RESEARCH BRIEF

Federated Deep Learning Enables Cancer Subtyping by Proteomics

Zhaoxiang Cai[1], Emma L. Boys[1], Zainab Noor[1], Adel T. Aref[1], Dylan Xavier[1], Natasha Lucas[1], Steven G. Williams[1], Jennifer M.S. Koh[1], Rebecca C. Poulos[1], Yangxiu Wu[1], Michael Dausmann[1], Karen L. MacKenzie[1], Adriana Aguilar-Mahecha[2], Carolina Armengol[3,4], Maria M. Barranco[3,4], Mark Basik[2], Elise D. Bowman[5], Roderick Clifton-Bligh[6,7,8], Elizabeth A. Connolly[1], Wendy A. Cooper[6,9,10], Bhavik Dalal[5], Anna DeFazio[11,12,13], Martin Filipits[14,15,16], Peter J. Flynn[17], J. Dinny Graham[11,18], Jacob George[6,19], Anthony J. Gill[6,7,20], Michael Gnant[15,16], Rosemary Habib[11,21,22], Curtis C. Harris[5], Kate Harvey[23], Lisa G. Horvath[6,23,24], Christopher Jackson[25], Maija R.J. Kohonen-Corish[23,26,27,28], Elgene Lim[23,29], Jia (Jenny) Liu[1,23,29], Georgina V. Long[6,30,31,32], Reginald V. Lord[33], Graham J. Mann[11,30,34], Geoffrey W. McCaughan[6,35,36], Lucy Morgan[6,17], Leigh Murphy[37], Sumanth Nagabushan[38], Adnan Nagrial[6,21,22], Jordi Navinés[39], Benedict J. Panizza[25], Jaswinder S. Samra[6,40], Richard A. Scolyer[6,9,30,31], John Souglakos[41], Alexander Swarbrick[23], David Thomas[42], Rosemary L. Balleine[1,11], Peter G. Hains[1], Phillip J. Robinson[1], Qing Zhong[1], and Roger R. Reddel[1]

ABSTRACT Artificial intelligence applications in biomedicine face major challenges from data privacy requirements. To address this issue for clinically annotated tissue proteomic data, we developed a federated deep learning approach (ProCanFDL), training local models on simulated sites containing data from a pan-cancer cohort (n = 1,260) and 29 cohorts held behind private firewalls (n = 6,265), representing 19,930 replicate data-independent acquisition mass spectrometry runs. Local parameter updates were aggregated to build the global model, achieving a 43% performance gain on the hold-out test set (n = 625) in 14 cancer subtyping tasks compared with local models and matching centralized model performance. The approach's generalizability was demonstrated by retraining the global model with data from two external, data-independent acquisition mass spectrometry cohorts (n = 55) and eight acquired by tandem mass tag proteomics (n = 832). ProCanFDL presents a solution for internationally collaborative machine learning initiatives using proteomic data, for example, for discovering predictive biomarkers or treatment targets while maintaining data privacy.

SIGNIFICANCE: A federated deep learning approach applied to human proteomic data, acquired using two distinct proteomic technologies from 40 tumor cohorts across eight countries, enabled accurate cancer histopathologic subtyping while preserving data privacy. This approach will enable the privacy-compliant development of large-scale proteomic artificial intelligence models, including foundation models, across institutions globally.

[1]ProCan, Children's Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, Westmead, Australia. [2]Lady Davis Institute at the Jewish General Hospital, McGill University, Montreal, Canada. [3]Childhood Liver Oncology Group, Germans Trias i Pujol Research Institute, Badalona, Spain. [4]Networking Biomedical Research Centre (CIBER) in Hepatic and Digestive Diseases, Barcelona, Spain. [5]Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland. [6]Faculty of Medicine and Health, The University of Sydney, Sydney, Australia. [7]Kolling Institute of Medical Research, Royal North Shore Hospital, St Leonards, Australia. [8]Department of Endocrinology, Royal North Shore Hospital, St Leonards, Australia. [9]Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital and NSW Health Pathology, Sydney, Australia. [10]School of Medicine, Western Sydney University, Sydney, Australia. [11]Centre for Cancer Research, The Westmead Institute for Medical Research, Faculty of Medicine and Health, The University of Sydney, Westmead, Australia. [12]The Daffodil Centre, The University of Sydney, a joint venture with Cancer Council NSW, Westmead, Australia. [13]Department of Gynaecological Oncology, Westmead Hospital, Westmead, Australia. [14]Center for Cancer Research, Medical University of Vienna, Vienna, Austria. [15]Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria. [16]Austrian Breast and Colorectal Cancer Study Group, Vienna, Austria. [17]Nepean Hospital, Kingswood, Australia. [18]Westmead Breast Cancer Institute, Westmead Hospital, Westmead, Australia. [19]Storr Liver Centre, The Westmead Institute for Medical Research, Westmead Hospital, Westmead, Australia. [20]NSW Health Pathology, Department of Anatomical Pathology, Royal North Shore Hospital, St Leonards, Australia.

INTRODUCTION

Artificial intelligence (AI) applications, driven by a wealth of online data, have gained traction as tools to enhance efficiency, convenience, and innovation across multiple sectors. The use of these applications in commercial products, ranging from personalized recommendations in streaming services to generative AI tools such as ChatGPT, has led to the widespread uptake of these technologies and ongoing discussions about their appropriate use and regulation (1). Within the biomedical domain, numerous applications of AI are undergoing rapid development, including diagnostic prediction tools, interpretation of radiological and histopathologic images, and methods of drug discovery (2). Although such tools offer promise, with efficiency gains and the potential for novel insights beyond those that can be achieved by traditional research studies, to date, large-scale AI modeling in most biomedical fields continues to be hindered by several substantive challenges (3).

These challenges include the privacy of data, especially personal clinical records, data ownership and governance, human research ethics, and intellectual property concerns (4). In transnational studies, compliance with laws and regulations that impose stringent standards on the collection, storage, sharing, and use of biomedical data may be complicated by differing requirements in the relevant jurisdictions (5, 6). Consequently, data sharing among collaborators within international consortia can be infeasible, hindering the assembly of heterogeneous and globally representative large-scale datasets and presenting a significant barrier to the development of practical and relevant AI tools in the biomedical field. This contrasts sharply with commercial AI products based on ubiquitous information regarded as nonsensitive. However, for biomedical applications such as cancer research, there is an urgent need to utilize all sources of good-quality data for purposes that include expediting the discovery of drug targets and predictive biomarkers and avoiding duplication of resources and therefore wastage due to data siloing and inaccessibility. To address these challenges, innovative solutions that balance the need for data accessibility and protection of clinical data are crucial.

Federated learning (FL) offers a promising solution to some of these challenges (7). As a distributed learning framework, FL permits the local training of sensitive data at participating sites, with only the local model updates being shared with a central server to create a global model. This approach ensures that local data remain protected and securely stored behind firewalls. The ability to protect confidential data and combine diverse and geographically distinct datasets has outstanding potential for the development of large-scale generative AI tools with utility in both biomedical research and healthcare settings (8–10).

Although genomic and transcriptomic studies have greatly advanced our understanding of cancer, proteomic data will play a crucial role in answering many unresolved questions about the molecular mechanisms of cancer (11–13) and in identifying predictive markers (13). However, as the scale of human proteomics research increases, so do the challenges related to data privacy. FL provides a promising solution to address this but so far has been applied only to nonhuman proteomic data (arXiv 2407.15220). A potential high-impact application of FL in proteomics would be to develop a federated global model for international proteomic consortia, as will be attempted by π-Hub (the proteomic navigator of the human body; ref. 14).

This study addresses these gaps by developing a federated deep learning (FDL)–based framework, ProCanFDL, for the analysis of proteomic data. The dataset, referred to here for brevity (and to distinguish it from external datasets) as the ProCan Compendium, includes 7,525 human biospecimens from 30 cohorts that were preserved and stored either by freezing or formalin fixation and paraffin embedding in pathology laboratories in multiple countries. There were sufficient samples to train the FDL proteomic model to recognize 14 cancer histopathologic subtypes; its accuracy, tested on a hold-out test set, consistently outperformed individual local models and was on par with the centralized model. The robustness of ProCanFDL was further validated using 10 external proteomic datasets, eight of which were generated by a different mass spectrometry (MS) technology, covering two additional cancer subtypes, to train a global model that can accurately recognize 16 histopathologic subtypes. These findings highlight the potential of FDL to advance global clinical proteomics research by enabling secure, integrative data analysis across institutions and jurisdictions.

[21]Crown Princess Mary Cancer Centre, Westmead Hospital, Westmead, Australia. [22]Blacktown Cancer and Haematology Centre, Blacktown Hospital, Blacktown, Australia. [23]Garvan Institute of Medical Research, Darlinghurst, Australia. [24]Chris O'Brien Lifehouse, Camperdown, Australia. [25]Department of Otolaryngology, Head and Neck Surgery, Princess Alexandra Hospital, Brisbane, Australia. [26]Woolcock Institute of Medical Research, Macquarie University, Sydney, Australia. [27]Macquarie Medical School, Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, Australia. [28]Sydney Local Health District, Sydney, Australia. [29]The Kinghorn Cancer Centre, St Vincent's Hospital, Sydney, Australia. [30]Melanoma Institute Australia, The University of Sydney, Sydney, Australia. [31]Charles Perkins Centre, The University of Sydney, Sydney, Australia. [32]Royal North Shore and Mater Hospitals, Sydney, Australia. [33]St.Vincent's Centre for Applied Medical Research, Darlinghurst, Australia. [34]John Curtin School of Medical Research, Australian National University, Canberra, Australia. [35]AW Morrow Gastroenterology and Liver Centre, Royal Prince Alfred Hospital, Camperdown, Australia. [36]Centenary Institute, Camperdown, Australia. [37]Manitoba Tumour Bank, University of Manitoba and Cancer Care Manitoba Research Institute, Winnipeg, Canada. [38]The University of Sydney, Sydney, Australia. [39]General Surgery Department, Hospital Germans Trias i Pujol, Barcelona, Spain. [40]Department of Upper GIT Surgery, Royal North Shore Hospital, St Leonards, Australia. [41]Department of Medical Oncology, Faculty of Medicine, University Hospital of Heraklion, University of Crete, Crete, Greece. [42]Centre for Molecular Oncology, University of New South Wales, Sydney, Australia.

Z. Cai, E.L. Boys, Z. Noor, and A.T. Aref contributed equally to this work.

Corresponding Authors: Peter G. Hains, ProCan, Children's Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, 214 Hawkesbury Road, Westmead 2145, Australia. E-mail: phains@cmri.org.au; Phillip J. Robinson, probinson@cmri.org.au; Qing Zhong, qzhong@cmri.org.au; and Roger R. Reddel, rreddel@cmri.org.au

Cancer Discov 2025;XX:1–16

doi: 10.1158/2159-8290.CD-24-1488

RESULTS

ProCan Compendium and Landscape Analysis

We first compiled the ProCan Compendium, quantifying proteomes from 7,525 tissue samples, including 5,982 tumors, 1,512 tumor-adjacent normal samples, and 30 benign samples from 4,954 individual patients (Supplementary Table S1). The data were generated in collaborative research projects involving 20 research groups across seven countries (Fig. 1A) who provided biospecimens, stored either fresh frozen (FF) or formalin-fixed and paraffin-embedded (FFPE), and the associated clinical data. Utilizing a high-throughput workflow with seven mass spectrometers, 19,930 data-independent acquisition mass spectrometry (DIA-MS) runs were used to obtain replicate proteomic data from the 7,525 samples (11, 15–17). Raw DIA-MS data were processed and normalized using DIA-NN with a DIA-NN–generated spectral library, quantifying a total of 9,102 proteins. The number of proteins quantified per sample, grouped by tissue of origin, cancer type, and cancer subtype, is presented in Supplementary Fig. S1A–S1C. These samples encompassed 31 tissues of origin, 29 cancer histopathology types, and more than 65 cancer subtypes, distributed across 30 cohorts (Fig. 1B and C; "Methods"). High correlations between replicates of individual samples were observed, with a sample-wise median Pearson's correlation coefficient (Pearson's r) of 0.96 and moderate correlations between samples of the same cancer and tissue of origin (0.84 and 0.81, respectively). Correlations between unmatched samples from the same instrument were equivalent to those of random sample pairings (median Pearson's $r = 0.75$), indicating that there were no instrument-specific batch effects (Fig. 1D).

