LINFENG WANG

Email: wanglinfeng1115@gmail.com
Mobile: +44 75 996 108 97
LinkedIn: https://www.linkedin.com/in/w15/
Github: https://github.com/linfeng-wang

PERSONAL SUMMARY

PhD candidate (2 months to completion) in computational genomics at LSHTM, with expertise in machine learning applied to. Author of 4 peer-reviewed papers and lead contributor to a TB diagnostic assay adopted in public-health settings with four more forthcoming. Presented work at 2 international conferences. Totalling 6 first author publication expected. I'm passionate about translating data-driven insights into innovation and products.

TECHNICAL SKILLS

Programming: Python, R, Bash, MySQL, HTML/CSS, C++
Data science: Pytorch, PyG, Fastai, scikit-learn, NumPy, Pandas, Matplotlib, Scipy, ggplot2, tidyverse, Jax
Specific genomic tools: BWA-mem, samtools, bcftools, Rxml, Freebayes, Beast2, Figtree, trimmomatic, iTol, Plink2, GATK, Nextflow

EDUCATION

London School of Hygiene and Tropical Medicine , DPhil Computational Genomics **Oct 2021 – Oct 2025**
Dissertation: *Machine Learning-Enhanced Drug Resistance and Bioinformatics Transmission Profiling of Tuberculosis Using Genome Sequencing*

Imperial College London, MRes Bioengineering (Hons), Merit **Oct 2019 – Oct 2020**
Modules: Computational & Statistical Methods for Research, Frontiers in Bioengineering, Biomaterials
Dissertation: *Design of an Artificial Bruch's Membrane from Synthetic Polyesters.*

King's College London, BSc Biochemistry (Hons), 1st **Sep 2016 – Jul 2019**
Module selected: Bioinformatics, Protein structure and design, Human genomics.
Dissertation: *Investigation of Concordance Between Molecular Dynamics Simulation and FRET Biosensor using Designed Protein Linker System.*

EXPERIENCE

PhD research, LSHTM (Clark Campino Phelan lab), London — Jul 2022 – Sept 2025

- **Statistics:** Use **python** to analyse data to create a statistical report on Philippine TB transmission and drug resistance (**Python data handling, data visualisation, linear regression,** GWAS, PCA, Odds ratio, Chi-square)**Deep learning** (**Pytorch, Jax**)**:** Design neural network models that predicts for drug resistance from DNA sequence; Explainable AI for drug resistance level prediction, and protein generation (**CNN, RNN, GNN, Transformers**).
- **Machine learning** (**Sklearn, Statsmodel**)**:** Create command-line based tool that analyse drug resistance based on **Gaussian Mixture** modelling and **XGBoost**.
- **Coding:** Genetic algorithm-based tool design for gene amplicon sequencing design using **python** & **Nextflow.**
- **Leverages resources:** While building neural networks, with GPU power and usage in mind I adjusted my input from DNA sequence to SNP frequency, effectively reducing model size and training time while preserving accuracy.

Machine learning consultant, Deep Science Venture, London — Mar 2025 -

- **Develop deep learning models: (CNN, RNN, VAE)** using **PyTorch** with data augmentation and hyperparameter tuning for sequence-based classification and generation on biological datasets.
- **Appy interpretability methods:** (**SHAP, LIME, DeepLIFT**) to extract biologically meaningful patterns and support AI-driven discovery**.**

Data study group hackathon, Turing Institute, London. — Jan 2024

- **Data-Driven Hazard Detection:** Built and validated deep learning models to identify shallow gas in 40+ years of seismic data, supporting offshore energy planning.
- **Computer Vision:** Deployed CNNs with contrastive learning for classification, detection, and segmentation on legacy geophysical imagery.
- **Team Communication & Adaptability:** Collaborated across disciplines in a high-pressure two-week sprint, balancing exploration with results.
- **Rapid Problem-Solving:** Tackled real-world geoscience challenges beyond core domain, delivering actionable ML solutions under tight deadlines.

