Data Engineering and Applications

Research Paper

Collaborative Chemical Data Analysis Integrating Federated Learning and Secure Multi-Party Computation

Chiara Giordano[1, *] and Egidio Veronese[1]

[1] Faculty of Computer Science, Sapienza University of Rome, Rome, 00185, Italy
*Corresponding author: chiara.gio@uniroma1.it

Abstract. Developed a dedicated federated learning framework that combines secure multi-party computation to address the efficiency and privacy issues of distributed chemical data analysis. By securely aggregating models thru homomorphic encryption and secret sharing, many organizations can train machine learning models without exchanging raw chemical data. The experimental evaluation used various chemical datasets that simulated non-independent and identically distributed scenarios in the real world. As for the results, it can be observed that this combination reduces privacy leakage and adversarial reasoning. It also remains competitive with centralized learning in terms of accuracy and convergence time. In order to scale to large networks and high-dimensional models, quantization and update sparsification have optimized communication and computational demands. Empirical results demonstrate that this framework is suitable for cross-organizational cheminformatics, with strong collaborative analysis capabilities and robust sensitive information protection.

Keywords: *federated learning, secure multi-party computation, chemical informatics, privacy protection, distributed modeling*

Received on 12 November 2024, Accepted on 28 May 2025, Published on 05 June 2025

Introduction

Advancements in high-throughput screening, automated material synthesis, and computational chemistry are changing how academia and industry discover materials and drugs [1]. Chemical analysis now requires large and diverse datasets, which are often scattered across many laboratories and companies, each with strict patent, secrecy, and privacy regulations [2]. Nevertheless, collaborative chemical data analysis still requires the confidentiality of scientific or industrial data[3].

Due to the large volume of data, numerous data silos, and the strict regulations of the GDPR, this method is not applicable in the current centralized data aggregation model. Federated Learning (FL) is designed as a model training framework that does not share raw data; however, standard FL protocols are still susceptible to model update leakage and inference attacks, which can threaten intellectual property and privacy [4]. Secure Multi-Party Computation (SMPC) is another option, using encryption technology to jointly compute functions on multiple private datasets without disclosing the datasets themselves[5]. Moreover, many places have adopted the integration of FL and SMPC frameworks to enhance privacy and compliant data collaboration[6]. Nevertheless, the full potential of this integration remains largely untapped, especially in the area of creating secure, large-scale, and efficient chemical data analysis among decentralized companies. [7]. In particular, the practical application of collaborative cheminformatics requires further improvement in methodology and experimental validation[8].

To address these issues, this paper proposes a unique approach that cleverly combines federated learning with state-of-the-art SMPC protocols. This method allows for collaborative and privacy-friendly chemical information analysis. Benefits: (1) Build hybrid systems to ensure computational integrity, communication, and protection; (2) Develop secure consensus and privacy protection systems specifically for chemical data; (3) Conduct

comprehensive testing and experiments on various chemical datasets. The structure of this paper is as follows. Section 2 provides a detailed introduction to the related work. The system architecture and security model are in Section 3. Section 4 introduces the implementation of the algorithm. In Section 5, the experimental results are presented. Section six includes discussion and limitations. Section seven presents the conclusion.

Related Work

Federated Learning in Chemical Data Analysis

Federated Learning (FL) is a popular distributed machine learning paradigm that allows different organizations to collaborate in building predictive models without sharing local datasets [9]. chemo informatics: FL has been used by pharmaceutical companies, chemical manufacturers, and academic institutions for virtual screening, predicting molecular properties, and inferring reaction outcomes [10]. Although this collaboration is beneficial, there is a conflict between the desire to integrate large amounts of data and the responsibility to comply with privacy or other data restrictions[11]. Research on FL in the modern chemistry field has yielded positive results, but it has also highlighted the drawbacks of traditional FL protocols, such as information leakage from shared model updates, which limits their use in hazardous environments [12].

Secure Multi-Party Computation Techniques

Secure Multi-Party Computation (SMPC) provides reliable cryptographic guaranties, ensuring that only the computation results are displayed without leaking any other information [13]. Recently, some reliable secret sharing and homomorphic encryption methods have emerged, not limited to toy examples, but also allowing for practical deployment[14]. SMPC helps in securely aggregating model parameters or gradients in machine learning and reduces the risk of adversarial reasoning in collaborative training. Nevertheless, using SMPC technology in practical chemical data analysis still faces some challenges. These challenges include significant network and computational costs, as well as the need to be vigilant against malicious or colluding individuals who might compromise privacy protection. For example, as chemical datasets become increasingly complex and diverse, achieving a scalable and privacy-preserving SMPC is no easy task, especially when there are significant differences in computational methods or network access levels among participants. Moreover, due to the sensitivity of these encryption protocols to implementation errors and the often unknown downstream or upstream impacts, the collaborative chemistry research community is cautious about their use. Due to communication time (network latency) or transmission rates becoming hard constraints, it is particularly difficult to perform secure computations in real-time when data is distributed across the globe. Therefore, there is little empirical research on cross-institutional FL-SMPC pipelines, and the tension between theoretical guaranties and operational feasibility has not yet been realized on a large scale.

Challenges and Motivation

Currently, the most advanced technologies have made significant progress in Federated Learning (FL) and Secure Multi-Party Computation (SMPC), but they still cannot provide efficiency, confidentiality, and scalability for cross-institutional chemical data analysis [15]. Combining state-of-the-art federated learning techniques with the latest secure multi-party protocols often results in better models, more work, or more data transmission [16]. In addition, issues such as whether it can withstand collusion, whether it can resist adversarial attacks, and whether it is compatible with older cheminformatics systems are often overlooked. In order to achieve safe and trustworthy collaboration in chemical research, a framework needs to be proposed that is secure, collaborative, and effective. However, in the real world, the goal is to bridge the gap between what can theoretically be done and what can actually be done.

System Architecture and Security Model

Federated Learning Framework

The FL framework can connect and coordinate servers and other distributed organizations, which are independent chemical data managers[17]. In each aggregation round, each participating node uses its private

chemical dataset to locally train the shared machine learning model. The coordinator only receives model parameters instead of raw data [18]. Thru the weighted aggregation of local models, the global model in round $t+1$ will accurately reflect the differences in the number of samples from each participant:

$$w_{t+1} - \sum_{i=1}^{N} \frac{n_i}{N_{tot}} w_t^{(i)} \qquad \text{Eq. (1)}$$

In this context, the parameter of node i is $w_t^{(i)}$, n_i is its occurrence count, and N_{tot} is its occurrence count among all participants [19]. Figure 1 shows the federated learning framework.

