Danny Antaki, PhD

San Diego, CA | +1 870-833-3770 | dannyantaki@gmail.com

QUALIFICATIONS

10 years of experience applying ML/AI to biological and business challenges. Data scientist using ML to make real-time production decisions at Twist Bioscience. Collaborated across cross-functional teams as bioinformatic lead to release the Twist RNA Library Prep product and bioinformatic pipeline. Used ML/AI in healthcare-focused research developing and improving next generation sequencing tools and advancing personalized medicine and diagnostics for Autism during postdoc and graduate school.

TECH SKILLS

python: pandas, numpy, sklearn, SHAP, pytorch, pyspark, mlflow, tensorflow, seaborn, pysam, pybedtools, matplotlib
samtools, bcftools, bedtools, GATK, Picard, VEP, plink, STAR, STAR-Fusion, QTL2
SQL, R, perl, bash, Snowflake, Databricks, AWS, github, docker, drone CI, tableau, shiny, jira, confluence, filemaker

EXPERIENCE

Twist Bioscience — Feb 2021 – Present

Senior Data Scientist (Oct 2023 – Present)
Senior Bioinformatics Scientist (Oct 2022 – Oct 2023)
Bioinformatics Scientist (Feb 2021 – Oct 2022)

- **ML/AI**
  - Trained a LLM using HyenaDNA architecture on 2M repeats, improving deletion risk prediction by 10%.
  - Fine-tuned pretrained DNABERT LLM on target enrichment probes to score off-bait effect.
  - Trained a CatBoost ML model on 1.8M examples to predict deletion risk, now in production.
    - Improved previous model with feature engineering and adding 9 features of imperfect repeats.
    - Validated on 250k genes using SHAP for model interpretation and MLFlow for tracking.
    - Designed in-silico experiments that projected a quarterly revenue increase of 5-10%.

- **Data Science and Engineering**
  - Designed an ETL pipeline collecting features and labels from manufacturing data to periodically update the deletion risk model.
  - Built feature store of repeats (500M+ entries) in Snowflake to test multiple training parameters for deletion risk modeling.
  - Set up ETL pipelines to monitor NGS outcomes of gene products, reporting plots and tables to a slack channel.

- **Bioinformatics and Next Generation Sequencing**
  - Improved design algorithm for synthetic Cot-1 blockers reducing material cost by 50-75% while increasing performance by 5-10%.
  - Bioinformatic lead for the Twist RNA Library Prep product, designed RNAseq analysis pipelines for internal QC use and field application scientist use on customer data.

Gleeson Lab for Pediatric Brain Disease — Sep 2018 – Feb 2021

Postdoctoral Researcher

- **ML/AI**
  - Conceptualized and developed visualization method for image-based AI somatic variant caller (Nature Biotechnology).

- **Bioinformatics and Clinical Genetics**
  - Analyzed deeply sequenced whole genomes for somatic variation in postmortem human brain (Nature) and in sperm (Cell, Nature Medicine), finding a ¼ recurrence risk of autism in fathers of autistic children.
  - Linked whole-exome and RNAseq data to patient records in database shared by the lab.

EDUCATION

PhD Biomedical Science – University of California San Diego — Sep 2013 – Sep 2018

- **ML/AI**
  - Designed a ML model to genotype structural variation, reducing the false discovery rate from 20% to below 1% (Bioinformatics)

- **Bioinformatics and Clinical Genetics**
  - Increased diagnostic yield for autism by 2% by combining de novo mutations, rare inherited variants, and polygenic risk in over 37,000 individuals (Science, Nature Genetics).

BS General Biology – Purdue University — Jan 2009 – May 2013