AI FOR BIOMEDICAL DATA SCIENCE ACROSS SCALES

A Statistical AI Symposium

TUESDAY, MAY 5, 2026 · 1–5 PM *Fred Hutch Cancer Center Steam Plant*

Agenda

Subject to Change

1:00 PM–1:15 PM

Introduction

Jingyi Jessica Li

Professor and Program Head, Biostatistics Program · Fred Hutch Cancer Center

---

1:15 PM–1:45 PM

The Office of the Chief Data Officer and Fred Hutch Data Supports

Jeff Leek

Chief Data Officer and AVP· Fred Hutch Cancer Center

Amy *Paguirigan*

Deputy Chief Data Officer · Fred Hutch Cancer Center

ABSTRACT

> We will provide an overview of the Office of the Chief Data Officer and the data systems, training, and governance support we provide to the Fred Hutch ecosystem.

1:45 PM – 2:15 PM

Toward a Holistic and Dynamic Stem Cell State Landscape

Susanne Rafelski

Senior Director, Quantitative Biology · Allen Institute

ABSTRACT

> Establishing a conceptual framework for holistic cell states and state transitions is an essential step towards building a multiscale understanding of cell biology. At the Allen Institute for Cell Science, we are working to define and map stem cell states in the context of organizational and transitional dynamics across diverse conditions. Over the past decade, we have developed experimental and analytical pipelines and tools to quantify and computationally represent the intracellular organization of cells from 3D microscopy images and applied these to a variety of biological questions. Building upon this foundation, our new CellScapes initiative aims to understand how cells organize themselves across

scales to form complex cell communities and tissues and use this understanding to build predictable and programmable tissue systems. The Institute has applied AI in various ways including the development of 3D bioimage analysis approaches, interpretable representation learning of intracellular structures, and extracting cell state transition dynamics from timelapse movies. At the core of our application of these modern AI approaches is creative framing of the question so it can be answered quantitatively and careful attention to the necessary application-appropriate validation toward biological interpretation of the data.

2:15 PM – 2:45 PM

Thinking Like a Biostatistician, Building Like an ML Researcher: A Case Study with Antibodies

Erick Matsen

Professor and Associate Program Head, Computational Biology · Fred Hutch Cancer Center

ABSTRACT

Modern AI/ML is a stunningly good toolbox. But the loss functions used often come from other fields, such as language and vision, and don't always fit the biology. Biomedical statisticians, by contrast, obsess over matching the model to the problem. In this talk I'll pitch "thinking like a biostatistician, building like an ML researcher" through a case study on antibodies.

Antibodies evolve by mutation and selection. The dominant tool for analyzing them is the protein language model: a transformer trained on hundreds of millions of sequences with a masked language modeling objective. But that objective is a density estimator over observed sequences, and density is not fitness! These models end up learning codon tables, hypermutation rates, and germline identity, none of which are relevant to function.

Our fix borrows from classical mutation-selection modeling. The Deep Amino acid Selection Model (DASM) factors the likelihood into a mutation term (small CNN, fit on neutrally evolving sequences) and a selection term (transformer, fit on millions of parent-child pairs from reconstructed phylogenies). Because the two are separated by construction, the selection model only quantifies function. It beats existing protein language models on experimental benchmarks while being an order of magnitude smaller and orders of magnitude faster — and surfaces real biology we didn't expect.

https://elifesciences.org/articles/109644

https://elifesciences.org/articles/111070

https://doi.org/10.1093/molbev/msaf186

2:45 PM

Break

3:00 PM – 3:30 PM

Foundation Models for Chromatin 3D Architecture and Proteomics Mass Spectrometry

William Noble

Professor, Department of Genome Sciences, Department of Computer Science and Engineering · University of Washington

ABSTRACT

In machine learning, a foundation model is a large-scale model that is trained in a self-supervised fashion on massive data and that can be easily fine-tuned to solve a variety of downstream tasks. In this talk, I will describe two recent efforts in my lab to develop foundation models. The first is designed to operate on 3D chromatin architecture data, represented as a DNA-DNA contact matrix produced by assays such as Hi-C. We show that the model, trained using techniques borrowed from image analysis, can be used for computing experimental similarity measures, improving effective sequencing depth, detecting chromatin loops, and predicting related, linear epigenomic measurements. The second model was initially designed for de novo sequencing of peptides from proteomics tandem mass spectrometry data. We show that the encoder learned by this model can be re-purposed for a variety of downstream tasks, including prediction of spectrum quality and chimericity, as well as presence of post-translational modifications.

