A Federated Learning Benchmark for Drug-Target Interaction

Gianluca Mittone*
gianluca.mittone@unito.it
University of Turin
Turin, Italy

Filip Svoboda*
fs437@cam.ac.uk
University of Cambridge
Cambridge, UK

Marco Aldinucci
marco.aldinucci@unito.it
University of Turin
Turin, Italy

Nicholas D. Lane
ndl32@cam.ac.uk
University of Cambridge
Cambridge, UK

Pietro Lio'
pl219@cam.ac.uk
University of Cambridge
Cambridge, UK

ABSTRACT

Aggregating pharmaceutical data in the drug-target interaction (DTI) domain can potentially deliver life-saving breakthroughs. It is, however, notoriously difficult due to regulatory constraints and commercial interests [5, 18]. This work proposes the application of federated learning, which is reconcilable with the industry's constraints. It does not require sharing any information that would reveal the entities' data or any other high-level summary. When used on a representative GraphDTA model and the KIBA dataset, it achieves up to 15% improved performance relative to the best available non-privacy preserving alternative. Our extensive battery of experiments shows that, unlike in other domains, the non-IID data distribution in the DTI datasets does not deteriorate FL performance. Additionally, we identify a material trade-off between the benefits of adding new data and the cost of adding more clients.

KEYWORDS

Federated Learning, Graph Neural Networks, Drug-Target Interaction, Benchmark

ACM Reference Format:
Gianluca Mittone, Filip Svoboda, Marco Aldinucci, Nicholas D. Lane, and Pietro Lio'. 2023. A Federated Learning Benchmark for Drug-Target Interaction. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3543873.3587687

1 INTRODUCTION

Federated learning (FL) is a privacy-preserving distributed learning that has gathered ground in healthcare applications over the past few years. Since it fits very well with the requirement of preserving patient data confidentiality, it saw considerable uptake in the analysis of Electronic Health Records and healthcare IoT, such as mobile health [7, 12, 15]. The closest application of Federated Learning to Drug-Target Interaction (DTI) was the solitary example of FL-QSAR, which presented the first federated model trained for a related drug discovery task, but stopped short of analyzing its performance beyond demonstrating its feasibility for up to 4 clients [3]. Instead, providing privacy and security to drug discovery in general, and DTI in particular, has been approached as a cryptography problem by obscuring the underlying data such that data itself and high-level statistics were rendered useless. However, a model was still trainable on it [10].

This paper delivers the first-ever Federated Learning benchmark for the DTI task, achieving up to a 15.53% reduction in MSE compared to a possible ensemble learning-based alternative. Furthermore, we develop a novel comprehensive analysis framework for FL applications letting us identify and explain a significant and material difference between the sensitivity of FL to non-IID data in the DTI task and sensitivity to it in any other task FL has been previously applied to, and discover and explore the importance of data ownership structure in FL for DTI as a major performance determinant and a key consideration when engaging real-world actors in the process of cooperatively training models. Ours is a novel and comprehensive analysis of FL in a critical and under-explored data domain.

This paper's scope is limited to the drug-target interaction DTI task of the drug discovery domain due to computational and other practical considerations. This task regresses the tuple protein-drug input onto a vector describing their interaction. In this domain, we chose to work with a single representative model. We chose the GraphDTA [13] model as it is the backbone of many current state-of-the-art models [4, 6, 14]. Our experiments aim to represent the complexities a federation of pharmaceutical labs would entail as realistically as possible. In particular, we deliberately explored the whole spectra of IID-ness and data ownership distribution. Finally, we only perform our experiments using the core algorithms in FL and distributed learning. This choice does not imply loss of generality, as any specific feature that might improve the performance of either one of them can be straightforwardly re-implemented for use by the other.

In summary, our contributions are the following:

- we deliver the first-ever Federated Learning benchmark for the DTI task;
- we achieve up to a 15.53% reduction in MSE when compared to a bagging-based alternative;

*Both authors contributed equally to this research.

