PPML-Omics: a Privacy-Preserving federated Machine Learning system protects patients' privacy from omic data

Juexiao Zhou[1,2†], Siyuan Chen[1,2†], Yulian Wu[1†], Haoyang Li[1,2], Bin Zhang[1,2], Longxi Zhou[1,2], Yan Hu[1], Zihang Xiang[1], Zhongxiao Li[1,2], Ningning Chen[1,2], Wenkai Han[1,2], Di Wang[1*] and Xin Gao[1,2*]

[1]Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955, Saudi Arabia.
[2]Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, 23955, Saudi Arabia.

*Corresponding author(s). E-mail(s): xin.gao@kaust.edu.sa; di.wang@kaust.edu.sa
Contributing authors: juexiao.zhou@kaust.edu.sa; siyuan.chen@kaust.edu.sa; yulian.wu@kaust.edu.sa; haoyang.li@kaust.edu.sa; bin.zhang@kaust.edu.sa; longxi.zhou@kaust.edu.sa; yan.hu@kaust.edu.sa; zihang.xiang@kaust.edu.sa; zhongxiao.li@kaust.edu.sa; ningning.chen@kaust.edu.sa; wenkai.han@kaust.edu.sa;
†These authors contributed equally to this work.

Abstract

Modern machine learning models towards various tasks with omic data analysis give rise to threats of privacy leakage of patients involved in those datasets. Despite the advances in different privacy technologies, existing methods tend to introduce too much noise, which hampers model accuracy and usefulness. Here, we built a secure and privacy-preserving machine learning (PPML) system by combining federated learning (FL), differential privacy (DP) and shuffling mechanism.

We applied this system to analyze data from three sequencing technologies, and addressed the privacy concern in three major tasks of omic data, namely cancer classification with bulk RNA-seq, clustering with single-cell RNA-seq, and the integration of spatial gene expression and tumour morphology with spatial transcriptomics, under three representative deep learning models. We also examined privacy breaches in depth through privacy attack experiments and demonstrated that our PPML-Omics system could protect patients' privacy. In each of these applications, PPML-Omics was able to outperform state-of-the-art systems under the same level of privacy guarantee, demonstrating the versatility of the system in simultaneously balancing the privacy-preserving capability and utility in omic data analysis. Furthermore, we gave the theoretical proof of the privacy-preserving capability of PPML-Omics, suggesting the first mathematically guaranteed model with robust and generalizable empirical performance.

Keywords: Differential privacy, Machine learning, Federated learning, PPML, Bulk RNA-seq, Single-cell RNA-seq, Spatial transcriptomics

1 Main

Individual privacy in biology and biomedicine is emerging as a big concern [1] with the development of biomedical data science in recent years. A deluge of genetic data from millions of individuals are generated from massive research projects in the past few decades, such as the Cancer Genome Atlas (TCGA) [2], the 100,000 genome project [3] and the Earth BioGenome Project (EBP) [4] from high-throughput sequencing platforms [5]. Those datasets may lead to potential leakage of genetic information and the privacy concern on ethical problems like genetic discrimination [6]. As a well-known case, genome-wide association study (GWAS) is one of the most common genomic studies that help uncover associations between single nucleotide polymorphisms (SNPs) and phenotypic traits like pathological status [7]. In addition, the aggregated statistics about SNPs are usually thought to be safe to publish publicly as they are pooled from thousands of individuals. Nevertheless, Homer et al.[8] showed that the attacker can determine whether the individual is in the case group of a GWAS with only minor allele frequencies of the SNPs used in the study and other publicly available information such as individual's genotype. As a result of such attacking, the NIH restricted the access to key results and data of GWAS to only trusted individuals [9]. In addition to SNPs, a large amount of potentially private genetic information from modern multi-modal sequencing platforms, including bulk RNA sequencing (RNA-seq) [10], single-cell RNA sequencing (scRNA-seq) [11] and spatial transcriptomics [12] might also be exposed as more and more data are being published (**Supplementary Section 1.1**). However, up to now, the community only has a limited view of such privacy leakage hidden behind.

In addition to the data itself, another risk factor to data privacy is the wide scope of applications of machine learning, especially deep learning, which evolves rapidly by taking advantage of large datasets. A massive number of models and applications, which require training on a large and diverse dataset (either public data or in-house data), are being created, shared and applied to various areas, such as the genomics [13], medical imaging [14] and healthcare [15]. Nevertheless, individual privacy is being exposed to high risk, leading to the new concern of privacy in modern artificial intelligence (AI) [16]. As shown in **Figure 1a**, typically, sensitive data may exist in a distributed manner, where the data owners do not want to share the raw data for privacy reasons, while the algorithm owners want to access enough data to improve model utility. To balance the needs of both, we need a machine learning framework for distributed data that can balance utility and privacy-preserving capabilities: a secure and privacy-preserving machine learning (PPML) system.

To alleviate the leakage of privacy, the most commonly used strategy is the anonymization or pseudonymization of sensitive data before transmitting to the data-sharing centre [17]. Unfortunately, recent studies showed that anonymization was insufficient towards re-identification attacks [18] and linking attacks [19]. To overcome the shortness of centralized data sharing and model training, federated learning (FL) was proposed in 2017 as a data-private collaborative learning method [20]. The collaborating institutions train a machine learning (ML) model with their own data in parallel and send the model updates to the central server, which can aggregate all model updates into a consensus model without accessing the raw data. Nevertheless, the distributed nature of FL gives rise to new threats of privacy leakage caused by potentially malicious participants [21–24], such as data poisoning attacks [25], membership inference attack [26–28], source inference attack (SIA) [29] and data reconstruction attack [30]. Hence, exposing the trained model to a non-trusted user may also cause privacy leakage [31].

To further strengthen FL's privacy guarantee to preserve privacy, additional privacy-enhancing modules are required. Thus, an innovative and state-of-the-art solution named differential privacy (DP) was proposed [32–34] (**Supplementary Section 1.2**). In addition to DP, multi-party computation (MPC) or multi-party homomorphic encryption (MHE) frameworks use cryptographic techniques to protect the data while enabling the training of ML models with perfect accuracy and these techniques have been used to secure FL training [35–37]. However, those methods introduce different trade-offs compared to DP-based solutions, such as bigger computation and communication overheads, which might be difficult to satisfy by some clients in practice, while DP yields a privacy-utility trade-off, under which we could protect privacy by sacrificing some data utility. Currently, many works in the literature utilized the DP notion to provide formal privacy guarantees for the participants in the research of SNPs and GWAS, which is a relatively narrow and specific problem in genomic studies and whose data is obtained by post-processing the raw sequencing data [9, 38–44] (**Supplementary Section 1.3**). In addition to

SNPs, raw sequencing data saved in matrix format contains much more sensitive information. However, there is no work studying privacy protection in sequencing data with DP from the bigger picture, even though raw sequencing data contains a large amount of private information of patients in addition to SNPs. Meanwhile, the state-of-the-art work related to applying DP in biology only reported the privacy protection in medical imaging with the FL-DP-only system [45, 46], which is unclear to be able to generalize to omic data analysis tasks, because omic data have very different characteristics to imaging data.

Here, we proposed a robust and powerful PPML-Omics system by integrating the FL, DP and the shuffling mechanism (**Figure 1b**). For the first time, we applied PPML-Omics to analyze data from three representative omic data analysis tasks, which were solved with three different but representative deep learning models. We demonstrated how to address the privacy concern in the cancer classification from TCGA with bulk RNA-seq [10], clustering with scRNA-seq [47], and the integration of spatial gene expression and tumour morphology with spatial transcriptomics [48, 49]. In addition, we examined in depth the privacy breaches that existed in all three tasks through privacy attack experiments and demonstrated that patients' privacy could be protected through our PPML-Omics system as shown in **Figure 1c**. In each of these applications, we showed that PPML-Omics was able to outperform state-of-the-art methods under the same level of privacy guarantee, demonstrating the versatility of the system in simultaneously balancing the privacy-preserving capability and utility in sequencing data analysis. Finally, we proved the privacy-preserving capability of PPML-Omics theoretically (**Supplementary Section 1.4**), suggesting the first mathematically guaranteed method with robust and generalizable empirical performance in protecting patients' privacy from omic data.

2 Results

2.1 Application 1. Cancer classification with bulk RNA-seq

Bulk RNA-seq can directly disclose patients' gene expression, while machine learning models trained with bulk RNA-seq can indirectly leak patients' signature expressed genes [43]. Here, we collected data from TCGA (**Methods and Supplementary Table 1**) and compared (i) computing resource requirements, (ii) privacy-preserving capabilities, and (iii) utility on data of different methods by working on the cancer classification task with gene expression as inputs.

To study the robustness and demonstrate the advantages of computational resources required by our method, we used this task as a benchmark for the profiling analysis to test different systems (level 1: with or without FL, DP and shuffling respectively together with a network specified in the level 2) against varied machine learning networks (level 2) (**Figure 2a**), a pair of which constitute a method/architecture. Depending on the differences in level 1, we

tested four systems, including (i) centrally trained system, which was trained with one model on the entire dataset pooled on a single machine; (ii) FL system, which was trained with separate models on the individual data owners' subsets of the dataset following the protocol of federated learning; (iii) FL + DP, in which the DP mechanism was integrated on (ii); (iv) PPML-Omics, in which the shuffling mechanism was further integrated on (iii). Furthermore, for each system in level 1, variants based on the fully connected neural network (FCN) (**Figure 2a**) were tested (level 2), including the use of different numbers of hidden layers, activation functions, dropout layers and values of ϵ for the 37-class classification task, where the privacy budget ϵ was a hyperparameter (Methods) and a smaller ϵ means that we have stricter requirements for privacy protection, resulting in the need to add more noise and lower model performance. Data visualization through t-SNE showed that the use of gene expression could effectively distinguish between different cancer types (**Figure 2b**). We trained all models to convergence on the same dataset, measured the privacy-preserving capability (**Methods**) and required computational resources, including time, RAM and GPU memory (**Supplementary Section 2.2**). Following the Li's definition of the privacy loss in [50], we used Jensen-Shannon (JS) divergence and the p-value of Kolmogorov–Smirnov (KS) test (P_{KS}) between two distributions to quantify privacy leakage, where a larger JS divergence and a smaller p-value indicate a more severe privacy leakage. Compared to the centrally trained system, the FL system did not exhibit significant privacy-preserving power from the JS divergence perspective (JS divergence >0.1 and similar to the centrally trained system), while the integration of the DP (FL + DP system) substantially improved the privacy-preserving power (JS divergence was much smaller than 0.1), and the JS divergence gradually decreased as the privacy budget ϵ decreased, as shown in **Figure 2c**, indicating that adding higher noise allowed for stronger privacy protection. As a trade-off between privacy protection and the data utility, $\epsilon = 5$ is usually chosen in practice (**Methods**). Therefore, with $\epsilon = 5$, PPML-Omics (JS divergence=0.00037, $P_{KS} = 0.91350$) showed a clear advantage over the centrally trained model (JS divergence=0.09712, $P_{KS} = 5.430 \times 10^{-5}$), the FL-only system (JS divergence=0.10797, $P_{KS} = 1.430 \times 10^{-5}$) and the FL-DP-only system (JS divergence=0.00063, $P_{KS} = 0.85142$) when we need to meet both acceptable computational requirements (**Supplementary Table 2**) and privacy protection capabilities (**Figure 2c**). Overall, the computational resource requirements of PPML-Omics were notably user-friendly.

