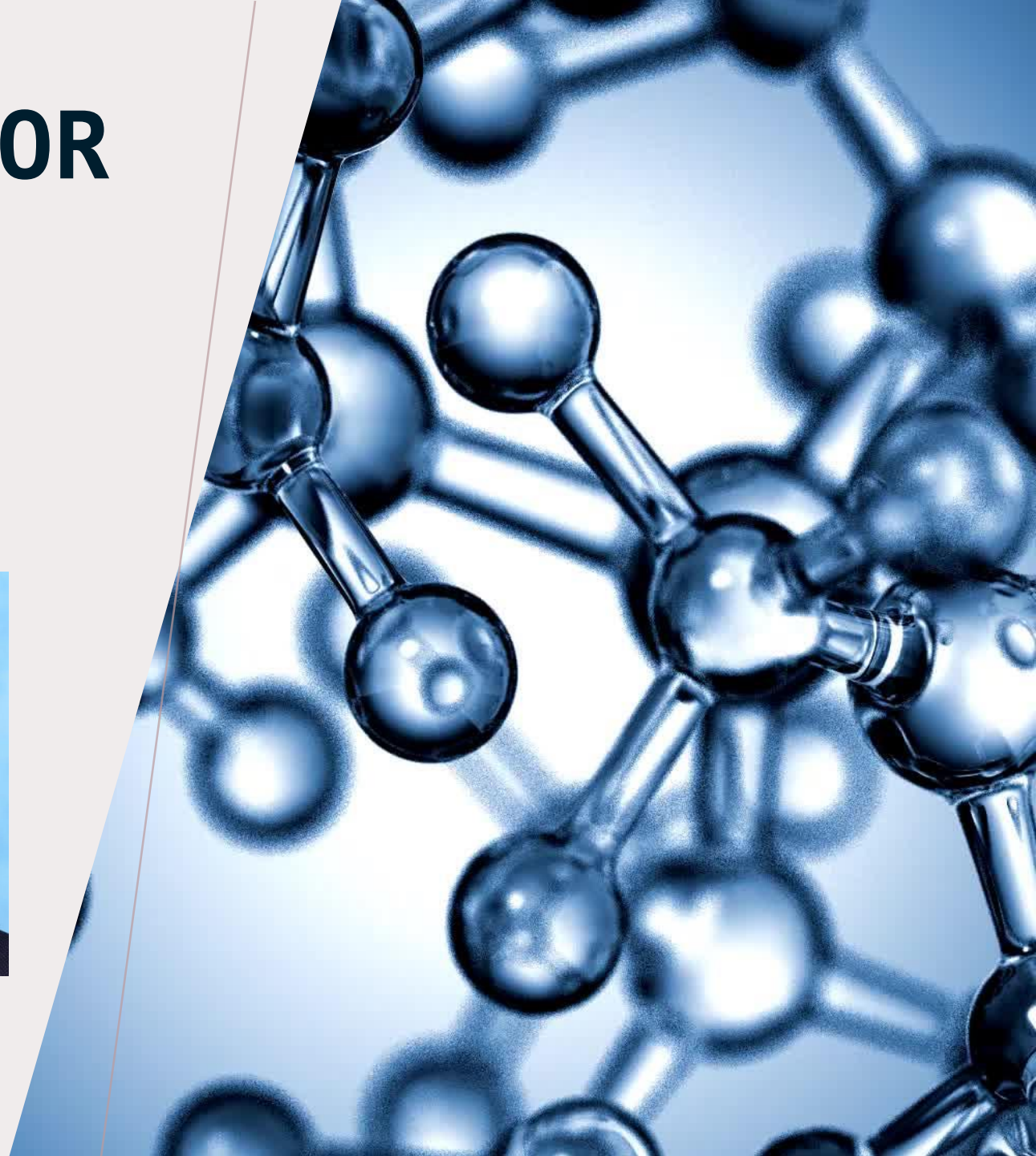
FEDERATED LEARNING FOR MATERIALS PROPERTY PREDICTION

CSCE585 – MACHINE LEARNING SYSTEMS

FINAL PROJECT PRESENTATION

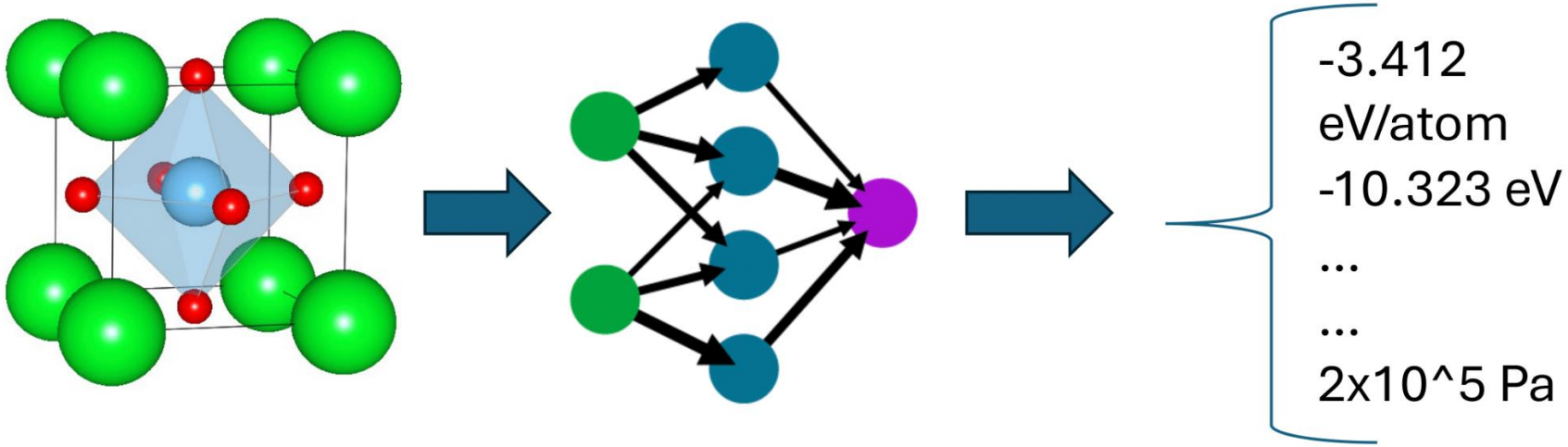
Project member 1: Sadman Sadeed Omee
Ph.D. Candidate (Computer Science)
Role: Machine Learning Researcher

