Yinkai Wang

E-mail:yinkai.wang@tufts.edu| Phone: (857)-366-1858 | Google Scholar

Personal Website:https://yinkaiw.github.io/

EDUCATION

Tufts University, the US **09/2022-07/2027**

PhD of Science in *Computer Science*

Tufts University, the US **09/2022-08/2024**

Master of Science in *Computer Science*

George Mason University, the US **08/2018-08/2021**

Bachelor of Science in *Computer Science*

- Honors/Awards: Dean's List (2018-2020)

Huaqiao University, China **08/2017-08/2022**

Bachelor of Science in *Computer Science*

SELECTED PUBLICATIONS

- **Yinkai Wang**, Jiaxing He, Yuanqi Du, Xiaohui Chen, Jianan Canal Li, Li-Ping Liu, Xiaolin Xu, Soha Hassoun. *Large Language Model is Secretly a Protein Sequence Optimizer*. Research paper on Arxiv.
- **Yinkai Wang**, Xiaohui Chen, Liping Liu, Soha Hassoun. *MADGEN - Mass-Spec attends to De Novo Molecular generation.* Research paper for ICLR.
- Xiaohui Chen, **Yinkai Wang**, Jiaxing He, Yuanqi Du, Xiaolin Xu, Soha Hassoun, Liping Liu. *Graph Generative Pre-trained Transformer.* Research paper for Arxiv.
- Xiaohui Chen, **Yinkai Wang**, Yuanqi Du, Soha Hassoun, Liping Liu. *On Separate Normalization in Self-supervised Transformers.* Research paper for NeurIPS.
- **Yinkai Wang***, Yanqiao Zhu*, Yuanqi Du*, Jieyu Zhang, Qiang Liu, Shu Wu. *A Survey on Deep Graph Generation: Methods and Applications.* Survey on LoG.
- Yuanqi Du, Xiaojie Guo, **Yinkai Wang**, Amarda Shehu, Liang Zhao. *Small Molecule Generation via Disentangled Representation Learning.* Research paper for Bioinformatics.
- Shiyu Wang, Xiaojie Guo, Xuanyang Lin, Bo Pan, Yuanqi Du, **Yinkai Wang**, Yanfang Ye, Ashley Ann Petersen, Austin Leitgeb, Saleh AlKhalifa, Kevin Minbiole, Bill Wuest, Amarda Shehu, Liang Zhao. *Multi-objective Deep Data Generation with Correlated Property Control.* Research Paper for NeurIPS 2022.
- Yuanqi Du, **Yinkai Wang**, Fardina Alam, Yuanjie Lu, Xiaojie Guo, Liang Zhao,Amarda Shehu. *Deep Latent-Variable Models for Controllable Molecule Generation.* Research Paper for BIBM 2021.
- Fahim Faisal, **Yinkai Wang**, Antonis Anastasopoulos. *Dataset Geography: Mapping Language Data to Language Users.* Research paper for ACL 2022 Theme Track.

SELECTED RESEARCH EXPERIENCES

Large Language Model is Secretly a Protein Sequence Optimizers **11/2024-01/2025**

Research Assistant(Advisor: Soha Hassoun) *Medford, US*

- Developed a novel evolutionary method using large language models (LLMs) for protein sequence optimization, demonstrating that LLMs can propose high-fitness protein variants without fine-tuning.
- Outperformed traditional evolutionary algorithms in multiple optimization tasks, including single-objective, constrained, and multi-objective optimization, across datasets like GB1, TrpB, and Syn-3bfo.
- Showed that LLMs can efficiently guide protein optimization by proposing new sequences through mutation and crossover, even under experimental budget constraints, enhancing the efficiency of protein engineering experiments.

On Separate Normalization in Self-supervised Transformers **06/2023-09/2023**

Research Assistant(Advisor: Soha Hassoun) *Medford, US*

- Unveiled a novel Separate Normalization (SepNorm) method, employing distinct normalization layers for tokens and the [CLS] symbol in self-supervised transformers, overcoming the constraints of the traditional Shared Normalization (ShareNorm).
- SepNorm boosted the encoding abilities of the [CLS] symbol, leading to an average 2.7% performance improvement across diverse domains. It ensured more uniformly distributed [CLS] embeddings, enhancing global contextual information encoding and representational power.
- Provided empirical evidence showcasing SepNorm's effectiveness and superiority over ShareNorm. It enhanced the uniformity of [CLS] embeddings, especially with Batch Normalization (BN), correlating to improved performance in downstream tasks.

Spectra Discovery with Deep Learning **09/2022-09/2023**

Research Assistant(Advisor: Soha Hassoun) *Medford, US*

- Developed MADGEN, a two-stage deep learning framework leveraging contrastive learning for scaffold retrieval and attention-based generative modeling for spectra-guided molecular generation.
- Improved scaffold prediction accuracy and molecular generation interpretability, achieving strong results on benchmark datasets such as NIST23, CANOPUS, and MassSpecGym.
- Engineered a novel graph-based neural network integrating mass spectra conditioning to reduce search complexity and enhance scaffold-based chemical space exploration.