Adding noise to gradients with DP has an amplification effect [51], meaning that the more times of adding noise, the greater the effect on model performance. Consequently, when applying FL systems to real-life scenarios, we need to handle the situation with multiple users. To study the impact of the number of clients (users) on the model performance under different systems and demonstrate the advantage of PPML-Omics over the FL-DP-only system, we tested the model performance by varying the number of clients on the same dataset. Though the number of clients does not have a significant impact on

the results of the FL-only system, since the integration of the DP mechanism requires adding noise to the gradient, increasing the number of clients had a significant impact on reducing the model performance of FL-DP-only system (**Figure 2d**). As a consequence, a large number of clients in real life would be the biggest challenge for applying the FL-DP-only system [46]. Surprisingly, the integration of the shuffling mechanism could effectively weaken the performance degradation caused by increasing the number of clients, indicating that PPML-Omics could maintain better performance with a larger number of clients, making it a feasible solution in practical applications.

Different cancer types have specifically expressed genes, which can be reconstructed by the attacker from the released models, thus compromising the potential privacy of the patient as part of the model training data. To investigate the privacy protected by PPML-Omics, we adopted the model inversion attack (MIA) for cancer classification models (**Methods**). In other words, we tried to optimize a gene expression vector that could give the highest prediction probability for a particular cancer type on the target model by using MIA. Then we analyzed the significantly expressed genes in the optimal gene expression vector. If the significantly expressed genes were cancer type-specific (with high expression in the target cancer type and low expression in other cancer types), then we could conclude that the target model leaked privacy. As shown in **Figure 2e**, the most significant genes reconstructed from the Baseline (centrally trained system), referring to the FCN with 3 hidden layers and the rectified linear unit (ReLU), showed significantly different distribution (JS divergence = 0.09712, $P_{KS} = 5.430 \times 10^{-5}$) on the z-score normalized real expression between two groups (solid lines for the target cancer type and dashed lines for other cancer types) compared to the one with Baseline + FL + DP ($\epsilon = 5$) + shuffling (JS divergence = 0.00037, $P_{KS} = 0.91350$), suggesting that the MIA on the centrally trained system could reconstruct genes with significantly different expression levels (privacy leakage) on the target cancer type and other cancer types. In other words, based on the published centrally trained models, the attacker could reconstruct the corresponding specifically expressed genes for each cancer type. In contrast, it was impossible to accurately reconstruct the corresponding specifically expressed genes for each cancer type from the published models using PPML-Omics. A significant difference in the expression distribution of genes specifically reconstructed by the centrally trained system in the target cancer type and other cancer types could be observed (COADREAD: $P_{KS} = 1.513 \times 10^{-7}$, KICH: $P_{KS} = 3.423 \times 10^{-7}$ and UVM: $P_{KS} = 1.430 \times 10^{-7}$) (**Figure 2f**), implying that genes reconstructed for each cancer type tended to have a higher expression in the corresponding cancer type. In contrast, the expression distribution of genes reconstructed by attacking PPML-Omics was very similar in the target cancer type and in other cancer types (COADREAD: $P_{KS} = 0.85142$, KICH: $P_{KS} = 0.77774$ and UVM: $P_{KS} = 0.52523$) (**Figure 2f**), implying that the genes reconstructed for each cancer type were not strongly correlated with the cancer type in terms of expression. The similar observation that reconstructed

genes by the centrally trained system showed significantly different expression between the target cancer type and other cancer types while reconstructed genes by the PPML-Omics did not show such significant difference could be drawn across all cancer types (**Supplementary Section 2.3, Supplementary Figure 1, 2 and 3 and Supplementary Tables 3-6**). Overall, we showed that PPML-Omics protects genomic privacy by obfuscating sensitive gene identification.

Compared to the centrally trained system, the FL-only system and the FL-DP-only system, another obvious advantage of PPML-Omics was the better data utility, meaning that PPML-Omics could retain a higher level of data utility while protecting the data privacy. To show the data utility of PPML-Omics in the cancer classification task, in the ablation study, we compared the accuracy and macro-F1 score on the testing dataset of the centrally trained system, the FL-only system, the FL-DP-only system and the PPML-Omics system (**Figure 2g and Supplementary Table 7**). The FL-only system achieved slightly worse performance (accuracy=0.723) than the centrally trained system (accuracy=0.744) ($P = 0.00201$ for one-sided Student's t-test). Meanwhile, the FL-DP-only system under $\epsilon = 5$ balanced both the utility (accuracy=0.722) and privacy guarantee (**Figure 2c**). With same privacy guarantee ($\epsilon = 5$), PPML-Omics showed even better performance (accuracy=0.748) than the FL-DP-only system (accuracy=0.722) ($P = 0.00024$ for one-sided Student's t-test). Since biological data is inherently noisy, it is possible that the noise we add with DP could potentially offset such noise, thus PPML-Omics (accuracy=0.748) was slightly but not significantly better than the centrally trained system (accuracy=0.744) in terms of accuracy. In summary, PPML-Omics achieved good utility while preserving the privacy of gene expression.

2.2 Application 2. Clustering with scRNA-seq data

scRNA-seq is a revolutionary technology to quantify gene expression in thousands and even millions of cells in parallel, which is powerful to identify cell populations and dissect the composition of each population in biological samples [52]. Applying scRNA-seq on primary tumours is not only able to decipher tumour cell heterogeneity [53] but can also uncover the specificity of tumour micro-environment [54] in each patient, which may lead to a personalized treatment strategy [55]. For instance, immune infiltration in the tumour that could be determined by scRNA-seq, is a good response indicator to immune blockade-based therapy. Thus, information of cell populations is critical in protecting patients' privacy when analyzing scRNA-seq data. Using the extracted features rather than the direct gene expression may avoid the leak of the specifically expressed gene while still harbouring the information for each cell population in the patients. Even so, obtaining accurate cell population results on patients' tissue could also violate privacy. Here, we used the low dimension features extracted from gene expression vectors of scRNA-seq data by Auto-encoder as input for k-means clustering. We then applied PPML-Omics to see how it protected patients' privacy from the results of clustering, by evaluating (i)

the data utility and (ii) the privacy-preserving capability in terms of cell type classification and composition quantification based on clustering (**Figure 3a**).

The number and composition of cells included in the scRNA-seq data vary greatly depending on the sample. To assess the robustness of PPML-Omics for different sample sizes in the clustering task, we evaluated it on four benchmark datasets (Yan, Pollen, Camp-Liver and Hrvatin) with varying cell numbers, ranging from 90 to 48,266. Following the Tran's work [56], we compared the Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Cluster Accuracy (CA), and Jaccard index (JI) (**Methods**) on the four datasets with different systems (**Table 1 and Supplementary Table 8**), where a larger value means better clustering results. For the dataset 'Yan', 'Camp-Liver' and 'Hrvatin', PPML-Omics achieved even better performance compared to the baseline on all four metrics ('Yan': $P = 1.840 \times 10^{-2}$, 'Camp-Liver': $P = 1.480 \times 10^{-3}$ and 'Hrvatin': $P = 4.880 \times 10^{-4}$), indicating that PPML-Omics even had the potential to surpass the centrally trained system in clustering task and noise added during training using the DP mechanism might be beneficial for unsupervised tasks. To ensure that the performance of PPML-Omics was at the same level as those commonly acknowledged tools in the clustering task with scRNA-seq data, we compared PPML-Omics with the existing state-of-the-art tools, including Seurat [57], SC3[58], CIDR[59] and SINCERA [60] (**Supplementary Section 3.2 and 3.3, Supplementary Figure 4 and Supplementary Tables 9-10**). PPML-Omics also achieved competitive utility as shown in **Supplementary Section 3.2**, proving that our method achieved the same performance as the most commonly used tools for the clustering task. To further investigate the effect of FL, DP and shuffling mechanism on the clustering task, we conducted an ablation study (**Figure 3b**)(centrally trained system at the first column, FL-only system at the second column, FL-DP-only system at the third column, and PPML-Omics at the fourth column) on these four datasets of different sizes and visualized the clustering results with uniform manifold approximation (UMAP) algorithm that projected the internal representations into a two-dimensional space. In conclusion, PPML-Omics could qualitatively and quantitatively compete or even outperform other systems in terms of utility.

When we obtain the scRNA-seq data of patients, we could learn the composition of the cells based on the clustering results. In extreme cases, suppose the clustering method is good enough and can 100% correctly distinguish different types of cells into clusters, it might potentially violate the patient's privacy by leaking some sensitive sub cell-types. Therefore, it is necessary to reasonably adjust the resolution of the clustering method to properly hide some small clusters (sub-types) from the final clustering results to achieve privacy protection according to the prior privacy requirements. In other words, for a system, if we cannot observe some sensitive small clusters (sub-types) in the final clustering results, then we could conclude that the system protects the patient's privacy. Additionally, if the clustering results for major cell types are reasonable, we could conclude that the system preserves an acceptable degree of data utility

while protecting privacy. To study the privacy-preserving power of PPML-Omics, we analyzed a public scRNA-seq dataset with 43,817 cells from 10 colon cancer patients [61], in which the cells were classified into 5 major cell types (hB: B cells, hI: Innate Iymphoid cells, hM: Myeloid cells, hT: CD4+ T cells and CD8+ T cells) and 38 sub-types **(Supplementary Section 3.5, Supplementary Tables 11-12**). We applied three systems for clustering, namely the centrally trained system and two of our systems with different noise levels ($\epsilon = 1$ and $\epsilon = 0.01$). With two patients (P0123 and P0408) as examples, the major cell types were successfully identified with all three systems (**Figure 3c**), indicating the overall clustering results of the three systems were reasonable, thus preserving the utility of the data. However, the clustering results from the three systems showed significant differences in the resolution of the sub-types. Comparing both our systems ($\epsilon = 1$ and 0.01) with the centrally trained system, 1 and 3 sub-clusters (sub-types) were undetected respectively for P0123 and 2 and 5 sub-clusters were undetected respectively for P0408. As expected, PPML-Omics with larger noise ($\epsilon = 0.01$) missed more sub-clusters (n = 3 for P0123 and n = 5 for P0408) **(Supplementary Tables 11-12**), suggesting that the local differences between sub-clusters (sub-types) could be diluted due to the DP and shuffling mechanisms in PPML-Omics and then those sub-types could be integrated into major clusters, thus achieving privacy protection as a consequence. For potential users of PPML-Omics, the value of ϵ could be adjusted according to the needs of the actual application scenario to achieve both good utility and privacy protection.

