University of Pittsburgh | School of Computing and Information

Proposal Defense

Doctor of Philosophy in Computer Science

"Machine Learning for Biomolecular Interactions Across Scales: Navigating Ambiguity from Molecular Properties to Protein Design"

by

Yue Wan

Date: May 19, 2026
Time: 10:30 a.m. – Noon
Place: 6106 Eli Lilly Room, Sennott Square, 210 South Bouquet Street, Pittsburgh, PA 15260

Committee:

- Xiaowei Jia, Associate Professor, Computer Science, University of Pittsburgh
- Milos Hauskrecht, Professor, Computer Science, University of Pittsburgh
- Lorraine (Xiang) Li, Assistant Professor, Computer Science, University of Pittsburgh
- Martin (Renqiang) Min, Department Head, Machine Learning Department, NEC Laboratories America

Abstract:
Biomolecular interactions are inherently ambiguous. The mapping from atomic identity to interaction outcome is many-to-many, context-dependent, and systematically underspecified by available experimental data, including the interaction of a small molecule with its biological target, the binding of a peptide within an immune receptor groove, or the biochemical interactions that govern the protein sequence identity and three-dimensional fold. This thesis develops machine learning methods that embrace and resolve this ambiguity across three molecular scales, using tailored representation learning and architectural designs.

At the scale of **small molecules**, we address molecular property prediction under the challenge of activity cliffs, where structurally similar molecules may elicit dramatically different biological activities given implicit target. We introduce a multi-channel self-supervised learning framework that encodes structural hierarchies within molecules through distinct learning channels and aggregates them adaptively for downstream tasks. By incorporating chemical heuristics, including scaffold and molecular similarity implication, the framework learns representations that are sensitive to the subtle structural variations underlying activity cliffs, achieving competitive performance across standard benchmarks and challenging cases of activity cliffs.

At the scale of **peptide-protein** interactions, we study the MHC class II (MHC-II) antigen presentation pathway, a key biological process in adaptive immunity, immunotherapy and vaccine design. MHC-II epitope prediction is significantly challenging due to the variable-length of peptides, open-ended grooves with ambiguous binding core alignments, multi-stage bioactivity signals, and the non-standardized data across alleles. We address these challenges in two complementary contributions. First, we construct a well-curated, multi-source dataset that standardizes existing peptide-MHC-II samples and introduces the novel antigen-MHC-II samples. We then formalize three prediction tasks of binding affinity, peptide presentation, and antigen presentation, to capture the biological ambiguity of the MHC-II pathway. Second, we present a structure-aware multi-instance learning framework to connect the sparse but hierarchical supervision signals from binding core estimation, peptide scoring, and antigen

presentation. The model consistently outperforms state-of-the-art methods, yields interpretable predictions, and generalizes to clinically relevant settings such as SARS-CoV-2 epitope analysis.

At the scale of **full proteins**, we explore the most fundamental biomolecular ambiguity between protein sequence (evolutionary and biological) and structure (chemical and physical). The bidirectional mapping between sequence and structure is both many-to-one, as vastly different sequences can fold into similar structures, and one-to-many, as a given structural scaffold is compatible with a large combinatorial space of sequences. We specifically study the protein sequence structure co-design problem to capture the joint probability between sequence and structure across the vast biochemical space. We pursue this through two approaches that differ in how they represent and navigate this joint space. First, we represent protein by interleaving residue identity and explicit structure tokens and treat co-design as a sequence generation problem amenable to the modern LLM reasoning capability. Second, we seek to learn a vector quantized latent representation that emerges directly from the joint sequence-structure space in an SE(3) equivariant manner. This latent tokenization allows the model to reason over a compact, interaction-aware codebook that encodes coupled sequence-structure information that neither modality alone can provide.

Together, these contributions explore the potential of machine learning models across a wide breadth of biomolecular scale. The resulting methods advance molecular property prediction, computational immunotherapy, and generative protein design, and collectively point toward a principled framework for machine learning over the biomolecular system.