In the ProCan Compendium, cohort 1 serves as the baseline pan-cancer cohort and its raw data, and the corresponding spectral library are made publicly available alongside this study as a resource for researchers in the field of cancer proteomics ("Data Availability"). This cohort was acquired from the Victorian Cancer Biobank, the Gynaecological Oncology Biobank (GynBiobank) at Westmead Hospital, and the Children's Medical Research Institute Legacy sample set and consists of 766 tumor samples from 638 patients. Similarly to the ProCan Compendium overall, high correlations were observed between replicates of individual samples across all cancer types in cohort 1 (Supplementary Fig. S2A). Protein intensities were visualized in tumor samples using Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP), revealing distinct clusters for several cancer types, including lymphoma and melanoma, with biologically related cancer types, such as neuroblastoma and ganglioneuroblastoma, forming neighboring clusters, indicating the robustness of this pan-cancer dataset (Fig. 1E). For broad cancer types, such as adenocarcinoma and carcinoma, the UMAP visualization of their subtypes provides a detailed representation of the subtypes and clearer insight into the diversity within the cancer types (Fig. 1F). Analysis of cancer subtype–enriched proteins showed that cohort 1 exhibited a pattern consistent with our previous study in cancer cell lines (13), with neuroblastoma showing the highest number of enriched proteins and quantification rate and lymphoma showing the second highest (Supplementary Fig. S2B and S2C). Cohorts 2 to 30 comprise 29 single-cancer cohorts, with a total of 5,217 tumor samples from 4,316 patients encompassing 42 cancer subtypes, which will be included in separate publications.

ProCanFDL Overview

The traditional method of machine learning is based on local learning (Fig. 2A), in which individual research groups independently train models on the data available to them. This approach preserves jurisdictional data control but limits the ability to generalize findings across diverse datasets. Centralized learning (Fig. 2B) improves predictive performance by aggregating data from multiple sites into a centralized model; however, it necessitates sharing sensitive data, which raises subsequent privacy concerns. FL (Fig. 2C) represents an evolution of these methodologies by enabling the training of a global model across decentralized data sources, updating both local and global model weights without the need to transfer raw data, thus preserving data privacy. FDL specifically refers to the implementation of deep learning techniques within this distributed setup.

The ProCanFDL framework employs a four-step FDL approach while maintaining data privacy, enabling collaborative research in an international consortium (Fig. 2C; "Methods"). In step 1 (initialization and local training), a global model is initialized with random weights and distributed to all participating local sites. Then, a local instance of a deep learning model is trained on its private proteomic data at each participating site. These models are trained independently, without sharing raw data across sites. In step 2 (global model aggregation), the trained model parameters are securely transferred to a central server, which aggregates these updates using a federated averaging algorithm. This process creates a global model reflecting the pooled knowledge from all local datasets, without the need for the server to access raw data. In step 3 (global model update), the newly aggregated global model is distributed back to all participating sites, where it serves as the starting point for the next round of local training. Finally, in step 4 (iteration and convergence), this process (steps 1–3) is repeated iteratively, with each cycle refining the global model until it converges. The resulting model becomes increasingly accurate and representative of the combined datasets, encapsulating the collective knowledge.

ProCanFDL on ProCan Compendium

To evaluate and benchmark ProCanFDL, we used proteomic data from the ProCan Compendium as input, training local, centralized, and ProCanFDL global models to assign each sample to its correct cancer subtype. As a proof of concept, we focused on 14 cancer subtypes, for each of which at least five samples were available in cohort 1 and at least 20 samples across cohorts 2 to 30 (Supplementary Fig. S3A and S3B). Additionally, only samples with replicate correlations greater than 0.9 were included in the analysis to ensure data quality and consistency. This filtering step resulted in a final subset of 4,558 samples, which was used in the subsequent analyses. No cohort-level normalization

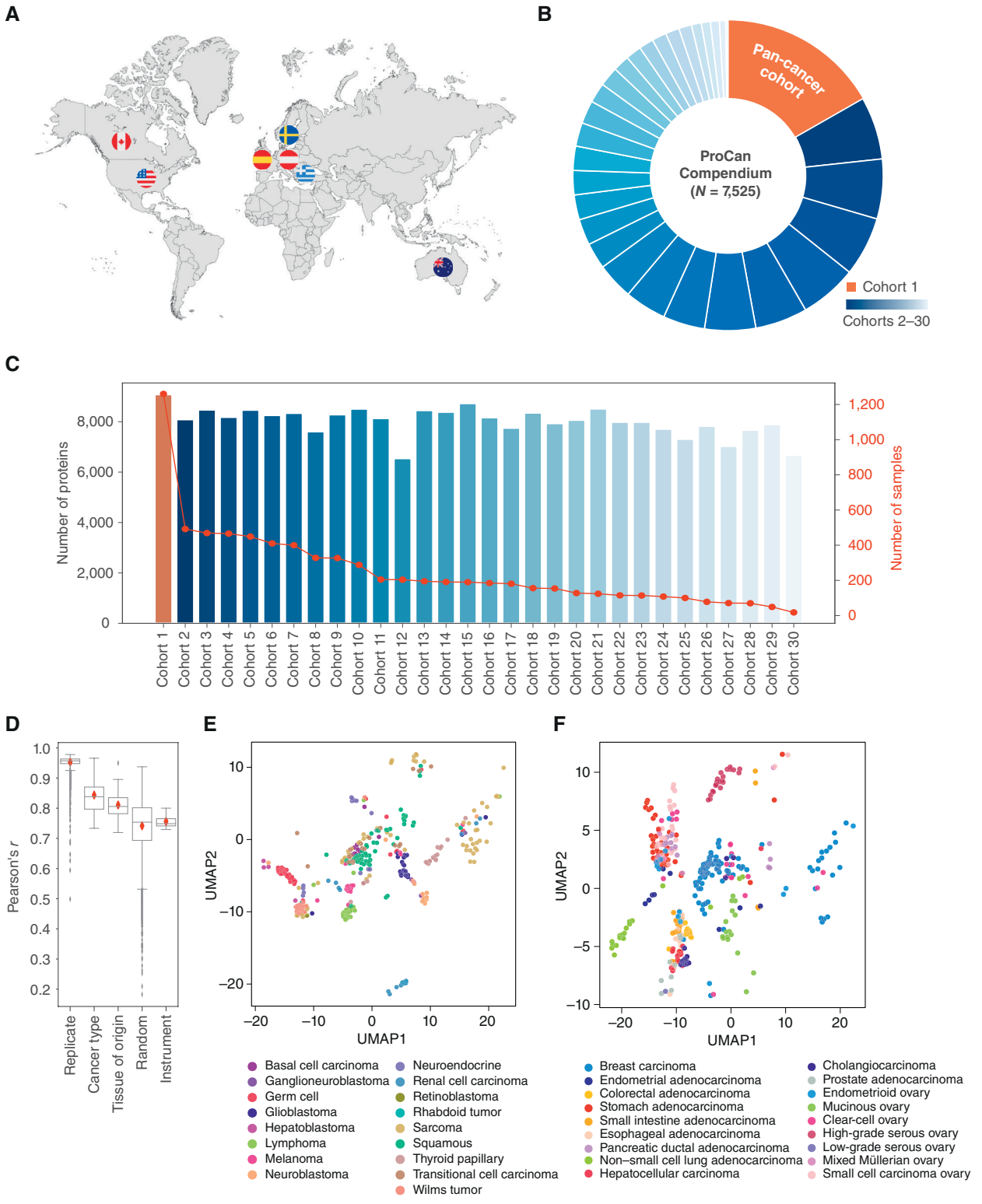

Figure 1. Overview of ProCan Compendium. **A,** The data were assembled from studies involving 20 collaborating cancer research groups across seven countries. **B,** The circular bar plot shows the sample sizes of the 30 cohorts, with the largest being the pan-cancer cohort (cohort 1). **C,** Bar plot showing the number of quantified proteins; the red dot indicates the number of samples. **D,** Box plot of mean Pearson's *r* for replicates, cancer types, tissue of origin, random sets of 10,000 samples, and different MS instruments. Box-and-whisker plots display 1.5× interquartile ranges, with centers indicating medians and red diamonds representing mean values. **E,** UMAP with samples colored by cancer types. Clusters of cancer subtypes as per histologic classification are evident. **F,** UMAP of carcinoma subtypes. Clusters of different subtypes as per primary tissue of origin are evident. (**A,** Created with BioRender.com.)