Project member 2: Md Hasibul Amin
Ph.D. Candidate (Computer Engineering)
Role: Machine Learning Researcher



THE PROBLEM

- **Problem statement:**
  - Predicting properties (**output**) of materials from its composition or structure (**input**) is a longstanding challenge (ex: formation energy, band gap, etc.).
  - The performance of machine learning (ML) models developed for this problem are largely limited due to the **scarcity of documented data** compared to the colossal possible materials space.
  - We want to perform "**federated learning (FL)"** to existing ML models for materials as a possible solution to this problem.



Input: material structure ML model **Output:** material properties

THE PROBLEM

- **Why is this problem interesting?**
  - **Scientific importance**: Accurate materials property prediction is crucial for accelerating the discovery and design of new materials with desired properties.
  - **Advantage of FL**: FL allows for the aggregation of models trained on distributed datasets across different institutions. This approach effectively enlarges the training data without sharing the raw data, improving model performance and generalization while preserving data privacy.
  - **Project type:** ***No in general FL for materials property prediction exists***, although FL is available for a similar problem called molecular property prediction.

MOTIVATING SCENARIOS

- Multiple companies are synthesizing drugs, vaccines, or synthesizing exceptional materials for commercial usage such as blue led with lower band-gap materials, wire made with high electrical conductivity material, etc.
  - **The challenge is to screen materials with desired properties from a colossal materials space.**
- Using density functional theory (DFT)-based solution is costly and time-consuming.
  - **So, researchers are employing ML for screening materials with desired property with a little accuracy-time trade-off.**
- FL allows an organization to help other organizations to develop their own ML model for property prediction for screen desired property materials (e.g., lower refractive index materials' glass, etc.) **without explicitly giving away the name of the materials they found experimentally and then synthesized in the lab.**

MOTIVATING SCENARIOS

1. **Drug Discovery:** Multiple pharmaceutical companies use FL to predict drug efficacy for novel compounds **while keeping their proprietary drug libraries private.**
2. **Renewable Energy:** Energy companies collaborate through FL to find high-efficiency solar panel materials with optimal electrical conductivity, **without sharing experimental data.**
3. **Vaccine Development:** FL helps pharmaceutical firms predict vaccine efficacy for new viruses, ensuring **confidential vaccine candidate** information is protected.
4. **Consumer Electronics:** Companies use FL to discover low refractive index materials for high-performance optical displays **without disclosing their unique material databases.**
5. **Finding Superconducting Materials:** Research labs collaborate via FL to identify materials with high superconductivity at room temperature, **while keeping the specific materials they test confidential.**

RELATED WORKS

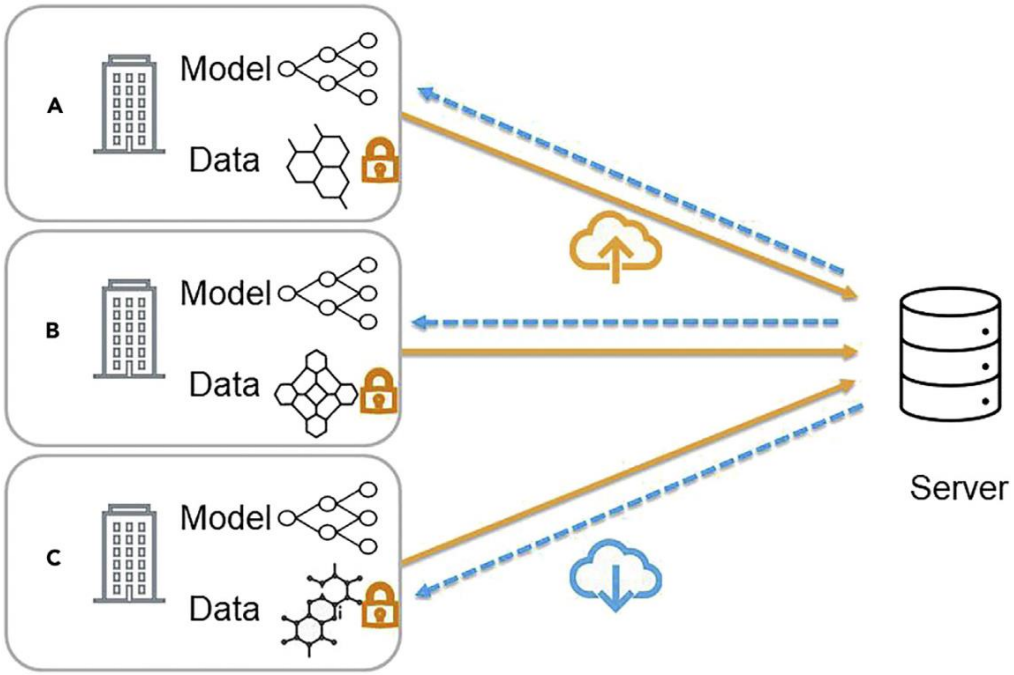
1. **Graph neural networks (GNNs):** GNNs are the state-of-the art performers for materials property prediction tasks [1]. Many GNNs such as CGCNN, MEGNet, DeeperGATGNN, SchNet, MPNN, ALIGNN are developed for this purpose.
2. **FL for Molecules:** Zhu et al. [2] explored FL for molecular property prediction focusing on the application of GNNs to learn molecular representations without sharing raw data.
3. **FL for Proteins:** Hausleitner et al. [3] applied FL to protein structure prediction, emphasizing its ability to integrate diverse data sources while maintaining confidentiality, which is crucial in fields with proprietary datasets.