PhD Placement, Linkgevity, London UK, — Aug 2024 – Oct 2024

- **Python and Machine Learning:** Built and refined MLP and **GNN** models in Python to predict drug-drug interactions.
- **Information Synthesis:** Reviewed and integrated literature to guide model design and optimization.
- **Accountability and Benchmarking:** Led model benchmarking to identify top-performing architectures and improve predictive accuracy.
- **Strategic Application:** Applied optimized models to real-world compounds, supporting pipeline decisions in drug development.
- **Adaptability and Industry Insight:** Operated effectively in a fast-paced startup, gaining exposure to diverse interaction mechanisms and development workflows.
- **Cloud Computing Efficiency:** Deployed and scaled ML pipelines on **Google Cloud Platform** clusters for efficient training and evaluation.

PhD Rotation Project, LSHTM, (Silver lab), London — Mar 2021 – July 2022

- **Statistics**: Analysed epigenetic data in R using PCR quantification, linear regression, correlation testing, and enrichment analysis.
- **Epigenomics**: Identified methylation patterns linked to foetal development and phenotypic variation through rigorous statistical profiling.
- **Data Quality & Reproducibility**: Conducted stringent quality control, plate adjustment, and data curation to ensure reproducibility and compliance with peer-review standards.

VoyagerX internship, ByteDance, London — Aug 2022 – Feb 2023

- **SQL and Python**: Built a chemical database and applied autoencoder models to integrated biological datasets, using visual analytics to reveal neoantigen-related insights.
- **Responsibility and Demonstrates Accountability**: Led a cross-disciplinary market research project, coordinating regular team check-ins and tracking deliverables.
- **Gather, Interpreting and Organizing Information**: Compiled and structured large datasets from literature and public sources to support machine learning-driven prediction tasks.
- **Presentation of complex ideas**: Authored bi-weekly scientific digests presentation linking research trends with broader implications, distilling complex findings into accessible insights.
- **Determines Tasks and Resources**: Scoped and resourced end-to-end data science workflows—data collection, integration, and ML deployment using Python.

Masters' Research project, Imperial College London (Stevens lab), London — Oct 2019 – Sep 2020

- **Polymer Chemistry**: Synthesized polymers via ROP and optimized electrospinning for fibrous scaffolds supporting RPE transplantation; validated structures with SEM and NMR.
- **Computational Simulation**: Simulated polymer degradation using **Multiphysics models;** visualized degradation kinetics in Python and cryoEM data via ImageJ.
- **Follows Procedures**: Executed polymer synthesis and scaffold fabrication under strict protocol adherence to ensure reproducibility and structural integrity.
- **Translational Teamwork / Involving Others**: Facilitated cross-disciplinary collaboration across chemistry, bioengineering, and cell biology to design clinically viable, degradable scaffolds.

Undergraduate Project, King's College London (Beavil lab), London — Sep 2018 – Jan 2019

- **Programming:** Simulated protein dynamics computationally using techniques including **R, python, molecular dynamic simulation, Bash script**
- **Protein design:** Designed proteins through molecular cloning, plasmid DNA purification, protein purification, TCSPC spectroscopy and molecular dynamic simulations, Gel electrophoresis, PCR
- **Proactively identifies problems and solutions:** During our **protein dynamics simulations**, I proactively developed a Bash script to automate the analysis of molecular dynamic simulation data.

