OPEN ACCESS

Artificial Intelligence ARTICLE

The Innovation Drug Discovery

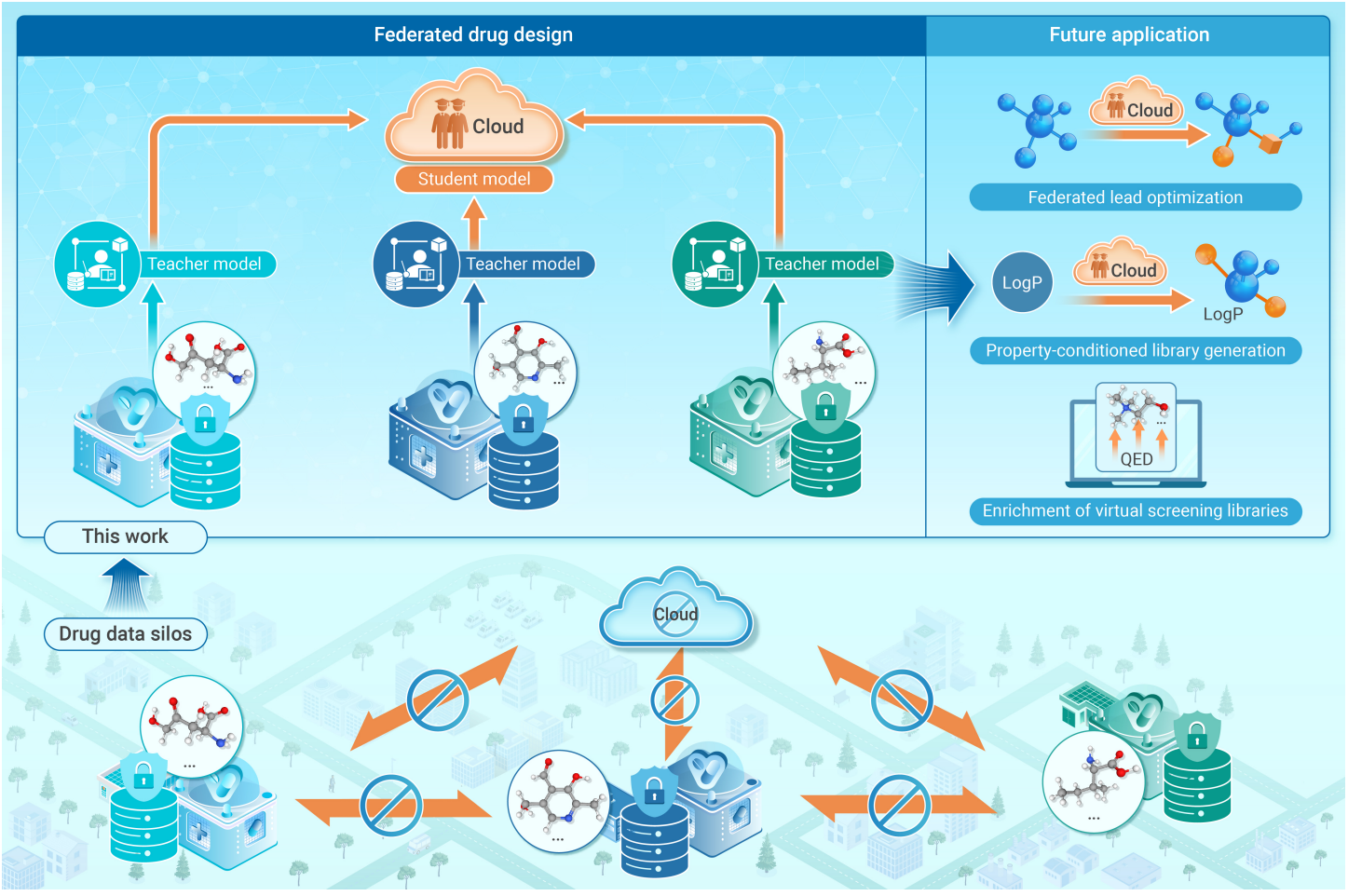
Cost-effective federated molecular design with FedMol

Tiejun Dong,[1] Linlin You,[1,4,*] and Calvin Yu-Chian Chen[2,3,*]

*Correspondence: youllin@mail.sysu.edu.cn (L.Y.); cy@pku.edu.cn (C.Y.C.)

Received: January 21, 2026; Accepted: June 2, 2026; Published Online: June 11, 2026; https://doi.org/10.59717/j.xinn-drugdisc.2026.100022

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- A federated framework for collaborative training of molecular generative models.
- Combining low rank adaptation (LoRA) with knowledge distillation (KD).
- FedMol achieves superior performance under high data heterogeneity, maintaining better distribution matching.

ARTICLE Artificial Intelligence

OPEN ACCESS

Cost-effective federated molecular design with FedMol

Tiejun Dong,[1] Linlin You,[1,4,*] and Calvin Yu-Chian Chen[2,3,*]

[1]Intelligent Medical Research Center, School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China

[2]School of AI for Science, Peking University Shenzhen Graduate School, Shenzhen 518055, China

[3]State Key Laboratory of Chemical Oncogenomics, Key Laboratory of Chemical Genomics, Peking University Shenzhen Graduate School, Shenzhen 518055, China

[4]Guangdong Provincial Key Laboratory of Intelligent Transportation System, School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 510275, China

*Correspondence: youllin@mail.sysu.edu.cn (L.Y.); cy@pku.edu.cn (C.Y.C.)

Received: January 21, 2026; Accepted: June 2, 2026; Published Online: June 11, 2026; https://doi.org/10.59717/j.xinn-drugdisc.2026.100022

Citation: Dong T., You L. and Chen C. (2026). Cost-effective federated molecular design with FedMol. *The Innovation Drug Discovery* **1**:100022.

Generating molecules with properties of interest is a fundamental challenge in computational molecular design. However, a significant portion of valuable chemical data resides within proprietary corporate data silos, limiting the potential of data-driven approaches for molecular design. Federated learning provides a framework for collaborative model training without sharing raw data, but existing methods face challenges in communication efficiency and robustness to heterogeneous data distributions. Here we introduce FedMol, a federated learning framework that combines pre-trained molecular language models with parameter-efficient fine-tuning (using low-rank adaptation) and knowledge distillation to enable privacy-preserving molecular generation. We evaluate FedMol on the GuacaMol distribution-learning benchmarks under varying degrees of data heterogeneity simulated via Dirichlet partitioning. Our results show that FedMol achieves competitive performance with state-of-the-art centralized models by updating only 2.4% of parameters through low-rank adaptation. Under high heterogeneity, knowledge distillation outperforms standard federated averaging in preserving chemical diversity. This work demonstrates a practical approach to collaborative molecular design that respects data privacy, with potential applications in drug discovery where proprietary data and multi-institutional collaboration are both critical.

INTRODUCTION

The discovery of novel therapeutic molecules remains a fundamental challenge in pharmaceutical research, traditionally relying on iterative cycles of hypothesis-driven design, synthesis, and biological testing.[1,2] While computational approaches have significantly accelerated this process, the emergence of deep generative models, particularly those based on transformer architectures, has revolutionized *de novo* molecular design by learning to generate chemically valid molecules directly from data.[3,4] These models, often pre-trained on millions of publicly available compounds, can capture intricate patterns of chemical space and propose molecules with desired properties. However, their full potential is constrained by a critical bottleneck: valuable data can be distributed across multiple institutions, each bound by strict privacy regulations and intellectual property protections that preclude data sharing.[5–8]

Federated learning offers a promising paradigm to overcome this limitation by enabling collaborative model training without centralizing raw data.[9,10] In typical federated learning, a global model is iteratively refined through updates computed on decentralized, private datasets. While federated learning has been explored for quantitative structure-activity/property relationship (QSAR/QSPR) modeling,[7,8,11,12] such as in toxicity prediction[13], its application to *de novo* molecular generation remains largely unexamined. Extending federated learning to large-scale molecular generation introduces three interconnected challenges: (1) the communication overhead of repeatedly transmitting millions of model parameters between clients and a central server[14], (2) the need to preserve data privacy not only at the raw data level but also during model updates,[15] and (3) the statistical heterogeneity (non-IID data) across clients, as different institutions often specialize in distinct therapeutic areas, leading to skewed local data distributions that can degrade global model performance.[16]

Recent advances in parameter-efficient fine-tuning, notably Low-Rank Adaptation (LoRA),[17] provide a compelling solution to the first challenge. By freezing pre-trained model weights and injecting trainable low-rank matrices into transformer layers, LoRA reduces the number of trainable parameters, dramatically cutting training costs while retaining most of the full fine-tuning performance. Concurrently, Knowledge Distillation (KD) has been explored as a privacy-enhancing technique in federated learning,[18] where local "teacher" models transfer knowledge to a global "student" through softened output distributions rather than sharing raw data or parameters, offering inherent privacy benefits.

In this work, we introduce FedMol, a framework that synergistically integrates a pre-trained molecular language model backbone with LoRA and KD to address the triad of training efficiency, privacy preservation, and robustness to data heterogeneity in federated molecular design. This integration yields cost-effectiveness through reduced model parameters, lower GPU memory consumption, and accelerated per-client training. FedMol operates in two stages: first, a transformer-based decoder is pre-trained centrally on a large corpus of public molecules to learn a general-purpose chemical language representation. Second, in a federated setting, each client fine-tunes this backbone on its private data using LoRA, then distills its specialized knowledge into a global student model via KD. This design ensures that clients never share raw data or even full model updates, only softened probability distributions that obscure the original inputs.