[1] Fung, V., Zhang, J., Juarez, E. and Sumpter, B.G., 2021. Benchmarking graph neural networks for materials chemistry. *npj Computational Materials*, *7*(1), p.84.

[2] Zhu, W., Luo, J. and White, A.D., 2022. Federated learning of molecular properties with graph neural networks in a heterogeneous setting. *Patterns*, *3*(6).

[3] Hausleitner, C., Mueller, H., Holzinger, A. and Pfeifer, B., 2024. Collaborative weighting in federated graph neural networks for disease classification with the human-in-the-loop. *Scientific Reports*, *14*(1), p.21839.

SOLUTION



Input: # clients L, # local updates T, # Comms round C.
Output: Global Model F^g

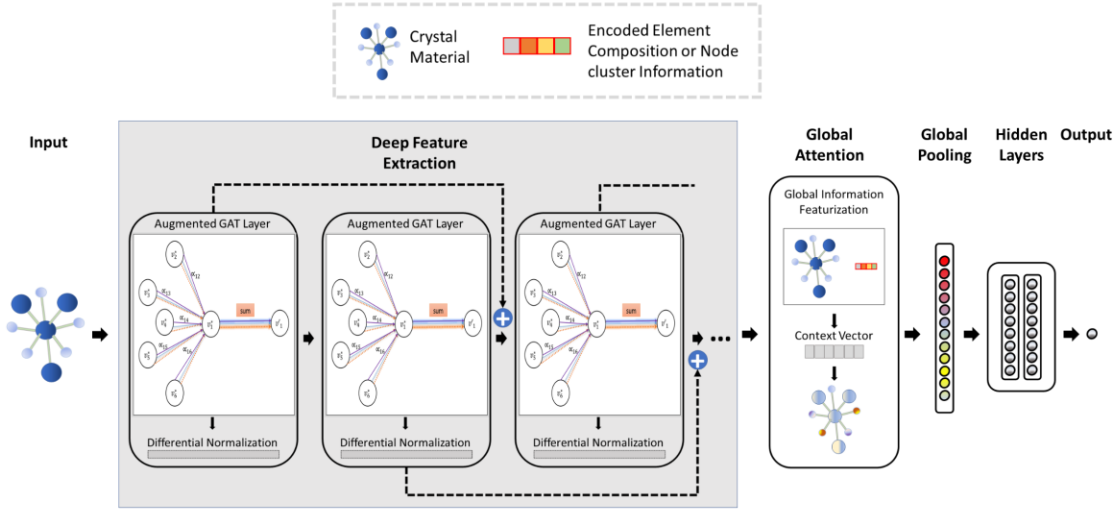
1: Server initialize a global model F^g ▷ Server init.
2: **while** Communication Round $< C$ **do**
3: Server broadcasts F^g to clients
4: $F^l \leftarrow F^g$ ▷ Client init.
5: **for** l : 1 to L in parallel **do** ▷ Client Update
6: **for** t : 1 to K **do** ▷ Update F^l for K steps
7: Sample a minibatch $\{g_i^l, y_i^l\}_{i=1}^{B} \sim X^l$
8: Update local model F^l by gradient descent
9: **end for**
10: Client sends updated model F^l to Server
11: **end for**
12: Server gets $F^g \leftarrow \sum_{l=1}^{L} \frac{|X^l|}{|X|} F^l$ ▷ Server Update
13: **end while**

Caption: (Left) An example FL system framework containing three clients and one "Server". Each client has one local dataset in this illustration, and the "Server" is used to receive and send the model parameters. (Right) The basic flow algorithm of an FL system.

