Robust DNA Image Storage Decoding with Residual CNN

Cihan Ruan*, Liang Yang†, Rongduo Han*‡, Shan Gao§, Haoyu Wu¶, and Nam Ling*
cruan@scu.edu, 2120140433@mail.nankai.edu.cn, a12910@qq.com gao_shan@mail.nankai.edu.cn. wuhaoyujerry@gmail.com, nling@scu.edu
*Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA, USA
†College of Software, Nankai University, Tianjin, China
‡School of Mathematical Sciences, Nankai University, Tianjin, China
§College of Life Sciences, Nankai University, Tianjin, China
¶Roku, Inc., San Jose, CA, USA

2024 IEEE International Symposium on Circuits and Systems (ISCAS) | 979-8-3503-3099-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISCAS58744.2024.10558127

***Abstract*—DNA storage is a promising data storage method with high density, durability, and easy maintenance, ideal for data archiving. However, wide-scale adoption is hindered by challenges like high synthesis costs, data loss, and I/O complexities. Addressing robustness is a primary concern in advancing DNA storage. Traditional strategies for robustness rely on increasing redundancy, replicas, and error-correction codes (ECC) for each DNA sequence strand. Given the unpredictable errors in DNA storage and the associated costs of absolute accuracy, we've embraced an error-tolerance approach. This paper introduces an innovative method utilizing the residual Convolutional Neural Network (CNN) during image decoding in DNA storage to combat noise and enhance robustness. We simulated compressed images in DNA sequences and restored them using our network, achieving commendable peak signal-to-noise ratios (PSNR) even with lower-quality images. Our method offers a balance between redundancy and image quality in DNA storage.**

***Index Terms*—DNA storage, Error tolerance, Robustness, Image decoding, Residual learning**

I. Introduction

The development of data storage has undergone a significant evolution throughout human history, from cave paintings to solid-state drives (SSDs). However, traditional storage techniques exhibit shortcomings such as limited storage capacity, susceptibility to mechanical damage, and high energy consumption. Over the past few years, the huge amount of data from short video platforms and the pandemic made those drawbacks more pertinent. According to projections from International Data Corporation (IDC), global data volume is expected to explode to an astonishing 175 zettabytes (ZB) by 2025 [1]. This unparalleled growth presents a significant challenge as it could potentially deplete the global supply of silicon used in computers by 2040 [2].

To address these challenges, researchers are seeking new data storage strategies to provide more efficient and stable solutions. Being the oldest yet most cutting-edge storage material, DNA has captured researchers' attention. Similar to computers using 0s and 1s for data storage, DNA employs four different elements, namely adenine (A), cytosine (C), guanine (G), and thymine (T). These four nucleotides can store the fundamental information of all living organisms throughout the world. Therefore, DNA storage can be viewed as a quaternary encoding scheme [3]. In comparison to silicon-based storage devices, DNA molecules have high-density storage potential. Furthermore, DNA storage has advantages such as high stability and low maintenance cost [4] [5]. Over the past decade, numerous efforts have focused on research on the use of DNA for digital file storage, with a key concern being how to apply nucleotides for data compression and encoding [6] [7] [8].

DNA storage also has several limitations, such as expensive costs and slow I/O speeds [9]. But the most significant drawback of DNA storage is the error proneness of insertions, deletions, and substitutions (IDS) in nucleotides, which causes data loss [10]. This drawback is attributed to the current limitations of existing genomics technologies and DNA's natural properties. Adhering to biologically-constrained encoding can enhance the stability of DNA chains to some extent. However, unavoidable errors still emerge at various stages of the experiments. Researchers have attempted to solve this problem by adopting robust coding schemes, making duplication of DNA strands, and designing error correction codes (ECC). Those efforts in the realm of DNA storage have focused on achieving 100% accurate recovery of stored digital files, with research primarily directed at how to use nucleotides for data compression and encoding. However, an overemphasis on perfect recovery might not always be the most practical approach, given the intrinsic challenges associated with DNA storage. Moreover, those attempts introduce additional redundancy and increase the storage cost.