We rigorously evaluate FedMol on the GuacaMol distribution-learning benchmarks,[19] simulating various degrees of client heterogeneity via a Dirichlet distribution. Our results demonstrate that under near-IID conditions, FedMol achieves competitive performance with state-of-the-art centralized models, while outperforming a standard FedAvg baseline. More importantly, under extreme heterogeneity, FedMol maintains superior distribution matching by effectively aggregating specialized client knowledge through distillation. Ablation studies confirm that LoRA applied to all linear layers offers an optimal trade-off, maintaining strong performance in comparison to full fine-tuning while only updating 2.4% of the parameters. Experiments show that FedMol can outperform parameter averaging (FedAvg) in preserving the global chemical distribution under high heterogeneity. By offering an efficient, privacy-preserving approach to collaborative molecular design, FedMol may help advance the application of federated learning in drug discovery contexts where data privacy and collaboration are both important considerations.

MATERIALS AND METHODS

Language model backbone

The network architecture of FedMol is a pre-trained transformer-style decoder that can autoregressive generate molecules by representing them as Simplified Molecular Input Line Entry System (SMILES), where the model predicts the next token (*i.e.*, SMILES character) in a sequence based solely on preceding tokens. The model was pre-trained on a large molecule dataset, allowing the model to be easily fine-tuned on local datasets (Figure 1).

The backbone is a stack of decoder-only Qwen2 network blocks,[20] where each block consists of a multi-head self-attention module followed by a feed-forward module. To preserve the autoregressive property, a causal (unidirectional) masking mechanism is strictly applied in the self-attention layers, preventing any token from attending to future tokens in the sequence. For an input sequence $\mathbf{s} \in \mathbb{R}^{N_{seq} \times d_{model}}$, each head projects it to queries ($\mathbf{q}$), keys ($\mathbf{k}$), and values ($\mathbf{v}$) with linear layers, where $\mathbf{q}, \mathbf{k}, \mathbf{v} \in \mathbb{R}^{N_{seq} \times d_{head}}$ and $d_{head} = d_{model}/N_{head}$. N_{seq} and N_{head} are the length of the sequence and the number of attention heads, respectively. For each head, we compute the scaled dot-product attention and the output $\mathbf{h}_i$ for head i is:

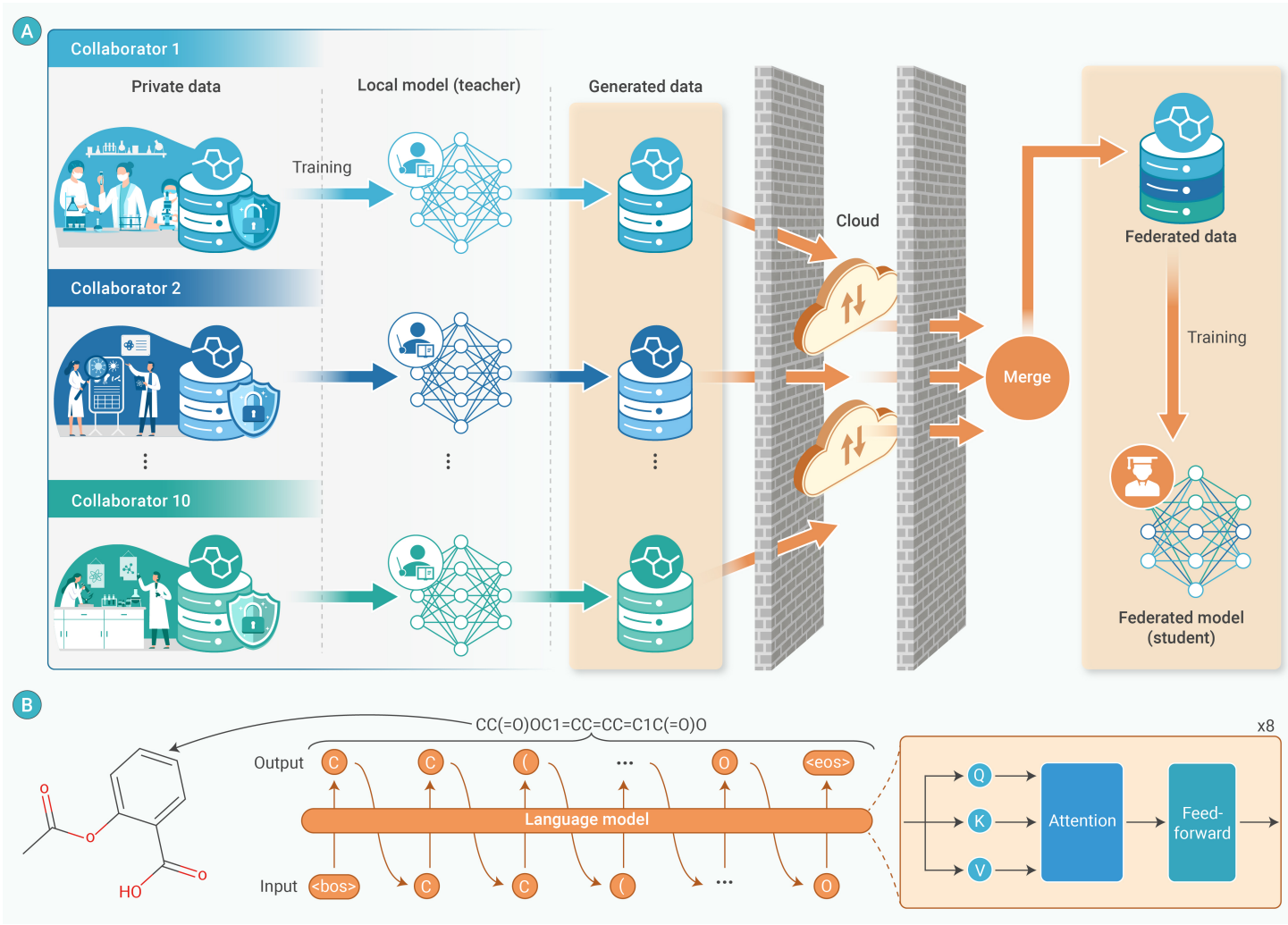


Figure 1. Overview of the FedMol framework (A) The federated KD process. Each client trains a local teacher model on its private data to generate a novel molecule set. These sets are aggregated on a central server to train a global student model, facilitating collaborative learning without raw data exchange. (B) The model architecture employs a transformer-based language model as its backbone, which autoregressively generates molecules represented as SMILES strings.

$$\mathbf{h}_i = \text{softmax}_i\left(\frac{\mathbf{q}_i\mathbf{k}_i^\top}{\sqrt{d_{head}}}\right)\mathbf{v}_i \quad (1)$$

The outputs of all heads are concatenated and then projected back to the model dimension via linear transformation. The feed-forward module that follows is a two-layer perceptron with SwiGLU activation. These network layers are integrated into each transformer block with residual connection and RMSNorm for layer normalization.

Besides, Qwen2 network blocks employ Grouped-Query Attention (GQA) and Rotary Position Embedding (RoPE) to improve model performance.[21,22] Unlike standard multi-head attention where query, key, and value heads are equal in number, GQA employs groups of key-value heads that are shared across multiple query heads. This design significantly reduces the memory footprint during autoregressive generation, particularly for long sequences. RoPE encodes the absolute position of a token in a sequence by rotating its query and key vectors in the complex plane with a frequency-based transformation, which has been shown to enhance model stability and performance on long sequences.

Four properties are considered for conditional generation: Quantitative Estimate of Drug-likeness (QED) score, Synthetic Accessibility (SA) score, LogP, Topological Polar Surface Area (TPSA). Each property value is encoded using a radial basis function (RBF) with d_{model} equally spaced Gaussian kernels. The encoding ranges for QED, SA, LogP, and TPSA are $[0,1]$, $[1,10]$, $[-2,10]$, and $[0,200]$, respectively. For every possible combination of properties (from none to all four), a dedicated condition token is introduced, resulting in a total of 16 extra tokens, including 'QED', 'SA', 'LogP', 'TPSA', and combinations such as 'QED_SA', 'QED_LogP', and so forth, up to 'QED_SA_LogP_TPSA'. At the beginning of each SMILES sequence, we prepend four copies of the token that matches the desired combination. The embeddings of these four tokens are then augmented by adding the RBF encoding of each property in a fixed order: the first token receives QED encoding (if QED is in the combination), the second receives SA (if included), the third receives LogP (if included), and the fourth receives TPSA (if included). Properties not in the combination contribute no addition. In this way, the four tokens collectively convey both the conditioned property set and their target values, enabling multi-property constrained generation.

Low-rank adaptation

To efficiently adapt the pre-trained language model to our federated molecular design while mitigating computational constraints, we employed LoRA.[17] This parameter-efficient fine-tuning strategy avoids the prohibitive cost of updating all model parameters by introducing a minimal set of trainable weights into the transformer architecture.