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04…$15.00
https://doi.org/10.1145/3543873.3587687

- we develop a novel comprehensive analysis framework for FL applications, allowing us to identify data ownership as a major performance determinant;
- we report almost 200 FL training results through many heatmaps characterising the performance of the final model when trained under a wide spectrum of non-IID-ness levels in data distribution and different federation sizes.

Each reported experiment needs approximately one GPU hour on a NVIDIA A40 or 7-8 on GTX-1080. We ran 197 experiments for 1,576 GPU hours spent on this research work.

2 METHODOLOGY OF THE PROPOSED APPROACH

This work proposes to use Federated Learning based on the FedAverage [11] algorithm to fit the GraphDTA model [13] on the KIBA [17] dataset split among multiple clients.

Federated learning is a distributed learning paradigm that shares model parameters at a much lower frequency than standard distributed learning. The exchanged model weights conceal each client's data sufficiently to preclude reconstruction. Straightforward extensions permit further increases in data protection, and defences against other potential interference [8, 9].

The KiBA dataset reports 246,088 Kinase Inhibitor BioActivity (KIBA) scores for 52,498 chemical compounds and 467 kinase targets, originating from three separate large-scale biochemical assays of kinase inhibitors. The score is a superior aggregate metric derived from a previously utilised battery of measurements such as IC_{50}, K_i, and K_d [17].

The GraphDTA model regresses the drug-target pair onto a continuous measurement of binding affinity for that pair, the KIBA score. It encodes the target as a 1D sequence and the drug as a molecular graph, making it possible for the model to capture the bonds among atoms directly [13]. To stay true to the simple the better ethos of this paper, we refrained from implementing fancy aggregation strategies or specialising too much the chosen model to fit the federated task. We limited ourselves to replacing stateful objects (outside the weights, clearly) with stateless ones: batch normalisation with layer normalisation and ADAM with SGD. These choices let us obtain more stable learning curves and cleaner convergence of the federated model without harming its performance. Both FL and non-FL ran with the same adjusted architecture.

Our implementation uses the open-source FLOWER[1] framework to implement the model federation and to simulate its running on multiple clients. We use the FedAverage aggregation algorithm, which combines local stochastic gradient descent (SGD) on each client with a server that performs model averaging [11].

The experimental setup builds on the experiments usually associated with Federated Learning benchmarks while substantially expanding them. First, the model is compared against a suitable alternative. Given the lack of prior work, there was no ready candidate for this comparison. A centralised model is unsuitable since its use is unrealistic due to the aforementioned regulatory and commercial considerations. The cryptographic approaches to data anonymisation would be usable in real life; however, they are not a direct competitor to Federated Learning. They can augment each other and provide joint solutions similar to what Federated learning with differential privacy does [19]. Ultimately, as a possible fair comparison, we chose a simple Bergman's ensemble [2] of models, each being trained separately on a different data split of the entire dataset. The data splits are maintained constants in comparing FL and Ensemble Learning. The choice of baseline algorithms for both FL and the ensemble is deliberate, as any extension applicable to one can be straightforwardly re-engineered for use with the other [16]. Therefore, working with simple implementations provides us with a fair, uncoloured comparison of the two approaches rather than of their two randomly chosen extensions. The metric used to evaluate each experiment is the Mean Squared Error (MSE); in the case of FL, the MSE of the global model is computed, while in the case of bagging, the MSE of the ensemble is taken into account. The test set is the same for all experiments, allowing for a fair comparison of different runs.

The code is made available on GitHub [1]. It can be used out of the box without the knowledge of distributed or Federated learning. It works with PyTorch deep models, but it will eventually be compatible with TensorFlow. It is being shared for the benefit of the Biologists working on DTI and those interested in proving and capitalising on Federated Learning's usefulness as a secure, privacy-preserving, and performance-conserving platform for sharing pharmaceutical data under regulatory and commercial constraints.

