William **Connell**

Research Scientist, Machine Learning · Ph.D.

(+1) 707-529-8516 | wconnell93@gmail.com | wconnell.github.io | wconnell | will-connell-26412352 | @wilstc

Summary

At Vevo Therapeutics, I led the initial machine learning initiatives from seed funding. I developed the core scientific hypotheses around discriminative and generative modeling for a phenotype-based drug discovery platform. I am interested in computational methods that practically influence drug discovery, particularly through a systems perspective. I enjoy working with quickly moving, collaborative teams and communicating with the broader community about industry developments.

Education

University of California, San Francisco — *San Francisco, CA*

PhD in Pharmaceutical Sciences and Pharmacogenomics — *2018 - 2022*

- Advisor: Michael Keiser
- Committee: Hani Goodarzi, Luke Gilbert

University of California, Los Angeles — *Los Angeles, CA*

BS in Microbiology, Immunology and Molecular Genetics — *2012 - 2016*

Experience

Vevo Therapeutics — *San Francisco, CA*

Research Scientist, Machine Learning — *September 2022 - January 2024*

- Developed and executed strategy for a scRNA-seq foundation model (scFM), expanding the pretraining corpus 65x and significantly improving target retrieval, sensitivity prediction, and biological recapitulation tasks
- Conceived and delivered a multimodal model aligning transcriptomes and chemical structures, achieving a 10x increase in top-k accuracy
- Initiated the ML strategy for chemical graph generation using transcriptional profiles for phenotype-guided chemical design
- Led a proposal accepted into the AWS Gen AI Accelerator program (<1%), securing $300k compute credits and industry recognition
- Managed contractors, recruited three FTEs, and mentored an intern, growing the ML team by 50%

Department of Pharmaceutical Chemistry, UCSF — *San Francisco, CA*

Graduate Researcher advised by Michael Keiser — *March 2019 - December 2022*

- Pioneered the application of self-supervised learning to scRNA-seq, advancing the pretrain/finetune framework in biological modeling
- Developed an *in silico* model for chemical probing (+10% baseline), finding new indication responses and potential ferroptosis drug targets
- Identified a genomic biomarker to stratify mAb ustekinumab response in psoriasis, advancing personalized treatment decisions

AI Research Group, Invitae — *Remote*

Computational Biology Research Intern — *May 2021 - August 2021*

- Aided in developing a hierarchical Bayesian model to enhance polygenic risk scoring accuracy
- Involved in comprehensive software and pipeline engineering for efficient management and analysis of large-scale datasets

UCSF Innovation Ventures — *San Francisco, CA*

Catalyst Awards Intern — *September 2018 - September 2019*

- Assessed scientific translation potential of a diagnostic gene expression biomarker panel, collaborating with corporate strategy experts
- Headed a team of four scientists to develop a Target Product Profile, resulting in a $100k translational funding award
- Performed product development feasibility studies, yielding a comprehensive report and pitch to venture investors

Datacamp — *Remote*

Project Developer — *November 2019 - January 2019*

- Crafted and delivered a data science project for a premier online educational platform
- Developed an interactive data analysis course in R, *Data Science for Social Good: Crime Study*
- Successfully engaged over 2000 learners, achieving a course rating of 4.7/5

Publications

Journal Articles

DNA-Diffusion: Leveraging Generative Models for Controlling Chromatin Accessibility and Gene Expression via Synthetic Regulatory Elements

Lucas Ferreira DaSilva, Simon Senan, Zain Munir Patel, Aniketh Janardhan Reddy, Sameer Gabbita, Zach Nussbaum, Cesar Miguel Valdez Cordova, Aaron Wenteler, Noah Weber, Tin M. Tunjic, Talha Ahmad Khan, Zelun Li, Cameron Ray Smith, Matei Bejan, Lithin Karmel Louis, Paola Cornejo, **William Connell**, Emily S. Wong, Wouter Meuleman, Luca Pinello

bioRxiv (2024). Cold Spring Harbor Laboratory, 2024. DOI: 10.1101/2024.02.01.578352