The core principle involves freezing the pre-trained model weights and injecting trainable low-rank matrices into the selected linear layers. This decomposes the weight update (ΔW) for a given layer into two small matrices ($B \cdot A$), where the rank (r) is significantly smaller than the original dimension. The forward pass becomes:

$$\mathbf{h} = W_0\mathbf{s} + \Delta W\mathbf{s} = W_0\mathbf{s} + BA\mathbf{s} \quad (2)$$

where $W_0 \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ represents the frozen pre-trained weights, and

$A \in \mathbb{R}^{r \times d_{model}}$, $B \in \mathbb{R}^{d_{model} \times r}$ are the trainable adapters. The rank $r \ll d_{model}$. A was initialized with a random Gaussian distribution and B was initialized to zeros, ensuring $\Delta \mathbf{W} = 0$ at the beginning of training.

Knowledge distillation

As shown in Figure 1A, We adopt a teacher-student paradigm where the global model acts as the student, and each client's local model serves as a teacher. The distillation process transfers not only the hard labels (ground-truth next tokens) but also the soft probabilistic knowledge (output distributions) from the teacher to the students, thereby regularizing global updates with the client consensus.

Let $\mathcal{D}_k$ denote the private molecular dataset on client k. For a given input sequence $\mathbf{s}$ with its corresponding ground-truth next token y, the local model produces a logit vector $\mathbf{z}_k \in \mathbb{R}^{|\mathcal{V}|}$, where $\mathcal{V}$ is the SMILES token vocabulary set. The model is trained with the standard cross-entropy loss for autoregressive language modeling:

$$\mathcal{L}_{\text{CE}} = -\log \frac{e^{z_{k,y}}}{\sum_{j \in V} e^{z_{kj}}}. \tag{3}$$

After local training, the kth teacher model generates a softened probability distribution over the vocabulary by applying a temperature scaling factor τ:

$$p_{\text{teacher}} = \frac{e^{z_{kj}/\tau}}{\sum_{j \in V} e^{z_{kj}/\tau}} \tag{4}$$

Similarly, the global student model produces its own softened distribution p_{student}. The distillation loss is then defined as the Kullback–Leibler (KL) divergence between the two softened distributions:

$$\mathcal{L}_{\text{kd}} = \tau_{KD}^2 \cdot D_{\text{KL}} \left(p_{\text{teacher}} \parallel p_{\text{student}} \right) \tag{5}$$

where the τ_{kd}^2 scaling compensates for the magnitude change of gradients when using high temperatures. This loss encourages the student model to retain the general knowledge embedded in the teacher while adapting to its specific data distribution.

Model training

The overall training of FedMol consists of two main stages: centralized pre-training on a large public molecular dataset to learn a general-purpose molecular language representation, followed by federated fine-tuning on distributed private datasets using a KD strategy to achieve privacy-preserving collaborative learning. This two-stage approach ensures the model first acquires a strong prior on chemical space before being efficiently and confidentially adapted to specialized, decentralized data distributions.

Pre-training. The network backbone was pre-trained in a centralized manner on the large-scale public dataset using a standard next-token prediction objective. Training was conducted using the AdamW optimizer with a constant learning rate. A linear learning rate warmup was applied for the first 100 steps to a peak learning rate of 5×10^{-4}, which reduced to 1×10^{-4} after 6×10^4 iterations. Training proceeded for a total of 7×10^4 iterations with a global batch size of 128.

Federated fine-tuning. We frame the federated learning process within a teacher-student paradigm to facilitate privacy-preserving knowledge transfer.

(1) Local teacher training. The server distributes the pre-trained model parameters to all participating clients. Each client fine-tunes it exclusively on its private dataset $\mathcal{D}_k$. Fine-tuning employs LoRA for parameter efficiency, updating only the injected low-rank adapters while keeping the pre-trained backbone frozen. The local optimization uses the cross-entropy loss $\mathcal{L}_{\text{CE}}$ with a learning rate of 1×10^{-3}. After local training, each teacher model k generates a set of molecules. Instead of sharing raw data, each client uploads the set of softened distributions p_{teacher} derived from its generated molecules. This design aligns with enhanced privacy principles, as transmitting vectorized generated molecules and soft labels inherently obscures the original raw data.

(2) Global student distillation. The server aggregates the softened distributions received from all clients. The global student model is then updated by minimizing a composite objective that combines the standard cross-entropy loss ($\mathcal{L}_{\text{CE}}$) and the KD loss ($\mathcal{L}_{\text{KD}}$), thereby learning from both the ground-truth labels and the consensus of the client teachers. During distillation, data are sampled with a probability weighted by the size of each client's local dataset. The student is optimized using a learning rate of 1×10^{-3}. During KD, each client generates a total of 2×10^4 distillation data samples, which takes approximately 80 minutes on a single NVIDIA RTX A4000 GPU device.

(3) FedAvg baseline. For comparative ablation studies, we implement a standard FedAvg algorithm. In this baseline, clients transmit their updated LoRA adapter parameters ΔW_k to the server after local training. The server updates the global model parameters θ_G via a weighted average: $\theta_G \leftarrow \theta_G + \sum_{k=1}^{K} w_k \Delta W_k$, where the weight w_k is proportional to the size of the kth local dataset. Communication and aggregation occur every 100 local training iterations. Learning rates for FedAvg were set as follows: 1×10^{-3} for LoRA-based fine-tuning, 1×10^{-4} for fine-tuning pre-trained models without LoRA, and 3×10^{-4} for training from scratch without LoRA.

(4) FedProx baseline. We also implement FedProx algorithm,[23] an extension of FedAvg that adds a proximal term to the local objective. FedProx modifies the local training loss as:

$$\mathcal{L}_{\text{Prox}} = \mathcal{L}_{\text{CE}} + \frac{\mu}{2} \|\theta - \theta_G\|^2, \tag{6}$$

where θ denotes the local model parameters, θ_G is the current global model, and μ is a hyperparameter controlling the strength of the proximal penalty. This term discourages local updates from deviating too far from the global model, mitigating client drift. In our experiments, FedProx follows the same training procedure as FedAvg (including LoRA-based fine-tuning, communication frequency, and learning rates), with the proximal coefficient set to $\mu = 0.001$.

Unless otherwise specified, all experiments were conducted for 1×10^4 training iterations with a global batch size of 128, using the AdamW optimizer with a warmup of 100 steps. In the distillation-based framework, the global student model is trained after all client teachers have completed their local training phase.

Molecule collection and preprocessing

Molecules were collected from the PubChem and ChEMBL databases.[24,25] Additionally, EGFR inhibitors were collected from the ChEMBL database by filtering with UniProt ID (P00533, EGFR_HUMAN) and a 'pChEMBL Value' greater than 6.0. Following ref.[3], we adopted the same regular expression to break the SMILES strings into "tokens", which are the basic units processed by the language model. Molecules containing more than 512 heavy atoms or 1024 tokens were excluded. Finally, 10,995,977 and 1,922,158 molecules from the PubChem and ChEMBL databases were used in this study, with 15,911 molecules treated as EGFR inhibitors. Figure 2A-E show the distributions of molecular weight (MW), partition coefficient (LogP), topological polar surface area (TPSA), synthetic accessibility (SA) score, and quantitative estimate of drug-likeness (QED) score.

Simulation of heterogeneous data

To rigorously evaluate the robustness of our federated learning framework under realistic and challenging conditions, we simulate non-independent and identically distributed (non-IID) data partitions across clients. This setup reflects practical scenarios in collaborative drug discovery, where participating institutions often possess molecular libraries that are biased toward specific chemical profiles or therapeutic targets.

We construct heterogeneous data distributions across clients by partitioning the dataset based on the SA score. To implement a label distribution skew, we first discretize the continuous SA scores into 16 bins of equal width, spanning the full range from 1.0 to 5.0. Each molecule is then assigned a categorical label corresponding to its SA score bin. Client data allocations are subsequently sampled from these labeled subsets according to a Dirichlet distribution,[16] with a parameter α that controls the degree of heterogeneity. A smaller α value induces a high degree of data skew, resulting in highly imbalanced label distributions across clients, whereas a larger α yields more uniform, near-IID partitions. Similarly, we further construct a two-dimensional heterogeneous setting by jointly considering SA score and TPSA. Specifically, we discretize SA scores into four equal-width bins (range 1.0-

