Yin Fang

yin.fang@nih.gov
fine.yinfang@gmail.com
fangyinfff.github.io
Postdoctoral Researcher
National Library of Medicine
National Institutes of Health

Research Profile

I am a biomedical AI researcher developing computational methods and benchmark resources for biomedical reasoning and data-driven biological discovery. My earlier work centered on molecular learning — knowledge graph-enhanced molecular representation, biomolecular instruction tuning, and molecular generation with feedback — bridging graph neural networks and large language models for chemistry and biology. My current research at NLM/NIH builds biomedical large language models for single-cell reasoning and clinical applications. I serve as project lead for TrialGPT 2.0, NLM/NIH's AI-powered patient-to-clinical-trial matching system.

Education

2020–2024 **Ph.D. in Computer Science**, Zhejiang University
Dissertation research focused on knowledge graph-enhanced molecular learning, biomolecular instruction tuning, and molecular generation.

2017–2020 **M.S. in Mathematics**, Sichuan University
Graduate training in machine learning.

2013–2017 **B.S. in Mathematics**, Sichuan University
Undergraduate training in mathematics.

Research Experience

2024–Present **Postdoctoral Researcher**
National Library of Medicine, National Institutes of Health
Supervisor: Dr. Zhiyong Lu

2020–2024 **Doctoral Researcher**
Department of Computer Science, Zhejiang University
Advisors: Dr. Huajun Chen and Dr. Xiaohui Fan.

Current Project

2026–Present **Project Lead, TrialGPT 2.0**
National Library of Medicine, National Institutes of Health
Lead NLM/NIH's TrialGPT 2.0 project, an AI-powered tool for matching patients to clinical trials. Oversee development and evaluation across retrieval, matching, and ranking modules for clinical trial recruitment support.

Awards & Honors

2022	Top 25 *Nature Communications* Life and Biological Sciences Articles of 2022	Nature Communications
2023	Best Paper Candidate	ISWC
2022	Best Application Paper	IJCKG

Selected Publications

All publications available at Google Scholar (2,046 citations)
† → Equal contribution

1. Molecular AI: Representation, Generation & Drug Discovery

Molecular/protein learning, generative chemistry, and property prediction.

1. **Yin Fang**, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, Huajun Chen. "Knowledge Graph-enhanced Molecular Contrastive Learning with Functional Prompt." *Nature Machine Intelligence*. 2023.
2. **Yin Fang**, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, Huajun Chen. "Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for Large Language Models." *International Conference on Learning Representations (ICLR)*. 2024.
3. **Yin Fang**, Ningyu Zhang, Zhuo Chen, Lingbing Guo, Xiaohui Fan, Huajun Chen. "Domain-Agnostic Molecular Generation with Self-feedback." *International Conference on Learning Representations (ICLR)*. 2024.
4. **Yin Fang**, Qiang Zhang, Haihong Yang, Xiang Zhuang, Shumin Deng, Wen Zhang, Ming Qin, Zhuo Chen, Xiaohui Fan, Huajun Chen. "Molecular Contrastive Learning with Chemical Element Knowledge Graph." *AAAI Conference on Artificial Intelligence*. 2022.
5. Jinlu Zhang†, **Yin Fang**†, Xin Shao, Huajun Chen, Ningyu Zhang, Xiaohui Fan. "The Future of Molecular Studies through the Lens of Large Language Models." *Journal of Chemical Information and Modeling*. 2024.
6. **Yin Fang**, Zhuo Chen, Xiaohui Fan, Ningyu Zhang, Huajun Chen. "Knowledge-informed Molecular Learning: A Survey on Paradigm Transfer." *International Conference on Knowledge Science, Engineering and Management (KSEM)*. 2024.
7. Xianrui Zhong, Bowen Jin, Siru Ouyang, Yanzhen Shen, Qiao Jin, **Yin Fang**, Zhiyong Lu, Jiawei

Han. "Benchmarking Retrieval-augmented Generation for Chemistry." *Conference on Language Modeling (COLM)*. 2025.

