Precision Medicine: Computational and AI Approaches for Understanding the Biological Basis of Health and Disease

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

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

Mohamed Omar
Cedars-Sinai Medical Center
Email: Mohamed.Omar@csmc.edu

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

Precision medicine increasingly depends on computational methods capable of transforming large, heterogeneous biological and clinical datasets into information that is meaningful at the level of individual patients. Recent advances in artificial intelligence, self-supervised learning, generative modeling, and biologically informed machine learning provide new opportunities not only to improve prediction, but also to interpret molecular mechanisms, characterize disease heterogeneity, quantify uncertainty, and connect biological variation to clinical outcomes. The 2027 Precision Medicine session of the Pacific Symposium on Biocomputing (PSB) highlights these developments across genomic variation, immune recognition, proteomics, transcriptomics, medical imaging, spatial omics, electronic health records, and clinical decision-making. The eleven accepted papers collectively demonstrate how computational and AI approaches can connect increasingly complex biomedical measurements to mechanistic understanding and patient-specific prediction and stratification.

Keywords: Precision medicine; Artificial intelligence; Multi-omics; Mechanistic modeling.

1. Introduction

Precision medicine seeks to tailor prevention, diagnosis, and treatment to individual patients by integrating information from their genetic, molecular, cellular, physiological, and clinical characteristics. The continued expansion of genomics, transcriptomics, proteomics, spatial profiling, medical imaging, and electronic health records has created unprecedented opportunities to characterize human disease at multiple biological scales. At the same time, the scale and heterogeneity of these data create substantial computational challenges involving representation, integration, generalization, interpretation, and translation into clinically meaningful information.

For PSB 2027, the Precision Medicine session places particular emphasis on emerging computational and artificial intelligence approaches for understanding the biological basis of health

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

and disease. These include foundation models, biological sequence models, generative AI, integration of multiple biological data types, patient-specific phenotype prediction, and mechanistic disease modeling. These methods offer opportunities to move beyond conventional predictive modeling by identifying biologically meaningful representations, elucidating mechanisms connecting molecular variation to disease, and developing computational approaches that remain useful despite heterogeneous or limited data.

The eleven papers accepted for the 2027 session illustrate the breadth of computational and AI approaches, including generative AI and foundation models, for translating large-scale biological data into insights on genomic variation, molecular function, disease risk, and therapeutic opportunities. The contributions span four broad themes: (i) AI-driven genomic and molecular interpretation; (ii) interpretable representation learning for molecular and multi-omic risk prediction; (iii) mechanistic and uncertainty-aware clinical stratification; and (iv) computational imaging, spatial omics, and generative pathology. Together, they highlight complementary paths toward linking biological understanding with individualized prediction and care.

2. Session Contributions

2.1. *AI-driven genomic and molecular interpretation*

Abderrazzaq et al. [1] address the persistent challenge of interpreting in-frame insertion and deletion variants in clinical genomics. The authors combine computational variant-effect predictions with calibrated high-throughput functional evidence across ten disease-associated genes and map both sources of evidence to standardized clinical classification criteria. Their integrated approach resolves up to 55% of existing ClinVar variants of uncertain significance in these genes and preclassifies as many as 5,500 previously unreported in-frame indels. The work demonstrates how experimentally derived and computational evidence can be systematically combined at scale while also revealing remaining challenges associated with indel length and insertion-versus-deletion direction.

Naz et al. [2] introduce GraphTCR-Mamba, a heterogeneous graph-learning framework for predicting T-cell receptor-epitope interactions. The method represents TCRs and epitopes as distinct node types, incorporates sequence similarity and known interaction structure, and uses a Mamba selective state-space component for efficient global information propagation. The authors additionally evaluate cold-start settings and show that performance is substantially stronger for unseen TCRs than for unseen epitopes or when both interaction partners are unseen. The study demonstrates both the potential of relational sequence modeling and the importance of stringent generalization testing for molecular interaction prediction.

2.2. *Interpretable representation learning for molecular and multi-omic risk prediction*

Nam et al. [3] develop a self-supervised protein-token Transformer for learning patient representations from plasma proteomic measurements when the set of available proteins differs across datasets. Using data from 53,014 UK Biobank participants and evaluating prospective predictions across 144 diseases, the model maps variable sets of measured proteins into a common

fixed-dimensional representation. Under partial coverage, the learned representations also outperformed coefficient-truncated LASSO models and provided additional predictive information beyond clinical covariates for most focused diseases. These findings address an important portability problem for proteomic risk models across heterogeneous measurement settings.