3:30 PM – 4:00 PM

Lightning Talk: From Histology to Molecular Insight: Transfer Learning for Molecular Biomarker Prediction in Prostate Cancer

Lucas Liu

Postdoctoral Fellow and AI Researcher · Fred Hutch Cancer Center

ABSTRACT

Artificial Intelligence (AI) has the great potential to advance pathology and oncology by enabling automated cancer detection, disease grading, and treatment response prediction directly from histopathological slides. However, training an AI model for specific clinical applications such as molecular alteration prediction from scratch is still challenging because labeled training data is often limited. Transfer learning and foundation models are the key to resolve this challenge. Foundation models capture rich and generalizable representations from pretraining on millions of histopathological images, and transfer learning adapts that knowledge to specific tasks.

In this talk, I will describe our transfer learning-based framework in which we fine-tune a sequence of foundation models for predicting clinical-actionable molecular biomarkers directly from histopathological images in prostate cancer. We started with predicting microsatellite instability high/mismatch repair-deficient (MSI-H/dMMR) status. We trained and validated the model across three independent datasets. The results suggest that our transfer learning-based approach has consistent performance across diverse clinical contexts, and it substantially outperforms the training-from-scratch approach. To improve the interpretation of the framework, we extended this framework with a clustering-based approach that incorporates spatial morphological structure through tile clusters and spatial statistics. Our preliminary results suggest that spatial representations achieve comparable performance to deep learning-based approaches and may be particularly informative for certain molecular alterations..

Lightning Talk: Transparent AI Agent for Differential Expression: Marrying Statistical Rigor with Natural Language Interfaces

Ziqi Rong
CS PhD Student · University of Washington

ABSTRACT

We present DE-Rigor-Agent, an AI-driven framework that enables differential expression analysis for bulk transcriptomics data through natural language input, while enforcing rigorous, expert-defined statistical assessment and decision-making. DE methods are context-dependent, varying with data properties, and there is no one-size-fits-all approach. Rather than operating as a black box that forces a single method pipeline, the agent systematically evaluates input data against methodological assumptions, screens out inappropriate methods, and executes statistically appropriate methods in parallel. By decoupling user interaction from method selection, our system eliminates subjective cherry-picking, delivering rigorous, reproducible, and fully auditable reports for differential expression analysis.

4:00 PM – 4:30 PM

From Models to Decisions: Bridging Statistical AI and Clinical Practice through Interpretable Systems and Workforce Readiness

Sonali Tamhankar
Senior Clinical Data Scientist · Fred Hutch Cancer Center

ABSTRACT

Achieving meaningful impact from healthcare AI requires more than strong models - it demands coordination across methodology, interpretability, deployment, and workforce readiness. In this talk, we articulate a layered framework for translating statistical innovation into clinical decision-making, illustrated through concrete applications and measurable outcomes.

At the foundation lies statistics: a rigorous mathematical core that comes into full expression through modern machine learning. We present a Statistical AI project predicting which sickle cell patients are at elevated risk for emergency visits, demonstrating how principled modeling can address clinically relevant and operationally important questions.

Prediction alone, however, is not enough. Interpretable AI enables clinicians and stakeholders to act on statistical findings, while generative AI methods can further contextualize results and surface clinically

meaningful insights. These layers are essential for transforming abstract models into tools that can credibly inform patient care.

Equally essential is the sociotechnical layer: healthcare AI is fundamentally a team science. Even well-designed, interpretable systems fail without an AI-literate workforce capable of engaging with models, questioning assumptions, and participating in organizational AI strategy. Workforce readiness is therefore not an implementation detail, but a prerequisite for sustainable AI impact.

To address this, we present our AI literacy initiative at Fred Hutch, the "AI: A-to-Z" course, designed to empower healthcare and research professionals as informed partners in AI-enabled decision-making. Using validated AI literacy instruments, we observed statistically significant improvements in participant comfort with AI. Notably, confidence in serving as partners in implementing organizational AI strategy increased from 2.4 pre-course to 4.7 post-course on a seven-point Likert scale.

Together, these examples illustrate that impactful healthcare AI emerges not from models alone, but from the deliberate integration of statistical rigor, interpretability, deployment strategy, and workforce empowerment.

4:30 PM

Panel Discussion and Q&A

Jonathan Bricker
Professor · Fred Hutch Cancer Center Cancer Prevention Program

Youyi Fong
Professor · Fred Hutch Cancer Center Biostatistics, Bioinformatics, and Epidemiology Program

Jeff Leek
Postdoctoral Fellow and AI Researcher · Fred Hutch Cancer Center Biostatistics Program

Lucas Liu
Postdoctoral Fellow and AI Researcher · Fred Hutch Cancer Center Biostatistics Program

William Noble
Professor · University of Washington Department of Genome Sciences and Department of Computer Science and Engineering

Ali Shojaie
Professor and Chair · University of Washington Department of Biostatistics

Wei Sun
Professor · Fred Hutch Cancer Center Biostatistics Program

This symposium was kindly supported by the Cancer Center Support Grant (NIH P30 CA015704).