香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

馮諾依曼研究院 Von Neumann Institute

北京中关村学院 x 中关村人工智能研究院

Zhongguancun Academy　Zhongguancun Institute of Artificial Intelligence

Pioneering AI for Scientific Discoveries: Zhongguancun x HKUST Innovations

SENG Commons ♦ 26 Sep 2025 14:00-16:05

*Artificial Intelligence is rapidly transforming the landscape of scientific research and societal systems. This workshop brings together leading scholars from the **Zhongguancun Academy (ZGCA), Zhongguancun Institute of Artificial Intelligence (ZGCI)**, and HKUST to explore cutting-edge AI applications across disciplines—from drug discovery and virus identification to LLM.*

This event aims to foster interdisciplinary collaboration, spark new research directions, and build a vibrant community of innovators working at the intersection of AI, science, and society. Join us to engage with pioneering ideas and help shape the future of AI-powered discovery.

Time	Title	Speaker
14:00-14:05	Opening Remarks	Ping TAN (HKUST) Tieyan LIU (ZGCA/ZGCI)
Workshop: 10-min presentation + 10-min discussion		
14:05-14:25	Drug Discovery: Tackling Diseases with hard-to-drug target and unknown target	Tao QIN (ZGCA/ZGCI)
14:25-14:55	AI for Chemistry via Multiscale Science Driven Modelling: From Models to Applications	Lixue CHENG (HKUST)
14:45-15:05	Virus Identification with a Protein Foundation Model	Haiguang LIU (ZGCA/ZGCI)
15:05-15:35	A Multi-Agent System for Complex Chemical Reaction Information Extraction	Hanyu GAO (HKUST)
15:25-15:45	Efficient and Robust Large Language Model (LLM) Inference Scheduling Optimization	Zijie ZHOU (HKUST)
15:45-16:05	Accommodating LLM Service over Heterogeneous Computational Resources	Binhang YUAN (HKUST)

RSVP: https://ust.az1.qualtrics.com/jfe/form/SV_1MkzOaqD8WaodOm

Enquiry: Ellen YAN　Email: ellenyan@ust.hk　Tel: 23588522

1) Drug Discovery: Tackling Diseases with Hard-to-drugTarget and Unknown Target

Mainstream AI-driven drug discovery technologies are predominantly designed for diseases with well-defined molecular targets. As a result, approximately 95% of diseases—characterized by either “undruggable targets” or the absence of clear targets—remain largely untouched, forming a vast “no-man's land” in AI pharmaceutical research.

To systematically address this challenge, ZGCA·ZGCI have developed a cross-domain biological foundation model capable of comprehensively understanding disease mechanisms through multi-modality data integration, thereby reducing reliance on predefined targets.

For diseases with undruggable targets, the team focused on malaria and successfully designed novel drug candidates capable of binding to disordered target proteins in Plasmodium species. These molecules demonstrated exceptional activity in wet-lab experiments. For diseases with unclear targets, the team focuses on fibrosis. Using cellular phenotypes as input, they designed a series of new drug candidates that significantly outperformed reference compounds in terms of biological activity.

These results highlight the potential of cross-domain biological foundation models to unlock new pathways in AI-driven drug discovery, offering promising strategies to explore previously inaccessible therapeutic landscapes.

Speaker: Dr. Tao QIN

Dr. Tao Qin is the vice president of Zhongguancun Academy (ZGCA) and oversees its AI4science division. He earned both his bachelor's and PhD degrees from Tsinghua University. Before joining ZGCA, he served as a partner research manager at Microsoft Research AI4Science Lab and led the Microsoft Research AI4Science Asia team. His team proposed dual learning in 2016, which helped Microsoft achieve human parity in the 2018 Chinese-English news translation and win 8 tasks at the WMT2019. In 2019, his team developed the most efficient speech synthesis model at the time, FastSpeech, achieving 100x acceleration and becoming a key component supporting hundreds of languages and voices in Microsoft's Azure cloud service. In the same year, his team developed the most powerful Mahjong AI ever, Suphx, achieving 10 DAN on the Tenhou platform, with a stable rank significantly superior to top human professionals. In 2020, he published the academic monograph 'Dual Learning'. Recently his team focuses on AI for scientific discovery, including science foundation models, drug discovery, materials design, biology research, etc.