3 RESULTS

Superior and privacy-preserving performance of our network is displayed in table 1. It reports the performance difference between the federation of deep model architectures and an ensemble of the same architecture. Based on our experimental setup (Section 2), all experiments in this section exploit the same GraphDTA [13] architecture, and we consider our model's performance to be successful if it can match that of the non-private distributed alternative.

The results in table 1 show that our approach can retain up to 15% better performance relative to the distributed alternative while ensuring that no data or any other high-level summary of it is revealed [1]. The general trend in the IID results points to a relative advantage for the ensembles at very low client counts that quickly dissipates, turns into parity, and from 16 clients up fully reverses as the client count increases. Second, the non-IID data display effective parity practically at all client counts, indicating that FL can deal with unequal data distributions much better than the distributed alternative. This matched performance, alongside FL's solid privacy and security guarantees [1] entirely lacking in the distributed alternative, makes it a clear favourite for future distributed learning research in the DTI domain. Furthermore, the results invite us to explore deeper. In particular, seeing that the IID and non-IID performances are effectively matched, we ask how the FL performance develops under varying non-IID conditions in the following subsection.

DTI is a data distribution-agnostic domain. Data non-IID-ness in DTI is two-dimensional as there are two model inputs. The protein and the chemical are jointly taken in to predict their interaction. Consequently, we can investigate the distribution one dimension at a time, either non-IID to the protein or chemical

[1] https://github.com/Giemp95/FedDTI

Table 1: Performance of our DTI-FL relative to ensemble alternative [2]. The % difference columns refer to the federated values compared to the ensemble ones; note that a positive difference in the MSE highlights worse learning performance, while negative differences indicate better MSE and, consequently, better learning performance.

	IID distribution			non-IID distribution		
Client count	**Ensemble MSE**	**Federated MSE**	**% difference**	**Ensemble MSE**	**Federated MSE**	**% difference**
2 clients	0.509	0.530	+4.08%	0.550	0.556	+1.19%
4 clients	0.563	0.577	+2.58%	0.556	0.556	-0.05%
8 clients	0.567	0.574	+1.30%	0.568	0.574	+1.20%
16 clients	0.576	0.578	+0.42%	0.573	0.578	+0.690%
32 clients	0.709	0.599	-15.53%	0.579	0.578	-0.024%

inputs, or explore it in both dimensions simultaneously. Neither of these three approaches can be ruled out as a priory as the input classes are statistically independent of each other. Consequently, the domain does not lend itself easily to the established notions of non-IID-ness in FL, and we have to test non-IID-ness under all three conditions.

Our experiments investigate the entire continuum of IID-ness rather than just its two extrema. IID data distribution is a random draw; each data point has an equal chance of being owned by each client. A non-IID distribution, on the other hand, assigns either proteins or drugs to specific clients, and these clients then own all experiments that contain said protein or drug. In the real world, these would be the laboratories looking for drugs targeting a specific protein or investigating the effects of a specific drug.

We obtained each row of each map in Figure 1 by first assigning to each client all experiments corresponding to an exclusive collection of either proteins or drugs. Then, at each step along the continuum, we let the clients exchange some of their data with their neighbours. This exchange follows a Gaussian curve, so we introduce an uneven representation of each data class outside its assigned client. This choice makes the distribution more realistic since it is unlikely that all clients but one would hold the same amount of data in any given class. We achieve the desired mix of protein- and drug-centric clients for the protein and drug experiments by splitting the data into two sub-datasets and then treating each as a separate one-class non-IID experiment. This scenario is closest to what we can expect in the real world. Each square in the figure reports the average over ten training iterations of the given model's loss performance relative to the centralized case. The client counts presented in these figures reflect the cross-silo setup of this domain.

