Xin (Allen) Wang

805-259-5932 | xin.wang.1@vanderbilt.edu | Personal Website

Education

Vanderbilt University Aug. 2023 – May. 2025
M.S. Computer Science, GPA: 4.0/4.0 *Nashville, TN*

- Thesis Track: *Towards Data-driven Machine Learning for Small Molecule Drug Discovery*

University of California, Santa Barbara Aug. 2019 – Jun. 2023
B.S. Statistics and Data Science, Overall GPA: 3.88/4.0, Major GPA: 3.92/4.0 *Goleta, CA*

Research Experience

Network and Data Science (NDS) Lab Aug. 2023 – Present
Graduate Research Assistant *Vanderbilt University*

- Advisor: Dr. Tyler Derr
- Research Interests: AI for Biochemistry, Data-driven ML, Deep Generative Model

Geometric Intelligence Lab Jan. 2023 – Jun. 2023
Undergraduate Researcher *UC, Santa Barbara*

- Mentor: Dr. Nina Miolane
- Research Interests: Geometric Machine Learning, Manifold Learning

Caves Lab/Data Science Capstone Nov. 2022 – Jun. 2023
Undergraduate Research Assistant *UC, Santa Barbara*

- Sponsor/Mentor: Dr. Eleanor Caves
- Research Interests: Image Processing, Computer Vision

Math Directed Reading Program Jan. 2022 – Jun. 2022
Mentee *UC, Santa Barbara*

- Mentor: Zach Wagner
- Research Interests: Universal Approximation Theory

Publications

1. **Xin Wang**†, Liu Yunchao (Lance)†, Ha Dong†, Rocco Moretti, Yu Wang, Zhaoqian Su, Jiawei Gu, Bobby Bodenheimer, Charles David Weaver, Jens Meiler, and Tyler Derr. "WelQrate: Defining the Gold Standard in Small Molecule Drug Discovery Benchmarking." In Proceedings of the Neural Information Processing Systems Conference, Datasets and Benchmarks Track (**NeurIPS '24**).

2. Wang Yu, Nedim Lipka, Ruiyi Zhang, Alexa Siu, Yuying Zhao, Bo Ni, **Xin Wang**, Ryan Rossi, and Tyler Derr. "Topology-aware Retrieval Augmentation for Text Generation." In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (**CIKM '24**), 2442–2452.

3. Xin Wang, Yu Wang, Liu Yunchao, Tyler Derr. "Data-driven Molecular Augmentation via Diffusion Models." to be submitted to **IJCAI 2025**.

Honors & Awards

Vanderbilt Graduate School Travel Grant	Nov. 2024
Vanderbilt Graduate Fellowship	Sep. 2023
UCSB's Arts and Sciences: Graduate with **College Honors**	Jun. 2023
UCSB Math DRP 2022: **People Choice Award**	Jun. 2022

Services

Subreviewer: NeurIPS 2024, CIKM 2024, SDM 2024, WWW 2024, ECML-PKDD 2024
PC Member: GTA^3-2024, WSDM 2024-MLoG
Board Member: Vanderbilt University Riichi Mahjong League
Mentor: UCSB Mentorship Program

Work Experience

Summer Researcher Jun. 2024 – Aug. 2024

Vanderbilt University, Computer Science Department

- Analyzed the ego-subgraph topology of anomalous nodes across datasets and tried to enhance graph anomaly detection by augmenting local topology using a diffusion model.
- Assisted writing the proposal of Amazon Research Awards: *Generative AI for Graph Anomaly Detection.*

Undergraduate Teaching Assistant Sep. 2022 – Jun. 2023

UC, Santa Barbara, PSTAT Department

- Assisted teaching in Statistical Machine Learning, Regression Analysis, and Big Data Analytics courses.
- Held lab/office hours to help students with homework and coding problems.

Algorithm Engineer Intern Jun. 2022 – Aug. 2022

PING AN TECH, Intelligent City Group

- Mined and analyzed monthly macro-data of cities and trained an LSTM-based model.
- Cleaned the Policy Database used for the Policy Recommendation System.

Algorithm Intern Jun. 2021 – Aug. 2021

BOSERA FUND, Index and Quantitative Department

- Collected and modeled the quantitative data on daily stock transactions.

