Optimising Scientific Collaboration with Federated Learning

FELIX O'BRIEN, Australian National University, Australia
LACHLAN STEWART, Australian National University, Australia
ERIC WANG, Australian National University, Australia

It is has been recognised by the materials science community for years that sharing data digitally paves a path to faster progress [9]. Also of relevance is the success of Machine Learning at predicting material properties for materials science research [10]. At the intersection of these facts lies a unique opportunity for Federated Learning. Of interest to collaborators in the wider scientific community, therefore, is the effect of different kinds of data sharing on the performance of models trained under contemporary Federated Learning pipelines. In this study, we apply such a pipeline, FedRED [5], and simulate different data privacy constraints through dataset segmentation. By doing this, our study makes the case that sample-space sharing is the most effective scientific collaboration strategy for Federated Learning. We fortify our results by exploring a wide range of parameters and models, and by using model regularisation. We also improve our results by exploring an alternative to standard data preprocessing practices which are prohibited in Federated Learning.

Additional Key Words and Phrases: Federated Learning, Materials Science, Scientific Collaboration, Regression

ACM Reference Format:
Felix O'Brien, Lachlan Stewart, and Eric Wang. 2025. Optimising Scientific Collaboration with Federated Learning. 1, 1 (November 2025), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Barriers to collaboration directly hinder scientific progress. One such barrier is the availability of data, which is essential to the foundation and legitimacy of any scientific endeavour. More data *exists* than ever before, but increasing data privacy concerns limit its availability for traditional Machine Learning (ML) applications at scale.

In general, data privacy concerns are varied and can be affirmed by law [4, 12]. They stand to protect the property/rights of the originators, and individuals whose information is included in the data. In the context of scientific collaboration, collaborators are concerned that the exclusivity of the proprietary, sometimes costly, sourcing of their data will be compromised. This is because, for traditional ML, it is necessary to have a unified sample and feature space represented by a single shared dataset. In such situations, collaborators are forced to train models on their data only. The fact this outcome is suboptimal is not a topic of debate; it is common knowledge within the ML community that the reliability and insight of models is inextricably dependent on the quality, quantity and diversity of the data they are trained on. Federated Learning (FL) offers an alternative where a model can be trained on the data of all collaborators, without the need for sharing. This presents an opportunity for FL to generate great value for the scientific community.

Authors' addresses: Felix O'Brien, u7490752@anu.edu.au, Australian National University, Canberra, Australian Capital Territory, Australia, 2601; Lachlan Stewart, u7284324@anu.edu.au, Australian National University, Canberra, Australian Capital Territory, Australia, 2601; Eric Wang, u7507270@anu.edu.au, Australian National University, Canberra, Australian Capital Territory, Australia, 2601.

XXXX-XXXX/2025/11-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Still, FL has to deal with the complications caused by having a distributed and divided dataset; some degree of data sharing will lead to better results. In some cases, collaborators may have the freedom to share some aspects of their data. Our study explores the different ways that collaborators can share *some* of their data to improve the performance of their models using FL.

As a word on terminology, we advise the reader that we represent both the wider Machine Learning field and the context where data privacy concerns do not constrain the training process by the words '*traditional ML*'. We represent the Federated Learning subfield and the context where data privacy concerns do impose constraints by '*FL*'.

Sharing data in different ways can lead to very different FL settings. These settings are categorized generally in the FL literature as horizontal, vertical, and hybrid [12], distinguished by the different interactions of the data that is shared (Figure 1). We simulate the effects of sharing material samples and/or scientific instruments in the context of materials science. We find that the contemporary FL pipeline FedRED [5] is applicable to all three FL settings. Informed by studies like ours, scientists can determine if and how they collaborate with others, with the assistance of novel FL paradigms.

Our study is restricted to materials science by our choice of dataset. Here we attend to the fact that the motivations behind FL are highly relevant in other fields also. In medicine, ML approaches often meet or exceed human performance and automate previously time-consuming tasks [1]. There, ethical and commercial concerns regarding data reduce the applicability of ML and halt progress. Sensitive information such as patient data and intellectual property, on which restrictions exist, are either incompatible with ML approaches or require significant work to be compatible; Information agreements and other legal processes are often required [8]. To introduce another example, this is also regularly the case at the intersection of industry and academia. Businesses often have superior ML applications and wish to protect their intellectual property. On the other hand, researches who wish to use these models often face the aforementioned ethical data concerns [8]. Regardless, collaboration between any parties which have obligations to their data is made more difficult.

