FedDNA: DNA Sequence Reconstruction via Deep Evidential Learning and Personalized Federated Aggregation

Haiyan Lin[1], Qi Shen[2], Fei Zhu[1, 3*], Zixuan Qin[4], Liu Yang[2], Yuping Duan[5]
[1]Center for Applied Mathematics, KL-AAGDM, Tianjin University
[2]School of Artificial Intelligence, Tianjin University
[3]State Key Laboratory of Synthetic Biology, Tianjin University
[4]School of Computer Science and Technology, Tianjin University
[5]School of Mathematical Sciences, Beijing Normal University
{linhaiyan, shenqi_405, fei.zhu, qinzixuan1958, yangliuyl}@tju.edu.cn, doveduan@gmail.com

Abstract

DNA-based data storage offers an attractive alternative to traditional media due to its exceptional density, durability, and sustainability. However, errors introduced across the DNA storage pipeline critically impede accurate sequence reconstruction from noisy sequencing reads. This paper addresses the DNA sequence reconstruction problem by proposing FedDNA, a novel Personalized Federated Learning (PFL) framework based on Evidential Deep Learning (DEL), designed for DNA storage environments. FedDNA quantifies robust predictive uncertainty through a novel evidence fusion mechanism that aggregates evidence from each noisy read in a cluster, thereby enhancing client-level prediction reliability. For efficient sequence modeling and reconstruction from these noisy clusters, its architecture employs a convolution-enhanced Mamba encoder and an LSTM decoder. To address prohibitive centralized training costs, privacy concerns, and data heterogeneity across diverse DNA storage data, FedDNA integrates PFL and designs an innovative uncertainty-driven personalized aggregation strategy based on epistemic and aleatoric decomposition, for which we also provide rigorous theoretical generalization bounds. Experimental results demonstrate FedDNA achieves superior reconstruction performance on DNA storage data with heterogeneity, highlighting its potential for secure and efficient DNA storage systems.

Introduction

DNA, as an information carrier, offers high storage density, exceptional durability, and low maintenance costs, making it a promising alternative to conventional storage media (Ceze, Nivala, and Strauss 2019; Meiser et al. 2022; Dou et al. 2024). The typical DNA data storage workflow involves encoding binary data into DNA sequences (known as *references*), synthesizing molecules, storing, retrieving via PCR amplification and sequencing, and finally decoding. Each reference produces numerous *reads*, which inevitably accumulate errors—primarily insertions, deletions, and substitutions (IDS errors)—during the pipeline of DNA storage. Recovering original information from these noisy reads is critical, often involving an initial clustering step to group related reads (Qu, Yan, and Wu 2022; Rashtchian et al. 2017). This study focuses on the subsequent *sequence reconstruction, aiming to accurately infer the original reference from its cluster of noisy reads*. Addressing these IDS errors during sequence reconstruction is paramount for reliable DNA data retrieval. While deep learning shows significant promise for DNA sequence reconstruction (Bar-Lev et al. 2025; Qin et al. 2024, 2025), its practical application faces critical bottlenecks:

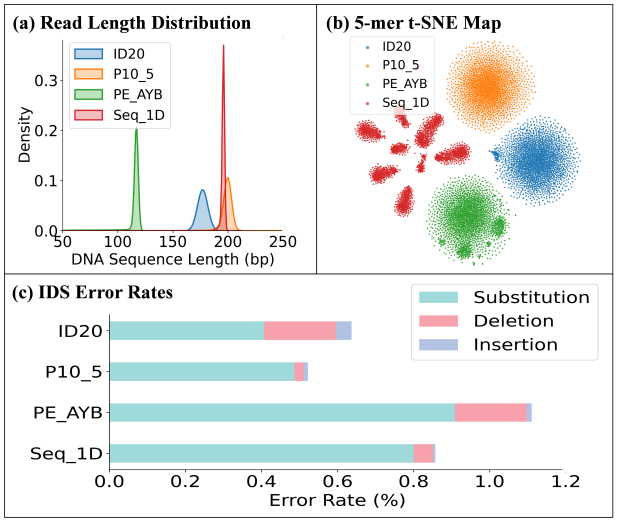


Figure 1: Illustration of **acute data heterogeneity** across diverse DNA storage clients. (a) Depicts varying distributions of read lengths using KDEs. (b) Reveals distinct clusters of 5-mer sequence patterns, indicating profound intrinsic content differences. (c) Presents probabilities of IDS errors, highlighting diverse data quality. Collectively, these panels provide compelling evidence of heterogeneity (in length, content, and quality) that significantly challenges FL for DNA sequence reconstruction

Problem 1: Data Scarcity and Privacy in DNA Sequence Reconstruction. Acquiring large-scale, high-quality real DNA data is prohibitively expensive and time-consuming (e.g., $0.8-5M per GB), limited by high synthesis and sequencing costs and experimental resource constraints (Yu

*Corresponding Author.

et al. 2024). Furthermore, stringent data privacy policies severely restrict data sharing and centralized model training among different DNA storage data centers. The inherent biological nature and extreme longevity of DNA as an information carrier significantly amplify these privacy concerns. These substantial impediments necessitate a novel, distributed paradigm. Federated Learning (FL), which aims to collaboratively train a shared model across multiple clients without directly exchanging local data while preserving data privacy, offers a compelling strategy for DNA sequence reconstruction. However, despite the inherent benefits for FL, its application to the DNA storage scenario introduces another critical challenge:

Problem 2: Data Heterogeneity in DNA Sequence Reconstruction via FL. Data heterogeneity poses a critical and multifaceted challenge in DNA sequence reconstruction via FL. Variances from diverse encoding, primer designs, synthesis and sequencing instruments across DNA storage datasets lead to significant distributional shifts. As illustrated in Figure 1, this severe heterogeneity fundamentally impairs global model performance and generalization capabilities in the FL process.