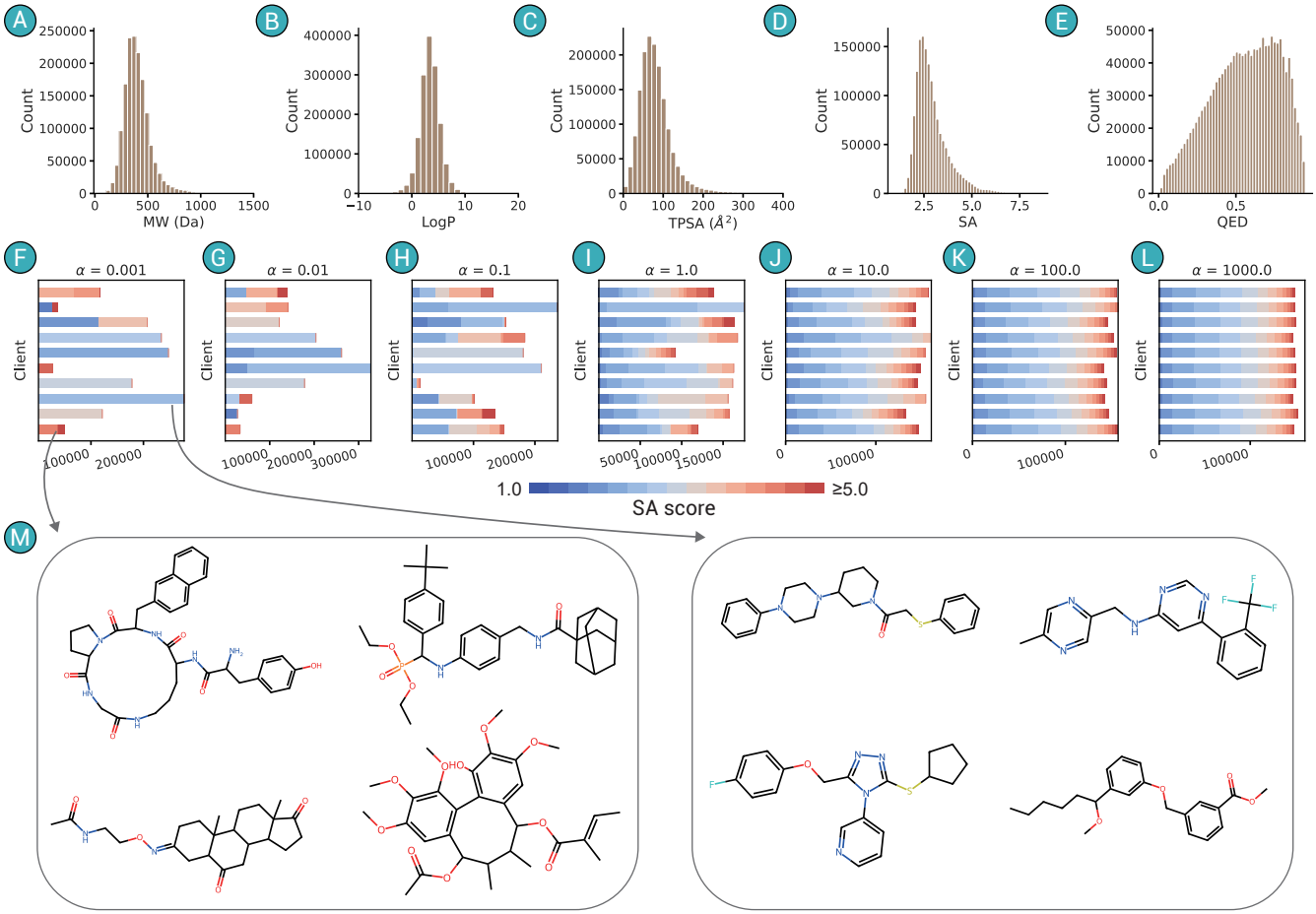


Figure 2. Data distribution and simulation of client heterogeneity (A-E) Distribution of molecular properties in the ChEMBL dataset: MW (A), LogP (B), TPSA (C), SA score (D), and QED score (E). (F-L) Simulation of non-IID data partitions across clients, generated by sampling from the SA-score distribution with a Dirichlet concentration parameter α ranging from 0.001 to 1000.0. Decreasing α increases SA distribution skew, resulting in more heterogeneous data allocations across clients. (M) Randomly sampled molecules derived from two clients under the high-heterogeneity setting ($\alpha = 0.001$).

5.0) and TPSA into four equal-width bins (range 0.0-240.0Å), forming a total of 16 combined categories. Each molecule is assigned a two-dimensional label based on its SA and TPSA bins. This approach allows us to systematically investigate the performance of our method across varying levels of data heterogeneity. In this study, each molecule is assigned to one of $K = 10$ clients based on the different α. The simulated one-dimensional heterogeneous data split for a range of α values is shown in Figure 2F–L.

Privacy leak analysis

To quantitatively assess the privacy guarantees of our framework, we conduct membership inference attacks (MIA) following a similar approach to previous studies. [26,27] MIA aims to determine whether a given data sample was part of a target model's training set. In our setting, a potential adversary could access the teacher or student models and the distillation set during the global training phase. We apply a threshold-based attack using perplexity.

The perplexity-based attack serves as an effective baseline for assessing privacy leakage. It exploits the observation that machine learning models typically assign higher likelihoods (lower perplexity) to samples encountered during training compared to unseen samples. For a given input sequence $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, where each x_t is a SMILES token, the model outputs a conditional probability distribution $p_\theta(x_t \mid x_{<t})$ over the vocabulary at each step. The perplexity (PPL) of the sequence is defined as:

$$\mathrm{PPL}(\mathbf{x}) = \exp\left(-\frac{1}{n}\sum_{t=1}^{n}\log p_\theta(x_t \mid x_{<t})\right). \quad (7)$$

Lower perplexity indicates that the model is more "confident" about the sequence, which is often correlated with membership in the training set.

The attack proceeds as follows: given a target sequence $\mathbf{x}$, the adversary computes its perplexity under the target model. If $\mathrm{PPL}(\mathbf{x}) < \tau$, the sample is classified as a member (i.e., part of the private data); otherwise, it is classified as a non-member. The threshold τ is selected on a held-out validation set by maximizing attack accuracy ($\tau = 1.70$ in this study).

Evaluation metrics

Model performance was quantified using five established metrics from the GuacaMol distribution-learning benchmark.[19] Validity: fraction of generated SMILES strings that correspond to chemically valid molecules. Uniqueness: proportion of unique molecules among valid generations. Novelty: fraction of valid, unique molecules not present in the training dataset. Kullback-Leibler Divergence (KLD): Kullback-Leibler divergence between the distributions of physicochemical descriptors (MW, LogP, TPSA, etc.) of generated versus training molecules. Higher KLD values indicate better distribution matching. Fréchet ChemNet Distance (FCD): distance between the hidden representations of the ChemNet for generated and training molecules.[28] Higher values indicate closer distributional similarity.

RESULTS

Centralized training baseline

To establish performance benchmarks and evaluate the effectiveness of our components in isolation, we first conducted experiments under centralized (non-federated) conditions to assess the individual contributions of pre-training and LoRA before their integration into the federated KD framework,

Table 1. Baseline performance on GuacaMol distribution-learning benchmarks under centralized condition

Pre-training	Fine-tuning	Validity	Uniqueness	Novelty	KLD	FCD
√	×	0.983	0.999	0.994	0.920	0.061
×	Full fine-tuning	0.922	0.999	0.973	0.969	0.825
√	Full fine-tuning	0.983	0.999	0.968	0.984	0.896
×	LoRA (all linear layers)	0.732	0.998	0.985	0.960	0.745
√	LoRA (all linear layers)	0.974	0.999	0.978	0.979	0.869
×	LoRA (query, key and value)	0.184	0.976	0.994	0.766	0.032
√	LoRA (query, key and value)	0.960	0.999	0.982	0.974	0.810

which involves training multiple models concurrently.

The results presented in Table 1 reveal the benefit of pre-training on molecular generation performance. The pre-trained model achieves near-perfect validity (0.983), uniqueness (0.999), and novelty (0.994), indicating that the pre-trained model has already learned a robust prior distribution over chemical space. However, this configuration exhibited suboptimal distribution matching, as evidenced by lower KLD (0.920) and FCD (0.061) scores, suggesting that while the pre-trained model generates high-quality molecules, their distribution does not perfectly match the target dataset.

Besides, the LoRA strategy shows efficient performance while maintaining small trainable model weights. We evaluated three fine-tuning strategies to characterize the efficiency-performance trade-off: Full parameter fine-tuning (Updates 100% parameters during fine-tuning, bandwidth cost: 95.6MB), LoRA on all linear layers (Updates only 2.4% of parameters, bandwidth cost: 2.32MB), LoRA on query, key, and value projections only (Updates only 0.8% of parameters, bandwidth cost: 0.79MB). When pre-trained models undergo full parameter fine-tuning, they achieve the performance ceiling for our architecture, with optimal distribution matching scores: KLD (0.984) and FCD (0.896). This configuration maintains excellent validity, uniqueness and novelty while significantly improving distribution alignment compared to the pre-trained-only baseline.

Remarkably, pre-trained models fine-tuned with LoRA on all linear layers achieve performance close to the upper bound while requiring updates to only 2.4% of parameters. Specifically, we observe minimal degradation in distribution matching. This demonstrates that the majority of specialized knowledge required for adapting to the target distribution can be captured through low-rank updates to linear transformations, validating LoRA's effectiveness for domain adaptation in molecular language models. The most parameter-efficient configuration, LoRA applied only to query, key, and value projections, maintains strong performance. With only 0.8% of parameters trainable, this configuration achieves 0.974 KLD, and 0.810 FCD, representing a viable option for scenarios with extreme computational or communication constraints. Based on this systematic analysis, which reveals an optimal balance between parameter efficiency and performance preservation, we selected the pre-trained backbone with LoRA applied to all linear layers as the default configuration for subsequent federated learning experiments.

Federated molecular generation under near-IID condition

We next evaluated our framework under near-IID distributed (near-IID) data conditions ($\alpha = 1000.0$). As shown in Figure 2L, this setting represents the idealized scenario where each client possesses data that closely approximates the overall population distribution, allowing us to assess the fundamental viability of our approach before addressing the challenges of statistical heterogeneity.