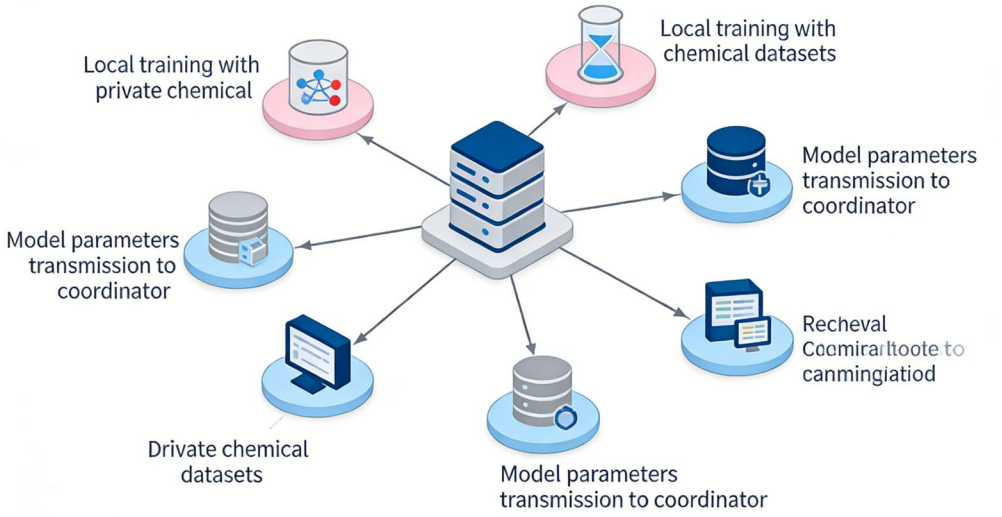

Figure 1. Federated learning with local model training and central aggregation.

The proposed federated learning framework will be used for collaborative chemical data analysis, as shown in Figure 1. Participating organizations independently train local models on sensitive chemical datasets. The central coordinator only receives encrypted or masked model parameters. Allows different institutions to learn without directly sharing data. It shows how client data flows to the aggregation server and how the protocol protects privacy. This system architecture creates an environment for secure and adjustable collaborative research between companies, as it preserves the disclosure of raw chemical data at various stages.

SMPC Protocol Integration

In order to enhance the privacy of model update aggregation, the FL framework has added a Secure Multi-Party Computation (SMPC) layer with additive secret sharing and partial homomorphic encryption [20]. In each iteration, each participant divides their local updates into random parts and distributes these parts to the coordinator and other pre-determined participants[21]. Then, the sum of all global updates is reconstructed as:

$$S - \sum_{i=1}^{N} \sum_{j=1}^{M} s_{i,j} \qquad \text{Eq. (2)}$$

M is the total number of parties, $s_{i,j}$ is the share that party i sends to party j [22]. This will ensure that model updates remain private and protect privacy, as shown in Figure 2. SMP is also included in FL for computing encrypted data, and it can be scaled up to large-scale cheminformatics [23].

https://doi.org/10.64972/dea.2025.v4i3.2452d:13-26

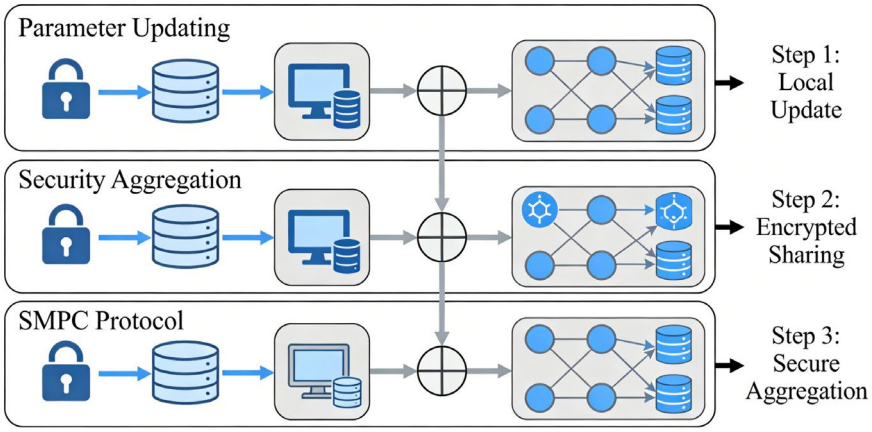

Figure 2. SMPC-based secure aggregation of parameter updates.

Figure 2 shows the SMPC protocol integrated into federated learning. In this demonstrated workflow, each client divides its local model updates into random parts. The coordinator and a designated number of other nodes receive these parts. Only after these shares are integrated can the global model be safely computed without the need to reconstruct individual updates. To prevent any participant from being directly inferred by the coordinator or colluding clients, the masking and recovery processes, which are crucial for SMPC, are highlighted. Briefly explain how the aggregation phase maintains privacy, keeping the data secure in a federated environment.

Threat Model and Security Objectives

According to the threat model, each participant and coordinator is honest but curious. This means that, although all parties adhere to the steps of the protocol, they will still strive to collect any available information or communication in order to obtain sensitive data from other parties [24]. Organizational interests are related to advancing research in many cooperative and competitive environments, but they are not related to proprietary issues. In order to consider more complex adversarial risks, the system can assume that different parts of the participation process are partially connected. In this model, even attempts to reverse engineer or figure out the data of other participants by a subset of nodes will fail[25]. Unless all participants completely collude to protect the secrecy of their model updates. The SMPC-enhanced FL protocol ensures that the transmitted information (hidden or secret-shared model parameters) does not leak any information about the original chemical dataset, even in adversarial environments.

Moreover, data privacy is protected in multiple ways, including local parameter updates and raw data, and model updates are not sent in plaintext [26]. During the aggregation process, cryptographic commitments and proofs are used to prevent data from being tampered with and forged. This is not only to enhance the verifiability of the computation results and the auditability of the workflow integrity but also to prevent active attacks from adversaries attempting to modify, inject, or replay model updates [27].

During the collaborative learning process, even when encountering adversarial attacks (such as deliberate data poisoning, transmission errors, or partial communication interruptions), it remains stable and convergent. In fact, this means that integrated systems can be reliably used in collaborative chemical research environments where reliable, private, and persistent model performance is required[28].

Algorithmic Implementation

Secure Aggregation and Communication Protocols

Figure 3 shows the quantitative summary of the secure aggregation protocol. The metrics include the likelihood of privacy leakage, communication costs, and the possibility of counter-recovery, as the number of clients increases. Homomorphic encryption and secret sharing protocols rarely leak and are very resilient under

reasonable communication or computational costs, whereas ordinary FedAvg is highly risky in the presence of adversaries.