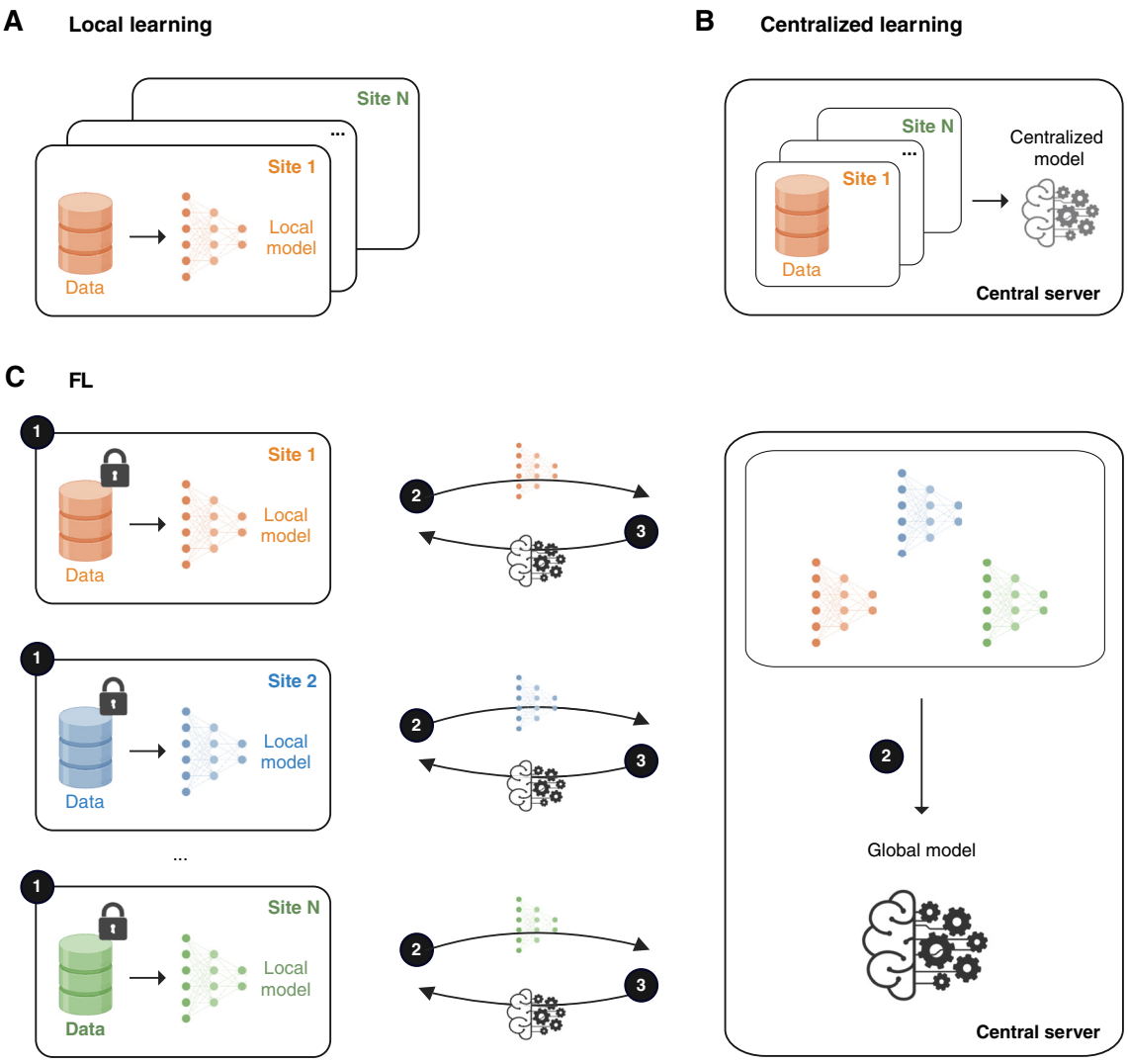

Figure 2. Local learning, centralized learning, and FL. **A,** Local learning refers to training machine learning models on local sites without sharing data with other sites or a central server. **B,** Centralized learning involves collecting data from multiple local sites and aggregating data on a central server where a centralized machine learning model is trained. **C,** In FL, local models are trained on decentralized sites, each residing behind their respective firewalls. Only the model parameters, and not the raw data, are shared with a central server. These model parameters are then aggregated to form a global model. The global model is sent back to all local sites for the next round of training. Numbers 1–3 correspond to the first three algorithmic steps in ProCanFDL. (Created with BioRender.com.)

was performed. The input data for all machine learning models were organized in a data matrix format, in which rows represent samples and columns correspond to protein abundances. We designated 10% of the patients from each of cohorts 2 to 30 as the fixed hold-out test set *T*. The training set consisted of the remaining 90% of data from cohorts 2 to 30 and all of the data from cohort 1 (Fig. 3A; "Methods"). To set up a curated baseline with cohort 1, we applied further quality control filtering steps, including histopathologic validation ("Methods"). The performance of the models for local, centralized, and ProCanFDL was evaluated using the same test set, *T*. Local, centralized, and ProCanFDL models employed identical deep learning architectures and hyperparameters, optimized through the cross-validation process using cohort 1 (Fig. 3B; "Methods").

Site Simulation

We created four local sites to simulate a FL scenario in which data from different institutions cannot be centrally combined due to privacy regulations. To obtain statistically robust and generalizable results, we ran the simulation 10 times. In each iteration, site 1 always included only data from cohort 1, whereas the other sites received a randomly selected, nonrepeating subset of cohorts from the remaining 29. This approach generated 10 unbiased cohort distributions across sites 2 to 4, ensuring that each site contained approximately 10 cohorts, representing a meaningful fraction of the 14 cancer subtypes (Fig. 3C). In addition, this setup demonstrates how different cancer subtypes are distributed across the sites in each experiment, highlighting the heterogeneous nature of data distribution in the FL setup (Supplementary Figs. S4 and S5).

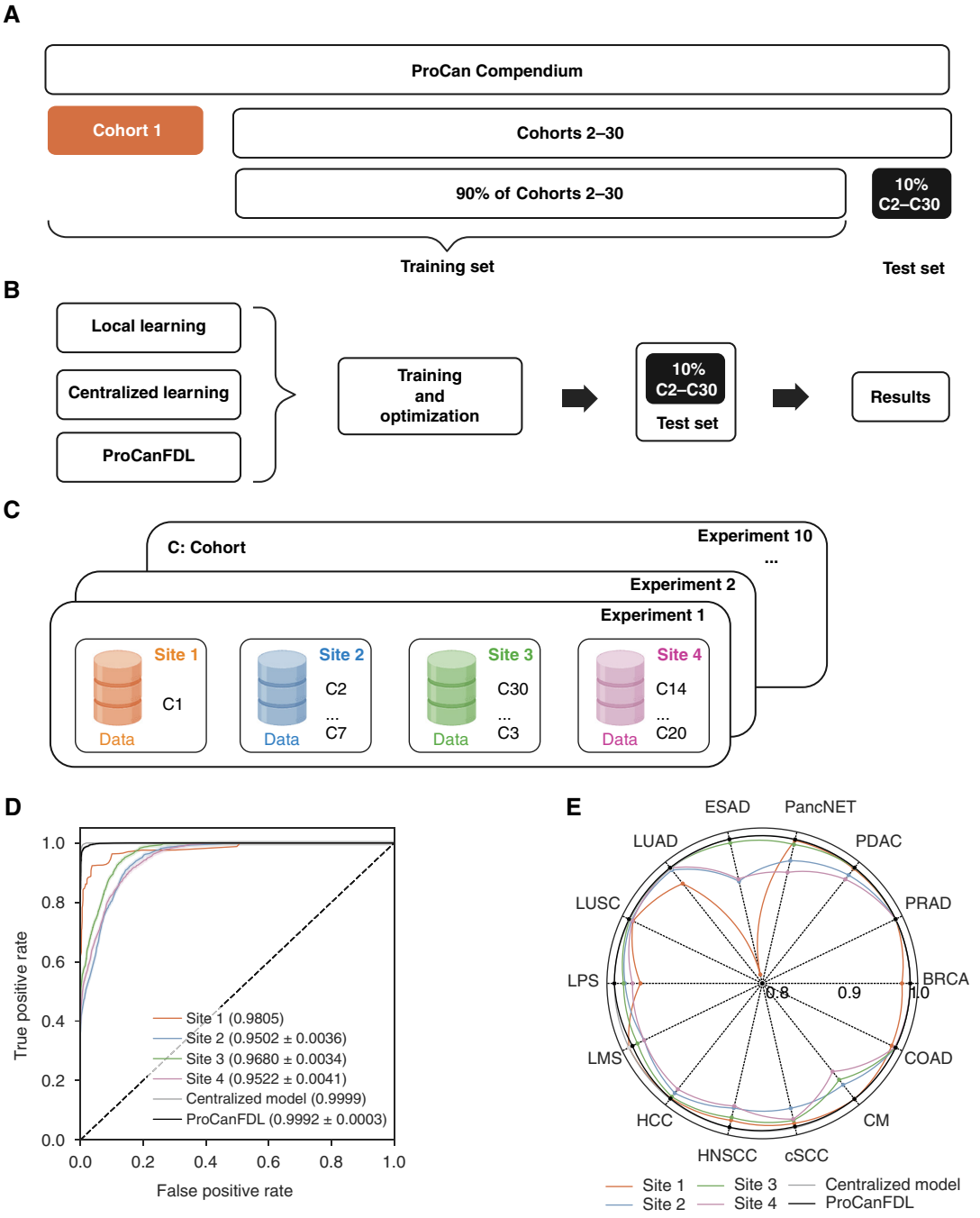

Figure 3. Experiment setup and model performance. **A,** The ProCan Compendium consists of cohorts 1–30. Cohort 1, along with 90% of cohorts 2–30, forms the training set, whereas the remaining 10% of cohorts 2–30 constitutes the hold-out test set. **B,** The final local, centralized, and ProCanFDL models were evaluated on the same test set. **C,** Ten experiments for evaluating ProCanFDL. Site 1 always contains cohort 1 (C1), whereas each site from sites 2–4 contains a random subset of cohorts. **D,** The performance of each model was benchmarked by macro-averaged AUROC for sites 2–4 and ProCanFDL. The AUROCs were annotated as the mean values ± the half-width of the 95% confidence intervals estimated from 10 experiments. **E,** AUROC of models across 14 cancer subtypes. The full names of the cancer subtypes can be found in "Methods." (**C,** Created with BioRender.com.)

Local Learning

We trained a local model for cohort 1 data at site 1. The performance of the final model achieved a macro-averaged area under the receiver operating characteristic curve (AUROC) of 0.9805 and an accuracy of 0.847 for classifying the 14 cancer subtypes (Fig. 3D; Supplementary Fig. S6A–S6C; "Methods"). We then evaluated the performance of the local models for sites 2 to 4 by averaging the results from 10 experiments for each site. Site 2 achieved a mean macro-averaged AUROC of 0.9502, and sites 3 and 4 recorded values of 0.9680 and

0.9522, respectively (Fig. 3D). These scores were lower than the mean macro-averaged AUROC of site 1, primarily due to the absence of data encompassing all 14 cancer subtypes, which was only available at site 1.

Centralized Learning

The centralized learning approach combines all data from the training set to train a single centralized model. The centralized model achieved a macro-averaged AUROC of 0.9999 (Fig. 3D) and an accuracy of 0.990 (Supplementary Fig. S6A–S6C), significantly outperforming the local models. These results demonstrate the expected benefits of data aggregation in improving predictive performance, but this approach requires the sharing of data between sites, which may be prevented by local laws and regulations.

ProCanFDL

Using the four-step algorithmic procedure of ProCanFDL, we developed a global model for each experiment ("Methods"). The performances of the global models were averaged across the 10 experiments. The ProCanFDL global model achieved a mean macro-averaged AUROC of 0.9992 (Fig. 3D) and a mean accuracy of 0.965 (Supplementary Fig. S6A–S6C). This represents a substantial improvement over the local models for sites 1 to 4 and approximates the performance of the centralized model. The AUROC for each cancer subtype is detailed in Fig. 3E and Supplementary Table S2. We next evaluated the number of true and false predictions across each class generated by the ProCanFDL global model with the highest macro-averaged AUROC from the 10 experiments. The global model achieved 100% sensitivity (true positive rate) in classifying 10 out of 14 cancer subtypes, plus sensitivity exceeding 90% for lung squamous carcinoma (Supplementary Fig. S6D), confirming the predictive power of the model.

Overall, these findings highlight the effectiveness of ProCanFDL in improving cancer subtyping performance. Notably, the federated approach not only surpasses the performance of local models but also delivers results comparable with the centralized model while providing the critical advantage of preserving data privacy, potentially enabling large-scale machine learning across sites worldwide.

Generalization and Integration

To rigorously assess the ProCanFDL model's performance on unseen data and provide a clear indication of its generalizability beyond the ProCan Compendium dataset and the DIA-MS method, we included datasets generated in other laboratories. We selected datasets encompassing subsets of 65 cancer subtypes, generated using two distinct MS technologies [DIA-MS and tandem mass tagging (TMT)]. For DIA-MS, we retrieved two cohorts from the Proteomics Identification database, one consisting of 40 colorectal adenocarcinoma samples from Spain and another with 15 pancreatic ductal adenocarcinoma samples from South Africa (18, 19). For TMT, we accessed data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) in the USA, which included 832 tumor samples across eight cohorts (Supplementary Table S3; ref. 20).

To ensure consistency across the different datasets, protein-wise *z*-score normalization was applied to both the DIA-MS proteomic data (the ProCan Compendium and two external DIA-MS cohorts) and the eight TMT cohorts separately. The two DIA-MS datasets were concatenated sample-wise into a single matrix before normalization, as were the eight TMT cohorts. The two normalized matrices were then concatenated sample-wise again to produce a final input matrix containing a set of 3,837 proteins that were quantified in common in these cohorts.

