A Federated Learning Benchmark for Drug-Target Interaction

Filip Svoboda[†], Gianluca Mittone[◇], Nicholas D. Lane[†], Pietro Lio'[†]

[†] University of Cambridge [◇] University of Turin

Abstract

Aggregating pharmaceutical data in the drug-target interaction (DTI) domain has the potential to 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 we argue to be reconcilable with the industry's constraints, as it does not require sharing of any information that would reveal the entities' data or any other high-level summary of it. 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.

1 Introduction

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

This paper (1) delivers the first-ever Federated Learning benchmark for the DTI task; (2) achieves up to a 15.53% reduction in error compared to the best available alternative; (3) develops a novel comprehensive analysis framework for FL applications. This let us to 3.a) 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 3.b) to 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 an important and under-explored data domain.

The scope of this paper is, due to computational and other practical considerations, limited to the drug-target interaction DTI task of the the drug discovery domain. 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 to a substantial degree the backbone of many current state of the art models [6, 14, 4], while still being reasonably computationally expensive to permit us running each of the many dozens of federated learning runs each taking approximately one GPU hour on a NVIDIA A40 or 7-8 on GTX-1080 for a total of

Machine Learning for Structural Biology Workshop (MLSB), NeurIPS 2022, New Orleans, USA.

197 or 1,576 hours in total respectively. In our experiments we aim to represent as realistically as possible the complexities a realistic federation of pharmaceutical labs would entail. In particular, we deliberately chose to explore the full spectra of IID-ness and data ownership distribution. Finally, we only perform our experiments using the core algorithms in FL and distributed learning. We consider this choice to be without the 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.

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 when split among multiple clients.

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

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

The GraphDTA model regresses the drug-target pair onto the 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 directly capture the bonds among atoms [13]. In order to stay true to the simple is better ethos of this paper we refrained from implementing fancy aggregation of internal state-dependent features of the model. Instead, we replaced them with suitable alternatives, namely batch norm with Layer norm and ADAM with simple SGD. 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 which combines local stochastic gradient descent (SGD) on each client with a server that performs model averaging [11].

The experimental setup builds on and substantially expands on the set of experiments usually associated with Federated Learning benchmarks. First, the model is compared against a suitable alternative. Given the lack of prior work, there was no ready candidate for this comparison. A centralized model is not a suitable candidate since its use is unrealistic due to aforementioned regulatory and commercial considerations. The cryptographic approaches to data anonymization would be usable in real life. They are, however, not a direct competitor to Federated Learning, and FL is not a direct competitor to them. They can augment each other and provide joint solutions similar to what Federated learning with differential privacy does [19]. In the end, we chose to compare against the simple Bergman's ensemble [2] of models trained separately on the same data splits the FL was used on. 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]. Working with the plain implementations, therefore, should provide us with a fair, and uncoloured comparison of the two approaches, rather than of their two randomly chosen extensions.

The code was packaged and made available on GitHub [1]. It can be used out of the box without any knowledge of distributed or Federated learning. At present it works with PyTorch deep models, but eventually it will be made compatible with TensorFlow too. 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. This reports the performance difference between the federation of deep model architectures and an ensemble of

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

the same architecture. Based on our experimental setup (section 2) all experiments in this section are based on the training of the same GraphDTA [13] architecture and we consider our model's performance to be a success if it is able to match that of the non-private distributed alternative.

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

Table 1: Performance of our DTI-FL relative to ensemble alternative [2]

