformatics conference, and a premier machine learning conference, respectively. In another instance, I mentored a doctoral student applying foundation models to cell segmentation. While he was an exceptionally strong coder, he tended to focus on minor implementation details at the expense of the project's higher-level research goals. I actively challenged his design choices and encouraged him to reconsider the overall framework. Drawing on my deep learning expertise, I guided him in exploring and implementing multiple high-level ideas. This shift in perspective was pivotal, and his work was ultimately accepted by a high-impact machine learning journal.

These examples reflect the core of my mentoring philosophy: I believe that successful mentorship requires a deep understanding of each student's unique strengths and challenges, and tailoring my approach accordingly. My goal is not only to help them solve technical problems but also to cultivate their high-level research thinking and independent problem-solving skills, guiding them from focusing on details to seeing the big picture, and ultimately achieving success in their respective fields.

Promoting Diversity through Teaching and Mentoring

A central tenet of my educational philosophy is the promotion of diversity and inclusivity. I am dedicated to creating a classroom and lab environment where every student feels valued, respected, and heard. In practice, this means actively incorporating course materials and research examples from a diverse range of scientists to show students that great ideas come from all backgrounds. By encouraging open discussion, highlighting contributions from researchers of all backgrounds, and being mindful of different learning styles, I aim to foster an atmosphere where diverse perspectives can flourish and drive innovation. In my mentoring, I am committed to creating an inclusive and supportive atmosphere for junior researchers, ensuring they have the resources and encouragement to thrive. I believe that fostering diversity is not just a moral imperative but also a catalyst for scientific excellence; diverse teams are more creative, ask better questions, and ultimately produce more robust and innovative research. My goal is to help make the fields of computer science and bioinformatics more accessible and welcoming to individuals from all walks of life. This commitment is an integral part of my educational philosophy, one that I will eagerly carry into my future endeavors..

Example Future Courses

My interdisciplinary background in computer science, deep learning, and bioinformatics equips me to teach a wide range of courses at both the undergraduate and graduate levels.

- Graduate Courses: I am prepared to teach advanced courses such as Machine Learning, Deep Learning, and specialized topics like "AI for Science" or "Large Language Models and Foundation Models"
- Undergraduate Courses: I am also enthusiastic about teaching fundamental computer science courses, including Data Structures and Algorithms, Introduction to Programming (Python) , Professional Practices in Computing, and a hands-on course like Applied Deep Learning.

Given the rapid advancements in the field, I am particularly excited about the prospect of developing and launching a new, cutting-edge course on "Generative AI for Biology." This course would explore how the latest generative models are being used to design novel proteins, discover new drugs, and generate synthetic biological data, equipping students with the skills to contribute to this exciting and fast-moving area of research.

Summary

My research interests primarily focus on two crucial topics: 1) **Large-Scale Artificial Intelligence (AI) Model Training**: Designing novel self-supervised pre-training models, and fine-tuning frameworks for biomedical foundation models. 2) **AI-Powered Drug Discovery**: Developing cutting-edge AI technologies to accelerate the discovery of new therapeutics and deepen our understanding of protein analysis.

1. Large-Scale AI Model Training

My research concentrates on developing large-scale AI models that learn powerful representations from complex biological data. This work addresses key challenges across the model development pipeline, from pioneering **self-supervised pre-training objectives on vast unlabeled data** and **enhancing low-quality data**, to designing **novel fine-tuning strategies that adapt foundation models for specific, high-impact biomedical tasks**.

1.1 Development of Novel Self-Supervised Pre-training Models

The power of deep learning in biology is unlocked by models that can learn the fundamental "language" of molecules, proteins, and genes from unlabeled data. To this end, my work has pioneered pre-training frameworks across multiple biological modalities. In chemistry, traditional property prediction relied on hand-crafted features or fixed molecular fingerprints, which often fail to capture the full complexity of chemical structures. Recognizing this limitation, my work on SMILES-BERT [1] was one of the first to successfully adapt the transformer architecture to learn the "language" of chemistry. My key contribution was to reformulate BERT's masked language model objective for SMILES strings, enabling the model to learn the underlying rules of chemical syntax and semantics directly from a massive corpus of unlabeled molecules, creating powerful and transferable representations.

