Avantika Lal

PhD biologist with expertise in machine learning for genetics and genomics.

Menlo Park, CA, USA | avantikalal02@gmail.com | linkedin.com/in/avantikalal

PROFESSIONAL EXPERIENCE

GENENTECH

Principal ML Scientist II — *Apr 2024 - present*
Principal AI Scientist — *Feb 2023 - Apr 2024*

- Designing, training and evaluating generative AI approaches for DNA and RNA
- Using deep learning to interpret human genetic data and discover novel drug targets.

INSITRO

Senior Data Scientist — *Aug 2021 - Feb 2023*

- Identifying novel target genes for neurodegenerative diseases by applying deep learning models to genomic and epigenomic data.
- Contributing to multiple therapeutic programs with biological data analysis, including bulk, single-cell, single-nucleus and spatial transcriptomics, epigenomics, CRISPR screens, imaging and proteomics.
- Evaluating state-of-the-art methods and developing robust, scalable analysis pipelines for bulk and single-cell sequencing data.

NVIDIA

Senior Scientist (Deep Learning & Genomics) — *2020 - 21*
Scientist (Deep Learning & Genomics) — *2018 - 20*

- First hire on NVIDIA genomics team. Led development and release of AtacWorks, a deep learning model that enhances epigenomic data, enabling previously impossible analyses.
- Led development and release of GPU-based single-cell genomic analysis software that was 25x faster than existing tools, resulting in adoption in industry and academia.
- Led development of deep learning models to improve short-read and long-read sequencing accuracy, e.g. by correcting upto 40% of DNA sequencing errors in PacBio HiFi sequencing reads.
- Strategic and technical exploration of new product areas in genomics. Mentored 4 interns on successful research projects.

STANFORD UNIVERSITY

Postdoctoral Fellow, Sidow lab, Departments of Genetics & Pathology — *2017 - 18*

- Built deep learning models to identify antimalarial drug targets from genomic and proteomic data. Increased accuracy by 15% over previous methods.
- Co-developed CIMLR, a clustering algorithm for multi-omic data. Analyzed data from thousands of human tumors and applied CIMLR to improve prediction of patient clinical outcomes.
- Co-developed SparseSignatures, an unsupervised learning method to identify causes of mutations in cancer, resulting in 4,000+ downloads and use worldwide.

NATIONAL CENTRE FOR BIOLOGICAL SCIENCES, INDIA

Graduate Student — *2010 - 16*

- Co-developed a new genomic assay to map genome-wide DNA supercoiling in bacteria for the first time.
- Designed and performed biochemical, genetic and genomic experiments, analyzed data, and built mathematical models to discover novel mechanisms of genome-wide transcriptional regulation in response to cellular stress.

ADVISORY / VOLUNTEER EXPERIENCE

CHAN ZUCKERBERG INITIATIVE (CZI) *2021 - present*

- Grant reviewer for single-cell computational biology proposals.

WHITE HOUSE COVID-19 HPC CONSORTIUM *2020*

- NVIDIA representative reviewing and supporting COVID-19 research proposals from academia and government.

THE CANCER GENOME ATLAS (TCGA) CONSORTIUM *2018*

- Invited member of the TCGA Tumor Molecular Pathology Working Group, using machine learning to develop a genomic classification of human cancers.

EDUCATION

TATA INSTITUTE OF FUNDAMENTAL RESEARCH, INDIA

M.Sc. + Ph.D., Biology *2010 - 16*

- 1st division with distinction (highest academic grade)
- Awards/fellowships from the Government of India, Biochemical Society, and Simons Foundation.

UNIVERSITY OF DELHI, INDIA

B.Sc. with Honors, Biochemistry *2007 - 10*

SKILLS

RESEARCH

- Technical project leadership
- Scientific software development
- Scientific communication

PROGRAMMING

- Python
- R
- Matlab
- Linux, Bash
- Git, GitHub
- Docker
- HPC and cloud environments (AWS, GCP)
- SQL

MACHINE LEARNING / DEEP LEARNING

- PyTorch
- Keras
- Scikit-learn
- Tensorflow

BIOINFORMATICS / COMPUTATIONAL BIOLOGY

- Secondary and tertiary analysis of sequencing data
  - Illumina, PacBio, ONT, 10X Genomics
  - DNA sequencing, RNA-seq, epigenomics, functional genomics, multi-omics
  - Single-cell and spatial omics
  - Single-cell CRISPR screens
- Bioinformatics tools and pipelines (e.g. BWA, GATK, Samtools, Bedtools, Seurat, Scanpy)
- Genomic databases (e.g. 1000 Genomes, UK Biobank, GTEx, TCGA)
- Statistical analysis
- Bioconductor

SELECTED PUBLICATIONS (Full list: http://bit.ly/avlalpapers)

- **Lal, A.**, *et al.* Deep learning-based enhancement of epigenomics data with AtacWorks. *Nature Communications* 12, 1507 (2021). https://doi.org/10.1038/s41467-021-21765-5
- Ramazzotti, D., **Lal, A.**, *et al.* Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nature Communications* 9, 4453 (2018). https://doi.org/10.1038/s41467-018-06921-8
- **Lal, A.** *et al.* Genome scale patterns of supercoiling in a bacterial chromosome. *Nature Communications* 7:11055 (2016). https://doi.org/10.1038/ncomms11055