Statement of Purpose

Yikun Han

At the intersection of natural language processing and graph learning, I believe AI is poised to revolutionize healthcare. My research aims to develop next-generation AI systems that empower medical discovery and enable personalized healthcare interventions. By integrating advanced computational methods with domain-specific medical knowledge, I strive to create trustworthy, interpretable, and impactful AI tools that enhance patient care and advance medical research.

Research in Graph ML. In machine olfaction, the goal is to predict the smell of unseen odorants—a complex task requiring models to translate molecular structure into sensory perception. During my first year of graduate study, under *Prof. Ambuj Tewari*. I worked on replicating the results of a principal odor map using Graph Neural Networks (GNNs). This experience honed my expertise in GNNs and deepened my understanding of olfactory science.

When the DREAM Olfactory Mixtures Prediction Challenge launched [1], I led a team tackling this unique problem: predicting perceptual distances between mixtures of molecules. To address the small dataset size, we combined pre-trained GNN embeddings with boosting for downstream metric learning. Our team secured first place, and I presented our work at RSGDREAM 2024 [2]. This collaboration has since expanded into a joint paper with top-ranking teams from Yale, Cornell Tech, and KTH.

Data scarcity emerged as a central challenge in machine olfaction. Building on this insight, I developed a novel approach combining transfer learning and weakly-supervised learning for odor descriptor prediction, yielding strong results. This project also gave me the opportunity to mentor a senior student, co-authoring an extended version of this work for journal submission.

Research in Natural Language Processing. My NLP research began with a project on LLM knowledge distillation for multiple-choice question answering, conducted under *Prof. Nitesh V. Chawla* in collaboration with researchers from *Notre Dame* and *UCLA*. The work addressed challenges in low-latency and domain-specific AI applications. I independently developed *TinyLLM*, a model leveraging reasoning capabilities from multiple large teacher LLMs via a Chain-of-Thought strategy, ensuring the student model absorbed accurate, contextually grounded rationales. This framework significantly improved prediction quality, outperforming benchmarks and state-of-the-art methods in knowledge distillation, leading to a publication at WSDM 2025 [3].

Iterative feedback during re-submissions strengthened the work and taught me resilience in navigating rejection. Following acceptance, we open-sourced our code to promote reproducibility and accessibility in the research community.

Research in Recommendation. As part of the Cell Maps for AI (CM4AI) program under *Prof. Ying Ding* and *Prof. Jiliang Tang*, I developed a recommendation system for fostering biomedical collaborations. Unlike traditional link prediction, our framework generates recommendations based on publications rather than authors, leveraging large language models and the PubMed knowledge graph.

We introduced nuanced personas for each author to reflect domain-specific expertise and implemented a retrieval-and-reranking pipeline, with LLMs improving results significantly on the Mean Reciprocal Rank (MRR) metric. This project highlighted my motivation to create solutions that address real-world challenges rather than merely pursuing state-of-the-art metrics. Currently, we are refining the system for a user study and plan to submit it to an AI conference's demo track.

Preparation and Future Directions. My diverse research experiences across machine olfaction, NLP, and recommendation systems have equipped me with transferable skills and the ability to learn

quickly. While my prior work focused on general machine learning, I am now tackling domain-specific challenges in bioinformatics and medical research.

Under the guidance of *Prof. Mengdi Wang*, I am currently working on designing reasoning pipelines for molecular cloning, which has expanded my understanding of biology and the practical tools used in the field. This project underscores the importance of grounding computational research in domain expertise. As *Prof. Fei-Fei Li* aptly stated, "AI is created by humans, intended to behave like humans, and, ultimately, to impact human lives and society." My aspiration is to develop biomedical AI that addresses real scientific challenges, advances discovery, and meaningfully impacts people.

At UIUC. I am particularly drawn to the *Information Sciences PhD program* at UIUC for its interdisciplinary emphasis and alignment with my research interests. Working with *Prof. Halil Kilicoglu* on connecting complementary medicine and biological knowledge to support integrative health would allow me to explore NLP approaches for mining COMB knowledge and constructing domain knowledge graphs.

I am also excited by *Prof. Yue Guo*'s work on lay language generation and jargon identification—both critical in biomedical AI—and would like to extend my expertise in LLMs to explore uncertainty and explainability in this domain. Similarly, *Prof. Ismini Lourentzou*'s research on AI for healthcare aligns with my interests in combining LLMs and graph learning to develop interpretable, impactful tools for the medical field.

UIUC's strong faculty, collaborative culture, and emphasis on impactful research make it an ideal environment for me to grow as a researcher and contribute meaningfully to biomedical AI.

Bibliography

[1] "DREAM Olfactory Mixtures Prediction Challenge." 2024.

[2] Y. Han, Z. Wang, S. Yang, and A. Tewari, "Advancing odor mixture discriminability with pretrained embeddings and boosting," in *Presented at the RSGDREAM 2024*, Madison, WI, USA, Oct. 2024.

[3] Y. Tian*, Y. Han*, X. Chen*, W. Wang, and N. V. Chawla, "Beyond answers: Transferring reasoning capabilities to smaller LLMs using multi-teacher knowledge distillation," in *Proc. ACM Int. Conf. Web Search and Data Mining*, 2025.