

LINFENG WANG

wanglinfeng1115@gmail.com ♦ +44 7599610897 ♦ linkedin.com/in/w15 ♦ linfeng-wang.github.io/

TECHNICAL STRENGTH

Language & Frameworks:	Python, PyTorch, PyG, R, Bash, MySQL, HTML/CSS, C++
Machine learning:	Scikit-learn, Transformers, LLMs, CNNs, RNNs, GNNs, VAEs, XGBoost
Development Tool:	Pandas, NumPy, Docker, Git, Nextflow, GoogleCloud, Jupyter, Huggingface, LangChain, Numba
Visualization:	OpenCV, Unity, Matplotlib, Plotly, ggplot2

EDUCATION

London School of Hygiene and Tropical Medicine, DPhil Computational Genomics	Oct 2021 – Oct 2025 (expected)
Dissertation: <i>Integrative Genomic Sequencing and Machine Learning Approaches for Tuberculosis Drug resistance, Diagnostic Tool Development and Transmission Analysis</i>	
Imperial College London, MRes Bioengineering (Hons), Merit	Oct 2019 – Oct 2020
Modules: Computational & Statistical Methods for Research, Frontiers in Bioengineering, Biomaterials Dissertation: <i>Design of an Artificial Bruch's Membrane from Synthetic Polyesters.</i>	
King's College London, BSc Biochemistry (Hons), 1st	Sep 2016 – Jul 2019
Module selected: Bioinformatics, Protein structure and design, Human genomics. Dissertation: <i>Investigation of Concordance Between Molecular Dynamics Simulation and FRET Biosensor using Designed Protein Linker System.</i>	

EXPERIENCE

PhD researcher - LSHTM (Clark Campino Phelan lab), London UK	Jul 2022 – Sep 2025
Project: <i>Deep Learning and Statistical Modelling for Tuberculosis Drug Resistance and Transmission</i>	
- Developed interpretable deep learning models (CNN, RNN, GNN, Transformer) in PyTorch and JAX to predict TB drug resistance from omic data - Built statistical pipelines on HPC in Python for GWAS, PCA, logistic regression, and odds ratio analysis on large omic data - Developed a web application and backend tool in Python for automating the design of targeted sequencing experiments (Amplicon primer design) - Developed an RNN-based generative model using NLP techniques for de novo antimicrobial peptide sequence design. - Built and deployed a RAG LLM chatbot integrating ChromaDB, LangChain, and OpenAI models for natural language interaction with course content	
Machine learning consultant - Deep Science Venture, London UK	Mar 2025 -
Project: <i>Sequence-Based Drug Discovery using Deep Generative Models</i>	
- Led CNN, RNN, and VAE models developments on biological sequences with data augmentation and hyperparameter optimization - Applied SHAP, LIME, and DeepLIFT to extract interpretable biological insights from deep models - Directed strategic model design for generative sequence discovery tasks	
Data study group hackathon participant – The Alan Turing Institute, London UK.	Jan 2024
Project: <i>Shallow Gas Hazard Detection in Offshore Seismic Data</i>	
- Led deep learning model development for detection and segmentation on geophysical imagery using CNNs with contrastive learning - Delivered >90% classification accuracy on legacy seismic datasets under 2-weeks - Collaborated across domain boundaries to translate geoscience challenges into ML solutions	
Machine learning Intern - Linkevity, London UK	Aug 2024 – Oct 2024
Project: <i>Graph-Based Drug-Drug Interaction Prediction</i>	
- Built optimized GNN models with various graph constructions and graphic input build with integration of BERT to predict compound interactions. - Trained drug-drug interaction model for drug design, improved predictive performance by 15%. - Scaled ML training workflows using Google Cloud Platform for high-throughput experimentation	
Data Science intern - ByteDance, London UK	Aug 2022 – Feb 2023
Project: <i>Neoantigen Insight Platform and Market Landscape Scouting</i>	
- Built SQL-integrated chemical database and applied autoencoders to reveal neoantigen patterns across biological datasets - Coordinated interdisciplinary market research on biologics and small molecules - Authored bi-weekly research briefings and developed end-to-end ML pipelines for drug discovery.	

PUBLICATIONS (8 First authored, 12 in total)

- **Wang, L.**, Thawong, N., Thorpe, J., Higgins, M., Tan, M., Sawaengdee, W., Mahasirimongkol, S., Perdigao, J., Campino, S., Clark, T. G. & Phelan, J. E. A novel tool for designing targeted gene amplicons and an optimised set of primers for high-throughput sequencing in tuberculosis genomic studies. *bioRxiv*, (Submitted to *BMC genomics*). doi: <https://doi.org/10.1101/2025.01.13.632698> (**presented at ESM2024-poster and ASMicrobe-talk**) - **Website**: <https://genomics.lshtm.ac.uk/webtoast/#/> **software package**: <https://pypi.org/project/toast-amplicon/>
- **Wang, L.**, Campino, S., Phelan, J. & Clark, T. G. Mixed infections in genotypic drug-resistant Mycobacterium tuberculosis. *Scientific Reports* **13**, 1–8 (2023). doi: <https://doi.org/10.1038/s41598-023-44341-x>
- **LSTM-Based Transfer Learning Models for Tuberculosis-Targeted Antimicrobial Peptide Classification and Generation.** (submitted for publication)
- **Decoding Positive Selection in Mycobacterium tuberculosis with Phylogeny-Guided Graph Attention Models.** (forth coming).
- Data Study Group Final Report: British Geological Survey - Detecting Shallow Gas from Marine Seismic Images". The Alan Turing Institute (2025)
- Phelan, J., Niaz, F., **Wang, L.**, Ngwana-Joseph, G. C., Sobkowiak, B., Cohen, T., Campino, S. & Clark, T. G. TGV: suite of tools to visualize transmission graphs. *NAR Genomics and Bioinformatics* **6**(4) (2024). doi: <https://doi.org/10.1093/nargab/lqae158>