1.1 Federated Learning

The prototypical Federated Learning methodology [7], consists of the following core loop:

1 Perform model training via Stochastic Gradient Descent separately on the local data of each collaborator (within the aforementioned privacy barriers)
2 Aggregate each model via a 'Federator' centrally into one global model, sending only the respective models to do so.
3 Redistribute the updated global model to the collaborators

In following this methodology, the central model is exposed to the data of all collaborators without any data ever being shared. From these initial explorations there have been efforts to improve the core aggregation method resulting in various Federator alternatives such as FedRED (Section 1.2).

The performance of Federated learning has been well underscored in literature. In an evaluation of FL using the popular Fashion MNIST Benchmark, the FL approach was shown to have an accuracy of 85.15%, only marginally lower than 86.23% for a distributed approach and 87% for a traditional (single full dataset) approach [2]. This is whilst ignoring the performance benefit in train time that the distributed training required for FL provides [2]. Federated approaches using sensitive patient data were shown to be 99% as effective compared to traditional alternatives [11]. Furthermore, errors that may be introduced by a Federated approach have been shown to be outweighed by increased access to data that comes as a result of multi institution collaborations [11].

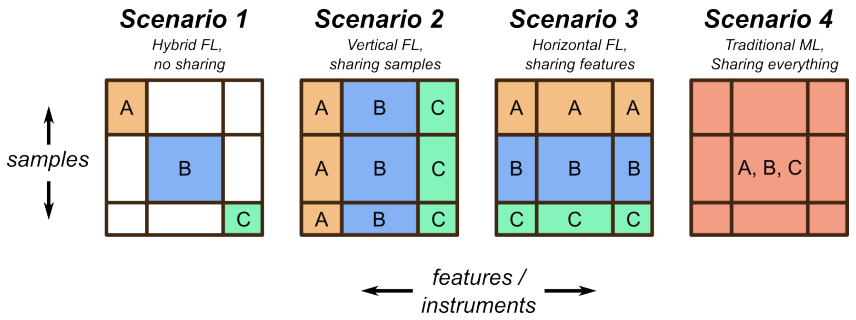

Fig. 1. The three FL settings, and the traditional whole-dataset setting, are termed 'scenarios' in our study

1.2 FedRED

The 'curse of dimensionality'[1] has specific significance in FL; having too many features across the collaborators means that it is more likely that a collaborator will under-represent at least one feature. Adding complication is that vertical and hybrid FL settings deal with collaborators that have unmatched feature spaces, both in dimension and shape. FedRED [5] presented the utility of Principal Component Analysis (PCA), a quintessential deterministic feature extraction method, in FL. PCA projects feature vectors to an orthogonal basis which explains the maximum amount of variance in the data. The key idea of FedRED consists of a three step process. First, the collaborators are instructed to perform PCA individually, resulting in projection matrices $U_1, ..., U_n$. Second, a convex combination U of the projection matrices of the collaborators is agreed upon via the Federator. Third, U is communicated back to the collaborators, and they each transform their datasets $X_1, ..., X_n$ using this:

$$\tilde{X}_i = X_i U, \quad \text{where:} \quad U = \sum_{i=1}^{n} \lambda_i U_i \quad \text{and} \quad \sum_{i=1}^{n} \lambda_i = 1, \quad \lambda_i \geq 0 \tag{1}$$

This provides a means of reducing the dimensionality of the data and also unifying the different distributions that each collaborator might have represented in their data.

2 METHODOLOGY

2.1 Dataset

To assess the efficacy of Federated Learning in situations with different levels of collaboration, our dataset of choice has been partitioned in the feature (column/instrument) and sample (row) dimensions, as illustrated in (Figure 1). From these partitions we derive four scenarios, each representing a different collaborative method and a different FL setting.