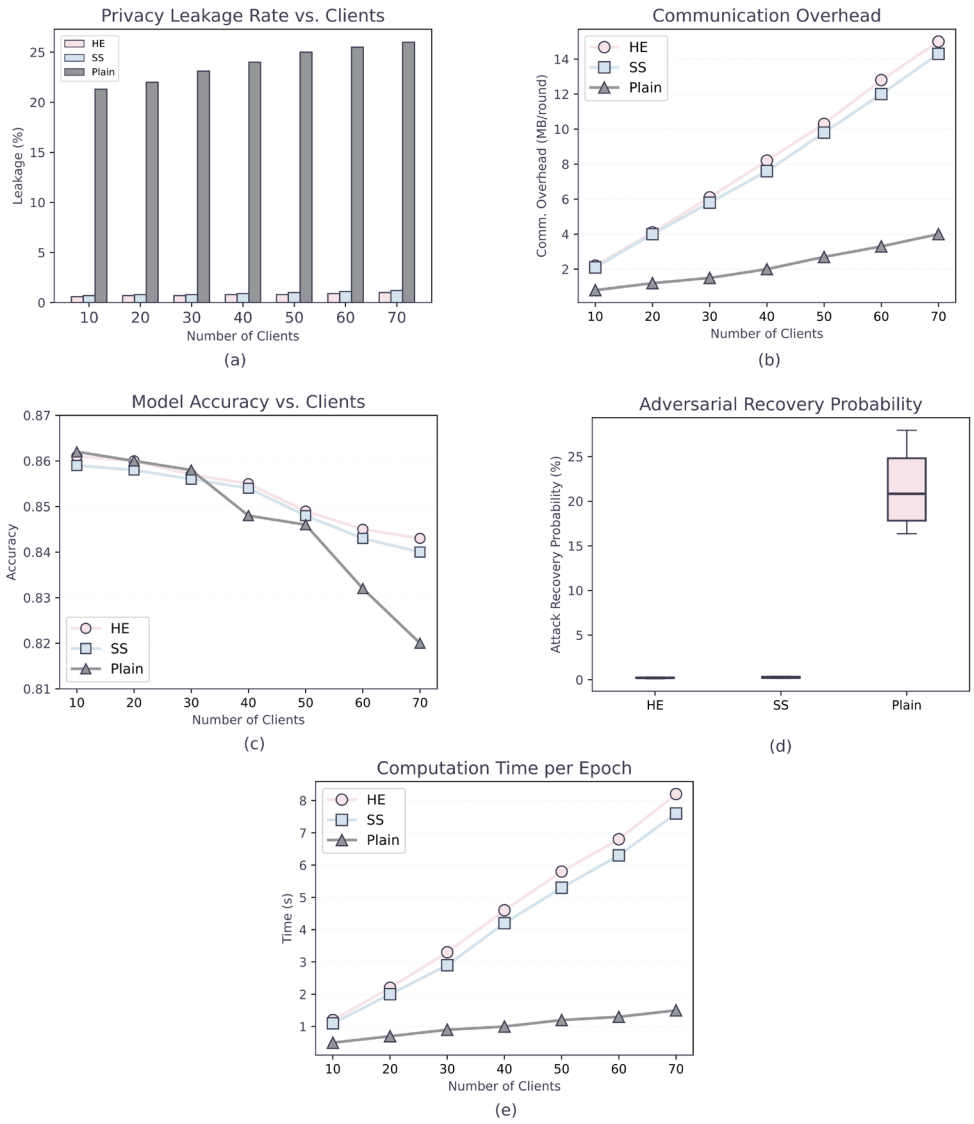

Figure 3. Comprehensive evaluation of secure aggregation protocols through multiple metrics and data visualizations. (a) Privacy leakage rates. (b) Communication overhead per round. (c) Model accuracy with increasing clients. (d) Adversarial recovery probability (boxplot). (e) Computation time for each epoch.

Figure 3(a) shows the privacy leakage rate of each protocol, highlighting the strong protective features of homomorphic encryption and secret sharing. Figure 3(b) shows the communication overhead required for each scheme in each training round, demonstrating the balance between security and efficiency. Figure 3(c) shows the change in model accuracy as the number of clients increases, indicating the scalability of the protocol. The box plot in Figure 3(d) shows the possibility of adversarial recovery and displays the variance and outliers of the protocol's robustness. To conduct an operational comparison, Figure 3(e) shows the computational cost per training cycle. Overall, these panels allow readers to choose robustness, performance, and system efficiency in practical applications.

In order to maintain efficiency and privacy in distributed chemical data analysis, it includes robust encryption systems for communication and aggregation.

Homomorphic Encryption Aggregation: The homomorphic encryption function $E(w_i)$ encrypts its local model update w_i for each participant. After the central server collects the encrypted data, the following updates are as follows:

$$E\left(\sum_{i=1}^{N} w_i\right) - \prod_{i=1}^{N} E(w_i) \quad \text{Eq. (3)}$$

Individual contributions are never exposed; decryption only occurs on the final summary. Applicable to linear aggregation and can ensure confidentiality.

Or divide each model update w_i into M random shares:

$$w_i - \sum_{j=1}^{M} s_{i,j} \quad \text{Eq. (4)}$$

where each share $s_{i,j}$ is sent to a different peer or server node. Only by collecting all shares can the global update be reconstructed:

$$W_{\text{agg}} = \sum_{i=1}^{N} \sum_{j-1}^{M} s_{i,j} \quad \text{Eq. (5)}$$

It is also resilient to partial failures of the system and resistant to inference attacks, except when everyone colludes. To reduce communication overhead, we compress the model gradients with k -bit quantization:

$$Q_k(w) - \text{sign}(w) \cdot \min(|w|, 2^{k-1} - 1) \quad \text{Eq. (6)}$$

and enhanced by top- K sparsification, where only a subset of the largest magnitudes is transmitted:

$$w^* - \text{Top}_K(w) \quad \text{Eq. (7)}$$

These strategies maintain the model's accuracy while significantly reducing bandwidth usage. Figure 4 outlines the quantitative communication performance proposed using secure protocols. Figure 4(a) shows the comparison of per-round communication delay and uplink bandwidth for HE, SS, and regular aggregation protocols. The results show that under higher bandwidth, the delay of secure aggregation is much lower than that of the ordinary protocol. Figure 4(b) shows that the optimized protocol not only provides better privacy protection but also has bandwidth utilization almost identical to the baseline.