I extended this concept beyond 1D sequences to the critical domain of protein structure. While sequence-based models were powerful, they ignored the 3D structural information that ultimately dictates biological function. To address this gap, I developed a novel self-supervised pre-training strategy to learn protein embeddings directly from their tertiary structures [2]. The core of my framework is a denoising score matching algorithm. The model is trained by systematically corrupting the 3D atomic coordinates of known protein structures with noise and then tasked with predicting the original, uncorrupted structure. By learning to effectively "denoise" these structures, my model is forced to implicitly learn the fundamental principles of protein folding—such as valid bond lengths, angles, and biophysically plausible conformations—resulting in rich embeddings that capture the complex spatial relationships invisible to sequence-only models. The foundational nature of this work and its powerful preliminary results were instrumental in securing a recent NIH R01 grant, demonstrating its direct impact on enabling future research.

My research has further advanced pre-training by injecting critical domain knowledge into the learning process. Standard sequence models treat all parts of a sequence equally, yet in genomics, specific motifs hold immense functional importance. In MoDNA [3], my contribution was to design a motif-oriented pre-training framework that explicitly teaches the model the grammar of DNA by rewarding the identification of conserved regulatory motifs. Similarly, while gene models often focused on sequence alone, genes function within complex biological networks. My work on GoBERT [4] tackled this by being the first model pre-trained directly on gene function. I designed a novel architecture that integrates the Gene Ontology, a massive graph of known gene functions, into the pre-training process. This forces the model to learn representations that respect the known biological context, allowing it to accurately predict functions for poorly understood genes.

1.2 Enhancing Low-Quality Data with Deep Learning Models

A significant real-world challenge in bioinformatics is the prevalence of noisy and incomplete data. For instance, Position-Specific Scoring Matrices (PSSMs), which are vital for many protein prediction tasks, are derived from Multiple Sequence Alignments (MSAs). However, their quality degrades significantly when few homologous sequences are available, leading to a loss of crucial evolutionary information. My research directly confronts this by using deep learning to enhance data quality. My BaggingMSA [5] method introduces a novel training paradigm to solve this problem. The core idea is to treat this as a data restoration task; I synthetically create low-quality training data by down-sampling high-quality MSAs, and then train a deep network to generate high-quality counterparts from these low-quality inputs. By learning this transformation, the model effectively learns to "in-paint" the missing evolutionary information. I further extended this work into a comprehensive semi-supervised framework [6], which demonstrated the approach's broad utility and enhanced performance on a variety of downstream tasks. This research culminated in the development of EPTool [7], a practical and highly accurate tool for protein secondary structure prediction, demonstrating how sophisticated modeling can overcome fundamental data limitations and produce tangible tools for the scientific community.

Complementary to enriching sparse homologous data, I also tackled the challenge of refining information from abundant homologous sequences. While having many homologs is beneficial, it often introduces significant redundancy and noise, where not all sequences contribute equally to the final prediction. To address this, my WeightAln [8] method introduces a learning-based framework to assign weights to each homologous sequence. By training a model to predict these weights, my approach learns to identify and emphasize the most informative sequences while down-weighting redundant or divergent ones, leading to a more refined and accurate representation of the protein's evolutionary profile.

1.3 Fine-Tuning and Adaptation of Large Foundation Models

The recent proliferation of powerful foundation models like the Segment Anything Model (SAM) has created immense opportunities, but their application to specialized scientific domains remains a significant hurdle. These generalist models lack the domain-specific precision required for complex tasks like medical image analysis. My research creates novel strategies to bridge this gap. For instance, SAM's performance on nuclei segmentation is highly dependent on precise, manually-provided prompts. My key contribution in "Segment Any Cell" [9] was to develop a novel auto-prompting fine-tuning framework. Instead of relying on manual intervention, my method trains a lightweight module that learns to automatically generate the optimal prompts for the foundation model, successfully adapting its powerful capabilities for a highly specific biomedical task. Furthermore, recognizing that no single foundation model excels at all tasks, my work on MFMF [10] addresses the challenge of multiple foundation model fusion. I designed a fusion network that combines the feature representations from multiple different foundation models to achieve a final prediction that is more robust and accurate than any single model could achieve on its own. Beyond these specific frameworks, my research has broadly explored the application and adaptation of large visual-language and multimodal models to a diverse set of biomedical challenges. This includes developing methods for contrastive pre-training in computational pathology [11], creating abnormality-aware models for whole slide image classification [12], advancing zero-shot medical image retrieval [13], building multimodal frameworks for tasks ranging from scoliosis screening from gait videos [14], integrating spatial transcriptomics with pathology images [15, 16], and exploring the robustness of unsupervised domain adaptation for semantic segmentation [17].

