Cancer Biomarkers AI and Bioinformatics Workshop
August 13-15, 2024
California Institute of Technology
Pasadena, California

Hameetman Auditorium, Caltech

The *Cancer Biomarkers AI and Bioinformatics Workshop* was held at the California Institute of Technology in Pasadena, California, on August 13-15, 2024. It was co-hosted by NASA's Jet Propulsion Laboratory and Caltech on behalf of the National Cancer Institute's Early Detection Research Network (EDRN). The workshop leveraged the latest advances in bioinformatics, machine learning (ML), and artificial intelligence (AI) to discuss enhancing cancer biomarker discovery and validation. With the rapid growth of data-intensive research, tools such as generative AI and foundation models provide new capabilities in areas like image segmentation and data analysis. The workshop gathered experts in cancer science, informatics, and AI to address key challenges, such as structuring diverse datasets for analysis and ensuring reproducibility, interpretability, and scalability of AI-driven methods. Participants explored how shared tools and datasets, particularly from consortia like the Early Detection Research Network (EDRN), can be effectively applied to advance scientific discovery in cancer research.

The workshop included defining use cases for AI in biomarker research, discussing the state of the art and existing gaps, and developing recommendations to tackle challenges in reproducibility, data sharing, and computational needs. The event featured keynotes, discussions, and a half-day hackathon focused on real-world data from the EDRN's LabCAS

biomarker data commons. Themes included AI's application in biomarker discovery, methodology considerations, data preparation, and emerging capabilities like large language models, digital twins, and federated learning. This workshop aimed to facilitate partnerships to drive innovation in the field of cancer biomarkers between academia and with industry.

Day 1 August 13, 2024

Welcome, Dan Crichton, NASA Jet Propulsion Laboratory

The "AI Workshop Welcome" presentation, led by Dan Crichton at NASA/JPL/Caltech, highlighted JPL's role in advancing data science and artificial intelligence (AI) for both space exploration and cancer research. Dan emphasized the exponential growth of data and the challenges science faces in data collection, storage, and analysis, particularly to support reuse and structure data for AI. He discussed NASA's interesting challenges such as onboard analytics and distributed data systems and analysis. He used that to then cover NASA's collaboration with biomedical researchers, applying big data management techniques from planetary science to cancer biomarker research, and set the stage for an AI workshop focusing on leveraging AI in biomarker discovery, with sessions and a hackathon aimed at advancing tools and methodologies in this field.

Dan thanked both the program committee and the local organizing committee, both who provided immense support in planning and preparing for the meeting.

AI for Improving Early Detection of Cancer, Sudhir Srivasava, Ph.D., National Cancer Institute

The "AI for Improving Early Detection of Cancer" presentation by Dr. Sudhir Srivastava highlighted the potential of AI in advancing early cancer detection and diagnostics. It introduced the Early Detection Research Network (EDRN) and its objectives of fostering collaboration across research institutions, establishing biomarker validation standards, and supporting regulatory processes for the rapid clinical use of biomarkers. AI applications in this field included risk stratification, classification of early-stage cancers, and improved interpretation of imaging scans. The presentation emphasized AI's potential to address challenges such as differentiating cancerous from non-cancerous tissues, detecting small lesions, and reducing unnecessary biopsies.

Additionally, the presentation underscored the need for large, well-defined datasets and reproducible AI-based prediction models for accurate early detection. It showcased examples such as pancreatic cancer, where current imaging techniques were insufficient, and outlined the integration of clinical, synthetic, and simulated data to train AI models. The presentation concluded by highlighting the importance of data sharing, synthetic data generation, and

collaborative efforts in AI model development to enhance early detection and improve patient outcomes.

Keynote: Making Sense of Biomedical Data Using the Tools of AI, Lior Pachter, Ph.D., Caltech

In the keynote "Making Sense of Biomedical Data Using the Tools of AI," Dr. Lior Pachter discussed the evolution of artificial intelligence and its growing impact on biomedical research. He began by tracing the origins of AI, from John McCarthy's formalization of the term to the development of early models like hidden Markov models for gene discovery. Pachter noted how foundational AI concepts, such as network structures and learning algorithms, enabled sophisticated machine learning techniques like neural networks. The presentation emphasized that Boltzmann Machines and their parameter-estimation techniques played a crucial role in the development of modern AI approaches used in biology today, highlighting the journey from AI's "winter" to its current explosion in utility.

Pachter then transitioned into explaining the vast scope of biomedical data that AI was now equipped to handle. This included data from genome sequencing, protein structures, imaging data, and electronic health records. With repositories like the Sequence Read Archive (SRA) holding more than 30 petabytes of sequencing data, the challenge lay in analyzing poorly structured data with missing metadata. AI tools, especially transformers and models like AlphaFold, had shown remarkable success, particularly in genomics and protein folding, enabling a deeper understanding of complex biological systems. The talk also discussed AI's role in sequence census experiments, where sequencing tools were repurposed to measure gene expression and detect conditions like trisomy 21.

In the final part of the presentation, Pachter explored the future of AI in biomedical research, focusing on areas like early virus detection, cancer research through a combination of image and sequence analysis, and microbiome studies. He envisioned AI becoming integral in uncovering patterns in vast datasets, helping researchers make sense of the complexity of biological systems. The presentation concluded with a vision of AI not only as a tool for data analysis but as a transformative force in understanding diseases, improving healthcare, and advancing biological research.

Session 1 - In Silico and Real-World Biomarker Discovery and AI, Matt Schabath, Ph.D., Moderator

The session "In Silico and Real-World Biomarker Discovery and AI," moderated by Dr. Matt Schabath, focused on the role of artificial intelligence (AI) in the identification of cancer-related biomarkers. The session opened with a historical context, tracing AI's roots back to the ideas of Alan Turing and John McCarthy, evolving from simple "if-then" rules to the current era of deep learning and high-performance computing. Today's AI systems were complex, leveraging multi-omics data and deep learning to support clinical decision-making. Two key definitions were

highlighted: machine learning (ML), which enabled computers to improve task performance without explicit instructions, and deep learning (a subfield of ML), which extracted complex features from raw data using neural networks with multiple layers.