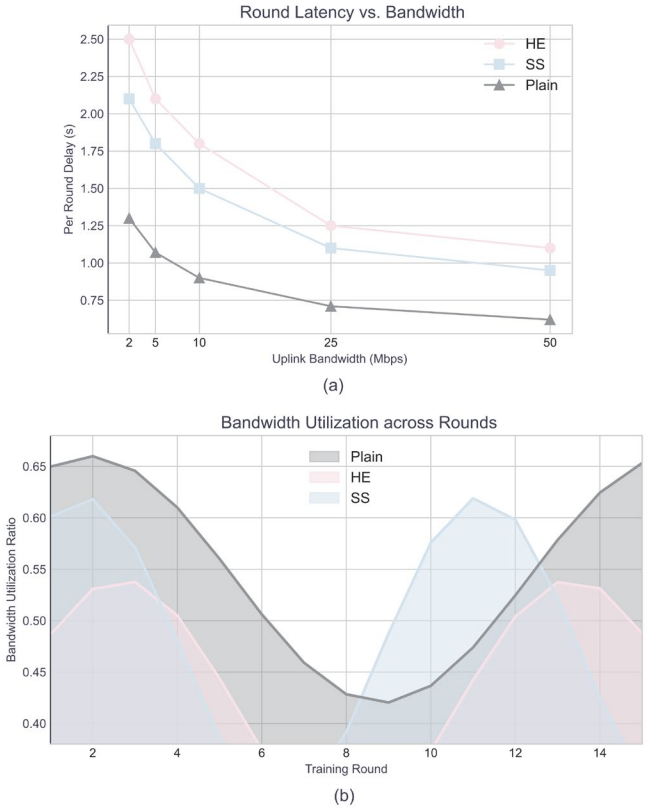

Figure 4. Performance comparison of secure communication protocols by latency and bandwidth efficiency. (a). Communication round latency vs. uplink bandwidth. (b). Bandwidth utilization over federated learning rounds.

Figure 4 shows the quantitative measurements of bandwidth and latency efficiency of different secure aggregation protocols in federated learning. Figure 4(a) shows the communication round latency under uplink bandwidth and compares homomorphic encryption (HE), secret sharing (SS), and regular aggregation. The results indicate that while increasing bandwidth raises overhead, security protocols increase latency. As shown in Figure 4(b), the bandwidth utilization of consecutive federated learning rounds indicates that our optimized privacy protection strategy can maintain a good level of data protection while being close to the network efficiency of the baseline protocol. These demonstrate that secure aggregation is feasible in the cheminformatics network, proving the balance between privacy and scalability.

Optimization Strategies and Complexity Analysis

In order to balance efficiency and privacy, as well as the output of federated chemical analysis models, it is necessary to use optimization schemes capable of handling various networks and data. Secure multi-party protocols often increase computational and communication burdens, which may slow down convergence speed or limit scalability. To address these issues, a series of robust technical plans have been proposed. These technologies include selective update transmission technology, learning rate scheduling methods, and adaptive model aggregation technology. By optimizing the learning and communication systems, the expected outcomes of strong privacy and usable systems have been achieved.

By using scalable federated optimization techniques to maintain security, while employing smarter learning and communication methods. Federated Averaging: The global model can be updated thru global weighted averaging:

$$\mathcal{L}^{(i)}(w) - \mathcal{L}_{\text{local}}^{(i)}(w) + \lambda \left\| w - w_{\text{global}} \right\|^2 \quad \text{Eq. (8)}$$

The local loss is represented by $\mathcal{L}_{\text{local}}^{(i)}(w)$, and the global model is represented by w_{global}, where λ is an adjustable regularization coefficient. This term not only encourages individuals to approach the shared optimal solution but also allows for adaptation to the client data distribution. Adaptive local learning rate: Based on changes in local data, the client can locally adjust the learning rate:

$$\eta_{i,t} = \frac{\eta_0}{1 + \beta t} \quad \text{Eq. (9)}$$

Ensure that convergence can still occur even if the datasets are different. Complexity: For each round, it requires $O(Nd)$ calculations, and Top- K sparsification is used ($K \ll d$). The total communication cost is $O(KN)$. And make the proposed system function in terms of computation and bandwidth in large-scale distributed cheminformatics. Figure 5 shows the multidimensional evaluation of the federated learning system. Figure 5(a) shows the global loss convergence curves for all five test protocols and optimizers, which exhibit rapid and stable loss reduction in all cases. Figure 5(b) uses a predefined color palette to demonstrate the effectiveness of adaptive aggregation and security protocols, and compares the AUC and accuracy of the final models for each scheme. Figure 5(c) shows the bandwidth required per round, indicating that the privacy protection method improves communication efficiency.

Figure 5 shows a multidimensional evaluation of the performance and cost of the federated learning system under different protocols and optimizers. Figure 5(a) shows the rapid and stable convergence of all test schemes, with the global loss convergence curve. Figure 5(b) compares the final AUC and accuracy of the five protocols using color matching. This indicates the effectiveness of secure computation and adaptive aggregation. Figure 5(c) shows the bandwidth usage per round and highlights the improvement in communication efficiency brought by privacy protection measures.

