Jacob Krol | curriculum vitae

Computational Biologist/Professional RA

GitHub: jakekrol
LinkedIn: jacob-krol
Email: jacob.krol@cuanschutz.edu
Denver, Colorado & Saline, Michigan; United States

Education

Non-degree seeking - University of Colorado Anschutz Medical Campus- Aurora, CO — Fall 2023

- BIOS 7747: Machine Learning for Biomedical Applications, graduate course offered by the University of Colorado School of Public Health

B.Sc. in Computational Neuroscience - Michigan State University - East Lansing, MI — 2020-2022

- GPA: 3.89/4.0
- Graduated 'With Honor'
- Semester awards: Dean's List

Math and Science Transfer Program - Washtenaw Community College - Ann Arbor, MI — 2018-2020

- GPA: 3.52/4.0
- Semester awards: Honor Roll
- Transferred

Publications

MolEvolvR: A web-app for characterizing proteins using molecular evolution and phylogeny.
Jacob D Krol*, Joseph T Burke*, Samuel Z Chen*, Lo M Sosinski*, Faisal S Alquaddoomi, Evan P Brenner, Ethan P Wolfe, Vincent P Rubinetti, Shaddai Amolitos, Kellen M Reason, John B Johnston, Janani Ravi
bioRxiv 2022.02.18.461833; doi: https://doi.org/10.1101/2022.02.18.461833 (* co-primary author)

Accurately predicting AR in ESKAPE pathogens using ML. **Jacob D Krol**, Ethan P Wolfe, Evan P Brenner, Keenan Manpearl, Vignesh Sridhar, Joe Burke, Jill Bilodeoux, Janani Ravi (*In preparation.*).

Presentations & posters

Great Lakes Bioinformatics Conference - MolEvolvR a web-app for protein characterization- McGill University, Montreal, CA — 2023

- Discussed the development and future directions of a web-app I develop: http://jravilab.org/molevolvr

Great Lakes Bioinformatics Conference - How and when to build a web-app or R package?- McGill Universtiy, Montreal, CA — 2023

- Hosted a 4 hour in-person workshop on how to build an R package using automation: devtools and usethis. A github repo for a sample R package I wrote is located at http://www.github.com/jravilab/iprscanr.

R/Bioconductor - MolevolvR a web-app for protein characterization - Boston University, Boston, MA — 2023

- Presented on MolEvolvR application methodology for the Cancer and Evolution talks section

CU DBMI Retreat - Robust machine learning-based classification of antimicrobial resistance in high-impact pathogens- CU Anschutz, Aurora, CO — 2023

- Presented on highly performant AMR ML classifcation models

Funding

- Submitted Fall 2023 NSF GRFP (in-review)
- Contributed to minor revisions NIH NIAID U01 grant submission (awarded; PI: Janani Ravi)
- Contributed revision and figures to NIH NIAID R01 grant submission (in-review; PI: Janani Ravi)
- Submitting Winter 2023 Department of Energy Computational Science Graduate Fellowship (in-preparation)

Professional Experience

Information Sciences Professional (PRA) - Dept. of Biomedical Informatics, Center for Health Artificial Intelligence, University of Colorado Anschutz School of Medicine - JRaviLab - Aurora, CO 2022-2023

- Redesigned the front and backend of MolEvolvR, a web app for protein characterization (http://jravilab.org/molevolvr; Krol, et al., 2023 _bioRxiv_; [DOI](https://doi.org/10.1101/2022.02.18.461833))
- Developed a machine learning (ML) pipeline for classifying antimicrobial resistance (AMR) in bacterial strains (publication in-prep for early 2024 submission)
- Listed in 'Acknowledgements' section of publication *provisionally accepted* at *mSystems ([DOI](https://doi.org/10.1101/2020.09.24.301986))
- Presented research talks and gave programming workshops at 2 international conferences (Bioconductor 2023 and GLBIO 2023)
- Mentored 3 PhD students and 1 undergraduate to assist in omics data featurization, model development, and model outcome analysis
- Developing two R packages for AMR project (source code for Bioconductor and data for ExperimentHub)
- Used University of Colorado Anschutz School of Medicine (CU Anschutz) Alpine high performance computing cluster (HPC) to aggregate, transform, and train ML models on over 100GB of omics data
- Containerization of AMR data collection code and the R packages I developed, and worked with Faisal Alquaddoomi on containerizing the front-end, back-end, and slurm instance for the MolEvolvR web-application
- Resolved over 50 GitHub issues for multiple lab repositories and signficantly cleaned up consolidated projects with multiple repositories
- Mentored by members of CU Anschutz Department of Biomedical Informatics (DBMI) software engineering team
- Assisted in hiring process (interviewing and feedback) for various lab positions: post-doctoral and research assistants
- Implemented various ML approaches and evaluation techniques: logistic regression, gradient boosting machines, random forests, linear discriminant analysis, stratified-cross-validation, class-weighting, hyperparamter searching, and evaluating auROC, balanced accuracy, etc., on hold-out dataset
- Hosted deparment wide workshops on Bash, Git, R package development, and ssh-workflow basics
- Assisted in writing and generating figures for lab grant proposals and publications
- Performed system administrator duties (e.g., server onboarding, dependency/user/data/resource management) for our lab's webserver