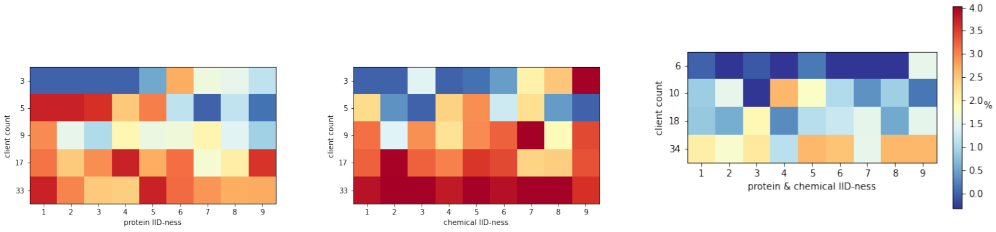
The results in table 1 show that our approach is able to retain up to 15% better performance relative to the distributed alternative, while at the same ensuring that no data or any other high-level summary of it is revealed [1]. The general trend in the IID results points to a relative advantage for the ensembles at very low client counts that quickly dissipates, turns into parity, and from 16 clients up fully reverses as client count increases. Second, the non-IID data display effective parity practically at all client counts, indicating that FL can deal with un-equal data distributions much better than the distributed alternative. This matched performance, alongside FL's very strong privacy and security guarantees [1], which are 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 and in particular, seeing that the IID and non-IID performances are effectively matched, we ask how does the FL performance develop under varying non-IID conditions in the next 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 either investigate the distribution one dimension at a time, that is either non-IID with respect to the protein or chemical inputs, or we can explore it in both dimensions at the same time. Neither of these three approaches can be ruled out 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.

In our experiments we investigate the full continuum of IID-ness, rather than just its two extrema. An IID data distribution is a random draw, that is one in which 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 drug to specific clients, and these clients then own all experiments that contain said protein or drug. In real world these would respectively be the labs that are looking for drugs targeting a specific protein or those that investigate 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 a portion of their data with their neighbours. This exchange follows a Gaussian curve - so that we introduce uneven representation of each data class outside its assigned client. This is to make the distribution more realistic since it is unlikely that all clients but one would hold the same amount of data in any given class. For the protein and drug experiments we achieve the desired mix of protein- and drug-centric clients by first 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 real world. Each square in the figure reports an average over 10 training iterations of the given model's loss performance relative to the centralized case. The client counts presented in these figures were chosen to, both, reflect the cross-silo setup of this domain and to best use the limited data available to us.

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 higher client count hurts the performance at all non-IID-ness levels. That is, the more fragmented is the dataset, the harder is the task of aggregating it, as the larger client count implies in this setup fewer data per client, 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 is exciting, as it implies that it does not matter whether the same combination of proteins is tested by all client labs, or if each client has their own or substantially similar portfolio. It also means that what is a major drain on FL's robustness in other domains, is not a factor in the DTI domain.



(a) Protein-based non-IID-ness. (b) Chemical-based non-IID-ness. (c) Combined non-IID-ness.

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 non-IID-ness, (b) chemical-based non-IID-ness, (c) protein and chamical-based non-IID-ness. Horizontal axis represents the two extrema and 7 equidistant points between them, vertical represents client count.

In summary, we can conclude that 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 major role, making the domain singularly unique among FL domains. This is particularly important observation as the resilience to non-IID data distribution is usually the chief robustness metric used to compare 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 major performance driver.

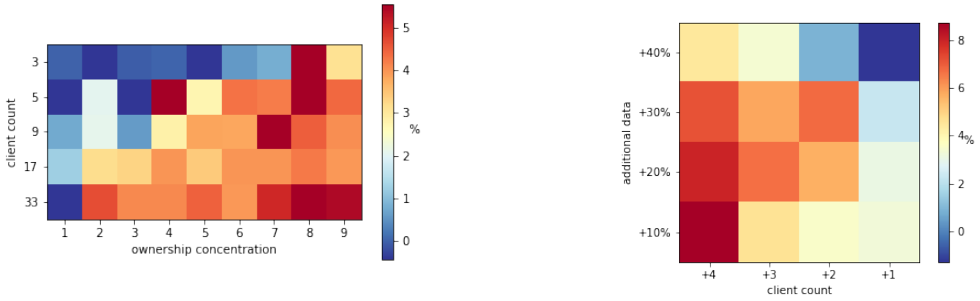
Data distribution imbalance plays a major role in Federated Learning's performance at DTI. The the role data distribution imbalance, unevenness in the data quantity among clients, plays is of particular concern in the DTI domain as the participant landscape is composed of a hodgepodge of big and small entities. The silent, but 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 let us speed up the training process, but ultimately data quantity distribution among the clients will have no impact on 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 variable percentage of the data. The rest of the data is distributed among the rest of the clients in an uneven fashion following the Gaussian curve. This is done to achieve a reasonable uneven distribution in line with our approach in the previous subsection.