2.3 Application 3. Integration of tumour morphology and gene expression with spatial transcriptomics

Sequencing data in applications 1 and 2 could reveal much sensitive information about patients, and the potential risk of privacy disclosure is also very high for spatial transcriptomics as one of the highest resolution sequencing technologies. In addition, spatial transcriptomics requires the usage of tissue images, which adds additional privacy leakage to some extent. ST-Net [49] is a machine learning model for predicting local gene expression from haematoxylin-and-eosin-stained pathology slides trained on a dataset of 30,612 spatially resolved gene expression data matched to histopathology images from 23 patients with breast cancer. Unlike the classical task of spatial transcriptomics, ST-Net predicts the spatially resolved transcriptome of a tissue directly from tissue images, while the gene expression measured from the spatial transcriptomics is used as the ground truth in the training phase. Here, (i) to show PPML-Omics realized competitive utility and (ii) to investigate how PPML-Omics protects privacy from those histopathology images on this task, we applied PPML-Omics to integrate tumour morphology and gene expression with spatial transcriptomics by incorporating the ST-Net model as the backbone network into PPML-Omics as shown in **Figure 4a**.

To demonstrate that PPML-Omics has the same level of utility compared to ST-Net on predicting local gene expression from pathology slides,

we visualized the prediction results and quantitatively compared the similarity between the predicted results and the ground truth by calculating the mean square error (MSE), where a larger MSE means a worse prediction. In order to amplify the difference between the predicted results, we added larger noise compared to applications 1 and 2 by further decreasing the value of ϵ to 0.01 to test the utility of PPML-Omics. From the visualization as shown in **Figure 4b**, both our system (at the fifth row) and the centrally trained system (at the fourth row) obtained good results from a visual perspective compared to the ground truth (at the third row), effectively predicting the local expression of *FASN* and other cancer marker genes, including *COX6C, GNAS, HLA-B, HSP90AB1, PABPC1* and *TFF3* (**Supplementary Section 4.2, Supplementary Figures 6-10**) with the histopathology images of 5 patients (BT23269_C1, BT23277_E1, BT23377_C1, BT23901_C2 and BT23944_E1) as inputs, which could be supported by the sequencing data from spatial transcriptomics (ground truth). Furthermore, the local gene expressions predicted by PPML-Omics and the centrally trained system were both aligned well to the tumour region (black) and normal region (white) annotated in the annotation (at the second row), indicating that the ST-Net integrated with PPML-Omics still had the power to predict the local gene expression accurately from tissue images. In addition, with the same small privacy budget $\epsilon = 0.01$ and 0.1, the prediction of PPML-Omics was more accurate (MSE=0.910782 at $\epsilon = 0.01$ and MSE=0.909360 at $\epsilon = 0.1$) than the FL-DP-only system (MSE=0.926142 at $\epsilon = 0.01$ and MSE=0.910072 at $\epsilon = 0.1$) with $P = 1.20169 \times 10^{-5}$ and $P = 2.45510 \times 10^{-9}$ for one-sided Student's t-tests respectively and achieved similar performance of the centrally trained system (MSE=0.908258) (**Table 2 and Supplementary Table 13**). Thus, PPML-Omics achieved good utility on this task.

It is well known that medical imaging data contains much potentially sensitive information [14], such as tissue patterns and lesions, which could compromise patients' privacy. Several studies have shown that machine learning models trained based on medical images remember much information about the images in the training dataset, and attackers could perform image reconstruction attacks with the published machine learning models to obtain the original training images [62–64]. To see whether PPML-Omics protects privacy from the histopathology images compared to the centrally trained system, we used the improved deep leakage from gradients method (iDLG) in **Methods** [65] to simulate an attacker reconstructing the training images by stealing the gradients passed between machines under the FL framework (approximating a centrally trained system when the number of clients is 1). In each attack, we initialized a noisy input (dummy data), used the local gene expression as the ground truth, computed the prediction of the model on the noisy input and calculated the gradient. Then, we updated the noisy input with the gradient. In other words, same as in application 1, we were trying to optimize an input image that could get the most similar prediction results to the ground truth. We performed the iDLG attack separately for the centrally trained system and

PPML-Omics and showed the input image every 1000 iterations, as shown in **Figure 4c**. Strikingly, in terms of visual results, the iDLG attack could effectively reconstruct the training images with the centrally trained system, while in PPML-Omics, the iDLG attack was effectively blocked due to the addition of noise to the gradients with the DP mechanism. Overall, PPML-Omics protected privacy by blocking the reconstruction of sensitive histopathology images.

2.4 Theoretical proof of the privacy-preserving power of PPML-Omics

With the three applications above, we have experimentally demonstrated our system's privacy-preserving capability and utility. We further theoretically proved the privacy-preserving capability of PPML-Omics with the DP notation.

Typically, there are 3 models: (i) the central DP (CDP) model, (ii) the local DP (LDP) model and (iii) the shuffling DP model as shown in **Figure 5**. In the CDP model, a server trusted by users collects users' raw data (e.g. local updates) and executes a private mechanism for deferentially private outputs. The privacy preserving goal is to achieve indistinguishability for any outputs w.r.t. two neighbouring datasets that differ by replacing one user's data. The definition of differential privacy (**Definition 4 in Supplementary Section 1.2**) requires that the contribution of an individual to a dataset has not much effect on what the adversary sees. Compared to CDP, LDP is a stronger notion of privacy. In LDP, each user's data is required to be perturbed to protect privacy before collection. The CDP model assumes the availability of a trusted analyzer to collect raw data, while LDP model does not rely on any trusted party because users send randomized data to the server. The CDP model protects the calculation result in the analyzer, so users need to trust the central server and send raw data to the server, which allows greater accuracy but requires a trusted analyzer which is impractical in most real cases, as shown in **Figure 5a**. In the LDP model, we protect the data information in the single local device, so users only need to trust their single device and randomize local data before sending them to the analyzer, thus a trusted analyzer is not required, but the accuracy is limited because we do lots of randomness on local data, as shown in **Figure 5c**. Therefore, the advantage of using the shuffling DP model (**Definition 6 in Supplementary Section 1.4**) in FL is that we could balance the strength in both CDP and LDP, i.e., good performance of accuracy in the central DP model and strong privacy in the LDP model without relying on any trusted central party (**Figure 5b**).

Under the FL, each client learns locally and sends the model weight updates to a central server that aggregates those updates into the central model. After updating the central model, the central server then broadcasts the new weight to all clients for updating (**Algorithm 1**). With the integration of DP in the FL framework, we manually chose several ϵ as discussed in **Methods** and set $\delta = 1e - 5$ by default. Given ϵ and δ, we applied the analytic Gaussian

mechanism to calculate the optimal σ for perturbation [66] as in algorithm 2. In other words, we could calculate the amount of noise that needs to be added based on the given ϵ and δ. The privacy guarantee of analytic Gaussian mechanism was given in (**Theorem 1 in Supplementary Section 1.4**). Thus, we applied the analytic Gaussian mechanism on a single client with the privacy guarantee proved to realize the DP mechanism.

Based on the post-processing property (**Lemma 1 in Supplementary Section 1.4**), the protocol $\mathcal{P}$ (defined in **Definition 6 in Supplementary Section 1.4**) achieves the same privacy level as $\mathcal{M}$ (defined in **Lemma 1 in Supplementary Section 1.4**) because $\mathcal{A}$ is executed by an untrusted analyzer, without protecting users' privacy. We wanted to obtain $\mathcal{M} = \mathcal{S} \circ \mathcal{R}^n$ that could satisfy (ϵ_c, δ_c)-DP, meaning we achieved the same privacy guarantee as the CDP. Thus, we focused on analyzing the indistinguishability for $\mathcal{M}(X)$ and $\mathcal{M}(X')$ where X and X' differ in one client's local vector such that we achieved privacy-preserving on client-level. [67] proved that the privacy of $\mathcal{M}$ could be amplified. In other words, when each user applies the local privacy budget ϵ_l in $\mathcal{R}$, $\mathcal{M}$ can achieve a stronger privacy of (ϵ_c, δ_c)-DP with $\epsilon_c < \epsilon_l$. Hence, the shuffling model has a larger privacy budget in local single client and needs less noise to achieve the same privacy model compared with the LDP model. In practice, we defined a privacy budget ϵ_c in a single epoch, then we amplified this budget to the local client model with the local privacy budget ϵ_l.

Finally, based on **Theorem 2 in Supplementary Section 1.4** and **Theorem 3 in Supplementary Section 1.4**, we showed that PPML-Omics owned a better theoretical privacy guarantee.

3 Discussion

Our PPML-Omics system is universal, meaning that PPML-Omics could be integrated with any machine learning model and applied to various biological problems. Thus, we applied PPML-Omics to three different machine learning models from simple to complex, including FCN, Anto-encorder, and DenseNet-121. Applying PPML-Omics to more complex machine learning models means that more computational resources are required to add noise, which may introduce more uncertainty in the performance of the models. Hence, whether PPML-Omics can be applied to more complex machine learning models and used to protect privacy in more complex data requires further research in the future.

It is commonly acknowledged that privacy protection and utility are two conflicting objectives. Thus it is challenging to find a balance that allows the model to protect privacy without overly damaging the usability of the data. Therefore, PPML-Omics balances the privacy-preserving capabilities with utility as much as possible, while also leaving some of the decision making to our potential users. Depending on the user's actual needs, the user can adjust the privacy budget (ϵ) in the system during the training phase to achieve different levels of privacy protection.

In summary, we have described a secure and privacy-preserving machine learning (PPML) system by combining federated learning, differential privacy and shuffling mechanism. For the first time, we have applied this system to analyze data from three representative omic data analysis tasks, which are solved with three different deep learning models, revisited and addressed the privacy concern in the cancer classification from TCGA with bulk RNA-seq, clustering with scRNA-seq, and the integration of spatial gene expression and tumour morphology with spatial transcriptomics. Moreover, we examined in depth the privacy breaches that existed in all three tasks through privacy attack experiments and demonstrated that patients' privacy could be protected by our PPML-Omics system. Additionally, we proved the privacy-preserving capability of PPML-Omics theoretically, suggesting the first mathematically guaranteed method with robust and generalizable empirical performance in protecting patients' privacy from omic data. We believe that PPML-Omics will raise the concern of future researchers on biological issues regarding data privacy, and also try to protect data privacy by applying PPML-Omics in research. Our system's modularized and extendable nature has great potential to be developed collaboratively for different biological tasks and deep learning models, and will shed light on the application and development of deep learning techniques in the privacy protection of biological data.