Figures 1a, 1b, and 1c show the heat maps exploring the IID-ness space along the protein, drug, and both dimensions, respectively. As expected, having a higher client count hurts the performance at all non-IID-ness levels. That is, the more fragmented the dataset is, the more challenging the task of aggregating it, as the larger client count implies fewer data per client in this setup, which hurts the individual client models. The different levels of IID-ness, however, do not appear to have a link to the model's performance. In other words, while we see a general trend towards worse performance in each column, we do not see any such trend in the rows. This property is exciting, as it implies that it does not matter whether all client labs test the same combination of proteins or if each client has their own or substantially similar portfolio. It also means that what is a significant drain on FL's robustness in other domains is not a factor in the DTI domain.

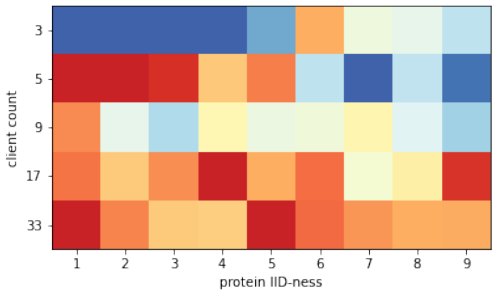
In summary, unlike in other domains in which Federated Learning has been investigated, in the Drug-Target interaction, due to its unique data structure, the input IID-ness does not play a significant role, making the domain singularly unique among FL domains. This observation is crucial as resilience to non-IID data distribution is usually the chief robustness metric for comparing different aggregation strategies in FL. With the data distribution eliminated as a major limitation to our implementation's robustness, we turn to data quantity distribution, i.e. uneven data ownership, as the next candidate for a significant performance driver.

Data distribution imbalance plays a major role in Federated Learning's performance at DTI. Data distribution imbalance and unevenness in the data quantity among clients are of particular concern in the DTI domain, as the participant landscape is composed of a hodgepodge of big and small entities. The often-made assumption that clients have access to about the same amount of data, while plausible in some domains, is contrary to the structure of the pharmaceutical industry. Moreover, when this assumption is relaxed, it is argued that exploiting client size will speed up the training process, while data quantity distribution among the clients will ultimately not impact the model performance [20]. Figure 2 challenges this assumption and examines data quantity distribution's impact on the model performance under varying client counts.

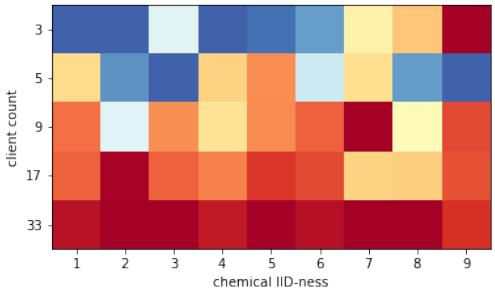
Figure 2a investigates the interplay between client count and data quantity distribution profile. The dataset is distributed among multiple clients. The same single client is designated as the dominant client and receives a variable percentage of the data. The rest of the data is distributed unevenly among the rest of the clients following the Gaussian curve. This is done to achieve a reasonable uneven distribution in line with our approach exposed in the previous subsection.

As before, increasing the client count makes the problem harder, increasing the error. This time, however, the rate of performance deterioration depends on the unevenness of data allocation among the clients. At each client count, irrespective of the ownership inequality level, it holds that moving to a more concentrated data ownership favours the model's performance. This effect is significant throughout the tested conditions but grows stronger the closer the tested setup is to the highly centralized data ownership.

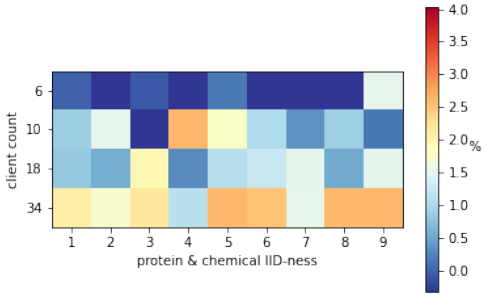
Crucially, the co-dependent effect is not only present in the overwhelmingly dominant client case (far left), where it could be



(a) Protein-based IID-ness variation.



