FedMDH: A Federated Learning Framework for Effective Sharing of Multi-Dimensional Heterogeneous Materials Data

Data-Driven Fourth Paradigm

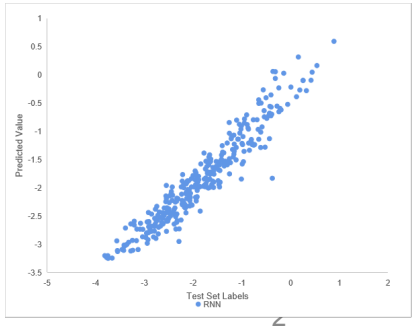
Materials research and development is moving towards a data-driven fourth paradigm

Input

Output

- **Material classification**

- **Material property prediction**



Large-scale Datasets

Machine Learning

As big data continues to flourish, numerous materials databases have been established globally to support new materials research and development.

Material Big-Data Platform

In the field of materials, some studies have already utilized data and services from existing materials big data platforms. However, the development of these platforms is still immature, with limited types of data services provided and a lack of security mechanisms to ensure the safety of materials data sharing.

Name	Data Source	Services	Security
AFLOW	Pauling File, ICSD Database, Navy Crystal, Lattice Database	Search and prediction	——
COD	File upload through script inspection, currently supports CIF files only	Search and download	Log recording, access control, data backup
MARVEL NCCR	Converts basic structural data files to AiiDA-compatible files for data model expansion	Search and knowledge building	Utilizes a source without environmental impact, system stability
MDF	Ingests user-uploaded data in specified formats, with data resources available for viewing or download	Search, data integration, automated analysis	Identity verification, content backup
Materials Project	Based on ICSD and other databases	Data creation, validation, search, download, analysis, and design	Identity verification, data integrity verification
NOMAD CoE	No limit on data upload, processed using DOI	Search and download; employs AI methods to filter available materials, recognize and process structure-related data	Identity verification, data access control
OQMD	Structural parameters based on customized ICSD	Search , data analysis (using DFT calculations for thermodynamic and structural properties of materials).	——
Open materials database	Based on COD database	Analysis with high-throughput tools	——
CISRI-New Materials	Customizable upload	Search, AI-based data identification, and cloud control/cloud service access	Identity verification

The Challenges of Material Data Sharing

The Material big-data platform facilitates data collection. However, a key issue that urgently needs to be addressed by the platform is how to fully utilize the data from various organizations and maximize the value of the materials data.

Public

Each materials organization can download the data for thorough analysis and exploration.

Confidential

Without aggregating these data for model training, the full value of the data cannot be realized.

Small sample data: The organization cannot perform model training with these data.

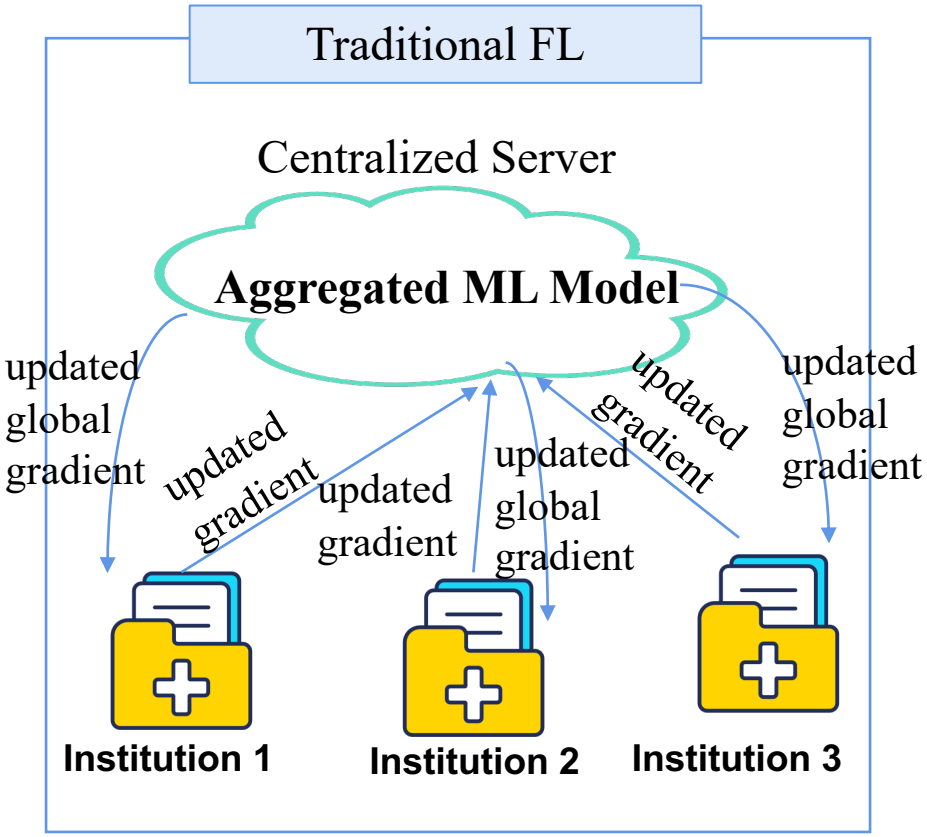
Data Augmentation: Only captures characteristics of the organization's dataset, lacking generalization capability.

Insufficient sample features: This leads to poor model generalization capability.

Transfer Learning: Enhances sample diversity but risks leaking original data.

Therefore, We need a secure distributed collaborative computing method to enable collaborative modeling of heterogeneous materials datasets while ensuring data safety. **Federated learning (FL)** is a distributed learning framework designed to train a model on data that cannot be centralized, without sharing raw data.

The Challenges of Fedrated Learning



Challenge 1：No Trusted Third-Party

The **centralized FL** setting is not suited to the multi-institutional collaboration problem, as it involves a centralized third party that controls a single model.

