Yi Yang

+1-412-736-3800 | yiy5@andrew.cmu.edu | linkedin.com/in/yiyang5 | github.com/Yi5817 | yi5817.github.io

Education

Carnegie Mellon University — Pittsburgh, PA
Doctor of Philosophy (Ph.D.) in Materials Science and Engineering — *Jan. 2023 – Fall 2026 (expected)*
- Research Focus: AI for molecular crystals, machine learning interatomic potentials.

Carnegie Mellon University — Pittsburgh, PA
Master of Science in Computational Materials Science and Engineering — *Aug. 2021 – Dec. 2022*

China University of Geosciences, Wuhan — Wuhan, China
Bachelor of Science in Materials Science and Engineering (experimental class) — *Sept. 2017 – July 2021*

Selected Publications

[1] Gharakhanyan, V.*, **Yang, Y.***, et al. FastCSP: Accelerated Molecular Crystal Structure Prediction with Universal Model for Atoms. *arXiv:2508.02641* (2025). [Under review, *Nature Machine Intelligence*; ∗ Equal contribution.]

[2] Gharakhanyan, V., **Yang, Y.**, et al. Open Molecular Crystals 2025 (OMC25) Dataset and Models. *Scientific Data*, **13**, 354 (2026).

[3] **Yang, Y.**, et al. Genarris 3.0: Generating Close-Packed Molecular Crystal Structures with Rigid Press. *J. Chem. Theory Comput.* **21**, 11318–11332 (2025).

[4] Nayal, K. S., **Yang, Y.**, et al. Efficient Molecular Crystal Structure Prediction and Stability Assessment with AIMNet2 Neural Network Potentials. *Crystal Growth & Design* **25**, 9092–9106 (2025).

[5] Hunnisett, L. M. et al. The Seventh Blind Test of Crystal Structure Prediction: Structure Ranking Methods. *Acta Cryst. B* **80**, 548–574 (2024).

Research Experience

Research Assistant, Carnegie Mellon University × Meta FAIR Chemistry — Pittsburgh, PA
ML Interatomic Potentials for Large-Scale Molecular Crystal Discovery — *2024 – Present*

Accelerated Molecular Crystal Structure Prediction with Machine Learning Atomic Potentials
- Constructed the **Open Molecular Crystals 2025 (OMC25)** dataset (**27M+** structures), supporting large-scale training and evaluation of deep learning models.
- Post-trained and benchmarked state-of-the-art MLIPs **Universal Models for Atoms (UMA)** for OMC task, achieving near **DFT-level accuracy**.
- Co-developed **FastCSP**, an open-source **Python** crystal structure prediction workflow integrating structure generation with MLIP-based high-throughput **geometry optimization** and **energy ranking**, reducing CSP from weeks to **hours** on GPUs.
- Validated **transferability and robustness** of UMA on organic molecular crystals in large-scale workflows, co-authoring two publications with the **Meta FAIR Chemistry** team.

Research Assistant, AI for Crystal Structure Prediction — Pittsburgh, PA
Carnegie Mellon University, Advisor: Noa Marom — *Sept. 2021 – Present*

Generative Model for Molecular Crystal Structure Prediction — *2026 – Present*
- Developing a **two-track flow matching generative model** for de novo molecular crystal structure prediction across **single-component, cocrystal, and salt** systems.
- Designing a **coarse-to-fine generation curriculum** that jointly generates over a **product manifold** of symmetry-aware lattice parameters, rigid-body rotations on **SO(3)**, and fractional translations on the 3-torus
- Aligning the generative prior to first-principles physics through **multi-fidelity reward fine-tuning** across increasing MLIP relaxation depths to improve physical validity.

Genarris 3.0: Random Molecular Crystal Structure Generator | (https://github.com/Yi5817/Genarris)
- Developed **Genarris 3.0**, an open-source **Python/C** package with **MPI-based parallelism** for large-scale molecular crystal structure generation on HPC clusters.
- Designed and implemented **Rigid Press**, a novel geometry optimization algorithm using a regularized hard-sphere potential for efficient unit cell compression without explicit energy evaluations.
- Integrated **GPU-accelerated MLIP inference** and **unsupervised clustering** for efficient down-selection, accelerating exploration of molecular crystal energy landscapes.
- Validated robustness across multiple **drug-like** molecular crystals and **energetic materials**, supporting applications in **AI-driven materials discovery**.

The 7^{th} Crystal Structure Prediction Blind Test, Cambridge Crystallographic Data Center
- Achieved a **67% success rate**, **best among academic teams** and **3rd overall**, pioneering early application of MLIPs with near-DFT accuracy at significantly reduced computational cost.
- Developed and trained **system-specific AIMNet2 MLIPs** using **active learning**, improving model accuracy while reducing cost.
- Designed **data sampling and augmentation strategies** to improve training data diversity and model robustness.

Course Projects

TimeAware CardAgent: Neuro-Symbolic Multi-Agent Planner | *11-766 LLM Applications, CMU Jan. 2026 – May 2026*

- Built a **neuro-symbolic multi-agent system** where an LLM planner proposes plans and a deterministic Python verifier enforces hard rules and computes expected value, eliminating a class of LLM failure modes by design.
- Developed a **temporal RAG pipeline** (**ChromaDB** + sentence embeddings) with pre-retrieval expiry filtering and daily refresh, so the planner reasons only over currently valid evidence.
- Implemented a 6-month **sequential decision-tree simulator** searching multi-card application sequences under 8 banking constraints, ranked by expected value.
- Evaluated the verified pipeline against baseline LLM planners, showing that deterministic constraint checks caught rule-violating plans produced by unconstrained model outputs.

AI Ramanujan: Formula Equivalence Discovery | *11-785 Introduction to Deep Learning, CMU* *Oct. 2022 – Dec. 2022*

- Developed **transformer-based models** with cross-attention mechanisms for sequence representation learning.
- Trained models using **contrastive learning**, improving embedding quality for downstream similarity tasks.

Time-Series Forecasting for Stock Ranking | *Kaggle Competition* *May 2022 – July 2022*

- Engineered temporal and statistical features from financial data for trend forecasting and ranking.
- Trained an **LSTM-based sequence model** in PyTorch for time-series prediction, ranking **top 13%** on Kaggle.

Honors & awards

Presidential Fellowship, College of Engineering, CMU	*Feb. 2026*
Machine Learning in Chemical and Materials Sciences (MLCM) Rising Star Award, LANL	*May 2025*
Frontera Computational Science Fellowship, TACC	*Apr. 2025*
Award for Research Excellence in the Master Program, CMU	*May 2023*
ATK-Nick G. Vlahakis Graduate Fellowship, CMU	*Feb. 2023*
Outstanding Undergraduate Award, CUG	*June 2021*

Technical Skills

Programming: Python, C, MPI
ML/AI: PyTorch, TensorFlow, Transformer, Graph Neural Networks, Diffusion Models, Flow Matching
Data & Infrastructure: PySpark, Linux, AWS, GPU Computing, HPC (Bridges2 (PSC), Frontera (TACC), Vista (TACC), Perlmutter (NERSC))
Libraries: ASE, Pymatgen, RDKit, OpenBabel, NumPy, Pandas, Scikit-learn
Computational Materials: FHI-aims, VASP, Orca
Languages: English (Fluent), Chinese (Native Speaker)