We applied the train–test split across the external data (21). In this method, 90% of the external data, comprising the 10 cohorts, was used to simulate additional local sites 5 and 6, whereas the remaining 10% was held out and combined with the existing hold-out test set T to form a new, unbiased evaluation set T' (Supplementary Fig. S7A). This test set T' was then used to evaluate and compare the generalizable performance of local, centralized, and ProCanFDL global models. The two external DIA-MS training cohorts were grouped as site 5, and the eight TMT training cohorts from CPTAC were grouped as site 6 (Supplementary Fig. S7B). The inclusion of these external datasets allowed us to extend the analysis to two additional cancer subtypes, high-grade serous ovarian carcinoma and clear-cell renal cell carcinoma, for which insufficient samples were available to meet the training cohort selection criterion of a minimum of 20 samples per cancer subtype in cohorts 2 to 30 of the ProCan Compendium. Thus, in the external validation analysis, the total number of cancer subtypes analyzed increased from 14 to 16, the total number of samples meeting the criteria for analysis in sites 1 to 4 increased by 195, and the number of samples from sites 1 to 4 included in the holdout test set increased by two.

The predictive performance of these local models was reflected by macro-averaged AUROCs of 0.8831 for site 1, 0.5162 for site 5, and 0.7294 for site 6. Additionally, the mean macro-averaged AUROCs over 10 experiments were 0.9133, 0.9038, and 0.8959 for sites 2, 3, and 4, respectively (Fig. 4A). Notably, local models from sites 5 and 6 exhibited lower performance across different cancer subtypes compared with those from sites 1 to 4, likely due to the limited coverage of the cancer subtypes. In contrast, the centralized model, trained on aggregated data from all six sites, exhibited significantly superior performance, achieving a macro-averaged AUROC of 0.9999 across all cancer subtypes (Fig. 4A). By fully integrating both internal data (sites 1–4) and external data (sites 5 and 6), the centralized model captured a more comprehensive range of features and cancer subtypes, enhancing its predictive accuracy.

The ProCanFDL global model, trained in a federated manner across sites 1 to 6, showed substantial improvements over the local models. The global model achieved a macro-averaged AUROC of 0.9987 (Fig. 4A). The predictive performance of the global model was similar to that of the centralized model while maintaining data privacy. Accuracy measures for local, centralized, and ProCanFDL models are provided in Supplementary Fig. S8A–S8C. The AUROC for each cancer subtype is detailed in Fig. 4B and Supplementary Table S4. A confusion matrix was generated for the global model with the highest macro-averaged AUROC, providing a detailed assessment of true and false predictions for each subtype. The model achieved

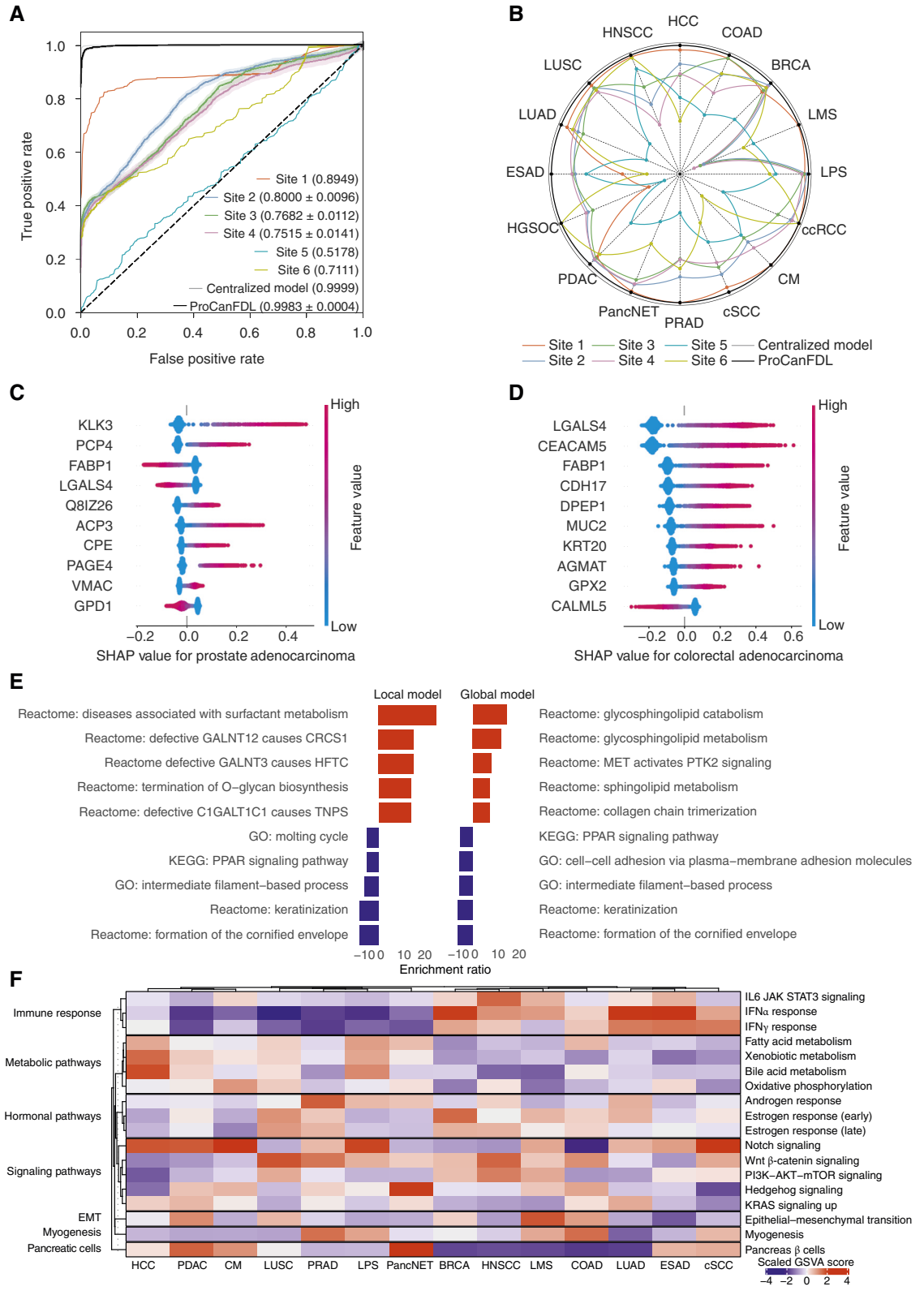

Figure 4. Generalization, integration, and model explanation. **A,** Performance of models was benchmarked by the AUROC with 95% confidence intervals estimated from the 10 experiments. **B,** AUROC of models across 16 cancer subtypes. **C,** Beeswarm plot demonstrating the top 10 proteins contributing to the prediction of prostate adenocarcinoma. **D,** Top 10 proteins contributing to the prediction of colorectal adenocarcinoma. **E,** Comparison of the top upregulated and downregulated pathways as identified by overrepresentation analysis using the top and bottom 200 proteins per SHAP value for lung adenocarcinoma between the global and local models. (*continued on following page*)

100% sensitivity for 9 out of 16 cancer subtypes, with particularly high sensitivity rates for subtypes such as pancreatic ductal adenocarcinoma, hepatocellular carcinoma, and head and neck squamous cell carcinoma (Supplementary Fig. S8D).

In summary, external validation of ProCanFDL, which also integrated datasets from two distinct MS platform types, showed that the global model consistently outperformed the local models. Importantly, the global model's performance matched that of the centralized model.

Model Explanation

For potential downstream clinical application and interpretation, understanding the learned relationships and key discriminatory proteins in the ProCanFDL global model is important. By leveraging Shapley Additive Explanation (SHAP) values (22), we identified the top features contributing to the cancer subtype predictions made by the ProCanFDL global model, which achieved the highest AUROC when trained with internal data (Supplementary Table S5). For example, proteins with utility in distinguishing between histologic types of cancer were identified. Desmoglein 3 (DSG3), a marker of squamous differentiation (23), was identified within the top 10 SHAP values for both cutaneous and head and neck squamous cell cancers, with SHAP values for this marker contributing negatively to the prediction of the other cancer subtypes (Supplementary Fig. S9A). Similarly, markers suggesting epithelial differentiation, such as anterior gradient 2 (AGR2), a protein disulfide isomerase family member (24), contributed positively to the identification of breast, colorectal, and pancreatic adenocarcinomas but negatively to the identification of tumors arising from other cell types, including pancreatic neuroendocrine tumors, sarcoma, and melanoma (Supplementary Fig. S9B). Finally, additional proteins were specific for identifying different tissues, including kallikrein-related peptidase 3 (prostate-specific antigen) and Purkinje cell protein 4 (PCP4) for prostate adenocarcinoma (Fig. 4C), as well as galectin 4 (LGALS4) and cadherin 17 (CDH17), which are known to be expressed in the intestinal epithelium, for colorectal adenocarcinoma (Fig. 4D; ref. 25). Cytokeratin 20 (KRT20) featured among the top 10 SHAP values for predicting colorectal adenocarcinoma and is known to have clinical utility for differentiating this cancer subtype from other subtypes of adenocarcinoma (26).

Overrepresentation analysis using the top 200 positive and negative SHAP values (Supplementary Fig. S10) for each cancer subtype revealed similar patterns with significant enrichment scores identified for specific or closely related cell types. Lung adenocarcinoma showed enrichment for alveolar type 2 cells driven by proteins, including napsin A (NAPSA), surfactant protein B (SFTPB), and lysophosphatidylcholine acyltransferase 1 (LPCAT2) (27). Squamous cancers were generally enriched for basal cells driven by proteins expressed in the squamous epithelium, including cytokeratin 6A (KRT6A), which has clinical utility in identifying squamous carcinoma (28). Esophageal and colorectal adenocarcinoma showed enrichment for enterocytes driven by proteins, including villin 1 (VIL1), as well as potential antibody–drug conjugate (ADC) targets, including carcinoembryonic antigen cell adhesion molecule 5 (CEACAM5; 29) and claudin 3 (CLDN3; 30, 31). This cell type contributed negatively to the identification of several cancer subtypes, including pancreatic neuroendocrine tumors, breast carcinoma, and leiomyosarcoma. Of note, in the FDL model, one liposarcoma case was misclassified as breast carcinoma due to the presence of proteins enriched in adipose tissue (fatty acid binding protein 4 (FABP4), perilipin 1 (PLIN1), and lipase E (LIPE); ref. 25) across both cancer subtypes.