2. AI-Powered Drug Discovery

The second pillar of my research program applies advanced AI models to address pressing needs in therapeutic development. My foundational work in this area established more robust methods for fundamental protein feature

prediction. In particular, I developed a Deep Ensemble Learning framework using Atrous Spatial Pyramid Networks [18, 19]. This architecture excels at capturing multi-scale contextual information within protein sequences, significantly improving the accuracy of secondary structure prediction by allowing the model to see both local residue interactions and broader structural motifs simultaneously. The novelty of these approaches for protein structure information prediction is further underscored by the issuance of a U.S. Patent [20], and the foundational work on advancing DNA language models was further applied to promoter prediction and transcription factor binding site prediction [21].

Building on this, my research has tackled central challenges in immunology and molecular property prediction. Predicting the binding between immune receptors and their corresponding epitopes is a notoriously difficult task critical for vaccine and therapy design. To address this, I developed GTE [22], a graph learning framework that models the TCR-epitope interaction as a complex graph of inter-atomic relationships, capturing nuanced biochemical properties that simpler models miss. More recently, my AlphaEpi [23] framework was one of the first methods to fully leverage the unprecedented structural accuracy of AlphaFold 3. My contribution was to design a deep learning architecture specifically tailored to ingest these high-fidelity 3D structures, leading to a new state-of-the-art in B-cell epitope prediction. The translational potential of this line of work is also demonstrated in our recent study published in Nature Cancer [24], where I developed and applied the core deep learning models that successfully predicted patient outcomes to immune-checkpoint inhibitors, forging a direct link between AI-driven molecular analysis and personalized immunotherapy. In addition, I have extensively researched robust methods for molecular property prediction, developing novel strategies such as robust self-training [25, 26], gradient-norm based attentive losses [27], semi-supervised frameworks [28], deep multi-label learning for limited-data scenarios [29], and TRIDENT framework integrates taxonomic annotations for richer embeddings [30]. The collective impact and translational potential of this research were pivotal in securing major funding from both CPRIT and Johnson & Johnson, enabling the continuation and expansion of these drug discovery efforts.

Future Research

My long-term research goal is to develop the next generation of AI systems that can reason across the full spectrum of biological data, from molecules to medicine, to accelerate scientific discovery and create personalized therapeutics. To achieve this, I will focus on two primary research directions in the coming years:

Multimodal Foundation Models for Integrated Biology (Years 1-5). The next frontier in biomedical AI is the integration of diverse data modalities. While my past work has established powerful representations for individual data types—molecules, proteins, genes, and medical images—true biological understanding requires models that can see the complete picture. To solve this, my primary focus in the upcoming years will be on developing unified multimodal foundation models. Building on my experience with large-scale model pre-training, fusing foundation models and aligning different data types, I will design novel architectures capable of jointly embedding and reasoning across genomics, transcriptomics, proteomics, and pathology imaging. The goal is to create models that can answer complex, cross-modal questions, such as predicting how a specific gene mutation (genomics) might alter protein structure (proteomics) and ultimately manifest in cellular morphology (imaging), enabling a deeper, systems-level understanding of disease.

Generative AI for De Novo Therapeutic Design (Years 3-10). This topic will pivot from predictive to generative modeling to design novel therapeutics from scratch. My previous work has focused on predicting molecular properties and interactions (e.g., GTE, AlphaEpi). The logical next step is to invert this problem: can we generate novel molecules and biologics with desired properties? To this end, I will develop structure- and function-conditioned generative models. Leveraging my expertise in 3D protein structure embedding, I will create generative frameworks that can design novel drug candidates or antibodies conditioned on the specific 3D geometry of a target binding site. Furthermore, by integrating

functional knowledge from frameworks like GoBERT, these models will be guided to produce therapeutics that modulate specific biological pathways, moving beyond simple binding affinity to designing for functional outcomes.

Funding Plan. I have gained significant grant writing experience as a postdoctoral associate. I was a main contributor to a successful NIH R01 grant ($3,136,243) and a CPRIT grant ($1,199,997), as well as a Johnson & Johnson award ($200,000). In the future, I plan to secure external funding to support this research agenda by targeting programs at the NIH (e.g., NIGMS for foundational model development, NCI and NIAID for therapeutic applications) and the NSF (e.g., CAREER awards and programs at the intersection of computing and life sciences).

Selected Publications

Development of Novel Self-Supervised Pre-training Models

[1] Wang, S., **Guo, Y.**, Wang, Y., Sun, H. and Huang, J., "SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction", In Proc. of The 10th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM BCB'19.