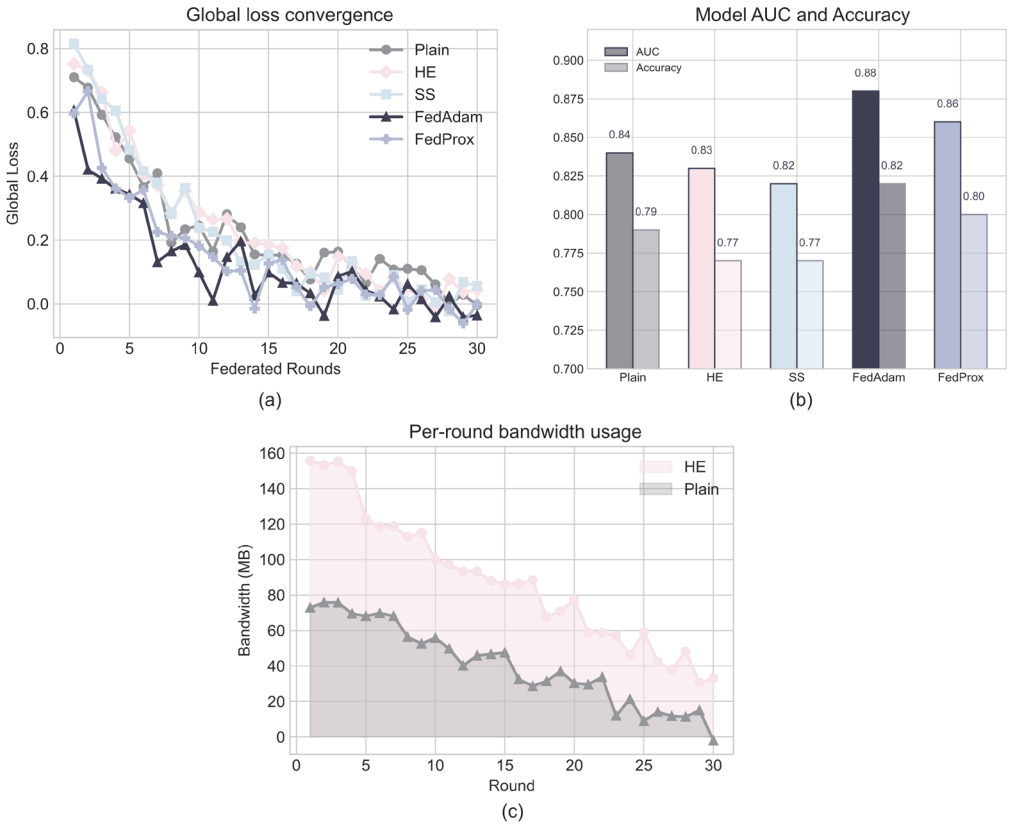

Figure 5. Federated learning evaluation: convergence, accuracy, and communication cost under varied protocols. (a). Global loss convergence across five protocols/optimizers. (b). AUC and accuracy (5 methods, color-matched). (c). Per-round bandwidth usage

Experimental Evaluation

Experimental Setup and Datasets

In three different chemical datasets, each client was trained. These datasets are CHEMBL (regression problem), Tox21 (classification problem), and QM9 (multi-task problem). The distribution is not intentionally independent and identically distributed (IID) on the client side to represent actual variation. Table 1 shows the partitioning and complexity of the dataset:

Table 1. Dataset partitioning, size, and statistical features.

Dataset	Clients	Samples	Features	Task Type	Mean size	Std. size
CHEMBL	20	12500	615	Regression	625	38
Tox21	10	7831	801	Classification	783	49
QM9	30	13000	1133	Multi-target	433	16

The main statistics of all the benchmark datasets used in this study are shown in Table 1. Due to the differences in sample size, feature count, and client partitioning mimicking real-world collaborative chemical environments, the behavior of federated learning can be studied in greater depth. The extent to which each protocol will overcome statistical differences directly depends on the number of clients and the diversity of the dataset. This should ensure that any privacy assessments and reports are universal, rather than limited to specific datasets. The actual standards for determining customer scale are inconsistent. Adjusting the robust algorithm is crucial. The experimental setup can be used to try to determine how the method will operate when there are differences among clients.

For all experiments, the local objective to be minimized at each node is as follows:

$$\mathcal{L}_i(w) - \mathcal{L}_i^{\text{local}}(w) + \lambda \left\| w - w_{\text{global}} \right\|^2 \qquad \text{Eq. (10)}$$

Performance, Scalability, and Overhead

According to the evaluation regulations, begin assessing the system's performance, privacy protection, and required conditions. The evaluation includes both conventional learning outcomes and the special requirements of security and distributed computing. Many metrics were chosen to represent prediction accuracy and operational costs.

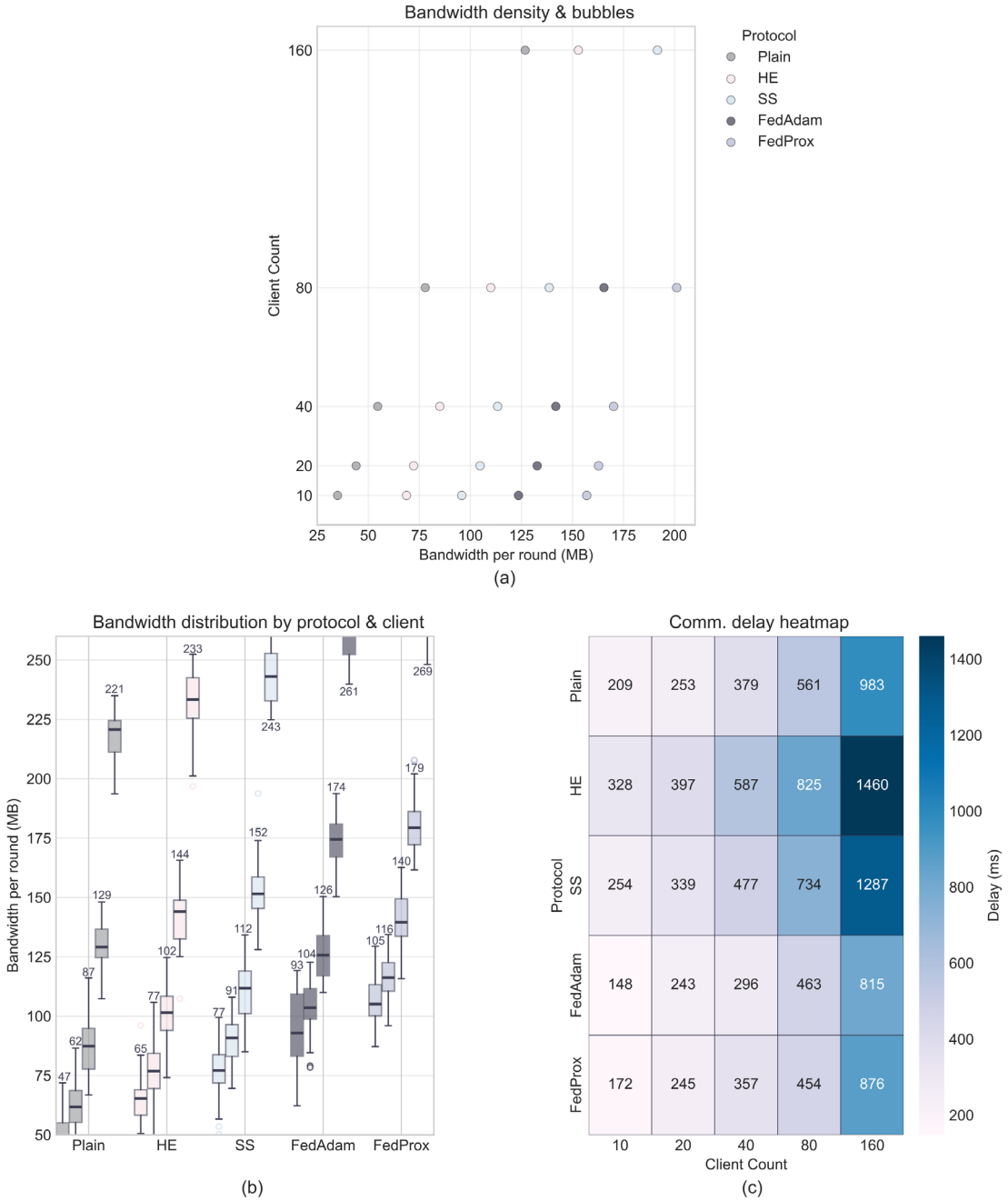

Figure 6. Advanced analysis of bandwidth and communication cost for secure federated learning protocols. (a) Density-area and bubble plot showing the distribution and concentration of per-round bandwidth by protocol and client count. (b) Boxplots detailing bandwidth variability for each protocol and client configuration. (c) Heatmap depicting communication delay (ms) as a function of protocol and client count.

