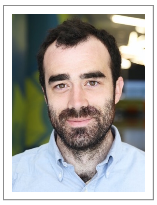
Rory Donovan-Maiye

Vashon, WA 98070
✉ donovanr@gmail.com
donovanr.github.io



Professional Profile

Delivering frontier results at the intersection of biology and machine learning

- I train foundation models in disease biology and cellular / molecular function using imaging, omics, and literature to solve real world biological problems and deliver value to scientists in the lab.
- Technical lead and company-wide key opinion leader for large model training and applying large/foundation models to biological data.
- Skilled data scientist and biology/physics/ML generalist, hands-on with analysis, partnering with SMEs and applying foundation models to specific research questions to create reproducible analysis workflows and user friendly apps.
- Effective communicator skilled at translating complex technical concepts across scientific disciplines and into tangible insights for both lay and professional audiences / collaborators.

Employment

2021– **Senior Scientist, Machine Intelligence**, *Novo Nordisk*
ML / GenAI driven multimodal data analysis and modeling applied to drug discovery and target identification.

2017–2021 **Research Scientist / Senior Scientist**, *Allen Institute for Cell Science*
Computational biology and machine learning – integrative methods for single cell modeling

2016–2017 **Postdoctoral Fellow**, *Institute for Systems Biology*
Hood–Price Lab – personalized medicine & wellness, mouse and human genomics, causal learning

Education

2011–2016 **Ph.D.**, *Carnegie Mellon–University of Pittsburgh Program in Computational Biology*
Systems Biology & Machine Learning

2004–2005 **M.S.**, *University of Washington*
Physics

2000–2004 **B.A**, *Reed College*
Physics

Recent Projects

Enhancing generative perturbation models with LLM-informed gene embeddings
We incorporate prior knowledge through embeddings derived from both text and biological sequence Large Language Models (LLMs), effectively informing our predictive models with a deeper biological context. Our models achieve state-of-the-art performance in predicting the outcomes of single-gene perturbations.

Multimodal generative models of *in vitro* cellular perturbations
Generating image-based and omics phenotypic responses to cellular perturbations in silico, to integrate multiple screening modalities and provide in silico predictions to efficiently guide further wet-lab experiments. “Stable diffusion but for biology,” if you like.

Multiscale modality-agnostic molecular embeddings using chemical language models
Motivated by the need for good embeddings/representations of modified peptides, we train an all-atom language model on a broad cross section of therapeutic modality space (proteins, RNA, peptides, small molecules) using a hierarchical / stochastic tokenization scheme for universally applicable molecular embeddings.

Universal cellular image embeddings
Self-supervised representation learning of cellular morphology / perturbation state across cell types and image modalities to build a SOTA foundation vision model capable of jointly embedding all *in vitro* cellular image modalities (Brightfield, Cell Painting, IHC, etc.).

Rescuing recombinant protein expression in mammalian cells using protein language models
Coupling inverse folding sequence generators with an in-house LLM-based expression prediction model, we rescue non- and low-expressing proteins of interest by proposing variant sequences that reliably express at 10-100x the level of the parent sequence.

Pharmacokinetic property prediction from molecular structure

Accurately predicting half-life and clearance rate for monomer and multimer peptide / small protein therapeutics from historical data to reduce the need for pre-clinical animal trials. We use multitask Gaussian processes to leverage and factor all high dimensional data relationships of interest, and make uncertainty-quantified predictions for novel molecular entities.

A byte-pair encoding (BPE) tokenizer for SELFIES strings

I implement a parallel / highly efficient BPE tokenizer in Rust with Python bindings exposed via PyO3/Maturin optimized for the SELFIES molecular representation syntax that outperforms off the shelf methods while producing significantly cleaner / more compact tokenized molecular representations.

A Python library for high throughput cellular image data analysis

An end-to-end Python library with templated HPC workflows, a modular CLI, pre-prepared analysis notebooks, and import/export image QC enables rapid, reproducible, and verifiable microscope → knowledge distillation, as well as efficient data curation for training large foundational models.