Student Research Assistant II - Computational Mathematics Science and Engineering program, Michigan State University - Krishnan Lab & Malmstrom Lab - East Lansing, MI 2022

- Learned, presented on, and implemented statistical methods on viral protein datasets: fisher test, logistic regression, & principal component analysis
- Trained machine learning classifiers to predict plant virus' host types; also, trained models to predict plant virus taxonomy.
- Featurization of protein sequences and data wrangling with Pandas, Biopython, NumPy, and R (Tidyverse + Bioconductor) packages for biological feature extraction
- Further preparing data with one-hot-encoding and z-score normalization
- Analyzing and visualizing model performance with Matplotlib/Seaborn & Scikit-learn performance metrics

Professional Summary

- Develop bioinformatics tools specializing in applying ML to large omics datasets using R, Python, and Bash/shell
- Use large, public databases (e.g., BV-BRC, NCBI, InterPro) with applied machine learning to study the relationship between genotypes and phenotypes
- Mentored 3 PhD students and 1 undergraduate student
- Developed a protein analysis web-app MolEvolvR
- Presented research talks and workshops at international conferences, hosted department-wide workshops, and presented research posters
- Exceptional IT knowledge: experienced with high performance computing (HPC), version control (GitHub), dependency management, containerization, package development, web-application hosting, using web APIs for data, etc.
- Basic familiarity of other common languages like Javascript, C, Java, and Perl
- Apply statistical methods to analyze omics data: supervised classification, fisher's test, under/over-sampling, cross-validation, rank-based hypothesis testing, etc
- Assist in various sub-tasks throughout the lab such as one-on-one mentorship with undergraduates and graduate students

Peer-Mentees

PhD Students

- Keenan Manpearl; 2023 (Computational Biosciences program; CU Anschutz)
- Jill Bilodeaux; 2023 (Computational Biosciences program; CU Anschutz)
- Charmie Vang; 2023 (Biophysical Sciences program; CU Anschutz)

Undergraduates

- Ethan Wolfe; 2023 (Biochemistry & Molecular Biology (BS) with CMSE and additional minors)

Project summaries

Machine learning classification of Antimicrobial Resistance: Classification antibiotic resistance of bacterial strains using supervised learning

- Project lead, first author
- Developed a computational pipeline to gather bacterial genomes and AMR data from public databases, featurize the genomes, develop ML models to classify AMR, and analyze model features to discover novel AMR genes
- Mentored 3 PhD students and 1 undergraduate to assist in omics data featurization, model development, and model outcome analysis
- Supervised learning with large (over 10k bacterial isolates) sample sizes and tackling class imbalance with weighted loss functions and undersampling
- Model evaluation using confusion matrix performance metrics such as balanced accuracy, auROC, etc.
- Implemented Fisher's Exact test for nearly a million genes to determine significant presence/absence in the binary classes (resistant/susceptible phenotypes) yielding ranked gene lists for AMR contribution
- Developed the omics data featurization, ML, and non-ML pipelines as Bioconductor R package with planned submission in early 2024

  - All source code for data wrangling, machine learning, and presentation is already installable as an R package on `https://github.com/jravilab/amR` (private until submission)
- Planned submission for datasets and results to Bioconductor's ExperimentHub in early 2024
- Future directions:
  - Build a sequence database of top resistance genes by clustering across species to address the uninterprebility of gene cluster assignment across species using species-wise pangenomics tools
  - Design a web-application to showcase results of ML models in classifying AMR for various drugs/species
  - Design a web-application which allows submission of bacterial genomes for AMR classification

References

Janani Ravi - University of Colorado Anschutz School of Medicine- Aurora, CO 2022-Present

- janani.ravi@cuanschutz.edu
- Assistant Professor, Principal investigator of JRaviLab

Faisal Alquaddoomi - University of Colorado Anschutz Anschutz School of Medicine- Aurora, CO 2022-Present

- faisal.alquaddoomi@cuanschutz.edu
- IT Principal Professional

Arjun Krishnan - Michigan State University- East Lansing, MI 2022

- arjun.krishnan@cuanschutz.edu
- Associate Professor, Principal invesigator of Krishnan Lab