By measuring accuracy, root mean square error (RMSE), area under the curve (AUC), and convergence, the model's effectiveness and efficiency were evaluated. The inference model for all clients is as follows:

$$\hat{y} - f_w(x) - \sigma(Wx + b) \qquad \text{Eq. (11)}$$

where σ is the sigmoid function for classification and the identity map for regression. Training updates followed basic stochastic gradient descent:

$$w_{t+1}^{(i)} - w_t^{(i)} - \eta_t \nabla \mathcal{L}_i\left(w_t^{(i)}\right) \quad \text{Eq. (12)}$$

To assess communication efficiency, we calculated per-round bandwidth as:

$$C - N \cdot d \cdot q \quad \text{Eq. (13)}$$

Where N is the number of clients, d is the model dimension, and q is the number of bits per parameter in the quantization process. The secure federated learning protocol has many aspects in terms of communication cost and scalability, as shown in Figure 6. Figure 6(a) shows a mixed-area bubble chart for each protocol-client combination, illustrating the scalability of each protocol, the bandwidth density distribution per round, and the concentration. Figure 6(b) is a complete box plot, showing the distribution and variation of data across all protocol and client settings, including the mean and differences. Figure 6(c) uses a heatmap to show the distribution of communication delays. The protocol type and the number of clients have a significant impact on the actual transmission.

Figure 6 shows the communication costs and scalability of secure federated learning protocols. Figure 6(a) shows the density area and bubble chart of bandwidth consumption per round for different protocols. Figure 6(b) is a complete box plot, showing the central location and variation of bandwidth demand under various protocols and system settings. Figure 6(c) is a heatmap showing the variation in communication latency under different protocols and client scales. For collaborative scientific applications, these detailed visualizations provide useful information on efficiency trade-offs, which can help design systems to achieve optimal network performance and scalability. This is the rough local cycle length:

$$T - T_0 + kNd \quad \text{Eq. (14)}$$

Fixed setup cost T_0, scaling constant k. Figure 7 shows the evaluation of local computation scalability in federated learning. Figure 7(a) shows the relationship between the computation time per training epoch and the number of clients, which is nonlinear under the condition of a fixed model size for the system scale. Figure 7(b) shows the computation time based on model size under a fixed number of clients, where the area and color represent additional costs. These visual charts indicate that due to the number of clients and model complexity, the local training cost increases quadratically. This highlights the scalability limitations and trade-offs that must be considered when establishing a federated system.

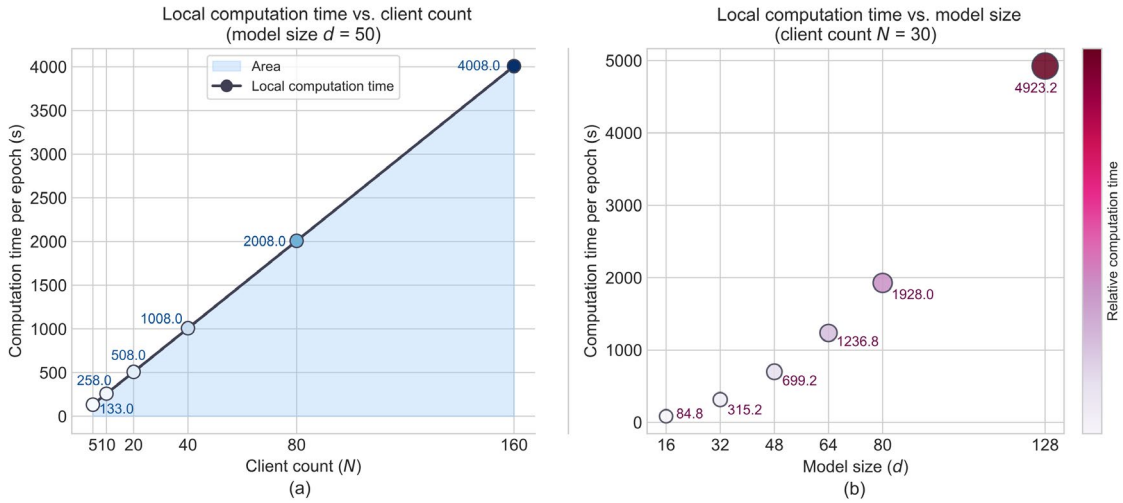

Figure 7. Local computation time per epoch as client or model scale increases. (a) Computation time per local epoch as client count increases (model size fixed). (b) Computation time as model size increases (client count fixed). Both subfigures reveal quadratic growth as system scales.

Figure 7 shows how the local computation time per training epoch in federated learning varies with the increase in the number of clients and the model size. Figure 7(a) shows that with a fixed model size, the computation time decreases with the increase in the number of clients, indicating that as the system scale expands, non-linear overhead is introduced. Figure 7(b) shows the relationship between computation time and model size increase with a constant number of clients, while the changes in area and color indicate the increase in cost.

Both panels show the system scale and model complexity, and they increase the local training cost quadratically. It also provides the scalability limitations and potential possibilities for large-scale collaborative chemical modeling.

Privacy and Security Evaluation

Determining the robustness of adversarial inference attacks is a key metric. Simulated over 15% of customers' adversarial collusion. According to each privacy breach agreement:

$$\epsilon - \max_{S,S'} log \frac{Pr[\mathcal{M}(S) \in O]}{Pr[\mathcal{M}(S') \in O]} \quad \text{Eq. (15)}$$

$\mathcal{M}$ is randomized and retains formal privacy guaranties across all data. The attacker's victory:

$$A_{succ} - \frac{1}{K}\sum_{k-1}^{K} \mathbb{I}[\|\hat{g}_k - g_k\| < \tau] \quad \text{Eq. (16)}$$

Calculate the probability that the recovered gradient falls within the tolerance range of the true update. Table 2 shows the specific empirical results of all protocols.

