ISHAAN GUPTA

Discovering Biology, optimizing therapies, and saving lives through diverse cutting-edge Bioinformatics, Machine Learning and Data Engineering skills, and dedication to fast-paced collaborative innovations

San Diego, CA
i3gupta@ucsd.edu
linkedin.com/in/IgAI
ishaanSD.github.io/home

EDUCATION

Ph.D, Computer Science (Research area: Bioinformatics), UC San Diego, La Jolla (2023 – present)
(simultaneously) MS, Computer Science, UC San Diego, La Jolla *(2023 – 2025)*
GPA: **3.92**, Received **Charles Lee Powell Research Fellowship**, Presenting poster at AI in Molecular Biology (Sep 25)
BS, Computer Science : specialization in Bioinformatics, UC San Diego, La Jolla *(2019 – 2022)*
GPA: **3.92**, Magna Cum Laude with CS Honors, IEEE Eta Kappa Nu Honors Society, Minor in General Biology

Research Experience

UCSD CSE Department, San Diego, CA

Graduate Student Researcher (Dr. Tiffany Amariuta) *08/2024 – Present*

- Implementing Deep Learning approaches to predict gene expression at population level and identify haplotype-driven regulation
- Developing statistical techniques for fine-mapping in eQTLs and TWAS for single-cell RNA-seq in multi-tissue polygenic traits like Lupus
- Programming Conda and Nextflow (nf-core) based pipelines for genetics workflows (GWAS, PRS, eQTL) commonly used in the lab

UCSD CSE Department, San Diego, CA

Research Assistant (Dr. Pavel Pevzner) *04/2021 – 08/2023*

Graduate Student Researcher (Dr. Pavel Pevzner) *08/2023 – 07/2024*

- Collaborated with T2T consortium on the evolution of immunology-related loci in humans and primates (***Nature* 2025**)
- Developed visualization tools for comparing repeats, synteny blocks, alignment and gene annotations across genomes (***Gen Res* 2025**)
- Optimized UniAligner (**C++** aligner for highly-repetitive regions that outperforms standard in accuracy and speed) - sped up by ~4 times
- Developed Unsupervised Learning and Signal Processing-based approach for Nanopore-based Peptide Identification

Industry Experience

Abterra Biosciences, San Diego, CA

Proteomics and Machine Learning Intern *07/2024 – 09/2024*

- Improved de novo peptide sequencing accuracy on proprietary long, modified peptides by fine-tuning Transformer-based models
- Automated large-scale hyperparameter sweeps using PyTorch Lightning and bash, programmed and compared search strategies
- Characterized common failure cases to identify model limitations, and guided the company to use the models reliably and efficiently

Illumina (Systems Integration), San Diego, CA

Software Engineer I *06/2022 – 12/2022*

- Performed Agile development for robust LIMS software central to Illumina's high-throughput Genotyping and Methylation pipelines
- Developed and deployed highly scalable microservices in Java (Spring+JDBC) + Angular integrated using REST APIs, AWS and AMQP

Abterra Biosciences, San Diego, CA

Proteomics and Machine Learning Intern *07/2021 – 09/2021*

- Achieved >95% accuracy in resolving ambiguous Antibody protein sequence through Mass-spec metrics and Deep learning approaches
- Implemented parallelized calculations in Java codebase; Trained models in Keras and integrated with Java (through DL4J)

Model Medicines, La Jolla, CA

Data Science Project Consultant *11/2020 – 03/2021*

- Engineered BigQuery-powered ETL pipeline supporting SQL queries and neural networks for screening COVID-19 drug candidates.
- Led software team of 4 to evaluate potential of drug repurposing and novelty by summarizing text from various scientific journals

Projects

- **PyFM: Python-based fine-mapper with efficient configuration search**
  Bayes Factor-based *Fine mapping of GWAS variants,* using Stochastic Search or Simulated Annealing, implementation in Python
- **BetaVAE for image generation of handwritten digits or celebrity faces (https://github.com/CSE203B-project)**
  (Group Leader) Implemented a modified version of *Variational Autoencoders* that results in more interpretable latent factors
- **GANs for Cancer Image Augmentation**
  UNET-based GAN model in *PyTorch* that simulates mammogram images with tumor for data augmentation, and explains tumor type