Challenge 2：Single Point of Failure

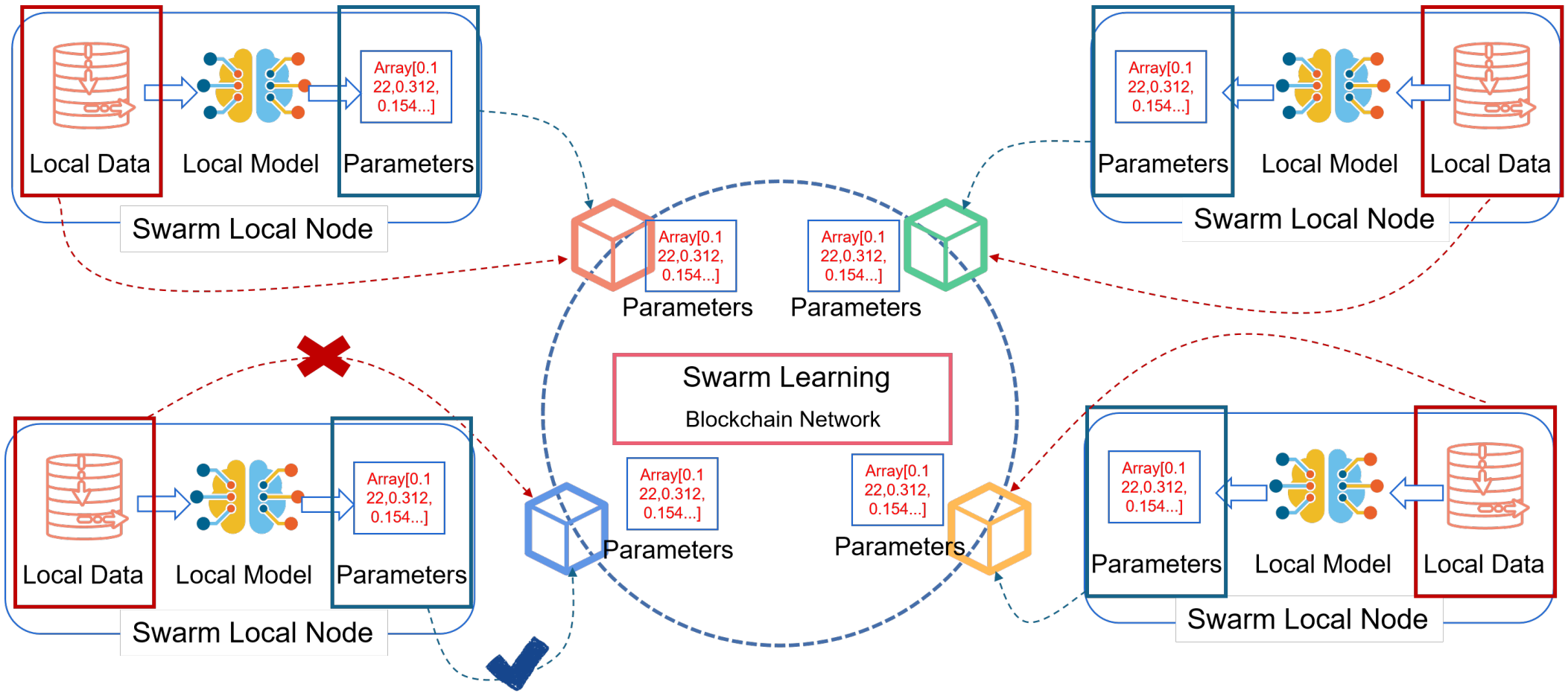
In **centralized FL**, the coordinator (usually a server) is responsible for receiving model updates from participants and aggregating them into a global model. If the coordinator fails, the entire federated learning process will be disrupted.

The **decentralized FL frameworks** are preferred under such settings.

Swarm Learning applies blockchain technology to promote a decentralized, secure network for collaborative training.

Swarm Learning

Unlike traditional federated learning, swarm learning uses blockchain technology to ensure secure and tamper-proof collaboration between institutions, allowing for data privacy, autonomy, and enhanced trust in multi-institutional collaboration without needing a central authority.

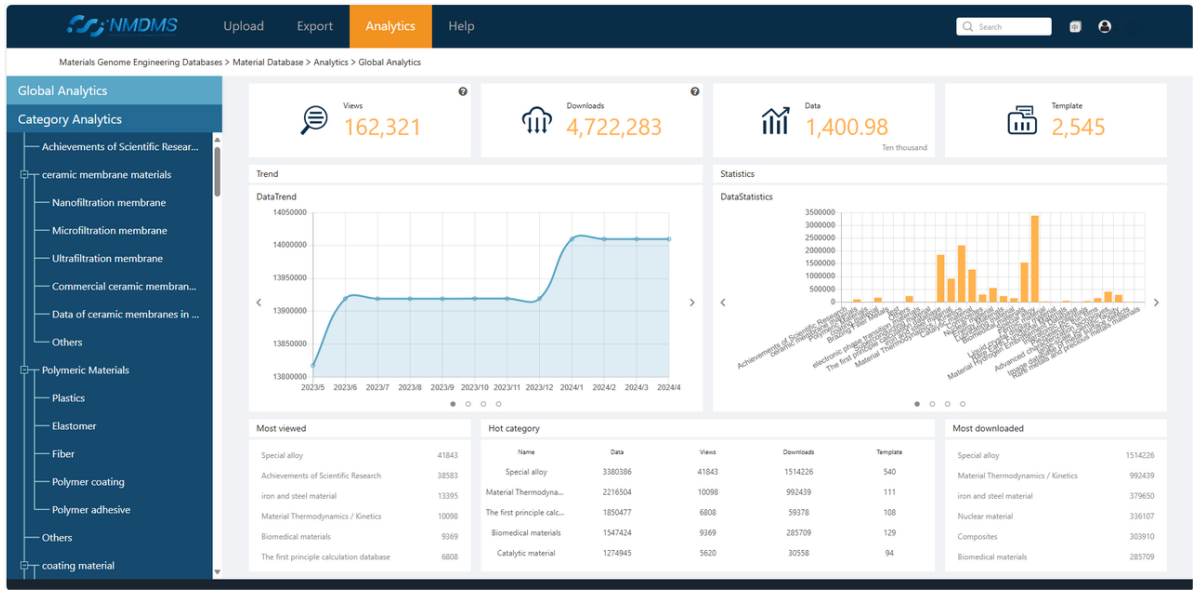


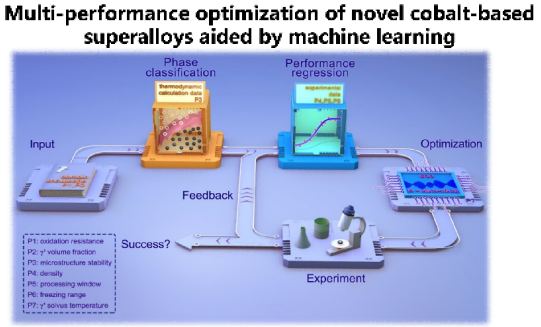
The NMDMS Platform

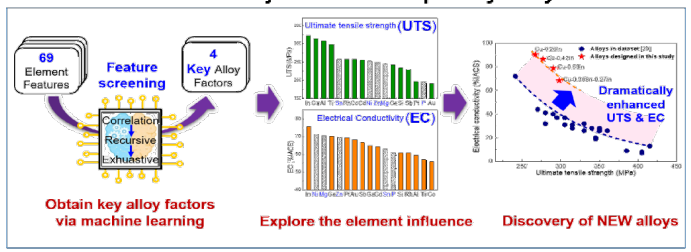
In order to provide an open and shared big data platform for Materials science researchers, relying on the Materials Genome Engineering Project of China, National Materials Data Management and Services (NMDMS) Platform has been established and put into use.

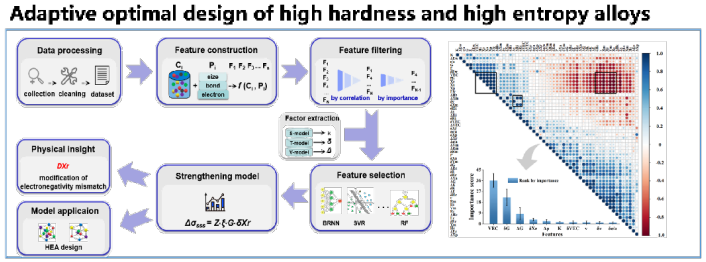
- *Currently, the platform has registered over **30** organizations in China and collected more than **14 million** valid materials data entries.*