4 Methods

4.1 Dataset preparation

For application 1, we used the Cancer Genome Atlas (TCGA) dataset, accessible through the Genomic Data Commons Data Portal and built our systems across 37 cancers with a total number of 13,057 patients. Each patient had the expression vector of 20,531 genes, and each expression vector was re-scaled in logarithm with 10 according to (1), to speed up learning and lead to faster convergence. Also, the normalization procedure ensured that the gene expression contributing to the model was on an equal scale. Patients were randomly divided into a training dataset (80%) and test dataset (20%) and the random split was repeated 5 times (**Supplementary Section 2.1**).

$$X_{input} = log_{10}(X_{original}) \tag{1}$$

For application 2, to assess the representation from our privacy preserved model on the scRNA-seq data, we tested our model on 4 published datasets with different samples sizes, all of which have expert-annotated labels from scDHA[56]. The Yan[68] dataset refers to the human preimplantation embryos and embryonic stem cells. In this dataset, 90 cells were sequenced with the Tang protocol. We first log-transformed the gene expression values selected for the highly variable genes. Finally, we scaled the dataset to unit variance and zero mean. The Pollen[69] dataset was sequenced with the SMARTer protocol. It contains 301 cells in the developing cerebral cortex from 11 populations. We downloaded it from the Hemberg Group's website (https://hemberg-lab.

github.io/scRNA.seq.datasets/human/tissues/pollen) and removed the low-quality data and cells with more mitochondrial genes and spike-in RNA. Then we selected highly variable genes after logarithmic transformation. The Camp Liver[70] dataset contains 777 cells from human liver tissues under the SMARTer protocol. The 3-D LBs dataset is benchmarked against fetal and adult human liver scRNA-seq data. We also selected the high variable genes before performing the logarithmic transformation. The Hrvatin[71] dataset contains 48,266 cells from 6-8 week old mice, which were sequenced by DropSeq. After filtering the low-quality data, we performed the aforementioned logarithmic transformation, normalization per cell count, and highly variable gene selection steps. Also, to further validate the privacy-preserving ability of PPML-Omics, we selected the scRNA-seq data on immune and stromal populations from colorectal cancer patients[61]. We visualized the global and local clusters regarding their specific macrophage and conventional dendritic cell (cDC) subsets. The single-cell data were selected from patients with informed mechanisms of Myeloid-Targeted therapies in colon cancer, which can be found in Data Availability. After quality control, each patient has around 5,000 cells and corresponding 13,538 genes (**Supplementary Section 3.1**).

For application 3, to test the capability of our method in spatial transcriptomics, we used a spatial transcriptomics dataset of 30,612 spots, including 23 patients with breast cancer with four subtypes: Luminal A, Luminal B, Triple-negative, and HER2-positive [49]. There were 3 high-resolution images of H&E staining tissue and their corresponding gene expression data for each patient. The number of spots for each replicate ranged from 256 to 712 depending on the tissue size and the diameter of a single spot was 100 μm arranged in a grid with a 200 μm as the centre-to-centre distance. There were 26,949 mRNA species detected across the dataset and each spot was represented as a 26,949-dimensional vector containing the elements denoted by the number of gene counts (**Supplementary Section 4.1**).

4.2 Deep learning model design and training

All codes were implemented using the PyTorch library [72] and processed on one machine with 2 NVIDIA V100 GPUs and 252G RAM. All experiments were repeated on the pre-split training and testing batches until fully converged on the testing dataset, and all reported performance was averaged over five random splits.

For application 1, we applied a fully connected neural network with 3 hidden layers as our benchmark machine learning network for PPML-Omics. The gene expression vector was fed into the network, and a ReLU layer was applied after each hidden layer to provide non-linearity. Finally, a Soft-max layer was attached to the last hidden layer to get the final prediction. With PPML-Omics, the network was trained in a federated, secure and privacy-preserving procedure to guarantee the utility while preserving privacy at the same time. The multi-class cross-entropy loss defined the loss function:

$$L = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic} log(p_{ic}) \tag{2}$$

where M represents the number of cancers (classes), which is 37 for this task. Correspondingly, $y_{ic} \in \{0, 1\}$ represents the classification result for the i-th sample, with 1 for a positive prediction, and 0 for a negative prediction. p_{ic} stands for the probability of the i-th sample to be in class c.

For application 2, the unsupervised framework for scRNA-seq clustering is composed of two steps. We first trained the Auto-encoder with the fixed number of gene expression value y as the input and outputed $\hat{y}$ with the same size of the input. We denoted d as the dimension of each patient's gene expression vector and N as the number of clients. We then optimized the Auto-encoder with the Mean squared error (MSE) loss:

$$L(y, \hat{y}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{d}(y - \hat{y})^2 \tag{3}$$

Then we extracted the central feature vector and clustered with k-means clustering. Specifically, in our PPML-Omics framework, the central shuffler should not know the data source of the gene expression value during the shuffled federated training procedure, thus preserving the client-level privacy of the scRNA-seq data source.

For application 3, we used ST-Net [49] as our baseline deep learning network and integrated it into PPML-Omics. In the ST-Net, we used DenseNet-121 to detect the fine-grained spatial heterogeneity within the tumour tissue. Small patches (224x224 pixels) were extracted from the whole-slide images (˜10000x10000 pixels) centred on the spatial transcriptomics spots, and each patch went through the pre-trained DenseNet-121 network for training and predicting the pre-selected 250 genes with the highest mean expression.

4.3 Performance assessment

To assess the performance of models in application 1, we adopted metrics including accuracy and macro-F1, which are defined as below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision_{macro} = \frac{1}{n}\sum_{i=1}^{n} Precision_i \tag{5}$$

$$Recall_{macro} = \frac{1}{n}\sum_{i=1}^{n} Recall_i \tag{6}$$

$$F1_{macro} = \frac{2 * Precision_{macro} * Recall_{macro}}{Precision_{macro} + Recall_{macro}} \tag{7}$$

where TP, TN, FP and FN stand for true positive, true negative, false positive and false negative, y_i is the predicted value, $\tilde{y}_i$ is the real value, and n is the total number of samples.

To evaluate PPML-Omics on application 2, we applied clustering Adjusted Rand Index (ARI) score, Normalized Mutual Information (NMI), Cluster Accuracy (CA), and Jaccard index (JI) to evaluate our method's utility comprehensively. The evaluation metrics between the predicted cluster B with the ground truth cluster A in ARI, NMI, CA, and JI are defined below:

$$ARI = \frac{\sum_{ij}\binom{n_ij}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}} \tag{8}$$

$$NMI(A,B) = \frac{2 \times I(A;\ B)}{H(Y) + H(C)} \tag{9}$$

$$CA(A,B) = \underset{perm \in P}{max} \frac{1}{n}\sum_{i=1}^{n-1}(perm(B_i) = A_i) \tag{10}$$

$$JI(A,B) = \frac{A \cap B}{A \cup B} \tag{11}$$

where A is the class labels, B is the cluster labels, $H(\cdot)$ is the Entropy and $I(A;\ B)$ is the Mutual Information between A and B. In the definition of Cluster Accuracy (CA): P stands for the set of all permutations in [1:K] where K is the number of clusters, and n is the sample size.

To measure the similarity between the predicted gene expression and the ground truth in application 3, we used the mean squared error (MSE) as the evaluation metric:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{12}$$

4.4 Implementation of federated learning framework

As shown in **Algorithm 1**, the federated learning framework aimed at ensuring data privacy and security along with the improvement of the AI model based on joint data sources from multiple clients around the world. We first initiated multiple data sources as our clients. Since the essence of federated learning is the union of samples, each client first needs to download the model from the server and initiate their client model with the server weight. Then each participant can use local data to train the model, encrypt their gradient, and upload it to the server. The server needs to aggregate the gradient of each client to update the server model parameters. In this process, each client is treated with the same and complete model, and there is no communication and no dependence between among clients. Therefore, each client can also make independent predictions during prediction.

Algorithm 1 FL

Require: K is the number of clients with data source $D_1, D_2, ..., D_K$; T is the number of epochs; η is learning rate; C is the parameter for l2-norm clipping;

procedure SERVER EXECUTION
  Initialize: w_0
  for $t = 1, \ldots, T$ **do**
    for each client $k : 1, \ldots, K$ in parallel **do**
      $\Delta w_{t+1}^k \leftarrow \text{ClientUpdate}(w_t, D_k)$
      $\Delta \tilde{w}_{t+1}^k \leftarrow \Delta w_{t+1}^k / \max(1, \frac{\|\Delta w_{t+1}^k\|_2}{C})$
    end for
    $w_{t+1} \leftarrow w_t + \eta(\frac{1}{K} \sum_{k=1}^{K} \Delta \tilde{w}_{t+1})$
  end for
end procedure

function CLIENTUPDATE(w_t, D_k)
  $w \leftarrow w_t$
  for each patient $p \in D_k$ **do**
    $w \leftarrow w - \eta \nabla \ell(w, p)$
  end for
  $\Delta w_{t+1} = (w - w_t)/\eta$
  return Δw_{t+1}
end function

4.5 Implementation of deferentially private model training

To add noise to the gradient, we used the calibrate analytic Gaussian mechanism to calculate the value of σ in the Gaussian distribution based on the values of ϵ and δ. In each epoch, after computing the gradient update for each client, we clipped the gradient so as to ensure a finite upper and lower bound and then added noise to each value in the gradient that fitted the previously defined Gaussian distribution as shown in **Algorithm 2**.

4.6 Implementation of shuffling model

In a practical scenario, before the end of each epoch, each client sends the gradient information to the server, while the server knows where each gradient information comes from, providing the possibility of privacy leakage. Therefore, the shuffling mechanism sits between clients and the server to disrupt the gradient information submitted by clients and eliminate the ID information carried in the updates. As shown in **Algorithm 2**, we collected updates from all clients in each epoch and then obtained a shuffled update after a random shuffling mechanism with mutable random seeds, and finally updated the server model with the shuffled updates.

Algorithm 2 PPML-Omics

Require: K is the number of clients with data source $D_1, D_2, ..., D_K$; T is the number of epochs; C is the parameter for l2-norm clipping; η is learning rate; standard deviation σ of Gaussian noise needed to achieve (ϵ,δ)-DP for a single iteration, where $\sigma = CalibrateAnalyticGaussianMechanism(\epsilon, \delta)$ by using the analytic Gaussian mechanism of [66], random shuffler S.

procedure SERVER EXECUTION
  Initialize: w_0
  for $t = 1, \ldots, T$ **do**
    for each client $k : 1, \ldots, K$ in parallel **do**
      $\Delta w_{t+1}^k \leftarrow \text{ClientUpdate}(w_t, D_k)$
      $\Delta \tilde{w}_{t+1}^k \leftarrow \Delta w_{t+1}^k / \max(1, \frac{\|\Delta w_{t+1}^k\|_2}{C}) + \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$
    end for
    Collect shuffled responses: $(\Delta \tilde{w}_{t+1}^{S(k)})_{k \in [K]}$
    $w_{t+1} \leftarrow w_t + \eta(\frac{1}{K} \sum_{k=1}^{K} \Delta \tilde{w}_{t+1}^{S(k)})$
  end for
end procedure

function CLIENTUPDATE(w_t, D_k)
  $w \leftarrow w_t$
  for each patient $p \in D_k$ **do**
    $w \leftarrow w - \eta \nabla \ell(w, p)$
  end for
  $\Delta w_{t+1} = (w - w_t)/\eta$
  return Δw_{t+1}
end function

4.7 σ selection with analytic Gaussian DP mechanism

For a given ϵ, we calculated the value of σ corresponding to the addition of Gaussian noise by using the analytic Gaussian DP mechanism [66].

