Precision Medicine: Integrating Large-Scale Data and Intermediate Phenotypes for Understanding Health and Treating Disease

Steven E. Brenner
University of California, Berkeley
Email: brenner@compbio.berkeley.edu

Nilah M. Ioannidis
University of California, Berkeley; University of California, Santa Cruz
Email: nilah@berkeley.edu

Tayo Obafemi-Ajayi
Missouri State University
Email: tayoobafemiajayi@missouristate.edu

Anne O'Donnell-Luria
Boston Children's Hospital; Harvard Medical School; Broad Institute of MIT and Harvard
Email: odonnell@broadinstitute.org

The field of precision medicine has undergone rapid development over the past three decades, driven by advances in high-throughput molecular profiling, large-scale electronic health data, and computational modeling. The central objective is to refine disease risk prediction, diagnosis, and treatment strategies by incorporating genetic, molecular, environmental, and clinical information into individualized care. However, the effective integration of these heterogeneous data sources presents substantial analytical challenges. The 2026 Precision Medicine session of the Pacific Symposium on Biocomputing (PSB) highlights computational methods that bridge large-scale biological data and intermediate phenotypes, emphasizing approaches that advance mechanistic understanding, risk prediction, and clinical utility. The contributions span multi-modal risk modeling, biomarker discovery, and causal inference frameworks, demonstrating the breadth and depth of research in computational precision medicine.

Keywords: Precision medicine; Multi-modal analysis; Biomarker discovery; Machine learning.

1. Introduction

Precision medicine seeks to tailor prevention, diagnosis, and treatment to individual patients by integrating biological and clinical information beyond traditional risk factors. Early progress emerged from the application of genomics to targeted therapy, most notably in oncology. More recently, the scope has expanded to include diverse high-throughput data modalities, including transcriptomics, proteomics, metabolomics, spatial omics, and single-cell technologies, combined with electronic health records (EHRs) and lifestyle data. The proliferation of these data provides an

unprecedented opportunity to identify intermediate phenotypes—molecular, cellular, or physiological traits that mediate between genotype and disease outcome—and to use them for predictive and mechanistic modeling. At the same time, the scale and complexity of such datasets introduce challenges in data harmonization, statistical power, and interpretability. Machine learning (ML) and deep learning (DL) methods have emerged as powerful strategies for extracting meaningful signals from heterogeneous data. Yet clinical translation requires models that are interpretable, mechanistically grounded, and generalizable across diverse populations.

The PSB has long provided a forum for the precision medicine community, and the 2026 session emphasizes computational and methodological innovations for integrating large-scale data and intermediate phenotypes to advance individualized healthcare. The accepted papers cover several broad themes: (i) integrating multi-modal data to improve disease risk prediction, (ii) modeling heterogeneity and disease subtypes, (iii) biomarker discovery, and (iv) modeling mechanism and causality. The papers in this session collectively advance the practice of precision medicine by integrating diverse data types to improve prediction, diagnosis, and mechanistic understanding. From generative transformers trained on millions of longitudinal health trajectories to multi-modal imaging-proteomics-genomics models and causal inference frameworks, the research highlights the power of computational biomedicine to uncover heterogeneity in disease processes and inform individualized care.

2. Session Contributions

2.1. *Multi-modal data integration for personalized risk prediction*

German et al. [1] introduce the Next Health Event (NHE) model, a generative transformer trained on the longitudinal health records of more than seven million participants. By incorporating demographic information, polygenic risk scores for nearly 300 traits, and longitudinal BMI into sequential health trajectories, the model can forecast disease progression at the individual level. Their ablation studies underscore the unique contributions of both genetic risk and longitudinal BMI, demonstrating the potential of integrated generative models for predictive health.

Venkatesh et al. [2] develop a multi-modal framework for coronary microvascular disease risk prediction. They integrate PET imaging-derived endotypes, plasma proteomics, and polygenic risk scores to classify patients more effectively. This approach captures disease heterogeneity, offering the first demonstration of imaging-based endotyping combined with genomic and proteomic data for this condition.

Cardone et al. [3] present an integrative framework that combines polygenic scores with clinical, lifestyle, and social determinants of health to predict heart failure. Their results show that combined predictors significantly outperform individual components, highlighting the clinical value of multi-domain integration for early detection and intervention.

2.2. *Modeling biological heterogeneity and disease subtypes*

Rifat et al. [4] present BioLM-NET, an interpretable deep learning framework that combines prior biological knowledge—including protein–protein and protein–DNA interactions and pathway

information—with large language model–derived gene embeddings to interpret gene expression and DNA methylation data. Applied to colorectal and breast cancer, glioblastoma, colon cancer, and Alzheimer's disease datasets, BioLM-NET improves prediction of subtype, progression and metastasis while enhancing interpretability.