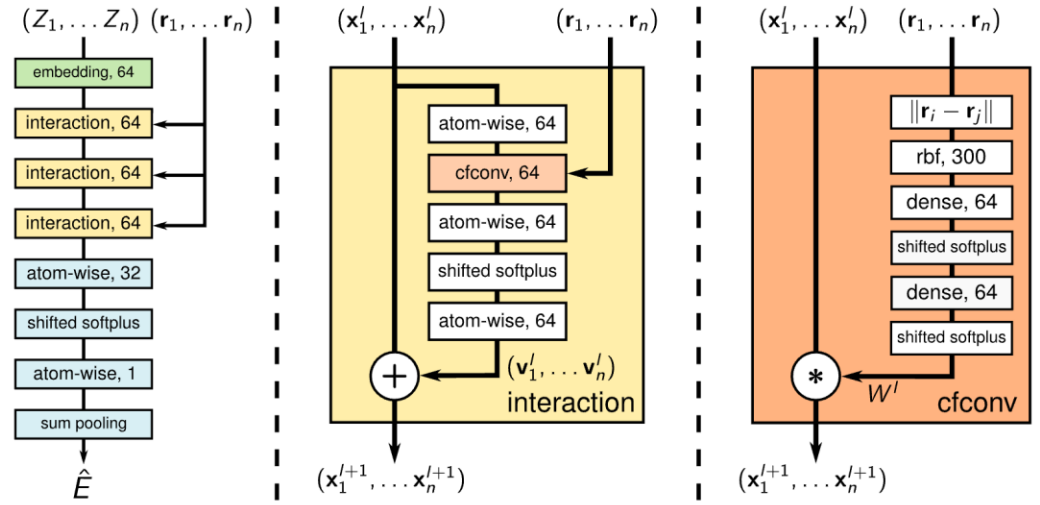
ML component: DeeperGATGNN, SchNet, MPNN.
Systems component: Pipeline design, server-client communication, parameter update scheme, client node scalability, etc, memory consumption, etc.

EXPERIMENTAL SETUP

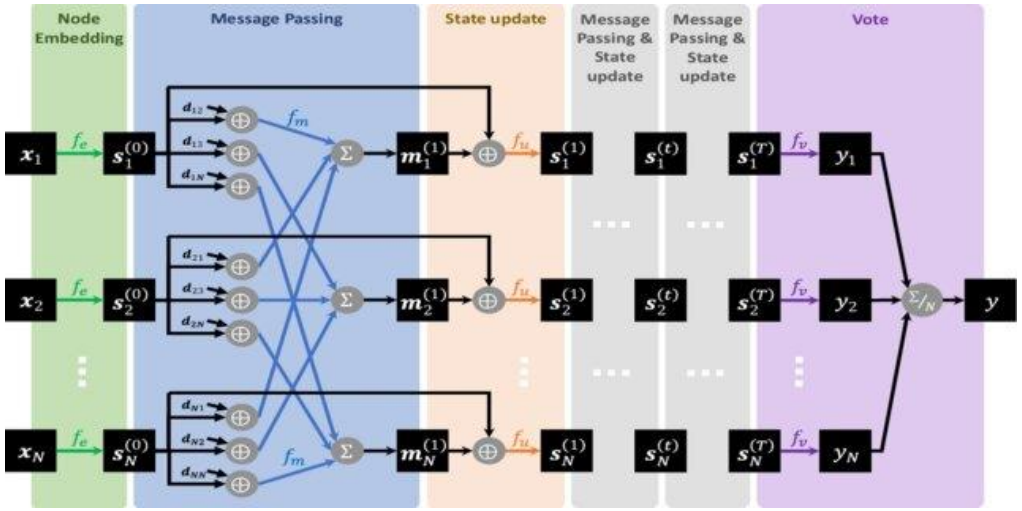
Model Architectures



DeeperGATGNN



SchNet



MPNN

EXPERIMENTAL SETUP

Datasets

Dataset	Property	# of Samples
Bulk Materials Formation energy	Formation Energy	36,839
Alloy Surface	Adsorption Energy	37,334
Pt-Cluster	Formation Energy	19,801
2D-Materials	Formation Energy	3,814
Band Gap	Band Gap	36,837

Table 1: Details of the five benchmark datasets used in this project.

Dataset	Property	# of Samples
Dielectric	Refractive Index	4764
Perovskites	Formation Energy	18,928
GVRH	Shear Modulus	10,987

Table 2: Details of the three OOD datasets used in this project.

Dataset	Property	# of Samples
Bulk Materials Formation energy	Formation Energy	36,839
Band Gap	Band Gap	36,837

Table 3: Details of the two datasets used for the scalability experiment in this project.

EXPERIMENTAL SETUP

FL Framework

- **Clients:** Default value 4 for benchmarking results, 5 for OOD results, and varying numbers (3, 4, 10... up to 500) for scalability testing.
- **Weight Update Scheme:** FedAvg.
- **Training Configuration:**
  - **Epochs:** 500
  - **Learning Rate:** 0.005
  - **Optimizer:** AdamW
  - **Batch Size:** 64
  - **Graph convolution layers:** 2-5 (SchNet, MPNN), 20 (DeeperGATGNN)
  - **Latent embedding dimension:** 64
  - **Activation function:** SoftPlus (DeeperGATGNN), ReLU (SchNet, MPNN)
  - **Communication round:** 10

EXPERIMENTAL SETUP

Implementation Platforms

- **Frameworks and Libraries:**
  - **Frameworks:** PyTorch [18], PyTorch-Geometric [19], FedML [20] frameworks were utilized for implementing the project. Overall, the FedChem [9] repository was used as a base for implementing our project.
  - **Libraries:** PyMatGen [17], ASE [16], Scikit-learn [21], etc. libraries were used to process the materials dataset. Matplotlib [22], Seaborn [23], etc. libraries were used for the visualization.
- **Hardware:**
  - The Hyperion cluster of the University of South Carolina were used to run the project. The code was run on the NVIDIA Tesla v100-32G GPUs .