To overcome these challenges, we propose FedDNA, a novel framework for DNA sequence reconstruction that uniquely combines Personalized Federated Learning (PFL) with an adaptive aggregation strategy. FedDNA is fundamentally guided by Deep Evidential Learning (DEL) (Sensoy, Kaplan, and Kandemir 2018), which supports key innovations at both the client and federated levels. *At the client level*, FedDNA integrates a convolution-enhanced Mamba encoder with a two-layer LSTM decoder. This efficient design supports effective sequence reconstruction, helping alleviate data scarcity. Through this architecture, we leverage DEL to model inherent IDS errors and their associated uncertainty, enabling robust prediction through evidence fusion. DEL quantifies prediction uncertainty based on Dempster-Shafer Theory (DST) (Sentz and Ferson 2002) and features a novel post-hoc evidence fusion for cluster of reads, enhancing local reconstruction. *At the federated level*, our PFL approach aggregates only encoder parameters, allowing clients to retain private decoders. Leveraging DEL's efficacy in mitigating data heterogeneity (Chen et al. 2024; Qin et al. 2023), we propose a novel adaptive aggregation strategy that dynamically adjusts federated weights via decomposed uncertainty (aleatoric and epistemic) (Malinin and Gales 2018) from local estimates, providing more informed and robust global model learning, supported by theoretical generalization bounds. The main contributions of this paper are summarized as follows:

- We propose the first end-to-end deep neural network framework for FL in DNA sequence reconstruction, pioneering this application domain. Our framework innovatively integrates a convolution-enhanced Mamba encoder with a two-layer LSTM decoder, an architecture chosen for its efficiency and ability to effectively process intricate DNA storage data.
- We introduce a novel application of DEL to DNA sequence reconstruction for predictive uncertainty quantification at the sequence level. Crucially, we design a post-hoc evidence fusion mechanism that effectively combine evidence from multiple reads within each cluster, enhancing prediction reliability.
- We propose an uncertainty-aware PFL framework, which leverages DEL's decomposed uncertainty (epistemic and aleatoric components) to dynamically adjust client aggregation weights. This enables more robust global model learning despite data heterogeneity, for which we further provide theoretical generalization error bounds.

Related Work

DNA Sequence Reconstruction

The problem of DNA-to-DNA reconstruction involves inferring the original DNA reference sequence from a cluster of its noisy reads. Traditional methods, including BMA Lookahead (Gopalan et al. 2018), BMA Divider (Sabary et al. 2024), and Iterative Reconstruction (Sabary et al. 2024), primarily employ symbol-wise majority voting, often with sequence alignment. Though effective for low IDS error rates, their performance critically depends on sufficient reads. More recently, deep learning has enabled neural network-based methods to directly map noisy reads to a reference. For instance, RobuSeqNet (Qin et al. 2024) used attention and Conformer blocks, demonstrating resilience to noisy reads, while DNAformer (Bar-Lev et al. 2025) proposed a twin-network combining convolution and Transformer modules with a learnable alignment component. However, these models often entail large parameter sizes, requiring extensive training data, typically mitigated by massive samples from DNA storage simulation tool (Bar-Lev et al. 2025).

Heterogeneous Federated Settings

Data heterogeneity is a key issue limiting FL performance. Various mitigation strategies have been proposed, primarily categorized into three directions: (1) federated averaging with optimized aggregation weights, exemplified by approaches like FedAWA (Shi et al. 2025), FedAW (Tang 2024), and ConFREE (Zheng et al. 2025); (2) regularization methods to constrain local update drift, such as FedProx (Li et al. 2020) and MOON (Li, He, and Song 2021); and (3) PFL with customized client models, such as FedPer (Arivazhagan et al. 2019) and pFedFDA (Mclaughlin and Su 2024), which tailor models to each client's specific data distribution. However, the aforementioned methods still face challenges in precisely evaluating the contribution of each client or the degree of local model drift. To this end, recent research has begun to explore the use of model uncertainty as an effective metric to provide more reliable guidance for federated aggregation.

DEL provides a feasible path for uncertainty-based aggregation by modeling the output as a Dirichlet distribution and jointly estimating the prediction and its uncertainty. RIPFL leverages DEL-based uncertainty to guide reliability-aware client selection, improving personalized model performance (Qin et al. 2023). Similarly, FedEvi uses decomposed uncertainties to estimate generalization gaps and client reliability, dynamically adjusting aggregation weights for im-

proved generalization (Chen et al. 2024). However, these approaches do not fully exploit the rich uncertainty decomposition provided by DEL, failing to comprehensively utilize the distinct contributions of epistemic and aleatoric uncertainties from both local and global model perspectives. Such partial uncertainty utilization becomes critically pronounced in tasks like DNA sequence reconstruction, where acute data heterogeneity across clients can severely impair personalized performance if full model sharing is enforced. For this reason, this paper introduces an uncertainty-decomposed aggregation to guide the encoder toward learning reliable shared representations, while keeping the decoder private for personalized adaptation to this task.

Methodology

To address the critical task of DNA sequence reconstruction, this paper proposes the FedDNA framework, as illustrated in Figure 2. Given the DNA alphabet $\Sigma = \{\text{A, C, G, T}\}$, an original reference sequence $y \in \Sigma^L$ generates n noisy copies. These noisy reads form a cluster $\mathcal{X} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{n}$. Here, $\boldsymbol{x}^{(i)}$ denotes the numerical representation of each read, which, to handle variable lengths from IDS errors, is one-hot encoded and uniformly zero-padded to a fixed length L_{pad}, resulting in a $4 \times L_{pad}$ matrix. The objective is to learn a hypothesis $h \in \mathcal{H}$, where $\mathcal{H}$ denotes the hypothesis class, that reconstructs the original reference from a noisy read cluster: $h : \mathcal{X} \to \Sigma^L$. Formally, the optimization goal is to find the optimal hypothesis h^* that minimizes the expected reconstruction error over the data distribution $\mathcal{D}$:

$$h^* = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{(y,\mathcal{X})\sim\mathcal{D}} \left[\ell(h(\mathcal{X}), y)\right], \tag{1}$$

where $\ell(\cdot,\cdot)$ is a function that quantifies the discrepancy between the reconstructed sequence and the reference.

Built on a PFL architecture, FedDNA involves K clients. A global model, comprising only a shared encoder $\boldsymbol{\theta}_G^E$, resides on the central server and is collaboratively updated through aggregation. For each client k, their personalized model $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_k^E, \boldsymbol{\theta}_k^D)$, comprises a private decoder $\boldsymbol{\theta}_k^D$ and a local copy of the shared encoder, $\boldsymbol{\theta}_k^E$. Each communication round, clients initialize their local encoder $\boldsymbol{\theta}_k^E$ with the current global encoder $\boldsymbol{\theta}_G^E$. Following local training, these updated local encoders $\boldsymbol{\theta}_k^E$ are then uploaded to the server for weighted aggregation, yielding the updated global encoder $\boldsymbol{\theta}_G^E$. This section details our methodology: an evidential deep learning framework for DNA sequence reconstruction, followed by its uncertainty-decomposed federated aggregation.

Evidential Deep Learning-based DNA Sequence Reconstruction Framework

We elaborate on the key components of FedDNA: a core Dirichlet-based evidential model, a novel encoder-decoder architecture, and a specialized evidential fusion module. Their detailed operation and training objectives are elaborated in the following.