Seagle et al. [5] explore the biological heterogeneity of apparent treatment-resistant hypertension by clustering associated variants into distinct cardiometabolic profiles, which may reflect underlying mechanisms such as metabolic dysregulation, inflammation, and vascular reactivity. This approach provides a route to disentangling heterogeneous mechanisms of complex traits and stratifying patients for tailored interventions.

2.3. *Diagnostics and biomarker discovery from digital data*

Sethi et al. [6] use ProtoECGNet, a prototype-based neural network trained for multi-label ECG classification, to learn physiologically meaningful prototypes that serve as interpretable digital biomarkers. Their results show that these prototypes correlate with clinical outcomes for both cardiovascular and systemic conditions, providing clinically meaningful and transferable intermediate phenotypes.

Choi et al. [7] develop a deep learning approach for diagnosing Postural Orthostatic Tachycardia Syndrome using continuously collected wearable ECG and accelerometer data. Their method captures physiological variability in daily life, offering a less burdensome alternative to traditional tilt-table tests and demonstrating the potential of wearable devices in digital phenotyping.

Tamura et al. [8] propose an approach for identifying sparse phenotype embeddings from histopathological images using weakly supervised neuron selection and sparse autoencoders applied to dense representations generated by pathology foundation models. Their approach results in interpretable phenotypes with strong performance on tumor patch identification.

2.4. *Causality and mechanistic modeling*

Sriram et al. [9] introduce DeepDiff-SHAP, an interpretable deep learning framework for subgroup-specific causal inference. Using conditional SHAP values, their method uncovers nonlinear and subgroup-specific causal changes in relationships that are obscured in population-averaged analyses. Applications to diabetes and sepsis datasets reveal clinically meaningful, comorbidity-specific mechanisms.

Jiang et al. [10] generate causal networks connecting genetic variants to biological processes and disease phenotypes based on mechanistic statements extracted from the literature. Their model reconstructs causal molecular mechanisms and predicts consequences for variants lacking prior annotation, enabling mechanistic interpretation of genomic variation.

References

1. C. German, S. Shringarpure, P. Dibaeinia, J. Ashenhurst, B. L. Koelsch, A. Auton, A. A. Khan. Integrating Polygenic Risk Improves Generative Forecasting of Disease Trajectories. *Pacific Symposium on Biocomputing* (2026).
2. R. Venkatesh, T. Cherlin, Penn Medicine BioBank, M. D. Ritchie, M. A. Guerraty, S. S. Verma. Integrating Imaging-Derived Clinical Endotypes with Plasma Proteomics and External Polygenic Risk Scores Enhances Coronary Microvascular Disease Risk Prediction. *Pacific Symposium on Biocomputing* (2026).
3. K. M. Cardone, D. Kim, M. D. Ritchie. Integrating Polygenic Scores with Clinical, Lifestyle, and Social Risk Factors to Improve Heart Failure Risk Prediction. *Pacific Symposium on Biocomputing* (2026).
4. J. I. M. Rifat, T. Tabashum, M. M. Rahman, M. F. Mokter, S. Engala, S. Bozdag. BioLM-NET: an interpretable deep learning model combining prior biological knowledge and contextual LLM gene embeddings on multi-omics data to predict disease. *Pacific Symposium on Biocomputing* (2026).
5. H. M. Seagle, J. Kim, A. T. Akerele, VA Million Veteran Program, A. Hung, J. N. Hellwege, T. L. Edwards. Polygenic partitioning of apparent treatment resistant hypertension implicates distinct biological processes in pathogenesis. *Pacific Symposium on Biocomputing* (2026).
6. S. Sethi, D. Chen, M. C. Burkhart, N. Bhandari, B. Ramadan, B. Beaulieu-Jones. Prototype Learning to Create Refined Interpretable Digital Phenotypes from ECGs. *Pacific Symposium on Biocomputing* (2026).
7. H. Choi, N. Matsumoto, X. Li, D. Teodorescu, A. Kote, M.-J. Yang, X. Liu, M. E. Hernandez, J. H. Moore, G. G. Hernandez, P.-S. Chen. Deep Learning-based Classification of Patients with Postural Orthostatic Tachycardia Syndrome using Wearable ECG and Accelerometer Data. *Pacific Symposium on Biocomputing* (2026).
8. K. Tamura, Y.-z. Zhang, Y. Okubo, S. Imoto. Patch-level phenotype identification via weakly supervised neuron selection in sparse autoencoders for CLIP-derived pathology embeddings. *Pacific Symposium on Biocomputing* (2026).
9. A. Sriram, S. Kim, J. A. Carcillo, H. J. Park. DeepDiff-SHAP: Interpretable deep learning for subgroup-specific causal hypothesis generation using conditional SHAP. *Pacific Symposium on Biocomputing* (2026).
10. J. Jiang, P. Radivojac, B. M. Gyori. Literature-driven extraction and computational prediction of causal statements linking genetic variants to biological processes, pathways and phenotypes. *Pacific Symposium on Biocomputing* (2026).