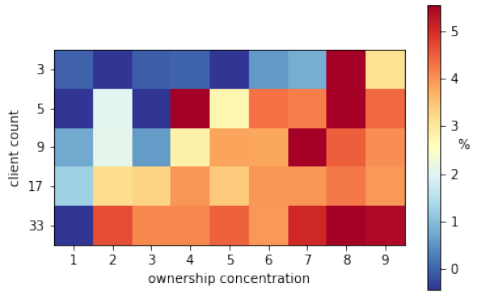
(b) Chemical-based IID-ness variation.



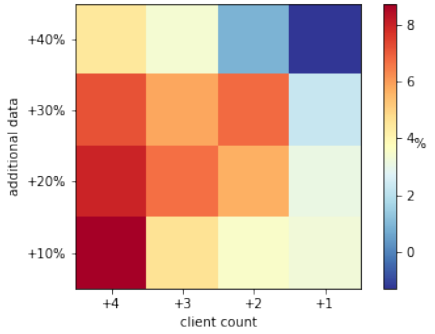
(c) Combined IID-ness variation.

Figure 1: A % change in MSE relative to the smallest client count and highest concentration in each setup is reported for a broad spectrum of client counts against (a) protein-based, (b) chemical-based, (c) protein- and chemical-based IID-ness variation. The horizontal axis represents the two extrema and seven equidistant points between them, vertical represents client count.

discounted as a case of mode collapse into a pseudo-centralized setup, but it holds throughout the tested conditions. This persistence makes our observation particularly salient. There is a cost to having a diluted client data ownership structure. Our next step is to investigate the interplay of this cost with the benefit of adding new data.



(a) Data quantity IID-ness variation.



(b) Data quantity IID-ness variation over different numbers of clients.

Figure 2: a): A % change in MSE relative to the smallest client count and the highest concentration is reported for a selection of client counts and a range of data quantity distributions sampled equidistantly. b): A % change in MSE relative to training solely based on the dominant client's (60% of the) data is reported for the combinations of adding up to 40% of extra data in increments of 10% and divided among 1 to 4 additional clients.

Figure 2b investigates the trade-off between the benefit of adding more data to an existing federation and the cost resulting from increasing the client count and thus diluting the client data ownership structure. We start with a single client allocated a 60% share of the data. Without the loss of generality, this can represent a preexisting federation of clients. The remaining 40% of the dataset is available for addition. The heat map reports the error implications from adding this data in increments of 10% distributed among 1 to 4 clients.

Predictably, increasing the amount of additional data and spreading this data among fewer clients improve model performance in Figure 2b. What is less predictable is that the rate of improvement is about the same in both of these dimensions, which is indeed remarkable. In the tested situation, increasing the concentration of data

ownership can, in some cases, have as strong a positive effect on the model's performance as adding 10% of the data. Consequently, we see that the benefit of additional data can be substantially offset by the cost due to the changed data ownership distribution. The symptom of this is that the top left to bottom right diagonal, where the forces work against each other, varies much less than the bottom left to top right diagonal, where they reinforce each other. The strength of this effect, and in particular its potential to overturn the benefits of substantial dataset increases, suggests questions beyond this paper's scope. Nevertheless, they are significant as they call for a re-think of our view of data imbalance as a mere convergence speed issue. The leveraging of this observation and its use in the design of superior aggregation strategies is left as future work.

4 CONCLUSION

This study delivered a privacy-preserving distributed learning implementation that both meets the limiting constraints of the industry's regulatory and commercial constraints and outperforms previously available alternatives by up to 15%. Furthermore, due to its unique data structure, our investigation demonstrated FL in DTI as the first identified data distribution-agnostic domain. Finally, we identified a material trade-off between the benefits of adding new data and the cost of introducing more clients. This observation is of particular relevance as it breaks the generally accepted maxim that more data is always better and thus motivates the need for further exploration to design superior federated learning algorithms.