Dirichlet-based Evidential Model Central to FedDNA's robust uncertainty quantification, each client's personalized model $\boldsymbol{\theta}_k$ employs a DEL(Sensoy, Kaplan, and Kandemir 2018) approach, grounded in DST (Sentz and Ferson 2002).

For a C-class classification problem, conventional deep neural networks typically use a softmax function, yielding single-point class probability estimates $\boldsymbol{\rho}$. This often leads to overconfident or miscalibrated predictions, especially with heterogeneous local data. Such point estimates struggle to express epistemic uncertainty, potentially yielding high confidence even for out-of-distribution or novel inputs. In contrast, DEL treats $\boldsymbol{\rho}$ as a random variable drawn from a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\rho}|\boldsymbol{\alpha})$ parameterized by $\boldsymbol{\alpha}$:

$$p(\boldsymbol{\rho}|\mathcal{X}, \boldsymbol{\theta}) = \mathrm{Dir}(\boldsymbol{\rho}|\boldsymbol{\alpha}) = \begin{cases} \dfrac{1}{B(\boldsymbol{\alpha})} \displaystyle\prod_{c=1}^{C} \rho_c^{\alpha_c - 1}, & \boldsymbol{\rho} \in \Delta^C \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $B(\boldsymbol{\alpha})$ is the C-dimensional multinomial Beta function, and $\Delta^C = \{\boldsymbol{\rho} | \sum_{c=1}^{C} \rho_c = 1 \text{ and } 0 < \rho_c < 1\}$ represents the C-dimensional unit simplex. In DEL, the network outputs non-negative evidence vectors $\boldsymbol{e}$, defining Dirichlet parameters $\boldsymbol{\alpha} = \boldsymbol{e} + 1$. Dirichlet strength $S = \sum \alpha_c$ quantifies total evidence and confidence. The belief mass vector $\boldsymbol{b} = \boldsymbol{e}/S$, and uncertainty $u = C/S$, satisfying $\sum b_c + u = 1$.

In this task, base prediction at each sequence position is a 4-class classification problem corresponding to the four nucleotides. Client k's model $\boldsymbol{\theta}_k$ reconstructs sequences from read cluster $\mathcal{X}$ by outputting $\boldsymbol{e}_l^{(i)} = f(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}_k)_l$ (via Softplus) for each base l in read $\boldsymbol{x}^{(i)} \in \mathcal{X}$. Its Dirichlet parameter is then $\boldsymbol{\alpha}_l^{(i)} = \boldsymbol{e}_l^{(i)} + 1$, thus the class probability vector $\boldsymbol{\rho}_l^{(i)}$ is drawn from $\mathrm{Dir}(\boldsymbol{\rho}_l^{(i)}|\boldsymbol{\alpha}_l^{(i)})$. These per-read, per-base evidence vectors quantify predictive uncertainty (given by $u_l = 4/S_l$ for each base l), foundational for subsequent sequence-level evidential fusion.

Encoder-Decoder Architecture The encoder adopts a hybrid architecture combining convolutional layers with Mamba, a recently proposed state space model capable of capturing global contextual dependencies while preserving linear computational complexity (Gu and Dao 2023). This design effectively leverages convolutional operations for local feature extraction and Mamba for long-range dependency modeling. The integrated approach robustly captures complex features in noisy DNA clusters, maintaining resilience against sequence errors, including base substitutions and positional shifts from insertions or deletions. Given the encoder's strong feature representation capability, we deliberately employ a lightweight two-layer LSTM (Hochreiter and Schmidhuber 1997) as the sequence generation module, avoiding overly deep or complex structures.

Evidential Fusion We propose a post-hoc evidential fusion strategy to enhance prediction performance and uncertainty representation. We obtain individual evidence vectors $\boldsymbol{e}_l^{(i)}$ by independently processing each read $\boldsymbol{x}^{(i)}$ through the encoder and decoder. Unlike early fusion, to accurately capture and aggregate each sequence's information, we average these vectors at each base position l, yielding the final averaged evidence $\overline{\boldsymbol{e}}_l = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{e}_l^{(i)}$. This fused evidence $\overline{\boldsymbol{e}}_l$ then directly informs the probabilistic base prediction for position l. From $\overline{\boldsymbol{e}}_l$, we obtain the fused Dirichlet parameter $\overline{\boldsymbol{\alpha}}_l = \overline{\boldsymbol{e}}_l + 1$, the fused Dirichlet strength $\overline{S}_l = \sum_{c=1}^{4} \overline{\alpha}_{lc}$, and

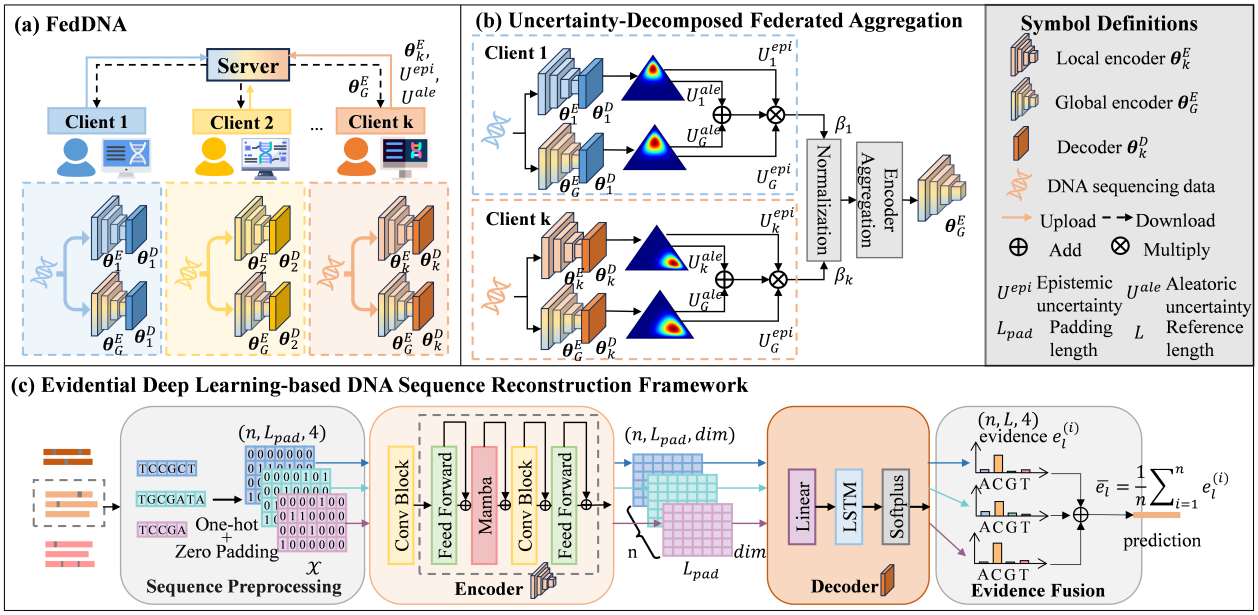


Figure 2: Overview of the proposed FedDNA framework. (a) FedDNA architecture; (b) Uncertainty-decomposed federated aggregation weights via evidential deep learning; (c) DNA sequence reconstruction framework.

the fused uncertainty $\overline{u}_l = 4/\overline{S}_l$. By integrating multi-source evidence, this strategy improves both model accuracy and uncertainty representation.