In Scenario 1, the collaborators do not share any of their samples nor their features. In Scenario 2 samples are shared across all collaborators, but instruments or features are associated with individual collaborators whilst Scenario 3 sees exchange of instruments or features across collaborators. In Scenario 4, collaborators share both the sample space and feature space. This represents traditional ML and collaboration without constraints caused by privacy concerns. The performance of the models trained in Scenario 4 is expected to be better than the performance of models from any other scenario, due to the lack of constraints, and serves as a baseline.

All data originates from several publicly available datasets [5]. Further information can be found in Appendix A.

[1] first termed by Richard. E. Bellman, this phrase refers to issues caused by having high-dimensional feature spaces

Specifically, our dataset is typical of materials science investigation with each of our collaborators given real roles and likely samples and features based on those roles. The breakdown of their profession, samples, and measurements can be found below.

Collaborator A	Collaborator B	Collaborator C
Phycicist	Chemist	Materials Engineer
Graphene Oxide Nanoflakes	Graphene Oxide Quantum Dot	Graphene Oxide Nanoflakes
STEM and AFM Imaging	IR, FTIR, XPS and XAS Spectroscopy	Raman Spectroscopy

2.2 Targets

The ML task we train our models on is a regression task. In the studied dataset, there are two targets, Fermi Energy and Thermodynamic Probability (Fermi and Thermo). These targets represent properties of a material that materials scientists wish to predict based on other measurements.

There are two available regression tasks here. One can choose single target regression for each target independently, resulting in two models, or one can choose to predict both targets simultaneously, resulting in one model. In traditional ML, it is a good decision to predict both simultaneously. This requires the maintenance of only one model. Additionally, learning to do multiple tasks at once has been shown to lead to more capable models, and the wider ML community is aware of this effect [3].

We have found that the choice to predict two targets simultaneously introduces new complications in the setting of FL, as the data privacy constraints prohibit conventional data standardization techniques prior to training. It is standard practice for regression tasks to transform the targets so that they are a mean centered with identity covariance. In FL, these conventional processes require privacy-compromising communication amongst the collaborators. For example, the number of samples belonging to each collaborator would need to be shared, at least.

A lack of standardization is problematic for multiple target regression. Model errors in the direction of one target may carry an unfair amount of weight compared to a different target (relative to the variance of the targets in their axes). There is generally no contextual reason to treat one target with more importance than the other, so this is an issue. In the studied data, we note that the variance in the Fermi target is much greater than the variance of the Thermo target by approximately 30x (Figure 2). Without a substitute for standardization, the model will be biased to predict more accurately (relative to the variances) on Fermi than Thermo. This exacerbates an issue already inherent in this data, which is the Thermo target has a heavy skew towards 0.00.

See 2.4 for our proposed solutions to this issue.

2.3 FedRED PCA revisited

In Section 1.2 we gave a brief explanation of the PCA aggregation strategy of FedRED. Here, we note additional consequences of this idea.

We note that if each collaborator had performed a full PCA transformation to their individual data beforehand, then the matrices $U_1, ..., U_N$, and finally U, would be identity matrices. Performing these PCA transformations individually does not compromise data privacy, and does not compromise the information held within each dataset, but it does make this PCA aggregation step (1) redundant; a convex combination of identity matrices gives the identity matrix. Instead, each collaborator can slice their dataset to the first D columns, where D is the minimum number of features required to reach the explanation ratio.

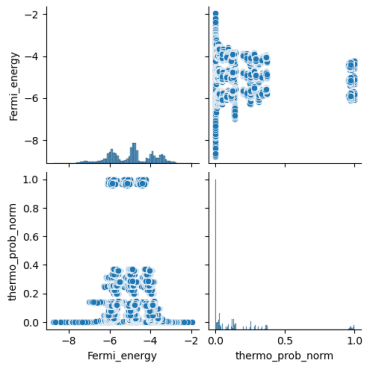

Fig. 2. Distribution of the Thermo and Fermi targets over the full data. Note the scaling of the axes.

In scenarios 1 and 2, the collaborators have different feature space dimensionalities. This is illustrated by the different widths of the highlighted regions in Figure 1. If we attempt to apply the PCA aggregation step proposed in FedRED, then the matrices $U_1, ..., U_N$ will not match in the size of the second dimension, however, if we apply the aforementioned PCA transformations to each client individually and take the first D columns, then the pipeline can proceed.