RESULTS AND EVALUATION

- We performed **3 sets of experiments** -
  - 1. Experiments on benchmark materials data.
  - 2. Experiments on out-of-distribution (OOD) data.
  - 3. Scalability experiment of FL for materials data.
- We used **Mean Absolute Error (MAE),** which is a widely used metric in regression tasks to measure the average magnitude of errors between predicted values and the actual ground truth. MAE is used as the **default performance metrics for materials property prediction tasks in most literature**, so we also used it here.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_{\mathrm{true},i} - y_{\mathrm{pred},i}|$$

Where:

- $y_{\mathrm{true},i}$: The actual or true value for the i-th sample
- $y_{\mathrm{pred},i}$: The predicted value for the i-th sample
- n: The total number of samples

RESULTS AND EVALUATION

Results on Benchmark Materials Datasets

Model	Band-gap	Bulk Materials Formation energy	2D Material	Alloy Surface	Pt-Cluster
DeeperGATGNN	Baseline: 0.24570 eV	Baseline: 0.02955 eV/atom	Baseline: 0.17185 eV/atom	Baseline: 0.04086 eV/atom	Baseline: 0.13210 eV/atom
	FL (4 clients): **0.2291 eV**	FL (4 clients): 0.07724 eV/atom	FL (4 clients): 0.2385 eV/atom	FL (4 clients): **0.03642 eV/atom**	FL (4 clients): 0.1468 eV/atom
SchNet	Baseline: 0.28168 eV	Baseline: 0.05 eV/atom	Baseline: 0.214 eV/atom	Baseline: 0.063 eV/atom	Baseline: 0.151 eV/atom
	FL (4 clients): 0.3005 eV	FL (4 clients): 0.108 eV/atom	FL (4 clients): 0.306 eV/atom	FL (4 clients): 0.09 eV/atom	FL (4 clients): 0.206 eV/atom
MPNN	Baseline: 0.26485 eV	Baseline: 0.046 eV/atom	Baseline: 0.204 eV/atom	Baseline: 0.058 eV/atom	Baseline: 0.182 eV/atom
	FL (4 clients): 0.2892 eV	FL (4 clients): 0.9293 eV/atom	FL (4 clients): 0.2974 eV/atom	FL (4 clients): 0.8805 eV/atom	FL (4 clients): 0.2655 eV/atom

Table: Performance comparison of baseline and federated version of DeeperGATGNN, SchNet, and MPNN for different benchmark datasets. The baseline results outperformed the baseline results in most cases. Cases where FL version improves benchmark results are marked in bold letters. This shows the promise of FL method for model training for materials property prediction. Further fine-tuning of the models and hyperparameter training might also improve the results.

RESULTS AND EVALUATION

Results on Benchmark Materials Datasets

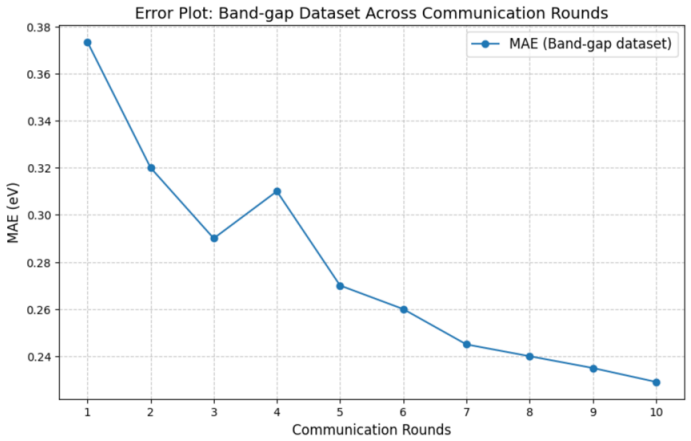
Model	Band-gap	Bulk Materials Formation energy	2D Material	Alloy Surface	Pt-Cluster
DeeperGATGNN	Baseline: 0.24570 eV	Baseline: 0.02955 eV/atom	Baseline: 0.17185 eV/atom	Baseline: 0.04086 eV/atom	Baseline: 0.13210 eV/atom
	FL (4 clients): **0.2291 eV**	FL (4 clients): 0.07724 eV/atom	FL (4 clients): 0.2385 eV/atom	FL (4 clients): **0.03642 eV/atom**	FL (4 clients): 0.1468 eV/atom
SchNet	Baseline: 0.28168 eV	Baseline: 0.05 eV/atom	Baseline: 0.214 eV/atom	Baseline: 0.063 eV/atom	Baseline: 0.151 eV/atom
	FL (4 clients): 0.3005 eV	FL (4 clients): 0.108 eV/atom	FL (4 clients): 0.306 eV/atom	FL (4 clients): 0.09 eV/atom	FL (4 clients): 0.206 eV/atom
MPNN	Baseline: 0.26485 eV	Baseline: 0.046 eV/atom	Baseline: 0.204 eV/atom	Baseline: 0.058 eV/atom	Baseline: 0.182 eV/atom
	FL (4 clients): 0.2892 eV	FL (4 clients): 0.9293 eV/atom	FL (4 clients): 0.2974 eV/atom	FL (4 clients): 0.8805 eV/atom	FL (4 clients): 0.2655 eV/atom

- **Observations:**
- The gap between FL and centralized baseline performance varied significantly across datasets, highlighting that FL may be more effective on some datasets than others, depending on the distribution and quality of the data.
- **DeeperGATGNN consistently outperformed the other model**s in the FL framework, making it the most promising candidate for this approach.
- **Fine-tuning FL hyperparameters** such as client selection, aggregation strategies, and communication rounds could further reduce the MAE for all models.