DEL-based Training Objectives To train our Dirichlet-based evidential model, which performs sequence-level predictions and aggregates losses from each DNA base position l, the overall loss function comprises two main components: a reconstruction loss and a Kullback-Leibler (KL) divergence regularization term. It is defined as $\ell(\boldsymbol{\theta}) = \ell_{\text{rec}}(\boldsymbol{\theta}) + \gamma_t \ell_{\text{KL}}(\boldsymbol{\theta})$, where γ_t is the annealing coefficient and t is the index of the current training epoch.

For the reconstruction loss, we adopt the Bayesian risk of the cross-entropy loss (Sensoy, Kaplan, and Kandemir 2018). Let $\text{CE}(\boldsymbol{y}_l, \boldsymbol{\rho}_l) = \sum_{c=1}^{4} -y_{lc} \log(\rho_{lc})$ denote the cross-entropy between the ground-truth vector $\boldsymbol{y}_l$ and the predicted probability vector $\boldsymbol{\rho}_l$ at position l. Here, $\boldsymbol{\rho}_l$ is drawn from a Dirichlet distribution parameterized by the fused Dirichlet parameters $\overline{\boldsymbol{\alpha}}_l$. This loss is:

$$\begin{aligned} \ell_{\text{rec}}(\boldsymbol{\theta}) &= \frac{1}{L}\sum_{l=1}^{L}\left[\int \text{CE}(\boldsymbol{y}_l, \boldsymbol{\rho}_l)\frac{1}{B(\overline{\boldsymbol{\alpha}}_l)}\prod_{c=1}^{4}\rho_{lc}^{\overline{\alpha}_{lc}-1}\, d\boldsymbol{\rho}_l\right] \\ &= \frac{1}{L}\sum_{l=1}^{L}\sum_{c=1}^{4} y_{lc}\left(\psi(\overline{S}_l) - \psi(\overline{\alpha}_{lc})\right), \end{aligned} \tag{3}$$

where $\boldsymbol{y}_l$ is the one-hot encoded ground-truth vector for position l, with $y_{lc} = 1$ if class c is the true class and $y_{lc} = 0$ otherwise. $\psi(\cdot)$ denotes the digamma function.

The model's Dirichlet output distribution is regularized using a Kullback-Leibler (KL) divergence term (Kullback and Leibler 1951; Sensoy, Kaplan, and Kandemir 2018). This regularization mitigates overconfident predictions on misclassified samples by promoting high uncertainty when the model lacks sufficient evidence:

$$\ell_{\text{KL}}(\boldsymbol{\theta}) = \frac{1}{L}\sum_{l=1}^{L} KL\left[\text{Dir}(\boldsymbol{\rho}_l|\tilde{\boldsymbol{\alpha}}_l) \,\|\, \text{Dir}(\boldsymbol{\rho}_l|\mathbf{1})\right], \tag{4}$$

where $\mathbf{1}$ is the vector of ones, $\tilde{\boldsymbol{\alpha}}_l$ is the adjusted Dirichlet parameter vector for position l, constructed as $\tilde{\boldsymbol{\alpha}}_l = \boldsymbol{y}_l \odot \mathbf{1} + (\mathbf{1} - \boldsymbol{y}_l) \odot \overline{\boldsymbol{\alpha}}_l$, and $\Gamma(\cdot)$ is the Gamma function.

Uncertainty-Decomposed Federated Aggregation

Our PFL model's design addresses acute heterogeneity in client DNA datasets (see Figure 1). Varying reference lengths necessitate client-specific decoders. To achieve this adaptability, while minimizing communication overhead, FedDNA globally shares encoder parameters during federated training, with each client retaining its personalized decoder—a crucial choice given the decoder's small parameter footprint compared to the dominant encoder.

We propose an adaptive weight allocation for federated aggregation that leverages epistemic and aleatoric uncertainties from our DEL framework to refine the global model. We denote the cluster sample as $\mathcal{X}$ and the model parameters as $\boldsymbol{\theta}$. Specifically, for each client, uncertainties are quantified by evaluating their local data $\mathcal{X}_k$ against two distinct model configurations: (i) their personalized model $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_k^E, \boldsymbol{\theta}_k^D)$; and (ii) a hybrid model formed by the global encoder and their local decoder $(\boldsymbol{\theta}_G^E, \boldsymbol{\theta}_k^D)$, with this quantification leveraging the Dirichlet-based evidential model's output.

For each position l, we compute the expected probability vector $P(\boldsymbol{y}_l|\mathcal{X}, \boldsymbol{\theta}) \triangleq \hat{\boldsymbol{\rho}}_l$, where

$$\begin{aligned} P(y_{lc}=1|\mathcal{X}, \boldsymbol{\theta}) &= \int p(y_{lc}=1|\boldsymbol{\rho}_l) p(\boldsymbol{\rho}_l|\mathcal{X}, \boldsymbol{\theta})\, d\boldsymbol{\rho}_l \\ &= \frac{\overline{\alpha}_{lc}}{\overline{S}_l} = \hat{\rho}_{lc}. \end{aligned} \tag{5}$$

Here, $\overline{\alpha}_{lc}$ is the fused Dirichlet parameter for class c at position l (an element of $\overline{\boldsymbol{\alpha}}_l$), and $\hat{\rho}_{lc}$ serves as the point estimate for the expected probability of class c. This predictive probability distribution, $P(\boldsymbol{y}_l|\mathcal{X}, \boldsymbol{\theta})$, quantifies total uncertainty (U^{total}) via its Shannon entropy (Shannon 1948). The total uncertainty is further decomposed into epistemic uncertainty (U^{epi}) and aleatoric uncertainty (U^{ale}) (Malinin and Gales 2018; Xie et al. 2023):

$$\underbrace{\mathcal{H}[P(\boldsymbol{y}_l|\mathcal{X}, \boldsymbol{\theta})]}_{U^{total}} = \underbrace{\mathcal{I}[\boldsymbol{y}_l, \boldsymbol{\rho}_l|\mathcal{X}, \boldsymbol{\theta}]}_{U^{epi}} + \underbrace{\mathbb{E}_{p(\boldsymbol{\rho}_l|\mathcal{X}, \boldsymbol{\theta})}[\mathcal{H}[P(\boldsymbol{y}_l|\boldsymbol{\rho}_l)]}_{U^{ale}}, \tag{6}$$

For a sequence of length L, total, epistemic, and aleatoric uncertainties are averaged across positions.

