VICTORIA CHEUNG

email: victoriakcheung@gmail.com website: https://vic-cheung.github.io/

TECHNICAL SKILLS

- **Programming Languages:** Python, R, MATLAB, SQL (PostgreSQL)
- **Machine Learning:** Supervised and unsupervised techniques (e.g., SVM, Random Forest, Boosted Trees, softmax, LDA, NMF), hyperparameter tuning, model evaluation
- **Data Science & Analytics:** Multi-omics data analysis, feature engineering, data preprocessing, statistical analysis, cross-validation, model deployment, data fusion
- **Bioinformatics:** CellRanger, BedTools, ScanPy, fgsea
- **Cloud Computing & Tools:** AWS, GCP, Azure, Flyte, Conda, Poetry, UV
- **Frameworks & Libraries:** PyTorch, scikit-learn, Pandas, NumPy, SHAP, Matplotlib, Seaborn
- **Operating Systems:** Linux (Ubuntu, CentOS), macOS

EDUCATION

University of California, San Francisco (UCSF) – PhD in Genetics with a concentration in Systems Neuroscience
University of California, San Diego (UCSD) – BS in Microbiology

PROFESSIONAL EXPERIENCE

Data Scientist–Computational Biology | *Freenome* | APR 2022 — PRESENT

- High-Throughput Data Science & Tool Development:
  - Designed pipelines and implemented computational tools (supervised and unsupervised ML approaches) for efficient analysis of NGS data, including RNA-Seq, WGS, targeted sequencing, and methyl sequencing.
  - Developed a software package to model cell-free DNA fragmentomic signatures and predict gene activation likelihood, incorporating these predictions as features for subsequent cancer detection tasks and disease monitoring models.
  - Applied wavelet-based signal processing to preprocess fragmentomics data, facilitating the development of robust machine learning models.
  - Engineered and deployed scalable data pipelines on distributed compute workflows (e.g., Flyte) to streamline NGS data processing, reducing analysis time from 8 days to half a day and enhancing reproducibility.
- Machine Learning and Multimodal Data Integration:
  - Led the integration of fragmentomics, methylomics, and proteomics data to develop classifiers for cancer detection, treatment response, and disease progression monitoring across diverse indications.
  - Partnership with Siemens Healthineers for early detection of breast cancer
  - Partnership with ADCT responders vs. nonresponders to loncastuximab in DLBCL: ASH 2022 abstract, poster **2nd author**, AACR 2023 abstract **co-first author**
  - Longitudinal monitoring of residual disease in colorectal cancer patients: AACR 2024 abstract, poster **co-first author**]
  - Applied SHAP analysis to interpret complex machine learning models and provide actionable insights for biopharma partners.
- Data Visualization and Communication:
  - Designed impactful data visualizations using Matplotlib and Seaborn to present complex findings to biopharma stakeholders and internal teams.
  - Created comprehensive slide decks and posters for scientific conferences, highlighting innovative methods and project outcomes.
- Leadership and Cross-Functional Collaboration:
  - Led cross-disciplinary teams to design analyses, interpret clinical data, and deliver multimodal analyses for biopharma partnerships to identify key biomarkers and optimize clinical trial strategies.
  - Built partnerships with key opinion leaders (KOLs) and biopharma collaborators to drive clinical impact and advance computational tool development.
  - Mentored junior scientists through guiding the development of machine learning models; fostering innovation and skill development.
  - Played a pivotal role in a successful biopharma collaboration, subtyping small cell lung cancer using liquid biopsy data and creating visual summaries of analyses, which secured a multi-million-dollar contract extension.

Data Scientist Intern–Bioinformatics Molecular Oncology | *Genentech* | SEP 2021 — APR 2022

- Contributed to early-stage drug development by characterizing gene signatures for T cell exhaustion after treatment in cancer models, providing insights into poorly understood biological mechanisms.
- Designed and implemented an efficient data processing pipeline for large-scale single-cell RNA-Seq analysis using Scanpy, Numpy, Pandas, scikit-learn, SciPy, and Matplotlib/Seaborn.

- Conducted comprehensive statistical analyses, including gene set enrichment analysis (pathway analysis) and differential gene expression, to identify key biological insights from T cells after administration of therapeutic agents.
- Applied supervised batch correction techniques and unsupervised clustering algorithms (UMAP, LDA, NMF) to reveal novel patterns in large-scale scRNA-Seq data.
- Collaborated effectively with cross-functional teams to deliver high-quality computational insights, driving the advancement of oncology research programs.