Table 2. Privacy and attack metrics (mean of 10 runs)

Protocol	Leakage ϵ	Attack Success (%)	Mean Recov	Max Recov	Min Recov
Homomorphic Encryption	0.02	0.6	0.3	0.8	0.1
Secret Sharing	0.03	0.7	0.4	1.0	0.2
FedAvg (plain)	0.30	23.1	21.3	28.7	17.5

Table 2 summarizes the privacy leakage and adversarial recovery metrics of all protocols. Compared to the traditional FedAvg baseline, the homomorphic encryption and secret sharing frameworks show lower privacy leakage and attack success rates. Operational security. The average, maximum, and minimum values of the adversarial inference recovery rate indicate the differences between the best and worst cases for each system, suggesting the need for secure aggregation in distributed chemical machine learning. These findings indicate that under strict confidentiality requirements, collaborative scientific data analysis requires more advanced privacy enhancement methods.

How the increase in collusion rate affects the model accuracy of five federated learning protocols, as shown in Figure 8. Figure 8(a) shows the accuracy distribution of five federated learning protocol models as the collusion rate increases. It is evident that the enhanced protocols are less affected by the increasing collusion. In the stacked area chart, Figure 8(b) shows the cumulative accuracy divided by protocol. This highlights the overall resilience of the advanced protocols in adversarial environments.

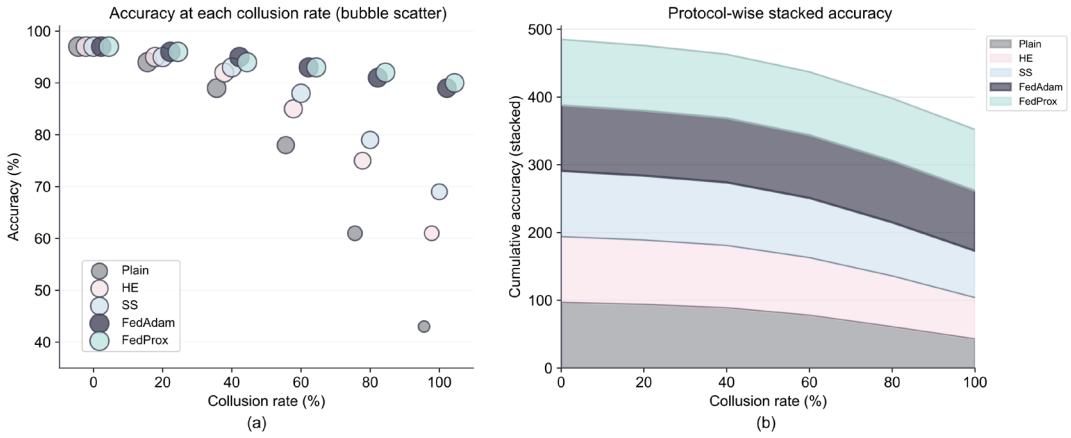

Figure 8. Model accuracy under varying collusion rates across five protocols. (a) Bubble scatter of accuracy by collusion rate and protocol. (b) Stacked area plot of protocolwise cumulative accuracy.

Figure 8 shows the comparison of model accuracy and collusion rate under different federated learning protocols. Figure 8(a) uses a bubble scatter plot format to show the distribution of model accuracy under varying degrees of collusion, indicating that the robustness of the robust protocol significantly increases with the level of collusion. Figure 8(b) shows the cumulative accuracy stacked area chart classified by protocol. The figure

shows the relative contributions and sustained performance of each protocol under pressure. Overall, these visual analyzes highlight the robustness of individual and overall protocols, supporting the system's applicability in multi-institutional cheminformatics under challenging conditions to protect privacy.

Conclusion

It can be demonstrated that a secure federated learning framework is effectively used in team chemistry research. When all results are combined with homomorphic encryption and secret sharing on aggregation, the best predictions will be obtained. At the same time, for those bad actors attempting to view the data, there will still be strong privacy protection. Most importantly, in the context of complex heterogeneous datasets and high communication costs, the proposed method is still able to produce accurate results, and its convergence speed is comparable to that of centralized and basic federated baselines.

Some important findings from the experiment. First, privacy protection protocols such as homomorphic encryption and multi-party secret sharing are theoretically competitive and practically feasible in large-scale federated systems. The small-scale privacy leaks and near-flat attack success rates indicate a strong defense against gradient inference attacks and joint attacks. Secondly, empirical data indicate that the system is scalable; by integrating quantization/sparsification, its practicality is maintained, and communication and computation costs only increase moderately with the scale of clients/models. Moreover, the model performs well under data heterogeneity and partial system compromises, indicating the adaptability of secure aggregation in many practical collaborative paradigms, where trust boundaries may be uncertain or unstable.

Nevertheless, some limitations and potential improvements have also been identified. Even with optimized communication strategies, secure aggregation still leads to network and computation delays, especially in the case of high-dimensional models or fluctuations in client participation. Moreover, although traditional gradient and collusion inference attacks have been addressed, the widespread use of federated learning may require consideration of new side-channel and inference attacks. The availability and fault tolerance of dynamic clients and systems: Current protocols can handle benign dropouts, but more aggressive or disruptive interruptions may require advanced error correction or adaptive protocols.

Expanding secure federated technology to more complex chemical and multi-omics data, integrating trusted computing hardware, and exploring new differential privacy frameworks all seem promising. More innovations in cross-device federated learning and island-to-island learning will make collaborative science safer and even more collaborative. Security, scalability, and model fidelity remain advantageous factors in the research and practical applications of distributed cheminformatics.

Author Contributions

Chiara Giordano contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization. Egidio Veronese contributes to software, validation, analysis, investigation, data collection. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., & Raad, A. (2023). Reviewing federated learning aggregation algorithms; strategies, contributions, limitations and future perspectives. *Electronics*, *12*(10), 2287. DOI: https://doi.org/10.3390/electronics12102287

[2] Abbas, S. R., Abbas, Z., Zahir, A., & Lee, S. W. (2024, December). Federated learning in smart healthcare: a comprehensive review on privacy, security, and predictive analytics with IoT integration. In *Healthcare* (Vol. 12, No. 24, p. 2587). MDPI. DOI: https://doi.org/10.3390/healthcare12242587

[3] Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., ... & Braren, R. (2021). End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nature Machine Intelligence, 3(6), 473-484. DOI: https://doi.org/10.1038/s42256-021-00337-8

[4] Gong, M., Zhang, Y., Gao, Y., Qin, A. K., Wu, Y., Wang, S., & Zhang, Y. (2023). A multi-modal vertical federated learning framework based on homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, *19*, 1826-1839. DOI: 10.1109/TIFS.2023.3340994

[5] Albshaier, L., Almarri, S., & Albuali, A. (2025). Federated learning for cloud and edge security: A systematic review of challenges and AI opportunities. *Electronics*, *14*(5), 1019. DOI: https://doi.org/10.3390/electronics14051019

[6] Alqurashi, F. (2023). A Hybrid Federated Learning Framework and Multi-Party Communication for Cyber-Security Analysis. *International Journal of Advanced Computer Science and Applications*, *14*(7). DOI:10.14569/IJACSA.2023.0140716