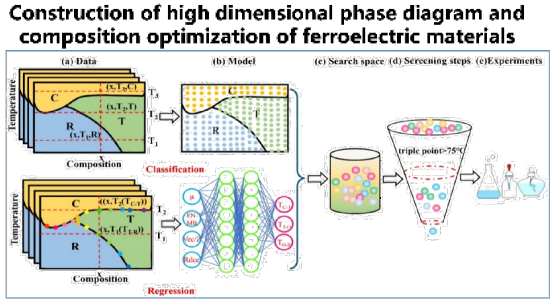
- *Relevant case studies include the **machine learning-assisted** multi-performance optimization of new cobalt-based superalloys, rational design of solution-strengthened copper alloys based on key elemental characteristics, and the construction and composition optimization of high-dimensional phase diagrams for ferroelectric materials.*









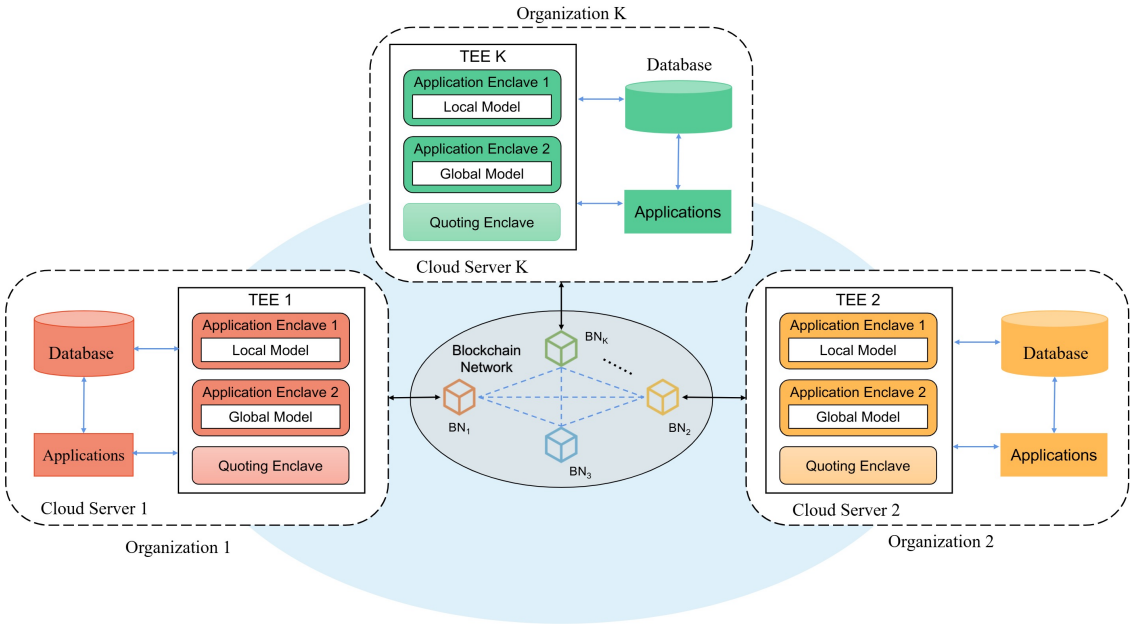


MatSwarm : Collaborative Materials Computation

The NMDMS platform has introduced the MatSwarm framework, integrating federated learning and blockchain technology to address data privacy and non-independent identical distribution (non-i.i.d.) issues in multi-agent collaboration in the materials field. By incorporating swarm transfer learning and Trusted Execution Environments (TEEs), the framework enhances model accuracy and generalization, ensuring secure data aggregation and collaborative training.

- **Secure Collaborative Computing Framework:** Proposes the MatSwarm decentralized collaborative computing framework for the materials science field, optimized for collaborative training of sensitive data sets. This is the first application in the materials domain, and its advantages over existing approaches have been validated.

- **Swarm Transfer Learning:** Addresses the non-independent distribution of materials data with an innovative swarm transfer learning approach. This method enables simultaneous preservation of data source characteristics while enhancing model accuracy and generalization, achieving high-precision models even with fundamentally distinct or sparse features.



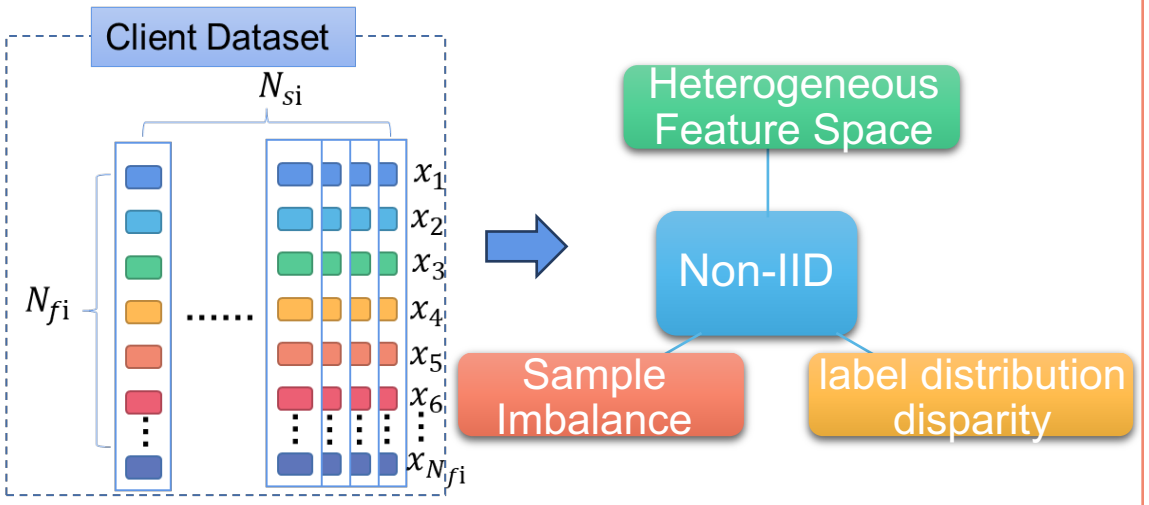
- **Enhanced Security through Trusted Execution Environments:** Introduces Intel SGX-based Trusted Execution Environments to ensure the confidentiality of model parameters during aggregation, preventing model tampering and ensuring secure and reliable model training and aggregation.

[1] Wang R, Xu C*, Zhang S, Ye F, Tang Y, Tang S, Zhang H, Du W, Zhang X*. MatSwarm: Trusted Swarm Transfer Learning Driven Materials Computation for Secure Big Data Sharing[J]. ***Nature Communications***, 2024.

Challenges of Heterogeneous Material Data in NMDMS

In materials science, variations in testing samples and equipment lead to non-IID data across organizations. This heterogeneity causes inconsistent model performance, slower global model convergence, and client-side bias.

Data Heterogeneity



Note: Each client's dataset can exhibit heterogeneity in all three aspects mentioned above.
N_{si} represents the number of samples for client i.
N_{fi} represents the number of features for client i.

Material Data Heterogeneity

➢ Characteristics of Heterogeneous Materials Datasets

Nearly all materials datasets exhibit **three major heterogeneity issues:** sample imbalance, heterogeneous feature space (feature attributes and dimensions), and label distribution disparity.

➢ Challenges of Heterogeneous Materials Datasets

① **Heterogeneous Feature Space :** Simple dimensionality reduction cannot address heterogeneity due to different feature attributes.

② **Sample Imbalance:** The diversity of samples generated by a single organization is often insufficient, which hinders the generalization ability of the model.

③ **Label Distribution Disparity :** Swarm learning is primarily used to solve regression problems in the material science.However, current solutions like knowledge distillation, used for classification problems, are not suitable for regression.

