David Lin

ddlin@berkeley.edu | (617) 412-5665 | U. S. citizen
https://tealeave.github.io/da-wei-lin-data/
5129 Vía Samuel, Yorba Linda, CA 92886

SKILLS

Data Science & Programming

- Adept in Python (NumPy, Pandas, Scikit-learn), SQL, R, Data Visualization (Plotnine, Seaborn, Plotly), Databrick workflow, and GitHub, EHR, TCGA
- Proficient in statistical Machine Learning and Big Data analysis using Spark (PySpark)
- Skilled in leveraging the PyTorch API for deep learning, image analysis, and developing large language models such as DNA-BERT
- Enthusiastic about Nextflow for achieving consistent results and efficient pipeline execution

Multi-Omic Data, Experimental Design, and Bioinformatics Expertise

- Experienced in transcriptomic and genomic data
  - Knowledge of germline, somatic, RNA-seq, and WES NGS workflows and pipelines
- Expertise in molecular assay development and proteomic assays
  - High-throughput screening platform for mutant p53 reactivating compounds
  - CRISPR-mediated gene editing: Rapid knockout and tagging tools
  - LC-MS-based metabolomic and proteomic data analysis, from raw data processing, QC, to downstream insight synthesis
  - Conversant in scRNA-seq, Methyl-seq, and spatial genomic analysis
- Broad domain knowledge in Cancer Metabolism. Navigator of the intricate pathways of Cell Cycle regulation. Explorer of Stem Cell Biology. CRISPR expert in the vast ocean of gene editing

Project Management

- Skilled in using JIRA, GitHub, Databrick,and Confluence for project organization, repository management, and documentation
- Supportive team player and mentor to junior scientists
- Responsible for cross-departmental communication with stakeholders in R&D, clinical diagnostics, assay validation, clinical super lab, grantee collaborators

EXPERIENCE

Data Scientist III

Sapient Bioanalytics **Oct 2023 - present**

- **Led the analysis on population-scale biomarker identification** and risk score machine learning for preeclampsia, stillbirth, small for gestational age, and preterm birth across five global locations as part of the Gates Foundation's MOMI project.
- **Spearheaded capacity-building initiatives in developing countries** by organizing workshops on metabolomics data analysis and offering mentorship to empower global researchers and healthcare professionals.
- **Collaborated with cross-functional teams** at Sapient Bioanalytics, including mass spectrometry, computational chemistry, and principal scientists from the MOMI sites, managing and cleaning data from 50,000 samples, each with over 40,000 metabolomic features, making it the largest metabolomics study to date.
- **Developed and Optimized a Three-Stage Analytical Pipeline:**
  1. **Data Cleaning and Stratification**: Streamlined the process for handling large-scale datasets, ensuring the integrity and quality of the data for subsequent analysis.
  2. **Feature Engineering and Regression Analysis with PySpark**: Leveraged advanced statistical learning techniques and conducted meta-analysis and pooled analysis to extract meaningful features, synthesize results, and develop predictive models, enabling scalable and efficient processing of vast datasets.

3. **Data Visualization and Pathway Enrichment Analysis**: Employed sophisticated data visualization tools and conducted pathway enrichment analysis to reveal significant biological pathways, providing deeper insights into the data's biological relevance.
4. **Metabolic Risk Score Development with XGBoost**: Applied cutting-edge machine learning techniques using XGBoost with derived biomarkers, combined with hyperparameter optimization via Optuna, to develop a metabolic risk scoring system.

- **Conducted mass-spectrometry-based metabolomic and proteomic analyses** on semaglutide-lead-compound-treated samples, employing advanced bioinformatics workflows and machine learning techniques to identify molecular signatures, potential biomarkers, and mechanistic insights supporting therapeutic development.

Bioinformatic Scientist III

Ambry Genetics **Dec 2019 - Oct 2023**

- Analyzed NGS panel probe coverage, sequencing run consistency, coverage uniformity, detection sensitivity and specificity for improved results.
- Contributed to the development of oncology panels, including high-volume CancerNext, RNA, Exome, and Somatic panels.
- Implemented strategies to reduce CNV calling QC failures and false positive rates during chemistry transitions, serving as a subject matter expert on the FDA application for the largest germline panel and contributing crucial technical and validation insights to support regulatory approval.
- Established statistical thresholds on allele frequency for reducing laboratory Sanger workload.
- Produced crucial data for variant assessment, including SNV, indel, CNV, Mobile Elements, and processed pseudogenes.
- Employed GradientBoostingRegressor for estimating the number of CNVs per sample and determining feature importance (including sequencing, coverage, and augmented metrics) in high false positive CNV investigations.
- Designed, developed, and validated a ctDNA-based liquid biopsy NGS pipeline, leveraging benchmark metrics from internal samples, GIAB standards, and TCGA datasets to ensure robust performance and adherence to product requirements.

Data Science Fellowship

The Data Incubator **Jun 2019 - Sep 2019**

Postdoctoral Researcher

UC Irvine **Jan 2014 - Dec 2019**

- Developed **CRISPR gene editing** (tagging and knockout) and viral derived vectors to investigate PP2A L309 methylation during SAM-checkpoint activation
- Purified HBTH-tagged PP2A (SILAC) for LC-MS. Followed by proteomic analysis and resulted in identifying PP2A interaction map during methionine stress
- Developed bio-screening platform in mammalian cells for p53 reactivating compounds, resulted in a publication in **Nature Communication**
- Analyzed RNA-seq data for hypothesis testing and discovered actionable targets for investigating SAM-checkpoint

EDUCATION

University of California, Berkeley
M.S. Information and Data Science **In progress**

University of California, Irvine **Sep 2007 – Jan 2014**
Ph.D. Biological Sciences

California Institute for Regenerative Medicine (CIRM) fellowship **2009 - 2011**

AWARDS & PUBLICATIONS

Molecular Metabolism | Nutrient control of splice site selection contributes to methionine addiction of cancer **2025**

Submitted | Sphingosine and anti-neoplastic sphingosine analogs activate PP2A and inhibit nuclear import in parallel by engaging PPP2R1A and importins **2024**

Cell Chemical Biology | Discovery of compounds that reactivate p53 mutants in vitro and in vivo **2022**

Journal of Lipid Research | Lipid remodeling in response to methionine stress in MDA-MBA-468 triple-negative breast cancer cells **2021**

Data Science Fellowship | The Data Incubator - San Francisco **2019**

Journal of Biological Chemistry | Microhomology based CRISPR tagging tools for protein tracking, purification, and depletion **2019**

Methods Mol. Biol. | Isolation and characterization of methionine-independent clones from methionine-dependent cancer cells **2019**

U.S. Patent | Chembridge Small Molecules that could enhance p53 activity **2015**

Journal of Cell Science | SAM limitation induces p38 mitogen-activated protein kinase and triggers cell cycle arrest in G1 **2014**

Nature Communication | Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53 **2013**

Cell Cycle | Downregulation of Cdc6 and pre-replication complexes in response to methionine stress in breast cancer cells **2012**

Journal of Biological Chemistry | Transforming growth factor β up-regulates cysteine-rich protein 2 in vascular smooth muscle cells via activating transcription factor 2 **2008**

Genes to Cells | Identification of a putative human mitochondrial thymidine monophosphate kinase associated with monocytic/macrophage terminal differentiation **2008**