8. Jian Liu, Xuan Huang, Zhuo Chen, **Yin Fang**. "DRAK: Unlocking Molecular Insights with Domain-specific Retrieval-augmented Knowledge in LLMs." *CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC)*. 2024.
9. Xiang Zhuang, Qiang Zhang, Bin Wu, Keyan Ding, **Yin Fang**, Huajun Chen. "Graph Sampling-based Meta-Learning for Molecular Property Prediction." *International Joint Conference on Artificial Intelligence (IJCAI)*. 2023.
10. Xiang Zhuang, Keyan Ding, **Yin Fang**, Huajun Chen, Qiang Zhang. "Prompting Meta-Learned Hierarchical Graph Network for Molecular Property Prediction." *IEEE International Conference on Medical Artificial Intelligence*. 2023.
11. Jiyu Cui, Fang Wu, Wen Zhang, Lifeng Yang, Jianbo Hu, **Yin Fang**, Peng Ye, Qiang Zhang, Yiming Mo, Xili Cui, Huajun Chen, Huabin Xing. "Direct Prediction of Gas Adsorption via Spatial Atom Interaction Learning." *Nature Communications*. 2023.

2. Single-Cell & Spatial Omics AI

Computational methods for single-cell and spatially resolved transcriptomics.

1. **Yin Fang**, Qiao Jin, Guangzhi Xiong, Bowen Jin, Xianrui Zhong, Siru Ouyang, Aidong Zhang, Jiawei Han, Zhiyong Lu. "Cell-o1: Training LLMs to Solve Single-Cell Reasoning Puzzles with Reinforcement Learning." *Bioinformatics*. 2026.
2. **Yin Fang**, Xue Deng, Kangwei Liu, Ningyu Zhang, Jingyang Qian, Penghui Yang, Xiaohui Fan, Huajun Chen. "A Multi-Modal AI Copilot for Single-Cell Analysis with Instruction Following." *arXiv*. 2025.
3. Jie Liao[†], Jingyang Qian[†], **Yin Fang**[†], Zhuo Chen, Xiang Zhuang, Ningyu Zhang, Xin Shao, Yining Hu, Penghui Yang, Junyun Cheng, Yang Hu, Lingqi Yu, Haihong Yang, Jinlu Zhang, Xiaoyan Lu, Li Shao, Dan Wu, Yue Gao, Huajun Chen, Xiaohui Fan. "De Novo Analysis of Bulk RNA-seq Data at Spatially Resolved Single-cell Resolution." *Nature Communications*. 2022.
4. Jingyang Qian, Xiaoyuan Shao, Hudong Bao, **Yin Fang**, Wenbo Guo, Chengyu Li, Anyao Li, Huan Hua, Xiaohui Fan. "Identification and Characterization of Cell Niches in Tissue from Spatial Omics Data at Single-cell Resolution." *Nature Communications*. 2025.
5. Jingyang Qian, Jie Liao, Ziqi Liu, Ying Chi, **Yin Fang**, Yanrong Zheng, Xin Shao, Bingqi Liu, Yongjin Cui, Wenbo Guo, Yining Hu, Hudong Bao, Penghui Yang, Qian Chen, Mingxiao Li, Bing Zhang, Xiaohui Fan. "Reconstruction of the Cell Pseudo-space from Single-cell RNA Sequencing Data with scSpace." *Nature Communications*. 2023.
6. Jingyang Qian, Hudong Bao, Xin Shao, **Yin Fang**, Jie Liao, Zhuo Chen, Chengyu Li, Wenbo Guo, Yining Hu, Anyao Li, Yue Yao, Xiaohui Fan, Yiyu Cheng. "Simulating Multiple Variability in Spatially Resolved Transcriptomics with scCube." *Nature Communications*. 2024.

3. Biomedical Foundation Models & LLM Reasoning

General-purpose LLM methods for biomedical reasoning, evidence, and retrieval—not tied to a single omics modality.

1. Qiao Jin, **Yin Fang**[†], Lauren He, Yifan Yang, Guangzhi Xiong, Zhizheng Wang, Nicholas Wan, Joey Chan, Donald C. Comeau, Robert Leaman, Charalampos S. Floudas, Aidong Zhang, Michael

F. Chiang, Yifan Peng, Zhiyong Lu. "Med-V1: Small Language Models for Zero-shot and Scalable Biomedical Evidence Attribution." *arXiv*. 2026.

2. Qingqing Zhu, Qiao Jin, Tejas S. Mathai, **Yin Fang**, Zhizheng Wang, Yifan Yang, Maame Sarfo-Gyamfi, Benjamin Hou, Ran Gu, Praveen T. S. Balamuralikrishna, Kenneth C. Wang, Ronald M. Summers, Zhiyong Lu. "CT-Bench: A Benchmark for Multimodal Lesion Understanding in Computed Tomography." *arXiv*. 2026.