Therefore, while the use of a swarm learning framework can ensure the security of sensitive materials data during collaborative computation, it is still necessary to address the heterogeneity of materials datasets. This is crucial for the practical application of the swarm learning framework in the field of materials science.

Existing Solution and Limitation

Nowadays , some solutions already exist to address data heterogeneity issues in federated learning. However, these solutions mainly focus on sample insufficiency and non-IID problems. There are fewer solutions addressing feature heterogeneity. **Existing solutions fail to provide a comprehensive approach to data heterogeneity that tackles sample imbalance, feature space heterogeneity, and label distribution disparity simultaneously.**

Challenge	Method	Limitation
Heterogeneous Feature Space	Domain Adaptation: Map image features to a shared space using local embedding functions.	Effective for image classification, but not directly applicable to non-linear regression in materials science.
Sample Insufficiency and Feature Missing	Data Augmentation: Add noise, fill mean values to generate samples and supplement missing features.	Affects model accuracy and does not improve generalization ability.
	GAN: Use distributed generative adversarial networks (FeCGAN, FedDA) to generate missing features for non-overlapping samples.	No overlapping samples in materials science, limiting the use of true data for generating missing features.
Label Distribution Disparity	Knowledge Distillation: Enable models to learn from the global model even with different architectures (FedKD).	Mainly applied to classification problems; not suitable for regression problems in materials science.
	Penalty Terms-based: Reduce local gradient variance by decoupling and correcting local drifts (FedDC).	Frequent parameter exchange affects communication efficiency and increases the risk of inferring original data.

The Contribution of FedMDH

Based on the characteristics of materials data and the application requirements in the materials field, we propose FedMDH algorithm suitable for heterogeneous materials data to address the practical application needs of collaborative prediction of new material properties.

- **How to Address Heterogeneous Feature Space in Materials Data?**

We propose Global Feature Anchor based Mapping (GFAM) model to map different feature spaces to a shared space with similar distributions, enhancing joint training effectiveness.

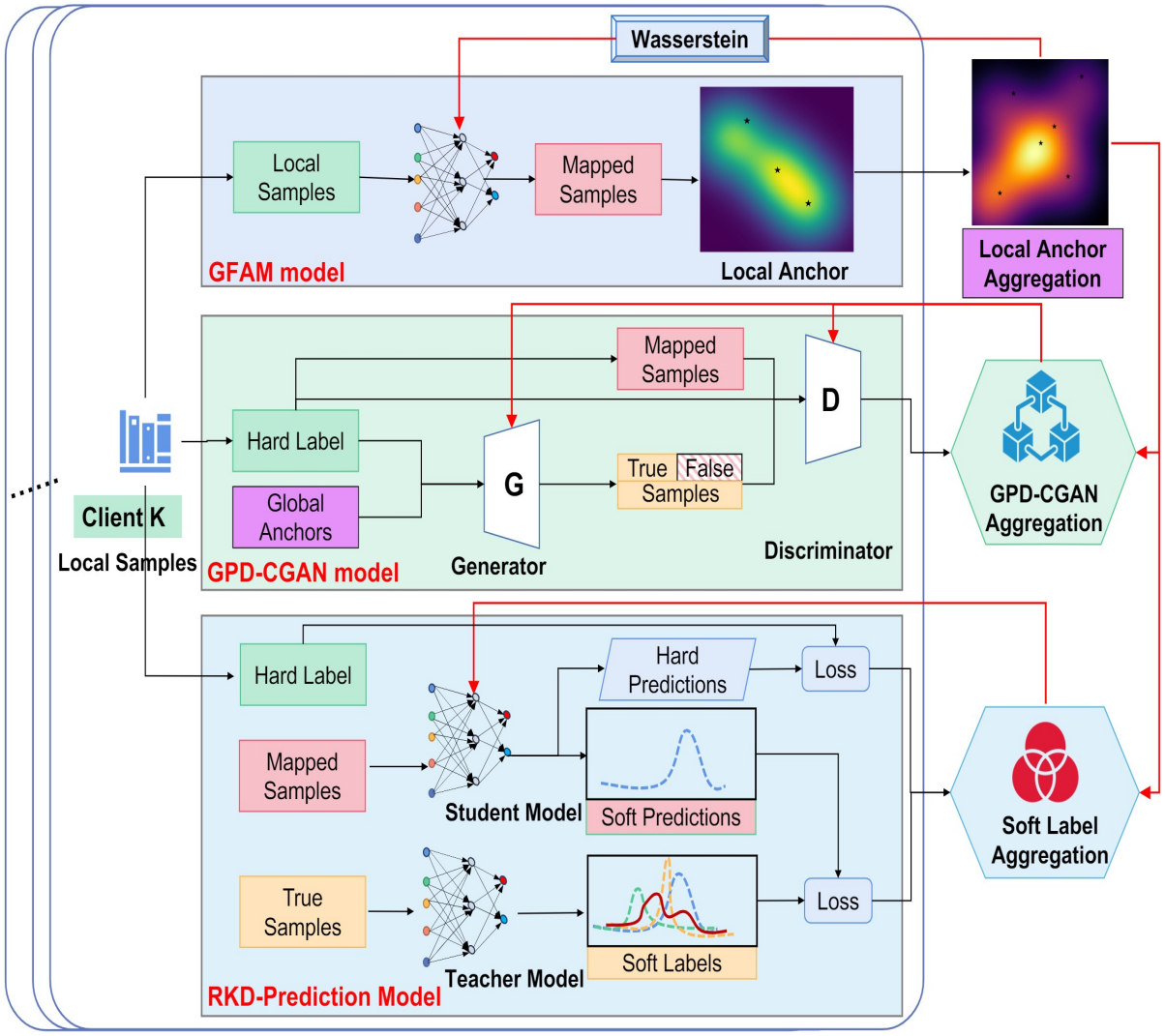
- **How to Address Sample Insufficiency and Imbalance in Materials Data?**

We use Global Prediction Values Distribution based Conditional Generative Adversarial Nets (GPD-CGAN) to generate labeled samples with a global feature distribution during the iterative optimization process, addressing sample quantity imbalances across organizations.

- **How to Address Label Distribution Disparity in Materials Data?**

We propose a Regression Knowledge Distillation based Prediction (RKD-Prediction) model to use predicted values as soft labels and aggregate these to obtain a global prediction distribution. This adjusts local training, addressing label distribution disparities.

FedMDH Framework



①Feature Dimensionality Reduction: Use a neural network-based mapping model GFAM to transform high-dimensional feature spaces of different dimensions into a low-dimensional space with uniform dimensions.

②Constructing a Shared Feature Space

- Map features to a shared low-dimensional space.
- Construct global feature distribution anchor points.
- Update GFAM model parameters using the Wasserstein barycenter method to align local feature distributions with the global distribution.

③Sample Quantity Alignment

- Train a global prediction values distribution-based conditional generative adversarial networks (GPD-CGAN) during the mapping model update.
- Achieve joint sample generation between clients, allowing each organization to generate samples with a global feature distribution.

④Prediction Model Training Based on RKD

- Use the shared model's prediction value distribution as soft labels.
- Share these labels via local blockchain nodes to the blockchain network.
- Obtain a global prediction value distribution through smart contracts' aggregation algorithms.
- Enable each organization to learn the global label distribution characteristics.

Construction of a Shared Feature Space(GFAM Model)

01 Design Mapping Model

Use MLP to map different dimensions d_i to a common k dimensional space.

