Xikun Zhang

443 Via Ortega, Stanford, CA 94305·xikunz2@cs.stanford.edu

RESEARCH GOALS

I work at the intersection of **biomedicine** and **AI**. I develop **machine learning models** for **multi-omics data**, with the goal of improving our understanding of the **mechanism of diseases**, like cancer, and enabling **personalized treatment**.

GOOGLE SCHOLAR CITATIONS

5766 (Up to Nov 19, 2024)

FIRST-AUTHOR PUBLICATIONS

1. ***Xikun Zhang***, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec, "*GreaseLM: Graph REASoning Enhanced Language Models for Question Answering,*" ICLR 2022 (**Spotlight**, Top 5%) [paper]
2. ***Xikun Zhang****, Deepak Ramachandran*, Ian Tenney, Yanai Elazar, and Dan Roth, "*Do Language Embeddings Capture Scales?*" Findings of EMNLP 2020 & EMNLP BlackboxNLP workshop 2020 [paper]

OTHER PUBLICATIONS

3. Charlotte Bunne, …, ***Xikun Zhang***, …, and Stephen R. Quake, "*How to Build the Virtual Cell with Artificial Intelligence: Priorities and Opportunities*", To appear in Cell [paper]
4. Jiayuan Mao*, Xuelin Yang*, ***Xikun Zhang,*** Noah Goodman, and Jiajun Wu, "*CLEVRER-Humans: Describing Physical and Causal Events the Human Way,*" NeurIPS Datasets and Benchmarks Track 2022 [paper]
5. Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, ***Xikun Zhang***, Christopher D Manning, Percy Liang*, and Jure Leskovec*, "*Deep Bidirectional Language-Knowledge Graph Pretraining,*" NeurIPS 2022 & AAAI DLG workshop 2023 (**Best Paper Award**) [paper]
6. Rishi Bommasani, …, ***Xikun Zhang***, …, and Percy Liang (116 authors), "*On the opportunities and risks of foundation models,*" arXiv 2021 [paper]
7. Srijan Kumar, ***Xikun Zhang***, and Jure Leskovec, "*Predicting Embedding Trajectories for Temporal Interaction Networks*," KDD 2019 (**Oral**, Top 9%) [paper]
8. Carl Yang, Mengxiong Liu, Frank He, ***Xikun Zhang***, Jian Peng, and Jiawei Han, "*Similarity Modeling on Heterogeneous Networks via Automatic Path Discovery*," ECML-PKDD 2018 [paper]
9. Qi Li*, Meng Jiang*, ***Xikun Zhang***, Meng Qu, Timothy Hanratty, Jing Gao, and Jiawei Han, "*TruePIE: Discovering Reliable Patterns in Pattern-Based Information Extraction*," KDD 2018 (**Long Presentation**, Top 11%) [paper]

EDUCATION

Stanford University
Ph.D. in Computer Science — Sep 2020 - present
M.S. in Computer Science — Sep 2020 - Aug 2023

- Research interests: machine learning, large language models, AI interpretability, AI safety, diffusion models, foundation models, computational biology, single-cell and spatial omics

University of Illinois at Urbana-Champaign
B.S. in Computer Science — Aug 2016 - May 2019

- GPA: 3.98/4.00, major GPA: **4.00/4.00**
- Honors: Bronze Tablet (highest undergraduate honor**, top 3%** in college, final year), James Scholar (top 5% in department, every semester), Dean's List (top 5% in college, every semester)
- Classes: **A+**'s in Machine Learning (**top 1** in a class of 300 people), Advanced Information Retrieval, Introduction to Data Mining, and Random Processes

Nanjing University
B.S. in Electronic Information Science (Transferred out) — Aug 2015 - Jul 2016

- GPA: 3.76/4.00, major GPA: 3.84/4.00
- Rank: **1/203**

RESEARCH EXPERIENCES

AI Resident, Chan Zuckerberg Initiative — Jan 2024 - present
AI-powered virtual cell models [Cell]

- Among one of the first two AI resident academic groups at CZI, with the goal of building AI-powered virtual cells to help scientists explore the molecular underpinnings of human health and disease.