The session also delved into important considerations for AI-driven biomarker discovery, such as ensuring clinical relevance, generalizability, integration into workflows, and fairness. Additionally, the need for explainable AI models was emphasized to build trust and manage AI's role in healthcare. Presentations included topics like using Vision-Transformer pipelines to explore cancer pathology, deep learning to predict platinum chemotherapy response, and innovative tools like OmicsCurveNet for transforming omics data into images for AI-driven biomarker discovery.

Incorporating Pathology Image and T-cell Receptor Repertoire Through AI Pipelines for Predicting Cancer Clinical Outcomes — Chad He, Ph.D., Fred Hutchinson Research Center

In the presentation "Incorporating Pathology Image and T-cell Receptor Repertoire Through AI Pipelines for Predicting Cancer Clinical Outcomes," Dr. He explored the integration of digital pathology and T-cell receptor (TCR) data using artificial intelligence (AI) to predict cancer outcomes. The presentation began by highlighting how digital pathology had advanced through deep learning (DL) applications, with successful use cases in tumor detection, cancer subtype classification, and clinical outcome prediction. Pathology images, coupled with TCR repertoire data, which was highly polymorphic and crucial to the adaptive immune response, provided a rich dataset for AI models. The potential of TCR data to predict the functional activity of tumor-infiltrating T cells was particularly emphasized as an emerging frontier in cancer prognosis.

The study analyzed data from The Cancer Genome Atlas (TCGA), specifically focusing on liver (LIHC) and skin (SKCM) cancer types, using AI models to predict new tumor events (NTE) and immune scores. Three model types were employed: pathology images alone, TCR data alone, and a combination of both, leveraging ResNet50 and Vision Transformer (ViT) architectures. The findings suggested that combining pathology images with TCR repertoire data often yielded more accurate predictions than using either alone. Additionally, the study demonstrated that models trained on common cancer types could effectively predict outcomes in rarer cancers, offering potential solutions for clinical decision-making in resource-limited settings.

Deep Learning AI Predicts HRD and Platinum Response from Histologic Slides, Erik N. Bergstrom, Ph.D., UC San Diego

The presentation "Deep Learning AI Predicts HRD and Platinum Response from Histologic Slides" by Dr. Erik N. Bergstrom discussed the use of artificial intelligence (AI) to predict Homologous Recombination Deficiency (HRD) and the response to platinum-based chemotherapy from histologic slides of cancer patients. The AI models analyzed mutational signatures found in the cancer cell's histology to identify HRD, which was often associated with better responses to platinum therapies. The method, called DeepHRD, significantly

outperformed traditional BRCA1/2 mutation tests by capturing a larger subset of responders, thus offering a more precise tool for predicting treatment outcomes in breast cancer patients. This approach could reduce the reliance on genomic testing while still providing accurate treatment predictions.

The model's effectiveness was validated across several large datasets, including TCGA and METABRIC, and was expanded to predict responses in ovarian cancer as well. The AI approach was able to identify approximately 2.5 to 4 times more responders than BRCA1/2 testing alone. DeepHRD specifically focused on predicting responses to platinum-based treatments but showed limitations in predicting responses to other chemotherapy types like taxanes. This research supported the potential for integrating AI models into clinical practice to complement existing pathology reports, speeding up diagnosis and treatment decisions, and improving accessibility to personalized cancer treatment.

OmicsCurveNet: Transforming Omics Data to Images for Biomarker Discovery Through Deep Learning, Zhen Zhang, Ph.D., Johns Hopkins University

The presentation "OmicsCurveNet: Transforming Omics Data to Images for Biomarker Discovery Through Deep Learning" by Drs. Shiyong Ma and Zhen Zhang introduced a novel approach to leverage deep learning for biomarker discovery by converting high-dimensional omics data into 2D images. The method, called OmicsCurveNet, used spatial mapping algorithms to transform omics data (such as RNA-Seq expression) into images, which were then processed by deep learning models, such as convolutional neural networks (CNNs), for classification. The CNNs identified important molecular features or "hot spots" through attention maps, which were projected back to specific molecular targets, functions, and pathways. This approach was designed to detect subtle, concordant signals from omics data that may have been missed by traditional statistical methods due to low prevalence in the study population.

The spatial mapping and deep learning model allowed OmicsCurveNet to capture biomarkers with functional significance that would otherwise fail conventional univariate global significance tests. The presentation highlighted its application using TCGA glioma data, identifying known cancer-related genes such as IDH1/2 and others. It also compared its discoveries with those from traditional methods, demonstrating its complementarity. While OmicsCurveNet showed strong potential in identifying novel biomarkers, the approach was continuously refined, particularly in enhancing spatial mapping algorithms to increase contrast and target identification. Overall, this method represented a significant advancement in using AI and bioinformatics to improve biomarker discovery in cancer research.

Keynote: Veridical Data Science and PCS Uncertainty Quantification, Bin Yu, Ph.D., University of California, Berkeley

The presentation "Veridical Data Science and PCS Uncertainty Quantification" by Dr. Bin Yu introduced the Veridical Data Science (VDS) framework, which emphasized producing

trustworthy and transparent data-driven results through rigorous documentation. The framework addressed the reproducibility crisis in scientific research, highlighting the need for proper uncertainty quantification (UQ) in data science. VDS integrated the PCS framework, which stood for Predictability, Computability, and Stability. The PCS framework promoted rigorous testing and evaluation of data science pipelines by using perturbations at various stages—such as data cleaning, modeling, and human judgment—allowing researchers to critically assess the stability and reliability of their results.

Through case studies, such as prostate cancer detection and genetic driver discovery in hypertrophic cardiomyopathy (HCM), the presentation demonstrated how PCS stress tests could improve predictive models and reveal stable biomarkers or genetic targets. These analyses leveraged deep learning and statistical models to aggregate results across perturbations, ensuring robustness in findings. PCS documentation was integral to this process, requiring transparency in decision-making throughout the data science life cycle (DSLC). Ultimately, VDS and PCS aimed to build trust in data-driven discoveries by offering a systematic approach to handling uncertainties and improving the interpretability of scientific results.

