Daniel (Xuechen) Kan | (858)5680019 |xkan@ucsd.edu | 8950 costa verde blv, San Diego, CA
Website: https://daniel-xkan.github.io/my_portfolio/

EDUCATION

University of California, San Diego | Computer Science | Jacob School of Engineering
Bachelor of Science in Computer Science
Graduation: June 2024. Graduated, GPA: 3.711
University of California, San Diego | Computer Science with Bioinformatics
Expected Graduation: June 2024. First-year master's student. GPA: 3.940

RELEVANT WORK (Labs, projects)

CCMS (UCSD Center for Computational Mass Spectrometry). – MassIVE-KB
Research Assistant on Human Proteome Project with Professor. Nuno Bandeira.
Work on: spectral dataset analysis and new protein identification | Code: Github Link here
High-Throughput Validation of Spectral Matches on Peptides → Identify Novel Proteins in Humans

- Python parallel processing tools to conduct **fast and space-efficient** searches on the human genome fasta file against **300,000**+ Prosit-predicted and MassIVE-KB spectra, resolving **1,200+** conflicting peptide identifications and providing insight on improving MS peak recognition accuracy by cross-referencing **18,000+** uniquely identified peptides and **1,000+** new proteins.

Automated Proteomics Data Extraction and Web Scraping:

- Developed Python scripts (Beautiful Soup) to web scrape and analyze proteins and peptides from PeptideAtlas.org. Comparing the data with the MassIVE database and identifying **1k+** missing proteins and **79k+** missing peptide**s**.
- Collect mass spectrometry scan identifiers in PeptideAtlas.org by automating through **12k** peptide sequences missing in the database, generating reports for **250k+** non-redundant spectrum scans.

Extended Proteomics Knowledge Base:

- Read Experimental papers and annotate **9** new mass spectrometry experiment datasets, each with **50-300** experiment identifiers, enhancing our database coverage and making spectrum-level runs and identification possible through annotation.

Technologies Used: Python, Parallel processing, Pandas, Excel, Beautiful Soup, HTTP Requests

Project: Deep Learning Models for θ Estimation in Population Genetics (GWAS)
With Professor Vineet Bafna | Project Report & Code: Link here

- Simulated Population Genetic Data: Data trained on 20,000 samples and testing on 1,000 samples datasets with randomized growth rates (0–200) and a recombination rate of 0.12, and ensuring realistic θ estimation scenarios.
- Baseline Model Evaluation: Implemented and compared classical estimation methods (Watterson's and Tajima's estimators) with machine learning baselines (Linear Regression and Random Forest), analyzing systematic biases and performance across θ values.
- Deep Learning Model Implementation: Developed and benchmarked MLP, CNN, and LSTM models for θ prediction using AFS and SNP matrices and explored the different convolutional layers of the CNN model to achieve the best CNN performance. The MLP model achieved the best performance.

Technologies Used: GWAS, Python, PyTorch, Scikit-learn, NumPy, Pandas, Deep Learning (MLP, CNN, LSTM), Data Simulation (MSMS), Statistical Analysis

RELEVANT SKILLS & COURSEWORK

Bioinformatics (2 years exp.): GWAS, Next Gen DNA/RNA Sequencing (NGS), PyTorch, BLAST, Mass Spectrometry analysis, protein/gene sequencing, fasta file handling, spectra run annotation, biological dynamic programming, and biological machine / deep learning
Programming Languages and systems(3-5 years experience): Python, R-studio, Linux, Java, C, C++, bash, x86, command line, PowerShell, ssh, React, HTML, CSS, React, JavaScript
Data (3-4 years exp.): Numpy, Pandas, Seaborn, SQL, GraphQL, BeautifulSoup, Data Parsing, Web Scraping, Matlab, MongoDB
Machine Learning (2 years exp.): PyTorch, NLP, TensorFlow, Scikit-learn, NLTK, Spark, deep learning, neural network, LLM
Collaboration (4 years exp.): Git, GitHub, Slack, Git Action
Other Tools: Microsoft Office, Excel, Word, PowerPoint, AWS, Docker