Graduate Research Assistant, Stanford Lundberg Lab, Prof. Emma Lundberg Dec 2022 - present

Generative models of spatial proteomics images

- Proposed and built a novel deep learning model based on state-of-the-art latent diffusion models and variational autoencoders to generate spatial proteomics images at the subcellular resolution, which enables in-silico profiling of subcellular protein localization to understand disease mechanisms and screen drugs in a high-throughput manner.

Graduate Research Assistant, Stanford Newman Lab, Prof. Aaron Newman Dec 2022 - present

Cellular deconvolution at the single-cell resolution

- Proposed and built a novel deep learning model that can deconvolve bulk tissue transcriptomics profiles into single-cell RNA sequencing data, which enables scaling up the power of single-cell biology in the clinic and downstream development of next-generation diagnostics and personalized treatment of diseases like cancer.

Graduate Research Assistant, Stanford Kundaje Lab, Prof. Anshul Kundaje Aug 2021 – Nov 2022

Chromosome scale deep learning models of gene expression

- Proposed graph neural networks to predict cell-type resolved gene expression from DNA sequences spanning megabase-scale regulatory domains by automatically learning cis-regulatory syntax of regulatory elements and long-range regulatory interactions.

Sequence-based imputation models of human epigenomic data at base resolution

- Developed novel deep learning architectures that map 1-5kb DNA sequence inputs to base-resolution TF ChIP-seq profiles jointly across different kinds of TFs and cell types in one model.

Graduate Research Assistant, Stanford Bassik Lab, Prof. Michael Bassik Aug - Nov 2021

Synthetic lethality prediction using the gene coessentiality network

- Designed new graph-neural-network-based deep learning architectures to learn gene embeddings using the gene coessentiality network.
- Used the learned model and the gene embeddings to predict which gene pairs are synthetic lethal.

Graduate Research Assistant, Stanford Network Analysis Project group, Prof. Jure Leskovec Apr - Oct 2021

GreaseLM: Graph REASoning Enhanced Language Models for Question Answering [ICLR 2022 (**Spotlight**)]

- Proposed GreaseLM, a new question-answering model that enhances pretrained language models with graph neural networks reasoning over an external knowledge graph and lets the two components interact with each other over multiple layers of modality interaction operations.
- Showed that GreaseLM boosted performance on answering questions in domains that need external knowledge, like commonsense reasoning, or biomedical question answering.

Deep Bidirectional Language-Knowledge Graph Pretraining [NeurIPS 2022 & AAAI DLG workshop 2023 (**Best Paper Award**)]

- Proposed Dragon, a self-supervised method to pretrain a deeply joint language-knowledge foundation model (GreaseLM) from text and knowledge graphs at scale.
- Pretrained this model by unifying two self-supervised reasoning tasks, masked language modeling and KG link prediction.
- Showed that Dragon outperforms existing language models and language model+knowledge graph models on diverse downstream tasks including question answering across general and biomedical domains, with +5% absolute gain on average.

Graduate Research Assistant, Stanford CogAI group & Stanford CoCoLab, Prof. Jiajun Wu & Prof. Noah Goodman Aug 2020 - Mar 2021

CLEVRER-Humans: Describing Physical and Causal Events the Human Way [NeurIPS Datasets and Benchmarks Track 2022]

- Constructed the CLEVRER-Humans benchmark, a video reasoning dataset for causal judgment of physical events with human labels and demonstrated its challengingness for existing question-answering models.
- Employed two techniques to improve data collection efficiency: first, a novel iterative event cloze task to elicit a new representation of events in videos, which we term Causal Event Graphs; second, a data augmentation technique based on neural language generative models.