2.4 Loss functions for training

MSE

Mean Squared Error loss (MSE) is a commonly applied loss function in regression problems. In the case of a Linear Regression model on one dataset, MSE provides a closed form analytic solution, making it particularly applicable to this study. Despite the strengths of MSE loss in traditional ML appplications, we contend that the lack of target standardisation makes this loss function less appropriate.

Stable Covariance-Adjusted Loss (SCAL)

Algorithm 1 SCAL

Input: $(\mathbf{t}, \mathbf{t}' \in \mathbb{R}^{B,T})$
$C \leftarrow \mathbf{cov}(\mathbf{t})$
$C^{-1} \leftarrow \mathbf{inv}(C)$
$\sigma_1^{C^{-1}} \leftarrow \mathbf{2\text{-}Norm}(C^{-1})$
$\Delta\mathbf{t} \leftarrow \mathbf{t} - \mathbf{t}'$
$S \leftarrow C^{-1}/\sigma_1^{C^{-1}}$
$L \leftarrow \Delta\mathbf{t}^T S \Delta\mathbf{t}$
Output: L

We propose an alternative loss function for computing the loss of the model which accounts for potentially skewed target distributions. This loss function is inspired by the Mahalonobis Distance [6]. With a sufficiently large batch size, the covariance matrix of the targets C is often a good approximation of the true covariance. However, it is sometimes the case that the batch of ground truth targets $\mathbf{t}$ lie on the same hyperplane. When this is the case, the *smallest* singular value of C,

$\sigma_{\min}(C)$, is close to zero. This causes the *largest* singular value of C^{-1} to be extremely large:

$$\sigma_{\max}(C^{-1}) = 1/\sigma_{\min}(C)$$

The largest singular value is equal to the matrix 2 norm:

$$||A||_2 = \sup_{x \neq 0} \frac{||Ax||_2}{||x||_2} = \sigma_{\max}(A)$$

By the above we see that $\sigma_{\max}(C^{-1})$ gives an upper bound on the amount that the transformation C^{-1} can increase the magnitude of a vector. For this reason, we divide C^{-1} by $\sigma_{\max}(C^{-1})$ before computing the final loss L. It was observed that without this division operation, training was completely unstable. We note that since $\sigma_{max}(S)$ is 1, this error function never exceeds MSE for the same inputs. We also note that if the covariance of the ground truth targets C is the identity, then this error is equal to the RMSE. This would be the case if a PCA transformation had been applied to the targets beforehand.

2.5 Federated model averaging strategies

A key component of FL is the way that the collaborator's models are averaged to maintain one centralised model throughout communication rounds. We investigate three weighting schemes which were already present in the FedRED pipeline: averaging, sample-size weighting and cross-validation score weighting.

3 FINDINGS AND DISCUSSION

Refer to the appendix for tables containing quantitative results

3.1 Sharing samples is the best collaboration strategy

The results in A.1 demonstrate that Scenario 2, sharing samples, consistently leads to the best model performance; this is true regardless of what underlying FL strategies are applied. We thought it was possible that this was due to Scenario 2 providing the models with more data overall. To eliminate this as a factor, we conducted an experiment where the overall dataset was randomly thinned beforehand to be match the instance size of the other scenarios. The results of A.2 show that the performance remains the same under these conditions. This is evidence that there is something intrinsically beneficial to sharing the sample space, even if features differ.

This result has significant practical ramifications in the context of materials science. It is often much easier to share samples than share features; sharing features requires that collaborators provide each other access to their instruments for collecting measurements. Of added significance is that if collaborators share features, then this access must be sustained during deployment. This is illustrated in 3.

With these observations in mind, the decision to share samples is a win on two fronts.

3.2 Linear Regression outperforming more complex models in FL tasks

The results in A.1 show that the Linear Regression model outperforms the more complex MLP and DNN models in most cases. This is unexpected, as the MLP and DNN models have more expressive power through their ability to handle non-linear relationships in the data. We found that the loss of these models converged, therefore this is not due to a lack of training. It is more likely that these models are exploiting their expressive power to over-fit to the training data. It is the subject of future work to investigate the use of loss functions with regularisation penalties, such as Ridge and Lasso loss.