Theorem 1 *Let $f : \mathcal{X}^n \to \mathbb{R}^k$ be a function with global ℓ_2-norm sensitivity Δ. Suppose $\epsilon > 0$ and $0 < \delta < \frac{1}{2} - e^{-3\epsilon/\sqrt{4\pi\epsilon}}$. If the Gaussian mechanism $A(D) = f(D) + Z$ with $Z \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_k)$ is (ϵ, δ)-DP then $\sigma \geq \frac{\Delta}{\sqrt{2\epsilon}}$.*

Proof When $\sigma = \frac{\Delta}{\sqrt{2\epsilon}}$, the mechanism will be (ϵ, δ)-DP with $\delta = \Phi(0) - e^{\epsilon}\Phi(-\sqrt{2\epsilon}) > \frac{1}{2} - e^{-3\epsilon/\sqrt{4\pi\epsilon}}$.Thus, it is impossible to achieve (ϵ, δ)-DP with $\delta = \Phi(0) - e^{\epsilon}\Phi(-\sqrt{2\epsilon})$ without increasing the variance of perturbation. □

We could use the above theorem to get the minimal noise to add for the given ϵ.

4.8 Hyperparameter selection in DP mechanism

Regarding the parameters in the DP mechanism, we selected the proper value **C** for l_2-norm clipping based on the distribution of gradients in the centrally trained model. The selection of the privacy budget ϵ was a tricky task, as commonly adopted, such as in the handbook 'Disclosure Avoidance for the 2020 Census: An Introduction' [73], we selected ϵ from 0.01, 0.1, 1, 5, 10, 15, where $\epsilon = 5$ was commonly used in practice. Since all 3 applications could converge in less than 20 epochs, we could ensure that the (ϵ, δ)-DP after the T epochs is acceptable using the composition theorem.

4.9 Hyperparameter and model selection in deep learning

To ensure that all models had a fair chance of learning a useful representation in all tasks, we trained multiple instances of each model (FCN for application 1, Auto-encoder for application 2 and ST-Net for application 3) using grid search of hyperparameter settings, including learning rate, epochs, number of hidden layers, number of filters, filter size, batch size, momentum, initial weight, weight decay and keep probability. Then, we selected model instances based on their training performance.

4.10 Model inversion attack for cancer classification models

ML model was abused to learn sensitive genomic information about individuals, which was shown in a case study of linear classifiers in personalized medicine by Fredrikson et al. [74]. Thus, we implemented the model inversion attack (MIA) proposed by Fredrikson et al. [75] to extract sensitive information from the trained models with our PPML-Omics system under different privacy budgets to address the privacy concerns. Regarding our tasks, we retreated the MIA as an optimization problem: to find the input that maximizes the returned class confidence score [75], which was achieved by using gradient descent along with modifications specific to this domain as shown in Algorithm 3 (**Supplementary Section 2.3**).

We reconstructed the original gene expression with the MIA for each cancer for all pre-trained models. For each cancer, we applied MIA to reconstruct the gene expression vector for the targeting cancer and selected genes with the highest reconstructed expression levels ($\geq$0.8) as the most significant reconstructed gene features specific to each cancer. Then we split all samples of 37 cancers into 2 groups (one only included the samples of the targeting cancer type and the other one included all remaining samples from other cancer types) and plotted the distribution of the z-score normalized expression of the previously selected genes of the real expression data in both groups (solid lines for samples of the target cancer type and dashed lines for other cancer types).

Algorithm 3 Model Inversion Attack (MIA)

Require: Pre-trained model M, target label $\mathbf{t}$, aim input to optimize $\mathbf{i}$, number of epochs α, learning rate η, early stopping parameter β and cost threshold γ.

```
Initialize i with 0
for epoch in {0...α} do
    Get prediction o = M(i)
    Calculate cost c = LossFunction(o, t)
    Calculate gradient Δ on c
    Update i = i − η ∗ Δ
    if c keeps increasing for β epochs then
        Halt
    end if
    if c ≤ γ then
        Halt
    end if
end for
```

4.11 iDLG reconstruction attack

Sharing gradient during the training of machine learning networks could potentially leak private data. Zhao et al. presented an approach named improved deep leakage from gradient (iDLG) as shown in **Algorithm** 4, which showed the possibility to obtain private training data from the publicly shared gradients. In integrating tumour morphology and gene expression with spatial transcriptomics, we synthesized the dummy data and corresponding labels with the supervision of shared gradients and optimized the dummy data with iDLG to obtain the one similar to the private training data.

Algorithm 4 iDLG

Require: Differentiable model $\mathbf{M}$, model parameters $\mathbf{W}$, private training data and labels $(\mathbf{x}, \mathbf{c})$, gradients produced by the private data $\nabla\mathbf{W}$,dummy data and labels $(\mathbf{x}', \mathbf{c}')$, number of iterations N, learning rate η and loss function l

```
Extract the target ground-truth label to initialize the dummy label c'
Initialize the dummy data x' ← N(0, 1)
for i ← 1 to N do
    Calculate the dummy gradients: ∇W ← ∂l(M(x', W), c')/∂W
    Calculate the loss: L_G = ‖∇W' − ∇W‖²_F
    Update the dummy data: x' ← x' − η∇_{x'} L_G
end for
```

4.12 Quantification of privacy leakage

For application 1, to investigate the significance of the results of gene reconstruction attacks, we used Kullback–Leibler (KL) divergence, Jensen-Shannon (JS) divergence and Kolmogorov–Smirnov (KS) test as shown below to characterize the degree of privacy leakage by calculating the divergence between two distributions of the expression level of the reconstructed genes in the target cancer type and other cancer types. Larger JS divergence values and smaller p-values of the KS test indicated more significant differences between the expression levels of the reconstructed signature genes in the target cancer type and other cancer types, suggesting a severe privacy leakage.

$$KL(P\|Q) = \sum p(x) log \frac{P(x)}{Q(z)} \tag{13}$$

$$JS(P\|Q) = \frac{1}{2} KL(P(x)\|\frac{P(x)+Q(x)}{2}) + \frac{1}{2} KL(Q(x)\|\frac{P(x)+Q(x)}{2}) \tag{14}$$

where P(x) and Q(x) are two probability distributions.

For application 2, given a system, if we cannot observe some targeted small clusters (sub-types) in the final clustering and the proportion results, then we could conclude that the system protects the patient's privacy. Additionally, if the clustering results for major cell types are correct, we could conclude that the system preserves a reasonable degree of usability while protecting privacy.

PPML-Omics avoids reconstruction attacks on medical images. Thus we evaluated the privacy-preserving capability by visually comparing the reconstructed image and the raw image after the iDLG attack on different systems in application 3.

5 Data availability

Data for the cancer classification with bulk RNA-seq is based upon the one generated by the TCGA Research Network: https://www.cancer.gov/tcga. The datasets for scRNA-seq clustering analysis are available at GEO accession ID: Yan (GSE36552), Pollen (SRP041736), Camp Liver (GSE81252), and Hrvatin (GSE102827). The rds files are also available on the SCDHA software availability page (https://bioinformatics.cse.unr.edu/software/scDHA/resource/Reproducibility/Data/). scRNA-seq data of patients are are available at GEO accession ID: GSE146771. For the spatial transcriptomics dataset, there are 30,612 spots, including 23 patients with breast cancer with four subtypes: Luminal A, Luminal B, Triple-negative and HER2-positive. There were 3 high-resolution images of H&E staining tissue and their corresponding gene expression data for each patient. All images and processed data are available at http://www.spatialtranscriptomicsresearch.org.

6 Code availability

Code for this paper is available at https://github.com/JoshuaChou2018/PPML-Omics.

7 Acknowledgement

We acknowledge funding from KAUST and funders played no role in the design of the study, the preparation of the manuscript or the decision to publish. We would also like to thank Lyu's article [76] for the reference of partial codes.

Juexiao Zhou, Siyuan Chen, Haoyang Li, Bin Zhan, Longxi Zhou, Zhongxiao Li, Ningning Chen, Wenkai Han and Xin Gao were supported in part by grants from Office of Research Administration (ORA) at King Abdullah University of Science and Technology (KAUST) under award number BAS/1/1624-01-01, FCC/1/1976-04-01, URF/1/4098-01-01, REI/1/0018-01-01, REI/1/4216-01-01, REI/1/4437-01-01, REI/1/4473-01-01, URF/1/4352-01-01, REI/1/4742-01-01, and URF/1/4663-01-01.

Yulian Wu, Yan Hu, Zihang Xiang and Di Wang were supported in part by the baseline funding BAS/1/1689-01-01 and funding from the AI Initiative REI/1/4811-10-01 of King Abdullah University of Science and Technology (KAUST).

8 Contributions

J.Z. and S.C. contributed the original idea, model design, experimental analysis and wrote the manuscript. Y.W contributed the theoretical proof and wrote the manuscript. The review and editing of the manuscript and some supplementary visualizations were carried out by H.L., B.Z., L.X., Y.H., Z.X., Z.L., N.C. and W.H. The entire project was supervised by X.G. and D.W..

9 Competing interests

The authors declare no competing interests.

References

[1] Joly, Y., Dyke, S. O. M., Knoppers, B. M. & Pastinen, T. Are Data Sharing and Privacy Protection Mutually Exclusive? *Cell* **167**, 1150–1154 (2016) .

[2] Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology* **19** (1A), A68 (2015) .

[3] Caulfield, M. *et al.* The 100,000 genomes project protocol. *Genomics Engl. P* **3**, 1–112 (2017) .

[4] Lewin, H. A. *et al.* Earth biogenome project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences* **115** (17), 4325–4333 (2018) .

[5] Alser, M. *et al.* Gatekeeper: a new hardware architecture for accelerating pre-alignment in dna short read mapping. *Bioinformatics* **33** (21), 3355–3363 (2017) .

[6] Joly, Y., Feze, I. N., Song, L. & Knoppers, B. M. Comparative approaches to genetic discrimination: chasing shadows? *Trends in Genetics* **33** (5), 299–302 (2017) .