In this paper, we present a paradigm shift by proposing a robust DNA image storage algorithm. Instead of solely aiming for perfect recovery, we focus on optimizing image quality through decoding. We employ a residual convolutional neural network (CNN) structure with a residual dense block (RDB) as a decoder. This network extracts features from compressed images and enhances image restoration efficiency. The unique strength of this approach lies in the residual CNN's ability to minimize image noise and highlight intricate details, directly addressing DNA storage's limitations.

The remainder of this paper is organized as follows. Sec-

Authorized licensed use limited to: Santa Clara University. Downloaded on August 31,2025 at 04:26:03 UTC from IEEE Xplore. Restrictions apply.

tion II introduces the background and related works on the robustness of DNA storage. Section III presents our proposed novel DNA decoder based on residual CNN. In Section IV, we present the results of the experiment and discuss the effectiveness of our approach to achieve reliable image data storage on DNA. Finally, Section V summarizes the key findings of our research and discusses the implications of our proposed DNA decoding strategy for image storage.

II. Background

Fig.1 illustrates the workflow of DNA image storage. When saving an image, we compress the image into binary bitstreams and convert them into nucleobase sequences. In a wet laboratory facility, these sequences are processed into DNA strands and stored in tubes. To recover the stored data, the polymerized chain reaction (PCR) and DNA sequencing are applied to obtain nucleobase sequences. After receiving the data from the wet lab, we applied a decoder to recover the binary bitstreams we stored. This workflow emphasizes the fundamental stages of encoding, synthesis, storage, and retrieval in DNA data storage.

In the context of DNA storage realities, the cumulative error rate throughout the experimental process is estimated to be around 1%-2% per nucleotide. This rate is derived from second-generation DNA synthesis and sequencing genomics techniques that are widely applied, coupled with the intrinsic mutation tendencies inherent in DNA replication. It's worth noting that this error rate can be influenced by the length of the DNA strand and specific experimental conditions. [11] [12] [13] [14].

Related Work: Examining the landscape of robust DNA storage methods, a significant portion of prior research primarily targets perfect data recovery. Such works employ a variety of strategies. Among these strategies, some utilize encoding methods inherently characterized by robustness, including the fountain [15] [16], Reed-Solomon code [17], and Hamming code [18]. When the error rate of IDS is minor, it's feasible for researchers to discard irregular DNA sequences during the encoding process. This approach, however, becomes inadequate when confronted with higher error rates.

Other traditional error correction approaches include the use of ECC [19] [20] [21] and multi-strand backups [13]. Although these methods can improve robustness, they also come with certain limitations. For example, ECC might not be able to fully rectify a significant number of errors, and backup strategies could increase considerable costs. In summary, these error correction methods, which are the pursuit of perfect data recovery, solve the error proneness in genomics processes by introducing redundancy into the original information chain, which impacts the efficiency and cost of the storage system.

Several deep learning-driven DNA storage encoding models have shown promising outcomes [22] [23]. However, these works allocate a significant portion of their works to describe the encoding process, while lacking consideration for robustness in the decoding process.

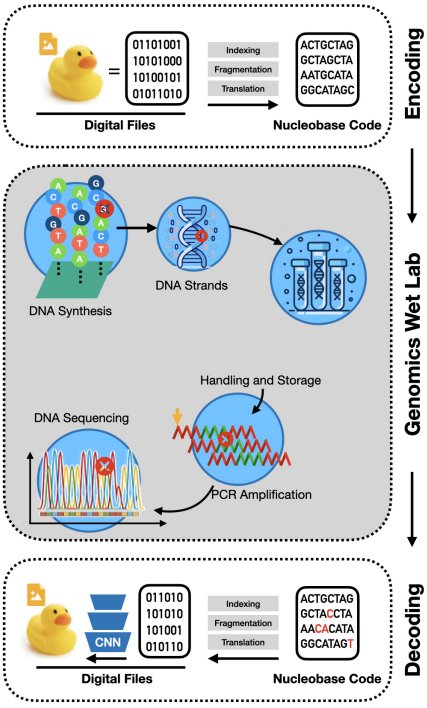

Fig. 1. Workflow of DNA data storage, indicating stages vulnerable to IDS errors. Potential points susceptible to IDS errors are indicated by red cross icons, emphasizing the challenges in maintaining data integrity during the storage process.