Figure 3A-E demonstrate that molecules generated by FedMol (green distributions) closely match the physicochemical property distributions of the target dataset (blue distributions) across five key metrics: MW, LogP, TPSA, SA score, and QED score. Figure 3F-M showcase our model's ability to generate molecules conditioned on single or multiple specific property values. This multi-constraint satisfaction capability is particularly valuable for practical drug discovery workflows where molecules must simultaneously satisfy multiple physicochemical and pharmacological criteria.

The results presented in Table 2 demonstrate that our federated learning framework achieves competitive performance with state-of-the-art centralized molecular generation models across all five evaluation metrics. When compared against sequential generation models, FedMol achieves superior validity (0.987) compared to LSTM (0.959), AAE (0.822), ORGAN (0.379), VAE (0.870), MolGPT (0.981), SmileyLlama (0.983), HYFORMER (0.979), and G2PT (0.953). In terms of distribution matching, FedMol attains KLD of 0.984 and FCD of 0.900, which are comparable to or exceed those of MolGPT (KLD: 0.992, FCD: 0.907), SmileyLlama (KLD: 0.985, FCD: 0.849), HYFORMER (KLD: 0.990, FCD: 0.916), and G2PT (KLD: 0.956, FCD: 0.927). Against graph-based generation models, FedMol demonstrates particularly strong performance, substantially outperforming VGAE (KLD: 0.554, FCD: 0.016), Graph MCTS (KLD: 0.522, FCD: 0.015), VGAE-MCTS (KLD: 0.659, FCD: 0.009), and RSG-VAE (KLD: 0.940, FCD: 0.800). This suggests that our language model-based approach, even when trained in a distributed manner, learns accurate representations of the underlying chemical distribution compared to many graph-based architectures trained in centralized settings.

FedMol consistently outperforms FedAvg across most metrics (Table 2, with particularly notable improvements in validity (0.987 vs. 0.977) and FCD (0.900 vs. 0.885). Besides, the full fine-tuning baseline achieves a validity of 0.983, KLD of 0.987, and FCD of 0.902. The validity gap is minimal, and distribution matching metrics remain highly comparable, demonstrating that our parameter-efficient approach with KD can nearly match the quality of full fine-tuning.

A finding emerges from the comparison between our proposed KD approach and the standard parameter averaging baselines. While FedAvg and FedProx both aggregate model parameters, with FedProx additionally introducing a proximal term to mitigate client drift, their performance under near-IID conditions is remarkably similar. As shown in Table 2, FedProx achieves a slightly higher validity (0.979/0.977) and FCD (0.891/0.885) than FedAvg, indicating comparable distribution matching. FedMol consistently outperforms FedAvg and FedProx across most metrics, with particularly notable improvements in validity and FCD. Besides, the performance achieved by FedMol under federated near-IID conditions approaches that of the centralized baseline with full parameter fine-tuning. The validity gap is minimal, while distribution matching metrics show comparable performance.

To further validate the generalizability of our framework, we evaluated FedMol on a more specialized dataset comprising EGFR inhibitors (Table 3). The pre-trained model without fine-tuning achieves high validity (0.981), uniqueness (0.999), and novelty (0.999) but exhibits poor distribution matching (KLD 0.604, FCD 0.006), highlighting the need for domain-specific adaptation. Under federated fine-tuning, both FedAvg and FedProx substantially improve distribution matching (KLD 0.97, FCD 0.66) at the cost of reduced uniqueness and novelty (0.57 and 0.40, respectively), indicating a trade-off between distribution alignment and generative diversity. Our proposed KD-based approach achieves slightly higher validity (0.989) and comparable distribution matching (KLD 0.974, FCD 0.673), closely approaching the performance of the centralized baseline, demonstrating that KD effectively transfers domain-specific knowledge while maintaining generative fidelity. These results are consistent with the observations from the ChEMBL dataset, suggesting that FedMol may be robust and generalizable across different molecular domains.

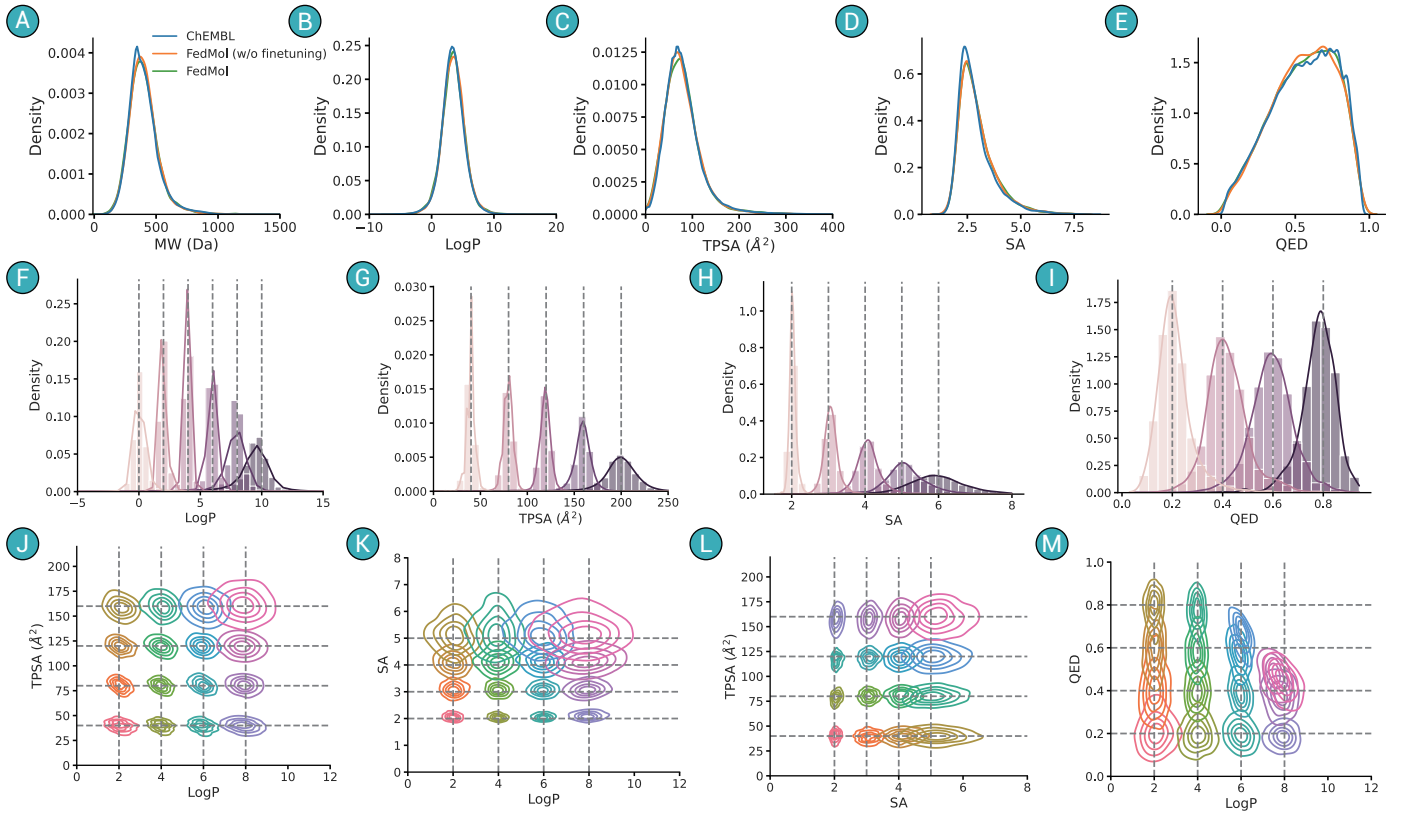


Figure 3. Molecuar generation with FedMol trained under near-IID condition (A-E) Comparative analysis of physicochemical property distributions between the target ChEMBL dataset and generated molecules. Blue distributions represent the reference ChEMBL dataset; orange distributions correspond to molecules generated by the pre-trained model without federated fine-tuning; green distributions show molecules generated by FedMol after federated training. Properties analyzed include: MW (A), LogP (B), TPSA (C), SA score (D), and QED score (E). (F-I) Single-property conditional generation performance, demonstrating the model's ability to generate molecules with specified target values (gray vertical lines) for: LogP (F), TPSA (G), SA score (H), and QED score (I), respectively. The input conditions are shown in grey lines. (J-M) Dual-property conditional generation performance, illustrating simultaneous constraint satisfaction for: TPSA-LogP (J), SA-LogP (K), TPSA-SA (L), and QED-LogP (M), respectively. In all conditional generation panels, input conditions are indicated by gray lines.

We next evaluated the privacy-preserving property of our framework by conducting MIA. Table 4 reports the attack performance against different model variants. The pre-trained model without fine-tuning achieves almost identical metrics (accuracy 0.499, AUC 0.499, TPR@1%FPR 0.010), indicating that the base model does not memorise the private data. In contrast, the FedAvg variant exhibits substantially higher vulnerability, with an accuracy of 0.814, an AUC of 0.894, and a TPR@1%FPR of 0.357. This result suggests that simply parameter averaging can inadvertently leak membership information, as the model updates are influenced by private data. Our proposed KD-based approach considerably reduces the attack success, achieving an accuracy of 0.768, an AUC of 0.827, and a TPR@1%FPR of 0.324. The lower metrics indicate that KD, which exchanges only softened probability distributions rather than model parameters, provides stronger protection against membership inference. This privacy benefit can be attributed to the soft-label smoothing effect inherent in distillation, where the softened outputs reduce overfitting to individual training samples and thereby mitigate membership leakage.[38] The reduction in attack performance highlights the inherent privacy advantages of the distillation paradigm within our federated molecular design framework.