2) AI for Chemistry via Multiscale Science Driven Modelling: From Models to Applications

In this talk, I will present my research journey in building multiscale, physics-driven AI frameworks to accelerate computational chemistry and chemical discovery. AI for Science is at an exciting crossroad: while deep learning has brought transformative advances in modeling complex systems, critical challenges remain in scalability, interpretability, and physical fidelity. My work integrates state-of-the-art machine learning, quantum chemistry, and quantum computing into a unified framework that spans multiple modeling scales—from deep quantum Monte Carlo (QMC) simulations to transferable orbital-based learning (MOB-ML), large-scale ML force fields, and AI-guided experimental design. I will first introduce recent developments in deep QMC, where neural networks are used to parameterize highly accurate wavefunctions for electronic structure calculations, and discuss how my methods extend QMC to extract molecular properties beyond energy. I will then describe MOB-ML, a molecular orbital–based machine learning approach that achieves wavefunction-level accuracy at DFT-level cost, supporting accurate property predictions and molecular dynamics simulations. Building on these models, I develop scalable ML force fields and close the loop between computation and experiment with Bayesian optimization pipelines for protein engineering and chemistry design. My ultimate goal is to construct a general-purpose “foundation model” for computational chemistry, one that bridges data-driven and physics-based methods, supports multiple fidelities of input and output, and enables applications in chemistry, biology, and materials science. This multiscale strategy offers a promising path toward accurate, interpretable, and efficient AI tools for understanding and manipulating the molecular world—empowering both academic research and practical applications in drug discovery, material design, and beyond.

Speaker: Lixue CHENG

Lixue Cheng (Sherry) is currently an Assistant Professor of The Hong Kong Science and Technology. Previously, she was a researcher Microsoft Research AI for Science Lab and Tencent Quantum Lab. She graduated with a PhD in theoretical chemistry in California Institute of Technology working with Prof. Thomas F. Miller III in 2022. Sherry received a B.S. degree with quadruple majors in Chemistry, Math, Biochemistry, and Molecular Biology and a minor in Computer Science from University of Wisconsin-Madison. She is interested in the interdisciplinary research between chemistry, physics, biology, and computer sciences, and passionate about bridging the mind gaps between different areas. Her current research focuses on interfaces of AI, quantum computing, and chemistry applications, such as molecular modelling by Orbital-Based Machine Learning, deep Quantum Monte Carlo (deep QMC), AI for quantum algorithms, and LLM for Chemistry.

RSVP: https://ust.az1.qualtrics.com/jfe/form/SV_1MkzOaqD8WaodOm

3) Virus identification with a Protein Foundation Model

Rapid identification of novel viruses and efficient protein engineering are central to pan-demic preparedness, therapeutic development, and biotechnological innovation. Yet, most viral and protein sequences remain poorly characterized, limiting our ability to predict function or evolution. Here, we leverage protein foundation models to capture structural, functional, and evolutionary signals beyond primary sequence similarity. In viral metagenomes, our framework models genomes as structured sequences of pro-tein-coding genes, enabling alignment-free detection of known viruses and discovery of candidate novel lineages. In protein engineering, the same embeddings guide predictive modeling of mutational effects and reinforcement-learning-driven exploration of the fit-ness landscape, efficiently identifying high-function multi-point mutants from limited experimental data. By unifying protein foundation modeling with genome- and fit-ness-aware strategies, this approach illuminates uncharted biological sequence space, offering scalable solutions for virus discovery, protein optimization, and next-generation biotechnology.

Speaker: Haiguang LIU

Dr. Haiguang Liu graduated from University of California, Davis with a Ph.D in Applied Sciences in 2009. Prior to that, Dr. Liu studied physics in Hong Kong Baptist University through the scholarship sponsored by Hong Kong Jockey Club and received Bachelor degree with first class honor in 2003. Dr. Liu worked in several prestigious institutes, including Lawrence Berkeley National Laboratory (LBNL), Arizona State University (ASU), and then started independent research group in Beijing Computational Science Research Center (CSRC) and tenured in 2020. He then joined Microsoft Research, mainly working in the field of AI for Science. Dr. Liu mainly engaged in biophysics research, including structure, function and interactions of protein molecules, drug discovery, protein design, and recently cell modeling with AI. Dr. Liu published 100 scientific papers in peer reviewed journals or conferences.

4) A Multi-Agent System for Complex Chemical Reaction Information Extraction

The extraction of structured chemical information from literature is essential for constructing reaction databases that drive data-driven and AI-powered chemical research. A major challenge lies in the multimodal and complex nature of chemical data. In order to convert the raw data into structured, machine readable datasets, a modeling framework is needed with the ability to comprehend text, tables, and graphical representations of molecular structures. Existing approaches predominantly focus on single tasks, limiting their ability to fully capture reaction details and leading to incomplete datasets. To overcome this, we present ChemEagle, a multimodal large language model (MLLM)-based multi-agent system that integrates diverse chemical information extraction tools. By integrating seven expert-designed tools and six chemical information extraction agents, ChemEagle not only processes individual modalities but also utilizes MLLMs' reasoning capabilities to unify extracted data, ensuring more accurate and comprehensive reaction representations. We demonstrated the model's capability in parsing chemical reaction diagrams, converting molecular structures to SMILES, and substituting functional groups or molecular fragments from tabulated lists. In all these tasks our model performed better than existing methods by a significant margin. Our approach presents a significant step towards automated chemical knowledge extraction, facilitating more robust AI-driven chemical research.