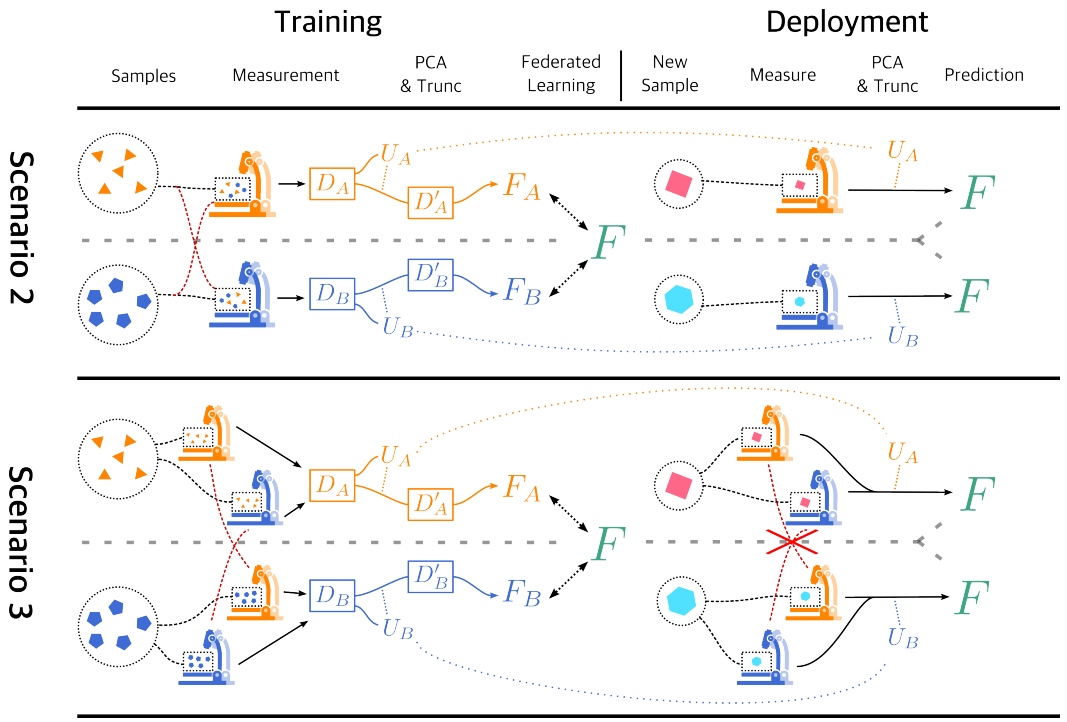

Fig. 3. Visualisation of the FedRED pipeline during training and deployment for Scenarios 2 and 3. Inconvenient sharing during deployment in Scenario 3 is marked with a cross.

3.3 SCAL does not outperform MSE but is feasible

The results show that SCAL does not outperform MSE in terms of the final MSE loss value of the models. Still, our results show that the loss values are highly comparable. This validates the use of SCAL as a loss function for future work in situations where its features as a loss function may be more relevant. It is important to acknowledge that the distribution of the targets in from our dataset is not particularly pathological. In our datset, the skew between each target could be much more severe, and the targets of each collaborator appear to originate from the same distribution. In future work it would be worthwhile to evaluate the use of SCAL where the skew of the targets is more extreme and where the distribution of the targets is not uniform across the different collaborators.

3.4 Aggregated PCs outperforms current FedRed

A.3 has shown that if PC aggregation methods from original FedRed are kept and the optimal aggregation method is chosen which in this case is straight averaging, resulting model slightly outperforms the current FedRed. It is expected as aggregation is able to capture information of distributions from different datasets and thus resulting in a better generalization performance. Hence, we conclude that our version of FedRed compromises some generalization performance, but offers greater flexibility for different FL situations.

3.5 Comparison between FL model and non-FL model

We experimented with applying individual collaborator's models on other datasets to assess their generalisation ability; the results are shown in A.4. These results demonstrate that non-FL models

have a extremely poor generalization performance for other local datasets. On the other hand, in A.1 we observe a consistently good performance for FL-model across all collaborators. Whilst non-FL models perform better for individual collaborators, the loss of generalisation to new and varying samples provides a concrete argument for choosing FL in collaborative settings.