ACKNOWLEDGMENTS

This work was supported by the UK's EPSRC with the OPERA (EP/R018677/1) and the MOA (EP/S001530/1) projects; the ERC via the REDIAL project (805194); the Project HPC-EUROPA3 (INFRAIA-2016-1-730897), with the support of the EC Research Innovation Action under the H2020 Programme; the European PILOT project via the EuroHPC JU (101034126); the CINECA award under the ISCRA initiative; the computer resources and technical support provided by ICHEC; the Spoke "FutureHPC & BigData" of the ICSC – Centro Nazionale di Ricerca in "High Performance Computing, Big Data and Quantum Computing", funded by European Union – NextGenerationEU.

REFERENCES

[1] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Titouan Parcollet, Pedro PB de Gusmão, and Nicholas D Lane. 2020. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390* (2020).

[2] Leo Breiman. 1996. Bagging predictors. *Machine learning* 24, 2 (1996), 123–140.

[3] Shaoqi Chen, Dongyu Xue, Guohui Chuai, Qiang Yang, and Qi Liu. 2021. FL-QSAR: a federated learning-based QSAR prototype for collaborative drug discovery. *Bioinformatics* 36, 22-23 (2021), 5492–5498.

[4] Pantelis Elinas, Edwin V Bonilla, and Louis Tiao. 2020. Variational inference for graph convolutional networks in the absence of graph data and adversarial settings. *Advances in Neural Information Processing Systems* 33 (2020), 18648–18660.

[5] Brian Hie, Hyunghoon Cho, and Bonnie Berger. 2018. Realizing private and practical pharmacological collaboration. *Science* 362, 6412 (2018), 347–350.

[6] Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. 2020. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics* 36, 22-23 (2020), 5545–5547.

[7] Madhura Joshi, Ankit Pal, and Malaikannan Sankarasubbu. 2022. Federated Learning for Healthcare Domain-Pipeline, Applications and Challenges. *ACM Transactions on Computing for Healthcare* (2022).

[8] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning* 14, 1–2 (2021), 1–210.

[9] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine* 37, 3 (2020), 50–60.

[10] Rong Ma, Yi Li, Chenxing Li, Fangping Wan, Hailin Hu, Wei Xu, and Jianyang Zeng. 2020. Secure multiparty computation for privacy-preserving drug discovery. *Bioinformatics* 36, 9 (2020), 2872–2880.

[11] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics AISTATS (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Xiaojin (Jerry) Zhu (Eds.). PMLR, Fort Lauderdale, FL, USA, 1273–1282.

[12] Dinh C Nguyen, Quoc-Viet Pham, Pubudu N Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia Dobre, and Won-Joo Hwang. 2022. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–37.

[13] Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. 2021. GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics* 37, 8 (2021), 1140–1147.

[14] Tuan Nguyen, Giang TT Nguyen, Thin Nguyen, and Duc-Hau Le. 2021. Graph convolutional networks for drug response prediction. *IEEE/ACM transactions on computational biology and bioinformatics* 19, 1 (2021), 146–154.

[15] Bjarne Pfitzner, Nico Steckhan, and Bert Arnrich. 2021. Federated learning in a medical context: a systematic literature review. *ACM Transactions on Internet Technology (TOIT)* 21, 2 (2021), 1–31.

[16] Mirko Polato, Roberto Esposito, and Marco Aldinucci. 2022. Boosting the Federation: Cross-Silo Federated Learning without Gradient Descent. *2022 International Joint Conference on Neural Networks (IJCNN)* (18-23 July 2022), 1–10. https://doi.org/10.1109/IJCNN55064.2022.9892284.

[17] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. 2014. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling* 54, 3 (2014), 735–743.

[18] Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* 25, 1 (2019), 44–56.

[19] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. 2020. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security* 15 (2020), 3454–3469.

[20] Hongda Wu and Ping Wang. 2021. Fast-convergent federated learning with adaptive weighting. *IEEE Transactions on Cognitive Communications and Networking* 7, 4 (2021), 1078–1088.