Wenduo Cheng

wenduoc@andrew.cmu.edu | wenduocheng.github.io | github.com/wenduocheng

Research Interests

My background spans bioinformatic sequence analysis, evolutionary biology, environmental science, GWAS, and deep learning for computational biology. My current research focuses on applying AI to accelerate biological discovery, particularly through biological foundation models and autonomous agents.

Education

Carnegie Mellon University — *Pittsburgh, PA*

PH.D. in Computational Biology — *Aug 2024 - Present*

- GPA: 4.00/4.33
- Advisor: Dr. Jian Ma

Carnegie Mellon University — *Pittsburgh, PA*

M.S. in Computational Biology — *Aug 2022 - May 2024*

- GPA: 4.06/4.33
- Thesis: Leveraging pre-trained language models to address genomics tasks
- Awards: Research Excellence and Honors; Outstanding Academic Achievement

Duke Kunshan University / Duke University — *Kunshan, China / Durham, NC*

B.S. in Biology / B.S. in Interdisciplinary Studies — *May 2018 - May 2022*

- GPA: 3.9/4.0, *Magna Cum Laude*

Publications

DNALONGBENCH: A Benchmark Suite for Long-Range DNA Prediction Tasks [paper] [code]

Wenduo Cheng†, Zhenqiao Song†, Yang Zhang†, Shike Wang, Danqing Wang, Muyu Yang, Lei Li, Jian Ma

In: Nature Communications *16, p. 10108. 2025.* ***(Editors' Highlights)***

L2G: Repurposing Language Models for Genomics Tasks [paper] [code]

Wenduo Cheng, Junhong Shen, Mikhail Khodak, Jian Ma, Ameet Talwalkar

In: Transactions on Machine Learning Research (TMLR) *2025.*

Specialized Foundation Models Struggle to Beat Supervised Baselines [paper] [code]

Zongzhe Xu, Ritvik Gupta, **Wenduo Cheng**, Alexander Shen, Junhong Shen, Ameet Talwalkar, Mikhail Khodak

In: Proceedings of the 12th International Conference on Learning Representations (ICLR) *2025.*

The Special and General Mechanism of Cyanobacterial Harmful Algal Blooms [paper]

Wenduo Cheng†, Somin Hwang†, Qisen Guo, Leyuan Qian, Weile Liu, Yang Yu, Li Liu, Yi Tao, Huansheng Cao

In: Microorganisms *11(4), p. 987. 2023.*

Draft Genome Sequence of an Epibiotic Bacterium, *Bacillus cereus*, Isolated from Cyanobacterial Blooms in Lake Taihu, China [paper]

Xiaoyuan Chen, Yinuo Yang, Yang Yu, Qisen Guo, Somin Hwang, **Wenduo Cheng**, Huansheng Cao

In: Microbiology Resource Announcements *12(3), e00936–22. 2023.*

Cyanobacterial Blooms Are Not a Result of Positive Selection by Freshwater Eutrophication [paper]

Yang Yu†, **Wenduo Cheng**†, Xiaoyuan Chen, Qisen Guo, Huansheng Cao

In: Microbiology Spectrum *12(3), e03194–22. 2022.*

(† Equal contribution)

Research Experience

The Computational Biology Department at Carnegie Mellon University *Pittsburgh, PA*

Advisor: Dr. Jian Ma *Jan 2023 – Present*

- **Project: Leveraging Pre-trained Language Models for Genomics Tasks**
  (Published in *TMLR*)
  - Designed L2G, a cross-modal fine-tuning framework that repurposes pre-trained language models (e.g., RoBERTa) for DNA sequence prediction without genomic pre-training.
  - Demonstrated that L2G matches or outperforms state-of-the-art DNA foundation models and AutoML baselines on GenomicBenchmarks, Nucleotide Transformer benchmarks, and regulatory activity tasks, while being significantly more data- and compute-efficient.
- **Project: DNALONGBENCH: A Benchmark Suite for Long-Range DNA Prediction Tasks**
  (Published in *Nature Communications*)
  - Built DNALONGBENCH, a benchmark suite covering five biologically meaningful long-range tasks. Curated unified datasets, splits, and evaluation metrics for both classification and regression tasks, enabling systematic comparison of long-range DNA models.
  - Benchmarked lightweight CNNs, task-specific expert models, and DNA foundation models (HyenaDNA, Caduceus), showing that expert models consistently outperform current foundation models and revealing key gaps in modeling long-range dependencies.
- **Project: TissueAgent – A Role-Based Multi-Agent Framework for Spatial Transcriptomics Analysis**
  (Preprint on arXiv, in preparation)
  - Designed TissueAgent, a role-based multi-agent system to automate spatial transcriptomics workflows such as cell type annotation, cell–cell communication analysis, hypothesis generation, and figure reproduction.
  - Showed that TissueAgent flexibly integrates internally developed and external domain-specific agents, enabling agent-agent collaboration for scientific discovery.