Although the molecules are vectorized when forming the distillation sets shared during federated KD, we performed a model-free analysis to assess whether an adversary could reconstruct private SMILES strings if the vectorization codebooks were known. Specifically, we calculated the exact match rate, the fraction of molecules in the distillation sets that also appear in the private dataset. This provides a lower bound on the difficulty of extracting private data without assuming a powerful attacker. As shown in Table 5, the exact match rate increases as the temperature decreases, indicating that a small number of private molecules can indeed be recovered from the distillation sets when low temperatures are used. evertheless, successful reconstruction critically requires prior knowledge of the vectorization codebook, as the mapping from token indices back to SMILES characters is not publicly disclosed in practice. When combined with moderate distillation temperatures, the probability of successful reconstruction becomes negligible, making the framework suitable for privacy-sensitive collaborations.

Federated molecular generation under non-IID condition

Having validated our framework under idealized near-IID conditions, we next evaluated its robustness across a spectrum of data heterogeneity, which more accurately reflects real-world collaborative drug discovery scenarios. Using the Dirichlet-based partitioning scheme described in methods (Figure 2F-L), we systematically varied the concentration parameter α from 0.001 (extreme heterogeneity) to 1.0 (moderate heterogeneity) to quantify the impact of statistical distribution skew on federated molecular generation.

We evaluated the performance of FedMol using KD alongside the standard FedAvg algorithm across this one-dimensional heterogeneity spectrum based on SA score (α ranging from 0.001 to 1.0). As shown in Figure 4A-E, increasing data heterogeneity progressively challenges both models, though to different degrees. Under extreme heterogeneity ($\alpha = 0.001$), FedMol maintains a validity of 0.971 compared to 0.985 for FedMol using FedAvg, while the distribution matching metrics (KLD and FCD) show a different trend: FedMol achieves a KLD of 0.9723 and FCD of 0.8166, whereas FedMol using FedAvg achieves 0.9327 and 0.7967, respectively. As heterogeneity decreases (α increases), both methods improve, with FedMol consistently outperforming FedMol using FedAvg in distribution matching. At $\alpha = 1.0$, FedMol achieves a KLD of 0.9774 and FCD of 0.8789, while FedAvg reaches 0.9674 and 0.8689, respectively. The validity for both methods remains high

Table 2. Results of GuacaMol distribution-learning benchmarks on the ChEMBL dataset under near-IID condition

Model	Validity	Uniqueness	Novelty	KLD	FCD
LSTM	0.959	1.000	0.912	0.991	0.913
AAE	0.822	1.000	0.998	0.886	0.529
ORGAN	0.379	0.841	0.687	0.267	0.000
VAE	0.870	0.999	0.974	0.982	0.863
MolGPT	0.981	0.998	1.000	0.992	0.907
SmileyLlama	0.983	0.999	0.962	0.985	0.849
HYFORMER	0.979	1.000	0.904	0.990	0.916
G2PT	0.953	1.000	0.995	0.956	0.927
VGAE	0.830	0.944	1.000	0.554	0.016
Graph MCTS	1.000	1.000	0.994	0.522	0.015
VGAE-MCTS	1.000	1.000	1.000	0.659	0.009
hG2G	1.000	0.995	0.940	0.888	0.506
MGM	0.849	1.000	0.722	0.987	0.845
MiCaM	1.000	1.000	0.986	0.989	0.731
RSG-VAE	1.000	1.000	0.996	0.940	0.800
FedMol (FedAvg)	0.977	0.999	0.976	0.982	0.885
FedMol (FedProx)	0.979	0.999	0.979	0.977	0.891
FedMol (Full fine-tuning)	0.983	0.999	0.959	0.987	0.902
FedMol	0.987	0.999	0.967	0.984	0.900

*All baseline models, with the exception of FedMol, were trained in a centralized manner. *Results of LSTM, AAE, ORGAN, VAE, Graph MCTS are obtained from ref.[19]. Results of MolGPT, SmileyLlama, HYFORMER, G2PT, VGAE, VGAE-MCTS, hG2G, MGM, MiCaM, RSG-VAE are obtained from ref.[3], ref.[29], ref.[30], ref.[31], ref.[32], ref.[33], ref.[34], ref.[35], ref.[36], ref.[37], respectively.

Table 3. Results of GuacaMol distribution-learning benchmarks on the EGFR dataset under near-IID condition

Model	Validity	Uniqueness	Novelty	KLD	FCD
FedMol (w/o fine-tuning)	0.981	0.999	0.999	0.604	0.006
FedMol (FedAvg)	0.978	0.569	0.393	0.972	0.661
FedMol (FedProx)	0.985	0.574	0.408	0.977	0.663
FedMol (centralized)	0.990	0.577	0.412	0.969	0.657
FedMol	0.989	0.576	0.391	0.974	0.673

Table 4. MIA against FedMol under near-IID conditions. Experiments were conducted on a balanced dataset containing both member and non-member samples. Non-member samples were randomly sampled from the PubChem dataset

Reference Model	Accuracy (τ = 1.70)	AUC	TPR@1%FPR
Random Guess	0.500	0.500	0.010
FedMol (w/o fine-tuning)	0.499	0.499	0.010
FedMol (FedAvg)	0.814	0.894	0.357
FedMol	0.768	0.827	0.324

(0.9777 and 0.9813), with FedAvg version showing a slight advantage in validity but at the cost of distribution matching. This indicates that KD better preserves the overall chemical distribution even when local data are highly skewed.

To further assess the robustness of our framework under more complex and realistic scenarios, we extended the heterogeneity simulation to a two-dimensional setting by jointly considering SA score and TPSA. As shown in Figure 4F–J, the two-dimensional heterogeneity introduces additional complexity, making the task more challenging compared to the one-dimensional case. Under this setting, both FedAvg and our KD-based approach exhibit reduced distribution matching performance across all levels of heterogeneity, reflected in lower KLD and FCD values relative to their one-dimensional counterparts. Nevertheless, our method consistently outperforms FedAvg in preserving the target distribution. For instance, under extreme heterogeneity ($\alpha = 0.001$), FedMol achieves KLD = 0.921 and FCD = 0.763, substantially higher than FedAvg (KLD = 0.878, FCD = 0.736). These

results mirror the trends observed in the one-dimensional setting, further demonstrating the potential robustness and generalizability of our approach across both heterogeneity dimensions in realistic federated molecular design scenarios.

FedAvg averages model parameters, implicitly assuming that the optimal global model lies in a convex combination of client updates. Averaging these divergent parameters leads to "client drift",[39] producing a global model that occupies an "average" region of chemical space that may not correspond to any valid mode of the true distribution. In contrast, KD aggregates softened probability distributions over the vocabulary space.[18] When each client teacher generates molecules and produces soft targets, the server-side student learns from the consensus of output distributions rather than from conflicting parameter updates, which makes FedMol achieve higher FCD and KLD. As shown in Figure 4F-O, each teacher model successfully specializes to its local SA distribution, indicating close adaptation to local chemical profiles. These results suggest that KD aggregates softened probability distributions, creating a consensus over the chemical space that is more robust to distributional skew.

Table 5. Quantifying potential private molecule leakage from distillation sets under near-IID conditions. The exact match rate is defined as the proportion of molecules in the distillation sets that are also present in the private dataset

Temperature	Exact match rate (%)
0.1	6.050
0.5	3.740
1.0	2.330
1.5	1.548
2.0	0.756

DISCUSSION

Our investigation establishes FedMol as an effective framework for privacy-preserving molecular generation that addresses the inherent tensions between collaborative drug discovery and data confidentiality. The framework stems from the synergistic integration of three key components: a pre-trained language model backbone, KD for privacy-preserving aggregation and critically, LoRA for parameter-efficient fine-tuning. Together, these elements enable federated learning to achieve molecular generation quality comparable to centralized approaches while maintaining strict data privacy across participating institutions.