Next, we evaluated biological associations using the top and bottom 200 proteins for each cancer subtype across several pathway databases (Supplementary Fig. S11). For pancreatic ductal adenocarcinoma, we identified pathways related to the extracellular matrix. Dense stroma is a key feature and a well-recognized barrier to efficient chemotherapeutic delivery in this cancer subtype. Pathways contributing positively to the identification of hepatocellular carcinoma highlighted a propensity for metabolic processes, including histidine and choline catabolism, consistent with the presence of hepatic tissue (32). We then compared the top identified pathways across the global and local models. The degree of overlap within the top and bottom 200 proteins, as identified by SHAP value for each model, was 221/400 (55%; Supplementary Fig. S12). Both models provided relevant biological information. For example, in both models, the top features contributing positively toward the identification of leiomyosarcoma were related to muscle development pathways, consistent with the known cell type of origin (33). Similarly, for squamous cancers, pathways relating to keratinization and the cornified envelope were identified, consistent with squamous epithelial origin (32). Further, these pathways contributed negatively to the identification of other cancer subtypes across both models, including colorectal cancer (Supplementary Fig. S11). However, there were notable differences between the global and local models, particularly for cancer subtypes underrepresented in the local model. For lung adenocarcinoma, pathways identified in the local model were related to underlying lung tissue, including surfactant metabolism and O-glycan biosynthesis due to the expression of multiple mucins, such as MUC5AC and MUC5B (34). In contrast, the global model identified pathways potentially more relevant to lung cancer biology, including the *MET* proto-oncogene. *MET* alterations are seen in more than 70% of lung adenocarcinoma tumor tissues (35), with multiple preclinical

Figure 4. ***(Continued)*** **F,** Heatmap showing the normalized enrichment scores for selected Hallmark gene sets across cancer subtypes in the global model. Pathways contributing positively toward the identification of each cancer subtype are shown in red, whereas pathways contributing negatively toward the identification of each cancer subtype are shown in blue. Plausible biological associations are identified within cancer subtypes. GO, Gene Ontology; GSVA, gene set variation analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes. BRCA, breast carcinoma; ccRCC, clear-cell renal cell carcinoma; COAD, colorectal adenocarcinoma; CM, cutaneous melanoma; cSCC, cutaneous squamous cell carcinoma; ESAD, esophageal adenocarcinoma; HNSCC, head and neck squamous carcinoma; HCC, hepatocellular carcinoma; HGSOC, high-grade serous ovarian carcinoma; LMS, leiomyosarcoma; LPS, liposarcoma; LUAD, lung adenocarcinoma; LUSC, lung squamous carcinoma; PancNET, pancreatic neuroendocrine tumor; PDAC, pancreatic ductal adenocarcinoma; PRAD, prostate adenocarcinoma.

and clinical studies currently investigating the role of c-MET–directed ADCs in this cancer subtype (36–38). Further, we identified several pathways relating to glycosphingolipid metabolism, which has been linked to lung cancer growth and progression (Fig. 4E; refs. 39–41). For colorectal cancer, the local model highlighted nonspecific pathways, including neutrophil degranulation and tissue-specific immune response. However, the global model highlighted pathways related to triglyceride metabolism and peroxisome proliferator–activated receptor signaling, which were driven by several fatty acid binding and perlipin proteins. FABP4 has been postulated as a diagnostic biomarker (42) for colorectal cancer and may have prognostic significance.

We then examined the protein features contributing positively and negatively to the prediction of each cancer subtype using the Hallmark gene set collection (Fig. 4F; ref. 43). For pancreatic adenocarcinoma and neuroendocrine tumors, we noted proteins specific to pancreatic beta cells, including glucagon (GCG), somatostatin (SST), and pyruvate kinase L/R (PKLR). For breast adenocarcinoma, we noted proteins related to estrogen response, including several *ETS* transcription factors (ETS1 and ELF1). We also noted plausible patterns related to cancer signaling pathways; for example, Notch signaling contributed positively toward the identification of pancreatic adenocarcinoma and melanoma (44). Several PI3K pathway targets were identified in head and neck squamous cancers, including EGFR, which is overexpressed in up to 90% of these cancers (45).

Finally, we looked at drug and immune treatment targets, including approved and emerging ADC targets (Supplementary Fig. S13). Again, plausible biological patterns were identified. For example, *ERBB2*, which is a well-established prognostic and therapeutic biomarker in breast cancer (46), showed a strong association with this cancer subtype. Trophoblast cell surface antigen 2 (TROP2), which is the target of the ADC sacituzumab govitecan, contributed positively to the identification of breast, head and neck squamous, lung squamous, and prostate adenocarcinoma. TROP2 contributed negatively to the identification of nonepithelial tumor entities, including liposarcoma, leiomyosarcoma, and melanoma, as well as colorectal adenocarcinoma, in keeping with published IHC and RNA data (47). Several markers involved in immune responses, such as signal transducer and activator of transcription 3 (*STAT3*), intercellular adhesion molecule 1 (*ICAM1*), and *CD274* (programmed death-ligand 1 (PD-L1), contributed positively to the identification of lung adenocarcinoma. Of note, PD-L1 expression has clinical utility in predicting the response of non–small cell lung cancer to checkpoint inhibitor therapy (48, 49).

Collectively, the proteins and pathways highlighted by model explanation analysis demonstrate that the FDL model predictions are plausibly related to known biological patterns that differ among cancer subtypes.

DISCUSSION

We have developed ProCanFDL, which enables accurate cancer subtyping using proteomic data derived from 7,525 FF and FFPE human biospecimens. The framework leverages data from 20 cancer research groups across seven countries and is processed in many different pathology laboratories, utilizing a privacy-preserving machine learning approach. Additionally, we demonstrated that ProCanFDL can integrate proteomic data generated via two different MS technologies.

The use of proteomic data in large-scale machine learning models has to date been limited by significant challenges, including the difficulties of integrating proteomic data from different platforms (50, 51). Moreover, data centralization is difficult to achieve with sensitive patient data, especially from different jurisdictions (7). FL offers a potential solution to these issues by allowing collaborative training across multiple sites without the need to transfer raw data. The use of FL with patient data, including chest X-rays and scans during the COVID-19 pandemic (10), illustrated the rapid gains that can be made by privacy-compliant data sharing.

Our model, which simulated FL by distributing data across four local sites behind private firewalls, demonstrated the feasibility of a real-world scenario whereby datasets are held across different institutions. By performing training independently at each local site and aggregating model updates centrally to form a global model, sensitive patient data can remain behind institutional firewalls. Notably, the model was enhanced by the incorporation of a distinct type of proteomic technology, TMT. Difficulties with the integration of data from different proteomic platforms have been a major issue in the aggregation of proteomic data. Therefore, ProCanFDL addresses multiple barriers to significantly scaling up proteomic machine learning analyses.

We used the macro-averaged AUROC as the primary model evaluation metric in this study. This method provides a more nuanced view of model performance compared with accuracy, especially in multiclass classification tasks in which class imbalances may exist. By assigning equal weighting to each class, regardless of sample size, it ensures that performance is assessed fairly across all cancer subtypes, preventing larger classes from dominating and biasing the evaluation metrics.

As with previous FL studies (10), the ProCanFDL global model demonstrated a 43% improvement over local models, highlighting the benefits of aggregating model parameters from local sites to enhance sample diversity and representation of cancer subtypes. This collaborative approach is particularly advantageous for sites with limited samples or underrepresented subtypes. By incorporating data from multiple local sites, the global model could accurately predict subtypes that were not present in the local datasets, thereby increasing its robustness and generalizability. Additionally, the global model achieved comparable evaluation metrics to the centralized model while offering the significant advantage of eliminating the need for data sharing between local sites. These attributes make ProCanFDL a practical and scalable solution for collaborative learning in proteomics, especially in situations in which data privacy is paramount. This advantage is further underscored when considering the limitations of "regulatorily clean" datasets from some local sites, where patient selection criteria often result in cohorts that are not representative of broader patient populations. This constrained representation can lead to reduced local predictive performance, yet our global model maintains robust accuracy through effective aggregation.

Unlike most studies that use independent datasets with uniform distribution for external validation, we applied a train–test split across external datasets to address the inherently nonuniform data distributions obtained from different proteomic platforms and to ensure model generalization and robustness. This approach allows the model to adapt to variations in data while maintaining privacy.

By model explanation analysis, we identified several important biological relationships, including the enrichment of proteins specific to cancer subtypes and relevant biological pathway information. This suggests that in addition to achieving high accuracy, the model can also capture meaningful biological signals. Such model explainability is vital for the acceptance of these technologies and their potential translation into clinical practice. We identified relevant ADC and immune targets, with patterns consistent with previous expression data. This highlights the unexplored potential of the ProCan Compendium to identify proteins with potential downstream prognostic and treatment applications. We anticipate that the application of ProCanFDL will facilitate our understanding of cancer biology and enable the use of proteomic data as an adjunct to histopathology in challenging diagnostic situations.

Although we demonstrated the utility of ProCanFDL, several limitations of the study provide scope for improvement and extension. One area for future research is the application of this FDL framework to more complex multi-institutional setups, in which even greater variations in data types, collection methods, and processing protocols could introduce additional data harmonization challenges. Our implementation used a simulated FL scenario in which global normalization was possible, but real-world FL deployments would require specialized techniques such as federated batch normalization (52) to harmonize data across sites without sharing raw values. This represents a significant technical challenge, as proteomics data normalization typically requires global statistics that cannot be directly shared in privacy-preserving scenarios. Another important direction is addressing the challenge of missing data by adapting FL frameworks such as FedIMPUTE (53) to handle missing data. Further, this study served as a proof of concept focused on cancer subtyping due to the availability of relevant annotations. Expanding its application to areas such as prognostic biomarker development and clinical outcome prediction will require datasets annotated with treatment outcomes, survival data, and other detailed metadata at each participating site.

ProCanFDL enables the development of foundation models for proteomics. Unlike large language models such as ChatGPT or foundation models for digital pathology, which benefit from large public datasets or readily available imaging data, the development of large proteomic models faces significant challenges due to restricted access to clinically annotated datasets. ProCanFDL will enable the gathering and utilization of the data needed to train proteomic foundation models without compromising data privacy. This federated approach is essential for creating the large, diverse datasets necessary for training robust and generalizable models. We envision that future iterations of ProCanFDL, trained on even larger and more diverse datasets through FL, will drive the development of pretrained foundation models for proteomics.

Overall, ProCanFDL represents a significant and tangible step toward applying FL to large-scale cancer proteomic datasets generated by different MS technologies. By balancing the need for robust and accurate model performance with data privacy, it fosters a practical and scalable solution for proteomic data analysis and collaborative biomedical research. We anticipate that this will create new opportunities in cancer research, for example, for the discovery of novel treatment targets and predictive biomarkers; accelerate the clinical application of proteomic technologies; and extend to multiomic data applications beyond proteomics.

METHODS

Biospecimen and Data Collection

FF and FFPE samples were obtained from malignant samples (tumor and premalignant samples) and nonmalignant tissues (benign tumors and tumor-adjacent normal samples). Ethics approval was obtained for the use of all patient samples. Cohort 1 consisted of FF samples (n = 766 primary tumor samples and n = 494 tumor-adjacent normal samples) obtained from the Victorian Cancer Biobank [2019/ETH02039 (HREC/17/WMEAD/63)], the Gynaecological Oncology Biobank (GynBiobank) at Westmead Hospital [2019/ETH02039 (HREC/17/WMEAD/63); 2019/ETH02043 (LNR/16/WMEAD/291)], and the Children's Medical Research Institute Legacy sample set [2019/ETH05866 (LNR/17/SCHN/291)]. Cohorts 2 to 30 included both FF and FFPE samples obtained from the following sites: Gynaecological Oncology Biobank (GynBiobank) at Westmead Hospital, Australia [2019/ETH02039 (HREC/17/WMEAD/63); 2019/ETH02043 (LNR/16/WMEAD/291)]; NCI Biobank, USA [2019/ETH02039 (HREC/17/WMEAD/63); 2019/ETH02075 (LNR/17/WMEAD/249)]; Westmead Institute for Medical Research, Australia [2019/ETH02039 (HREC/17/WMEAD/63); 2019/ETH10764 (LNR/19/WMEAD/39)]; Germans Trias i Pujol Research Institute, Badalona, Spain [2019/ETH06112 (HREC/17/SCHN/63), PI-17-079]; Royal Prince Alfred Hospital and Centenary Institute, Australia [2019/ETH02039 (HREC/17/WMEAD/63); 2021/ETH11460]; Royal Prince Alfred Hospital and Woolcock Institute, Australia [2019/ETH02039 (HREC/17/WMEAD/63); 2020/ETH01304 (X20-0223)]; St Vincent's Hospital, Department of Surgery, Sydney, Australia, and Institute of Clinical Sciences, Lund University Hospital, Sweden [2019/ETH02039 (HREC/17/WMEAD/63); 2021/ETH11590]; Garvan Institute of Medical Research (APGI), Australia [2019/ETH02039 (HREC/17/WMEAD/63); X16-0293 (HREC/11/RPAH/329)]; Garvan Institute of Medical Research [2019/ETH02039 (HREC/17/WMEAD/63)]; Melanoma Institute Australia [2019/ETH02039 (HREC/17/WMEAD/63); X15-0454 (HREC/11/RPAH/444); X17-0312 (HREC/11/RPAH/32); X15-0311 (HREC/10/RPAH/530)]; International Sarcoma Kindred Study [2019/ETH02039 (HREC/17/WMEAD/63); SVH 16/126; PMCC 09/11]; University of Manitoba Tissue Biobank, Canada [2019/ETH02039 (HREC/17/WMEAD/63); HS14811 (H2001-083)]; Australian Breast Cancer Tissue Bank [2019/ETH02039 (HREC/17/WMEAD/63); 2019/ETH02413 (LNR/16/WMEAD/93)]; Nepean Research Biobank, Australia [2019/ETH02039 (HREC/17/WMEAD/63)]; Princess Alexandra Hospital, Queensland Medical Labs, Mater Hospital, Queensland, Australia [2019/ETH02039 (HREC/17/WMEAD/63); PR/2022/QMS/8692 (HREC/03/QPAH/197)]; Institute of Cancer Research, Comprehensive Cancer Centre, Medical University of Vienna, Austria [2019/ETH02039 (HREC/17/WMEAD/63); 1312/2022]; the University of Sydney, Royal North Shore Hospital, NSW Health Pathology, Australia [2019/ETH02039 (HREC/17/WMEAD/63); 2019/ETH08639 (HREC/16/HAWKE/105)]; Jewish General Hospital Breast Cancer Biobank, Montreal, Canada [2019/ETH02039

