Anabel Yong

LinkedIn: anabelyong
Website: anabelyong.github.io

Email: ucabyon@ucl.ac.uk
Mobile: 07845636286

Summary

MSc Computational Statistics & Machine Learning @ UCL, with special interests in bioinformatics, structural biology & computational proteomics/ genomics. BSc in Mathematics & Biology, with extensive bioinformatics research & data science internships. Passionate about data, & machine learning!

Technical Skills

- **Languages:** R, Python, MATLAB, Octave, Stata, SQL, Go
- **Developer Tools:** Git, VS Code, PyCharm, Google Cloud Platform, Jupyter Notebook
- **Data Repositories:** NCBI, UniProt, Protein Data Bank(PDB), dbSNP, SFARI, ClinVar, BioBank, PantherDB
- **Database Management Tools:** MongoDB, MySQL, Neo4j
- **Bioinformatics Tools:** BLAST, PantherDB, Biopython, Bioconductor, GSEApy, MaxQuant, Percolator, Fragpipe

Education

- **University College London** — London, England
  MSc Computational Statistics & Machine Learning — October 2023 - October 2024

  Relevant Modules: Probabilistic & Unsupervised Learning (Gatsby PhD Module), Statistical Natural Language Processing (Facebook AI Research), Reinforcement Learning (Deepmind's David Silver)

- **University of Edinburgh** — Edinburgh, Scotland
  BSc Mathematics & Biochemistry, First Class Honours — October 2019 - May 2023

  Specialisations: Structural Bioinformatics, Numerical Optimisation, Machine Learning, Statistical Biology

Experiences

Research Assistant in Kustatscher Lab, Welcome Centre for Cell Biology @ University of Edinburgh
Edinburgh, Scotland
Computational Biologist (under supervision of Professor Georg Kutstatcher) — Sept 2022 - July 2023

- Optimised state-of-the-art data preprocessing pipeline, (Link to pgFDR), maximizing human protein detection by 58%, which let to discovering 7000+ novel microproteins
- Benchmarked softwares, MaxQuant & Fragpipe, identifying parameters required for curation of mass spectrometry (MS) proteomics dataset, ProteomeHD2 (Link to Paper)
- Analysed false discovery rate (FDR) strategies for ability to detect microproteins in human proteome.

Data Science @ SAINS Kuching — Sarawak, Malaysia
Part-Time Data Scientist — June 2023 - Sept 2023

- Deployed supervised learning algorithms to predict consumer behaviors in pharmaceutical companies in Kuching

Data Scientist @ Graphcore — London, England
Intern — June 2022 - September 2022

- Implemented A/B testing models, (multivariate A/B tests), to compare and evaluate different IPU-powered generative models tailored to dynamics of consumer markets for laboratory automation.

Research Internship @ UCL Centre of Computational Biology — London, United Kingdom
Statistical Geneticist (under supervision of Professor ZiHeng Yang & Tomas Fluori) — March 2022 - May 2022

- Developed statistical models for phylogenetic analysis on blood allele frequencies.
- Implemented Bayesian MCMC algorithm to sample probability distributions in p, q, and r frequencies, for ABO Blood Frequency Modelling in C# language

Research Internship @ Centre for Systems Biology Edinburgh — Edinburgh, Scotland
Computational Biologist - (under supervision of Professor Ramon Grima) — June 2020 - January 2022

- Developed stochastic models (stochastic simulation algorithm) & deterministic models to analyse and predict complex data patterns in mRNA decay biochemical kinetics.
- Conducted extensive research in Continuous Time Markov Chain processes for sampling functional Kolmogorov Forward Equations (chemical Master Equation).
- Applied Gibbs sampling (MCMC) and Metropolis Hastings algorithm to investigate consistent genes in the bipartite network.

Data Analyst @ International Medical University, Malaysia Kuala Lumpur, Malaysia
Data Analyst Intern June 2020 - July 2020

- Implemented logistic regression in R for feature importance and selection in EAWAG-BBD database for biodegradation genes.
- Applied data normalization, data cleaning and reduction techniques for data preprocessing in biodegradation database with 25000 observations and 25 parameters.

Personal Projects

Random Forests Bayesian Optimisation for HeartFailurePred (Link)

- Implemented Random Forests with Bayesian Optimisation hyper-parameter tuning on heart failure dataset, increasing 5.6% in AUC and accuracy than standard Random Forests
- Applied several feature engineering techniques on heart failure dataset to reduce dataset dimensionality, which improved ML performance by 5% compared to standard unprocessed dataset (clinical dataset) on Kaggle.
- Increased ML performance from 89% to 96% with grid search optimisation compareed to Bayesian Optimisation.

Using structural bioinformatics approach for GULO functionality (Link)

- Conducted a comprehensive structural bioinformatics analysis using PYMOL and CONSURF softwares to investigate the functional domains of the GULO enzyme, contributing to a deeper understanding of the enzymatic conversion crucial for Vitamin C biosynthesis and its association with scurvy.
- Leveraged functional genomics and a suite of bioinformatics tools including NCBI Blastn, Blastp, and tBlastn to investigate the GULO gene functionality across various species, integrating statistical analysis and hypothesis testing to reinforce gene ontology and biochemical pathway insights.

Positions Held

- **EdIntelligence - Artificial Intelligence & Data Science Society** Edinburgh, Scotland
  Vice President Sept. 2021 - May 2022
  - Organising committee for ClimateHackAI datathon where participants develop machine learning models for data provided by OpenClimateFix to reduce industrial carbon emissions
  - Managed social media platforms which has a substantial following of 2000+ followers on Facebook (EdIntelligence-AI& Data FB page and EdIntelligence Machine Learning Society)
- **CompSoc Edinburgh** Edinburgh, Scotland
  Treasurer September 2020 - May 2021
  - Organising committee of HACKTHEBURGH 2021(largest hackathon in Scotland, sponsored by Google, Optiver and MajorLeagueHacking
  - Advised management within subsocieties such as SigCOIN, SigINT and SigWEB (cybersecurity & AI) regarding significant expenditures to keep operations in line with budget limitations

Accomplishments

- **Structures and Functions of Proteins 3 Award, 2022** Highest overall mark in my university cohort for SFP3 in biophysical, & experimental methods component
- **Edinburgh Biological Chemistry 2 Award, 2020** Highest overall mark in my university cohort for quantitative methods in biochemistry.