Yang Ouyang

Raleigh, NC 27606, United States

(984) 325-2686 youyang7@ncsu.edu LinkedIn GitHub Personal Website

Education Experience

North Carolina State University **Aug. 2024 – Present**

Doctor of Philosophy in Electrical and Computer Engineering *Raleigh, U.S.A*

- **Advisor: Kaixiong Zhou**
- **Research Interests:** AI for biology (Interpretable RNA), Trustworthy AI, Large Language Model Quantization/Pruning

Duke University **Aug. 2022 – May 2024**

Master of Engineering in Electrical and Computer Engineering *Durham, U.S.A*

- GPA: 3.83 / 4.0
- **Teaching Assistant** of ECE 551K: Programming, Data Structures, and Algorithms in C++

Shenzhen University **Sep. 2018 – July 2022**

Bachelor of Engineering in Computer Science and Technology *Shenzhen, China*

- GPA: 3.75 / 4.0
- Honors/Awards: Two times winner of The Second Award of Studying Star in 2020 & 2021 (Ranked in 4 & 6); Outstanding Graduate of the Year 2022

Selected Publication

- **[Proceeding to NAACL 2025] Yang Ouyang**, Hengrui Gu, Shuhang Lin, Wenyue Hua, Jie Peng, Bhavya Kailkhura, Meijun Gao, Tianlong Chen, Kaixiong Zhou. "Layer-AdvPatcher: Layer-Level Self-Exposure and Patch for Jailbreak Defense"
- [**Proceeding to ICLR 2025**] Jingyang Zhang, Jingwei Sun, Eric Yeats, **Yang Ouyang**, Martin Kuo, Jianyi Zhang, Hao Frank Yang, Hai Li. "Min-K%++: Improved Baseline for Detecting Pre-Training Data of LLMs"
- [**Submitted to ICML 2025**] Shuhang Lin, Wenyue Hua, Lingyao Li, **Yang Ouyang**, Jie Peng, Bhavya Kailkhura, Kaixiong Zhou, Tianlong Chen. "Skeletonic Speculative Decoding"

Project Experience

Weakly-supervised Motif Segmentation of RNA Sequence **Oct. 2024 – Present**

Collaborated with MIT Institute for Medical Engineering and Science *Raleigh, U.S.A*

- Developed a **five-step** weakly supervised approach for motif segmentation: i) trained a classification/regression head upon **RNA-FM large pre-trained model** using **Mean Ribosome Load Prediction (MRL)** datasets, ii) used **Class Activation Maps** to generate pseudo-masks, iii) trained a **dedicated segmentation model** using refined pseudo-masks, vi) **validated** the final segmented motifs against existing motif datasets to confirm the method's effectiveness.
- Achieved up to **87% segmentation accuracy** on the **MRL** dataset, with strong generalization to **uORF** and **IRES** datasets, showing **40% Jaccard similarity**. Outperformed **baseline methods (MEME, Homer)** by over **50% in F1 score**.

Translation Efficiency Prediction Using RNA-FM Embedding **Oct. 2024 – Dec. 2024**

Collaborated with Prof. Cranos Williams *Raleigh, U.S.A*

- Developed a classification framework for predicting translation efficiency (TE) from 5' UTR sequences using **RNA-FM embeddings** with a simple Linear head. Evaluated both pretrained and fine-tuned embeddings on the full dataset.
- Designed controlled experiments to compare against traditional handcrafted feature selection pipelines:
  - ∗ Experiment 1–5: selected windows around uORF or main ORF, with/without shuffling or random offsets
- Compared against three baselines:
  - ∗ **Baselines 1:** Feature-engineered inputs from Peiran's window-based strategy
  - ∗ **Baselines 2:** Traditional CNN / LSTM models trained on full 5' UTR sequences
  - ∗ **Baselines 3:** Word2Vec-based embeddings with attention classifier
- Achieved best performance with **RNA-FM + Linear Head** trained on full 5' UTRs (**63.1% accuracy, 0.6799 ROC-AUC**), outperforming all baselines by up to **15% ROC-AUC**, without relying on complex sequence selection.
- Results support the effectiveness of large foundation model embeddings in capturing biologically relevant signals end-to-end, simplifying the traditional pipeline design.