(HREC/17/WMEAD/63); 2023-3377]; Laboratory of Translational Oncology, Medical School, University of Crete, and Laboratory of Pathology, University Hospital of Heraklion, Greece [2019/ETH02039 (HREC/17/WMEAD/63)]; and University of Crete (ref 27/February 17, 2020. University Hospital ref 9920).

Samples were sectioned as follows: 30-micron curls (at least one) for FF tissue and 10- to 20-micron curls (at least one) for FFPE tissue. A small proportion of samples underwent tissue punching (using a 1 mm biopsy punch tool) or macrodissection to enhance the proportion of tumor content.

Specimens were annotated according to the provided histologic cancer type and subtype diagnosis. For each sample, an adjacent hematoxylin and eosin slide was reviewed by a specialist pathologist to confirm that the section received in the proteomic laboratory was consistent with the diagnosis and to evaluate the percentage of tumor content and necrosis, the extent of lymphocytic infiltration, and the presence of additional tissue elements. Samples in cohort 1 that were not consistent with the provided diagnosis ($n = 84$) and/or contained a significant presence of atypical or normal tissue elements ($n = 116$), had a percentage tumor content <20% ($n = 57$), or had a percentage necrosis >80% ($n = 14$) were excluded from the FDL analyses.

Sample Preparation and Mass Spectrometric Acquisition

All samples were prepared using the Heat 'n Beat (17) method, and MS data were acquired from technical duplicate or triplicate MS runs using seven different SCIEX TripleTOF 6600 mass spectrometers.

Spectral Library Generation

To generate the spectral library, 19,930 DIA-MS runs from the ProCan Compendium were collected in .wiff file format and processed using DIA-NN software. MS/MS spectra were referenced to the UniProt human proteome (RRID:SCR_002380). The spectral library, containing 193,354 peptides, including retention time peptides and peptides from commonly occurring microbial and viral proteins, and corresponding to a total of 15,306 proteins, was used to search the entire set of 19,930 sample and quality control runs to extract the DIA data.

Data Extraction

Raw DIA-MS data were processed using DIA-NN software, implementing retention time–dependent normalization and the DIA-NN–generated spectral library. The input parameters are given below:

-report-lib-info --out step3-out.tsv --qvalue 0.01 --pg-level 1 --mass-acc-ms1 40 --mass-acc 40 --window 9 --int-removal 1 --matrices --temp . --smart-profiling --peak-center

Data were filtered to retain only precursors from proteotypic peptides with Global.Q.Value ≤0.01. Protein abundance was calculated using maxLFQ, with default parameters and implemented using the DIA-NN R package (https://github.com/vdemichev/diann-rpackage). Data were then $\log_2$-transformed.

Proteomic Profiling

All data were processed using custom R/Python scripts (RRID: SCR_001905, SCR_008394). For the entire set of samples from the ProCan Compendium, the Pearson's correlation coefficient was calculated using the corr function in the Python (RRID: 008394) package pandas (v2.0.3) to analyze the technical reproducibility of the data. In addition to the correlation among replicates from each sample, the mean correlation among samples from different cancer types, tissues of origin, MS instruments, and randomly selected samples was also calculated. Batch effects and clustering for these multicohort data were visualized using the UMAP dimensionality reduction tool. Moreover, the numbers of proteins quantified across tissues of origin and cancer types were assessed using box plots. The box plots showed the range of proteins quantified in each tissue and cancer type, along with the median protein count in each class.

To further investigate the proteomic profiles of cancer subtypes from various origins, we selected cell type–enriched proteins using previously defined thresholds (13). Cell type–enriched proteins were defined as proteins quantified in at least 50% of samples from no more than one cancer subtype and in ≤35% of samples from all other subtypes, considering only subtypes represented by at least 10 samples. For this, only tumor samples were used.

Preprocessing and Statistical Analysis

For downstream analysis, the sample replicates were merged, and a final protein matrix with only tumor samples was used. The protein matrix showed an average of 57% missingness per individual sample. Missing values were imputed with zero. No additional normalization or preprocessing was performed. In addition to filtering by replicate correlation, we also filtered samples in cohort 1 to include only those samples in the FDL analyses that met the following criteria: (i) consistent with the histopathologic diagnosis, (ii) adequate percentage of tumor content, and (iii) low percentage of necrosis (see "Methods" and "Biospecimen and Data Collection").

Preparation of Training and Test Sets

The train–test split of 90% training and 10% testing was performed at the patient level, ensuring that multiple samples from the same patient were consistently assigned to the training set. For patients assigned to the test set, one random sample was selected to simulate real-world conditions.

Hyperparameter Tuning

Hyperparameter tuning was conducted via a threefold cross-validation on cohort 1. Based on the cross-validation results, which helped identify the optimal architecture for model performance, the final architecture includes an input layer, a hidden layer, a Rectified Linear Unit (ReLU) activation function, a dropout layer with a probability of 0.2, and an output layer. The hyperparameters for the training process were set as follows: a learning rate of 1×10^{-4}, a weight decay of 1×10^{-4}, a hidden dimension size of 256, a batch size of 100, and a total of 200 epochs. The Adam optimizer was utilized to update the model parameters during training. Default settings were used for other hyperparameters that are not specifically mentioned above. These hyperparameters were then used for all models in local, centralized, and ProCanFDL learning.

ProCanFDL

ProCanFDL is a deep learning–based FL framework with a neural network architecture. FDL is conducted through iterative communication rounds between the central server and the participating local sites. The training procedure comprises the following four steps.

Step 1: Initialization and Local Training. A global model was first initialized with random weights and distributed to all participating local sites. This initial global model served as the starting point for the subsequent rounds of FL, ensuring that each site began the process with a common starting point. The model architecture was consistent across all sites, ensuring uniformity in training and subsequent aggregation. Each participating site locally trained its own instance of the deep learning model on its private proteomic data. PyTorch (v2.3.0) was utilized as the deep learning framework to

implement and train the model. During this phase, model parameters (weights and biases) were optimized using the Adam optimizer. To prevent overfitting and enhance generalization, techniques such as early stopping and dropout were applied, as detailed in the model's hyperparameter setup. The local models were trained independently, capturing unique proteomic signatures relevant to specific cancer types and tissues of origin. No raw data were shared between sites, ensuring data privacy.

Step 2: Global Model Aggregation. Following local training, the optimized model parameters, specifically the weights and biases, were securely transferred from each site to a central server. The server aggregated these updates using the federated averaging algorithm. The aggregation involved averaging the weights from all participating sites, resulting in a new global model that reflected the pooled knowledge from all local datasets, without the central server accessing any raw data. This step allowed the global model to capture the diversity of the proteomic data from all sites.

Step 3: Global Model Update. Once the aggregation was complete, the updated global model parameters were distributed back to all participating sites. Each site received the updated global model, which served as the starting point for the next round of local training. This iterative exchange allowed the model to progressively improve and adapt to the heterogeneous data across the sites.

Step 4: Iteration and Convergence. Steps 1 to 3 were repeated for a total of 10 iterations. This fixed number of iterations allowed the model to progressively refine its performance by incorporating data from all local sites. After the 10 iterations, the global model was evaluated on a hold-out test set to assess its generalizability across cancer subtypes. All global models in this study converged within 10 iterations, but the number of iterations may need to be increased for other data and tasks.

A pseudocode for this four-step algorithm is described below.

```
// Initialization
Initialize global_model with random weights
Distribute global_model to all local sites

// Iterative process with 10 iterations (Step 4)
for iteration in range(10):
    // Step 1: Local Training
    for each site in participating_sites:
        // Train the local model using the given data and hyperparameters
        local_model = train_model(global_model, site_data, hyperparameters)
        // Optimize the model parameters using the Adam optimizer
        local_weights = optimize(local_model, 'Adam', hyperparameters)
    end
    // Step 2: Global Model Aggregation
    // Send local model parameters (weights) to the central server
    local_weights = send_to_server(local_model_parameters)
    // Central server aggregates all local weights
    global_weights = federated_average(local_weights)
    // Step 3: Global Model Update
    // Update global model with aggregated weights from all sites
    global_model.update(global_weights)
    // Distribute updated global model back to the local sites
    distribute(global_model, participating_sites)
end
// Evaluation of the final global model on the test set
evaluate(global_model, hold_out_test_set)
```

Evaluation Metrics

The performance of ProCanFDL was measured using the following two metrics. The first was the AUROC, which is calculated for multiclass classification using the one-vs-rest approach. For each cancer subtype successively, that subtype is treated as the positive class, whereas the remaining subtypes are considered to be negative, allowing for the calculation of per-class AUROC. To more comprehensively evaluate the model's ability to discriminate between multiclass cancer subtypes, the macro-averaged AUROC is computed by averaging the AUROC scores across all classes without class-size weighting. This macro-average provides an overall measure of model performance across all classes, treating each class equally regardless of its prevalence in the dataset. The second metric is the multiclass accuracy, which measures the proportion of correctly classified cancer subtypes among the total instances:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

This provides a single value that summarizes the model's performance across all cancer subtypes.

Model Explanation Analysis

Feature importance scores from the FDL model were calculated using SHAP values with the Python package, SHAP (v0.45.1; ref. 22). In the beeswarm plots, features contributing positively to class prediction are shown on the right-hand side, and features contributing negatively are shown on the left-hand side. Overrepresentation analysis using the 200 top and bottom proteins, as indicated by SHAP values, was performed using the WebGestalt package (54) using the Human Cell Landscape, Reactome, Kyoto Encyclopedia of Genes and Genomes, and Gene Ontology: Biological Pathway databases. For all overrepresentation analyses, the q-value cutoff was set as 0.05, and the input background gene set encompassed all proteins used for building the relevant global or local models. Feature importance scores were also used to evaluate proteins for potential biological relevance using the Hallmark gene set collection (23). The ranked SHAP values were used to calculate normalized enrichment scores using the gsva function from the R GSVA package (v1.50.5; ref. 55) for the Hallmark gene sets obtained from the Molecular Signatures Database via the R msigdbr package (v7.5.1). Plots were generated using the R ComplexHeatmap (v2.18.0) and ggplot2 packages (v3.5.1).