Our approach introduces an error-tolerant algorithm for DNA image storage. Instead of pursuing perfect recovery, the algorithm is designed to manage the inevitable disruptions caused by IDS in the DNA strands. Drawing inspiration from Zan et al. [24] and their exploration into DNA textual data storage, we specifically zoom in on DNA image data storage. Within image data, disruptions from IDS manifest as noise. Our research aims to manage this noise and improve robustness without heavy redundancy. We use the power of residual learning within the network architecture to better extract complex image features, perform super-resolution, and reduce noise, thus achieving enhanced robustness in image recovery. This innovation offers a new avenue in the DNA storage domain.

III. Structure of the Proposed Residual CNN

The structure of our super-resolution residual CNN is shown in Fig.2. This network architecture transforms low-resolution input into high-resolution output. A detailed explanation of this structure is provided in the subsequent sections.

A. Main Idea

In the context of DNA storage, images demonstrate a superior error tolerance when compared to other data formats. This is primarily because image-related errors can manifest as noise, which can subsequently be minimized. In our methodology, to emulate the noise observed in actual images derived from DNA strands, Gaussian noise is introduced to the images.

Authorized licensed use limited to: Santa Clara University. Downloaded on August 31,2025 at 04:26:03 UTC from IEEE Xplore. Restrictions apply.

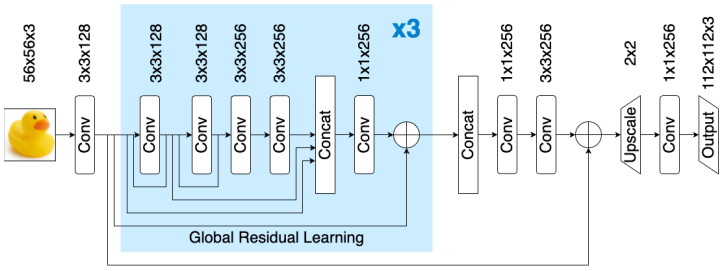

Fig. 2. The architecture of our proposed residual dense network

We opt for Gaussian noise to model the errors, as single-point errors commonly arise in experiments when encoding DNA bases adhering to biological constraints such as GC-content, avoiding homopolymers, and maintaining a certain short-chain length. Such errors can be aptly abstracted as Gaussian noise.

We leverage a residual dense network (RDN) as the foundation for our super-resolution technique. Within the RDN architecture, the fusion of Dense Blocks and Residual Learning mechanisms extract and restore image features, resulting in exceptional performance across various processing tasks.

B. RDN Structure

Our deep neural network is designed primarily for image super-resolution using the residual dense network (RDN) architecture. It comprises 33 convolutional layers followed by a linear output layer that maps into a softmax. Initially, it extracts shallow features from the input low-resolution image. This is followed by a series of residual dense blocks (RDB) and feature fusion blocks (FFB). The final phase involves an upsampling mechanism to produce the desired high-resolution result. Let's denote the network's input and output as $Input_{LR}$ and $Output_{HR}$, respectively. The primary convolutional layers are tasked with extracting rudimentary image features, represented as:

$$Feature_l = C_{Shallow^l}(Input_{LR}) \tag{1}$$

where C symbolizes convolutional operations and in our model, the number of layers dedicated to shallow image feature extraction, l, is 3.

Subsequent to this, the extracted shallow features undergo further processing via the next set of residual blocks. To represent the flow of these operations in a concise manner, let's define the following notations:

- R_m: Denotes the residual operation at the m-th layer.
- D_i: Denotes the operation within the i-th dense block.

Given these definitions, assuming there are m layers of RDBs, the progression of features through the residual dense blocks is succinctly described as:

$$Feature_{m+l} = R_m(D_{m-1} \dots D_1(Input_{LR})) \tag{2}$$

where R_m means the operations of the m-th RDB. The contiguous memory mechanism ensures that the previous RDB's state is passed on to each layer of the current RDB. This merging of states is facilitated by the local feature fusion process. It amalgamates the states from both the preceding RDB and all convolutional layers within the current RDB. Moreover, local residual learning is integrated into the RDB, enhancing the information flow. This becomes considering the multitude of convolutional layers housed within a single RDB.

