Erol **Kavvas**

DATA SCIENTIST · SYSTEMS BIOLOGIST · MACHINE LEARNING RESEARCHER

☐ (+1) 530-220-2141 | ✉ eskavvas@gmail.com | 🏠 erolkavvas.github.io | 🐙 erolkavvas | 💼 erolkavvas

Summary

Data scientist with +10 years of experience developing AI and mechanistic models for functional genomics. Proven track record in building and deploying scalable, interpretable ML solutions for genomics, drug/target 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, Neural Nets, Linear models, SVMs, Interpretable ML, LIME, SHAP, Tree-based Methods, Decomposition Methods, Ensemble Methods, Network Modeling, Recommender Systems, Model Optimization
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

- 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 of 20+ drug targets.
- Led development of a graph interpretation algorithm for *SplicePath* to explain transcript-pathway effects via molecular interactions; cellular network was built using COBRApy and results visualized with NetworkX.
- Designed a no-code interface via Microsoft Fabric for non-technical users and improved internal adoption of *SplicePath* across oncology teams.
- Developed *SpliceImpact*, a transformer-based LLM model for predicting per-residue functional effects from protein sequence; input embeddings generated by ESM2 and optimized with Azure ML GPUs.
- Designed an algorithm simulating alternative splicing to generate 70,000 non-functional protein sequences used to train the *SpliceImpact* LLM model; optimized using PySpark.

Integrative Biology and Predictive Analytics Lab
UC Davis, CA

POSTDOCTORAL RESEARCHER
Sep 2020 - July 2022

- Developed ML models for personalized dietary recommendations by integrating chemical composition data from FoodAtlas, 5000+ 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.
- Applied ML to multi-omics data to identify genetic determinants of isobutyl acetate (IBA) tolerance, in collaboration with an experimental evolution lab.
- 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

- Preprocessed and batch-corrected drug response metabolomics data from 1,000+ samples.
- Built ML models to identify metabolic signatures distinguishing drug action.

Systems Biology Research Group
UC San Diego, CA

GRADUATE STUDENT RESEARCHER
Sep 2015 - Sep 2020

- Wrote 4 first-author publications and coauthored 10 papers; cited 900+ times.
- Research received an F1000 recommendation, a UCSD press release, and two invited talks at AI conferences.
- Built pangenome from 1,595 pathogens and trained SVMs to predict drug resistance; identified 33 known and 24 novel resistance genes.
- Pioneered first hybrid FBA–ML framework to model biochemical effects of mutations; matched SOTA performance while improving interpretability and experimental hypotheses for 3 anti-tuberculosis drugs.
- Applied ICA and statistical modeling to link *E. coli* adaptive mutations to transcriptomic and flux-level constraints; identified 6 conserved transcriptomic strategies, 5 regulatory tradeoffs, 4 mutation-flux correlates, and 8 mutation-transcriptome correlates.

Selected Publications ————————————— (*Full list at my Google Scholar*)

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

Thesis: *Biologically-Interpretable Machine Learning for Microbial Genomics* *2015 - 2020*

University of California, Davis *Davis, CA*

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

Invited Talks & Posters

Artificial Intelligence in Genomics *Potsdam, Germany*

Presenter for <Association of Gene Diagnostics Meeting> *Sep, 2019*

Tuberculosis and Machine Learning *Oslo, Norway*

Presenter for <4th Turning the Tide of Antimicrobial Resistance Meeting> *Nov. 2019*

COBRA 2018 *Seattle, Washington*

Poster for <5th Conference on Constraint-Based Reconstruction and Analysis> *Oct. 2018*

NIH NIAID UO1 Conference *Ann Arbor, Michigan*

Poster for <NIH NIAID UO1 Antimicrobial Resistance Conference> *Oct. 2017*

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