Department of Data Science
香港城市大學
City University of Hong Kong

DS SEMINAR

Data-Centric Machine Learning and Foundation Models for Molecular Discovery

Date: 10 March 2026 (Tuesday)

Time: 9:30am - 10:30am

Zoom: https://cityu.zoom.us/j/84206715622

ABSTRACT

Foundation models and generative AI are changing how we search and design molecules across chemistry, biology, and materials. However, progress is limited by a basic mismatch: chemical space is enormous (often estimated to exceed 10^{63} candidates), while labeled measurements are scarce (often only a few hundred to a few thousand per property). In addition, real applications require optimizing multiple, sometimes conflicting, properties at once, such as potency and toxicity for drugs or permeability and selectivity for gas-separation membranes.

In this talk, I present a data-to-discovery workflow for molecular virtual screening and a foundation model for inverse molecular design under multi-property constraints. First, I develop data-centric learning methods for small and imbalanced datasets. By learning interpretable subgraph rationales and using them for data augmentation and confidence-based self-training, my models improve prediction accuracy while giving structure-level explanations that scientists can validate. Second, I introduce Graph Diffusion Transformers (Graph DiTs) for multi-conditional molecular generation, and show how combining Graph DiTs with large language models leads to multimodal foundation models that can interleave text, molecules, and multi-step reactions for controllable design and retrosynthesis. Third, I translate these advances into practical tools and shared resources, including the open-source library torch-molecule and an open polymer challenge that connects machine learning researchers with domain scientists.

I conclude with case studies in sustainable materials, including gas-separation membranes, where these methods helped drive experimentally validated discoveries, and I outline a roadmap toward multi-scale, multi-modal molecular foundation models and agent systems that work in tighter loops with experiments.

Mr. Gang LIU

GUEST SPEAKER'S PROFILE

Mr. Gang LIU is a fifth-year Ph.D. student at the University of Notre Dame, working on generative AI and foundation models for molecular discovery. He has published as (co-) first author in NeurIPS, KDD, ICLR, IEEE TKDE, ACM TKDD, and Cell Reports Physical Science. His work has been supported by an IBM Ph.D. Fellowship and featured by MIT News, Notre Dame Engineering News, and Snap Research News. He is the author of two books on deep learning for polymers and the creator of torch-molecule, an open-source toolkit for molecular discovery. He led the NeurIPS 2025 Open Polymer Challenge, which attracted more than 10,000 registrations and 50,000 submissions from over 100 countries.

Enquiries: ds.go@cityu.edu.hk

All are welcome