Speaker: Hanyu GAO

Hanyu is joining HKUST as an assistant professor in the Spring term 2021. Hanyu obtained his Bachelor's degree in chemical engineering from Tsinghua University, China, where he was Magna Cum Laude in 2012. He then went to the U.S. and completed his PhD in the Department of Chemical and Biological Engineering at Northwestern University with Prof. Linda Broadbelt in 2017. After that he worked as a postdoctoral associate at MIT in Prof. Klavs Jensen's group. Hanyu's research interest lies in using modeling techniques, including simulation, optimization and machine learning, to solve chemical engineering problems ranging from polymer reaction engineering to organic synthesis design.

RSVP: https://ust.az1.qualtrics.com/jfe/form/SV_1MkzOaqD8WaodOm

5) Efficient and Robust Large Language Model (LLM) Inference Scheduling Optimization
We study the problem of optimizing Large Language Model (LLM) inference scheduling to minimize total completion time. LLM inference is an online and multi-task service process and also heavily energy consuming by which a pre-trained LLM processes input requests and generates output tokens sequentially. Therefore, it is vital to improve its scheduling efficiency and reduce the power consumption while a great amount of prompt requests are arriving. There are two key challenges: (i) each request has heterogeneous prefill and decode lengths. In LLM serving, the prefill length corresponds to the input prompt length, which determines the initial memory usage in the KV cache. The decode length refers to the number of output tokens generated sequentially, with each additional token increasing the KV cache memory usage by one unit. We show that minimizing total completion time is NP-hard due to the interplay of batching, placement constraints, precedence relationships, and linearly increasing memory usage. We then analyze commonly used scheduling strategies in practice, such as First-Come-First-Serve (FCFS) and Shortest-First (SF), and prove that their competitive ratios are unbounded. To address this, we propose a novel algorithm based on a new selection metric that efficiently forms batches over time. We prove that this algorithm achieves a constant competitive ratio. (ii) the output length, which critically impacts memory usage and processing time, is unknown. We first design a conservative algorithm, Amax, which schedules requests based on the upper bound of predicted output lengths to prevent memory overflow. However, this approach is overly conservative: as prediction accuracy decreases, performance degrades significantly due to potential overestimation. To overcome this limitation, we propose Amin, an adaptive algorithm that initially treats the predicted lower bound as the output length and dynamically refines this estimate during inferencing. We prove that Amin achieves a log-scale competitive ratio.

Speaker: Zijie ZHOU
I am an assistant professor at HKUST IEDA, starting August 2025. My current research mainly focuses on making Large Language Model (LLM) inference faster and more cost-efficient using operations research—including online optimization, scheduling, queueing, and resource allocation. I also work on classical OR topics like revenue management, online learning, and experiment design. Before joining HKUST, I earned my PhD degree from MIT's Operations Research Center (ORC) and Laboratory for Information & Decision Systems (LIDS) in 2025. During my PhD, I interned at Microsoft Research and Microsoft Azure in 2024, working on LLM inference and cloud computing. I also interned at Oracle Lab in 2023 working on hospitality optimization.

6) Accommodating LLM Service over Heterogeneous Computational Resources

Deploying a large-scale language model service is crucial to contemporary AI applications. We focus on deploying such services in a heterogeneous and potentially decentralized setting to mitigate the substantial costs typically associated with centralized data centers. Our work relies on carefully designed scheduling algorithms where we model the computation capacity and inter-machine connection precisely and propose an efficient searching algorithm to find the optimal allocations for different LLM serving paradigms. Our empirical study suggests that the proposed method can efficiently reduce service costs while preserving service quality.

Speaker: Binhang YUAN
Binhang Yuan is an Assistant Professor at the Department of Computer Science and Engineering (CSE), the Hong Kong University of Science and Technology (HKUST). He received his Ph.D. and master's degrees from Rice University and his bachelor's degree from Fudan University. Before joining HKUST, he was a Postdoc at the Swiss Federal Institute of Technology Zurich (ETH Zurich). His main research interests are in data management systems for machine learning and distributed and decentralized machine learning systems.

RSVP: https://ust.az1.qualtrics.com/jfe/form/SV_1MkzOaqD8WaodOm