Validation by External Datasets

To provide consistency in the underlying model structure and training process, all local models, the centralized model, and the ProCanFDL global model were trained using the same model architecture and hyperparameter configuration as previously applied in local, centralized, and FL. Z-score normalization was applied using the StandardScaler from scikit-learn (v1.4.2; RRID:SCR_002577) to both the DIA proteomic data (18, 19) and the eight TMT datasets from CPTAC (20) separately. Specifically, we first concatenated the sample-wise DIA proteomic data (ProCan Compendium and two external DIA datasets) into one matrix and then applied z-score normalization to this matrix. Similarly, the eight TMT cohorts were concatenated sample-wise into a single matrix, and z-score normalization was applied. Finally, the two normalized matrices were concatenated sample-wise to serve as the input for ProCanFDL.

Data Availability

The raw DIA-MS data and processed data of cohort 1 and the corresponding spectral library have been deposited in the Proteomics Identification database (PRIDE) (RRID:SCR_003411) under the dataset identifier PXD056810. The two external DIA-MS datasets have the identifiers PXD019549 and PXD007810, respectively.

Proteomics data for CPTAC datasets are available at the Proteomic Data Commons (PDC) at https://proteomic.datacommons.cancer.gov/pdc/cptac-pancancer. The PDC accession numbers for CPTAC datasets are as follows: BRCA, PDC000120; CCRCC, PDC000471; COAD, PDC000109; HNSCC, PDC000221; LUAD, PDC000153; PDAC, PDC000270; LSCC, PDC000234; and OV, PDC000250. The software code is available on GitHub at https://github.com/CMRI-ProCan/ProCanFDL.

Authors' Disclosures

E.L. Boys reports other support from Merck Sharp & Dohme and grants from Royal Australasian College of Physicians (RACP) and Institute of Clinical Pathology and Medical Research (ICPMR) Westmead, NSW Health Pathology outside the submitted work. C. Armengol reports grants from the Scientific Foundation of the Spanish Association Against Cancer, Fight Kids Cancer Funding Programme (BT4ChildLC project supported by Imagine for Margo, Fondation KickCancer, Foundation Kriibskrank Kanner, Federazione Italiana Associazioni Genitori e Guariti Oncoematologia Pediatrica, Cris Cancer Foundation), and the Agency for Management of University and Research Grants outside the submitted work. R. Clifton-Bligh reports personal fees from Kyowa Kirin, Lilly, Specialized Therapeutics, and the Endocrine Society outside the submitted work. A. DeFazio reports grants from the Australian Cancer Research Foundation, Cancer Institute NSW, and the National Health and Medical Research Council of Australia during the conduct of the study as well as grants from Cancer Council NSW, Australia; Medical Research Future Fund, Australia; Department of Defense, U.S. Army Medical Research and Materiel Command Congressionally Directed Medical Research Programs Ovarian Cancer Research Program Proteogenomics Research Award; Cancer Institute NSW; National Health and Medical Research Council of Australia; NSW Ministry of Health; Sydney Cancer Partners; Australia and New Zealand Gynaecological Oncology Group; and U.S. DoD Ovarian Cancer Research Program; nonfinancial support from Illumina Singapore Pte. Ltd.; and other support from AstraZeneca outside the submitted work. M. Filipits reports personal fees from AstraZeneca and Eli Lilly outside the submitted work. M. Gnant reports personal fees from Amgen, AstraZeneca, Bayer, Daiichi Sankyo, Eli Lilly, EPG Health (IQVIA), Menarini-Stemline, Merck Sharp & Dohme, Novartis, Pierre Fabre, and Veracyte outside the submitted work. L.G. Horvath reports grants and personal fees from Astellas and Bayer; personal fees from Janssen, Amgen, and Imagion Biosystems; and other support from Merck Sharp & Dohme outside the submitted work. J. Liu reports personal fees from Specialised Therapeutics and Taiho Therapeutics; other support from Starpharma, Innovent Biologics, ImmVirx, Merck Sharp & Dohme, Regeneron, Bristol Myers Squibb, AbbVie, AVEO, Virocure, Covus Pharmaceuticals, Relay Therapeutics, ALX Oncology, and IDEAYA Biosciences; and nonfinancial support from Greywolf Therapeutics outside the submitted work. G.V. Long reports personal fees from Agenus, Amgen, Array Biopharma, AstraZeneca, Bayer HealthCare, BioNTech, Boehringer Ingelheim, Bristol Myers Squibb, Evaxion Biotech, Fortiva, GI Innovation, Hexal AG, Highlight Therapeutics, IO Biotech, Immunocore, Innovent Biologics, Iovance Biotherapeutics, Merck Sharp & Dohme, Novartis Pharma, OncSec, PHMR, Pierre Fabre, Regeneron, Scancell Ltd, and SkylineDX B.V outside the submitted work. L. Morgan reports personal fees from GSK, BI, ReCODE, AstraZeneca, and Sanofi outside the submitted work. L. Murphy reports grants from CancerCare Manitoba Foundation and the Canadian Institute for Health Research during the conduct of the study. R.A. Scolyer reports grants from the National Health and Medical Research Council of Australia during the conduct of the study as well as personal fees from SkylineDx BV, IO Biotech ApS, MetaOptima Technology Inc., F. Hoffmann-La Roche Ltd, Evaxion, Provectus Biopharmaceuticals Australia, Qbiotics, Novartis, Merck Sharp & Dohme, NeraCare, AMGEN Inc., Bristol Myers Squibb, Myriad Genetics, and GSK outside the submitted work and that he has a patent pending. A. Swarbrick reports personal fees from Myeloid Therapeutics and grants from the National Breast Cancer Foundation and BCRF outside the submitted work. D. Thomas reports personal fees from Omico; grants, nonfinancial support, and other support from Roche; grants from AstraZeneca, Bayer, and Pfizer; other support from Medicenna, Boehringer Ingelheim, Servier, Hummingbird, Beigene, Merck Sharp & Dohme, and PMV Pharma; and grants and nonfinancial support from Sunpharma, Illumina, Eisai, and Elevation Oncology during the conduct of the study. R.L. Balleine reports grants from the National Health and Medical Research Council of Australia, the Cancer Institute New South Wales (NSW), the Cancer Institute NSW Sydney West Translational Cancer Research Centre, the National Breast Cancer Foundation Australia, the Australian Cancer Research Foundation, the NSW Ministry of Health, the Cancer Council NSW, the Ian Potter Foundation, and the Medical Research Future Fund and other support from the University of Sydney and the Department of Gynaecological Oncology at Westmead Hospital during the conduct of the study as well as other support from Illumina outside the submitted work. R.R. Reddel reports grants from the Australian Cancer Research Foundation, Cancer Institute NSW, NSW Ministry of Health, the University of Sydney, Cancer Council NSW, Ian Potter Foundation, Australian Government (Medical Research Future Fund), National Health and Medical Research Council of Australia, and National Breast Cancer Foundation during the conduct of the study as well as other support from Tessellate Bio BV outside the submitted work. No disclosures were reported by the other authors.

Authors' Contributions

Z. Cai: Conceptualization, software, formal analysis, investigation, visualization, methodology, writing–original draft, writing–review and editing. **E.L. Boys:** Conceptualization, software, formal analysis, investigation, visualization, methodology, writing–original draft, writing–review and editing. **Z. Noor:** Software, formal analysis, investigation, methodology, writing–original draft, writing–review and editing. **A.T. Aref:** Data curation, software, formal analysis, methodology, writing–review and editing. **D. Xavier:** Resources, data curation, formal analysis, writing–review and editing. **N. Lucas:** Resources, data curation, writing–review and editing. **S.G. Williams:** Resources, data curation, writing–review and editing. **J.M.S. Koh:** Resources, data curation, writing–review and editing. **R.C. Poulos:** Resources, data curation, writing–review and editing. **Y. Wu:** Writing–review and editing. **M. Dausmann:** Data curation, software, writing–review and editing. **K.L. MacKenzie:** Data curation, software, writing–review and editing. **A. Aguilar-Mahecha:** Resources, writing–review and editing. **C. Armengol:** Resources, writing–review and editing. **M.M. Barranco:** Resources, writing–review and editing. **M. Basik:** Resources, writing–review and editing. **E.D. Bowman:** Resources, writing–review and editing. **R. Clifton-Bligh:** Resources, writing–review and editing. **E.A. Connolly:** Resources. **W.A. Cooper:** Resources. **B. Dalal:** Resources. **A. DeFazio:** Resources. **M. Filipits:** Resources. **P.J. Flynn:** Resources. **J.D. Graham:** Resources. **J. George:** Resources. **A.J. Gill:** Resources. **M. Gnant:** Resources. **R. Habib:** Resources. **C.C. Harris:** Resources. **K. Harvey:** Resources. **L.G. Horvath:** Resources. **C. Jackson:** Resources. **M.R.J. Kohonen-Corish:** Resources. **E. Lim:** Resources. **J. Liu:** Resources. **G.V. Long:** Resources. **R.V. Lord:** Resources. **G.J. Mann:** Resources. **G.W. McCaughan:** Resources. **L. Morgan:** Resources. **L. Murphy:** Resources. **S. Nagabushan:** Resources. **A. Nagrial:** Resources. **J. Navinés:** Resources. **B.J. Panizza:** Resources. **J.S. Samra:** Resources. **R.A. Scolyer:** Resources. **J. Souglakos:** Resources. **A. Swarbrick:** Resources. **D. Thomas:** Resources. **R.L. Balleine:**

Resources, writing–review and editing. **P.G. Hains:** Resources, data curation, methodology, writing–review and editing. **P.J. Robinson:** Resources, supervision, funding acquisition, writing–review and editing. **Q. Zhong:** Conceptualization, formal analysis, supervision, investigation, visualization, methodology, writing–original draft, project administration, funding acquisition, writing–review and editing. **R.R. Reddel:** Resources, Conceptualization, supervision, funding acquisition, writing–review and editing.

Acknowledgments

ProCan is supported by the Australian Cancer Research Foundation, Cancer Institute New South Wales (NSW; 2017/TPG001, REG171150), NSW Ministry of Health (CMP-01), the University of Sydney, Cancer Council NSW (IG 18-01), Ian Potter Foundation, the Medical Research Future Fund, National Health and Medical Research Council (NHMRC) of Australia European Union grant (GNT1170739, a companion grant to support the European Commission's Horizon 2020 Program, H2020-SC1-DTH-2018-1, iPC: individualized pediatric cure), and the National Breast Cancer Foundation (IIRS-18-164). The work at ProCan was done under the auspices of a Memorandum of Understanding between Children's Medical Research Institute and the U.S. National Cancer Institute's International Cancer Proteogenomics Consortium that encourages cooperation among institutions and nations in proteogenomic cancer research, in which datasets are made available to the public. The Victorian Cancer Biobank, through the Cancer Council Victoria as Lead Agency, is supported by the Victorian Government through the Victorian Cancer Agency, a business unit of the Department of Health and Human Services. G.V. Long is supported by an NHMRC Investigator Grant (2021/GNT2007839) and by the University of Sydney Medical Foundation. R.C. Poulos and P.J. Robinson are supported by NHMRC Fellowships (GNT1138536 and GNT1137064, respectively). R.C. Poulos is supported by a Sydney Cancer Partners Translational Partners Fellowship with funding from a Cancer Institute NSW Capacity Building Grant (grant ID: 2021/CBG0002). Z. Cai is the recipient of a PhD Scholarship from Sydney Cancer Partners with funding from Cancer Institute NSW (2021/CBG0002). R.A. Scolyer is supported by a National Health and Medical Research Council of Australia (NHMRC) Investigator Grant (GNT2018514). Research conducted at the Westmead Institute for Medical Research (WIMR) was supported by Tour de Cure and the Cancer Institute NSW through the Sydney West Translational Cancer Research Centre (SW-TCRC, 15/TRC/1-01). Tissues were received from the Australian Breast Cancer Tissue Bank, which is generously supported by the NHMRC, the Cancer Institute NSW, and the NBCF. The Gynaecological Oncology Biobank at Westmead was funded by the NHMRC (ID310670, ID628903), the Cancer Institute NSW (12/RIG/1-17, 15/RIG/1-16), and the Department of Gynaecological Oncology, Westmead Hospital, and acknowledges financial support from the SW-TCRC funded by the Cancer Institute NSW (15/TRC/1-01). A. Aguilar-Mahecha is supported by the Guerrera Family Cancer Scientist Award. We gratefully acknowledge the contributions of Catherine Kennedy, Jessica Boros, Yoke-Eng Chiew, the Westmead Hospital Department of Gynaecological Oncology, clinical collaborators, and all the women who have consented to participate in research through the Westmead GynBiobank. The JGH Breast Biobank is supported by the RRCancer and the Quebec Breast Cancer Foundation. This work was partially supported by the Scientific Foundation of the Spanish Association Against Cancer (IGTP-AECC_2022-042/PRYCO223102ARME), the Fight Kids Cancer Funding Programme (BT4ChildLC project supported by Imagine for Margo, Fondation KickCancer, Foundation Kriibskrank Kanner, Federazione Italiana Associazioni Genitori e Guariti Oncoematologia Pediatrica, Cris Cancer Foundation), and AGAUR (2021-SGR-01186). Support for title page creation and format was provided by AuthorArranger, a tool developed at the NCI.