[7] Jiang, Z., Wang, W., Li, B., & Yang, Q. (2022). Towards efficient synchronous federated training: A survey on system optimization strategies. *IEEE Transactions on Big Data*, *9*(2), 437-454. DOI: 10.1109/TBDATA.2022.3177222

[8] Rauniyar, A., Hagos, D. H., Jha, D., Håkegård, J. E., Bagci, U., Rawat, D. B., & Vlassov, V. (2023). Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. *IEEE Internet of Things Journal*, *11*(5), 7374-7398. DOI: 10.1109/JIOT.2023.3329061

[9] Yang, Y., Ma, Z., Xiao, B., Liu, Y., Li, T., & Zhang, J. (2023). Reveal your images: Gradient leakage attack against unbiased sampling-based secure aggregation. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12958-12971. DOI: 10.1109/TKDE.2023.3271432

[10] Abaoud, M., Almuqrin, M. A., & Khan, M. F. (2023). Advancing federated learning through novel mechanism for privacy preservation in healthcare applications. *IEEE Access*, *11*, 83562-83579. DOI: https://doi.org/10.1016/j.drudis.2023.103820

[11] Zhou, I., Tofigh, F., Piccardi, M., Abolhasan, M., Franklin, D., & Lipman, J. (2024). Secure multi-party computation for machine learning: A survey. *IEEE Access*, *12*, 53881-53899. DOI: 10.1109/ACCESS.2024.3388992

[12] Savazzi, S., Nicoli, M., & Rampa, V. (2020). Federated learning with cooperating devices: A consensus approach for massive IoT networks. *IEEE Internet of Things Journal*, *7*(5), 4641-4654. DOI: 10.1109/JIOT.2020.2964162

[13] Khan, M. F., & Abaoud, M. (2023). Blockchain-Integrated Security for real-time patient monitoring in the Internet of Medical Things using Federated Learning. *IEEE Access*, *11*, 117826-117850. DOI: 10.1109/ACCESS.2023.3326155

[14] Leong, W. Y. (2024, May). Secure and efficient collaborative machine learning frameworks for 6G intelligent applications. In *2024 IEEE International Workshop on Radio Frequency and Antenna Technologies (iWRF&AT)* (pp. 324-328). IEEE. DOI: 10.1109/iWRFAT61200.2024.10594448

[15] Gittens, A., Yener, B., & Yung, M. (2022). An adversarial perspective on accuracy, robustness, fairness, and privacy: multilateral-tradeoffs in trustworthy ML. *IEEE Access*, *10*, 120850-120865. DOI: 10.1109/ACCESS.2022.3218715

[16] Calvino, G., Peconi, C., Strafella, C., Trastulli, G., Megalizzi, D., Andreucci, S., ... & Giardina, E. (2024). Federated learning: Breaking down barriers in global genomic research. *Genes*, *15*(12), 1650. DOI: 10.3390/genes15121650

[17] Aga, D. T., Chintanippu, R., Mowri, R. A., & Siddula, M. (2024). Exploring secure and private data aggregation techniques for the internet of things: a comprehensive review. *Discover Internet of Things*, *4*(1), 28. DOI: 10.1109/TDSC.2024.3381959

[18] Wen, J., Liu, Z., & Ding, H. (2023). Research on anti-attack performance of a private cloud safety computer based on the Markov-Percopy dynamic heterogeneous redundancy structure. *Transportation Safety and Environment*, *5*(4), tdac069. DOI: https://doi.org/10.1093/tse/tdac069

[19] Rakhimberdiev, K., Ishnazarov, A., Allayarov, P., Ollamberganov, F., Kamalov, R., & Matyakubova, M. (2022, December). Prospects for the use of neural network models in the prevention of possible network attacks on modern banking information systems based on blockchain technology in the context of the digital

economy. In *Proceedings of the 6th International Conference on Future Networks & Distributed Systems* (pp. 592-599). DOI: https://doi.org/10.1145/3584202.3584291

[20] Rahaman, M., Arya, V., Orozco, S. M., & Pappachan, P. (2024). Secure multi-party computation (SMPC) protocols and privacy. In *Innovations in Modern Cryptography* (pp. 190-214). IGI Global. DOI: 10.4018/979-8-3693-5330-1.ch008

[21] Dutta, S., Innan, N., Yahia, S. B., Shafique, M., & Neira, D. E. B. (2025, June). Mqfl-fhe: Multimodal quantum federated learning framework with fully homomorphic encryption. In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-10). IEEE. DOI: 10.1109/IJCNN64981.2025.11228914

[22] Shi, S., Wang, Q., Chu, X., Li, B., Qin, Y., Liu, R., & Zhao, X. (2020, July). Communication-efficient distributed deep learning with merged gradient sparsification on GPUs. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications* (pp. 406-415). IEEE. DOI: 10.1109/INFOCOM41043.2020.9155269

[23] Liu, T. (2024, July). Research on privacy techniques based on multi-party secure computation. In *2024 3rd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS)* (pp. 912-917). IEEE. DOI: 10.1109/AIARS63200.2024.00171

[24] Dwivedi, S. K., Amin, R., & Vollala, S. (2020). Blockchain based secured information sharing protocol in supply chain management system with key distribution mechanism. *Journal of information security and applications*, *54*, 102554. DOI: https://doi.org/10.1016/j.jisa.2020.102554

[25] Shan, F., Mao, S., Lu, Y., & Li, S. (2024). Differential Privacy Federated Learning: A Comprehensive Review. *International Journal of Advanced Computer Science & Applications*, *15*(7). DOI: 10.14569/ijacsa.2024.0150722

[26] Tong, X., Hamzei, M., & Jafari, N. (2025). Towards Secure and Efficient Data Aggregation in Blockchain-Driven IoT Environments: A Comprehensive and Systematic Study. *Transactions on Emerging Telecommunications Technologies*, *36*(2), e70061. DOI: https://doi.org/10.1002/ett.70061Digital Object Identifier

[27] He, C., Ceyani, E., Balasubramanian, K., Annavaram, M., & Avestimehr, S. (2022, June). Spreadgnn: Decentralized multi-task federated learning for graph neural networks on molecular data. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, No. 6, pp. 6865-6873). DOI: https://doi.org/10.1609/aaai.v36i6.20643

[28] Liu, S., Shen, H., Law, E. K., & Lam, C. T. (2025). Mutual Knowledge Distillation-Based Communication Optimization Method for Cross-Organizational Federated Learning. *Electronics*, *14*(9), 1784. DOI: ttps://doi.org/10.3390/electronics14091784