Session 2 - Biomarker Computation and Methodology Considerations for Applying AI — Moderator: Steven Skates, Ph.D., Massachusetts General Hospital, Harvard Medical School

Dr. Steve Skates introduced the session, "Biomarker Computation and Methodology Considerations for Applying AI," which focused on the challenges of using AI for biomarker discovery and emphasized the importance of robust computational techniques to handle complex and noisy omics data. He highlighted how AI models could enhance the detection of subtle, low-frequency molecular changes critical to early disease development, especially in cancer. The presentation also discussed the use of weakly supervised learning and advanced data transformation techniques, such as converting omics data into images, to efficiently identify biomarkers with less manual annotation. This approach enabled the identification of novel biomarkers by creating attention maps that highlighted significant molecular features, offering complementary insights beyond traditional statistical methods. A panel discussion was included at the end to address audience questions and discuss optimal methods when applying statistics vs machine learning.

Machine Learning and Biomedical Applications: Promise and Peril, Padhraic Smyth, Ph.D., UC Irvine

In the presentation "Machine Learning and Biomedical Applications: Promise and Peril," Dr. Padhraic Smyth explored both the exciting potential and significant challenges of applying machine learning, particularly deep learning, in biomedical contexts. He emphasized that deep learning had achieved remarkable success in fields like image classification, speech recognition, and natural language processing. In healthcare, deep learning could extract valuable feature representations from complex biomedical data (e.g., patient records, medical

imaging, and omics data), allowing for improved predictions in diagnosis and treatment. However, the success of these models often relied on the availability of large datasets and high computational power, and their utility in clinical settings remained to be fully validated.

Smyth also addressed the limitations and risks associated with deep learning in healthcare. While these models could achieve high accuracy, they often suffered from overconfidence in predictions, even when they were wrong, and were vulnerable to distribution shifts—where the model's performance deteriorated when applied to data from different settings (e.g., across hospitals). These models also lacked robustness compared to human experts and required more data than what was typically available in clinical practice. Additionally, challenges like algorithmic fairness, handling missing data, and managing commercial hype complicated the deployment of AI in medicine. Smyth concluded by stressing the importance of careful model evaluation, transparency, and skepticism in applying deep learning to healthcare.

Evaluating the Robustness of Features Generated by a Foundation Model from Lung Nodule Regions in CT with Different Reconstruction Parameters, Stephen Park, Ph.D., UCLA

The presentation "Evaluating the Robustness of Features Generated by a Foundation Model from Lung Nodule Regions in CT with Different Reconstruction Parameters" by Dr. Stephen Park and colleagues focused on assessing the consistency of imaging features extracted from lung nodule CT scans under varying reconstruction conditions. The team compared features generated by a deep learning-based foundation model with traditional handcrafted radiomic features (extracted using Pyradiomics) across six different CT image settings, including variations in slice thickness and reconstruction kernel. The aim was to evaluate how these different settings affected the robustness of the extracted features, which were critical for downstream tasks such as classifying nodules as cancerous or non-cancerous.

The findings revealed that the foundation model consistently extracted features with higher agreement across different image conditions compared to Pyradiomics. Although post-processing techniques like ComBat harmonization could reduce the variability of Pyradiomics features, their impact on predictive performance was inconsistent. Despite the foundation model generally performing better in classifying nodule outcomes, it also exhibited variability across different image conditions. The study highlighted the need for continued efforts to improve the robustness of imaging features to enhance their generalizability and clinical utility, particularly in the application of AI models in medical imaging.

No Winners: Performance of Lung Cancer Prediction Models in Screening-Detected, Incidental, and Biopsied Pulmonary Nodule Use Cases, Thomas Li, B.S., Vanderbilt University

The presentation "No Winners: Performance of Lung Cancer Prediction Models in Screening-Detected, Incidental, and Biopsied Pulmonary Nodule Use Cases" by Thomas Li evaluated various AI models used to predict lung cancer across different clinical scenarios. It focused on lung nodules detected through screening, incidental findings, and biopsies. The models examined included both traditional statistical models (like Brock and Mayo) and more advanced AI models (like Sybil and DLSTM). The results showed that no single model consistently outperformed the others across all settings. While AI models like Sybil performed well in screening scenarios, the performance declined when applied to incidental or biopsied nodules. This variability was attributed to differences in image acquisition settings, lung cancer subtypes, and patient characteristics across study sites.

The study also highlighted the challenges of achieving reproducibility and generalizability in AI models due to site-specific factors like CT scanner settings and population characteristics (e.g., smoking habits or environmental factors). To address this, the presentation suggested harmonizing imaging data and incorporating multimodal approaches that integrated clinical data with imaging for better predictions. The researchers emphasized that while retrospective studies offered insight into model performance, randomized controlled trials were still necessary to validate the clinical utility of these AI models in real-world practice.

Day 2 August 14, 2024

Keynote Address: AI and Computational Microscopy: Visual Biomarkers for Disease Progression Predictions — Yang Changhuei, Ph.D., Caltech

The presentation by Dr. Changhuei Yang focused on the intersection of AI, deep learning, and computational microscopy in advancing disease progression predictions, specifically through visual biomarkers. Yang highlighted the application of Fourier Ptychography (FP), a high-resolution imaging technique, in digital pathology. FP allowed for high-quality, all-in-focus imaging that significantly enhanced the potential of deep learning models. These models could detect complex patterns in medical images, such as predicting cardiovascular risks or detecting early-stage non-small cell lung cancer (NSCLC) progression, surpassing human capabilities. However, challenges such as focus and stain variations in digital pathology persisted, which Yang's team continued to address.

A key aspect of Yang's work was the integration of deep learning with AI-designed imaging systems, rather than adapting human-oriented designs. He argued that human vision and interpretation limited the potential of these technologies, while AI systems could handle the complexity of phase data and other imaging modalities that were challenging for human observers. His team demonstrated AI's ability to predict brain metastasis in NSCLC patients with a high level of specificity. Yang emphasized the paradigm shift towards leveraging

computational techniques to correct physical limitations in microscopy, thus enabling the use of less expensive equipment while improving accuracy and scalability in clinical settings.

Session 3 - Considerations in Data Preparation, Sharing, and Analysis — Moderator: Jennifer Beane, Ph.D., Boston University

Dr. Jennifer Beane introduced the session by addressing key considerations and needs in data sharing, preparation, and analysis in the context of AI and cancer research. She highlighted the significant advances in computational tools, driven by innovations in GPUs, cloud computing, and electronic medical records (EMR) systems. The presentation emphasized the increasing use of digitized radiology and pathology images, along with routine genomic profiling, to support cancer detection, diagnosis, and treatment. AI applications, such as lesion detection, risk prediction, diagnosis of benign or malignant conditions, and identification of treatment responses, were transforming cancer research by improving prognostic accuracy and personalizing patient care.