[2] **Guo, Y.**, Wu, J., Ma, H. and Huang, J., "Self-supervised Pretraining for Protein Embeddings Using Tertiary Structures", In Proc. of the Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI'22.

[3] An, W., **Guo, Y.**, Bian, Y., Ma, H., Yang, J., Li, C. and Huang, J., "MoDNA: Motif-Oriented Pre-training for DNA Language Model", In Proc. of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, ACM BCB'22.

[4] Miao, Y., **Guo, Y.**, Ma, H., Yan, J., Jiang, F., Liao, R. and Huang, J., "GoBERT: Gene Ontology Graph Informed BERT for Universal Gene Function Prediction", In Proc. of the Thirty-Ninth AAAI Conference on Artificial Intelligence, AAAI'25.

Enhancing Low-Quality Data with Deep Learning Models

[5] **Guo, Y.**, Wu, J., Ma, H., Wang, S. and Huang, J., "Bagging MSA Learning: Enhancing Low-quality PSSM with Deep Learning for Accurate Protein Structure Property Prediction", In Proc. of The 24th International Conference on Research in Computational Molecular Biology, RECOMB'20.

[6] **Guo, Y.**, Wu, J., Ma, H., Wang, S. and Huang, J., "Comprehensive Study of Bagging MSA Learning for Protein Structure Property Prediction", Journal of Computational Biology, Volume 28, Number 4, April 2021.

[7] **Guo, Y.**, Wu, J., Ma, H., Wang, S. and Huang, J., "EPTool: A New Enhancing PSSM Tool for Protein Secondary Structure Prediction", Journal of Computational Biology, Volume 28, Number 4, April 2021.

[8] **Guo, Y.**, Wu, J., Ma, H., Yang, J., Zhu, X., and Huang, J., "WeightAln: Weighted Homologous Alignment for Protein Structure Property Prediction", IEEE International Conference on Bioinformatics and Biomedicine, BIBM'20.

Fine-Tuning and Adaptation of Large Foundation Models

[9] Na, S., **Guo, Y.**, Jiang, F., Ma, H., Gao, J. and Huang, J., "Segment Any Cell: A SAM-based Auto-prompting Fine-tuning Framework for Nuclei Segmentation", IEEE Transactions on Neural Networks and Learning Systems, August 2025.

[10] Dang, T. M., **Guo, Y.**, Ma, H., Zhou, Q., Na, S., Gao, J. and Huang, J., "MFMF: Multiple Foundation Model Fusion Networks for Whole Slide Image Classification", In Proc. of the 15th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM BCB'24.

[11] Zhou, Q., Dang, T. M., **Guo, Y.**, Ma, H., Zhong, W., Na, S., Gao, J. and Huang, J., "Contrastive Pretraining for Computational Pathology With Visual Language Models", In Proc. of IEEE International Symposium on Biomedical Imaging, ISBI'25.

[12] Dang, T. M., Zhou, Q., **Guo, Y.**, Ma, H., Na, S., Dang, T. B., Gao, J. and Huang, J., "Abnormality-Aware Multimodal Learning for WSI Classification", Frontiers in Medicine, Volume 12, February 2025.

[13] Zhong, W., Barton, R., An, W., Jiang, F., Ma, H., **Guo, Y.**, Dan, A., Sam, S., Bouyarmane, K. and Huang, J., "Zero-Shot Composed Image Retrieval via Dual-Stream Instruction-Aware Distillation", In Proc. of International Conference on Computer Vision, ICCV'25.

[14] Li, H., **Guo, Y.**, Jiang, F., Dang, T., Ma, H., Zhou, Q., Gao, J. and Huang, J., "Text-Guided Multi-Instance Learning for Scoliosis Screening via Gait Video Analysis", In Proc. of the 28th Annual International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'25.

[15] Dang, T., Li, H., **Guo, Y.**, Ma, H., Jiang, F., Miao, Y., Zhou, Q., Gao, J. and Huang, J., "HAGE: Hierarchical Alignment Gene-Enhanced Pathology Representation Learning with Spatial Transcriptomics", In Proc. of the 28th Annual International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'25.

[16] Zhou, Q., Zhong, W., **Guo, Y.**, Xiao, M., Ma, H. and Huang, J., "PathM3: A Multimodal Multi-Task Multiple Instance Learning Framework for Whole Slide Image Classification and Captioning", In Proc. of the 27th Annual International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'24.

