Abdoulfatah Abdillahi

abdoulsfsu@gmail.com | 317-828-1996 | Linkedin | GitHub | Portfolio | San Jose, CA

EDUCATION

Master of Science (MS), Data Science and Artificial Intelligence **Jan 2024 - May 2026**
San Francisco State University — San Francisco, CA
Thesis: Predicting E. coli Drug Resistance Using Traditional Machine Learning and Deep Learning

***Key Coursework*:** Artificial Intelligence, Deep Learning, Natural Language Processing, Probability, Statistical Learning & Data Mining, Probability and Statistic I & II, Data Visualization, Human-Computer Interaction

Bachelor of Science (BS), Computer Science ***Dec 2023***
San Francisco State University — San Francisco, CA

***Key Coursework*:** Data Structures & Algorithms, Operating Systems, Software Engineering, Data Science and Machine Learning, Software Development

INDUSTRY EXPERIENCE

Bioinformatic Intern, Genentech, South San Francisco, CA **June 2023 – Sep 2023**
Project Intern: ***Unveiling Treatment Effects from Digital Health Data through Pharmacology-Informed Neural-SDE***

- Continued a research project (originally developed by a PhD student at Genentech) applying pharmacology-informed Neural Stochastic Differential Equations to model treatment effects on simulated wearable cardiac signals.
- Migrated the existing TensorFlow research codebase to PyTorch to enable downstream research iteration; gained hands-on experience with SDE solvers, custom training loops, and autograd, documenting technical progress and blockers in weekly mentor reviews.
- Presented project progress as a poster at Genentech's end-of-summer intern symposium; collaborated with ML engineers and pharmacology scientists through weekly mentor meetings and ad-hoc design discussions.

Data Collection Contractor, Applied Intuition, Sunnyvale CA **Apr 2026 – Present**

- Operate instrumented vehicles equipped with LiDAR, multi-camera, and radar sensors using Applied Intuition's STACK platform to capture training data for autonomous vehicle perception models; perform pre-mission hardware verification and post-mission data offload to storage servers.
- Diagnose and troubleshoot sensor and capture issues during 8-hour field collection shifts, running diagnostic scripts independently and escalating hardware faults to engineering teams to minimize downtime and data loss.

RESEARCH EXPERIENCE

Graduate ML Research, CodeLab, SF State, San Francisco ***Jan 2024 - May 2026***

- Built end-to-end ML pipeline predicting antibiotic resistance from 588 E. coli genomes, benchmarking 6 models across two engineered feature representations; achieved 0.94 balanced accuracy and 0.989 ROC-AUC on a held-out test set with Random Forest on gene-presence features.

- Designed feature engineering pipeline converting raw 5.7 Mb genome alignments into ML-ready matrices: coordinate-mapping algorithm with gap-thresholding produced a 588 × 6,092 gene-presence matrix; MAF + gap filtering produced a 67,259-feature SNP matrix — no off-the-shelf tooling.
- Architected parallel HPC infrastructure on SLURM (job arrays, partition optimization, monitoring dashboards) for full-pangenome BLAST annotation across ~40K gene clusters; pivoted to a Snippy-based shell-genome pipeline after benchmarking revealed the full approach was computationally infeasible.
- Implemented Bayesian hyperparameter optimization (Optuna + TPE sampler, 100 trials per fold) across 5 classical models with 3-fold stratified cross-validation; integrated TabR (retrieval-augmented deep learning) to compare deep vs. classical approaches on tabular genomic data, demonstrating classical methods outperform DL at the 588-sample scale.
- Conducted model interpretability analysis across Mutual Information and MultiSURF feature selection methods, identifying parC and gyrA SNPs as top resistance drivers — consistent with established quinolone resistance biology — and communicating findings to a multidisciplinary team of ML and biology collaborators.

