Harshita Saha
408-839-5639 • San Diego, CA • hsaha@ucsd.edu

SKILLS

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

Tools: pandas | numpy | scikit-learn | nltk | matplotlib | Snowflake | Power BI | Git | Jupyter | HPC

EDUCATION

University of California San Diego | **Bachelor of Science** Sep. 2020 - Present (Expected June 2024)

Major: Bioinformatics | **Minor: Data Science** Cumulative GPA: **3.96/4.00**

Relevant Coursework: Advanced Data Structures, Design and Analysis of Algorithms, Data Analysis and Inference, Data Management, Recommender Systems, Biological Databases, and Advanced Bioinformatics.

EXPERIENCE

Bioinformatics Research Assistant Jan. 2023 - Present
Rana Lab - UC San Diego School of Medicine | San Diego, CA

- Developed **scRNA-seq**, **Bulk RNA-Seq,** and **visualization** pipelines for immunology applications.
- Developed, optimized, and automated pipelines in **Python** and **R** to analyze responses to treatments.
- Pipelines conducted immune cell **clustering**, type **labeling**, and **differential expression analysis**.
- Identified **novel COVID-19 mRNA vaccines** that increased immune cell diversity and population.

Data Science Instructional Assistant Sep. 2022 - Present
UC San Diego Halıcıoğlu Data Science Institute | San Diego, CA

- IA for the courses **Principles of Data Science** and **Theoretical Foundations of Data Science I**.
- 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 Engineering 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 relevant business insights.
- Identified critical **KPIs**, metrics, and **visualization** methods based on client data collection practices.

PROJECTS

Computational Drug Discovery for HIV Sep. 2023 – Present
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.
- Developing **random forest** 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.