The Laboratory of Environmental Sciences at Duke Kunshan University *Kunshan, China*

Advisor: Dr. Huansheng Cao *Dec 2020 - May 2022*

- **Project: The Special and General Mechanism of Cyanobacterial Harmful Algal Blooms**
  (Published in *Microorganisms*)
  - Conducted literature review and wrote a review paper on the formation of cyanobacterial harmful water blooms.
  - Proposed a synthesis of cyanobacterial harmful algal blooms as a result of the synergistic interactions between cyanobacterial superior functions and elevated nutrients.
- **Project: The Role of Positive Selection in the Formation of Cyanobacterial Algal Blooms**
  (Published in *Microbiology Spectrum*)
  - Studied whether cyanobacterial harmful algal blooms are a result of positive selection by water eutrophication.
  - Conducted molecular evolutionary analyses on the genomes of 9 bloom-forming cyanobacteria, combined with pangenomics and metatranscriptomics.
- **Project: Draft Genome Assembly of Bacteria Isolated from Lake Taihu**
  (Published in *Microbiology Resource Announcements*)
  - Isolated and sequenced the whole genome of a bacterial strain collected from Lake Taihu, China.
  - Assembled the draft genomes of isolated bacteria through quality control, read trimming, de novo assembly, genome annotation, and species identification.

The Statistical Genetics Laboratory at Westlake University *Hangzhou, China*

Advisor: Dr. Jian Yang *Jul 2021 - Sep 2021*

- **Project: Genetic variant detection from scRNA-seq data**
  - Built a pipeline to detect single-nucleotide variants from single-cell and bulk RNA sequencing data and DNA sequencing data.
  - Established the feasibility of calling and genotyping SNPs from 10x Genomics scRNA-seq data, laying a solid foundation for future research on finding eQTLs from scRNA-seq.

The Biology Department at Washington and Lee University *Remote*

Advisor: Dr. Natalia Toporikova *May 2020 - Jun 2021*

- **Project: Statistical Analysis of Gene Expression in Spider Using RNA-Seq**
  (Preprint in *Research Square*)
  - Conducted differential expression analysis and functional analysis using R studio to determine pattern of gene expression in spiders in response to collection time and light pulse.
  - Identified 528 differentially expressed transcripts with a flipped pattern of expression, representing prime candidates for light-sensitive circadian genes.

Professional Service

Reviewer	RECOMB 2024, RECOMB 2025, ACM BCB 2025, ISMB/ECCB 2025
Teaching Assistant	Computational Medicine (CMU 02718, 23 fall), Computational Genomics (CMU 02710, 24 spring)

RELATED COURSEWORK

CS	Deep Learning, Natural Language Processing, Multimodal Machine Learning, Algorithms and Data Structure
Math	Probability and Statistics, Linear Algebra, Multivariable Calculus, Random Variables and Stochastic Processes
Biology	Cell and Molecular Biology, Genetics and Evolution, Biophysics, Biochemistry, Microbiology, Population Genomics, Functional Genomics, Molecular Genetic Analysis
Comp Bio	Computational Genomics, Computational Medicine, Biological Modeling and Simulation, Automation of Scientifit Research

Skills

Programming	Proficient in Python, R and Bash; Familiar with Java, Golang, MATLAB, Mathematica and SQL
Miscellaneous	Git, LaTeX, Linux, Shell, Slurm, PyTorch, Scikit learn, VS Code, Jupyter Notebook and Markdown