Publications & Presentations

Machine Learning-Driven Pangenome Pipeline for Predicting E. coli Drug Resistance ***Presented at SACNAS Conference, Phoenix Arizona | Nov 2024***

- Architected an end-to-end pangenome data pipeline to process and aggregate three distinct genomic datasets (Mills, Gladstone, Kallonen) for downstream analysis.
- Engineered a novel feature-extraction process to unify multiple data types (SNPs, Gene Presence/Absence) from the pangenome framework, enhancing the input for predictive models.
- Benchmarked a suite of machine learning models (Logistic Regression, Random Forest, GBDT) to validate the hypothesis that a combined pangenome approach yields more generalizable resistance predictions.
- Spearheaded future-work planning to integrate advanced NLP models (e.g., Nucleotide Transformer, BioBERT) for sequence-based resistance prediction.
- Co-presented team poster on behalf of CodeLab at the 2024 NDiSTEM National Conference; contributed to dataset integration across three published E. coli bloodstream-infection studies and pangenome construction pipeline development.

Predicting Amoxicillin (AMC) Resistance in E. coli with CNN and VAE Architectures ***Presented at The Allied Genetics Conference, Washington D.C. | March 2024***

- Led the preprocessing of dual genomic modalities (pangenome alignments, gene presence/absence) to convert raw data files into deep-learning-ready image tensors.
- Benchmarked a Variational Autoencoder (VAE) against a CNN for predicting AMC resistance, with the VAE demonstrating stronger F1-score and accuracy across evaluated configurations.
- Engineered a custom, color-map-based data pipeline, which enabled 40% faster model training and informed the adoption of the VAE for larger-scale analysis.
- **My role:** Contributed to VAE pipeline development; teammates implemented CNN comparison and additional models.

Mentorship & Leadership

Research Peer Mentor — PINC Summer Program, SF State ***Summer 2024***

Advisor: Dr. Pleuni Pennings

- Mentored five undergraduate researchers students in R programming and guided them through a full research pipeline on quinolone use & ST131 E. coli across ten European countries.

- Supported data collection from European CSC & Enterobase and statistical analysis workflow.
- Presented findings and reinforced that multifactorial approaches are needed to address antibiotic resistance.

Grader, San Francisco State University
- CSC 665/865: Artificial Intelligence **Sep 2025 - Dec 2025**
- CSC 408: Data Science for Personalized Medicine (Python) ***Sep 2024 – Dec 2024***
- CSC 219: Data Structures for Data Science Application Development (Python) ***Jan 2024 – May 2024***
- CSC 101: Introduction to Computing and Programming (Java) ***Jan 2024 – May 2024***

Braven (Remote), *Grader* **Sep 2024 – May 2025**
- Graded assignments and projects for a cohort of 24 students using AU 220 rubrics.
- Provided detailed feedback to enhance students' career and leadership development.
- Held weekly virtual office hours to support learning, address concerns, and foster academic success.
- Reviewed instructional materials including slide decks, projects, and rubrics to ensure alignment with learning outcomes.

HONORS & AWARDS

NIH – UCSF/SFSU Bridges to the Doctorate Fellowship, SFSU ***2023–2025***
Awarded ~$32,000 in a two-year research training fellowship to master's students from underrepresented groups in science aimed at transition into biomedical research-focused doctoral programs.

BMC Scholarship, SFSU ***September 2024***
Awarded $1,750 by the Computer Science Department for academic excellence and commitment to research.

National Diversity in STEM (NDiSTEM) Travel Scholarship, SACNAS Conference, Phoenix, AZ ***October 2024***
Received full travel support including roundtrip flight, lodging, and conference registration to attend the 2024 NDiSTEM Conference.

Award of Recognition, SFSU PINC Program ***May 2024***
Recognized for exemplary performance and service to students and faculty, contributing significantly to the success of the PINC community.

Certificate

Data Science and Machine Learning for Biotechnology Certificate ***2022–2023***
San Francisco State University (in collaboration with Genentech)
- Completed a 12-unit, two-semester program designed in partnership with Genentech covering machine learning, data science, and computational methods for biotechnology and pharmaceutical applications.