After extracting local dense features using a series of RDBs, our methodology introduces Feature Fusion Blocks (FFBs). These FFBs are designed to exploit hierarchical features on a broader scale by amalgamating the features from every individual RDB. Central to our FFB strategy are two primary mechanisms: global feature fusion and global residual learning. Through this, the process is empowered to seamlessly use the features drawn from all the earlier layers:

$$Feature_{fin} = C_{feature}(\sum_{i=0}^{n} Feature_i) \tag{3}$$

Subsequently, our approach deploys global residual learning. This is key for securing feature maps right before the commencement of the up-scaling operation. With both local and global features extracted from the low-resolution (LR) domain, our next step is the integration of an Upsampling Network (UPNet) within the high-resolution (HR) dimension. Assuming our whole network as C_{RDN}, the complete process can be abstracted by:

$$Output_{HR} = C_{RDN}(Input_{LR}) \tag{4}$$

Furthermore, the loss function governing the entire RDN is defined by the L_2 norm.

C. Data Augmentation for Enhanced Robustness and System Scalability

We initially trained our residual CNN network on a standard image dataset. Recognizing the inherent noise introduced during the DNA storage process, we incorporated various levels and intensities of Gaussian noise into the dataset to further train and refine our network. This augmented training strategy emulates the real-world uncertainties encountered in DNA storage, enhancing the model's resilience and adaptability.

Our network functions primarily as the decoder end of the entire DNA image storage system. Interestingly, while the exact encoder algorithm doesn't remain our primary focus, ensuring the network's input as a compressed image is crucial.

Authorized licensed use limited to: Santa Clara University. Downloaded on August 31,2025 at 04:26:03 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARISON OF DNA STORAGE METHODS BASED ON DIFFERENT LITERATURE SOURCES.

Features/Methods	Our Method	DNA-Fountain Code [7]	ECC Design [21]	Strand Overlapping [25]	Duplicate & Voting [26]
Redundancy Level	Low	Medium	High	High	Low-Medium
Error Handling	Tolerates minor errors	Error correction	Error correction	Error mitigation	Error consensus
Decoding Complexity	Simplified with CNN	Moderate	Potentially complex	Moderate	Moderate
Cost Implications	Low	Moderate-High	Moderate-High	High	High

Given our network's inherent error tolerance, this decoder can cooperate with other robust encoder algorithms to further elevate the overall system's robustness, making it an innovative cornerstone for long-term data archiving in the DNA storage.

IV. EXPERIMENTAL METHODOLOGY AND ANALYSIS

We conducted our experiments using a data set comprising a total of 8,600 images from the ECG dataset. Among these, 70% was allocated for training, 10% for validation, and the remaining 20% for testing purposes. Within each training batch, we extracted 16 LR RGB patches as input data in a random manner. These patches went through random augmentation, involving horizontal or vertical flips as well as rotations.

Initially, our experiment centered on noise-free images to gauge the inherent capabilities of the RDN network. To corroborate the effectiveness of our RDN design, we utilized a bicubic model for simulating LR images with scaling factors of ×2 and ×4. Fig.3 presents visual comparisons at a scale of ×4. Evidently, our RDN excels in restoring crisp and defined edges, mirroring the ground truth with remarkable accuracy.

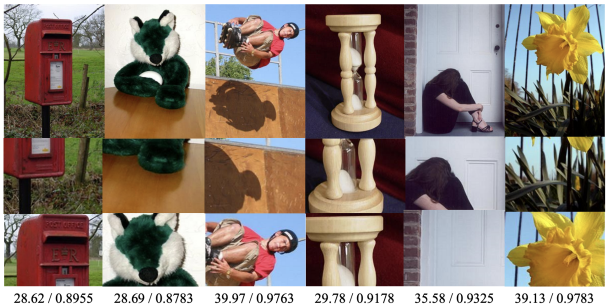

Fig. 3. Visual comparisons at a scale of ×4 using the proposed work in PSNR/SSIM.

In later experiments, we introduced varying noise levels to input low-resolution images to evaluate the restoration prowess of the network. Fig.4 presents visual comparisons at scaling factors ×2 and ×4 across distinct Gaussian noise rates, with PSNR and SSIM data comparisons for restored images in Fig.5. Impressively, our network maintains strong restoration performance and a high PSNR value even when noise levels reach 10%. Performance at ×2 scaling consistently outperforms ×4, a trait credited to our RDN's unique architecture.

Our results underline the efficiency of our deep learning approach for DNA-based image storage and retrieval. Even with constrained information storage, our network achieves notable image restoration against substantial noise disruptions.