3. Yichi Zhang, Zhuo Chen, **Yin Fang**, Lingbing Cheng, Yanchao Lu, Fang Li, Wen Zhang, Huajun Chen. "Knowledgeable Preference Alignment for LLMs in Domain-specific Question Answering." *Findings of ACL*. 2024.

4. Guangzhi Xiong, Qiao Jin, Xiang Wang, **Yin Fang**, Haitao Liu, Yifan Yang, Fan Chen, Zhen Song, Dong Wang, et al.. "RAG-Gym: Optimizing Reasoning and Search Agents with Process Supervision." *arXiv*. 2025.

5. Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, **Yin Fang**, Jinhua Zhu, Shufang Xie, Tao Qin, Rui Yan. "BioT5+: Towards Generalized Biological Understanding with IUPAC Integration and Multi-task Tuning." *Findings of ACL*. 2024.

4. Knowledge Graphs & Multimodal Learning

KG representation, entity alignment, and cross-modal grounding.

1. Zhuo Chen, Yichi Zhang, **Yin Fang**, Yuxia Geng, Lingbing Guo, Jiaoyan Chen, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, Huajun Chen, Wen Zhang. "Knowledge Graphs for Multi-modal Learning: Survey and Perspective." *Information Fusion*. 2025.

2. Zhuo Chen, **Yin Fang**, Yichi Zhang, Lingbing Guo, Jiaoyan Chen, Jeff Z. Pan, Huajun Chen, Wen Zhang. "Noise-powered Multi-modal Knowledge Graph Representation Framework." *International Conference on Computational Linguistics (COLING)*. 2025.

3. Lingbing Guo, Zhuo Chen, Jiaoyan Chen, Yichi Zhang, Zhiqiang Sun, Zhenyu Bo, **Yin Fang**, Xiaowang Liu, Huajun Chen, et al.. "Distributed Representations of Entities in Open-world Knowledge Graphs." *Knowledge-Based Systems*. 2024.

4. Lingbing Guo, Zhuo Chen, Jiaoyan Chen, **Yin Fang**, Zhihao Wen, Huajun Chen. "Revisit and Outstrip Entity Alignment: A Perspective of Generative Models." *International Conference on Learning Representations (ICLR)*. 2024.

5. Zhuo Chen, Lu Guo, **Yin Fang**, Yichi Zhang, Jiaoyan Chen, Jeff Z. Pan, Yuanzhi Li, Huajun Chen, Wen Zhang. "Rethinking Uncertainly Missing and Ambiguous Visual Modality in Multi-Modal Entity Alignment." *International Semantic Web Conference (ISWC)*. 2023.

6. Zhuo Chen, Yuxia Huang, Jiaoyan Chen, Yuxia Geng, Wen Zhang, **Yin Fang**, Jeff Z. Pan, Wenting Song, et al.. "DUET: Cross-modal Semantic Grounding for Contrastive Zero-shot Learning." *AAAI Conference on Artificial Intelligence*. 2023.

7. Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, **Yin Fang**, Yuxia Huang, Yuxia Geng, Jeff Z. Pan, et al.. "MEAformer: Multi-modal Entity Alignment Transformer for Meta Modality Hybrid." *ACM Multimedia*. 2023.

8. Zhuo Chen, Yuxia Huang, Jiaoyan Chen, Yuxia Geng, **Yin Fang**, Jeff Z. Pan, Ningyu Zhang, Wen Zhang. "LaKo: Knowledge-driven Visual Question Answering via Late Knowledge-to-Text Injection." *International Joint Conference on Knowledge Graphs*. 2022.

Selected Open-Source Resources

Dataset	**Mol-Instructions** – biomolecular instruction dataset	46,813 downloads
Model	**MolGen-large** – molecular language model for generation	34,160 downloads

Academic Service

Journals Reviewer for *Nature Communications*, *Bioinformatics*, *Communications Chemistry*, *IEEE Transactions on Knowledge and Data Engineering*, *Molecular Informatics*, *Pattern Recognition* and *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Conferences Reviewer for ACL Rolling Review, NeurIPS, ICML, AAAI, and ICLR.