- The mapping model: $\phi_i: x_i^{d_i} \rightarrow x_i^k$

02 Local Feature Space Approximation

Approximate the local feature space probability distributions.

- Use multivariate Gaussian distributions $v_{\phi_{L_i}(x_{i;k})} = N(m_i, \Sigma_i); i \in K$ for each local feature space.

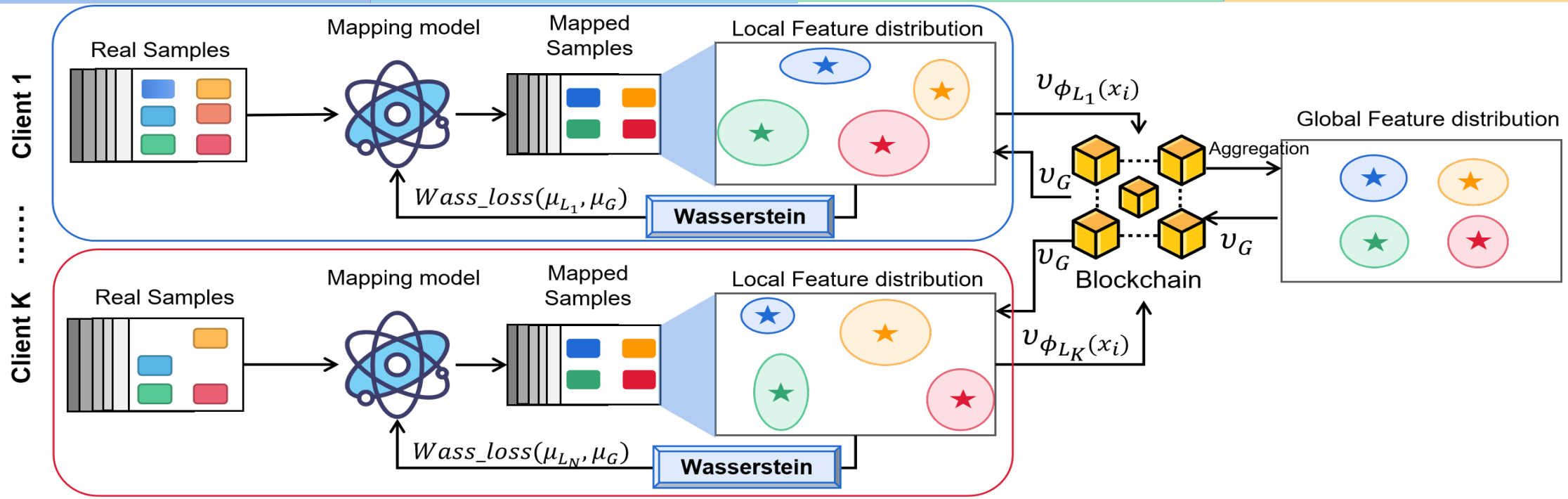
03 Creat a Global Anchor Point

Introduce a global feature anchor point with an initial distribution $\{v_{G_0} = N(m_{G_0}, \Sigma_{G_0})\}$, updated by aggregating local distributions.

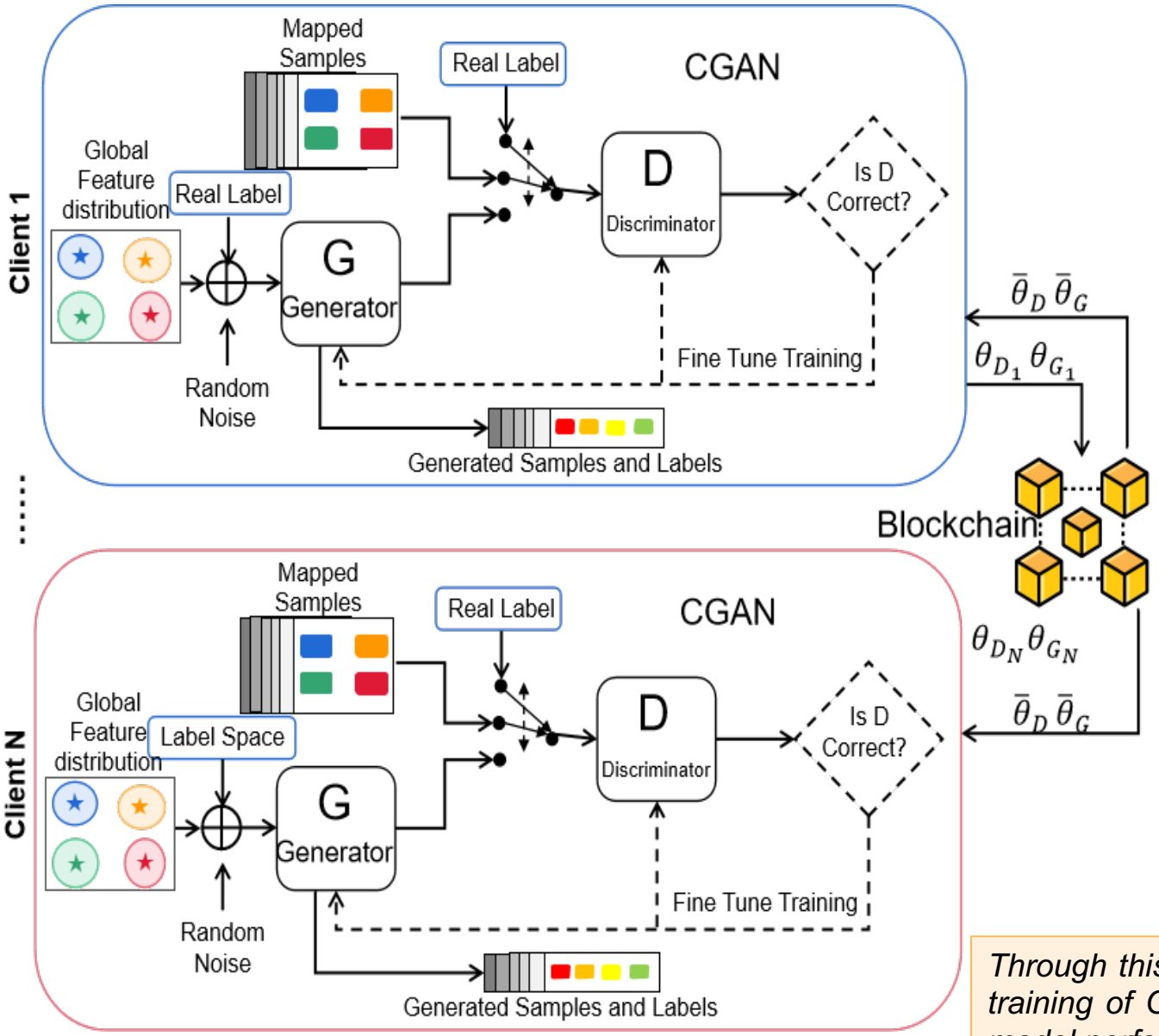
04 Minimizing Distribution Discrepancies

Align local feature distributions with the global feature distribution.

- Optimize the objective $\mathcal{L}_w$ to make the local feature distributions as close as possible to the global distribution.



Samples Generation based on GPD-CGAN



Implementation Principle of GPD-CGAN

GPD-CGAN leverages a collaborative approach to train conditional generative adversarial networks (CGANs) across multiple clients.