4 LIMITATIONS

Due to the PCA aggregation step in the FedRED pipeline, our study of traditional FL includes scenarios where hybrid or vertical FL approaches may be more appropriate. We do not consider where our study fits in to these categories, and we do not compare the performance of FedRED to appropriate vertical / hybrid FL algorithms.

The collaboration techniques explored in this study are relevant only to materials science and similar domains. For instance, in medicine, the strategy of sharing samples in diagnostic machine learning tasks is analogous to sharing patients which is often impossible. Furthermore, the partitioning of our dataset simulates only three clients. Whilst this is a reasonable number to expect from collaboration in materials science or similar fields, in other FL applications and particularly industry, the number of clients is likely to be much larger and may produce different results. This is whilst ignoring the increased infeasibility that certain scenarios, such as sharing expertise or instruments, bring to the data collection stage of a federated pipeline.
The optimization algorithm used throughout this study is Stochastic Gradient Descent, and no experimentation with hyperparameters such as learning rate or model structure was conducted. Although this simplified our study, these choices are not representative of contemporary ML literature, where more recent optimizers and model architectures are used.

5 CONCLUSION

By evaluating the implementation of Federated Learning across various methods of collaboration we were able to determine that FL can be best applied in situations where the feature space differs across collaborators, but the instance space remains constant. This form of collaboration is arguably the most convenient and realistic in material science and similar fields, in that other collaborative environments demand increased collaboration to provide features across new instances and during inference. We ensure that no particular scenario is significantly harmed by the application of Federated Learning, affirming its ability to perform in varying situations. Ideally, this will encourage researchers in materials science and related fields to consider collaborating with each-other to train models in new situations using new Federated Learning technologies.

ACKNOWLEDGMENTS

Supervisation on this project was provided by Jacob Huang and Amanda Barnard

REFERENCES

[1] Omar Ali, Wiem Abdelbaki, Anup Shrestha, Ersin Elbasi, Mohammad Abdallah Ali Alryalat, and Yogesh K Dwivedi. 2023. A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *Journal of Innovation Knowledge* 8, 1 (2023), 100333. https://doi.org/10.1016/j.jik.2023.100333

[2] Kunal Chandiramani, Dhruv Garg, and N Maheswari. 2019. Performance Analysis of Distributed and Federated Learning Models on Private Data. *Procedia Computer Science* 165 (2019), 349–355. https://doi.org/10.1016/j.procs.2020.01.039 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] European Parliament and Council of the European Union. [n. d.]. Regulation (EU) 2016/679 of the European Parliament and of the Council. https://data.europa.eu/eli/reg/2016/679/oj
[5] W Huang and A S Barnard. 2022. Federated data processing and learning for collaboration in the physical sciences. *Machine Learning: Science and Technology* 3, 4 (dec 2022), 045023. https://doi.org/10.1088/2632-2153/aca87c
[6] Gabriel Martos, Alberto Muñoz, and Javier González. 2013. On the Generalization of the Mahalanobis Distance. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, José Ruiz-Shulcloper and Gabriella Sanniti di Baja (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 125–132.
[7] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2023. Communication-Efficient Learning of Deep Networks from Decentralized Data. arXiv:1602.05629 [cs.LG]
[8] Blake Murdoch. 2021. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Medical Ethics* 22, 1 (15 Sep 2021), 122. https://doi.org/10.1186/s12910-021-00687-3
[9] Brian Puchala, Glenn Tarcea, Emmanuelle A Marquis, Margaret Hedstrom, HV Jagadish, and John E Allison. 2016. The materials commons: a collaboration platform and information repository for the global materials community. *Jom* 68 (2016), 2035–2044.
[10] Marques M.R.G. Botti S. et al. Schmidt, J. 2019. Recent advances and applications of machine learning in solid-state materials science. https://doi.org/10.1038/s41524-019-0221-0
[11] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, and Spyridon Bakas. 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* 10, 1 (July 2020), 12598.
[12] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. *ACM Trans. Intell. Syst. Technol.* 10, 2, Article 12 (jan 2019), 19 pages. https://doi.org/10.1145/3298981

A RESULTS TABLES

A.1 Overall Performance (Mean square error)