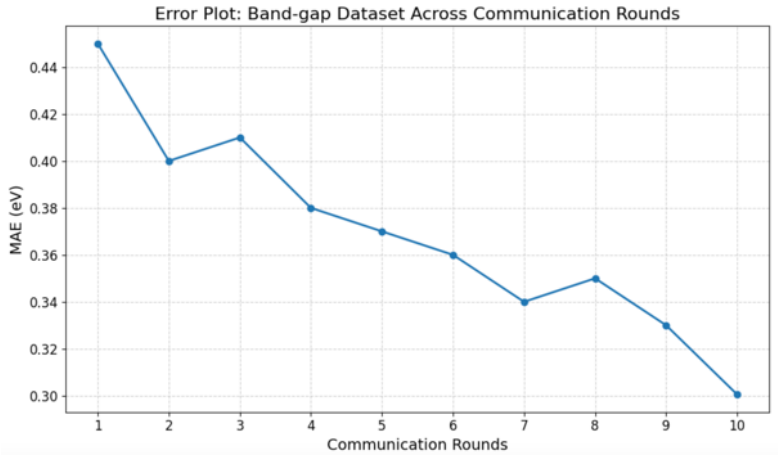
RESULTS AND EVALUATION

Results on Benchmark Materials Datasets

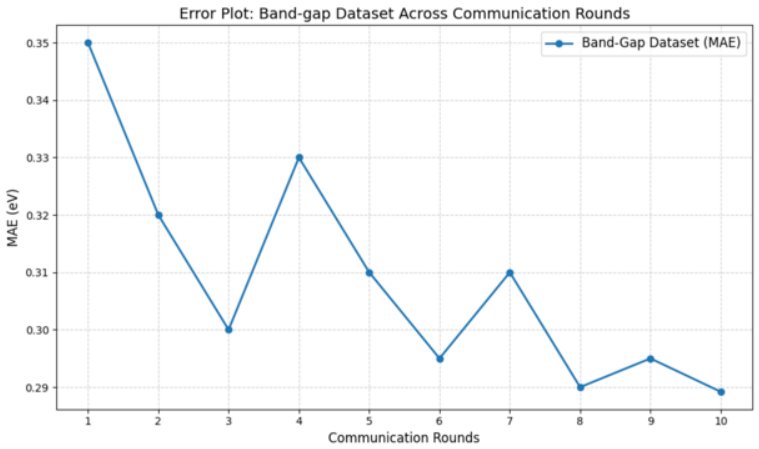
- **MAE vs communication rounds:**



DeeperGATGNN



SchNet



MPNN

- The error **decreases on average with each communication round** for each model, suggesting that **more communication rounds might be needed for fine-tuning** the results in the future.

RESULTS AND EVALUATION

Results on OOD Datasets

Dataset	DeeperGATGNN	SchNet	MPNN
Dielectric	ID: 0.3355 (unitless)	ID: 0.3277 (unitless)	ID: 0.4682 (unitless)
	OOD: 0.5079 (unitless)	OOD: 0.6621 (unitless)	OOD: 0.587 (unitless)
GVRH	ID: 0.0903 log10(GPa)	ID: 0.0796 log10(GPa)	ID: 0.1206 log10(GPa)
	OOD: 0.1319 log10(GPa)	OOD: 0.1882 log10(GPa)	OOD: 0.1756 log10(GPa)
Perovskites	ID: 0.0288 eV/unit cell	ID: 0.0342 eV/unit cell	ID: 0.0621 eV/unit cell
	OOD: 0.0671 eV/unit cell	OOD: 0.0906 eV/unit cell	OOD: 0.1205 eV/unit cell

Table: Performance comparison of ID and OOD data of DeeperGATGNN, SchNet, and MPNN for different datasets. Although FL has shown good promise in improving OOD results in other tasks in the past, it did not however improve performances of OOD materials dataset results, demonstrating the complexity of material data and hardness of various distribution material data to cover

RESULTS AND EVALUATION

Results on OOD Datasets

Dataset	DeeperGATGNN	SchNet	MPNN
Dielectric	ID: 0.3355 (unitless)	ID: 0.3277 (unitless)	ID: 0.4682 (unitless)
	OOD: 0.5079 (unitless)	OOD: 0.6621 (unitless)	OOD: 0.587 (unitless)
GVRH	ID: 0.0903 log10(GPa)	ID: 0.0796 log10(GPa)	ID: 0.1206 log10(GPa)
	OOD: 0.1319 log10(GPa)	OOD: 0.1882 log10(GPa)	OOD: 0.1756 log10(GPa)
Perovskites	ID: 0.0288 eV/unit cell	ID: 0.0342 eV/unit cell	ID: 0.0621 eV/unit cell
	OOD: 0.0671 eV/unit cell	OOD: 0.0906 eV/unit cell	OOD: 0.1205 eV/unit cell

- **Observations:**
  - **Overall, did not improve OOD performance.**
  - This underscores the **inherent complexity and heterogeneity of materials datasets**, which pose significant challenges for machine learning methods.