Epistemic Uncertainty Epistemic uncertainty (model uncertainty) stems from insufficient evidence, reflecting uncertainty in model parameter estimation. Crucially, this uncertainty is reducible, diminishing with increasing training samples as the model gains more evidence. Quantified by the mutual information $\mathcal{I}[\boldsymbol{y}_l, \boldsymbol{\rho}_l|\mathcal{X}, \boldsymbol{\theta}]$, it measures uncertainty reduction in true labels $\boldsymbol{y}_l$ if the true $\boldsymbol{\rho}_l$ were known, thereby directly reflecting the Dirichlet distribution's spread on the simplex: higher values indicate broader opinions due to insufficient evidence (Xie et al. 2023). For a cluster sample $\mathcal{X}$ and model parameters $\boldsymbol{\theta}$, epistemic uncertainty is:

$$U^{epi}(\mathcal{X}, \boldsymbol{\theta}) = \frac{1}{L}\sum_{l=1}^{L}\left(\sum_{c=1}^{4}\hat{\rho}_{lc}\left[\psi(\overline{\alpha}_{lc}+1) - \psi(\overline{S}_l+1)\right] - \sum_{c=1}^{4}\hat{\rho}_{lc}\log\hat{\rho}_{lc}\right). \tag{7}$$

Aleatoric Uncertainty Aleatoric uncertainty, or data uncertainty, originates from the inherent ambiguity and irreducible noise within the data itself. It is quantified as the expected entropy over all possible predictive outcomes, thereby reflecting the intrinsic complexity of the data and representing an error source that cannot be eliminated through learning. The aleatoric uncertainty for a cluster sample $\mathcal{X}$ under model parameters $\boldsymbol{\theta}$ is:

$$U^{ale}(\mathcal{X}, \boldsymbol{\theta}) = \frac{1}{L}\sum_{l=1}^{L}\sum_{c=1}^{4}\hat{\rho}_{lc}\left[\psi(\overline{S}_l+1) - \psi(\overline{\alpha}_{lc}+1)\right]. \tag{8}$$

Adaptive Aggregation Weights Building on the decoupled uncertainties quantified in Eq. (7) and (8), we propose an adaptive aggregation weighting strategy that assigns a weight β_k to each client k as follows:

$$\beta_k = \frac{U_G^{epi}}{U_k^{epi}\cdot(U_G^{ale}+U_k^{ale})}, \tag{9}$$

where all uncertainty terms are computed as averages across the batch of clusters in client k's local data. Specifically, U_k^{epi} and U_k^{ale} are derived using model configuration (i), while U_G^{epi} and U_G^{ale} are derived using the same local data with model configuration (ii), as previously defined. Subsequently, these weights β_k are normalized to obtain the final aggregation coefficients $\lambda_k = \frac{\beta_k}{\sum_{j=1}^{K}\beta_j}$. During the encoder aggregation phase, the global encoder parameters θ_G^E are updated via a weighted sum of the local encoder parameters θ_k^E from each client: $\theta_G^E = \sum_{k=1}^{K}\lambda_k\theta_k^E$.

Each of the three uncertainty components in Eq. (9) carries a modeling significance: A high U_G^{epi} indicates a global epistemic gap that client k's local model might address, thus warranting a higher aggregation weight. Conversely, a low U_k^{epi} signifies client k's local model has effectively adapted and offers highly confident predictions, which increases its weight. Finally, a small combined aleatoric uncertainty ($U_G^{ale}+U_k^{ale}$) reflects low inherent randomness and noise in client k's batch of local clusters, ensuring more robust updates and thereby a larger aggregation weight. In summary, this aggregation mechanism assigns higher weights to clients whose local model is confident (low local epistemic uncertainty), whose data is clean (low aleatoric uncertainty), and where the global model is uncertain (high global epistemic uncertainty). By dynamically adjusting aggregation weights based on these average uncertainty evaluations, this strategy enhances global aggregation quality and improves generalization robustness in heterogeneous environments.

Generalization Bound

We consider K clients, each with local distribution $\mathcal{D}_k$ and $\mathcal{S}_k$ is a sample of size m_k, with $m = \sum_{k=1}^{K} m_k$, $\boldsymbol{m} = (m_1, \ldots, m_K)$. Let $\mathcal{G}$ denote the family of the losses associated to a hypothesis set $\mathcal{H} : \mathcal{G} = \{(\mathcal{X}, y) \mapsto \ell(h(\mathcal{X}), y) : h \in \mathcal{H}\}$. Assume the loss function $\ell(h(\mathcal{X}), y) \in [0, M]$ is bounded. According to Eq. (9), the aggregation weights are normalized as λ_k, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$. Based on the *digamma difference inequality*, we can assume that the uncertainty upper bounds satisfy: $U_k^{\text{epi}} \le \frac{A_1}{m_k}$ and $U_k^{\text{ale}} \le \frac{A_2}{m_k}$.

Lemma 1 (Approximation of Aggregation Weights)**.** *Given the upper bounds $U_k^{epi} \le \frac{A_1}{m_k}$ and $U_k^{ale} \le \frac{A_2}{m_k}$, the normalized aggregation weight λ_k satisfies the approximation $\lambda_k = \frac{m_k}{m} + \epsilon_k$, where $|\epsilon_k| \le B \cdot \frac{m_k}{m^2}$, for some constant $B > 0$ that depends only on A_1, A_2, and the class count C.*

Lemma 2 (Bound on Aggregation Skewness)**.** *The skewness $s(\boldsymbol{\lambda}\|\boldsymbol{m}) := 1 + \sum_{k=1}^{K}\frac{(\lambda_k - m_k/m)^2}{m_k/m}$ satisfies $s(\boldsymbol{\lambda}\|\boldsymbol{m}) = 1 + \mathcal{O}\left(\frac{1}{m^2}\right)$.*

Theorem 1 (Generalization Bound with Uncertainty-based Aggregation)**.** *With probability at least $1-\delta$, the following bound holds for all $h \in \mathcal{H}$:*

$$\ell_{\mathcal{D},\lambda}(h) \le \hat{\ell}_{\mathcal{S},\lambda}(h) + 2\,\mathfrak{R}_m(\mathcal{G}, \lambda) + M\cdot\sqrt{\frac{1}{2m}\log\left(\frac{1}{\delta}\right)} + \mathcal{O}\left(\frac{M}{m^{2.5}}\right). \tag{10}$$

Where $\ell_{\mathcal{D},\boldsymbol{\lambda}}(h)$ is the true weighted risk, $\hat{\ell}_{\mathcal{S},\boldsymbol{\lambda}}(h)$ is the empirical weighted risk, and $\mathfrak{R}_m(\mathcal{G}, \boldsymbol{\lambda})$ is the empirical Rademacher complexity.