Venkatesh et al. [4] present a pathway-based multi-omic framework for atrial fibrillation risk that transforms individual genomic information into biologically interpretable pathway representations and incorporates tissue-specific transcriptomic signals. Pathway-level modeling substantially improves discrimination over conventional polygenic risk scores, while individual-level analyses reveal heterogeneous contributions from pathways involving electrical signaling, cardiac development, and DNA repair. Comparison with the AlphaGenome foundation model provides orthogonal support that highly ranked pathways are enriched for regulatory effects in cardiac tissues. Rather than representing inherited susceptibility as a single polygenic score, the framework provides patient-specific biological mechanisms underlying risk.

Yasenpoor et al. [5] introduce a supervised Bayesian Poisson factorization approach for jointly learning disease-relevant gene programs (DRGPs) and predicting personalized disease risk from RNA-seq data. The framework decomposes expression into latent gene programs, optionally incorporates prior pathway knowledge, and links program activity and other risk covariates to disease using sparse logistic regression. Evaluations using single-cell RNA-seq data for COVID-19 and whole-blood GTEx data show competitive predictive performance while recovering biologically coherent gene programs associated with disease. The study reveals that molecular risk prediction can be constructed around inherently interpretable biological representations rather than relying only on post-hoc explanations of complex predictors.

2.3. *Mechanistic and uncertainty-aware clinical stratification*

Keat et al. [6] develop a computational and statistical framework that uses AI-derived computed tomography phenotypes as intermediate traits linking genetic variation to disease. Using genomics-linked CT scans from the Penn Medicine Biobank, the authors conduct genome-wide association analyses of organ volume and attenuation phenotypes and then use mediation analysis to examine whether these imaging-derived phenotypes explain genetic associations with disease. Their analyses identify bone attenuation as a mediator connecting genetic variants to fracture risk and support liver attenuation as an intermediate phenotype linking variation near PNPLA3 to metabolic dysfunction-associated steatotic liver disease. The framework illustrates how routinely acquired clinical imaging can help move from genetic association toward mechanistic hypotheses about disease biology.

Priya et al. [7] address heterogeneity in metabolic dysfunction-associated steatotic liver disease by developing a partitioned machine-learning framework for predicting rapid fibrosis progression. The approach identifies clinically meaningful patient subgroups and then develops subgroup-specific models integrating longitudinal laboratory measurements, diagnostic histories, and genetic information. These models improve discrimination compared with globally trained models. The authors further convert model explanations into an interpretable Partitioned Rapid Fibrosis Progression Score (pRFPS), which identifies 22-25% of rapid progressors classified as low risk by

conventional clinical tools and stratifies patients into groups associated with subsequent hepatocellular carcinoma and liver transplantation.

Sanchez et al. [8] broadened the concept of precision medicine by focusing on uncertainty in estimated kidney function rather than treating estimated glomerular filtration rate as a precise point measurement. Using nationally representative NHANES data from 44,360 adults, the authors propagate known uncertainty in creatinine-based eGFR through clinically relevant decision thresholds. They further show that simple distance from a threshold can approximate decision uncertainty for direct GFR boundaries but becomes less adequate for multivariable, patient-specific risk thresholds. The work emphasizes that individualized medicine may require personalizing the uncertainty surrounding a measurement as well as the measurement itself.

2.4. *Computational imaging, spatial omics, and generative pathology*

Prathifkumar et al. [9] investigate predictive uncertainty in virtual spatial transcriptomics of colorectal cancer using Bayesian deep learning. A heteroscedastic ResNet-34 trained on 45 Visium colorectal cancer tissue sections predicts expression for 991 spatially variable genes, while Monte Carlo dropout separates uncertainty into epistemic, or model-related, and aleatoric, or intrinsic, components. Variation in uncertainty across genes, biological pathways, and tissue architectures provides a way to distinguish predictions that may improve with additional data or better models from molecular programs whose relationship to tissue morphology may be intrinsically more difficult to recover.

Takahashi et al. [10] apply self-supervised and transfer learning to the problem of causative-gene identification in lissencephaly, where the rarity of the disease makes large, labeled imaging cohorts difficult to obtain. The authors pretrain a three-dimensional image representation using 3D-SimCLR on FOMO300K, a dataset of approximately 300,000 unlabeled brain MRI scans, and subsequently fine-tune the model using a much smaller lissencephaly cohort to distinguish DCX, LIS1, and other causative genes. The study illustrates how large-scale unlabeled imaging data can be leveraged to improve rare-disease modeling despite severe limitations in labeled sample availability.