RESULTS AND EVALUATION

Results on OOD Datasets (Fold-wise comparison)

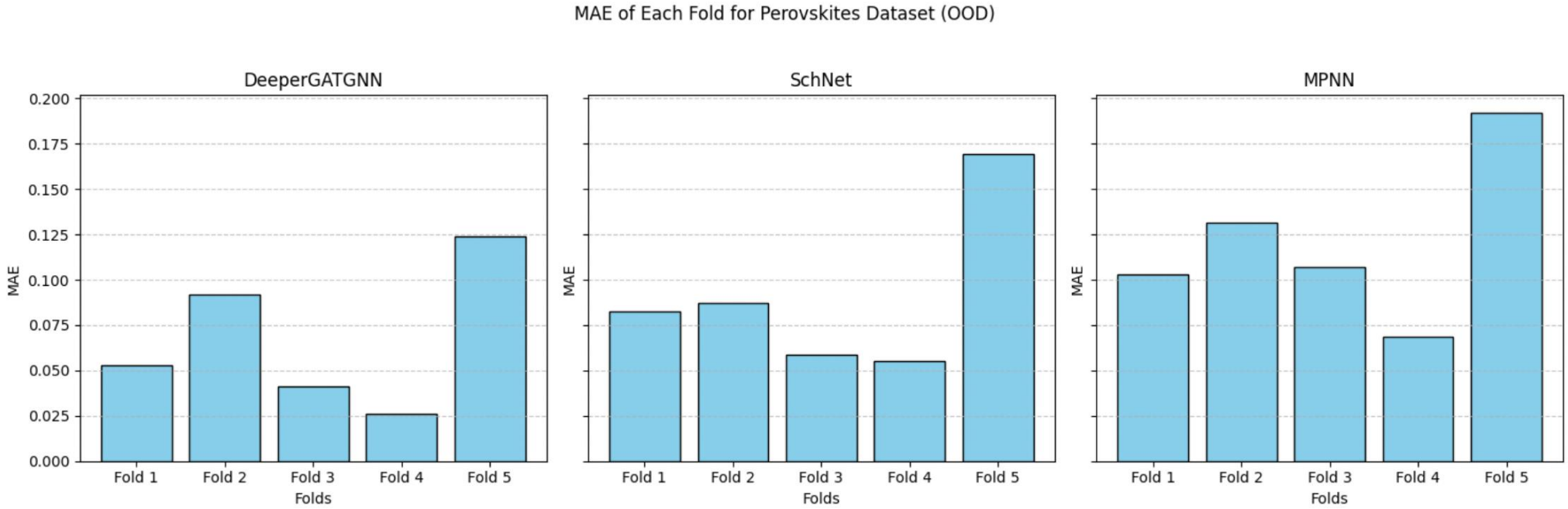


Figure: Fold-wise error of federated DeeperGATGNN, SchNet, and MPNN for the perovskites OOD dataset. It shows that the deterioration of results is mostly achieved because of subpar performances in 1-2 folds, indicating more fine-tuning is needed on those OOD folds **.**

- **Observations:**
  - The poor overall MAE observed in some datasets was **primarily driven by poor results in 1-2 specific folds**, suggesting that **targeted fine-tuning and data-specific optimizations are necessary** to improve performance on those folds.

RESULTS AND EVALUATION

Results on Scalability Experiments

# of Clients	Convergence Time (hr)	Formation Energy (MAE)
3	≈ 28	0.1844 eV/atom
4	≈ 35	0.0772 eV/atom
10	≈ 42	**0.0734 eV/atom**
50	≈ 66	0.1508 eV/atom
100	≈ 79	0.1562 eV/atom
200	≈ 151	0.5780 eV/atom
500	memory overload	-

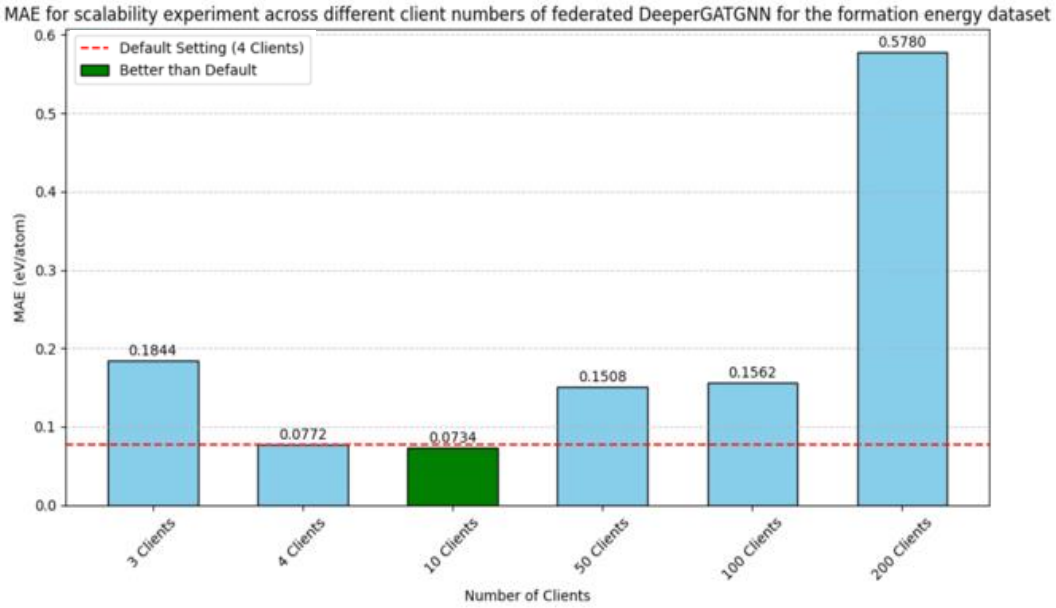
# of Clients	Convergence Time (hr)	Band Gap (MAE)
3	≈ 25	**0.2106 eV**
4	≈ 38	0.2291 eV
10	≈ 47	0.2784 eV
50	≈ 60	**0.1907 eV**
100	≈ 69	0.2652 eV
200	≈ 134	0.3319 eV
500	memory overload	-