Strategy	Scenario	linreg	mlp	dnn	mean loss	best loss	best model
Average (MSE)	Sc1	0.703	1.259	0.765	0.909	0.703	linreg
	Sc2	0.582	0.646	0.694	**0.640**	**0.582**	linreg
	Sc3	0.831	0.824	0.743	0.799	0.743	dnn
Average (SCAL)	Sc1	0.723	0.804	0.789	0.772	0.723	linreg
	Sc2	0.620	0.726	0.739	0.695	0.620	linreg
	Sc3	0.714	0.790	0.771	0.758	0.714	linreg
Size (MSE)	Sc1	0.829	0.780	0.969	0.859	0.780	mlp
	Sc2	0.595	0.698	0.739	0.677	0.595	linreg
	Sc3	0.778	0.839	0.747	0.788	0.747	dnn
Cross-Val (MSE)	Sc1	0.830	1.329	1.113	1.091	0.830	linreg
	Sc2	0.589	0.968	1.063	0.874	0.589	linreg
	Sc3	0.883	0.844	1.061	0.929	0.844	mlp
Full (MSE)	Sc4	0.534	0.521	0.532	0.529	0.521	mlp

Under the strategy column, the unbracketed term denotes the model averaging strategy and the bracketed term denotes the loss function employed.

A.2 Scenario 2 performance with thinned dataset

	linreg	mlp	dnn	mean loss	best loss	best model
Fold 1	0.586	1.290	0.563	0.813	0.563	dnn
Fold 2	0.591	0.692	0.659	0.647	0.591	linreg
Fold 3	0.592	0.639	0.723	0.652	0.592	linreg
Fold 4	0.586	1.530	0.695	0.937	0.586	linreg
Fold 5	0.581	0.803	0.632	0.672	0.581	linreg

A.3 Comparison between aggregated and non-aggregated version of FedRed on scenario 3