Thiriveedi et al. [11] examine the reverse of the more common virtual spatial-transcriptomics problem by generating histopathologic images from spatial molecular measurements. Their conditional latent diffusion model uses spatial gene expression, tissue architecture, and cell-type composition to generate skin histology from 39,579 Visium spots across sixteen patients. Generated tissues preserve measurable characteristics of epithelial, stromal, and immune composition. When used as generative augmentation, the synthetic samples produce only modest overall improvements in spatial gene-expression prediction and tissue-architecture classification, but larger benefits emerge within particular tissue microenvironments. The results provide an initial benchmark showing that biologically conditioned generative pathology can preserve tissue information while its downstream benefits are strongly context dependent.

References

1. H. Abderrazzaq, R. Stewart, M. Tejura, A. E. McEwen, S. Casadei, M. W. Snyder, N. Smith, A. F. Rubin, D. M. Fowler, L. M. Starita, P. Radivojac. Classification of In-frame Indels at Scale Using Calibrated Functional and Computational Evidence. *Pacific Symposium on Biocomputing* (2027).
2. E. Naz, A. Fan, W. Lin, P. Luo. GraphTCR-Mamba: Heterogeneous Graph Learning with Mamba for TCR-Epitope Binding Prediction. *Pacific Symposium on Biocomputing* (2027).
3. Y. Nam, T. M. Westbrook, J. Woerner, J. Joo, M. E. Lee, M. McKeague, A. E. Baxter, Shwetank, M. Ionita, S. A. Apostolidis, A. R. Greenplate, E. J. Wherry, D. Kim. Self-supervised Plasma Proteomic Representations for Prospective Disease Prediction Across Varying Protein Availability. *Pacific Symposium on Biocomputing* (2027).
4. R. Venkatesh, R. Deo, T. P. Cappola, Penn Medicine BioBank, M. D. Ritchie, D. Kim. Pathway Modeling of Genomic and Tissue-Specific Transcriptomic Architecture Identifies Personalized Mechanisms of Atrial Fibrillation Risk. *Pacific Symposium on Biocomputing* (2027).
5. T. Yasenpoor, M. Hosseini, A. J. Matz, B. Zhou, D. Aguiar. Learning Interpretable Disease-relevant Gene Programs for Predicting Personalized Disease Risk. *Pacific Symposium on Biocomputing* (2027).
6. K. Keat, D. Y. Zhang, L. Carruth, J. Duda, C. Beeche, C. Kripke, H. Sagreiya, W. Witschey, Penn Medicine Biobank, Regeneron Genetics Center, D. Rader, S. S. Verma, A. Verma. A Computational and Statistical Framework Leveraging AI-Derived CT Phenotypes for Causal Mediation Effects Between Genetic Variants and Disease. *Pacific Symposium on Biocomputing* (2027).
7. T. S. Priya, H. Yan, T. L. Haynes, K. J. Wangensteen, A. J. Acosta, K. N. Lazaridis, E. W. Klee, S. Tian, D. Yao. pRFPS: A Subgroup-Specific Interpretable Risk Score for Predicting Rapid Fibrosis Progression in Metabolic Dysfunction-Associated Steatotic Liver Disease. *Pacific Symposium on Biocomputing* (2027).
8. M. Sanchez, J. Cummings, J. A. Diao, B. Y. Reis, R. D. Balicer, N. Dagan, A. K. Manrai. Quantifying the Clinical Implications of Imprecision in Kidney Function Estimates. *Pacific Symposium on Biocomputing* (2027).
9. T. Prathifkumar, N. Mohamed, M.-K. Le, A. Kakkera, Z. Azher, X. Liu, F. Kolling IV, L. Perreard, S. Palisoul, L. J. Vaickus, K. Coleman, M. Omar, A. Gertych, M. Guindi, K. Yao, J. C. Figueiredo, J. Levy. Decomposing Predictive Uncertainty in Virtual Spatial Transcriptomics of Colorectal Cancer Using Bayesian Deep Learning. *Pacific Symposium on Biocomputing* (2027).
10. N. Takahashi, Y. Sato, M. Kato, A. Yamaguchi. Application of Transfer Learning for Lissencephaly Gene Identification Using a Large-Scale Brain MRI Dataset and Self-Supervised Learning. *Pacific Symposium on Biocomputing* (2027).
11. V. Thiriveedi, A. Chaphalkar, G. Srinivasan, C. Miles, Z. Azher, A. Kakkera, F. Kolling IV, L. Perreard, S. Palisoul, L. J. Vaickus, M. Davis, M. LeBoeuf, I.-C. Liao, M.-K. Le, K. Yao, J. Levy. Spatial Transcriptomics-Informed Histopathologic Image Generation for Modeling the Skin Tissue Microenvironment. *Pacific Symposium on Biocomputing* (2027).