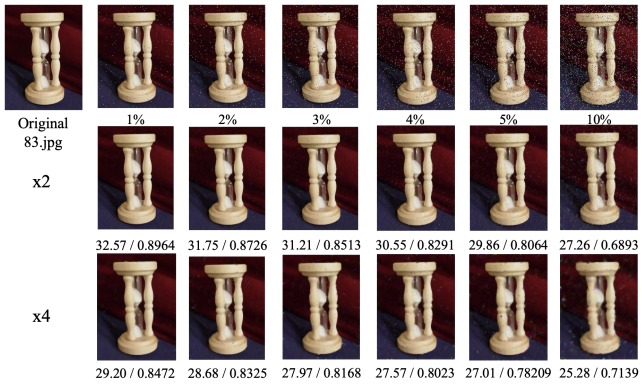

Fig. 4. The comparison of final recovery results by the proposed network in PSNR/SSIM under different noise rates.

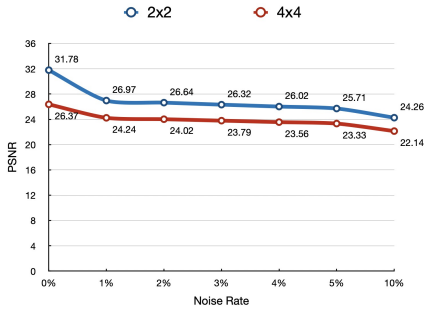

Fig. 5. The comparison of PSNR results for the final recovery outcomes achieved by the proposed network is conducted across varying noise rates and scaling factors.

Table I outlines a comparative view of various robust DNA storage strategies. Given the unique emphasis of our work on the decoding aspect of DNA image storage, our comparisons with existing studies remain at a conceptual level.

V. CONCLUSION

In conclusion, this study presents a unique approach applying residual CNN to boost the robustness and efficiency of DNA-based image storage and recovery. Leveraging the RDN network's capabilities, we showed the potential to store compressed image data in DNA while ensuring accurate image recovery, even amidst noise. Key to our method is its distinct error tolerance, safeguarding data integrity amid uncertainties. The results highlight its versatility across scenarios and its ability to reduce redundancy and optimize data use. The applied neural network also hints at broad applicability, potentially extending to audio and video data recovery.

Authorized licensed use limited to: Santa Clara University. Downloaded on August 31,2025 at 04:26:03 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] J. R. David Reinsel, John Gantz. (2017) Data age 2025. [Online]. Available: https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf

[2] A. Extance, "How DNA could store all the world's data," *Nature*, vol. 537, no. 7618, 2016.

[3] Y. Dong, F. Sun, Z. Ping, Q. Ouyang, and L. Qian, "DNA storage: research landscape and future prospects," *National Science Review*, vol. 7, no. 6, pp. 1092–1107, 2020.

[4] A. Swati, F. Mathuria, S. Bhavani, E. Malathy, and R. Mahadevan, "A review on various encoding schemes used in digital DNA data storage," *Int. J. Civ. Eng. Technol*, vol. 8, no. 12, pp. 108–114, 2017.

[5] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.

[6] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *nature*, vol. 494, no. 7435, pp. 77–80, 2013.

[7] Y. Erlich and D. Zielinski, "DNA fountain enables a robust and efficient storage architecture," *science*, vol. 355, no. 6328, pp. 950–954, 2017.

[8] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen *et al.*, "Scaling up DNA data storage and random access retrieval," *BioRxiv*, p. 114553, 2017.

[9] D. Panda, K. A. Molla, M. J. Baig, A. Swain, D. Behera, and M. Dash, "DNA as a digital information storage device: hope or hype?" *3 Biotech*, vol. 8, pp. 1–9, 2018.

[10] R. Xie, X. Zan, L. Chu, Y. Su, P. Xu, and W. Liu, "Study of the error correction capability of multiple sequence alignment algorithm (mafft) in DNA storage," *BMC bioinformatics*, vol. 24, no. 1, pp. 1–11, 2023.

[11] E. LeProust, "Agilent's microarray platform: How high-fidelity DNA synthesis maximizes the dynamic range of gene expression measurements," *Agilent Technologies, Santa Clara, CA*, p. 45, 2008.