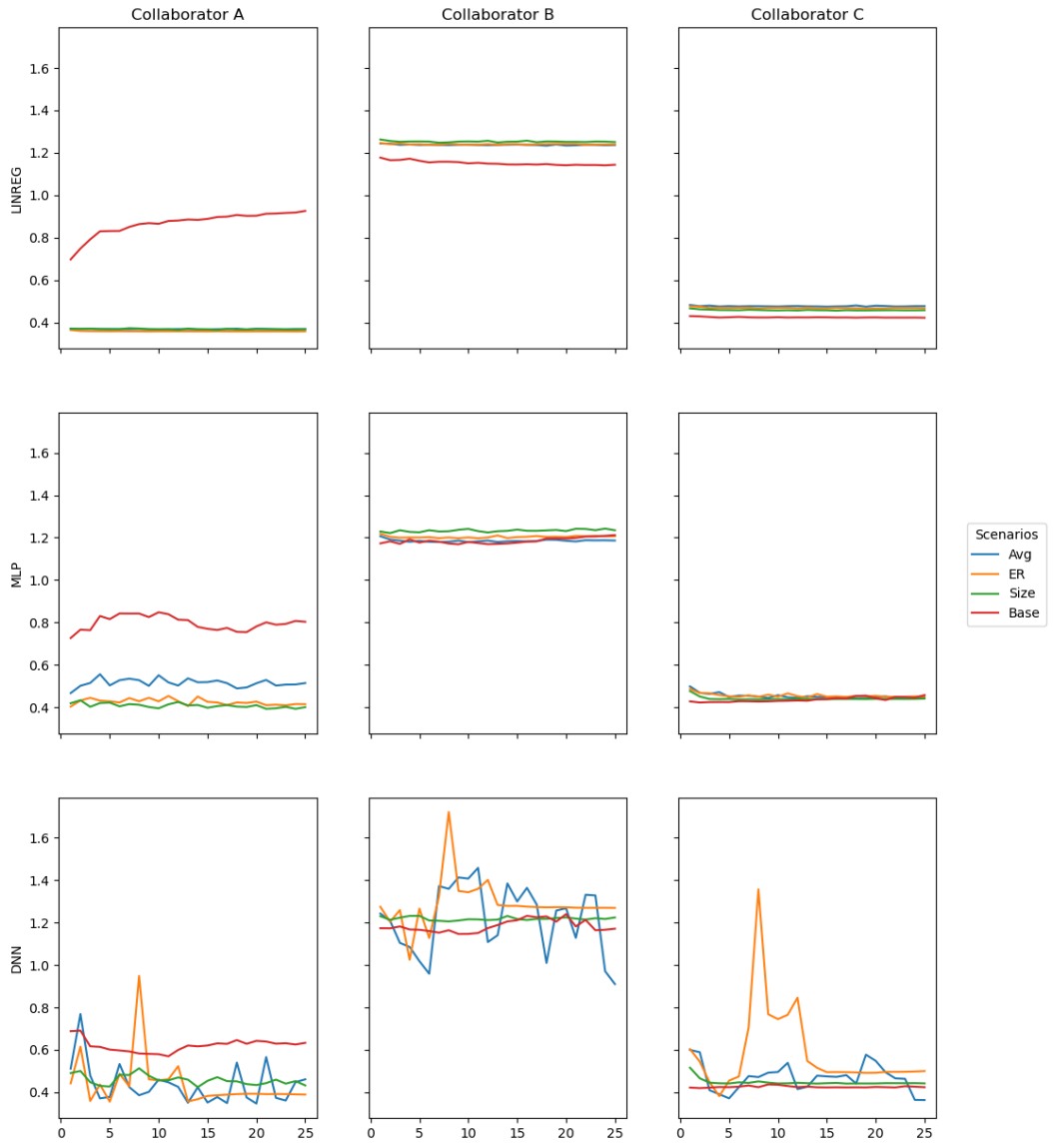

Fig. 4. Caption

Experiment Base is the algorithm we are using throughout the project, while avg, er and size are different aggregation methods for projection matirx which corresponds use straight averaging, explain ratio and size as aggregated weight for projection matrix.

A.4 Generalization performance of non-FL model

Scenario	Collaborator	linreg	mlp	dnn	mean loss	best loss	best model
Sc1	**A**	0.175	0.190	0.186	0.184	0.175	lr
	B	33.71	34.27	33.75	33.91	33.71	lr
	C	31.29	31.43	31.21	31.31	31.21	dnn
	A	31.98	35.14	36.78	34.63	31.98	lr
	B	0.524	0.034	0.029	0.196	0.029	dnn
	C	24.85	26.03	26.38	25.75	24.85	lr
	A	30.53	29.82	31.94	30.76	29.81	mlp
	B	26.14	25.91	26.73	26.26	25.91	mlp
	C	0.420	0.424	0.421	0.422	0.420	lr
Sc2	**A**	0.593	0.618	0.590	0.600	0.590	dnn
	B	0.650	0.661	0.673	0.645	0.650	lr
	C	24.50	25.10	24.63	24.74	24.50	lr
	A	24.78	25.44	27.39	25.87	24.78	lr
	B	0.296	0.071	0.054	0.140	0.054	dnn
	C	24.64	25.12	26.60	25.45	24.64	lr
	A	25.33	24.76	26.21	25.43	24.76	mlp
	B	0.640	0.666	0.624	0.643	0.624	mlp
	C	0.600	0.597	0.597	0.598	0.597	mlp/dnn
Sc3	**A**	0.108	0.034	0.059	0.067	0.034	mlp
	B	33.75	33.97	33.09	33.60	33.09	dnn
	C	31.10	31.33	30.35	30.93	30.35	dnn
	A	30.02	30.56	31.38	30.65	30.02	lr
	B	1.116	1.130	0.935	1.060	0.935	dnn
	C	24.12	23.78	23.70	23.87	23.70	dnn
	A	29.91	29.58	31.77	30.42	29.58	mlp
	B	25.87	25.78	26.82	26.16	25.78	mlp
	C	0.414	0.419	0.416	0.416	0.414	lr

The bold letter in collaborator indicates that the model is trained based on this collaborator's local dataset.

B DATA AVAILABILITY STATEMENT

All data used in this study can be found at the following DOI locations:
http://doi.org/10.25919/5d395ef9a4291;
http://doi.org/10.25919/5e30b5231c669;
http://doi.org/10.25919/5d3958d9bf5f7;
http://doi.org/10.25919/5d3958ee6f239;
http://doi.org/10.25919/5d22d20bc543e.