However, the integration of diverse data sources, such as histopathology slides, genomic data, and medical imaging, posed significant challenges, particularly in terms of data curation, harmonization, and ensuring patient privacy. The session explored various models for data sharing, including federated learning and multi-institutional agreements, which enabled collaborative cancer research across different institutions while protecting sensitive information. Speakers from institutions including the University of Chicago, Dana Farber Cancer Institute and UCLA presented case studies and practical lessons, providing insights into the creation of AI-ready datasets and the operational complexities of multi-center research initiatives. A panel discussion was included which discussed several questions from the audience.

Curating AI-Ready Datasets: The Pediatric Cancer Data Commons — Kaitlyn Ott, M.S., University of Chicago

The presentation titled "Curating AI-Ready Datasets: The Pediatric Cancer Data Commons" highlighted the challenges and opportunities of preparing datasets for AI in pediatric oncology. It underscored the importance of large, diverse, and interoperable data to facilitate AI research. The Pediatric Cancer Data Commons (PCDC) played a pivotal role in capturing the data by serving as a hub for researchers across various pediatric cancers, helping harmonize data from clinical trials, registries, and electronic health records. However, challenges remained, including the fragmentation of data, non-standardized formats, and the global spread of data sources, which complicated AI-ready dataset preparation. The PCDC addressed these challenges through the development of a common data model and consensus dictionaries, making data more accessible for AI applications.

The presentation also discussed the role of AI in pediatric oncology, particularly in diagnostic classification, risk stratification, and predicting treatment responses. A case study was presented on using convolutional neural networks (CNNs) to predict chemotherapy responses in

neuroblastoma patients using diagnostic imaging. The PCDC not only facilitated scientific research but also aimed to make high-quality data useful for patients and families by driving AI innovations that could improve outcomes for rare pediatric cancers. By increasing data interoperability and promoting international collaboration, the PCDC continued to advance AI-driven discoveries in pediatric oncology.

Lessons Learned in Federated Cancer Research Pilots — Michael Rosenthal, Ph.D., Dana Farber Cancer Institute, Harvard University

The presentation titled "Lessons Learned in Federated Cancer Research Pilots," delivered by Sahil Nalawade and Dr. Michael Rosenthal from the Dana-Farber Cancer Institute, outlined the implementation and benefits of federated learning (FL) in cancer research. Federated learning allowed researchers to train AI models across multiple institutions without sharing sensitive patient data by keeping data decentralized at the local institutions. This approach ensures data privacy while enhancing the generalizability of models trained on diverse datasets from different sites. The presentation emphasized the motivation for federated learning, including overcoming the limitations of centralized data lakes and ensuring robust, scalable solutions for early cancer detection, particularly in projects like the pancreatic cancer risk model.

The presentation also provided insights into setting up a federated learning system, the frameworks used, and pilot applications in cancer research. Examples included federated training models for clinical risk prediction and image segmentation across institutions like DFCI and MD Anderson. Results from these pilots demonstrated that federated models often outperformed centralized models in terms of accuracy and privacy preservation. The conclusion highlighted the benefits of federated networks, such as reducing administrative barriers, maintaining local control of data, and fostering collaboration across institutions, while also addressing challenges like data standardization and site compliance.

Practical Considerations in Preparing and Sharing AI-Ready Medical Images — William Hsu, Ph.D., UCLA

The presentation "Practical Considerations in Preparing and Sharing AI-Ready Medical Images" by Dr. William Hsu focused on the challenges and key factors involved in preparing large-scale, multimodal medical imaging data for AI applications. Hsu highlighted the importance of transparency in data collection, emphasizing that imaging data must align with the intended use and target population for AI models. The presentation discussed practical considerations, such as providing detailed metadata to maintain the quality and utility of the imaging data, addressing privacy concerns while ensuring data fidelity, and ensuring that the temporal relationships between different imaging datasets were maintained. Examples of large-scale imaging datasets, such as those used in prostate MRI studies, illustrated how these datasets could be leveraged for precision oncology.

Another key aspect of the presentation was the need for standardization and harmonization across institutions to ensure that AI models could work with diverse datasets. Hsu stressed the importance of collaboration and overcoming institutional barriers, such as approvals for data

sharing. The presentation also advocated for the use of "data cards" to systematically document the sources, collection methods, and metadata of imaging datasets to enhance transparency and reproducibility. These practical steps were essential for preparing AI-ready datasets that could support innovative research and drive advancements in early detection, diagnosis, and treatment of cancer and other diseases.

Session 4 - Emerging Capabilities and Methodologies in AI — Moderator: Ashish Mahabal, Ph.D., Caltech

The session "Emerging Capabilities and Methodologies in AI" was moderated by Dr. Ashish Mahabal. In introducing the session, he discussed key advancements in artificial intelligence (AI), focusing on various models and techniques. It covered tree-based models, neural networks, and foundational models like Generative Adversarial Networks (GANs), Variational AutoEncoders (VAEs), and Large Language Models (LLMs). Notable developments such as the Segment Anything Model (SAM), the role of the internet, processing power, and the growing importance of open-source libraries were emphasized. His presentation also touched on the increasing reliance on data and the importance of understanding probabilistic graphical models, deep learning, and transformers in the AI landscape.

Additionally, he discussed the benefits and risks of synthetic data, including improved data distributions but also the potential for biases and information leakage. The session also covered transparency in AI development, advocating for tools like data sheets and model cards to ensure ethical AI progress. The session included a discussion on distributed intelligence with privacy preservation, digital twins for personalized patient management, and hybrid extended reality data visualization, concluding with a panel discussion.