TEACHING

- **Python coding** – (London School of Hygiene and Tropical Medicine)
  - Python coding and use of data science packages, teaching master's degree modules
- Pathogen computational Genomics workshop (UK, Philippines, Indonesia, Thailand)
  - Extensive use of terminal command (**Bash**) and tools spanning all bioinformatics processes in DNA
  - **MCMC, Bayesian, Smith-waterman, Needleman-Wunsch algorithm**
  - **Design and teaching workshop for Deep learning**
- *Sysmic* statistics with python course teaching assistant
  - Running Q&A tutorial workshops

PUBLICATIONS

- *Mixed infections in genotypic drug-resistant Mycobacterium tuberculosis*. Nature *Scientific Reports*. (2023).
  - Applied custom Python pipelines and **Gaussian mixture model** to detect multi-strain *Mtb* infections from whole-genome sequencing data.
- *TB-ML: A framework for comparing machine learning approaches to predict drug resistance of Mycobacterium tuberculosis*. *Bioinformatics Advances*. (2023).
  - Built a **modular evaluation framework for ML classifiers** (Random Forest, CNN, GCN) using **PyTorch and scikit-learn**; optimized performance on sparse mutation matrices.
- *Whole genome sequencing analysis of Mycobacterium tuberculosis*: *strain types and resistance mutations in the Philippines*. Nature *Scientific Reports*. (2024).
  - Led data processing for 729 TB genomes; built phylogenetic trees, extracted resistance markers, and implemented **metadata integration in Python**.
- *TGV: suite of tools to visualize transmission graphs*. *NAR Genomics and Bioinformatics*. (2024)
  - Developed a **Python-based visualization toolkit** for transmission graph analytics using network theory and metadata overlays.
- Data Study Group Final Report: British Geological Survey - Detecting Shallow Gas from Marine Seismic Images". The Alan Turing Institute (2025)
  - Built a **computer vision** model for detecting shallow gas pockets from legacy seismic images, performing tasks including **classification, object detection and segmentation**, achieve classification accuracy of 90%.
- *A novel tool for designing targeted gene amplicons and an optimised set of primers for high- throughput sequencing in tuberculosis genomic studies, BioRxiv, (submitted for publication).*
  - Designed and automated primer selection **software using Python**, thermodynamic modelling, and alignment scoring; validated in vitro.
  - **Website:** https://genomics.lshtm.ac.uk/webtoast/#/ **software package**: https://pypi.org/project/toast-amplicon/
- *LSTM-Based Transfer Learning Models for Tuberculosis-Targeted Antimicrobial Peptide Classification and Generation*. (forth coming).
  - Developed **LSTM-based transfer learning** models to **classify** and **generate** antimicrobial peptides, combining pretrained protein embeddings with custom fine-tuning workflows in **PyTorch**.
- *Deep Learning Approaches for MIC Prediction in Tuberculosis: Addressing Cryptic Resistance and Data Imbalance*. (forth coming).

- Built **CNN** to predict minimum inhibitory concentrations (MICs) in **Pytorch**, using **binary supported learning** to handle imbalance and predict continuous drug resistance rather than binary.

- *Stepwise Prediction of Tuberculosis Treatment Outcomes Using XGBoost and Feature-Level Analysis: A Multi-Stage Approach to Clinical Decision Support*. (forth coming).
  - Applied **XGBoost** models with staged feature integration to predict patient outcomes from clinical and demographic data; emphasized interpretability through feature importance metrics.

- *Decoding Positive Selection in Mycobacterium tuberculosis with Phylogeny-Guided Graph Attention Models*. (forth coming).
  - Designed a **graph neural network (Graph attention neural network)** using phylogenetic trees as input graphs to identify adaptive genetic signals, integrating evolutionary structure into attention-based learning in **pyG**.

LEADERSHIP

LiDo Student Committee – LIDo London **April 2024 – April 2025**

- **Survey design:** Gather student opinion and come up with questionaries for student well-being for programme improvement
- **Event Organising**: Designing activities and inviting speakers for annual 3-day programme retreat, comprising of ~300 attendants

AWARDS

- UKRI – BBSRC – LiDo PhD scholarship (2021-2025)
- Wellcome Trust Biomedical Vacational Studentship. (2018)
- Imperial Award for personal development (2020)

INTERESTS

- Judo, diving, basketball, guitar, coding,