[7] McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics* **9** (5), 356–369 (2008) .

[8] Homer, N. *et al.* Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics* **4** (8), e1000167 (2008) .

[9] Almadhoun, N., Ayday, E. & Ulusoy, Ö. Differential privacy under dependent tuples—the case of genomic privacy. *Bioinformatics* **36** (6), 1696–1703 (2020) .

[10] Wang, Z., Gerstein, M. & Snyder, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* **10** (1), 57–63 (2009) .

[11] Tang, F. *et al.* mrna-seq whole-transcriptome analysis of a single cell. *Nature methods* **6** (5), 377–382 (2009) .

[12] Marx, V. Method of the year: spatially resolved transcriptomics. *Nature Methods* **18** (1), 9–14 (2021) .

[13] Zou, J. *et al.* A primer on deep learning in genomics. *Nature genetics* **51** (1), 12–18 (2019) .

[14] Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* **2** (6), 305–311 (2020) .

[15] Esteva, A. *et al.* A guide to deep learning in healthcare. *Nature medicine* **25** (1), 24–29 (2019) .

[16] Al-Rubaie, M. & Chang, J. M. Privacy-preserving machine learning: Threats and solutions. *IEEE Security & Privacy* **17** (2), 49–58 (2019) .

[17] Sheller, M. J. *et al.* Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports* **10** (1), 1–12 (2020) .

[18] Schwarz, C. G. *et al.* Identification of anonymous mri research participants with face-recognition software. *New England Journal of Medicine* **381** (17), 1684–1686 (2019) .

[19] Harmanci, A. & Gerstein, M. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature methods* **13** (3), 251–256 (2016) .

[20] McMahan, B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. *Communication-efficient learning of deep networks from decentralized data*, 1273–1282 (PMLR, 2017).

[21] Hitaj, B., Ateniese, G. & Perez-Cruz, F. *Deep models under the gan: information leakage from collaborative deep learning*, 603–618 (2017).

[22] Melis, L., Song, C., De Cristofaro, E. & Shmatikov, V. *Exploiting unintended feature leakage in collaborative learning*, 691–706 (IEEE, 2019).

[23] Nasr, M., Shokri, R. & Houmansadr, A. *Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning*, 739–753 (IEEE, 2019).

[24] Zhu, L., Liu, Z. & Han, S. Deep leakage from gradients. *Advances in Neural Information Processing Systems* **32** (2019) .

[25] Tolpegin, V., Truex, S., Gursoy, M. E. & Liu, L. *Data poisoning attacks against federated learning systems*, 480–501 (Springer, 2020).

[26] Rahman, M. A., Rahman, T., Laganière, R., Mohammed, N. & Wang, Y. Membership inference attack against differentially private deep learning model. *Trans. Data Priv.* **11** (1), 61–79 (2018) .

[27] Lu, H., Liu, C., He, T., Wang, S. & Chan, K. S. Sharing models or coresets: A study based on membership inference attack. *arXiv preprint arXiv:2007.02977* (2020) .

[28] Salem, A. *et al.* Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246* (2018) .

[29] Hu, H., Salcic, Z., Sun, L., Dobbie, G. & Zhang, X. Source inference attacks in federated learning. *arXiv preprint arXiv:2109.05659* (2021) .

[30] Geiping, J., Bauermeister, H., Dröge, H. & Moeller, M. Inverting gradients – how easy is it to break privacy in federated learning? (2020). 2003.14053.

[31] La, H. J., Kim, M. K. & Kim, S. D. *A personal healthcare system with inference-as-a-service*, 249–255 (IEEE, 2015).

[32] Dwork, C. *Differential privacy: A survey of results*, 1–19 (Springer, 2008).

[33] Wei, K. *et al.* Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security* **15**, 3454–3469 (2020) .

[34] Rodríguez-Barroso, N. *et al.* Federated learning and differential privacy: Software tools analysis, the sherpa. ai fl framework and methodological guidelines for preserving data privacy. *Information Fusion* **64**, 270–292 (2020) .

[35] Bonawitz, K. *et al.* *Practical secure aggregation for privacy-preserving machine learning*, 1175–1191 (2017).

[36] Wagh, S., Gupta, D. & Chandran, N. Securenn: 3-party secure computation for neural network training. *Proc. Priv. Enhancing Technol.* **2019** (3), 26–49 (2019) .

[37] Sav, S. *et al.* Poseidon: privacy-preserving federated neural network learning. *arXiv preprint arXiv:2009.00349* (2020) .

[38] Constable, S. D., Tang, Y., Wang, S., Jiang, X. & Chapin, S. *Privacy-preserving gwas analysis on federated genomic datasets*, Vol. 15, 1–9 (BioMed Central, 2015).

[39] He, Z. *et al.* Achieving differential privacy of genomic data releasing via belief propagation. *Tsinghua Science and Technology* **23** (4), 389–395 (2018) .

[40] Cho, H., Wu, D. J. & Berger, B. Secure genome-wide association analysis using multiparty computation. *Nature biotechnology* **36** (6), 547–551

(2018) .

[41] Yilmaz, E., Ji, T., Ayday, E. & Li, P. Genomic data sharing under dependent local differential privacy. *arXiv preprint arXiv:2102.07357* (2021) .

[42] Gürsoy, G., Brannon, C. M., Navarro, F. C. & Gerstein, M. Fancy: fast estimation of privacy risk in functional genomics data. *Bioinformatics* **36** (21), 5145–5150 (2020) .

[43] Gürsoy, G. *et al.* Functional genomics data: privacy risk assessment and technological mitigation. *Nature Reviews Genetics* 1–14 (2021) .

[44] Froelicher, D. *et al.* Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. *Nature communications* **12** (1), 1–10 (2021) .

[45] Li, W. *et al.* *Privacy-preserving federated brain tumour segmentation*, 133–141 (Springer, 2019).

[46] Kaissis, G. *et al.* End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence* **3** (6), 473–484 (2021) .

[47] Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics* **14** (9), 618–630 (2013) .

[48] Burgess, D. J. Spatial transcriptomics coming of age. *Nature Reviews Genetics* **20** (6), 317–317 (2019) .

[49] He, B. *et al.* Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering* **4** (8), 827–834 (2020) .

[50] Li, T. & Li, N. *On the tradeoff between privacy and utility in data publishing*, 517–526 (2009).

[51] Feldman, V., Mironov, I., Talwar, K. & Thakurta, A. *Privacy amplification by iteration*, 521–532 (IEEE, 2018).

[52] Cristea, S. & Polyak, K. Dissecting the mammary gland one cell at a time. *Nature communications* **9** (1), 1–3 (2018) .

[53] González-Silva, L., Quevedo, L. & Varela, I. Tumor functional heterogeneity unraveled by scrna-seq technologies. *Trends in cancer* **6** (1), 13–19 (2020) .

[54] Zhang, Q. *et al.* Establishment and optimization of scrna-seq assay to find the mechanism of immune therapy against tumors. *Cell* **8**, 9 (2021) .

[55] Ding, S., Chen, X. & Shen, K. Single-cell rna sequencing in breast cancer: Understanding tumor heterogeneity and paving roads to individualized therapy. *Cancer Communications* **40** (8), 329–344 (2020) .

[56] Tran, D. *et al.* Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nature communications* **12** (1), 1–10 (2021) .

[57] Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36** (5), 411–420 (2018) .

[58] Kiselev, V. Y. *et al.* Sc3: consensus clustering of single-cell rna-seq data. *Nature methods* **14** (5), 483–486 (2017) .

[59] Lin, P., Troup, M. & Ho, J. W. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome biology* **18** (1), 1–11 (2017) .

[60] Guo, M., Wang, H., Potter, S. S., Whitsett, J. A. & Xu, Y. Sincera: a pipeline for single-cell rna-seq profiling analysis. *PLoS computational biology* **11** (11), e1004575 (2015) .

[61] Zhang, L. *et al.* Single-cell analyses inform mechanisms of myeloid-targeted therapies in colon cancer. *Cell* **181** (2), 442–459 (2020) .

[62] Song, C., Ristenpart, T. & Shmatikov, V. *Machine learning models that remember too much*, 587–601 (2017).

[63] Raj, A., Bresler, Y. & Li, B. *Improving robustness of deep-learning-based image reconstruction*, 7932–7942 (PMLR, 2020).

[64] Cappelli, R., Maio, D., Lumini, A. & Maltoni, D. Fingerprint image reconstruction from standard templates. *IEEE transactions on pattern analysis and machine intelligence* **29** (9), 1489–1503 (2007) .

[65] Zhao, B., Mopuri, K. R. & Bilen, H. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610* (2020) .

[66] Balle, B. & Wang, Y.-X. *Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising*, 394–403 (PMLR, 2018).

[67] Erlingsson, Ú. *et al.* *Amplification by shuffling: From local to central differential privacy via anonymity*, 2468–2479 (SIAM, 2019).

[68] Yan, L. *et al.* Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology* **20** (9), 1131–1139 (2013) .

[69] Pollen, A. A. *et al.* Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology* **32** (10), 1053–1058 (2014) .

[70] Camp, J. G. *et al.* Multilineage communication regulates human liver bud development from pluripotency. *Nature* **546** (7659), 533–538 (2017) .

[71] Hrvatin, S. *et al.* Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nature neuroscience* **21** (1), 120–129 (2018) .

[72] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32**, 8026–8037 (2019) .

[73] Bureau, U. C. Disclosure avoidance for the 2020 census: An introduction (2021). URL https://www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.html.

[74] Fredrikson, M. *et al.* *Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing*, 17–32 (2014).

[75] Fredrikson, M., Jha, S. & Ristenpart, T. *Model inversion attacks that exploit confidence information and basic countermeasures*, CCS '15, 1322–1333 (Association for Computing Machinery, New York, NY, USA, 2015). URL https://doi.org/10.1145/2810103.2813677.

[76] Lyu, L. *Dp-signsgd: When efficiency meets privacy and robustness*, 3070–3074 (IEEE, 2021).

