PANKAJ PRADEEP

65 West 106th Street, New York, NY 10025 | 332-250-4539 | pp2831@columbia.edu | LinkedIn | GitHub Pages

EDUCATION

Columbia University — **New York, NY**
Master of Science in Biomedical Engineering (Bioinformatics Focus) — *Dec 2023*

Vellore Institute of Technology (VIT University) — **Vellore, India**
Bachelor of Technology in Biotechnology — *Jul 2022*

EXPERIENCE

Columbia Irving Medical Centre - Gertrude H. Sergivesky Centre — **New York, NY**
Graduate Student Researcher - Bioinformatics — *Sep 2023 to Dec 2023*

- Performed quality control analysis of 2500 FASTQ files to investigate Alzheimer's Disease in the MHAS study, leveraging fq2vcf to convert samples to VCF, enhancing the accuracy of analysis by30%
- Designed an R package to visualize genetic data, enabling clear distinction between three local ancestries, clustering samples of Alzheimer's disease risk factors.
- Collaborated with post-docs to troubleshoot errors in pipeline set-up and multi-tasking jobs (qsub, Unix) on HPC cluster, reducing processing time and improving parallel computing efficiency.
- Conducted deconvolution of RNAseq data to single cell compositions to enhance biomarker discovery by 40% using CIBERSORTx

GlaxoSmithKline Pharmaceuticals — **Collegeville, PA**
Summer Intern – Computational Biology Oncology — *Jun 2023 to Aug 2023*

- Utilized Python and R to investigate clinical signatures for gene and protein expression from public databases (CPTAC), performing statistical analyses to identify key biomarkers predictive of patient outcomes, and increasing predictive accuracy by 20%.
- Implemented a correlation analysis pipeline in Python to visualize and distinguish patterns across transcriptomics (mRNA) and proteomics (mass spectrometry) in 947 cancer cell lines from the Cancer Cell Line Encyclopedia, based on different subgroups.
- Collaborated with remote researchers at GSK Germany to mine and ingest 12 scRNAseq cancer datasets for Chronic Lymphocytic Leukemia (CLL), from GEO using Python, enhancing the applicability of in-house bioinformatics HTML pipelines.
- Proposed the usage of a novel metric 'Earth Mover Distance' during a quarterly company wide ideathon 'GSK Science' participated by 200+ researchers, which was implemented in the Target Evaluation Framework by Computational Biology team.

Columbia University - Herbert and Florence Irving Institute for Cancer Dynamics — **New York, NY**
Bioinformatics Research Intern — *Oct 2022 to Jun 2023*

- Leveraged AmpliconArchitect to generate amplicons visualizations annotated with regions of probable extrachromosomal DNA.
- Analyzed 80 whole genome samples in the HPC (bedtools, samtools, bwa) to locate and identify regions of 50 potential hits of extrachromosomal DNA using Integrated Genome Viewer, UCSC Browser for refining amplicon images.
- Worked independently and as a team in setting up Python packages for further analyzing genome data for esophageal cancer.
- Organized literature surveys (PubMed) with detailed reports to further understand mechanism of ecDNA occurrence.

TECHNICAL SKILLS

- **Bioinformatics:** Omics Data Analysis, Samtools, GATK, Seurat, GEO, TCGA, CCLE, UCSC, NCBI, AWS Cloud Practitioner
- **Programming:** R/RStudio, Python, Machine Learning, Bash, Unix, MATLAB, HPC, Intermediate C++/Java
- **Research/Biology:** scRNAseq, bulk RNAseq, RNA/DNA Microarray, Molecular Biology, Genomics, DNA/RNA Extraction, PCR
- **Packages:** Bioconductor, Biopython, Seaborn, Monocle3, PyData stack
- **Relevant Coursework:** ML for Functional Genomics, Statistical ML for Genomics, Applied Data Science, Computational Modeling of Physiological Systems, Immunology and Immunotechnology, Bioinformatics (Genetic Alignment and Protein Engineering)

RESEARCH PROJECTS

Columbia University — **New York, NY**
Enhancing TIGER Model for Cas13 Off-Target Prediction for Indel gRNA Data — *Sep 2023 to Dec 2023*

- Re-architected CNN model to accept gRNA data containing up to 3 insertions/deletions by leveraging Levenshtein distance, resulting in an increase in model applicability and a 60% increase in accuracy of predicting off-target scores for diverse input data.

Multi-Omics Data Integration BRCA-SAE using Stacked Autoencoders — *Jan 2023 to May 2023*

- Pre-processed and normalized transcriptomic, proteomic, and methylation data to ensure accuracy in subsequent analysis.
- Invented a novel Machine Learning Stacked Autoencoder Algorithm (97% accuracy) to infer driving biomarkers for cancer progression by learning patterns across datasets for 1000 samples ingested from TCGA and GEO for breast invasive carcinoma.
- Collaborated effectively with a teammate to troubleshoot errors in model set-up and training for multi-omics analysis.
- Performed Feature Attribution to discover the top 20 biomarkers which are drivers of cancer progression across all datasets.

Predictive Machine Learning Model for Alzheimer's Disease Prediction — *Jan 2023 to May 2023*

- Acquired and pre-processed RNA microarray data for 800 Alzheimer's patients from public databases (GEO, TCGA).
- Performed extensive comparative analysis of gene expression between normal and diseased states using ANOVA, providing insights into differential expression patterns to extract key drivers of disease state.
- Led model training and testing by performing One-Hot encoding for features to feed into regression and machine learning models.
- Assessed various predictive models, prioritizing those with the lowest false negative rates to enhance diagnostic accuracy (>96%).