Multilingual Geospatial Language Expression Discovery **05/2021-03/2022**

Research Assistant (Advisor: Antonios Anastasopoulos) *Fairfax, US*

- Spearheaded the development of a sophisticated system leveraging parallel multilingual datasets. Utilized the mBert model in conjunction with alignment techniques to predict the consistency across multilingual datasets. This approach ensured that the system was robust and adaptable to various linguistic nuances, enhancing its accuracy and reliability.
- Introduced the concept of cross-lingual consistency, elucidating it as the essential characteristic where parallel sentences across two languages, ideally using identical named entities, are indeed tagged with the same entities. This definition not only provided clarity but also underscored the pivotal role of consistency in Named Entity Recognition (NER) tasks, establishing a benchmark for evaluating multilingual systems.
- Delved deep into the realm of cross-lingual consistency within the context of multilingual NER models. Through rigorous research and experimentation, highlighted the paramount importance of parallel data spanning multiple languages. Demonstrated that such data is indispensable for a comprehensive and accurate evaluation of NER models, ensuring they are truly multilingual in their capabilities.

Deep Latent-Variable Models for Controllable Molecule Generation **08/2021-09/2021**

Research Assistant (Advisor: Dr. Amarda Shehu) *Fairfax, US*

- Championed the proposal and development of cutting-edge deep latent-variable models specifically tailored for the generation of small molecules. These models were meticulously designed to produce molecules with targeted molecular properties, ensuring that the generated compounds met specific criteria and exhibited desired characteristics.
- Innovatively structured the models to operate under the principles of supervised and disentangled representation learning. Integrated graph representation learning techniques to effectively capture the inherent constraints present within the vast chemical space. Furthermore, introduced an inductive bias, a strategic approach that bridges the gap between the chemical and biological domains, ensuring a holistic understanding and representation of molecules.
- Undertook a comprehensive evaluation process to validate the efficacy and potential of the developed models. The results underscored their promise in revolutionizing controllable molecule generation. These findings highlighted the models' potential impact in various fields, notably cheminformatics and drug discovery, paving the way for advancements in these critical domains.

SELECTED PROFESSIONAL EXPERIENCE

Tufts University **09/2022-Present**

PhD Researcher

- Applied advanced Machine Learning algorithms to address complex biological challenges. This involved leveraging state-of-the-art techniques to decipher intricate biological phenomena, ensuring a seamless integration of computational methodologies with biological insights.
- Collaborated with a multidisciplinary team of biologists, data scientists, and software engineers. Played a key role in project design and implementation, fostering a cohesive team environment and ensuring that research objectives were met in a timely and efficient manner.
- Developed custom ML models tailored to specific biological datasets. These models were meticulously designed to handle the nuances of biological data, ensuring accurate predictions and providing meaningful insights into the underlying biological systems.

JD.com **12/2021-06/2022**

Research Intern

- Worked on text generation with a primary focus on creating text outputs that seamlessly blend with human-written content. This involved utilizing advanced algorithms and iterative training processes to refine the generated text, ensuring it met the desired quality standards.
- Engaged with the Unilm and LDA models to extract topics from customer reviews. This extracted information was then utilized to craft personalized written content for each customer, enhancing the user experience and ensuring content relevancy.
- Collaborated closely with Dr. Xiaojie Guo and Dr. Lingfei Wu on various research initiatives. This collaboration involved regular brainstorming sessions, project discussions, and feedback loops, ensuring the alignment of research objectives and outcomes.

Peking University **11/2021-08/2022**

Researcher in VDIG lab

- Delved deep into object detection, a pivotal computer vision technique aimed at pinpointing instances of objects within images or videos. This research not only involved the application of established methodologies but also led to innovative outcomes, such as the introduction of the "objection" concept within object detection algorithms. This concept enhanced the interpretability of detection results.
- Leveraged the power of self-supervised learning (SSL) combined with transformer architectures to enhance object detection in images. This approach capitalized on unlabeled data, harnessing its potential to train models more effectively. The integration of transformer structures further improved the spatial understanding of the model, leading to more accurate and granular object detections.

Bytedance **04/2021-07/2021**

Intern in DevEco

- Concentrated on the foundational layer of Bytedance's host Android app, a critical component with intricate coupling relationships to a majority of Bytedance's applications. This involved understanding the app's architecture, dependencies, and ensuring seamless integration and communication between the host app and other dependent apps.
- Pioneered the creation of a mock setting environment, a simulated platform tailored to facilitate QA testing. This innovative approach streamlined the testing process, reducing potential bottlenecks and significantly accelerating the publishing timeline, leading to more efficient release cycles.
- Championed the enhancement of the development environment, specifically catering to developers working on microapps within Bytedance. This initiative involved optimizing tools, workflows, and collaboration platforms, ensuring a productive and comfortable workspace that fostered creativity and efficiency.

SERVICES

- **Reviewer**: AAAI-DLG'22, KDD- DLG'22, BIOKDD'22, NeurIPS Datasets and Benchmarks'22, ICML-SPIGM'23, NeurIPS'23, NeurIPS Datasets and Benchmarks'23, TGL'23, ICLR'24, ICML'24, ICML-SPIGM'24, NeurIPS'24, NeurIPS Datasets and Benchmarks'24

SKILLS

- **Programming skills**: Python, Java, C, Kotlin, MIPS, Julia
- **Language**: Chinese (native), English
- **Hobbies**: Basketball, Movies
- **Research Interests**: Machine Learning, Deep Learning, Computational Biology, AI for Science, Deep Graph Learning, Natural Language Processing.