SIZHUANG HE

Email: sizhuang.he@yale.edu | Homepage: https://sizhuang.org | Github: https://github.com/SizhuangHe

RESEARCH INTEREST

Generative Modeling: Flow Matching, Diffusion, Discrete Diffusion, **Operator Learning**: Modeling Continuous Spatiotemporal Dynamics, Integral Equations, **Computational Biology**: Single-cell Transcriptomics Data Analysis, **LLMs and Agentic AI**: Autonomous Systems for Biological Discovery

Currently, I work on **discrete diffusion models on the finite symmetric group** and develop **LLM multi-agent systems** for single-cell perturbation response prediction and DNA methylation data curation.

EDUCATION

Yale University   New Haven, CT
Ph.D. in Computer Science   *Aug. 2024 – Present*

- Advisor: Dr. David van Dijk
- Research Focus: Machine Learning for Computational Biology

University of Michigan, Ann Arbor   Ann Arbor, MI
Bachelor of Science in Honors Mathematics (Minor in Computer Science)   *Sep. 2019 – May 2023*

- Graduated with Highest Distinction
- GPA: 4.0 / 4.0

RESEARCH EXPERIENCE

Generative Models and LLMs

- Developed **CaDDi**, a Non-Markovian Discrete Diffusion framework that unifies discrete diffusion with causal LLMs—two core paradigms in natural language generation previously viewed as orthogonal.
  - By breaking the longstanding Markovian assumption in discrete diffusion models, **CaDDi** achieves **more coherent text generation**—demonstrated by lower generative perplexity compared to prior discrete diffusion approaches.
  - Finetuned a Qwen-3 1.5B LLM to perform diffusion-style denoising, achieving improved performance on multiple reasoning benchmarks.
  - Published as a co-first author at NeurIPS 2025.
- Developed **CaLMFlow**, a flow-matching paradigm that reformulates the problem from solving an ODE to a Volterra integral equation, mitigating **stiffness** issues commonly encountered in ODE-based formulations.
  - Applied **CaLMFlow** to single-cell perturbation response prediction, demonstrating **superior ability to model the underlying data distribution** and **significantly improved extrapolation to OOD perturbations** compared to specialized baselines.

LLM Agentic Systems for Biological Data Curation

- Developing an **LLM-based multi-agent system** for large-scale, automated curation of DNA methylation datasets.
  - Designed a **fully automated** and **generalizable** pipeline that adapts to unseen datasets and aligns with the workflow needs of biological researchers.
  - The pipeline **scales to thousands of publicly available datasets** and reduces curation workloads that previously required years of manual effort by multiple biologists to just a few hours.
  - Designed for **deployment as a continuous system** that automatically downloads, curates, and updates relevant public datasets, providing biologists with clean, standardized data to accelerate downstream research.
  - Developing a **graph-based RAG system** for advanced agent memory storage and retrieval, enabling more accurate generation and allowing agents to self-evolve by learning from past experience.
  - Building a **LLM conversational assistant** that helps users explore and understand curated datasets

PUBLICATIONS

Non-Markovian Discrete Diffusion with Causal Language Models
Y. Zhang*, S. He*, et al. (NeurIPS 2025)

TANTE: Time-Adaptive Operator Learning via Neural Taylor Expansion
Z. Wu, S. Wang, S. Zhang, S. He, et al. (In Review)

Intelligence at the Edge of Chaos
S. Zhang*, A. Patel*, S. Rizvi, N. Liu, S. He, et al. (ICLR 2025)

COAST: Intelligent Time-Adaptive Neural Operators
Z. Wu, S. Zhang, S. He, et al. (AI4MATH Workshop at ICML 2025)

Scaling Large Language Models for Next-Generation Single-Cell Analysis
S. Rizvi*, D. Levine*, A. Patel*, S. Zhang*, E. Wang*, S. He, et al. (In Review)

CaLMFlow: Flow Matching using Causal Language Models
S. He*, D. Levine*, et al. (arXiv)

Operator Learning Meets Numerical Analysis: Improving Neural Networks through Iterative Methods
E. Zappala, D. Levine, S. He, et al. (arXiv)

* denotes equal contribution

HONORS & AWARDS

- **Fan Family Fellowship**, Yale University (2025)
- **Outstanding Achievement in Mathematics Award**, University of Michigan, Ann Arbor (2023)
- **James B. Angell Scholar**, University of Michigan, Ann Arbor (2023)
- **University Honors**, University of Michigan, Ann Arbor (2022, 2023)

SERVICES

Journal Reviewer
- Transactions on Machine Learning Research (TMLR)

Conference Reviewer
- International Conference on Learning Representations (ICLR), 2026
- AI4MATH Workshop at International Conference on Machine Learning (ICML), 2025