Note

Supplementary data for this article are available at Cancer Discovery Online (http://cancerdiscovery.aacrjournals.org/).

Received October 18, 2024; revised March 4, 2025; accepted May 30, 2025; posted first June 9, 2025.

REFERENCES

1. Makridakis S. The forthcoming Artificial Intelligence (AI) revolution: its impact on society and firms. Futures 2017;90:46–60.
2. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022;28:1773–84.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56.
4. Price WN 2nd, Cohen IG. Privacy in the age of medical big data. Nat Med 2019;25:37–43.
5. Marelli L, Testa G. Scrutinizing the EU general data protection regulation. Science 2018;360:496–8.
6. Marks M, Haupt CE. AI chatbots, health privacy, and challenges to HIPAA compliance. JAMA 2023;330:309–10.
7. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med 2020;3:119.
8. Ogier du Terrail J, Leopold A, Joly C, Béguier C, Andreux M, Maussion C, et al. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. Nat Med 2023;29:135–46.
9. Zhou J, Chen S, Wu Y, Li H, Zhang B, Zhou L, et al. PPML-Omics: a privacy-preserving federated machine learning method protects patients' privacy in omic data. Sci Adv 2024;10:eadh8601.
10. Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med 2021;27:1735–43.
11. Tully B, Balleine RL, Hains PG, Zhong Q, Reddel RR, Robinson PJ. Addressing the challenges of high-throughput cancer tissue proteomics for clinical application: ProCan. Proteomics 2019;19:e1900109.
12. Poulos RC, Cai Z, Robinson PJ, Reddel RR, Zhong Q. Opportunities for pharmacoproteomics in biomarker discovery. Proteomics 2022;23:e2200031.
13. Gonçalves E, Poulos RC, Cai Z, Barthorpe S, Manda SS, Lucas N, et al. Pan-cancer proteomic map of 949 human cell lines. Cancer Cell 2022;40:835–49.e8.
14. He F, Aebersold R, Baker MS, Bian X, Bo X, Chan DW, et al. π-HuB: the proteomic navigator of the human body. Nature 2024;636:322–31.
15. Poulos RC, Hains PG, Shah R, Lucas N, Xavier D, Manda SS, et al. Strategies to enable large-scale proteomics for reproducible research. Nat Commun 2020;11:3793.
16. Manda SS, Noor Z, Hains PG, Zhong Q. PIONEER: pipeline for generating high-quality spectral libraries for DIA-MS data. Curr Protoc 2021;1:e69.
17. Xavier D, Lucas N, Williams SG, Koh JMS, Ashman K, Loudon C, et al. Heat 'n Beat: a universal high-throughput end-to-end proteomics sample processing platform in under an hour. Anal Chem 2024;96:4093–102.
18. López-Sánchez LM, Jiménez-Izquierdo R, Peñarando J, Mena R, Guil-Luna S, Toledano M, et al. SWATH-based proteomics reveals processes associated with immune evasion and metastasis in poor prognosis colorectal tumours. J Cell Mol Med 2019;23:8219–32.
19. Nweke EE, Naicker P, Aron S, Stoychev S, Devar J, Tabb DL, et al. SWATH-MS based proteomic profiling of pancreatic ductal adenocarcinoma tumours reveals the interplay between the extracellular matrix and related intracellular pathways. PLoS One 2020;15:e0240453.
20. Li Y, Dou Y, Da Veiga Leprevost F, Geffen Y, Calinawan AP, Aguet F, et al. Proteogenomic data and resources for pan-cancer analysis. Cancer Cell 2023;41:1397–406.

21. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag 2020;37:50–60.
22. Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. J Comput Aided Mol Des 2020;34:1013–26.
23. Viehweger F, Azem A, Gorbokon N, Uhlig R, Lennartz M, Dwertmann Rico S, et al. Desmoglein 3 (Dsg3) expression in cancer: a tissue microarray study on 15,869 tumors. Pathol Res Pract 2022;240:154200.
24. Wang Z, Hao Y, Lowe AW. The adenocarcinoma-associated antigen, AGR2, promotes tumor growth, cell migration, and cellular transformation. Cancer Res 2008;68:492–7.
25. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science 2015;347:1260419.
26. Oien KA, Dennis JL. Diagnostic work-up of carcinoma of unknown primary: from immunohistochemistry to molecular profiling. Ann Oncol 2012;23(Suppl 10):x271–7.
27. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, et al. Construction of a human cell landscape at single-cell level. Nature 2020;581:303–9.
28. Kaufmann O, Fietze E, Mengs J, Dietel M. Value of p63 and cytokeratin 5/6 as immunohistochemical markers for the differential diagnosis of poorly differentiated and undifferentiated carcinomas. Am J Clin Pathol 2001;116:823–30.
29. Gazzah A, Bedard PL, Hierro C, Kang Y-K, Abdul Razak A, Ryu M-H, et al. Safety, pharmacokinetics, and antitumor activity of the anti-CEACAM5-DM4 antibody-drug conjugate tusamitamab ravtansine (SAR408701) in patients with advanced solid tumors: first-in-human dose-escalation study, Ann Oncol 2022;33:416–25.
30. Büyücek S, Schraps N, Menz A, Lutz F, Chirico V, Viehweger F, et al. Prevalence and clinical significance of Claudin-3 expression in cancer: a tissue microarray study on 14,966 tumor samples. Biomark Res 2024;12:154.
31. Romani C, Comper F, Bandiera E, Ravaggi A, Bignotti E, Tassi RA, et al. Development and characterization of a human single-chain antibody fragment against claudin-3: a novel therapeutic target in ovarian and uterine carcinomas. Am J Obstet Gynecol 2009;201:70.e1–9.
32. Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, et al. The reactome pathway knowledgebase 2024. Nucleic Acids Res 2024;52:D672–8.
33. Lacuna K, Bose S, Ingham M, Schwartz G. Therapeutic advances in leiomyosarcoma. Front Oncol 2023;13:1149106.
34. Groneberg DA, Eynott PR, Oates T, Lim S, Wu R, Carlstedt I, et al. Expression of MUC5AC and MUC5B mucins in normal and cystic fibrosis lung. Respir Med 2002;96:81–6.
35. Ichimura E, Maeshima A, Nakajima T, Nakamura T. Expression of c-met/HGF receptor in human non-small cell lung carcinomas in vitro and in vivo and its prognostic significance. Jpn J Cancer Res 1996;87:1063–9.
36. Mer AH, Mirzaei Y, Misamogooe F, Bagheri N, Bazyari A, Keshtkaran Z, et al. Progress of antibody-drug conjugates (ADCs) targeting c-Met in cancer therapy; insights from clinical and preclinical studies. Drug Deliv Transl Res 2024;14:2963–88.
37. Gymnopoulos M, Betancourt O, Blot V, Fujita R, Galvan D, Lieuw V, et al. TR1801-ADC: a highly potent cMet antibody-drug conjugate with high activity in patient-derived xenograft models of solid tumors. Mol Oncol 2020;14:54–68.
38. Tian Y, Sun X, Yang C, Liao C. WS02.11 HRA00129-C004, a novel c-met ADC with promising preclinical anti-tumor activity and expanded therapeutic window. J Thorac Oncol 2023;18:S38.
39. Petrache I, Berdyshev EV. Ceramide signaling and metabolism in pathophysiological states of the lung. Annu Rev Physiol 2016;78:463–80.
40. Chen Y, Ma Z, Min L, Li H, Wang B, Zhong J, et al. Biomarker identification and pathway analysis by serum metabolomics of lung cancer. Biomed Res Int 2015;2015:183624.
41. Meng Q, Hu X, Zhao X, Kong X, Meng Y-M, Chen Y, et al. A circular network of coregulated sphingolipids dictates lung cancer growth and progression. EBioMedicine 2021;66:103301.
42. Zhang Y, Zhao X, Deng L, Li X, Wang G, Li Y, et al. High expression of FABP4 and FABP6 in patients with colorectal cancer. World J Surg Oncol 2019;17:171.
43. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 2015;1:417–25.
44. Shi Q, Xue C, Zeng Y, Yuan X, Chu Q, Jiang S, et al. Notch signaling pathway in cancer: from mechanistic insights to targeted therapies. Signal Transduct Target Ther 2024;9:128.
45. Kalyankrishna S, Grandis JR. Epidermal growth factor receptor biology in head and neck cancer. J Clin Oncol 2006;24:2666–72.
46. Swain SM, Shastry M, Hamilton E. Targeting HER2-positive breast cancer: advances and future directions. Nat Rev Drug Discov 2023;22:101–26.
47. Dum D, Taherpour N, Menz A, Höflmayer D, Völkel C, Hinsch A, et al. Trophoblast cell surface antigen 2 expression in human tumors: a tissue microarray study on 18,563 tumors, Pathobiology 2022;89:245–58.
48. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, et al. Pembrolizumab versus chemotherapy for PD-L1–positive non–small-cell lung cancer. N Engl J Med 2016;375:1823–33.
49. Mok TSK, Wu Y-L, Kudaba I, Kowalski DM, Cho BC, Turna HZ, et al. Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): a randomised, open-label, controlled, phase 3 trial. Lancet 2019;393:1819–30.
50. Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience 2022;25:103798.
51. Cai Z, Apolinário S, Baião AR, Pacini C, Sousa MD, Vinga S, et al. Synthetic augmentation of cancer cell line multi-omic datasets using unsupervised deep learning. Nat Commun 2024;15:10390.
52. Du Z, Sun J, Li A, Chen P-Y, Zhang J, Li "Helen" H, et al. Rethinking normalization methods in federated learning. In: Proceedings of the 3rd International Workshop on Distributed Machine Learning [Internet]. New York (NY): ACM; 2022 [cited 2025 May 2]. Available from: http://dx.doi.org/10.1145/3565010.3569062.
53. Li S, Yan M, Yuan R, Liu M, Liu N, Hong C. FedIMPUTE: privacy-preserving missing value imputation for multi-site heterogeneous electronic health records. J Biomed Inform 2025;165:104780.
54. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 2005;33:W741–8.
55. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 2013;14:7.