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

SUMMARY

Ph.D.-trained data scientist with 8+ years of hands-on laboratory and bioinformatics experience in drug discovery settings. Expert in deriving statistical insights and developing AI/ML models, alongside designing and executing molecular and cellular assays (CRISPR screens, high-throughput compound screening, LC-MS metabolomics/proteomics) within scalable Linux/Python pipelines. Demonstrated track record driving biomarker identification, assay development, and cross-functional projects from R&D through clinical validation.

CORE COMPETENCIES

Drug Discovery & Development

- **Assay Design & Execution:** Development and optimization of cell-based (CRISPR knockout/tagging, high-throughput p53 reactivation screens) and biochemical assays (enzyme activity, binding studies) in 96-/384-well formats
- **Lead Identification & Optimization:** High-throughput screening (HTS) workflows, hit triage, SAR analysis, and structure–activity relationship support for compound advancement
- **Omics-Driven Target Validation:** LC-MS/MS–based metabolomics and proteomics workflows for mechanism-of-action studies and biomarker discovery
- **NGS-Based Profiling:** RNA-seq, whole-exome, and targeted panel sequencing for pharmacogenomic insights and off-target effect assessment
- **Translational Biomarker Development:** Machine-learning model development (XGBoost, Optuna tuning) for predictive biomarker signatures in preclinical and early-phase studies
- **Regulatory & Validation Support:** Generation of assay validation reports and statistical thresholds to satisfy FDA In Vitro Diagnostic submissions

Multi-Omics & Bioinformatics

- **Bioinformatics:** Designed and integrated end-to-end pipelines for bulk and single-cell data, including RNA-seq, WES, scRNA-seq, methylation sequencing, germline and somatic variant workflows, CNV detection, and comprehensive variant pipeline benchmarking.
- **Proteomics & Metabolomics:** Managed LC-MS data processing through to biomarker discovery, translating findings into clear visualizations and crafting detailed reports and presentations.
- **Statistical Analysis & Modeling:** Selected and refined statistical approaches, performing hypothesis testing and diagnostic checks, evaluating model fit and residuals, and applying causal inference techniques where appropriate.

Advanced Machine Learning & AI

- **Sequence Modeling & Forecasting:** Developed ARIMA–LSTM–CNN hybrid pipelines and attention-based transformer models (TimeXer) for multivariate time-series, integrating calendar and exogenous features.
- **Deep Architectures & Embeddings:** Implemented hybrid LSTM–CNN architectures, built custom positional-embedding modules, and visualized high-dimensional embeddings using Manim.
- **Optimization & Scalability:** Designed and executed Optuna-driven tuning of hyperparameters, training features, regularization strategies, and loss functions to maximize model performance and ensure robust generalization.
- **Dynamic Risk-Scoring Pipelines:** Engineered end-to-end XGBoost and PyTorch workflows for real-time risk prediction, automated via Databrick MLflow.

Programming & Data Management:

- Advanced skills in Python (NumPy, Pandas, Scikit-learn, PyTorch) and R for data analysis and model development
- Proficient in SQL and Spark (PySpark) for managing and processing large-scale datasets including high-dimensional clinical data

EXPERIENCE

Data Scientist III

Sapient Bioanalytics **Oct 2023 - present**

- **Population-Scale Biomarker Discovery & Risk Modeling:** Led the Gates Foundation MOMI project's analysis of 50,000 samples (40,000+ metabolites per sample) to identify biomarkers and train ML risk-score models for preeclampsia, stillbirth, small-for-gestational-age, and preterm birth across five global cohorts.
- **Three-Stage Analytical Pipeline Architecture:** Designed and implemented a scalable workflow comprising (1) large-scale data cleaning and stratification, (2) PySpark-driven feature engineering and regression/meta-analysis, and (3) interactive data visualization coupled with pathway enrichment to uncover biological insights.
- **Metabolic Risk-Score Development:** Built an XGBoost-based scoring system with Optuna-tuned hyperparameters, deployed on HPC to process terabyte-scale datasets and deliver robust, reproducible predictive models.
- **Transformer-Based Longitudinal Modeling:** Pioneered the use of transformer architectures on longitudinal metabolomics data to enhance risk-score prediction performance for adverse pregnancy outcomes compared to traditional methods.
- **Mass-Spec Metabolomic & Proteomic Analysis:** Conducted advanced bioinformatics and ML workflows on semaglutide-lead-compound-treated samples, uncovering molecular signatures, candidate biomarkers, and mechanistic insights to support therapeutic development.
- **Capacity Building & Collaboration:** Organized hands-on workshops and mentorship programs in developing countries to elevate local multi-omics and AI expertise, and coordinated with mass spectrometry, computational chemistry, and clinical teams to ensure data integrity, reproducibility, and alignment with project milestones.

Bioinformatic Scientist III

Ambry Genetics **Dec 2019 - Oct 2023**

- **Oncology Panel Development**: Led bioinformatic design and validation of high-volume NGS panels (CancerNext, RNA, WES, somatic). Ensured ≥99% sensitivity/specificity for SNVs, indels, CNVs under FDA regulatory guidelines.
- **QC & Statistical Thresholding**: Established robust statistical cutoffs for allele-frequency and coverage metrics, reducing false positives by 30% during chemistry transitions.
- **Pipeline Automation**: Built and maintained Git-versioned, Nextflow pipelines on HPC, streamlining end-to-end data processing from raw FASTQ to annotated VCF.
- **Machine Learning Applications**: Implemented GradientBoostingRegressor models to predict CNV counts and prioritize QC investigations, improving workflow efficiency.

Data Science Fellowship

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

Postdoctoral Researcher

UC Irvine **Jan 2014 - Dec 2019**

- **CRISPR & Proteomics**: Developed CRISPR tagging/knockout tools and purified tagged PP2A complexes for SILAC-LC-MS; mapped protein–protein interactions under metabolic stress.
- Developed HTS bio-screening platform in mammalian cells for p53 reactivating compounds, resulting in a publication in **Nature Communications.**

- **RNA-seq Analytics**: Designed pipelines for differential expression and time-series analysis, identifying novel therapeutic targets.
- **Metabolomics Integration**: Processed and analyzed raw LC-MS metabolomic profiles; performed pathway enrichment to elucidate methionine addiction mechanisms in cancer.

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 **2025**

Cancer Research | Role of PP2A methylation on methionine dependence of cancer **2023**

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**