Harnessing Distributed Intelligence with Privacy Preservation: Federated Learning in Oncology — A Preliminary Simulation Study — Ghulam Rasool, Ph.D., Moffitt Cancer Center

The presentation titled "Harnessing Distributed Intelligence with Privacy Preservation: Federated Learning in Oncology" by Dr. Ghulam Rasool and his team explored how Federated Learning (FL) could address the challenges of data privacy in medical AI applications, particularly in oncology. Traditional AI models benefited from large datasets, but in medical domains, privacy concerns limited data sharing across institutions. FL enabled training of AI models across multiple institutions without sharing sensitive data. It worked by allowing each institution to train its own model locally and then send updates to a global model without transferring raw data. The presentation highlighted the use of homomorphic encryption (HE) to protect the process, allowing computations on encrypted data for enhanced security. Moreover, uncertainty estimation and model calibration techniques were emphasized to improve trustworthiness and performance, especially when deploying AI models in real-world, high-stakes environments like cancer screening.

The study's experimental setup used data from the National Lung Cancer Screening Trials (NLST) and tested three different data configurations—No Split (NS), Random Split (RS), and Cohort Split (CS)—to simulate realistic FL development. Results showed that FL models performed well, though homomorphic encryption slightly decreased accuracy in exchange for greater privacy protection. The NVIDIA FLARE platform was used for FL workflows, and the study demonstrated that FL, combined with encryption and uncertainty estimation, could be a viable solution for developing large-scale AI/ML models in healthcare while preserving patient privacy. Future work would focus on refining these techniques and expanding their application.

Digital Twin for Personalized Management of Patients with Prostate Cancer — Radka Stoyanova, Ph.D., University of Miami

The presentation titled "Digital Twin for Personalized Management of Patients with Prostate Cancer" by Dr. Radka Stoyanova and colleagues focused on utilizing a digital twin model to enhance personalized treatment for prostate cancer patients. A digital twin is a virtual representation of a patient, created using multimodal and multiscale data such as clinical parameters, imaging, blood biomarkers, and genomics. This model aimed to predict the aggressiveness of cancer progression and guide treatment strategies, particularly for those undergoing Active Surveillance. By leveraging various data sources, including MRI scans and pathology slides, the digital twin allowed clinicians to make more informed decisions regarding the timing and type of treatment, aiming to improve patient outcomes while minimizing unnecessary interventions.

The presentation also highlighted two trials, the Miami MAST trial and the MDSelect trial, which used MRI-guided selection for prostate cancer treatment. These trials compiled a comprehensive longitudinal dataset that combined clinical data with advanced imaging and genomic analyses. The digital twin model was designed to integrate these diverse data streams, employing AI-driven analysis to support real-time decision-making. The model's ability to simulate patient outcomes and predict disease progression made it a promising tool for improving personalized cancer management and tailoring treatments to individual patient profiles.

Hybrid Extended Reality Data Visualization for Medicine — George Djorgovksi, Ph.D., Santiago Lombeyda, Caltech

The presentation titled "Hybrid Extended Reality Data Visualization for Medicine" by Dr. George Djorgovski and Santiago Lombeyda explored the application of advanced visualization techniques, combining 2D and 3D extended reality (XR), to enhance data interpretation in medical contexts. The presenter, Santiago Lombeyda, demonstrated how visualization tools could help make complex data—such as gene expression and tumor detection—more accessible and interpretable for both researchers and clinicians. By integrating visualization into machine learning (ML) and artificial intelligence (AI) workflows, the presentation showed how these technologies could be leveraged to improve medical outcomes, particularly in cancer detection and treatment. Key examples included interactive tools for gene expression

visualization and 3D genome structure maps that helped unravel genetic information, as well as applications in brain imaging and surgical planning.

The presentation highlighted several visualization projects, including tools for training tumor detection and segmentation, which incorporated AI models to analyze and classify medical data. These tools were built upon collaborations between institutions like Caltech, JPL, and medical research centers, showcasing practical implementations of XR in real-time analysis. Lombeyda's work underscored the value of user-centric design in creating visualization platforms that enabled researchers and clinicians to interact with data more intuitively, ultimately leading to more precise medical interventions and diagnostics. The integration of AI and ML further enhanced the capabilities of these visualization systems, making them powerful tools in the evolving field of medical technology.

Keynote Address: Multimodal and Generative AI for Pathology, Faisal Mahmood, Ph.D., Harvard Medical School

Dr. Fasial Mahood discussed advances in digital pathology and artificial intelligence and potential to build assistive tools for objective diagnosis, prognosis and therapeutic-response and resistance prediction. In his presentation he discussed: (1) Data-efficient methods for weakly-supervised whole slide classification with examples in cancer diagnosis and subtyping), identifying origins for cancers of unknown primary and allograft rejection; (2) Discovering integrative histology-genomic prognostic markers via interpretable multimodal deep learning; (3) Building unimodal and multimodal foundation models for pathology, contrasting with language and genomics; (4) Developing a universal multimodal generative co-pilot and chatbot for pathology; (5) 3D Computational Pathology, and finally (6) Bias and fairness in computational pathology algorithms.

Session 5 - Academic-Industry Partnerships in AI and Bioinformatics — Moderator: Eugene Koay, M.D., MD Anderson

Dr. Eugene Koay introduced the session on "Academic–Industry Partnerships in AI and Bioinformatics." He discussed the dynamics and challenges of collaborations between academia, industry, and government in the fields of artificial intelligence (AI) and bioinformatics. It highlighted the tensions that arose due to differing motivations, timelines, and standards for scientific rigor between academic institutions and industrial partners. For example, while academia often prioritized scientific discovery and long-term research, industry focused on commercialization and shorter timelines. The government played a critical role in these collaborations by setting strategic research directions, establishing policies on data sharing, intellectual property, and regulations that governed commercialization and marketing. The presentation also prompted reflection on key issues such as ensuring adherence to data-sharing policies like those of the National Institutes of Health (NIH), the importance of data standards and interoperability, and best practices for evolving commercial biomarker assays.

The session's agenda featured talks from prominent figures in AI and bioinformatics, including Hoifung Poon from Microsoft Research, Ittai Dayan from Rhino Health, and Amoolya Singh from Grail Bio. These experts addressed critical topics such as leveraging AI for healthcare advancement, federated computing to balance data privacy and sharing, and the use of machine learning (ML) and genomics for early cancer detection. The session underscored the importance of fostering productive academic-industry-government partnerships to drive innovation in bioinformatics and AI while navigating the challenges related to privacy, data sharing, and commercialization. A panel discussion was held at the end where each presenter shared their perspectives and engaged with the audience.

