

MATTHEW ZHANG

✉ mz473@cantab.ac.uk  github.com/MatthewZhang473  matthewzhang473.github.io

EDUCATION

University of Cambridge

Cambridge, UK

BA & MEng Engineering (Distinction & First Class Honours)

October 2021 – June 2025

- Specialized in Information and Computer Engineering with a research focus on Bayesian ML.
- Consistently ranked in the top 5% of the cohort; awarded the Clough Scholarship.
- **Key Modules:** Deep Learning & Structured Data; Probabilistic Machine Learning; Computational Statistics & Machine Learning; Computer Systems & Concurrency.

PUBLICATION

Graph Random Features for Scalable Gaussian Processes

Accepted at ICLR 2026.

Matthew Zhang, Jihao Andreas Lin, Krzysztof Choromanski, Adrian Weller, Rich Turner, Isaac Reid

TL;DR: We enabled graph Gaussian process inference in $\mathcal{O}(N^{3/2})$ time using graph random features (GRFs), making Bayesian optimization feasible on million-node graphs on a single computer chip.

<https://arxiv.org/abs/2509.03691>

RESEARCH EXPERIENCE

Myrtle.ai

Cambridge, UK

Research Engineer

November 2025 – Present

- Developed a Sequential Monte Carlo sampler for the intractable RNN-T alignment posterior, using a cheap surrogate proposal to produce an unbiased, differentiable loss estimator.
- Designed fused GPU kernels (TileLang/CUDA) for the RNN-T joint network that avoid materialising the dominant $[B, T, U, V]$ activation tensor, cutting its memory footprint by $1000\times$ and substantially reducing HBM traffic.
- Contributed to Myrtle's production ASR training & inference pipeline: [Caiman-ASR](#).
- **Keywords:** ASR, ML infra, Monte Carlo methods.

University of Cambridge, Computational and Biological Learning Lab

Cambridge, UK

Research Assistant, supervised by Prof. Rich Turner

October 2024 – October 2025

- Designed a scalable inference algorithm for GPs on graphs, reducing complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^{3/2})$ and enabling inference on networks with 10M+ nodes.
- Integrated sparse linear algebra and randomized feature maps into the GP framework, achieving over $100\times$ speedups versus standard implementations.
- Drafted a first-author manuscript, accepted at ICLR26.
- **Keywords:** GPs, Monte Carlo methods, Random features.

ACADEMIC SERVICE

Conference Reviewer

- Reviewer for NeurIPS 2026, evaluating submissions in probabilistic ML and generative modeling.

WORK EXPERIENCE

Microsoft

Cheltenham, UK

Software Engineering Intern

July – September 2024

- Researched and deployed a scalable graph ML algorithm for threat detection, processing data from 30M+ Azure cloud apps daily.
- Developed distributed ML pipelines using Azure ML to accelerate large-scale data processing.
- **Tech Stack:** C#, .NET, KQL, PySpark, Pandas, Graph neural networks (GNNs).

Playfair Technologies

Remote

Machine Learning Engineer

November 2022 – May 2023

- Built end-to-end NLP pipelines for financial sentiment analysis on news; fine-tuned open source Transformer models, improving accuracy by 20%+.
- Developed CV models to extract sentiment signals from trading-chart images, expanding training data 30× via augmentation (image transforms, noise injection) to improve robustness.
- **Tech Stack:** PyTorch, Transformers, CNNs, OpenCV.

Huawei

Cambridge, UK

CPU Architecture / Software Engineering Intern

June – September 2022

- Automated generation of machine-readable ISA specifications in [Sail](#).
- Deployed a Kubernetes cluster to run containerized applications and streamline CI/CD pipelines.
- **Tech Stack:** Kubernetes, Docker, Jenkins, CI/CD, Linux, Git.