① **Individual Client Models:** Each client i has its own generator G_i and discriminator D_i.

② **Model Training and Updates**

- Train generator $G_i: (z_{v_G}, y_i) \rightarrow (\hat{x}, y_i)$ and get generator parameters θ_{G_i} ;
- Train discriminator $D_i: (\hat{x}, x, y_i) \rightarrow P_{real}$ and get discriminator parameters θ_{D_i}.

Clients send these updated parameters to the blockchain network.

③ **Blockchain-Based Aggregation:** Smart contracts on the blockchain network aggregate the parameters from all clients

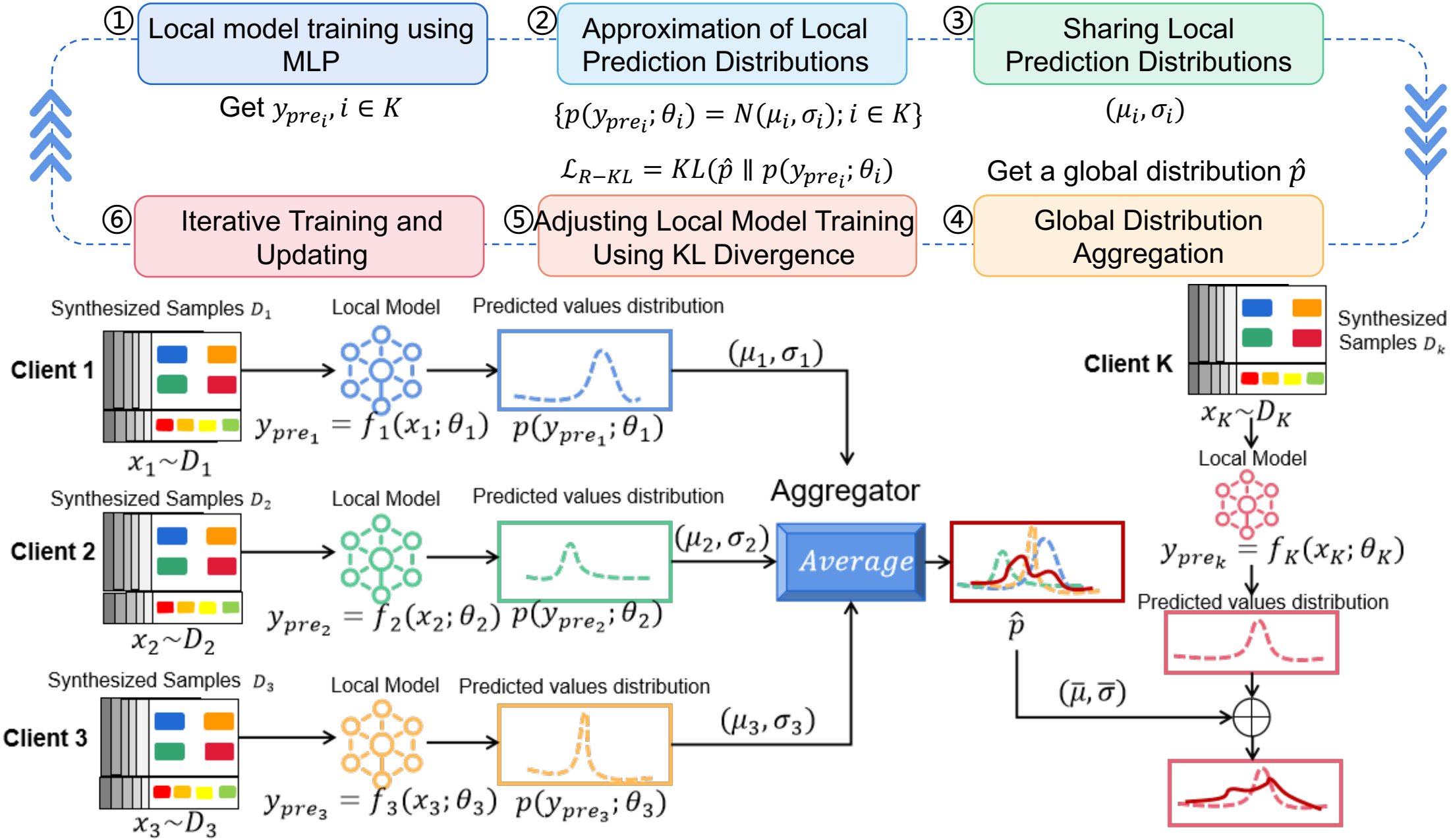
$$\bar{\theta}_G = \frac{1}{K}\sum_{i=1}^{K} \theta_{G_i} \qquad \bar{\theta}_D = \frac{1}{K}\sum_{i=1}^{K} \theta_{D_i}$$

④ **Global Model Distribution:** These aggregated parameters $(\bar{\theta}_G, \bar{\theta}_D)$ are then distributed back to the clients.

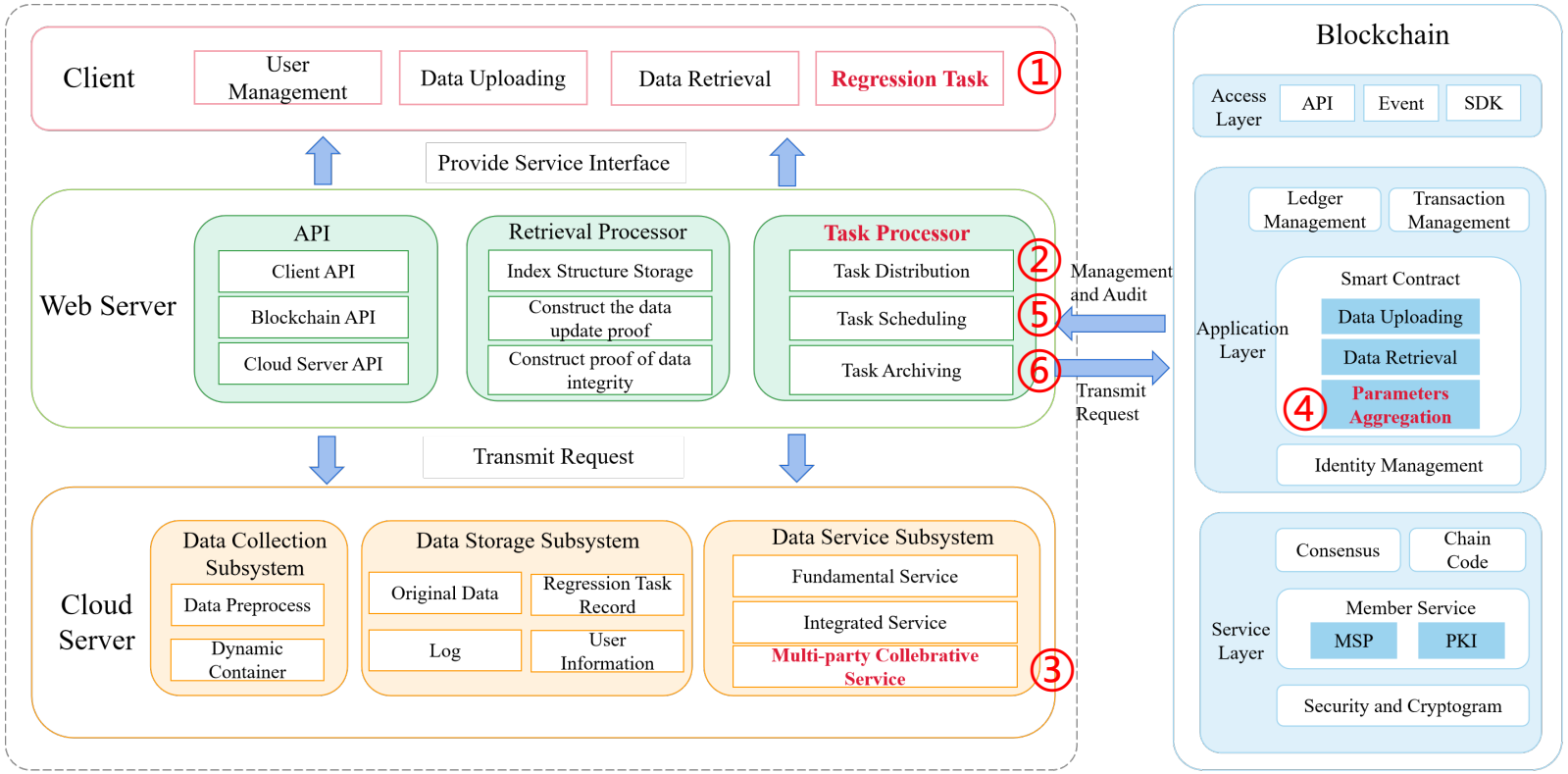
⑤ **Iterative Training:** The process of local training, parameter update, aggregation, and distribution is repeated iteratively.