Multimodal Generative AI for Precision Health — Hoifung Poon, Ph.D., Microsoft Research

The presentation "Multimodal Generative AI for Precision Health" by Dr. Hoifung Poon from Microsoft Health Futures highlighted the transformative potential of AI in healthcare, particularly in addressing the inefficiencies of current medical practices. Today, many treatments are imprecise, with high non-responder rates and substantial wasted healthcare spending. For example, while immunotherapy like Keytruda showed promise in cancer treatment, it only worked for a minority of patients. The presentation emphasized the need for complex biomarkers and precision health solutions to better stratify patients and tailor treatments. Poon introduced the concept of a "multimodal patient journey," where various data sources, such as electronic health records (EHRs), real-world data (RWD), and genomics, were integrated to create a comprehensive view of patient health. This multimodal approach enabled the development of more precise models to track disease progression and treatment response.

Poon also discussed the role of real-world evidence (RWE) and advanced AI models like multimodal generative AI in improving patient care and accelerating medical discoveries. The presentation delved into the use of "patient embedding" techniques, which created digital twins of patients by combining multimodal data streams. These AI-driven models could help simulate clinical trials and predict patient outcomes on a population scale. The presentation further highlighted the growing area of frontier models, such as LLaVA-Med and BiomedCLIP, that leveraged vast datasets of biomedical image-text pairs for advanced analysis. Poon showcased how these technologies were applied in complex scenarios like immunotherapy response, digital pathology, and tumor modeling, pointing to the critical role AI would play in advancing precision medicine and improving health outcomes.

Federated Computing: Strike the Balance Between Data Privacy and Data Sharing to Promote Open Science — Ittai Dayan, MD, MPH, Rhino Health

The presentation "Activating the World's Health Data with Federated Computing" by Dr. Ittai Dayan, co-founder of Rhino Health, explored the potential of federated computing to balance data privacy and sharing in healthcare. Federated learning (FL) was introduced as a solution to ease the tension between the need for large-scale data to train AI models and the challenges of data sharing due to privacy concerns. The EXAM consortium's 2020 initiative demonstrated the scalability of FL in hospitals, and since then, it gained significant traction with numerous

collaborations worldwide. FL allowed multiple institutions to collaborate on research without transferring sensitive data outside their firewalls, which was particularly useful in a healthcare environment where privacy was paramount. The Rhino Federated Computing Platform (FCP) expanded on this by enabling seamless distributed execution while maintaining rigorous security, privacy, and compliance standards, addressing the challenges posed by the NIH's Data Management and Sharing Policy.

The presentation highlighted the advantages of federated computing in addressing data centralization issues, such as security risks, high costs, and governance challenges. Rhino FCP supported flexible data-sharing models, including fully federated, hybrid, and centralized approaches, making it adaptable to various healthcare needs. By enabling federated statistics, machine learning operations (MLOps), and trusted research environments, the platform facilitated secure, privacy-preserving collaboration across global networks. Examples of successful implementations included federated datasets and training across institutions like Massachusetts General Hospital, Tel Aviv Sourasky Medical Center, and Assuta Medical Centers. The Rhino platform aimed to accelerate research, enhance data accessibility, and support real-time analysis while eliminating the need for data centralization.

ML and Genomics for Detecting Cancer Early — Amoolya Singh, Ph.D., Grail Bio

The presentation "ML and Genomics for Detecting Cancer Early" by Dr. Amoolya Singh, Chief Scientific Officer at GRAIL, highlighted the critical role of machine learning (ML) and genomics in advancing early cancer detection. Singh began by addressing the global impact of cancer, with 18 million new cases and 9.6 million deaths annually, emphasizing that most cancer deaths occurred due to late-stage diagnosis. The presentation pointed out that approximately 80% of cancer deaths were from cancers without recommended screening protocols. Singh introduced the potential of combining genomics and ML, showcasing how the discovery of circulating tumor DNA (ctDNA) from non-invasive prenatal testing led to the development of early cancer detection techniques by identifying cancer signals in the blood before symptoms appeared.

The presentation further explained how DNA methylation patterns—common across various cancer types—had been identified as particularly sensitive biomarkers for detecting cancer and determining the tissue of origin. Singh also described GRAIL's large-scale clinical programs, including the Circulating Cell-free Genome Atlas (CCGA), aimed at establishing the safety and effectiveness of methylation-based cfDNA tests. The program enrolled over 100,000 participants and continued to generate data that deepened the understanding of cancer biology. Singh concluded by emphasizing the importance of rigorous bioinformatics and ML practices in ensuring reliable cancer classification, outlining a path for future advancements in early cancer detection through the integration of genomics and AI technologies.

Day 3 - Hackathon

Hackathon Overview - Dr. Ashish Mahabal, Ph.D.

Dr. Ashish Mahabal provided an overview of the hackathon. Participants were tasked with analyzing imaging and metadata to develop machine learning (ML) models for breast cancer biomarker analysis. Three main datasets from the H. Lee Moffitt Cancer Center were provided, with a focus on FAIR (Findable, Accessible, Interoperable, Reusable) data practices. The goal was to run ML models, such as convolutional neural networks (CNNs), using tools like Jupyter notebooks and Python libraries for segmentation, detection, and classification.

The hackathon encouraged participants to start with simple models using provided code and then explore more complex tasks like running methods on larger datasets or combining datasets for multi-modal analysis. Techniques such as Bayesian CNNs, GANs, VAEs, and transformers were encouraged. Participants were also expected to follow good practices, including modular coding, visualizations, interpretability, and scalability.

The hackathon aimed to foster brainstorming and ideation, serving as a stepping stone for future research, publications, and the development of advanced models. Each session involved team collaboration, brief reporting, and discussions on future plans for extending the work beyond the hackathon.

Hackathon Preparations and Setup — Ashish Mahabal, Ph.D., Caltech

Dr. Ashish Mahabal described the preparation and setup for the hackathon including setting up a tutorial to ensure participants could access the data and use machine learning libraries. The tutorial provided a step-by-step guide for using Google Colab to run a convolutional neural network (CNN) on breast cancer data from the Early Detection Research Network (EDRN). It began by instructing users on how to download the dataset via Aspera, upload it to Google Drive, and then copy a demonstration Jupyter Notebook into their own drive. The final steps involved running the notebook on a TPU v2 accelerator in Colab, executing the CNN, and analyzing the results. Full instructions were detailed on the EDRN site.