P300 Acetylation Substrate Classification **Oct. 2024 – Present**

Collaborated with NC State Chemical Engineering Department *Raleigh, U.S.A*

- Improved the **large foundation model PeptideBERT** for binary substrate classification by implementing balanced batch sampling, ranking loss, and replacing the classification head with LSTM/ResNet, raising accuracy **from 80.3% to 80.85%** on 5-fold validation. Combined these enhancements with **retrieval augmented generation** to achieve an **additional 1% accuracy improvement.**
- Leveraged the **larger ESM2 650M model** to boost validation accuracy to **88.47%**, and realized further marginal gains by integrating 3D structural data from ESMFold with EGNN embeddings.

Layer-AdvPatcher: Layer-Level Self-Exposure and Patch for Jailbreak Defense **Aug. 2024 – Oct. 2024**

Proceeding to NAACL 2025 *Raleigh, U.S.A*

- Developed the Layer-AdvPatcher framework to defend against jailbreak attacks in LLMs, including a three-step pipeline for defense: i) toxic layer identification, ii) adversarial augmentation, and iii) localized toxic layer editing.
- Achieved a **25% reduction** in Attack Success Rate using our method across models including Mistral-7B and Llama2-7B compared to modification-based defense methods.

Skeletonic Speculative Decoding — May. 2024 – Oct. 2024

Submittd to ICML 2025 — *Raleigh, U.S.A*

- Introduced Skeletonic Speculative Decoding (SSD), a divide-and-conquer approach to enhance speculative decoding efficiency by decomposing complex queries into manageable sub-questions, improving the acceptance rate and enabling parallel processing.
- Achieved up to a **2.8x speedup** over standard speculative decoding methods across various model configurations by leveraging parallelization and increased acceptance rates for simplified sub-questions.

Min-K%++: Improved Baseline for Detecting Pre-Training Data of LLMs — Jan. 2024 – Apr. 2024

Proceeding to ICLR 2025 — *Durham, U.S.A*

- Proposed a theoretically motivated methodology, Min-K%++, for pre-training data detection in LLMs, leveraging local maxima of the modeled distribution to identify training data effectively.
- Achieved new state-of-the-art performance, surpassing existing methods **by 6.2% to 10.5%** in AUROC on the WikiMIA benchmark and performing competitively on the challenging MIMIR benchmark.

Internship Experience

Trip.com Group Ltd | *Java, Spring Framework* — May 2023 – Aug. 2023

Back End Developer Intern, Flight Ticket Department — *Shanghai, China*

- Contributed to the optimization of MegaSearch which serves as an aggregation and cache layer for Trip's international ticket responses using **Java**.
- Optimized the response size to fit AWS's smaller bandwidth while saving some storage costs. Reduced the **Protobuf response size by 50%** in total using a variety of methods.
- Compared a variety of serialization and deserialization means using **JMH**: including the latest open source Fury, Kryo, and ultimately found that Protobuf is the most efficient serialization, but Kryo in the serialization of the size of a small advantage.

Amazon Web Services | *Java, K8s* — July 2022 – Oct. 2022

Back End Developer Intern, DeepJavaLibrary Department — *Mountain View, U.S.A (remote)*

- Integrated DeepJavaLibrary Model Server with KServe by developing 3 robust HTTP APIs in **Java** for KServe's inference engine, supporting DJL-Serving health checks, model information retrieval, and inference result processing with request data, each passing unit tests and providing clear response codes.
- Deployed containerized DJL-Serving on KServe using **yaml** files to configure ports and parameters, facilitating model deployment and testing within the KServe framework.

Tencent Music Entertainment Group | *Javascript, Vue* — May 2021 – Sep. 2021

Front End Developer Intern, Security Center — *Shenzhen, China*

- Applied **Vue2.0** framework based on JavaScript to develop the inner front-end of content audit security platform.
- Developed search, collection, and recently used functions for the middle ground management system.
- Utilised Least Recently Used (LRU) to design a cache that was able to clear the cache efficiently.
- Configured Webpack to optimize the local development and deployment **increased the packaging speed by 75% and decreased the packaging size by 10%.**

Technical Skills

- Programming Languages: Python, Java, C++, Javascript
- Deep Learning Frameworks: PyTorch, Huggingface