As before, increasing client count makes the problem harder, and so the error increases. 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 true 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 extremely centralized data ownership.

Crucially, the co-dependent effect is not only present in the overwhelmingly dominant client case (far left), where it could be 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. It suggests that there is a cost to having diluted client data ownership structure. Our next step is to investigate the interplay of this cost with the benefit of adding new data.

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. This is allocated a 60% share of the data. Without the loss of generality, this can be seen as representing a preexisting model 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.



(a)

(b)

Figure 2: a): A % change in MSE relative to the smallest client count and 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 on the basis of 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.

Predictably, increasing the the amount of of additional data, as well as spreading this data among fewer clients both improve 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 and that is indeed very remarkable. It means that in the tested situation increasing the concentration of data ownership can in some cases have as strong 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, the one where the forces work against each other, varies much less than the bottom left to top right diagonal, the one where they reinforce each other. The strength of this effect, and in particular it potential to overturn the benefits of substantial dataset increases, suggests questions that go beyond the scope of this paper. They are, nevertheless very important, as they call for a re-thing 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, our investigation demonstrated FL in DTI to be the first identified data distribution-agnostic domain due to the its unique data structure. Finally, we identified a material trade-off between the benefits of adding new data, and the cost of introducing more clients. This is of particular relevance as it breaks the generally accepted maxim that more data is always better and thus motivates the need for its further exploration for the purposes of designing superior federated learning algorithms.

Acknowledgements

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; and the computer resources and technical support provided by ICHEC.

References

[1] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, P. P. de Gusmão, and N. D. Lane. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020.

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

[3] S. Chen, D. Xue, G. Chuai, Q. Yang, and Q. Liu. Fl-qsar: a federated learning-based qsar prototype for collaborative drug discovery. *Bioinformatics*, 36(22-23):5492–5498, 2021.

[4] P. Elinas, E. V. Bonilla, and L. Tiao. Variational inference for graph convolutional networks in the absence of graph data and adversarial settings. *Advances in Neural Information Processing Systems*, 33:18648–18660, 2020.

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

[6] K. Huang, T. Fu, L. M. Glass, M. Zitnik, C. Xiao, and J. Sun. Deeppurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23):5545–5547, 2020.

[7] M. Joshi, A. Pal, and M. Sankarasubbu. Federated learning for healthcare domain-pipeline, applications and challenges. *ACM Transactions on Computing for Healthcare*, 2022.

[8] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[9] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[10] R. Ma, Y. Li, C. Li, F. Wan, H. Hu, W. Xu, and J. Zeng. Secure multiparty computation for privacy-preserving drug discovery. *Bioinformatics*, 36(9):2872–2880, 2020.

[11] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas. Communication-efficient learning of deep networks from decentralized data. 2016.

[12] D. C. Nguyen, Q.-V. Pham, P. N. Pathirana, M. Ding, A. Seneviratne, Z. Lin, O. Dobre, and W.-J. Hwang. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(3):1–37, 2022.

[13] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh. Graphdta: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.

[14] T. Nguyen, G. T. Nguyen, T. Nguyen, and D.-H. Le. Graph convolutional networks for drug response prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(1):146–154, 2021.

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

[16] M. Polato, R. Esposito, and M. Aldinucci. Boosting the federation: Cross-silo federated learning without gradient descent. 2022.

[17] J. Tang, A. Szwajda, S. Shakyawar, T. Xu, P. Hintsanen, K. Wennerberg, and T. Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, 2014.

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

[19] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

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