Owais Siddiqi

✉ omsiddiqi01@gmail.com | 🌐 owaissiddiqi.co.uk | 📍 London, United Kingdom

Professional Summary

Data Scientist with experience in **Bioinformatics**, **Machine Learning**, and **Clinical Data Analysis**. Proven ability to design and deploy scalable RNA-seq and variant-calling workflows using **Python**, **Nextflow**, and **Docker** on **HPC** environments. Skilled in developing **audit-ready, FAIR-compliant ETL pipelines** and leveraging **AI/ML models** to identify biomarkers, model treatment response, and predict patient outcomes. Strong record of translating complex multi-omics datasets into actionable insights that drive oncology R&D and support regulatory-grade analyses (**ISO 13485**, **GDPR**).

Technical Skills

- **Programming & Data:** Python, R, SQL, Bash, MATLAB, Git — proficient in data wrangling, automation, and analytics using Pandas, NumPy, and Jupyter
- **Machine Learning & AI:** Scikit-learn, LightGBM, TensorFlow, PyTorch — applied to predictive modelling, clustering (HDBSCAN, GMM), anomaly detection, and survival analysis (Cox PH, time-varying, Kaplan–Meier)
- **Bioinformatics & Genomics:** RNA-seq (QC, alignment, differential expression, fusion detection) and variant calling pipelines using STAR-Fusion, Salmon, DESeq2, EdgeR, and Nextflow
- **Data Visualisation & Reporting:** Streamlit, Plotly Dash, Matplotlib, Seaborn, Excel — development of interactive dashboards and reports for biological and clinical insights
- **Cloud & Workflow Automation:** Google Cloud Platform (GCP), Docker, CI/CD, Linux/HPC — deployment of scalable, containerised, and audit-ready bioinformatics workflows

Work Experience

CelLBxHealth plc *Jan. 2025 – Present*

R&D Data Analyst

- **Clinical Data Engineering & ETL:** Developed and maintained automated ETL pipelines (Python, Nextflow, Docker, GCP) integrating imaging, NGS, and clinical metadata — reducing onboarding and QC turnaround by **50%**.
- **NGS & RNA-Seq Analytics:** Built scalable, high-sensitivity RNA-seq workflows for pre-processing, differential expression, and fusion detection (STAR-Fusion, Salmon, DESeq2, EdgeR); optimised for low-input rare-cell RNA and deployed on HPC using Docker for reproducibility.
- **AI & Bioinformatics:** Applied unsupervised learning (HDBSCAN, GMM, autoencoders) and risk-modelling pipelines to circulating tumour cell (CTC) and transcriptomic data, identifying high-risk phenotypes and biomarker signatures.
- **Machine Learning & Survival Modelling:** Delivered predictive ML pipelines (LightGBM, scikit-learn) and survival analysis frameworks (Cox PH, time-varying Cox, Kaplan–Meier) to forecast disease progression and therapy response.
- **Integrated Multi-Omics Analytics:** Merged imaging-derived and molecular features to characterise CTC phenotypes and gene-expression programmes, enabling data-driven biomarker discovery and patient stratification.

Genevation Ltd. *Apr. 2024 – Dec. 2024*

Junior Genomic Data Scientist

- **AI-Driven Biomarker Discovery:** Trained and optimised GPU-accelerated PyTorch models for neoantigen prediction, advancing personalised oncology vaccine development.
- **NGS Workflow Automation:** Engineered containerised, cloud-deployable pipelines (Docker, Nextflow, Bash) converting FASTQ to VCF, reducing analysis turnaround by **60%**.
- **RNA-Seq Variant Analysis:** Executed RNA-seq alignment and variant-calling workflows on GCP and HPC environments, delivering reproducible, audit-ready genomic data outputs.
- **Cross-Functional Collaboration:** Worked closely with immunologists and software engineers to refine biomarker hypotheses, producing visual and statistical summaries for internal reports and funding proposals.

PDUK Ltd. *Aug. 2023 – Jan. 2024*

Data Scientist Intern

- **Data Engineering & ETL:** Optimised MySQL-based data cleansing workflows, reducing processing time by **25%** across **3,000+** customer records.
- **Business Intelligence Dashboards:** Built interactive analytics dashboards (Pandas, Matplotlib) surfacing real-time KPIs that enabled faster, data-driven decisions for sales leadership.
- **Forecasting & Predictive Modelling:** Deployed PyTorch RNN models that improved demand forecast accuracy by **20%** and reduced inventory costs by **10%**.
- **Customer Segmentation:** Applied K-means clustering to identify high-value customer groups, increasing campaign conversion rates by **15%**.

Imperial College London (MSc Project) *Mar. 2023 – Sept. 2023*

AI Researcher – Deep Learning

- **Deep Learning for Cell Phenotyping:** Designed and trained a custom TensorFlow/Keras CNN for cell-differentiation scoring, achieving **80%** predictive accuracy.
- **Model Robustness Enhancement:** Implemented advanced data augmentation (elastic deformations, intensity jitter, mix-up), improving generalisation by **30%**.
- **Reproducible AI Workflow:** Developed an end-to-end Jupyter pipeline integrating OpenCV, NumPy, and Matplotlib, enabling seamless reuse and adoption across the research group.

Education

Imperial College London *Oct. 2022 – Oct. 2023*

MSc **Biomedical Engineering**

- Specialised in **Reinforcement Learning** (DQN, PPO, SAC) and **Bayesian Decision Theory**; applied RL to simulated prosthetic-control optimisation tasks.
- Built end-to-end **Computer Vision pipelines** (CNNs for classification, detection, segmentation) using TensorFlow and PyTorch for biomedical imaging.

Queen Mary University of London *Sept. 2019 – Sept. 2022*

BEng (Hons) **Biomedical Engineering — First Class**

- Developed strong foundation in multivariable calculus, linear algebra, and differential-equation modelling for biomedical signal processing and analysis.