Learning chemical sensitivity reveals mechanisms of cellular response

William Connell, Kristle Garcia, Hani Goodarzi, Michael J. Keiser
bioRxiv (under review) (Aug. 2023). Cold Spring Harbor Laboratory, 2023. DOI: 10.1101/2023.08.26.554851

Genome-wide association study of ustekinumab response in psoriasis
William Connell, Julie Hong, Wilson Liao
Frontiers in Immunology 12 (Jan. 2022). 2022. DOI: 10.3389/fimmu.2021.815121

A single-cell gene expression language model
William Connell, Umair Khan, Michael J. Keiser
Learning Meaningful Representations of Life Workshop, NeurIPS (Oct. 2022). 2022. DOI: 10.48550/arXiv.2210.14330

Predicting cellular drug sensitivity using conditional modulation of gene expression
William Connell, Michael J. Keiser
Learning Meaningful Representations of Life Workshop, NeurIPS 2020 (Dec. 2020). 2020. DOI: 10.1101/2021.03.15.435529

Helical antimicrobial peptides assemble into protofibril scaffolds that present ordered dsDNA to TLR9
Ernest Y. Lee, Changsheng Zhang, Jeremy Di Domizio, Fan Jin, **William Connell**, Mandy Hung, Nicolas Malkoff, Veronica Veksler, Michel Gilliet, Pengyu Ren, Gerard C.L. Wong
Nat. Commun. 10.1 (Mar. 2019) p. 1012. 2019. DOI: 10.1038/s41467-019-08868-w

A single-cell transcriptomic atlas of human neocortical development during mid-gestation
Damon Polioudakis, Luis Torre-Ubieta, Justin Langerman, Andrew G. Elkins, Xu Shi, Jason L. Stein, Celine K. Vuong, Susanne Nichterwitz, Melinda Gevorgian, Carli K. Opland, Daning Lu, **William Connell**, Elizabeth K. Ruzzo, Jennifer K. Lowe, Tarik Hadzic, Flora I. Hinz, Shan Sabri, William E. Lowry, Mark B. Gerstein, Kathrin Plath, Daniel H. Geschwind
Neuron 103.5 (Sept. 2019) 785–801.e8. 2019. DOI: 10.1016/j.neuron.2019.06.011

Presentations

Quantifying the similarity of transcriptomic states in cancer
William Connell, Michael J. Keiser
CZI Neurodegeneration Challenge Network 2020 Annual Meeting. Poster. 2020. Virtual.

Target deconvolution across phenotypic space
William Connell, Garrett Gaskins, Michael J. Keiser
Northern California Computational Biology Symposium. Oral. 2019. Davis, CA.

Extracurricular Activity

Behind BioML *Remote*
Substack, writing *Current*
- An emerging business model in drug development
- Scaling biology – part 1
- Computation through the lens of biology
- ML for target identification
- The evolution of ML in early-stage drug development

SynBioBeta *San Jose, CA*
Foundation Models for Bio Panel *2024*
- Moderator for panel discussion: technical overview, role of data, model evals, and impact on industry verticals

ICLR *Remote*
MLGenX Workshop Reviewer *2024*
- Machine Learning for Genomics Explorations workshop (reviewer)

OpenBioML *Remote*
DNA Diffusion, Contributing Member *August 2023 - Present*
- Contributed to OS research focusing on *in silico* generative sequence design
- Led the creation of an oracle model for assessing DNA sequences, improving baseline model 18x
- Coordinated scientific and technical efforts among a diverse group of global collaborators

Skills

programming	python [numpy, pandas, pytorch, scikit-learn, huggingface, wandb, lightning, pytorch-geometric, rdkit, hydra], AWS, R, bash, git, plink, snakemake
OS contributions	enformer-finetune, pytorch-metric-learning, scikit-learn sprint (WiMLDS)
model repos	ChemProbe, Exceiver