Erol **Kavvas**

Data Scientist · Systems Biologist · Machine Learning Researcher

eskavvas@gmail.com | erolkavvas.github.io | erolkavvas | erolkavvas

Summary

Data scientist with +10 years of experience developing AI and mechanistic models for biotechnology. Proven track record in building and deploying scalable, interpretable ML solutions for genomics, drug discovery, and precision medicine. Skilled in leading cross-functional collaborations, mentoring teams, and optimizing cloud-based pipelines for high-dimensional biomedical data. Passionate about translating biological complexity into therapeutic insight through AI-driven modeling.

Skills

Machine Learning	PyTorch, scikit-learn, Deep Learning, Large Language Models (LLMs), Neural Nets, Linear models, Interpretable ML, LIME, SHAP, Tree-based Methods, Decomposition Methods, Ensemble Methods, Network Modeling, Recommender Systems, Model Optimization, Gradient Boosting, SVMs
Programming	Python (Expert), R, SQL, Linux, C, Visual Basic, Fortran, Bash
Data Analysis & Visualization	Pandas, Numpy, Matplotlib, Seaborn, NetworkX, Statistics, A/B Testing, Experimental Design
Big Data Technologies	PySpark, Microsoft Azure (Synapse, ML Studio), Microsoft Fabric, Docker, GPU Optimization
Other Skills	Git, NGS Analysis (DNA, RNA, ^{13}C flux, 16s-rDNA), Flux Balance Analysis, Affinity Designer

Experience

Envisagenics (RNA Therapeutics Startup)

New York City, NY

Data Scientist - Functional Genomics

July 2022 - Feb 2025

- **Developed LLM for RNA therapeutics**: Built *SpliceImpact*, a transformer-based large language model using ESM-2 embeddings for predicting per-residue functional effects of alternative splicing on proteins, with GPU-accelerated training on Azure ML
- **Scaled synthetic data generation**: Designed novel algorithm simulating alternative splicing to generate 70,000 non-functional protein sequences for LLM training, optimized using PySpark for distributed processing
- **Deployed production ML platform**: Built and deployed *SplicePath*, an Azure cloud-based ML pipeline for modeling exon/transcript impact on pathway activity, adopted by cross-functional oncology teams for experimental design and hypothesis generation across 20+ drug targets
- **Advanced model interpretability**: Led development of graph interpretation algorithm for *SplicePath* to explain transcript-pathway effects via molecular interactions using COBRApy, with results visualized through NetworkX
- **Democratized ML access**: Designed no-code interface via Microsoft Fabric for non-technical users, improving adoptation of *SplicePath* adoption across oncology teams

Integrative Biology and Predictive Analytics Lab

UC Davis, CA

Postdoctoral Researcher in Dr. Tagkopoulos Lab

Sep 2020 - July 2022

- **Pioneered personalized nutrition ML**: Developed machine learning models for personalized dietary recommendations by integrating chemical composition data from FoodAtlas, 5,000+ samples from the American Gut Project, and microbiome metabolic models from AGORA2; performed community-scale metabolic modeling using MICOM to simulate individual-specific microbial responses
- **Led multi-omics drug tolerance research**: Applied ML to multi-omics data to identify genetic determinants of isobutyl acetate (IBA) tolerance, in collaboration with an experimental evolution lab
- **Cross-disciplinary collaboration**: Led ML development on a systems biology project in collaboration with the veterinary school to uncover how bats tolerate viral infections using proteomics data; co-authored the successful grant proposal

Sinopia Biosciences (Drug Discovery Startup)

San Diego, CA

Machine Learning Consultant

Oct 2019 - Jan 2020

- **Optimized drug discovery pipeline**: Preprocessed and batch-corrected drug response metabolomics data from 1,000+ samples, implementing robust quality control procedures
- **Developed predictive biomarkers**: Built ML models to identify metabolic signatures distinguishing drug mechanisms of action, enabling more efficient compound screening

Systems Biology Research Group

UC San Diego, CA

Graduate Student Researcher in Dr. Palsson's Lab

Sep 2015 - Sep 2020

- **Achieved exceptional research impact**: Published 4 first-author papers—including 2 in Nature Communications—and co-authored 10 additional papers with 900+ total citations
- **Gained international recognition**: Research received F1000 recommendation, UCSD press release, and two invited talks at international AI conferences
- **Advanced antimicrobial resistance research**: Built pangenome from 1,595 bacterial pathogens and trained SVMs to predict drug resistance, successfully identifying 33 known and 24 novel resistance genes
- **Pioneered interpretable ML in systems biology**: Developed first hybrid FBA-ML framework to model biochemical effects of mutations, matching state-of-the-art performance while improving biological interpretability for 3 anti-tuberculosis drugs

