Harshita Saha

408-839-5639 • Boston, MA • hsaha@hsph.harvard.edu

SKILLS

Languages: Python | R | SQL | Java | C++ | C | Bash | JavaScript | HTML | LaTeX

Tools: Git | Jupyter | HPC | Docker | Snowflake | Bioconductor

Packages: pandas | numpy | matplotlib | scikit-learn | networkx | ggplot2 | dplyr | tidyr | caret

EDUCATION

Harvard University | Cambridge, MA — Sep. 2024 - May 2026
M.S. Computational Biology and Quantitative Genetics — Cumulative GPA: **4.00/4.00**

University of California San Diego | San Diego, CA — Sep. 2020 - June 2024
B.S. Bioinformatics w/ Data Science minor — Cumulative GPA: **3.96/4.00**

Relevant Coursework: Advanced Data Structures, Algorithm Design and Analysis, Data Analysis and Inference, Data Management, Recommender Systems, Computing for Big Data, Advanced Bioinformatics.

EXPERIENCE

Bioinformatics Research Assistant — Jan. 2023 - June 2024

Rana Lab - UC San Diego School of Medicine | San Diego, CA

- Designed and developed **scRNA-seq**, **Bulk RNA-Seq**, and **visualization** pipelines for immunology.
- Created and optimized scripts in **Python**, **R**, and **Bash** to analyze **immune responses** to treatment in **cancer** and **COVID-19** via cell **clustering**, type **labeling**, and **differential expression analysis**.
- Identified **novel COVID-19 mRNA vaccines** that increased immune cell diversity and population.
- Discovered novel vaccine effects on **immune cells** in the **bone marrow** and **spleen** for publication.
- Used packages **DESeq2**, **edgeR**, **topGO**, **clusterProfiler**, **Seurat**, **sctype**, **SoupX**, and **Harmony**.

Data Science Instructional Assistant — Sep. 2022 - June 2024

Halıcıoğlu Data Science Institute - UC San Diego | San Diego, CA

- Instructed **Principles**, **Practice and Application**, and **Theoretical Foundations of Data Science**.
- Provided support to over **2000 students** across **6 quarters** with an average approval rating of **95%**.
- Assisted students by applying understanding of **Python**, **data science**, and **statistical data analysis**.
- Tutored for topics including **machine learning**, hypothesis testing, bootstrapping, and **A/B testing**.
- Held office hours and worked with staff to curate data and course materials using **Jupyter** and **Git**.

Data Science Intern — June 2023 - Oct. 2023

Infometry Inc. | Fremont, CA

- Created **Snowflake stored procedures** using **SQL**, **Python**, and **JS** to automate **ELT** workflows.
- Created pipelines using **Python** to clean and load data into Snowflake from local **Postgres** databases.
- Conducted ad-hoc **data analysis** using **Python** and **SQL** to provide actionable business insights.
- Identified critical **KPIs**, metrics, and **visualization** methods based on client data collection practices.
- Used **Python**, **SQL**, and **Regex** to clean, transform, and structure raw data to **tidy data** for analysis.

PROJECTS

Pancreatic Cancer Differential Expression and Network Analysis Nov. 2024 – Dec. 2024
Harvard University | Cambridge, MA

- Analyzed **differential gene expression** in tumors vs. normal tissue using **TCGA** and **GTEx** data.
- Performed **GSEA** with **GO** and **KEGG** to identify dysregulated biological processes and pathways.
- Conducted **network analysis** with **PANDA** and **KEGG** to explore gene interactions and pathways.
- Identified **prognostic signatures** and potential therapeutic **targets** for early stage pancreatic cancer.

Heart Attack and BMI Associations with Health Factors Nov. 2024 – Dec. 2024
Harvard University | Cambridge, MA

- Analyzed **450,000** CDC 2022 **BRFSS** survey records using a variety of **machine learning** models.
- Conducted data wrangling and **multiple imputation** with chained equations to handle missing data.
- Identified lifestyle factors and comorbidities associated with heart attack, BMI, and both conditions.
- Applied **R** for statistical modeling, visualization, and ensuring reproducibility throughout the analysis.

Computational Modeling of HIV Drug Efficacy Sep. 2023 – Nov. 2023
UC San Diego | San Diego, CA

- Investigated compounds targeting CCR5 as part of HIV treatments using **Python** and **ChEMBLdb**.
- Calculated Lipinski Molecular Descriptors to indicate **bioactivity** and pIC50 to indicate **efficacy**.
- Used **PaDEL** descriptors to identify properties and fingerprints of CCR5 targeting drug molecules.
- Developed **machine learning** models to predict pIC50 and bioactivity to gauge structural efficacy**.**

Machine Learning Pipeline for Recipe Interaction Prediction Nov. 2022 – Dec. 2022
UC San Diego | San Diego, CA

- Predicted user interaction and rating left by user given a user-recipe id pair, using **880,000** data points.
- Conducted **EDA**, feature engineering, and made models using **heuristics**, **regression**, and **NLP**.
- Resulted in accuracy of **0.977** and **0.711**, from baselines **0.457** for interaction and rating respectively.
- Utilized Python and tools including **pandas**, **numpy**, **scipy**, **sklearn**, **nltk**, **seaborn**, and **matplotlib**.

VOLUNTEER EXPERIENCE

Lead Volunteer and Ambassador July 2016 – Present
Save The Child Foundation | Frisco, TX

- Developing a **coding** and **data science** program for women and girls in South Asia using **Python**.
- Creating data driven outreach and intervention programs to address **female health** in South Asia.
- Led the creation and distribution of **biodegradable pads** in underserved South Asian communities.
- Helped raise over **$2M** for female health, education, and sustainability initiatives focused on India.
- Worked with Indian organizations to care for children with special needs and **survivors of violence**.

AWARDS

Halıcıoğlu Data Science Institute Undergraduate Tutor Excellence Award May 2024
UC San Diego | San Diego, CA