TECHNICAL SKILLS

Languages: Python (advanced), SQL, Bash, R, Java, C++/C, HTML, CSS
ML / DL Frameworks: scikit-learn, PyTorch, XGBoost, Optuna, MLflow, TensorFlow/Keras, TabR
Data & Compute: Pandas, NumPy, Dask, Matplotlib, Seaborn, Plotly, HPC/SLURM, parallel computing, Apache Spark
Statistics & Modeling: Probability, statistical inference, hypothesis testing, regression, regularization, experimental design, cross-validation
Infrastructure & DevOps: Docker, Git/GitHub, GitHub Actions (CI/CD), PyPI packaging, Linux/Bash scripting, SLURM/HPC clusters
Web & Backend: Flask, REST APIs, SQLAlchemy, HTML/CSS
Bioinformatics: Snippy, Roary, Prokka, BLAST, FastQC, pangenome analysis
Visualization & BI: Matplotlib, Seaborn, Plotly, Tableau

ACADEMIC PROJECTS

Energy Consumption Modeling – Regression & Clustering Project
MATH 448: Statistical Learning | San Francisco State University ***Feb 2025 – May 2025***

- Built end-to-end regression and forecasting pipeline on UCI Appliances Energy dataset (19,735 observations × 29 features); benchmarked Linear/Ridge/Lasso, Decision Trees, Random Forest, and tuned tree ensembles with ANOVA-based feature selection. Random Forest delivered best performance ($R^2 = 0.60$, MAPE reduced from 67% to 30% vs. linear baseline).
- Applied clustering algorithms to uncover patterns in consumption behavior, delivering insights into sustainable energy management.

MongoDB Database Python Package **Feb 2025 – May 2025**

- Built and published a Python package automating MongoDB connections and bulk CSV/Excel ingestion; configured full CI/CD (GitHub Actions, unit tests, tox), packaging (setuptools, pyproject.toml), and shipped 11 versions to PyPI.

Data Visualization Project – Shopping Trends and Customer Behavior ***Jan 2024 – May 2024***
CSC 805: Data Visualization

- Built an interactive web dashboard using Dash, Plotly, and Python to visualize and explore customer behavior from a Kaggle dataset (~3,900 records).
- Engineered modular visualizations including a product insights bar chart, seasonal trends graph, U.S. heatmap of popular items, and demographic filters (age, gender) for purchase behavior.
- Delivered a fully functional data exploration tool that enables users to dynamically filter by age, gender, and product type, offering real-time insights into customer trends and purchase patterns.

Alzheimer's Disease Classification Using PET Images **2024**

- Implemented and compared transfer learning models (VGG16, InceptionV3, ResNet50) to classify PET scans into cognitive stages; achieved 95%+ accuracy with VGG16 after augmentation and hyperparameter tuning on the ADNI dataset.

Full-Stack Developer – Educlouds App ***Jan 2024 - May 2024***
AI Entrepreneurship Course Project

- **Frontend Development**: Designed and implemented a responsive user interface using modern web technologies, enabling students to easily access and request computational resources.
- **Backend Engineering**: Developed RESTful APIs and backend logic using Flask, Python, and SQLAlchemy to manage data flow and facilitate secure transactions.
- **Project Showcase**: Successfully presented the completed Educlouds platform during the end-of-semester CSC 890 showcase, demonstrating full-stack functionality and real-world applicability

REFERENCES

Dr. Sara El Alaoui
Assistant Professor, Department of Computer Science, San Francisco State University
Thesis Supervisor
elalaoui@sfsu.edu

Dr. Pleuni Pennings
Professor, Department of Biology, San Francisco State University
Co-Thesis Supervisor
pennings@sfsu.edu

Dr. Ilmi Yoon
Teaching Professor and Director of Computing Programs, Khoury College of Computer Sciences, Northeastern University
i.yoon@northeastern.edu