*Through this mechanism, **GPD-CGAN** enables secure and efficient collaborative training of CGANs across multiple clients, ensuring data privacy and enhancing model performance.*

Model Training based on RKD (RKD-Prediction Model)



System Design based on FedMDH



1. The client submits a task via the "Regression Task" interface.

2. The Task Distribution module assigns tasks based on participant information.

3. All participants invoke and deploy Docker containers provided by the Multi-party Collaborative Service to train their local models for FedMDH.

4. Participants return the locally trained model parameters to the Task Scheduling, which then invokes the "Parameters Aggregation" smart contract for aggregation.

5. The Task Scheduling returns the aggregated parameters to update each local model.

6. This process is repeated until the model converges, at which point the final global model is archived and stored on the blockchain to facilitate future model retrieval.

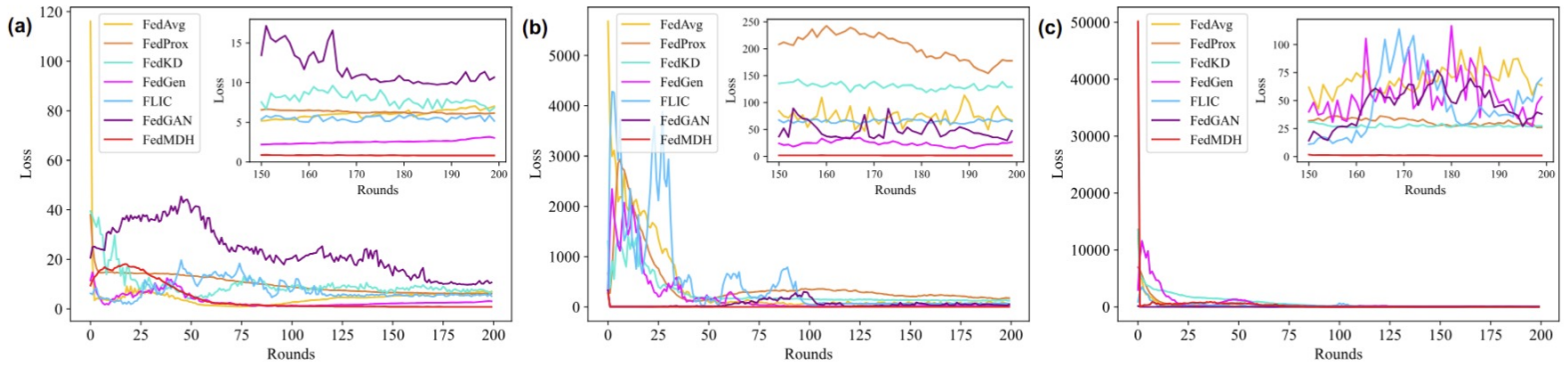
Experiment Evaluation——Different Dataset

The comparison of different performance metrics between FedMDH and other existing federated learning algorithms.

Dataset	Methods	RMSE	MAE	R2
Perovskite	FedAvg	2.6492±0.16	2.4946±0.17	-6.7806±0.67
	FedProx	2.4813±0.44	2.1497±0.45	-5.8258±2.10
	FedKD	2.3204±0.32	2.0347±0.33	-6.5623±1.38
	FedGen	1.7439±0.82	1.2505±0.83	-2.3738±3.81
	FLIC	2.2716±0.08	1.9152±0.08	-4.8803±0.25
	FedGAN	3.2653±1.20	3.1757±1.24	-10.8201±3.93
	FedMDH	**0.9006±0.04**	**0.7352±0.05**	**0.0880±0.07**
Silicon	FedAvg	7.8942±0.09	5.5254±0.05	-39.5136±0.16
	FedProx	13.3319±0.08	9.9623±0.05	-105.6403±0.16
	FedKD	10.3141±0.04	8.6408±0.04	-76.3445±0.09
	FedGen	4.9359±0.03	4.1281±0.01	-15.0191±0.05
	FLIC	8.0048±0.16	6.1365±0.18	-37.8688±0.49
	FedGAN	6.8925±0.08	5.9507±0.04	-27.5028±0.14
	FedMDH	**0.9535±0.03**	**0.8364±0.02**	**0.0956±0.06**
Ferroalloy	FedAvg	7.9698±0.04	7.3643±0.07	-51.2793±0.07
	FedProx	5.1159±0.04	4.0385±0.07	-20.5413±0.08
	FedKD	5.1978±0.99	4.6955±1.02	-21.4532±4.72
	FedGen	5.2449±0.05	4.4994±0.08	-22.6871±0.10
	FLIC	7.1022±.02	5.8519±0.04	-45.4407±0.05
	FedGAN	6.1769±1.43	5.9712±1.47	-30.4036±4.53
	FedMDH	**0.9285±0.03**	**0.7254±0.03**	**0.1407±0.06**