Technical Skills

Programming Languages: Python, R, C/C++, SAS, SQL
Libraries/Softwares: PyTorch, PyTorch-Geometric, PyMOL, AutoDock, AmberTools
Relevant Coursework: Computational Structural Biochemistry, Graph ML, Geometric ML, Advanced ML, Representation Learning, Stochastic Process, Statistical Computing, Real Analysis

Projects

WelQrate: Defining the Gold Standard in Small Molecule Drug Discovery (Accepted) *Link*

- *Problem*: Existing HTS datasets like MoleculeNet and TDC suffer from issues such as inconsistent chemical representations, undefined stereochemistry, and noisy data, which hinder the effective training of deep learning models for drug discovery.
- *Overview*: Proposed a new standard for benchmarking small molecule drug discovery, with contributions including a rigorously **curated dataset collection**, an **evaluation framework**, and comprehensive **benchmarking**.
- 1) Developed an automatic hierarchical curation pipeline, including multiple filters, follow-up experimental screening to ensure data quality and in-depth assessment of bioassay metadata from the PubChem database.
  2) Generated various data formats (e.g., InChI, SDF, 3D graphs) and introduced specialized split schemes and metrics to assess model performance in prioritizing active molecules under realistic conditions.
  3) Thoroughly benchmarked multiple models and demonstrated that the high-quality WelQrate data improves model identification for unknown, active molecules.

BioML Challenge 2024: Bits to Binders *Link*

- *Problem*: How to Design the binding domains for a Chimeric Antigen Receptor (CAR) to interact with the extracellular region of cancer antigen CD20 and trigger the immune response to kill the cancer cell?
- *Overview*: Developed a large-scale binder design pipeline that integrates backbone generation, inverse folding, and validation steps using SOTA generative models (e.g., RFDiffusion, Chai-1)
- The pipeline includes careful tuning of the generative models and algorithmic implementation to ensure that:
  1) the validation model produces protein-binder complexes with correct binding sites and good iPTM scores, and
  2) the final submission includes binders with diverse 3D structures and sequences.

Data-driven Molecular Augmentation via Diffusion Models (in preparation)

- *Problem*: The HTS dataset faces distribution shifts from scaffold splits and severe class imbalance due to the low hit rate, hindering model generalization and biasing predictions towards inactive compounds.
- *Overview:* Designed a novel learning framework that augments HTS datasets with the awareness of diverse scaffolds and imbalanced class, enhancing the robustness and generalizability of the activity predictors.
- 1) Pre-trained a diffusion model on an unlabeled dataset that covers diverse chemical spaces of drug-like molecules.
  2) Designed a novel algorithm to select training molecules with under-represented scaffold structures and classes.
  3) Slightly perturbed the selected molecular graphs (e.g., edge dropping, node adding) and used the pre-trained diffusion model to generate new, valid molecules.
  4) Augmented the training set with generated molecules using semi-supervised strategies (e.g., pseudo-labeling), enhancing the predictor's decision boundary under the class imbalance and distribution shift conditions.

Learning Molecules as Cellular Complexes (in preparation)

- *Problem*: Can higher-order representations, which incorporate ring structures in molecular graphs, improve molecular property prediction?
- *Overview*: Developed a training pipeline with *TopoModelX* to easily apply cellular complex neural networks, such as the *Cell Attention Network* and *Cellular Isomorphism Networks*, for molecular property prediction tasks.
- We are trying to answer the following research questions:
  1) Do ring-level features enhance the performance of molecular property prediction models?
  2) Do molecular graph classification and regression tasks benefit from the higher-order Weisfeiler-Lehman test?
  3) How do cellular complex neural networks compare to traditional GNNs in terms of performance on molecular datasets?

Identifying and Measuring Visual Acuity Features from 2D Bee Images *Link*

- *Problem*: The traditional method requires labor-intensive manual measurement of ommatidia diameter via microscopy and calculating interommatidial angles from 3D μCT images, which are often unavailable.
- *Overview*: Designed an automated, fast pipeline for measuring ommatidia diameter and interommatidial angles to calculate visual acuity from 2D bee images as part of the NSF-funded Big Bee Project.
- 1) Developed an algorithm to detect ommatidia coordinates and calculated diameters using 2D Fast Fourier Transform, autocorrelation, and image filtering.
  2) Utilized the Segment-Anything model to extract eye masks, followed by an algorithm to derive the eye contour as a geodesic of the eye surface, enabling local curvature modeling and calculation of interommatidial angles.
  3) Applied the pipeline to measure visual acuity in Apidae samples and derived meaningful correlations between visual acuity and biological features, such as sex and ecological roles.

Proving Universal Approximation Theorem *Link*

- Conducted a thorough study on Moshe Leshno et al. *"Multilayer feedforward networks with a nonpolynomial activation function can approximate any function"* and related math problems