- **Discovered adaptive evolution principles**: Applied Independent Component Analysis (ICA) and statistical modeling to link *E. coli* adaptive mutations to transcriptomic and metabolic constraints, identifying 6 conserved transcriptomic strategies, 5 regulatory trade-offs, and multiple mutation-phenotype correlations

Selected Publications

2022	**Laboratory evolution reveals unifying systems-level principles of adaptation**, [*1st Author]	*mSystems*
2022	Adaptive laboratory evolution for improved tolerance of isobutyl acetate in *E. Coli*, [5th Author]	*Metab. Eng.*
2021	Independent component analysis recovers consistent regulatory signals..., [5th Author][*Github link*]	*PLoS Comp. Bio*
2020	**A biochemically-interpretable machine learning classifier for microbial GWAS**, [*1st Author][*Github link*]	*Nature Comm.*
2020	Machine learning with random subspace ensembles identifies antimicrobial..., [2nd Author]	*PLoS Comp. Bio.*
2019	Systems Biology and Pangenome of *Salmonella* O-Antigens, [4th Author]	*mBio*
2018	*Escherichia coli* B2 strains prevalent in inflammatory bowel disease patients..., [6th Author]	*BMC Syst. Bio.*
2018	**Machine learning and structural analysis of *M. tuberculosis* pangenome...**, [*1st Author] [*Github link*]	*Nature Comm.*
2018	**Updated and standardized genome-scale reconstruction of *M. tuberculosis*...**, [*1st Author] [*Github link*]	*BMC Syst. Bio.*
2018	Genome-scale metabolic reconstructions of multiple Salmonella strains..., [2nd Author]	*Nature Comm.*
2018	ssbio: a Python framework for structural systems biology, [6th Author] [*Github link*]	*Bioinformatics*
2018	The *Staphylococcus aureus* Two-Component System AgrAC Displays Four Distinct..., [4th Author]	*Front. in Microbio.*
2018	iCN718, an Updated and Improved Genome-Scale Metabolic Network Reconstruction..., [2nd Author]	*Front. in Genetics*

Education

University of California, San Diego *San Diego, CA*

PhD in Bioengineering, Advisor: Bernhard Ø. Palsson *2015 - 2020*

Thesis: *Biologically-Interpretable Machine Learning for Microbial Genomics*

University of California, Davis *Davis, CA*

B.S. in Civil and Environmental Engineering *2010 - 2015*

Conferences

2019	**Invited Talk:** Artificial Intelligence in Genomics, *<Association of Gene Diagnostics Meeting>*	*Oslo, Norway*
2019	**Invited Talk:** Tuberculosis and Machine Learning, *<4th Turning the Tide of AMR Meeting>*	*Potsdam, Germany*
2018	**Poster:** COBRA 2018, *<5th Conference on Constraint-Based Reconstruction and Analysis>*	*Seattle, Washington*
2017	**Poster:** NIH NIAID UO1 Conference, *<UO1 Antimicrobial Resistance Conference>*	*Ann Arbor, Michigan*

Leadership & Mentorship

2024	Mentored UVA undergraduate, Saumil Chukka, during his data science internship	*Envisagenics*
2021	Mentored 3rd year Veterinary Science PhD student on her analysis of bat proteomics data	*UC Davis*
2018	Mentored 1st year PhD student, Jason Hyun, on his first thesis project	*UC San Diego*
2018	Managed 16 students and 5 group projects for a graduate level systems biology course	*UC San Diego*

Teaching & Training

2018	BENG 125. Modeling and Computation in Bioengineering	*UC San Diego*
2018	BENG 123: Dynamic Simulation in Bioengineering	*UC San Diego*
2017	BENG 212: Systems Biology and Bioengineering	*UC San Diego*
2012	ENGR 422: Engineering Mechanics, Statics	*Sac. City College*
2010	Private math tutor for Davis Senior High School students	*Davis*

References

Bernhard Ø. Palsson, Ph.D.
Distinguished Professor
Bioengineering
Jacobs School of Engineering
University of California, San Diego
La Jolla, CA 92093-0411
palsson@ucsd.edu

Martin Ackermann, Ph.D.
Co-Founder and CTO
Envisagenics
30-02 48th Ave, Suite 140,
Long Island City, NY 11101
makerman@envisagenics.com

Ilias Tagkopoulos, Ph.D.
Associate Professor
Computer Science
College of Engineering
University of California, Davis
1 Shields Ave, Davis, CA 95616
itagkopoulos@ucdavis.edu