[17] Yang, J., Li, C., An, W., Ma, H., **Guo, Y.**, Rong, Y., Zhao, P., Huang, J., "Exploring Robustness of Unsupervised Domain Adaptation in Semantic Segmentation", In Proc. of International Conference on Computer Vision, ICCV'21

AI-Powered Drug Discovery

[18] **Guo, Y.**, Wu, J., Ma, H., Wang, S. and Huang, J., "Deep Ensemble Learning with Atrous Spatial Pyramid Networks for Protein Secondary Structure Prediction", Biomolecules, 12, 774, June 2022.

[19] **Guo, Y.**, Wu, J., Ma, H., Wang, S., and Huang, J., "Protein Ensemble Learning with Atrous Spatial Pyramid Networks for Secondary Structure Prediction", IEEE International Conference on Bioinformatics and Biomedicine, BIBM'20.

[20] Wu, J., **Guo, Y.**, and Huang, J. "Protein structure information prediction method and apparatus, device, and storage medium." U.S. Patent 12,288,599, issued April 29, 2025.

[21] An, W., **Guo, Y.**, Bian, Y., Ma, H., Yang, J., Li, C. and Huang, J., "Advancing DNA Language Models through Motif-Oriented Pre-training with MoDNA", Biomedinformatics, Volume 4, Number 2, pp.1556-1571, June 2024.

[22] Jiang, F., **Guo, Y.**, Ma, H., Na, S., Zhong, W., Han, Y., Wang, T. and Huang, J., "GTE: A Graph Learning Framework for Prediction of T-Cell Receptors and Epitopes Binding Specificity", Briefings in Bioinformatics, Volume 25, Issue 4, July 2024.

[23] Jiang, F., **Guo, Y.**, Ma, H., Na, S., An, W., Song, B., Han, Y., Gao, J., Wang, T. and Huang, J., "AlphaEpi: Enhancing B Cell Epitope Prediction with AlphaFold 3", In Proc. of the 15th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM BCB'24.

[24] Song, B., Wang, K., Na, S., Yao, J., Fattah, F. J., Martin, A. L., von Itzstein, M. S., Yang, D. M., Liu, J., Xue, Y., Liang, C., **Guo, Y.**, Raman, I., Zhu, C., Dowell, J. E., Homsi, J., Rashdan, S., Yang, S., Gwin, M. E., Wu, T., Hsiehchen, D., Gloria-McCutchen, Y., Lu, C. P., Raj, P., Bai, X., Wang, J., Conejo-Garcia, J., Xie, Y., Huang, J., Gerber, D. E., Wang, T.*, "Profiling Antigen-Binding Affinity of B Cell Repertoires in Tumors by Deep Learning Predicts Immune-Checkpoint Inhibitor Treatment Outcomes", Nature Cancer, June 2025.

[25] Ma, H., Jiang, F., **Guo, Y.** and Huang, J., "Towards Robust Self-training for Molecular Biology Prediction Tasks", Journal of Computational Biology, Volume 31, Issue 3, pp. 213-228, March 2024.

[26] Ma, H., Jiang, F., Rong, Y., **Guo, Y**. and Huang, J., "Robust Self-training Strategy for Various Molecular Biology Prediction Tasks", In Proc. of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, ACM BCB'22.

[27] Ma, H., Rong, Y., Liu, B., **Guo, Y.**, Yan, C., and Huang, J., "Gradient-Norm Based Attentive Loss for Molecular Property Prediction", In Proc. of IEEE International Conference on Bioinformatics and Biomedicine, BIBM'21.

[28] Zhong, W., Ma, H., Rong, Y., Bian, Y., Huang, L., **Guo, Y.**, Zhao, P. and Huang, J., "CoSSL: A Context-based Semi-Supervised Framework for Molecular Property Prediction", ICML Workshop on Computational Biology, July 2023.

[29] Ma, H., Yan, C., **Guo, Y.**, Wang, S., Wang, Y., Sun, H. and Huang, J., "Improving Molecular Property Prediction on Limited Data with Deep Multi-Label Learning", IEEE BIBM Workshop on Machine Learning and Artificial Intelligence in Bioinformatics and Medical Informatics, MABM'20.

[30] Jiang, F., Prakash, M., Ma, H., Deng, J., **Guo, Y.**, Mollaysa, A., Mansi, T., Liao, R. and Huang, J., "TRIDENT: Tri-Modal Molecular Representation Learning with Taxonomic Annotations and Local Correspondence", In Proc. of the 39th Annual Conference on Neural Information Processing Systems, NeurIPS'25.