Weichi Yao

Address 500 S State St, Ann Arbor,MI 48109
Email weichi.yao@gmail.com
Homepage www.weichiyao.com

Updated in December 2025

Education

2023- **Postdoctoral Research Fellow**, Schmidt AI in Science - University of Michigan

AI Mentor Prof. Yixin Wang
Science Mentor Prof. Bryan R. Goldsmith

2017-2023 **Ph.D.**, Statistics - New York University
M.Phil, Statistics

Advisor Prof. Halina Frydman
Thesis: Machine Learning for Science

2015-2016 **M.A.**, Statistics - Columbia University

2011-2015 **B.A.**, Finance - Wuhan University
B.A., Applied Mathematics

Grants

2025-2027 ACED: Tail-aware Generative Modeling for Inverse Discovery of Molecules (Co-PI).

National Science Foundation (Total: $500K).

PI: Y. Wang (Statistics); Co-PI: B. R. Goldsmith (Chemical Engineering).

2025-2026 Reasoning-Augmented Large Language Models for Molecule Design and Discovery.

University of Michigan-OpenAI Research Grants (Total: $50K).

PI: B. R. Goldsmith (Chemical Engineering).

Ongoing Projects

2024- W. Yao, C. Gruich, B. R. Goldsmith and Y. Wang.

Large foundational model for conditional molecule generation.

Developing large foundational models to generate novel molecules with desired HOMO-LUMO gaps, for advanced energy storage and optoelectronic applications discovery.

2024- W. Yao, B. R. Goldsmith and Y. Wang.

Extrapolative conditional molecule generation.

Developing methods to adapt pretrained conditional molecular generative models to low-data, out-of-distribution regimes, with applications in drug and materials discovery.

2025- W. Yao, V. Madhavan , B. R. Goldsmith and Y. Wang

Reasoning-augmented property prediction in material science.

Leveraging large language models (e.g., ChatGPT) to generate structured chain-of-thought chemical reasoning that augments base property predictors, improving molecular property prediction extrapolative accuracy–especially for out-of-distribution molecules.

2025- V. Madhavan & W. Yao, et al.

Catalyst-Copilot: An autonomous catalyst analysis agent for transition state search.

Developing an autonomous LLM-driven "catalyst copilot" that manages end-to-end transition-state search workflows from user problem specification to converged structures. The agent tailors methods and settings to each reaction environment, reducing expert effort and improving robustness in computational catalysis.

Preprints

2025 W. Yao, B. Dumitrascu, B. R. Goldsmith and Y. Wang. Goal-oriented influence-based active data acquisition. *In preparation for submission to Journal of Machine Learning Research.*

2024 W. Yao, C. Gruich, B. R. Goldsmith and Y. Wang. Tail extrapolative conditional molecule generation. *In Proceedings of the International Conference on Machine Learning (ICML) AI4Science Workshop.*

2021 W. Yao, K. Storey-Fisher, D. W. Hogg and S. Villar. A simple equivariant machine learning method for dynamics based on scalars. *In Proceedings of the Advances in Neural Information* Processing Systems *(NeurIPS) Workshop on Machine Learning and the Physical Sciences.*

Publications

2024 S. Villar, D. W. Hogg, W. Yao, G. A. Kevrekidis and B. Schölkopf. Towards fully covariant machine learning. *Transactions on Machine Learning Research.*

2023 S. Villar, W. Yao, D. W. Hogg, B. Blum-Smith and B. Dumitrascu. Dimensionless machine learning: Imposing exact units equivariance. *Journal of Machine Learning Research, 24.*

2022 W. Yao, H. Frydman, D. Larocque and J. S. Simonoff. Ensemble methods for survival function estimation with time-varying covariates. *Statistical Methods in Medical Research*, 31(11): 2217-2236.

2021 W. Yao, H. Frydman and J. S. Simonoff. An ensemble method for interval-censored time-to-event data. *Biostatistics*, 22(1): 198-213.

2021 H. Moradian, W. Yao, D. Larocque, J. S. Simonoff and H. Frydman. Dynamic estimation with random forests for discrete-time survival data. *The Canadian Journal of Statistics*, 50(2): 533-548.

2021 S. Villar, D. W. Hogg, K. Storey-Fisher, W. Yao and B. Blum-Smith. Scalars are universal: Equivariant machine learning, structured like classical physics. *In Proceedings of the Advances in Neural Information Processing Systems*

2019 W. Yao, A. S. Bandeira and S. Villar. Experimental performance of graph neural networks on random instances of max-cut. *In Proceedings of the Society of Photographic Instrumentation Engineers.*

2017 J. H. Lee, D. E. Carlson, H. S. Razaghi, W. Yao, G. A. Goetz, E. Hagen, E. Batty, E. J. Chichilnisky, G. T. Einevoll and L. Paninski. YASS: Yet Another Spike Sorter. *In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 4005-4015.

Internship

12/2022-05/2023 Science Intern. Amazon Development Center, Germany
Transfer learning with deep tabular models.

05/2022-09/2022 Research Intern. Applied Machine Learning Team, TikTok at ByteDance
Maximum likelihood estimation with full-likelihood formulation for click-through rate prediction on nonuniform subsampled data.

12/2016-07/2017 Research Assistant. Grossman Center for the Statistics of Mind, Columbia University
Statistical models and pipelines for spike sorting on dense multi-electrode arrays.

Teaching experience

2020 Course Instructor. STAT-UB Statistics for Business Control, New York University.

Prizes or Awards

2014 Honorable Mention. Mathematical Contest in Modeling (MCM)

2013 Second Prize. Symphony Orchestra, National College Students Art Exhibition, China.

2009 Gold Award. Symphony Orchestra, Australian International Music Festival, Australia.