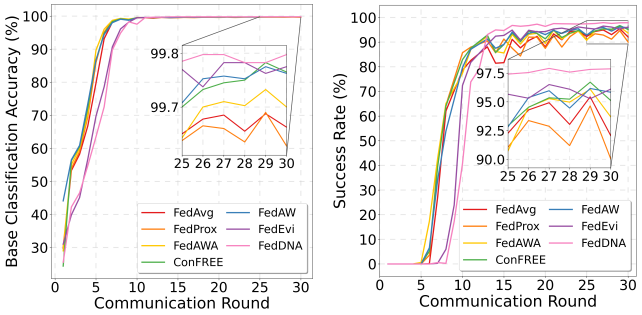
Building upon Lemma 1 and Lemma 2, alongside the established FedAvg generalization bound (Mohri, Sivek, and Suresh 2019), we derive Theorem 1.

Experiments

Datasets Our approach is evaluated for DNA sequence reconstruction on four real-world DNA storage datasets: Id20 (Organick et al. 2018), P10_5_BDDP210000009 (P10_5)

Methods	↑ Success Rate (%)					↓ Edit Error Rate (%)				
	ID20	P10_5	PE_AYB	Seq_1D	Avg.	ID20	P10_5	PE_AYB	Seq_1D	Avg.
FedAvg	97.07±0.69	88.43±1.50	95.92±0.54	87.78±0.55	92.30±0.55	0.10±0.01	0.22±0.02	0.20±0.03	0.36±0.00	0.22±0.01
FedProx	97.00±0.43	90.95±1.59	95.72±0.90	89.73±0.51	93.35±0.53	0.10±0.01	0.18±0.02	0.19±0.02	0.35±0.02	0.21±0.01
FedAWA	97.05±0.56	96.09±0.70	96.53±0.63	92.20±0.26	95.47±0.40	0.11±0.01	0.12±0.02	0.20±0.04	0.18±0.02	0.15±0.02
ConFREE	98.22±0.19	96.72±0.06	96.62±0.10	91.83±0.69	95.85±0.20	0.09±0.01	0.08±0.01	**0.12±0.02**	0.25±0.02	0.13±0.00
FedAW	98.08±0.22	96.24±0.98	96.51±0.36	94.07±0.90	96.23±0.61	0.09±0.01	0.10±0.02	0.14±0.01	0.26±0.12	0.15±0.03
FedEvi	99.26±0.26	97.82±0.06	96.09±0.43	95.09±0.45	97.07±0.22	**0.05±0.00**	0.09±0.02	0.30±0.14	0.16±0.03	0.15±0.04
–FL	98.97±0.10	97.06±0.43	94.53±0.50	95.01±0.17	96.39±0.51	0.08±0.02	0.10±0.00	0.22±0.01	0.19±0.04	0.15±0.01
FedDNA	**99.50±0.02**	**97.96±0.06**	**96.89±0.19**	**97.26±0.26**	**97.91±0.10**	**0.05±0.01**	**0.06±0.05**	0.23±0.04	**0.15±0.02**	**0.12±0.02**

Table 1: Performance comparison of FedDNA against six federated learning methods on four real-world DNA storage datasets, showing Success Rate and Edit Error Rate. '–FL' denotes the centralized variant of FedDNA.



(a) Base Classification Acc (b) Success Rate

Figure 3: Average client performance over 30 communication rounds. (a) Base Classification Accuracy. (b) Success Rate.

(Song et al. 2022), PE_AYB (Goldman et al. 2013), and Sequencing_data_first_dimension (Seq_1D) (Pan et al. 2022). As illustrated in Figure1, high data heterogeneity, especially prominent within Seq_1D, poses significant challenges for federated sequence reconstruction. Data preprocessing involved aligning reads with BWA (Li and Durbin 2009) for ground-truth labeling, then filtering those whose lengths deviated from their reference by over ±5 bases. Finally, for each DNA cluster, 5 to 30 reads were randomly sampled as reconstruction input.

Metrics To evaluate the performance of DNA sequence reconstruction, we adopt Success Rate (Sabary et al. 2024; Qin et al. 2024) as the primary metric:

$$\text{Success Rate} = \frac{\#\{\text{prediction} == \text{reference}\}}{\#\{\text{reference}\}}. \quad (11)$$

A successful prediction is defined by an exact, nucleotide-level match between the predicted and reference sequences. This strict, widely used criterion reflects information integrity pivotal for DNA storage.

Additionally, Edit Error Rate is reported as a supplementary metric, quantifying reconstruction accuracy via normalized edit distance (Levenshtein 1966):

$$\text{Edit Error Rate} = \frac{\text{Edit Distance}(\text{prediction}, \text{reference})}{\text{Length}(\text{reference})}. \quad (12)$$

Methods	ID20	P10_5	PE_AYB	Seq_1D	Avg.
BMA Look.	99.47	97.40	95.82	96.13	97.21
BMA Divider	99.48	97.70	95.70	96.62	97.38
Iter. Recon.	99.45	97.85	96.14	**97.45**	97.72
RobuSeqNet	95.90	96.04	94.02	94.52	95.12
FedDNA	**99.50**	**97.96**	**96.89**	97.26	**97.90**

Table 2: Success Rate(%) of FedDNA and four comparing algorithms on four real-world DNA datasets.

