# Avantika Lal

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

Menlo Park, CA, USA +1 (415) 244-9688 + avantikalal02@gmail.com + linkedin.com/in/avantikalal

## **PROFESSIONAL EXPERIENCE**

# GENENTECH, INC. Principal ML Scientist II Principal Al Scientist

2024 - present 2023 - 24

- Leading development of foundational deep learning models to learn the regulatory language of human DNA for applications including target discovery, gene therapy and cell therapy.
- Leading development of software to train and deploy deep learning models for genomics research across Genentech and Roche.
- Developing and applying generative AI methods to design therapeutic DNA and RNA molecules.

# INSITRO, INC.

Senior Data Scientist 2021 - 23

- Discovering drug targets for neurological diseases using deep learning and single-cell genomics.
- Contributed to multiple therapeutic programs with data analysis, including single-cell and spatial transcriptomics, epigenomics, CRISPR screens, imaging and proteomics.
- Developed robust, scalable analysis pipelines for bulk and single-cell sequencing.

#### **NVIDIA CORPORATION**

# Senior Scientist (Deep Learning & Genomics) Scientist (Deep Learning & Genomics)

2020 - 21

2018 - 20

- First hire on NVIDIA genomics team. Led development 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 tools 10-100x faster than existing methods, resulting in adoption in industry and academia.
- Led development of deep learning models to improve short-read and long-read DNA sequencing accuracy, e.g. by correcting up to 40% of errors in PacBio sequencing reads.

#### 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 10,000+ downloads and use worldwide.

# **ADVISORY / VOLUNTEER EXPERIENCE**

#### 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

UNIVERSITY OF DELHI, INDIA

B.Sc. with Honors, Biochemistry 2007 - 10

# **SKILLS**

## **PROGRAMMING**

Pvthon R Bash SOL

• Git, GitHub Docker • HPC and cloud environments (AWS, GCP)

## MACHINE LEARNING

PyTorch PyTorch Lightning • Scikit-learn Keras Tensorflow

#### **COMPUTATIONAL BIOLOGY**

• Secondary and tertiary analysis of sequencing data

- o Illumina, PacBio, ONT, 10X Genomics
- o DNA sequencing, RNA-seq, epigenomics, functional Genomic databases (e.g. 1000 Genomes, UK genomics, multi-omics
- Single-cell and spatial omics
- o Perturb-sea

- Bioinformatics tools and pipelines (e.g. BWA, GATK, Samtools, Bedtools, Seurat, Scanpy)
- Biobank, GTEx, TCGA)
- Statistical analysis
- Bioconductor

# **SELECTED PUBLICATIONS** (Full list: <a href="http://bit.ly/avlalpapers">http://bit.ly/avlalpapers</a>)

- Lal, A., Gunsalus, L., Nair, S., Biancalani, T., & Eraslan, G. gReLU: A comprehensive framework for DNA sequence modeling and design. Nature Methods, Accepted.
- Lal, A., et al. Designing realistic regulatory DNA with autoregressive language models. Genome Research 34.9: 1411-1420 (2024).
- Wang, C., Uehara, M., He, Y., Wang, A., Biancalani, T., Lal, A., Jaakkola, T., Levine, S., Wang, H. and Regev, A. Fine-tuning discrete diffusion models via reward optimization with applications to DNA and protein design. ICLR 2025, Accepted.
- Uehara, M., Zhao, Y., Hajiramezanali, E., Scalia, G., Eraslan, G., Lal, A., Levine, S. and Biancalani, T. Bridging model-based optimization and generative modeling via conservative fine-tuning of diffusion models. Advances in Neural Information Processing Systems, 37, pp.127511-127535 (2024).
- Lal, A., et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nature Communications 12, 1507 (2021).
- Ramazzotti, D., Lal, A., et al. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nature Communications 9, 4453 (2018).

#### **PATENTS**

Israeli, J., Lal, A., et al., Nvidia Corp, 2022. Genetic mutation detection using deep learning. U.S. Patent 11,443,832.

## **PEER REVIEW**

Journals: Nature Genetics, Nature Methods, Nature Computational Science, Nucleic Acids Research, Bioinformatics, PLoS Computational Biology

Conferences: NeurIPS, ICML, RECOMB, ISMB

**Grants**: Chan Zuckerberg Initiative

Page 2/2