Result Analysis

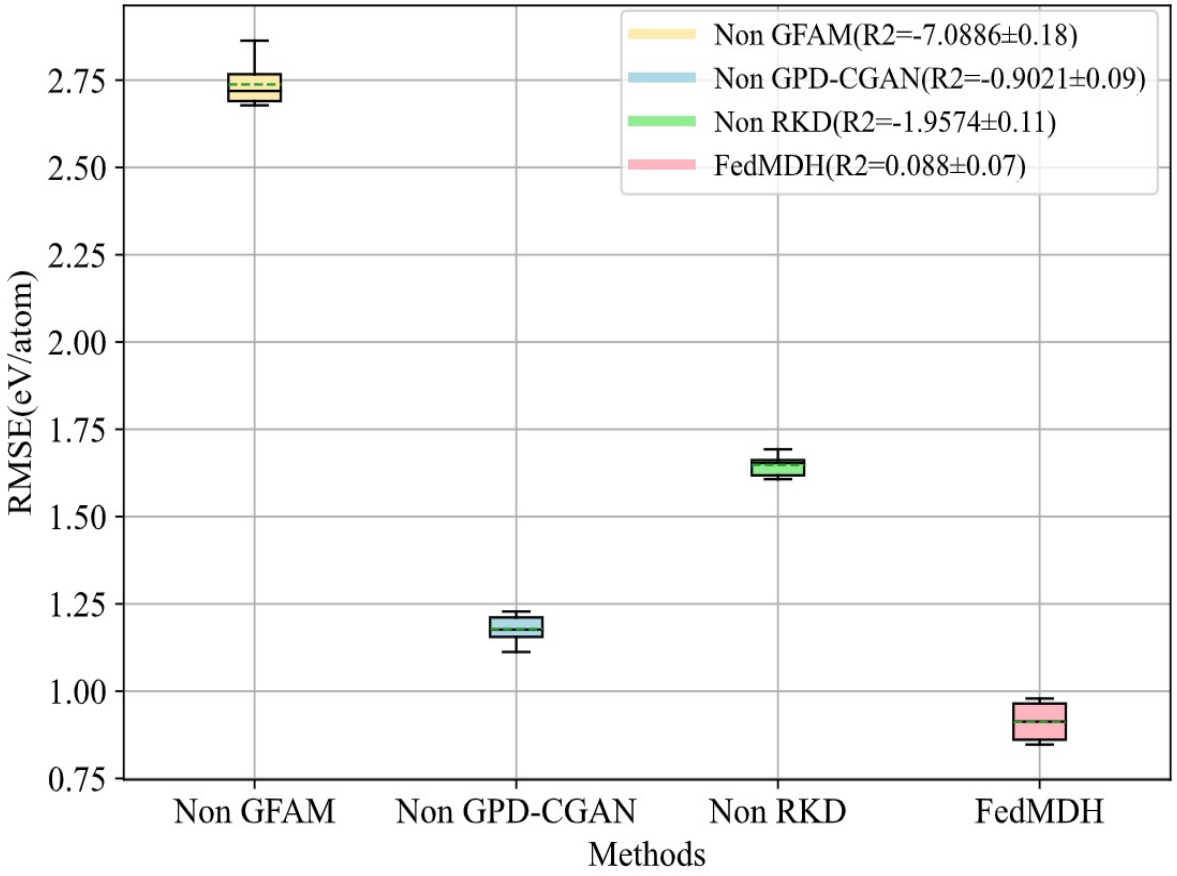
- ***FedMDH achieved the lowest RMSE within the different datasets, which loss curve showed stronger generalization and faster stabilization in environments with multi-dimensional heterogeneity.***
- ***Based on the R2 values from the three datasets, all comparative methods exhibit negative R2 values, indicating that these methods fail to effectively capture the characteristics of the data, resulting in poor predictive performance.***



The comparison of loss curves between FedMDH and other existing federated learning algorithms. (a) Perovskite Dataset; (b) Silicon Material Dataset; (c) Ferroalloy Dataset.

Experiment Evaluation——Ablation

Experiment Results



Results Analysis

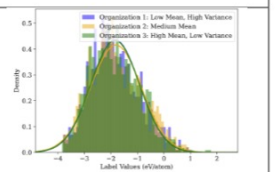
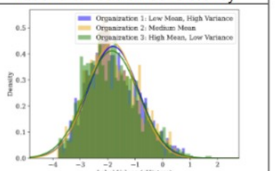
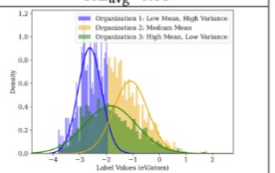
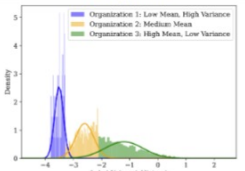
GFAM Model: Removing the GFAM model resulted in an average increase of 1.82 eV/atom in RMSE, highlighting its importance in aligning feature spaces across clients.

GPD-CGAN Model: The GPD-CGAN model generates synthetic samples that supplement the local datasets, providing diversity and helping to mitigate the impact of data scarcity or imbalance.

RKD-Prediction Model: The RKD component helps to ensure that local models not only align their predictions with the global model but also adjust to the global label distribution.

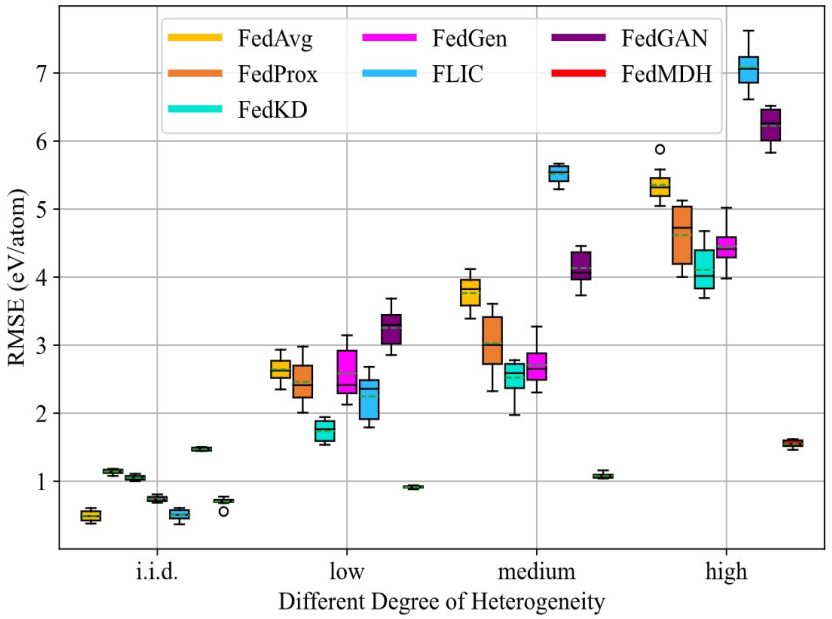
Experiment Evaluation——the Degree of MDH

Degree of MDH

Multi-dimensional Heterogeneity Degree	Feature Space Disparity	Sample Imbalance	Label Distribution Variance
I.I.D.	feature space consistency	sample balance	 label distribution consistency
Low Degree of Heterogeneity	$\mathcal{H}$=0.81	α = 1.09, and the sample size difference does not exceed 10%.	 KL_{avg}=0.06
Medium Degree of Heterogeneity	$\mathcal{H}$=0.31	α = 3.02, and the sample size ratio: 1:2:3.	 KL_{avg}=0.52
High Degree of Heterogeneity	$\mathcal{H}$=0.03	α = 15.01, and the sample size ratio: 1:10:15.	 KL_{avg}=0.88

The summary table of the degree of multi-dimensional heterogeneity in the training set.

Experiment Results



Results Analysis

- ***I.I.D:*** *FedMDH performs on par with the other methods in this ideal setting, demonstrating that it is well-equipped to handle uniform data distributions without sacrificing performance.*
- ***Low:*** *FedMDH starts to distinguish itself, maintaining lower RMSE compared to other methods.*
- ***Medium:*** *FedMDH continues to achieve significantly lower RMSE values, underscoring its superior capability in data multi-dimensional heterogeneity across clients.*
- ***High:*** *FedMDH continues to maintain a much lower RMSE compared to all other methods. This underscores FedMDH′ s strength in handling high levels of multi-dimensional heterogeneity, making it one of the most effective models for complex regression task environments.*

Welcome to NMDMS

NMDMS

Swarm Learning-Driven Material Genome Engineering Big-Data Sharing