Figures and Tables

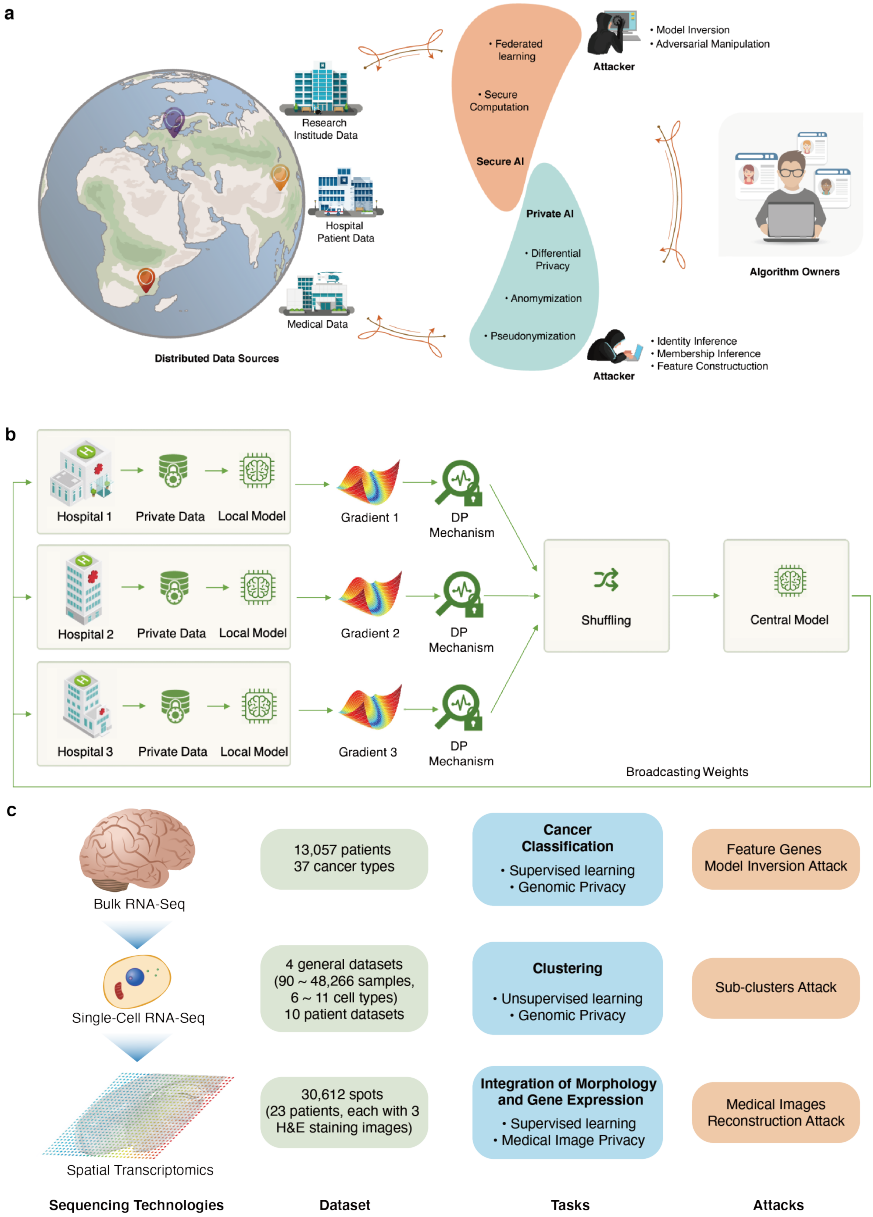
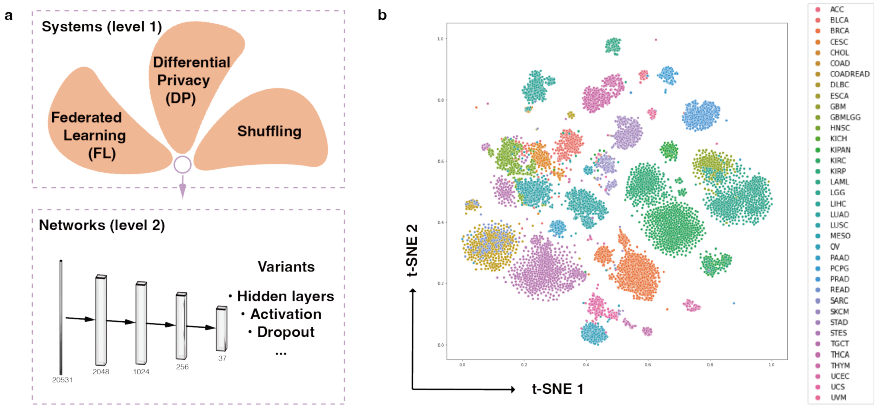
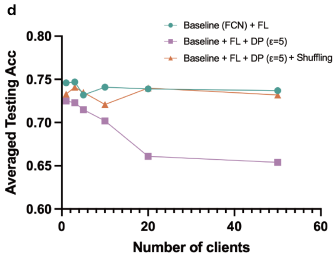


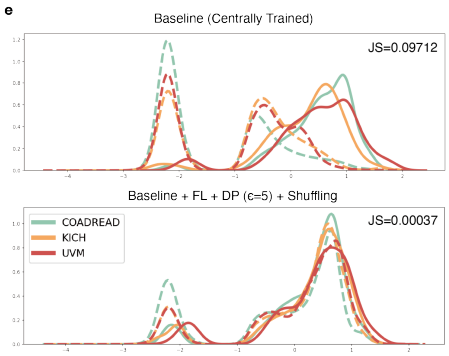
Fig. 1 PPML-Omics: a Privacy-Preserving federated Machine Learning system protects patients' privacy from omic data a, Schematic overview of the relationships and interactions between distributed data owners, algorithm owners, attackers and techniques in the field of secure and private AI. **b**, Schematic overview of PPML-Omics system design, in which the clients train the local model with local private data and the noised gradients with DP mechanism goes through the shuffling mechanism. Then all noised and shuffled updates are sent to the central server and used for updating the central model. **c**, Illustration of 3 representative tasks, datasets and attacks of omic data in this paper for demonstrating the utility and privacy-preserving capability of our PPML-Omics system, including the 1) cancer classification with bulk RNA-seq, 2) clustering with scRNA-seq and 3) integration of morphology and gene expression with spatial transcriptomics.

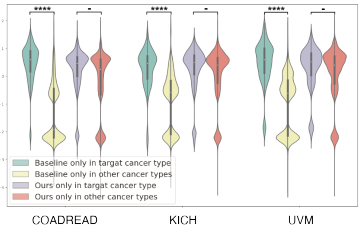


c

Architecture	Number of Hidden layers	Activation Function	Dropout	ϵ	Privacy-preserving Power (JS Divergence)
FCN (Centrally Trained)	3	/	/	/	0.13399
FCN (Centrally Trained)	3	Sigmoid	/	/	0.10466
FCN (Centrally Trained)	3	ReLU	/	/	0.09712
FCN (Centrally Trained)	3	ReLU	0.5	/	0.14202
FCN (Centrally Trained)	5	ReLU	/	/	0.08564
FCN + FL	3	/	/	/	0.16899
FCN + FL	3	Sigmoid	/	/	0.11389
FCN + FL	3	ReLU	/	/	0.10797
FCN + FL	3	ReLU	0.5	/	0.14303
FCN + FL	5	ReLU	/	/	0.10760
FCN + FL + DP	3	ReLU	/	1	0.00018
FCN + FL + DP	3	ReLU	/	5	0.00063
FCN + FL + DP	3	ReLU	/	10	0.00264
FCN + FL + DP	3	ReLU	/	15	0.00307
FCN + FL + DP + Shuffling	3	/	/	5	0.04276
FCN + FL + DP + Shuffling	3	Sigmoid	/	5	0.00092
FCN + FL + DP + Shuffling	3	ReLU	/	5	0.00037
FCN + FL + DP + Shuffling	3	ReLU	0.5	5	0.01548
FCN + FL + DP + Shuffling	5	ReLU	/	5	0.00966







g

Architecture	Testing Acc	Testing Macro-F1
Baseline (FCN, Centrally Trained)	0.744	0.757
Baseline + FL	0.723	0.755
Baseline + FL + DP (ϵ=5)	0.722	0.715
Baseline + FL + DP (ϵ=5) + Shuffling	0.748	0.750

Fig. 2 Results of cancer classification with bulk RNA-seq. a, Illustration of the relationship between Systems (level1) and Networks (level2). In level 1, we studied the effect of systems with or without federated learning (FL), differential privacy (DP) and shuffling mechanism. In level 2, we studied the effect of different machine learning networks. **b**, t-SNE plot on all patients' data from TCGA, each data point represents one patient and colours represent cancer types. **c**, Profiling analysis of different systems, including the centrally trained system, the FL system, the FL-DP system and PPML-Omics on FCN with a varying number of hidden layers, activation function, the drop out layer and values of ϵ. **d**, The effect of the number of clients on the classification task of 3 architectures. **e**, The distribution of z-score normalized expression of reconstructed signature genes in the target cancer type (solid lines) and other cancer types (dashed lines) by the baseline (centrally trained system) and PPML-Omics from three representative cancers (COADREAD: Colorectal adenocarcinoma, KICH: Kidney Chromophobe, UVM: Uveal Melanoma). **f**, The violin plot of the distribution of z-score normalized expression of reconstructed signature genes in the target cancer type and other cancer types specifically identified by the baseline only (centrally trained system) and PPML-Omics only. Test performed was two-sided Kolmogorov–Smirnov test, p-value annotation legend is the following: $^{****}P \leq 0.0001$, $^{-}P>0.1$. Exact P values are the following: Baseline only on COADREAD, $P = 1.513 \times 10^{-7}$; Ours only on COADREAD, $P = 0.85142$; Baseline only on KICH, $P = 3.423 \times 10^{-7}$; Ours only on KICH, $P = 0.77774$; Baseline only on UVM, $P = 1.430 \times 10^{-7}$; Ours only on UVM, $P = 0.52523$. **g**, The utility of different architectures.