Hackathon LabCAS Training — Heather Kincaid, Jet Propulsion Laboratory

Heather Kincaid provided an overview for accessing and using scientific data through the LabCAS (Laboratory Catalog and Analysis System) platform. LabCAS is part of a larger initiative from the NASA JPL and NCI partnership, aimed at creating a reproducible, data-driven research environment for biomarker studies. The platform is designed to support cancer biomarker discovery by offering secure, organized access to data, integrating analysis tools, and supporting reproducible research.

The training emphasized the LabCAS hierarchical data structure, where data is organized into collections, datasets, and files, allowing researchers to navigate through different layers of scientific data. It supports the sharing, downloading, and processing of data, with functionalities like APIs for automation, Jupyter notebooks, and public libraries. The session also covered the

use of IBM Aspera for efficient data downloads and highlighted FAIR (Findable, Accessible, Interoperable, Reusable) principles, encouraging participants to make the data AI-ready for future research.

The goal of this training was to equip researchers with the tools and knowledge to effectively access, analyze, and share biomarker data. Feedback from participants is crucial to refining the platform, ensuring that LabCAS continues to be a useful resource for cancer biomarker research and advances in AI-driven health studies.

Overview of Hackathon Data Collections — Erin Fowler, Moffitt Cancer Center

Erin Fowler from Moffitt Cancer Center provided an overview of three datasets that were structured for use in the hackathon. These included:

1. **Automated System for Breast Cancer Biomarker Analysis:**
    - 2D mammograms
    - 180 case-control pairs
2. **Automated Quantitative Measures of Breast Density:**
    - 2D mammograms
    - 319 case-control pairs
3. **Moffitt Imaging Biomarker Validation Center:**
    - 348 case-control pairs
    - 2D mammograms
    - Extra: DBT volumetric, C-View

This data collection and analysis was part of the National Cancer Institute's Early Detection Research Network (EDRN) along with several other NIH grants. The team compiled three main collections of mammogram data from 2006 to 2022 for studies including breast density and cancer risk. These mammograms were captured using different technologies, including full-field digital mammography (FFDM) and digital breast tomosynthesis (DBT), along with detailed case-control data. The research aimed to develop automated systems for breast cancer biomarker analysis, measure breast density, and create quantitative metrics to predict cancer risk from imaging data.

The study focused on calibrating mammographic data using odds ratios to assess the relationship between image metrics and cancer risk. The collections include hundreds of case-control pairs of mammograms, supplemented by demographic and pathological data. Advanced imaging processing techniques, such as Fourier and local spatial correlation, were applied to standardize measurements across different mammography systems. Despite challenges related to demographic factors, x-ray detector design, and case-control matching, the research demonstrated important associations between breast density and cancer risk. This work contributed to improved methods for early breast cancer detection and helped identify imaging biomarkers for cancer risk prediction.

Tutorial: CNN Example Notebook — Ashish Mahabal, Ph.D., Caltech

Dr. Ashish Mahabal provided an overview of "Running the Basic CNN Colab", a step-by-step guide for participants, especially beginners, to run a basic Convolutional Neural Network (CNN) during the Cancer Biomarkers AI Hackathon. The tutorial included downloading a small dataset (AQMBDsmall.tar.gz), using Google Colab, and running the model with a TPU V2 runtime. Participants followed specific instructions to mount Google Drive, change file paths, and run cells in the notebook. The goal was to familiarize attendees with CNN workflows, though no complex optimizations were applied in this basic run.

While the initial model performance was intentionally simple, the purpose of this tutorial was to ensure that participants understood the process of handling data, setting up environments, and running a CNN. The tutorial encouraged further experimentation and model improvements, setting the stage for more advanced tasks and brainstorming during the hackathon. This introduction to CNNs was seen as a foundation for deeper exploration and learning throughout the event.

Charge to the Hackathon Teams — Ashish Mahabal, Ph.D., Caltech

Dr. Ashish Mahabal provided direction to each hackathon group to pick a joint problem and to ideate about how the supplied datasets could be used to further the application of AI to cancer biomarker research science challenges. Three hackathon groups were identified as follows:

1. **Breast Segmentation:**
    - Alekhya Vittalam (University of California, Los Angeles), Jennifer Beane-Ebel (Boston University), Heather Kincaid (JPL) , Ghulam Rasool (Moffitt Cancer Center), Giresh Yemparala (FHCRC), Yuesong Wu (Michigan State University)
2. **Breast Cancer Stage Classification:**
    - Stephen Park (UCLA), Luoting (Lottie) Zhuang (UCLA), Irene C (UCLA), Alexander Chowdhury (DFCI), Sahil Nalawade (DFCI), Soujanya Samineni (DFCI), Zhiwei L (DFCI)
3. **Mammo-CLIP:**
    - Radka Stoyanova (University of Miami), Adrian Breto (University of Miami), Wenxin Zhang (UC Berkeley), Peter Yu (UCLA), David Wong (UCLA), Dan Crichton (NASA/JPL), Steve Skates (MGH), Shirley Li (Abbott), Ben Jacob (RCSI University – Ireland)

Hackathon Final Reports — Each Hackathon Group

Hackathon groups met and then provided a report, as follows:

The "*Breast Segmentation Group*" hackathon project aims to develop an AI model to segment dense breast regions in mammograms for early cancer detection. The team worked with Moffitt's Hologic Dimensions 3D Case-Control Mammography Study, which included 348 case-control pairs of mammographic images and clinical data. They proposed fine-tuning the Segment Anything Model 2 (SAM2) by annotating a small set of images, having a radiologist verify these annotations, and using the verified data to improve the model.

The approach emphasized collaboration between the team and radiologists to ensure accurate segmentation. Resources required included cloud-based GPU access for processing and storage for both pre- and post-processed data. The hackathon served as a platform for brainstorming and testing new solutions for breast cancer detection, with the ultimate goal of improving early diagnosis through AI-driven segmentation models.

The "*Breast Cancer Stage Classification*" hackathon project focuses on building a model to classify breast cancer into five stages (0 to IV) based on key features such as tumor size, lymph node location, and extracted radiomic features. The project involved using segmentation and pre-trained models to identify tumor size, followed by feature selection through Grid Search Cross-Validation (CV). For classification, Random Forest (RF) or Support Vector Machine (SVM) models were employed to predict the stage of cancer based on the selected features.