Computational Skills

Strong abilities and interests:

- Pytorch-focused machine learning / Generative AI on biological data: cellular images, sequencing data, protein sequence and structure.
- Cloud and on-prem GPU-centric model training / data and compute orchestration + scaling (AWS, NVIDIA DGX Cloud, Azure Machine Learning, large-scale Slurm-based on-prem)
- Open data & open science: reproducible and efficient data analysis / ML pipelines: novel algorithms, data wrangling and code/data/model versioning, machine provisioning and distributed compute.
- Python ML stack: NumPy / SciPy / pandas / Polars / Numba / PyTorch / JAX / Hugging Face / Lightning / Hydra / scikit-learn / Dask / Prefect / Quilt / Snakemake / Hatch / uv / Ruff / Pyright / Black / Click / Typer

Selected Publications

Kaspar Märtens, Rory Donovan-Maiye, and Jesper Ferkinghoff-Borg. Enhancing generative perturbation models with LLM-informed gene embeddings. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

Matheus P Viana, Jianxu Chen, Theo A Knijnenburg, Ritvik Vasan, Calysta Yan, Joy E Arakaki, Matte Bailey, Ben Berry, Antoine Borensztejn, Eva M Brown, et al. Integrated intracellular organization and its variations in human ips cells. *Nature*, 613(7943):345–354, 2023.

Rory M Donovan-Maiye, Jackson M Brown, Caleb K Chan, Liya Ding, Calysta Yan, Nathalie Gaudreault, Julie A Theriot, Mary M Maleckar, Theo A Knijnenburg, and Gregory R Johnson. A deep generative model of 3d single-cell organization. *PLOS Computational Biology*, 18(1):e1009155, 2022.

Kaytlyn A Gerbin, Tanya Grancharova, Rory M Donovan-Maiye, Melissa C Hendershott, Helen G Anderson, Jackson M Brown, Jianxu Chen, Stephanie Q Dinh, Jamie L Gehring, Gregory R Johnson, et al. Cell states beyond transcriptomics: Integrating structural organization and gene expression in hipsc-derived cardiomyocytes. *Cell Systems*, 12(6):670–687, 2021.

Cory C Funk, Alex M Casella, Segun Jung, Matthew A Richards, Alex Rodriguez, Paul Shannon, Rory Donovan-Maiye, Ben Heavner, Kyle Chard, Yukai Xiao, et al. Atlas of transcription factor binding sites from encode dnase hypersensitivity data across 27 tissue types. *Cell reports*, 32(7), 2020.

Rory M Donovan-Maiye, Christopher J Langmead, and Daniel Zuckerman. Systematic testing of belief-propagation estimates for absolute free energies in atomistic peptides and proteins. *Journal of chemical theory and computation*, 2017.

Ramu Anandakrishnan, Zining Zhang, Rory Donovan-Maiye, and Daniel M Zuckerman. Biophysical comparison of atp synthesis mechanisms shows a kinetic advantage for the rotary process. *Proceedings of the National Academy of Sciences*, 113(40):11220–11225, 2016.

Rory M. Donovan, Jose-Juan Tapia, Devin P. Sullivan, James R. Faeder, Robert F. Murphy, Markus Dittrich, and Daniel M. Zuckerman. Unbiased rare event sampling in spatial stochastic systems biology models using a weighted ensemble of trajectories. *PLoS Computational Biology*, 2016.

Andrew J. Sedgewick, Ivy Shi, Rory M. Donovan, and Panatiotis V. Benos. Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Genomics*, 2014.

Rory M. Donovan, Andrew J. Sedgewick, James R. Faeder, and Daniel M. Zuckerman. Efficient stochastic simulation of chemical kinetics networks using a weighted ensemble of trajectories. *The Journal of Chemical Physics*, 139(11):115105, 2013.