[12] K. Reinert, B. Langmead, D. Weese, and D. J. Evers, "Alignment of next-generation sequencing reads," *Annual review of genomics and human genetics*, vol. 16, pp. 133–151, 2015.

[13] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen *et al.*, "Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, pp. 242–248, 2018.

[14] P. McInerney, P. Adams, and M. Z. Hadi, "Error rate comparison during polymerase chain reaction by DNA polymerase," *Molecular biology international*, vol. 2014, 2014.

[15] Y.-J. Chen, C. N. Takahashi, L. Organick, C. Bee, S. D. Ang, P. Weiss, B. Peck, G. Seelig, L. Ceze, and K. Strauss, "Quantifying molecular bias in DNA data storage," *Nature communications*, vol. 11, no. 1, p. 3264, 2020.

[16] C. Ruan, R. Han, Y. Li, S. Gao, H. Wu, and N. Ling, "Efficient DNA-based image coding and storage," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2023, pp. 1–5.

[17] L. C. Meiser, P. L. Antkowiak, J. Koch, W. D. Chen, A. X. Kohll, W. J. Stark, R. Heckel, and R. N. Grass, "Reading and writing digital data in DNA," *Nature protocols*, vol. 15, no. 1, pp. 86–101, 2020.

[18] C. N. Takahashi, B. H. Nguyen, K. Strauss, and L. Ceze, "Demonstration of end-to-end automation of DNA data storage," *Scientific reports*, vol. 9, no. 1, p. 4998, 2019.

[19] L. Song, F. Geng, Z. Song, B.-Z. Li, and Y.-J. Yuan, "Robust data storage in DNA by de Bruijn graph-based decoding," 2021.

[20] T. Xue and F. C. Lau, "Construction of gc-balanced DNA with deletion/insertion/mutation error correction for DNA storage system," *Ieee Access*, vol. 8, pp. 140 972–140 980, 2020.

[21] W. H. Press, J. A. Hawkins, S. K. Jones Jr, J. M. Schaub, and I. J. Finkelstein, "Hedges error-correcting code for DNA storage corrects indels and allows sequence constraints," *Proceedings of the National Academy of Sciences*, vol. 117, no. 31, pp. 18 489–18 496, 2020.

[22] P. Li, K. Cai, G. Song, W. Song, Z. Mei, and X. Zhong, "Neural network-based decoding of constrained codes for DNA data storage," in *2020 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*. IEEE, 2020, pp. 1–4.

[23] A. Rasool, Q. Qu, Y. Wang, and Q. Jiang, "Bio-constrained codes with neural network for density-based DNA data storage," *Mathematics*, vol. 10, no. 5, p. 845, 2022.

[24] X. Zan, X. Yao, P. Xu, Z. Chen, L. Xie, S. Li, and W. Liu, "A hierarchical error correction strategy for text DNA storage," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 14, no. 1, pp. 141–150, 2022.

[25] Z. Ping, D. Ma, X. Huang, S. Chen, L. Liu, F. Guo, S. J. Zhu, and Y. Shen, "Carbon-based archiving: current progress and future prospects of DNA-based data storage," *GigaScience*, vol. 8, no. 6, p. giz075, 2019.

[26] C. Xu, B. Ma, Z. Gao, X. Dong, C. Zhao, and H. Liu, "Electrochemical DNA synthesis and sequencing on a single electrode with scalability for integrated data storage," *Science Advances*, vol. 7, no. 46, p. eabk0100, 2021.

[27] L. Sun, J. He, J. Luo, and D. H. Coy, "DNA and the digital data storage," *Health Science Journal*, vol. 13, no. 3, pp. 1–7, 2019.

[28] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[29] M. Schwarz, M. Welzel, T. Kabdullayeva, A. Becker, B. Freisleben, and D. Heider, "MESA: automated assessment of synthetic DNA fragments and simulation of DNA synthesis, storage, sequencing and pcr errors," *Bioinformatics*, vol. 36, no. 11, pp. 3322–3326, 2020.

[30] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the DNA data storage channel," *Scientific reports*, vol. 9, no. 1, p. 9663, 2019.

Authorized licensed use limited to: Santa Clara University. Downloaded on August 31,2025 at 04:26:03 UTC from IEEE Xplore. Restrictions apply.