AI Resident, Google Research, Prof. Michael Collins Oct 2019 - Aug 2020
Modeling Non-linguistic Contexts for Language Understanding
- Pretrained multimodal models via self-supervised learning on large-scale image caption datasets (scale of millions) on hundreds of TPUs, which powers heterogeneous downstream vision-and-language tasks with the learned joint multimodal embeddings.
- Applied this model to significantly improve matching between dish photos and menu items on Google Maps restaurants through jointly modeling dish photos, user reviews, and menus.

Research Intern, Google Research, Prof. Dan Roth, Ian Tenney Jun - Sep 2019
Do Language Embeddings Capture Scales? [Findings of EMNLP 2020 & EMNLP BlackboxNLP workshop 2020]
- Showed that pretrained language models capture a significant amount of information about the scalar magnitudes of real-world objects but are short of the capability required for general common-sense reasoning.
- Identified contextual information in pre-training and numeracy as two key factors affecting their performance and showed that a simple method of canonicalizing numbers can significantly affect the results.

Research Assistant, Stanford Network Analysis Project group, Prof. Jure Leskovec May - Aug 2018
Predicting Embedding Trajectories for Temporal Interaction Networks [KDD 2019 (**Oral**)]
- Proposed a coupled recurrent neural network model called JODIE that learns dynamic embeddings of users and items from a sequence of temporal interactions.
- Proposed the t-Batch algorithm which can create temporally coherent mini-batches of interactions, making JODIE highly scalable and 9 times faster than existing co-evolution models.
- Conducted experiments to show that JODIE achieved state-of-the-art results on two prediction tasks---future interaction prediction and state change prediction.

Research Assistant, UIUC Data Mining Group, Prof. Jiawei Han Sep 2017 - Apr 2018
Similarity Modeling of Heterogeneous Networks via Automatic Path Discovery [ECML-PKDD 2018]
- Proposed an algorithm called Autopath that can automatically discover useful paths to find similar pairs of nodes using both structure and content information from a heterogeneous network by combining reinforcement learning and deep embedding into a novel semi-supervised joint learning framework.

TruePIE: Discovering Reliable Patterns in Pattern-Based Information Extraction [KDD 2018 (**Long Presentation**)]
- Proposed a novel method, called TruePIE, that finds reliable text patterns to not only extract information related to a certain task based on the statistics of the patterns' individual contents (e.g., length, frequency), but also correct information from a text corpus.
- Adopted a self-training framework that repeats the training-predicting-extracting process to gradually discover more and more reliable patterns.

Research Assistant, UIUC KnowEnG BD2K center, Prof. Saurabh Sinha May - Aug 2017
Genomic Knowledge Network Construction
- Automatically constructing the KnowEnG Knowledge Network, a massive heterogeneous network composed primarily of genes and their annotations as well as their mutual relationships, by identifying online relevant knowledge bases, and automatically cleaning and consolidating those disparate resources.
- Integrating and fusing heterogeneous, high-dimensional, and noisy biological data.

PROFESSIONAL ACTIVITIES

- Top 5% teaching assistant across the Stanford computer science department during fall 2023 (Head TA in CS224W)
- Organizing a workshop "Tools for assembling the cell: Towards the era of cell structural bioinformatics" at the Pacific Symposium on Biocomputing (PSB) 2024
- Invited talk at Amazon AI Lab on the GreaseLM paper (Apr 2022)
- Conference reviewer: AAAI, AACL-IJCNLP
- Journal reviewer: BMC Bioinformatics, IEEE Transactions on Computational Social Systems
- Workshop reviewer: ACL CSRR workshop, CIKM FedGraph workshop, AAAI-GCLR workshop, AAAI-UDM workshop

SKILLS

- Programming languages: Proficient in Python, C / C++, Java, OCaml, Javascript, R
- Machine learning frameworks and data science packages: PyTorch, PyTorch Lightning, TensorFlow, Caffe, Keras, scikit-learn, numpy, scipy, pandas
- Database Management Systems: MySQL
- Web development technologies: HTML, CSS, Javascript, JQuery, Bootstrap, Python Flask