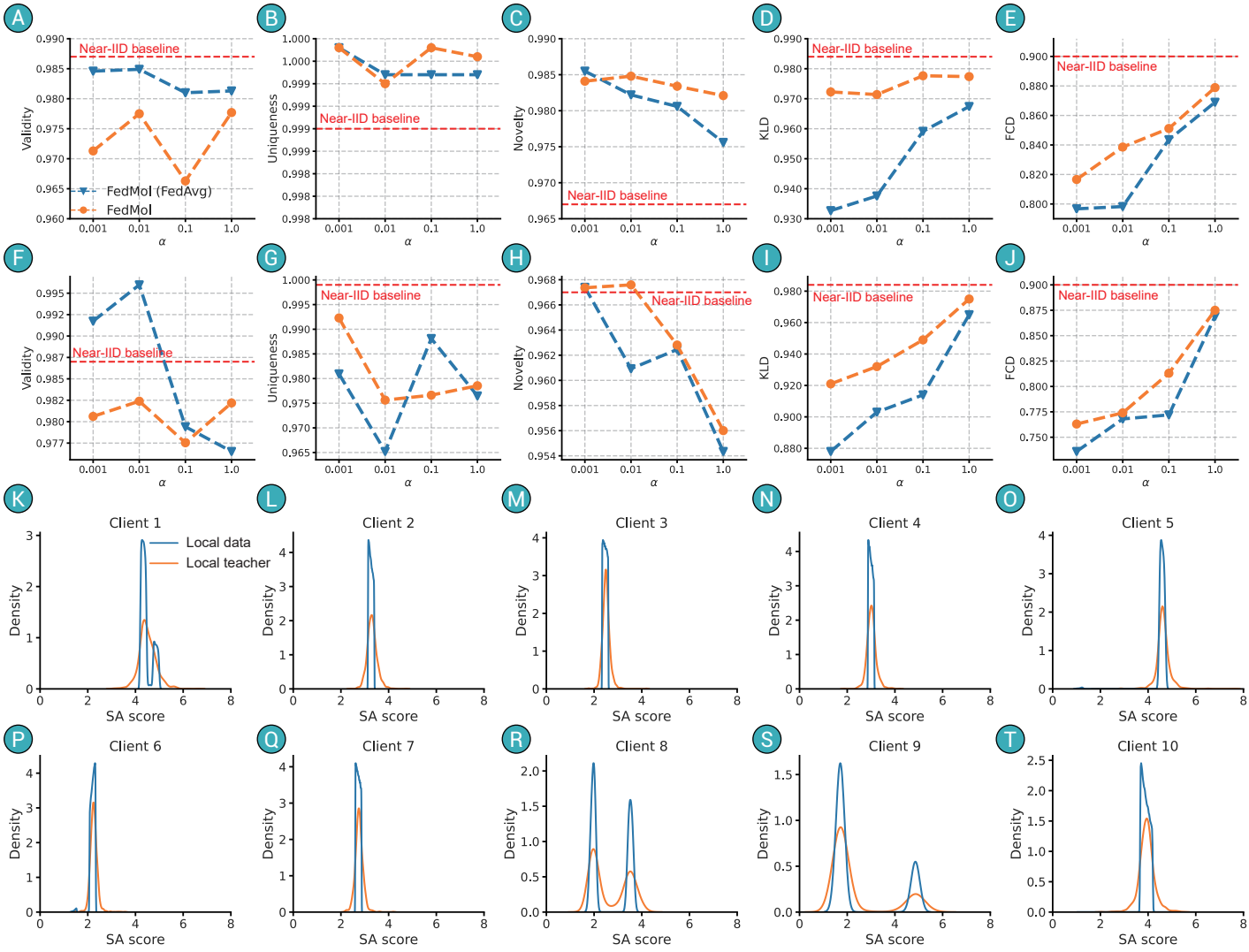


Figure 4. Performance of FedMol under varying degrees of data heterogeneity (A-J) Performance on GuacaMol distribution-learning benchmarks under one-dimensional (A–E) and two-dimensional (F–J) data heterogeneity, with the parameter α ranging from 0.001 to 1.0. Metrics include: validity (A,F), uniqueness (B,G), novelty (C,H), KLD (D,I) and FCD (E,J). Performance using the FedAvg algorithm is shown in blue, while the KD approach is shown in orange. Horizontal red lines indicate the baseline performance achieved under near-IID conditions. (K-T) Teacher specialization in extreme heterogeneity conditions ($\alpha = 0.001$) for the one-dimensional setting. For each of the ten clients, distributions of SA scores are shown for: (blue) the client's local private data, and (red) molecules generated by the client's locally trained teacher model.

Our work suggests that LoRA can play an important role in making federated molecular generation practically feasible. Traditional federated learning approaches face prohibitive communication costs when fine-tuning language models, as each round of communication requires transmitting millions or billions of parameters. LoRA fundamentally addresses this challenge by reducing the number of trainable parameters by 97.6% to 99.2% while preserving performance comparable to that achieved through full parameter fine-tuning (Table 1). This parameter efficiency translates directly into a commensurate reduction in communication overhead, making frequent model updates across distributed clients economically viable. Our results demonstrate that LoRA applied to all linear layers represents an optimal efficiency-performance trade-off for federated molecular generation. This configuration achieves a KLD of 0.979 and FCD of 0.869 under centralized conditions, closely approaching the performance ceiling of full fine-tuning (KLD: 0.984, FCD: 0.896). More importantly, this parameter efficiency directly translates to practical communication advantages in federated settings, where bandwidth constraints often limit the feasibility of collaborative model development.

Besides, our framework enables knowledge transfer without exposing proprietary molecular structures. The KD process aggregates probabilistic representations rather than raw data, providing inherent privacy protection while facilitating the integration of specialized chemical expertise from multiple sources. This approach aligns with emerging regulatory frameworks for data governance while enabling collaborative innovation. The combination of LoRA with KD creates a powerful synergy for federated molecular generation. LoRA's parameter efficiency enables frequent communication rounds with manageable bandwidth requirements, while KD's robustness to heterogeneity ensures that this communication leads to meaningful model improvement rather than optimization conflicts. The reduced parameter in LoRA fine-tuning makes the computation of softened probability distributions more efficient, lowering the computational burden on client devices. Next, the LoRA structure may provide implicit regularization that complements the explicit regularization provided by temperature scaling in KD. Furthermore, by reducing the dimensionality of the adaptation space, LoRA may help align client models, making the KD process more stable and effective. While the generation of distillation samples introduces additional computation, the process is highly parallelizable and can be accelerated through optimized inference strategies. In our implementation, this step increases the total training time by approximately 30% compared to a standard FedAvg baseline, a trade-off we consider acceptable given the substantial gains in privacy preservation, communication efficiency, and robustness to data heterogeneity.

While FedMol demonstrates strong performance across established evaluation metrics, several limitations merit consideration and suggest directions for future research: First, our current implementation employs SMILES-based molecular representation, which captures two-dimensional structural information but lacks explicit three-dimensional conformation data. Future work could extend the framework to incorporate three-dimensional representations through several promising ways: (1) integrating graph neural networks (GNNs) as the molecular generation backbone to explicitly model atomic connectivity and incorporate spatial features such as bond angles and interatomic distances;[40] (2) adopting 3D-equivariant architectures such as SchNet or equivariant GNN to directly operate on 3D coordinates while maintaining rotational and translational equivariance;[41,42] and (3) exploring hybrid pretraining strategies that combine large-scale SMILES corpora with smaller 3D datasets to achieve conformational awareness.[43] Importantly, these extensions remain compatible with our core privacy-preserving components, as teachers would share probabilistic information about generated molecules rather than explicit coordinate data. Second, while our Dirichlet-based heterogeneity simulation provides systematic control over distribution skew, real-world pharmaceutical heterogeneity may involve more complex, multi-dimensional distributions across therapeutic targets, synthesis methods, and biological activities. More sophisticated heterogeneity models could provide additional insights into framework robustness. Third, although our KD framework provides inherent privacy benefits by transmitting only softened probability distributions rather than raw data or model parameters, it does not offer formal privacy guarantees. Recent work has shown that distillation-based approaches can still leak sensitive information under certain attack models.[44,45] Future iterations of FedMol could incorporate differential privacy (DP) mechanisms during local training or employ secure aggregation protocols.[46,47] For instance, DP can be applied during the generation of distillation samples (e.g., by perturbing the soft labels or the sampled molecules themselves). Adding DP would introduce a trade-off between privacy and utility: stronger privacy guarantees require larger noise, which may degrade the quality of generated molecules, potentially reducing validity or distorting the learned chemical distribution. Finding the optimal balance between privacy loss and molecular fidelity is an important direction for future work, particularly in high-stakes pharmaceutical collaborations where both confidentiality and molecular quality are important.