Baselines To evaluate the effectiveness of FedDNA, we conduct comparative experiments against two classical FL baselines FedAvg (McMahan et al. 2016) and FedProx (Li et al. 2020), alongside four aggregation-based approaches FedAWA (Shi et al. 2025), ConFREE (Zheng et al. 2025), FedAw(Tang 2024) and FedEvi (Chen et al. 2024). Direct model aggregation is infeasible for DNA sequence reconstruction due to varying client-specific reference lengths (L), which necessitates variable-length decoder outputs. Therefore, all methods utilize the same underlying network architecture and operate within a personalized FL framework, where only encoder parameters are globally aggregated to ensure fairness and comparability. Notably, non-evidential methods employ conventional softmax-based classification loss, whereas FedDNA and FedEvi utilize loss functions from the DEL framework.

Experimental Setup During federated training, each client performs one local epoch per communication round, for a total of 30 communication rounds. The local dataset for each client is partitioned into a training set (1,000 clusters), a validation set (1,000 clusters), and a test set (10,000 clusters). All experiments are repeated three times, and we report the mean and standard deviation of the performance.

Performance Analysis

Comparison with FL Algorithms Table 1 reports comparative experiments with the comparing FL methods; results for each are from the model achieving minimum validation loss over 30 federated training rounds.

For the primary metric Success Rate, FedDNA consistently achieves the best results across all datasets, averaging 97.91%. This marks a 0.84% improvement over the second-best FedEvi (97.07% average), an uncertainty-

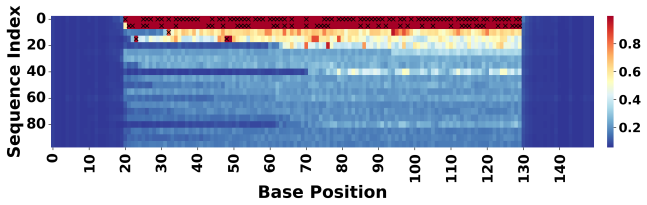


Figure 4: Heatmap of fused base-wise uncertainty in predicted DNA sequences. Color intensity ranges from blue (low uncertainty) to red (high uncertainty). × indicate incorrect base predictions.

based method. FedDNA substantially outperforms classical FL methods like FedAvg and FedProx by over 4.5%, and other advanced aggregation methods. Particularly on two challenging datasets (PE_AYB and Seq_1D), FedDNA significantly improves the success rate by 0.80% and 2.17%, respectively, compared to FedEvi. For Edit Error Rate, FedDNA likewise demonstrates top-tier accuracy, with the lowest average rate (0.12%), followed by ConFREE (0.13%). While FedEvi shows excellent results on ID20 (0.05%), FedDNA matches it on ID20 and achieves the lowest rate on P10_5 (0.06%). Notably on PE_AYB, while FedDNA's edit error rate (0.23%) is not lowest, its success rate (96.89%) is highest. This indicates FedDNA's prediction error distribution is more concentrated, prioritizing complete reconstruction for improved DNA storage integrity. Overall, these results demonstrate FedDNA's superior evidential federated aggregation strategy in DNA sequence reconstruction. Given shared underlying network architecture and PFL framework across methods, these significant gains are directly attributable to our novel strategy, effectively handling noise, uncertainty, and data heterogeneity for robust and precise sequence recovery. Moreover, FedDNA significantly surpasses the local training baseline (–FL) across all tested datasets, with performance gains reaching 2.25% on challenging Seq_1D, highlighting substantial benefits of the FL framework.

Figure 3 shows average client performance (30 communication rounds). Figure 3(a) shows most methods achieve very high base classification accuracy, directly driven by low training loss and suggesting strong local model performance. However, a notable divergence emerges in the more critical success rate (Figure 3(b)). FedDNA surpasses all baselines; FedEvi also performs strongly. These evidence-based methods show a slight convergence delay in initial 10 rounds, yet ultimately achieve superior performance.

Comparison with Sequence Reconstruction Algorithms We also compare FedDNA with SOTA DNA sequence reconstruction algorithms: BMA Lookahead (Gopalan et al. 2018), BMA divider (Sabary et al. 2024), Iterative Reconstruction (Sabary et al. 2024), and RobuSeqNet (Qin et al. 2024). Due to its large model size, RobuSeqNet tends to overfit on small-scale datasets; we thus trained it using 10,000 clusters. As Table 2 shows, FedDNA achieves the best average reconstruction performance, obtaining the highest success rate on Id20, P10_5, and PE_AYB, while Iterative Reconstruction slightly outperforms it on Seq_1D.

PHF	DEL	FAW	ID20	P10_5	PE_AYB	Seq_1D	Avg.
-	-	-	78.20	67.80	79.15	50.71	68.97
✓	-	-	97.07	88.43	95.92	87.78	92.30
✓	✓	-	99.07	97.86	95.81	93.81	96.64
✓	✓	✓	**99.50**	**97.96**	**96.89**	**97.26**	**97.91**

Table 3: Ablation study on Success Rate(%) of key components within FedDNA. Components include: PHF (Post-hoc Fusion module), DEL (Deep Evidential Learning), and FAW (Federated Aggregation Weighting).

Uncertainty-Based Analysis of Prediction Quality Using the ID20 dataset, we investigate the correlation between fused base uncertainty ($\overline{u}_l$) and prediction errors. From 10,000 FedDNA-reconstructed sequences on the test set, sequences were ranked by summing their base-wise uncertainties, and the top 100 were then systematically displayed by sampling every fifth sequence, yielding 20 representative visualizations. Figure 4 reveals a strong positive correlation: regions with higher uncertainty frequently correspond to incorrect base predictions. This evidence supports DEL's effectiveness and rationale for DNA sequence reconstruction.

Ablation Study Table 3 presents a systematic ablation study validating FedDNA's component efficacy. The baseline model, directly processing noisy clusters, achieved a mere average Success Rate of 68.97%, underscoring the task's inherent difficulty with small, high-noise real-world data. Implementing Post-hoc Fusion led to a significant performance leap, demonstrating the effectiveness of aggregating independent evidence for noise mitigation and reliable sequence reconstruction. Integrating DEL further enhanced model robustness against conflicting or low-quality data by quantifying prediction uncertainty, boosting Success Rate by 4.34%. The complete model, with its decomposed-uncertainty federated weighting, surpasses simple averaging, confirming the superiority of our adaptive weighting.

Conclusion

This work proposes FedDNA, a novel deep learning framework for DNA sequence reconstruction in privacy-sensitive federated environments. FedDNA pioneers the application of DEL, enabling precise reconstruction from noisy read clusters and robust uncertainty quantification via a convolutional-Mamba encoder with an LSTM decoder architecture. Leveraging DEL's ability to decompose predictive uncertainty, we propose a novel uncertainty-aware adaptive aggregation strategy. This mechanism dynamically adjusts federated weights to mitigate client data heterogeneity, for which we provide theoretical generalization bounds. Through extensive experiments, we demonstrate FedDNA significantly enhances DNA sequence reconstruction performance and reliability in federated settings, validating its efficacy in this challenging real-world application.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant 2020YFA0712100 and 2025YFC3409900, the National Natural Science Foundation of China under Grant 62476194 and U23B2049 and the Emerging Frontiers Cultivation Program of Tianjin University Interdisciplinary Center.

