Kangjie Zheng

✉ Kangjie.zheng@gmail.com   📞 (+86)15550386688   🔗 Google Scholar   in Linkedin

Education

Peking University   *Aug 2020 - June 2025*
Ph.D. in Computer Science
- ○ **Supervisor:** Prof. Ming Zhang ☒
- ○ **Research Interests:** AI for Science, Sequence Modeling (e.g., natural language, genome, proteins, SMILES).

Harbin Institute of Technology   *Aug 2016 - June 2020*
B.Eng. in Computer Science
- ○ **College:** The Honors School of HIT (Top 10 graduates out of over 150 students)

Research Experience

Research Intern   *Beijing, China*
Tsinghua University, Institute for AI Industry Research   *Aug 2022 – Nov 2024*
- ○ Mentors: Prof. Wei-Ying Ma ☒ and Prof. Hao Zhou ☒.
- ○ Field of Research: Language models for protein and drug molecule modeling.
- ○ Developed the ESM All-Atom ☒ model to better understand multi-scale molecular data including drug molecules and protein molecules, achieving the state-of-the-art results on multiple protein-molecule tasks. (e.g., drug-target affinity prediction, enzyme–substrate affinity prediction, virtual screening, etc.)
- ○ Developed the Mol-AE ☒ model to better understand 3D molecular structural data, achieving the state-of-the-art results on multiple molecular property prediction tasks.

Research Intern   *Shenzhen, China*
Tencent AI Lab   *Aug 2021 - Aug 2022*
- ○ Mentors: Dr. Longyue Wang ☒ and Dr. Zhaopeng Tu ☒
- ○ Field of Research: Non-autoregressive generation models for text generation.
- ○ Designed a high-performance edit-based generative model, Dual-LevT ☒, achieving the state-of-the-art performance on multiple text generation tasks such as machine translation and text summarization.

Selected Publications

AI for Science:

1. **Kangjie Zheng***, Siyu Long*, Tianyu Lu[†], Junwei Yang, Xinyu Dai, Ming Zhang, Zaiqing Nie, Wei-Ying Ma, Hao Zhou. *ESM All-Atom: Multi-scale Protein Language Model for Unified Molecular Modeling,* ☒ *International Conference on Machine Learning* (ICML 2024).

2. **Kangjie Zheng**, Siyue Liang[†], Junwei Yang, Bin Feng, Zequn Liu, Wei Ju, Zhiping Xiao, Ming Zhang. *SMI-Editor: Edit-based SMILES Language Model with Fragment-level Supervision,* ☒ *International Conference on Learning Representations* (ICLR 2025).

3. Junwei Yang*, **Kangjie Zheng***, Siyu Long, Zaiqing Nie, Ming Zhang, Xinyu Dai, Wei-Ying Ma, Hao Zhou. *Mol-AE: Auto-Encoder Based Molecular Representation Learning With 3D Cloze Test Objective,* ☒ *International Conference on Machine Learning* (ICML 2024).

Language Modeling:

1. **Kangjie Zheng**, Junwei Yang, Siyue Liang[†], Bin Feng, Zequn Liu, Wei Ju, Zhiping Xiao, Ming Zhang. *ExLM: Rethinking the Impact of* `[MASK]` *Tokens in Masked Language Models,* ☒ *International Conference on Machine Learning* (ICML 2025).

2. **Kangjie Zheng**, Longyue Wang, Zhihao Wang, Binqi Chen[†], Ming Zhang and Zhaopeng Tu. *Towards A Unified Training for Levenshtein Transformer,* ☒ *IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP 2023).

3. Chenyang Huang, Hao Zhou, Cameron Jen, **Kangjie Zheng**, Osmar Zaiane, Lili Mou. *A Decoding Algorithm Based on Directed Acyclic Transformers for Length-Control Summarization,* ☑ *The 2024 Conference on Empirical Methods in Natural Language Processing Findings* (EMNLP 2024).

Others:

1. Chen Ye, Hongzhi Wang, **Kangjie Zheng**, Youkang Kong, Rong Zhu, Jing Gao, and Jianzhong Li. *Constrained Truth Discovery,* ☑ *IEEE Transactions on Knowledge and Data Engineering* (2020).

2. Wei Ju, Yifang Qin, Siyu Yi, Zhengyang Mao, **Kangjie Zheng**, Luchen Liu, Xiao Luo, Ming Zhang. *Zero-shot Node Classification with Graph Contrastive Embedding Network,* ☑ *Transactions on Machine Learning Research* (2023).

3. Chen Ye, Hongzhi Wang, **Kangjie Zheng**, Jing Gao, Jianzhong Li. *Multi-Source Data Repairing Powered by Integrity Constraints and Source Reliability,* ☑ *Information Sciences*, 507:386-403(2020).

* *Equal contribution.* † *Undergraduate students I supervised.*

Presentations

1. Poster: SMI-Editor: Edit-based SMILES Language Model with Fragment-level Supervision ☑ (ICLR'25).

2. Poster: Multi-scale Protein Language Model for Unified Molecular Modeling ☑ (ICML'24).

3. Poster: Auto-Encoder Based Molecular Representation Learning With 3D Cloze Test Objective ☑ (ICML'24).

4. Poster: Towards A Unified Training for Levenshtein Transformer ☑ (ICASSP'23).

5. Invited Talk for SAS Company: The Era of Large Models: An Introduction to Large Foundation Models For Language and Bio ☑.

6. Invited Talk for SAS Company: Towards A Unified Training for Levenshtein Transformer ☑.

Academic Service

○ Reviewer for Conference on Neural Information Processing Systems (NeurIPS'25).

○ Reviewer for International Conference on Learning Representations (ICLR'24, 25).

○ Reviewer for International Conference of Machine Learning (ICML'24, 25).

○ Reviewer for Annual Meeting of the Association for Computational Linguistics (ACL ARR).

Scholarships and Awards

○ **Merit Student of Peking University**, Oct. 2024

○ **Luoyuehua Scholarship of Peking University**, Oct. 2024

○ **Top 10 Graduates of the Honors School of HIT (Top 3%)**, Jun. 2020

○ **Merit Student of HIT**, Dec. 2018

○ **Nubiya Scholarship of HIT**, Mar. 2019

○ **First Class Renmin Scholarship (Top 10%) of HIT**, 2019, 2018, 2017

Skills

○ **Programming language:** C/C++, Python, Java, Matlab, CUDA

○ **Software:** Linux, MacOS, Git, GNU GCC, Visual Studio, VS Code

○ **Packages:** Pytorch, Fairseq, Numpy, Scipy, Scikit-Learn, Numba, PaddlePaddle, Matplotlib, Cupy