The project emphasized the importance of accurate feature extraction and selection to improve the performance of the stage classifier. By focusing on tumor size and lymph node data, the model aimed to provide a reliable system for classifying the progression of breast cancer, helping clinicians in early and precise diagnosis. This hackathon project combined medical imaging techniques and machine learning models to tackle one of the most critical aspects of breast cancer management—accurate staging.

The "*Mammo-CLIP*" hackathon project focuses on evaluating the performance of MammoCLIP, a Vision Language Model (VLM), in classifying cancer and normal cases using the Moffitt dataset. The team formatted the Moffitt data for fine-tuning MammoCLIP and trained it on 80% of the dataset, with evaluation on the remaining 20%. This approach aimed to explore the zero-shot capabilities of the model and its effectiveness in medical imaging classification tasks.

The team also explored the model's ability to generalize across different data sources, assessing how well it performed when trained on breast cancer imaging data. Resources required included access to cloud-based GPU clusters for training the large models and computing power to handle the extensive dataset. The hackathon project highlighted the potential of VLMs like MammoCLIP in medical imaging, showing that these models could contribute to more accurate classifications in cancer detection.

Final Remarks — Dan Crichton, Jet Propulsion Laboratory

Dan Crichton concluded the hackathon and the overall workshop by thanking the participants for their engagement and contributions. He expressed gratitude to the National Cancer Institute (NCI) for their continued support of initiatives like the Early Detection Research Network (EDRN) and the integration of AI in cancer biomarker research. Crichton emphasized the importance of continued collaboration between institutions and the need to explore innovative applications of AI in cancer research.

He encouraged participants to continue developing their hackathon projects beyond the event and to consider future opportunities for partnership and funding. Crichton reaffirmed the value of AI in advancing the field of cancer biomarkers and highlighted the role of platforms like LabCAS in enabling reproducible, data-driven science. The event served as a foundation for future efforts aimed at enhancing early cancer detection and leveraging AI to improve patient outcomes. He closed by urging attendees to stay connected and continue working together on AI-driven cancer research.

Summary

The Cancer Biomarkers AI and Bioinformatics Workshop brought together leading experts from academia, industry, and government to explore the transformative potential of artificial intelligence and cancer bioinformatics in cancer biomarker discovery and validation. It provided a multidisciplinary platform to address the pressing challenges and opportunities in the field.

Key outcomes of the workshop include the following:

- **Advancements in AI Applications**: Participants highlighted how AI and ML techniques are pioneering cancer biomarker discovery, particularly through deep learning models that can analyze complex, high-dimensional data such as genomics and imaging.
- **Data Preparation and Sharing**: The importance of high-quality, curated datasets was emphasized. Discussions revolved around data harmonization, standardization, and the creation of AI-ready datasets to enhance reproducibility and generalizability of AI models.
- **Emerging Technologies**: The workshop showcased cutting-edge technologies like federated learning, digital twins, and multimodal generative AI. These tools offer new ways to handle data privacy concerns, personalize patient care, and integrate diverse data types for more robust predictive modeling.
- **Methodological Considerations**: Robustness, interpretability, and fairness of AI models is as necessary as the FAIR principles for raw data access. The need for explainable AI

and proper uncertainty quantification was underscored to build trust in AI-driven healthcare solutions.
- **Academic-Industry Partnerships**: The event fostered dialogue on the dynamics of collaborations between academic institutions and industry partners—but also highlighted data sharing policies, intellectual property, and regulatory compliance. There are important opportunities for continued partnerships and leveraging of AI capabilities from industry in areas such as foundation models, federated learning, and analytical tools.
- **Hackathon Insights**: The half-day hackathon allowed participants to apply AI tools to real-world data from the LabCAS cancer biomarker data commons, promoting hands-on experience in tackling practical challenges in data analysis and model development. It is recommended that future hackathons be organized and the teams continue to work on the ideas, which should be presented at future meetings.

To build upon the insights and progress from the workshop, the following recommendations are proposed as next steps:

- **Continue establishing a community of practice in EDRN and with the cancer biomarker community in AI:** EDRN should continue building an AI and data community, connecting leaders from across institutions, and developing plans to share expertise, capabilities, tools, and best practices to support cancer biomarker research.
- **Provide additional workshops and hackathons:** EDRN's experience, along with others within NCI, have recognized the importance of bringing the community together to support use of data and tools. EDRN should continue to hold workshops and support hackathons going forward.
- **Pursue a special journal issue on cancer biomarkers and AI:** Given the rising importance of AI and biomarkers, it is recommended that a special issue be established on cancer biomarkers and AI.
- **Drive forward one or more projects as an EDRN use case from the hackathon:** Several projects and teams were proposed. These also provide a template for working AI projects across EDRN. The hackathon teams should continue to work and consider presenting at a future EDRN meeting.
- **Work with with the Informatics and Data Sharing Subcommittee within the EDRN to enhance data usability and reuse:** through further use of LabCAS, there is a landscape that includes the creation of AI-ready datasets with comprehensive metadata and documentation; this will enable rigorous data harmonization and preprocessing techniques to ensure consistency and reliability of inputs for AI models
- **Increase access to shared computation for joint AI projects:** providing tools and environments which support the application of AI and the use of common tools and data along with scalable computation.

Appendix A - Program Committee

Jennifer Ellen Beane-Ebel (Boston University)
Paul Boutros (UCLA)
Dan Crichton (NASA JPL)
Ziding Feng (Fred Hutch Cancer Center)
Chad He (Fred Hutch Cancer Center)
Heather Kincaid (NASA JPL)
Eugene Koay (MD Anderson Cancer Center)
Ashish Mahabal (Caltech)
Anirban Maitra (MD Anderson Cancer Center)
Christos Patriotis (NIH/NCI)
Matthew B. Schabath (Moffitt Cancer Center)
Amanda Skarlupka (NIH/NCI)
Steven Skates (MGH)
Sudhir Srivastava (NIH/NCI)
Zhen Zhang (JHMI)

Acknowledgements

The authors would like to acknowledge the National Cancer Institute for their support of the workshop. A portion of this report was generated as follows: OpenAI. (2024). *ChatGPT (October 2024 version)* [Large language model]. https://www.openai.com.