Data Scientist–Health Data Analytics | *Insight Data Science* | MAY 2020 – JULY 2020

- Developed a predictive clinical tool using supervised machine learning models (scikit-learn, XGBoost) to forecast Acute Kidney Injury (AKI) with ~91% accuracy, enhancing care management and potentially reducing hospital length of stay.
- Queried and processed over 3 million rows of data from the MIMIC-III database (publicly accessible collection of de-identified medical data from critically ill patients) using PostgreSQL and Python (Pandas), extracting 70 unique features from 25 tables and 46,000 patients. Features included lab tests, demographic information, thousands of diagnoses, and other clinical documentation.
- Leveraged AWS EC2 for efficient data processing and analysis, demonstrating proficiency in cloud computing.
- Published work in ***Towards Data Science***: Predicting Acute Kidney Injury in Hospitalized Patients Using Machine Learning

PhD Researcher–Neuroinformatics and Systems Neuroscience | *UCSF @Evan Feinberg Lab* | JUL 2016 — SEP 2021

- Innovative Method Development**:**
  - Developed VECTORseq, a high-throughput single-cell sequencing method for neurons, enabling viral barcoding to preserve connectivity information and allowing pooling of multiple projection populations in a single sequencing run, significantly reducing costs and labor.
  - Engineered a scalable data processing pipeline using Python, integrated with genome alignment tools (e.g., Cellranger, 10x Genomics) on AWS EC2, ensuring computational efficiency for large-scale RNA sequencing datasets.
  - Applied unsupervised machine learning techniques (e.g., t-SNE, UMAP, Leiden) to analyze and map molecular identities of neuronal clusters, linking gene expression to neuronal function and behavior.
  - Validated VECTORseq by identifying both well-characterized and novel neuronal subtypes, such as visual cortical projections and deep SC populations, highlighting its utility for discovering new projection populations.
  - Applied VECTORseq to integrate findings across studies, identifying a GABAergic population in the zona incerta (ZI) expressing Pax6 that innervates the ventral midbrain (VM), linking gene expression to connectivity and functional properties.
  - Optimized brain dissociation protocols based on clustering-derived insights, enhancing neuron survival yield by 100x, which directly improved the quality and consistency of samples for sequencing.
- Signal Processing and Data Automation: Designed vision and audition-based behavioral paradigms to investigate sensorimotor integration in mice, combining custom software with experimental frameworks.
  - Engineered a rotary encoder system with Arduino to track mice responses to auditory cues, serially communicating data to MATLAB for recording positional coordinates and downstream decision-making analysis.
  - Automated and parallelized data acquisition pipelines with MATLAB and Arduino, improving productivity by 6-fold.
  - Analyzed sensory input representations and behavioral command transformations using fiber photometry and custom-built software to enhance understanding of neural-behavioral interactions.
  - Conducted signal processing on fiber photometry and calcium imaging data, applying advanced techniques like noise filtering and event detection to uncover neuronal dynamics.
  - Designed and engineered systems to track decision-making by implementing rotary encoder systems with Arduino-MATLAB serial communication.

SELECT PUBLICATIONS [* denotes equal contribution]

Herault, A., et al., **Cheung, V.** "NKG2D-bispecific enhances NK and CD8+ T cell antitumor immunity." *Cancer Immunology, Immunotherapy*, 73.10 (2024): 1-16. DOI

Tang, A.D.*, Gupte, R.*, **Cheung, V.*,** Qing, T.*, et al. "Noninvasive longitudinal monitoring of residual disease in chemotherapy-treated colorectal cancer patients." *Cancer Research* 84(6): 6258. DOI

Vallania, F.*, **Cheung, V.***, et al. "Discovery of plasma protein biomarkers associated with overall survival in R/R DLBCL patients treated with loncastuximab tesirine." *Cancer Research*, 83(7_Supplement), 2023, 5387. DOI

Vallania, F., **Cheung, V.**, Zamba, MD., Liu, J., Pasupathy, A., et al. "Identification of Predictive Biomarkers for Response of R/R DLBCL Patients Treated with Loncastuximab Tesirine Using Low Pass Whole-Genome Sequencing (WGS)." *Blood*, 140(Supplement 1), 2022, 3551-3552. DOI

Cheung, V., et al. "Transcriptional profiling of mouse projection neurons with VECTORseq." *STAR Protocols*, 3(3), 2022, 101625. DOI

Cheung, V., et al. "Virally Encoded Connectivity Transgenic Overlay RNA sequencing (VECTORseq) defines projection neurons involved in sensorimotor integration." *Cell Reports*, 37(12), 2021, 110131. DOI