- **Autocorrelation for ecDNA hubs (Graduate-level course project)**
  Analyzed spatial organization of *extrachromosomal DNA*, and confirmed the formation of accumulation hubs that help co-promote oncogenic expression by visualizing the *Autocorrelation* function for *image processing* of nuclei-stained cell images
- **COVID Mutations Analysis (https://youtu.be/C27B4mYRpXg)**
  Automatic pipeline to find and analyze *phylogeny* of mutations in SARS-CoV-2 proteins from *NCBI database* that increase infectivity
  Led project on phylogeny of Sars-CoV-2 variants to understand the evolution of different strains with *geographical visualization*
- **Transcriptomic analysis pipeline for three cell types using EBSeq**
  Well-documented pipeline to process RNAseq reads from three neuron types and predict equally/differentially expressed genes using R
- **Notes2Map (Python Flask API + React.js Web app)**
  NLP app to generate an interactive network of concepts by processing lecture notes and transcripts, extracting and ranking keywords

Teaching Experience

CSE 181 Bioinformatics Algorithms *01/2021 – 03/2022*

CSE 182 Biological Databases *04/2021 – 06/2022*

Tutored students in understanding, implementing, and applying graph algorithms, Dynamic Programming, and HMMs in Bioinformatics
Created weekly quizzes, engaging puzzles, review cards, and model solutions to exams; aided students in programming; graded exams

BICD 140 Immunology (https://miro.com/app/board/uXjVOAJBvdo=/?share_link_id=954883948124) *04/2021 – 06/2022*

Tutored students in understanding complex concepts like B cell and T cell development/activation, Graft rejection, hypersensitivity etc.
Created interactive graphical board for visualizing concepts, practice problems, and model solutions; graded weekly quizzes and exams

BILD 4 Introductory Biology Lab *04/2021 – 06/2022*

Led Biology Lab sessions, taught lab (pipetting, gel electrophoresis, PCR), and research (literature reading/writing, statistics) skills

Mentoring/Leadership

- **Vice President, Undergraduate Bioinformatics Club** *08/2021 – 07/2022*
  Led a team of 15 to build a thriving community of Bioinformatics students with access to mentors, research talks, and workshops
- **Data Science Student Society Workshop Chair and Bioinformatics Bootcamp Chair** *08/2020 – 07/2021*
  Developed 11 dry lab workshops on Bioinformatics skills, ML, Shell, Python, R, SQL; hosted labs on an AWS EC2 instance

SKILLS

- Python (Pandas, PyTorch, TensorFlow, SciPy), C++, Shell, R, SQL
- Java 7/8/11 (Stream, lambda, DL4J, JUnit), Spring, JDBC, RabbitMQ
- Agile (Scrum) practices, Webdev, Git, CI/CD, Docker, AWS
- Hadoop, Apache Spark (MLlib), Data Pipeline: Airflow, Kafka
- Transcriptomics, single-cell analysis, GWAS, ChIP Seq, Bioconductor

COURSEWORK

- Software Engineering, Parallel Computing (CUDA); Design Patterns
- Deep Learning, Statistical NLP, Recommender Systems, Algorithms
- Bioinformatics Algorithms, Population Genetics, Proteomics Data
- Databases, Convex Optimization, High-dimensional stats
- Spark and Hadoop; NoSQL (MongoDB); Spring; Integration Testing

Publications

Zhenmiao Zhang, **Ishaan Gupta**, Pavel A Pevzner, GenomeDecoder: inferring segmental duplications in highly repetitive genomic regions, *Bioinformatics*, Volume 41, Issue 2, February 2025, btaf058 [Paper] [Code]
DongAhn Yoo *et al.* Complete sequencing of ape genomes. *Nature* 641, 401–418 (2025). [Paper] (Contrib Author: **Ishaan Gupta**)