REFERENCES

1. Kang, N.Y., Ha, H.H., Yun, S.W. et al. (2011). Diversity-driven chemical probe development for biomolecules: beyond hypothesis-driven approach. *Chem. Soc. Rev.* **40**:3613–3626. DOI:10.1039/C0CS00172D
2. Honarparvar, B., Govender, T., Maguire, G.E. et al. (2014). Integrated approach to structure-based enzymatic drug design: molecular modeling, spectroscopy, and experimental bioactivity. *Chem. Rev.* **114**:493–537. DOI:10.1021/cr300314q
3. Bagal, V., Aggarwal, R., Vinod, P. et al. (2021). MolGPT: molecular generation using a transformer-decoder model. *J. Chem. Inf. Model.* **62**:2064–2076. DOI:10.1021/acs.jcim.1c00600
4. Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D. et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**:268–276. DOI:10.1021/acscentsci.7b00572
5. Li, C. (2025). Breaking data silos in drug discovery with federated learning. *Nat. Chem. Eng.* **2**:288–289. DOI:10.1038/s44286-025-00213-x
6. Durant, G., Boyles, F., Birchall, K. et al. (2024). The future of machine learning for small-molecule drug discovery will be driven by data. *Nat. Comput. Sci.* **4**:735–743. DOI:10.1038/s43588-024-00699-0
7. Hanser, T., Ahlberg, E., Amberg, A. et al. (2025). Data-driven federated learning in drug discovery with knowledge distillation. *Nat. Mach. Intell.* 1–14. DOI:10.1038/s42256-025-00991-2
8. Chen, S., Xue, D., Chuai, G. et al. (2020). FL-QSAR: a federated learning-based QSAR prototype for collaborative drug discovery. *Bioinformatics* **36**:5492–5498. DOI:10.1093/bioinformatics/btaa1006
9. Li, L., Fan, Y., Tse, M. et al. (2020). A review of applications in federated learning. *Comput. Ind. Eng.* **149**:106854. DOI:10.1016/j.cie.2020.106854
10. Huang, T., Xu, H., Wang, H. et al. (2023). Artificial intelligence for medicine: Progress, challenges, and perspectives. *Innov. Med.* **1**:100030. DOI:10.59717/j.xinn-med.2023.100030
11. Zhu, W., Luo, J., and White, A.D. (2022). Federated learning of molecular properties with graph neural networks in a heterogeneous setting. *Patterns* **3**. DOI:10.1016/j.patter.2022.100521. DOI:10.1016/j.patter.2022.100521
12. Cozac, R., Hasic, H., Choong, J.J. et al. (2025). kMoL: an open-source machine and federated learning library for drug discovery. *J. Cheminform.* **17**:22. DOI:10.1186/s13321-025-00967-9
13. Bassani, D., Brigo, A., and Andrews-Morger, A. (2023). Federated learning in computational toxicology: an industrial perspective on the effiris hackathon. *Chem. Res. Toxicol.* **36**:1503–1517. DOI:10.1021/acs.chemrestox.3c00137
14. Luo, B., Li, X., Wang, S. et al. (2021). Cost-effective federated learning design. In IEEE INFOCOM 2021. pp. 1–10. DOI: 10.1109/INFOCOM42981.2021.9488679.
15. Zhu, Z., Hong, J., and Zhou, J. (2021). Data-free knowledge distillation for heterogeneous federated learning. In ICML. pp. 12878–12889. DOI: 10.48550/arXiv.2105.10056.
16. Lu, Z., Pan, H., Dai, Y. et al. (2024). Federated learning with non-IID data: A survey. *IEEE Internet Things J.* **11**:19188–19209. DOI:10.1109/JIOT.2024.3376548
17. Hu, E.J., Shen, Y., Wallis, P. et al. (2022). LoRA: Low-rank adaptation of large language models. In ICLR. pp. 3. DOI: 10.48550/arXiv.2106.09685.
18. Gou, J., Yu, B., Maybank, S.J. et al. (2021). Knowledge distillation: A survey. *Int. J. Comput. Vis.* **129**:1789–1819. DOI:10.1007/s11263-021-01453-z
19. Brown, N., Fiscato, M., Segler, M.H. et al. (2019). GuacaMol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**:1096–1108. DOI:10.1021/acs.jcim.8b00839
20. Yang, A., Yang, B., Hui, B. et al. (2024). Qwen2 technical report. *ArXiv.* DOI: 10.48550/arXiv.2407.10671.
21. Ainslie, J., Lee-Thorp, J., De Jong, M. et al. (2023). GQA: Training generalized multi-query transformer models from multi-head checkpoints. In Proc. EMNLP 2023. pp. 4895–4901. DOI: 10.18653/v1/2023.emnlp-main.298.
22. Su, J., Ahmed, M., Lu, Y. et al. (2024). Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **568**:127063. DOI:10.1016/j.neucom.2023.127063
23. Li, T., Sahu, A.K., Zaheer, M. et al. (2020). Federated optimization in heterogeneous networks. In Proc. Mach. Learn. Syst. pp. 429–450. DOI: 10.48550/arXiv.1812.06127.
24. Kim, S., Chen, J., Cheng, T. et al. (2023). PubChem 2023 update. *Nucleic Acids Res.*

51:D1373–D1380. DOI:10.1093/nar/gkac956
25. Gaulton, A., Bellis, L.J., Bento, A.P. et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40:D1100–D1107. DOI:10.1093/nar/gkr777
26. Shokri, R., Stronati, M., Song, C. et al. (2017). Membership inference attacks against machine learning models. In Proc. IEEE Symp. Secur. Priv. pp. 3–18. DOI: 10.1109/SP.2017.41.
27. He, Y., Li, B., Liu, L. et al. (2025). Towards Label-Only membership inference attack against pre-trained large language models. In Proc. 34th USENIX Secur. Symp. pp. 1609–1628. DOI: 10.48550/arXiv.2502.18943.
28. Preuer, K., Renz, P., Unterthiner, T. et al. (2018). Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *J. Chem. Inf. Model.* 58:1736–1741. DOI:10.1021/acs.jcim.8b00234
29. Cavanagh, J.M., Sun, K., Gritsevskiy, A. et al. (2024). Smileyllama: Modifying large language models for directed chemical space exploration. *arXiv.* DOI: 10.48550/arXiv.2409.02231.
30. Izdebski, A., Olszewski, J., Gawade, P. et al. (2025). Synergistic benefits of joint molecule generation and property prediction. *arXiv.* DOI: 10.48550/arXiv.2504.16559.
31. Chen, X., Wang, Y., He, J. et al. (2025). Graph generative pre-trained transformer. *arXiv.* DOI: 10.48550/arXiv.2501.01073.
32. Kwon, Y., Yoo, J., Choi, Y. et al. (2019). Efficient learning of non-autoregressive graph variational autoencoders for molecular graph generation. *J. Cheminform.* 11:70. DOI:10.1186/s13321-019-0396-x
33. Iwata, H., Nakai, T., Koyama, T. et al. (2023). VGAE-MCTS: A new molecular generative model combining the variational graph auto-encoder and monte carlo tree search. *J. Chem. Inf. Model.* 63:7392–7400. DOI:10.1021/acs.jcim.3c01220
34. Wu, J.N., Wang, T., Chen, Y. et al. (2024). t-SMILES: a fragment-based molecular representation framework for de novo ligand design. *Nat. Commun.* 15:4993. DOI:10.1038/s41467-024-49388-6
35. Mahmood, O., Mansimov, E., Bonneau, R. et al. (2021). Masked graph modeling for molecule generation. *Nat. Commun.* 12:3156. DOI:10.1038/s41467-021-23415-2
36. Geng, Z., Xie, S., Xia, Y. et al. (2023). De novo molecular generation via connection-aware motif mining. *arXiv.* DOI: https://doi.org/10.48550/arXiv.2302.01129.
37. Huang, Y., Zeng, Y., Dwivedi, V.P. et al. (2026). Higher-order grammar representations for molecular generation and learning. In ICLR.
38. Ueda, R., Nakai, T., Yoshida, K. et al. (2025). Mitigation of membership inference attack by knowledge distillation on federated learning. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* 108:267–279. DOI:10.1587/transfun.2024CIP0004
39. Karimireddy, S.P., Kale, S., Mohri, M. et al. (2020). Scaffold: Stochastic controlled averaging for federated learning. In ICML. pp. 5132–5143. DOI: 10.48550/arXiv.1910.06378.
40. You, J., Ying, R., Ren, X. et al. (2018). Graphrnn: Generating realistic graphs with deep auto-regressive models. In ICML. pp. 5708–5717. DOI: 10.48550/arXiv.1802.08773.
41. Schütt, K., Kindermans, P.J., Sauceda Felix, H.E. et al. (2017). Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. vol. 30. DOI: 10.48550/arXiv.1706.08566.
42. Satorras, V.G., Hoogeboom, E., and Welling, M. (2021). E(n) equivariant graph neural networks. In ICML. pp. 9323–9332. DOI: 10.48550/arXiv.2102.09844.
43. Stärk, H., Beaini, D., Corso, G. et al. (2022). 3d infomax improves gnns for molecular property prediction. In ICML. pp. 20479–20502. DOI: 10.48550/arXiv.2110.04126.
44. Carlini, N., Chien, S., Nasr, M. et al. (2022). Membership inference attacks from first principles. In Proc. IEEE Symp. Secur. Priv. pp. 1897–1914. DOI: 10.1109/SP46214.2022.9833649.
45. Shi, S., Zhang, J., Jiang, S. et al. (2025). DP-GENG: Differentially private dataset distillation guided by dp-generated data. DOI: 10.1609/aaai.v40i11.37854.
46. Abadi, M., Chu, A., Goodfellow, I. et al. (2016). Deep learning with differential privacy. In ACM CCS. pp. 308–318. DOI: 10.1145/2976749.2978318.
47. Bonawitz, K., Ivanov, V., Kreuter, B. et al. (2017). Practical secure aggregation for privacy-preserving machine learning. In ACM CCS. pp. 1175–1191. DOI: 10.1145/3133956.3133982.

FUNDING AND ACKNOWLEDGMENTS

This work was supported by Prevention and Control of Emerging and Major Infectious Diseases-National Science and Technology Major Project (2025ZD01901502), National Natural Science Foundation of China (62541160275, 62576366), Natural Science Foundation of Yunnan (202504BW050004) and the Guangdong Basic and Applied Basic Research Foundation (2026A1515011721). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, methodology, investigation, writing, T.D.; Review & editing C.Y.C. and L.Y. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

Not Applicable

DATA AND CODE AVAILABILITY

All datasets used in this study are publicly available. Molecules are obtained from the PubChem (https://pubchem.ncbi.nlm.nih.gov) and GuacaMol (https://github.com/BenevolentAI/guacamol), respectively.