References

Arivazhagan, M. G.; Aggarwal, V.; Singh, A. K.; and Choudhary, S. 2019. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*.

Bar-Lev, D.; Orr, I.; Sabary, O.; Etzion, T.; and Yaakobi, E. 2025. Scalable and robust DNA-based storage via coding theory and deep learning. *Nature Machine Intelligence*, 7(4): 639–649.

Ceze, L.; Nivala, J.; and Strauss, K. 2019. Molecular digital data storage using DNA. *Nature Reviews Genetics*, 20(8): 456–466.

Chen, J.; Ma, B.; Cui, H.; and Xia, Y. 2024. FedEvi: Improving Federated Medical Image Segmentation via Evidential Weight Aggregation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 361–372.

Dou, C.; Yang, Y.; Zhu, F.; Li, B.; and Duan, Y. 2024. Explorer: efficient DNA coding by De Bruijn graph toward arbitrary local and global biochemical constraints. *Briefings in Bioinformatics*, 25(5): bbae363.

Goldman, N.; Bertone, P.; Chen, S.; Dessimoz, C.; LeProust, E. M.; Sipos, B.; and Birney, E. 2013. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 494(7435): 77–80.

Gopalan, P. S.; Yekhanin, S.; Ang, S. D.; Jojic, N.; Racz, M.; Strauss, K.; and Ceze, L. 2018. Trace reconstruction from noisy polynucleotide sequencer reads. US Patent App. 15/536,115.

Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8): 707–710.

Li, H.; and Durbin, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14): 1754–1760.

Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10713–10722.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. In *Proceedings of Machine learning and systems*, volume 2, 429–450.

Malinin, A.; and Gales, M. 2018. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, volume 31.

Mclaughlin, C.; and Su, L. 2024. Personalized federated learning via feature distribution adaptation. In *Advances in Neural Information Processing Systems*, volume 37, 77038–77059.

McMahan, H. B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2016. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Artificial intelligence and statistics*.

Meiser, L. C.; Nguyen, B. H.; Chen, Y.-J.; Nivala, J.; Strauss, K.; Ceze, L.; and Grass, R. N. 2022. Synthetic DNA applications in information technology. *Nature Communications*, 13(1): 352.

Mohri, M.; Sivek, G.; and Suresh, A. T. 2019. Agnostic federated learning. In *International Conference on Machine Learning*, 4615–4625.

Organick, L.; Ang, S. D.; Chen, Y.-J.; Lopez, R.; Yekhanin, S.; Makarychev, K.; Racz, M. Z.; Kamath, G.; Gopalan, P.; Nguyen, B.; Takahashi, C. N.; Newman, S.; Parker, H.-Y.; Rashtchian, C.; Stewart, K.; Gupta, G.; Carlson, R.; Mulligan, J.; Carmean, D.; Seelig, G.; Ceze, L.; and Strauss, K. 2018. Random access in large-scale DNA data storage. *Nature Biotechnology*, 36(3): 242–248.

Pan, C.; Tabatabaei, S. K.; Yazdi, S. M. H. T.; Hernandez, A. G.; Schroeder, C. M.; and Milenkovic, O. 2022. Rewritable two-dimensional DNA-based data storage with machine learning reconstruction. *Nature Communications*, 13(1): 2984.

Qin, Y.; Zhu, F.; Xi, B.; and Duan, Y. 2025. TransDNA: A Deep Transfer Learning Network for Sequence Reconstruction in DNA-Based Data Storage. *IEEE Transactions on Computational Biology and Bioinformatics*, 1–12.

Qin, Y.; Zhu, F.; Xi, B.; and Song, L. 2024. Robust multi-read reconstruction from noisy clusters using deep neural network for DNA storage. *Computational and Structural Biotechnology Journal*, 23: 1076–1087.

Qin, Z.; Yang, L.; Wang, Q.; Han, Y.; and Hu, Q. 2023. Reliable and interpretable personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20422–20431.

Qu, G.; Yan, Z.; and Wu, H. 2022. Clover: tree structure-based efficient DNA clustering for DNA-based data storage. *Briefings in Bioinformatics*, 23(5): bbac336.

Rashtchian, C.; Makarychev, K.; Racz, M.; Ang, S.; Jevdjic, D.; Yekhanin, S.; Ceze, L.; and Strauss, K. 2017. Clustering Billions of Reads for DNA Data Storage. In *Advances in Neural Information Processing Systems*, volume 30.

Sabary, O.; Yucovich, A.; Shapira, G.; and Yaakobi, E. 2024. Reconstruction algorithms for DNA-storage systems. *Scientific Reports*, 14(1): 1951.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. In *Advances in Neural Information Processing Systems*, volume 31.

Sentz, K.; and Ferson, S. 2002. Combination of evidence in Dempster-Shafer theory.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3): 379–423.

Shi, C.; Zhao, H.; Zhang, B.; Zhou, M.; Guo, D.; and Chang, Y. 2025. FedAWA: Adaptive Optimization of Aggregation Weights in Federated Learning Using Client Vectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 30651–30660.

Song, L.; Geng, F.; Gong, Z.-Y.; Chen, X.; Tang, J.; Gong, C.; Zhou, L.; Xia, R.; Han, M.-Z.; Xu, J.-Y.; Li, B.-Z.; and Yuan, Y.-J. 2022. Robust data storage in DNA by de Bruijn graph-based de novo strand assembly. *Nature Communications*, 13(1): 5361.

Tang, Y. 2024. Adapted weighted aggregation in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23763–23765.

Xie, M.; Li, S.; Zhang, R.; and Liu, C. H. 2023. Dirichlet-based uncertainty calibration for active domain adaptation. *arXiv preprint arXiv:2302.13824*.

Yu, M.; Tang, X.; Li, Z.; Wang, W.; Wang, S.; Li, M.; Yu, Q.; Xie, S.; Zuo, X.; and Chen, C. 2024. High-throughput DNA synthesis for data storage. *Chemical Society Reviews*, 53(9): 4463–4489.

Zheng, H.; Hu, Z.; Yang, L.; Zheng, M.; Xu, A.; and Wang, B. 2025. ConFREE: Conflict-free Client Update Aggregation for Personalized Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22875–22883.