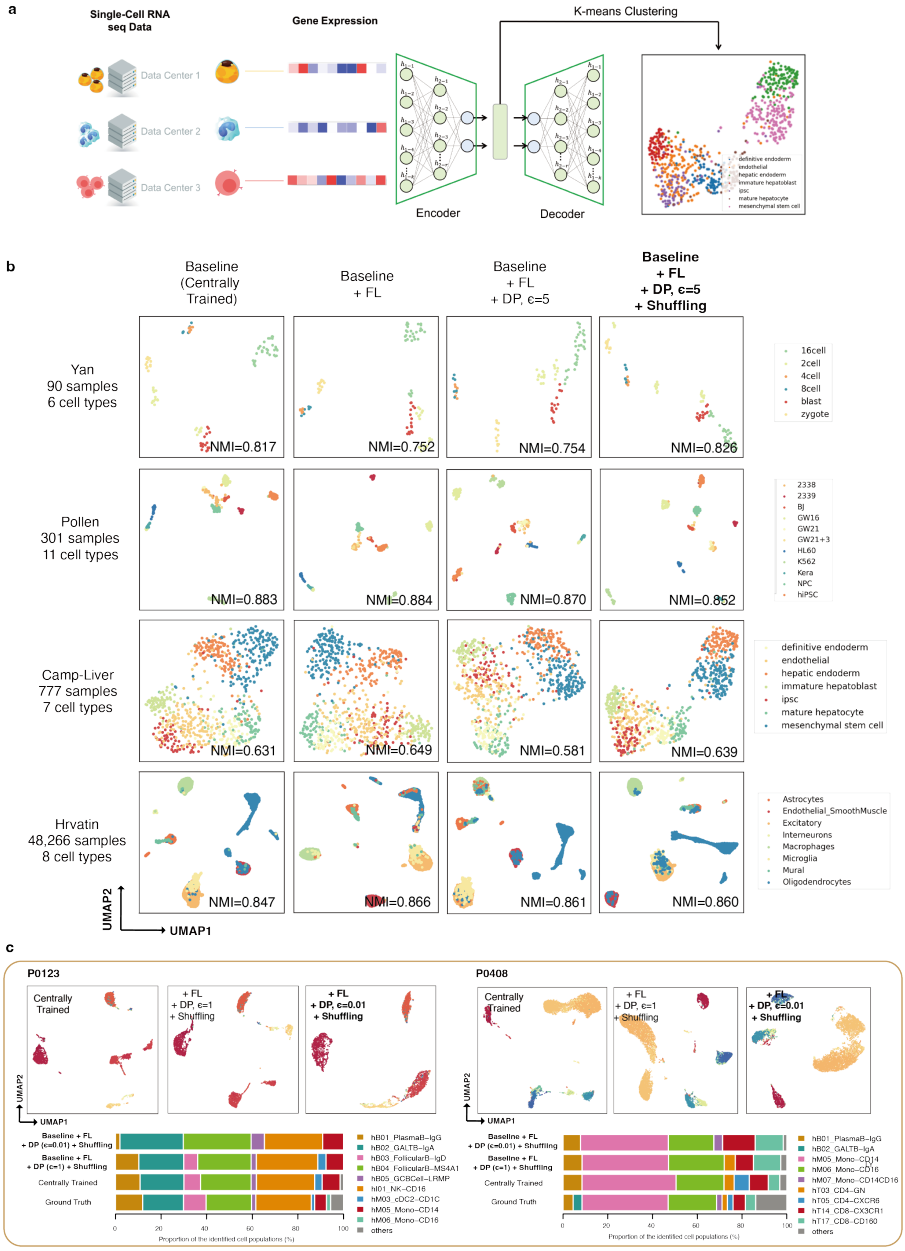


Fig. 3 Results of clustering with scRNA-seq. a, Architecture of the baseline machine learning network (Auto-encoder). Gene expression vectors were fed into the Auto-encorder for feature reduction and selection, after which the low dimensional features were used for the K-means clustering. **b**, The clustering results visualized by the UMAP of four datasets with an increasing number of samples (Yan with 90 samples from 6 cell types, Pollen with 301 samples from 11 cell types, Camp-Liver with 777 samples from 7 cell types and Hrvatin with 48,266 samples from 8 cell types). For these four datasets, all clustering results given by PPML-Omics showed a similar visual pattern with the centrally trained system, indicating the perfect utility of PPML-Omics. **c**, The distribution and proportion of major clusters and sub-clusters of the same patient clustered by different systems, including the centrally trained system, and our systems with $\epsilon = 1$ and $\epsilon = 0.01$, indicating that PPML-Omics could protect patients' privacy by removing the local information of sub-clusters.

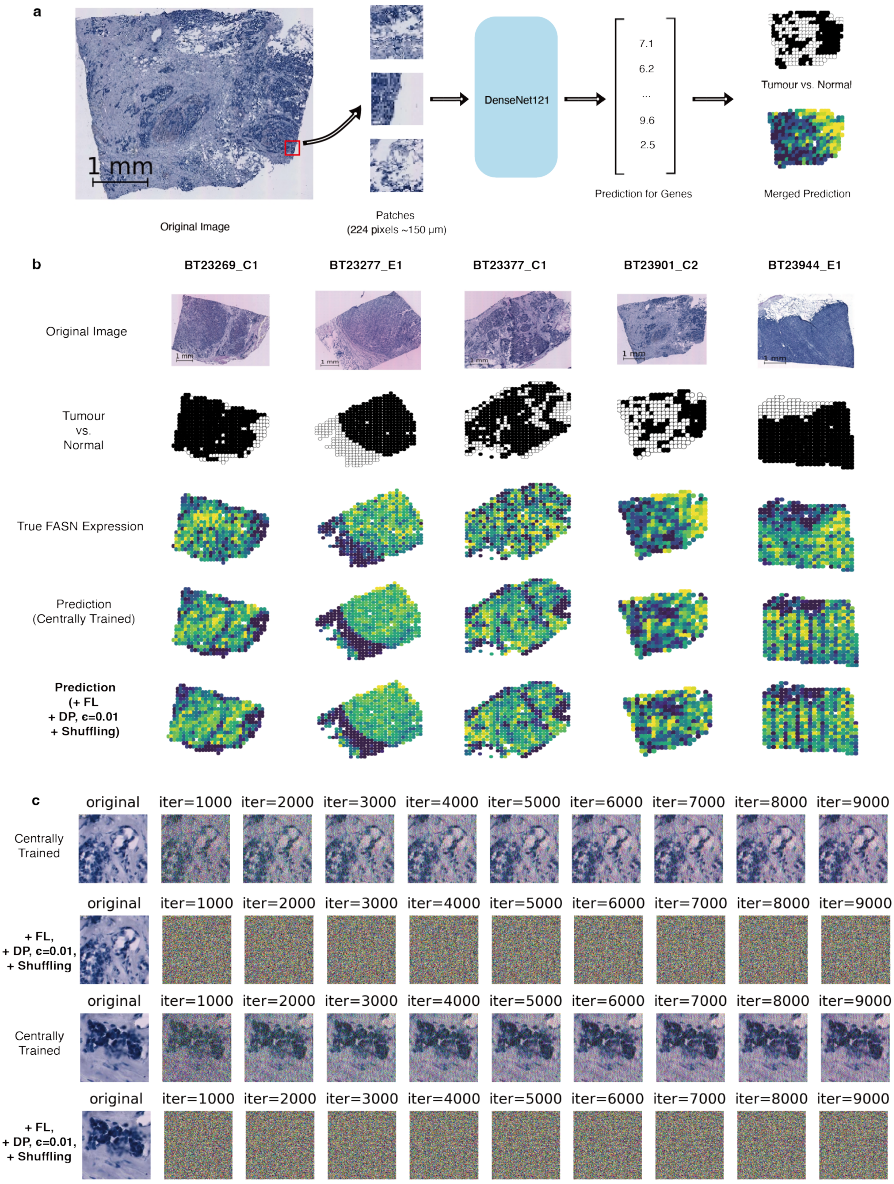


Fig. 4 Results of integration of tumour morphology and gene expression with spatial transcriptomics. a, Pipeline for predicting local gene expression from high-resolution tissue image referred by ST-Net. The patches (224x224) extracted from the original image were fed into the DenseNet121 and the local expression of selected cancer marker genes was predicted and compared with the real local expression (ground truth) from the spatial transcriptomics data. **b**, The results of five samples (BT23269_C1, BT23277_E1, BT23377_C1, BT23901_C2, BT23944_E1) showed the original image, the binary labels of the tumour (black) and normal regions (white), the real expression of *FASN* and predicted results by the centrally trained system and PPML-Omics (FL + DP with $\epsilon = 0.01$ + shuffling). All predictions show a similar visual pattern with the ground truth. **c**, Image reconstruction attack with improved deep leakage from gradients (iDLG) on the centrally trained system and PPML-Omics (FL + DP with $\epsilon = 0.01$ + shuffling).

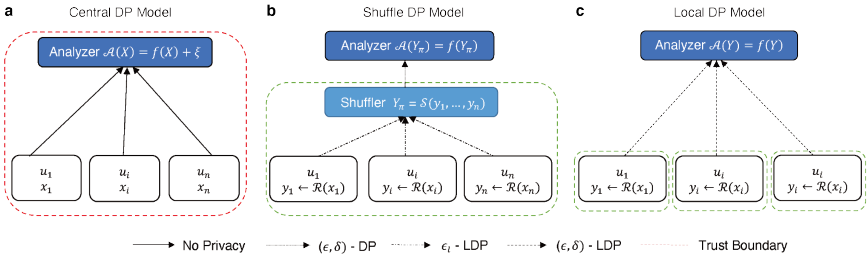


Fig. 5 Trust boundaries in the central, shuffling and local model of differential privacy. The central DP model relies on a trusted analyzer. The local DP only trusts a single local device. Moreover, the shuffling model needs to trust the local single device and the shuffler.

Table 1 Comparison of clustering result on multiple scRNA-seq datasets[56]

Dataset	Architecture	ARI	NMI	CA	JI
Yan (90 samples 6 cell types)	Baseline (Centrally Trained)	0.672	0.817	0.766	0.584
	Baseline + FL	0.628	0.752	0.722	0.536
	Baseline + FL + DP (ϵ=5, δ=10e-5)	0.609	0.754	0.711	0.519
	Baseline + FL + DP + shuffling (ϵ=5, δ=10e-5), PPML-Omics	0.685	0.826	0.777	0.598
Pollen (301 samples 11 cell types)	Baseline (Centrally Trained)	0.819	0.883	0.867	0.721
	Baseline + FL	0.807	0.884	0.837	0.705
	Baseline + FL + DP (ϵ=5, δ=10e-5)	0.810	0.870	0.853	0.709
	Baseline + FL + DP + shuffling (ϵ=5, δ=10e-5), PPML-Omics	0.763	0.852	0.817	0.651
Camp-Liver (777 samples 7 cell types)	Baseline (Centrally Trained)	0.517	0.631	0.674	0.424
	Baseline + FL	0.512	0.649	0.666	0.418
	Baseline + FL + DP (ϵ=5, δ=10e-5)	0.472	0.581	0.652	0.383
	Baseline + FL + DP + shuffling (ϵ=5, δ=10e-5), PPML-Omics	0.539	0.639	0.693	0.440
Hrvatin (48,266 samples 8 cell types)	Baseline (Centrally Trained)	0.772	0.847	0.832	0.684
	Baseline + FL	0.794	0.866	0.832	0.710
	Baseline + FL + DP (ϵ=5, δ=10e-5)	0.848	0.861	0.857	0.783
	Baseline + FL + DP + shuffling (ϵ=5, δ=10e-5), PPML-Omics	0.859	0.860	0.867	0.797

Table 2 Comparison of different settings in the integration task of spatial gene expression and tumour morphology with spatial transcriptomics

Architecture	Averaged Training MSE	Averaged Testing MSE
Baseline (ST-Net, Centrally Trained)	0.853899	0.908258
Baseline + FL	0.887772	0.908397
Baseline + FL + DP (ϵ=0.01, δ=10e-5)	1.012351	0.926142
Baseline + FL + DP (ϵ=0.1, δ=10e-5)	0.906163	0.910072
Baseline + FL + DP + shuffling (ϵ=0.01, δ=10e-5), PPML-Omics	0.934313	0.910782
Baseline + FL + DP + shuffling (ϵ=0.1, δ=10e-5), PPML-Omics	0.896146	0.909360