Figure: Convergence time and global-model MAE performance across different client configurations for the formation energy (left), and the band-gap (right) dataset of the federated DeeperGATGNN model. The client number changed improved the performance of the model in one case (marked in bold letter).

- **Observations:**
  - **Improvement with More Clients:** Increasing the number of clients initially enhanced model performance as more data diversity and heterogeneity contributed to better generalization for the global model.
  - **Performance Degradation with Too Many Clients:** Beyond a certain threshold (e.g., 50 clients), the dataset size for each client became too small, leading to overfitting of local models .
  - **Resource Constraints:** Experiments with 500 clients encountered memory limitations, emphasizing the need for more efficient aggregation algorithms and better computational resource management in large-scale FL setups.

RESULTS AND EVALUATION

Results on Scalability Experiments



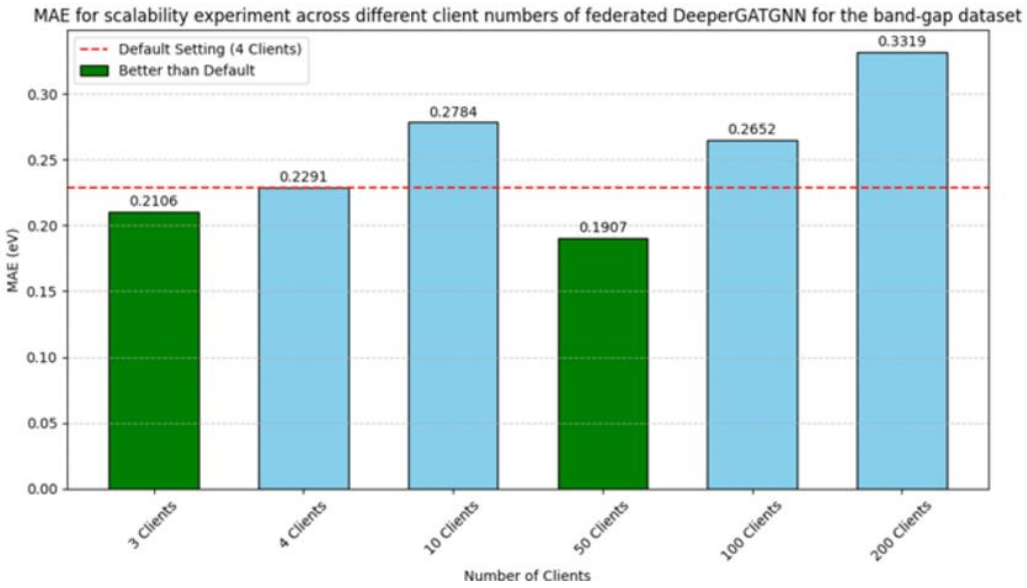


Figure: Convergence time and global-model MAE performance across different client configurations for the formation energy (left), and the band-gap (right) dataset of the federated DeeperGATGNN model. This shows that given more client nodes, that results can be improved as demonstrated for 10 clients for the formation energy dataset, and 50 clients for the band-gap dataset.

Observations:

- **Improvement with More Clients:** Increasing the number of clients initially enhanced model performance as more data diversity and heterogeneity contributed to better generalization for the global model.
- **Performance Degradation with Too Many Clients:** Beyond a certain threshold (e.g., 100 clients), the dataset size for each client became too small, leading to overfitting of local models .
- **Resource Constraints:** Experiments with 500 clients encountered memory limitations, emphasizing the need for more efficient aggregation algorithms and better computational resource management in large-scale FL setups.

BROADER IMPACT

- **Impact of the Work**
  - First federated learning (FL) for materials property prediction, addressing privacy and enabling collaboration.
  - Provides a framework for use in materials discovery and other scientific domains.
- **How Others Can Use It**
  - Adapt FL for other data-scarce and decentralized domains.
  - Develop model for real life materials science applications such as drug discovery, or vaccine development, without explicitly sharing the data.
- **Limitations**
  - Struggles with generalization to out-of-distribution datasets.
  - Scalability issues due to memory limits and communication overhead.
- **Future Improvements**
  - Develop better aggregation methods and task-specific optimizations.
  - Use data augmentation and invariant learning for better OOD results.
  - Explore resource-efficient FL frameworks and extend communication rounds.
  - Explore multiple models for multiple clients for collective improvement of the global model.

Thank you !!!