HybridFlow-DNA: A Deep Generative Compression Framework for DNA Storage of Images

Cihan Ruan*, Rongduo Han*†, Shan Gao‡, Lei Lu§¶, Wei Jiang¶, Wei Wang¶, Haoyu Wu‖ and Nam Ling*
Email: cruan@scu.edu, hrd12910@gmail.com, gao_shan@mail.nankai.edu.cn, lu.lei1@northeastern.edu
{wjiang, rickweiwang}@futurewei.com, 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
‡College of Life Sciences, Nankai University, Tianjin, China
§Electrical and Computer Engineering, Northeastern University, MA, USA
¶Futurewei Technologies Inc., Santa Clara, CA, USA
‖Roku, Inc., San Jose, CA, USA

2025 IEEE International Symposium on Circuits and Systems (ISCAS) | 979-8-3503-5683-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISCAS56072.2025.11044311

***Abstract*—DNA storage has emerged as a promising solution to address the exponentially growing demand for storage capacity, offering advantages in density, stability, and long-term preservation potential. Currently, image compression for DNA storage has evolved into learned image compression (LIC), particularly through the application of deep learning methods based on artificial neural networks. The present study proposes a novel image compression framework for DNA Storage, named HybridFlow-DNA. HybridFlow-DNA is established by integration of VQGAN and MLIC with the adaptive dynamic DNA fountain encoding scheme. Experimental results demonstrate that HybridFlow-DNA achieves a high virtual information capacity while effectively maintaining the fidelity of the reconstruction of images.**

***Index Terms*—LIC, VAE, VQGAN, JPEG-DNA, Goldman**

I. Introduction

The annual volume of the global datasphere will increase to 175 ZB by 2025 [1], a substantial expansion that presents a multitude of challenges [2]. In response, DNA storage has emerged as a promising solution, characterized by its ultrahigh information storage density of 215 petabytes per gram (PB/g) and exceptional durability [3]. Early research work established the foundations for DNA storage. In particular, Goldman et al. [4] developed the first large-scale DNA storage system with a ternary coding scheme, achieving an information storage density of 2.2 PB/g. Subsequently, the emergence of Fountain DNA encoding (Fountain-DNA) [5] advanced the field by achieving a potential information capacity of 1.98 bits per nucleotide (bits/nt), close to the theoretical upper limit of 2 bits/nt, established by the Shannon-Hartley theorem when treating DNA storage as a communication channel.

Theoretically, a DNA storage system can achieve a Shannon information capacity of 2 bits/nt, which is twice that of a binary digital storage system. However, the effective information capacities in practical DNA storage systems are lower because of biological constraints on encoding and the necessity for error correction or tolerance. Both Shannon and effective information capacities are traditionally defined without taking into account data compression. As data compression is a routine procedure in DNA storage for images, we propose the concept of virtual information capacity, defined as the product of Shannon information capacity and the data compression ratio. This metric facilitates a direct comparison of various DNA storage systems by indicating their information capacities in a more meaningful manner.

As image compression increases the virtual information capacity, many existing methods have been introduced. For example, Pan et al. introduced quaternary Huffman coding to convert RGB channel-wise bitstreams into DNA codes [6]. Another notable example is IMG-DNA, which was designed to optimize DNA structures for synthesis [7]. These image compression methods are classified into conventional methods, in contrast to the learned image compression (LIC) methods. According to this classification criterion, more advanced methods such as JPEG [8] and VVC [9], which have been developed in the JPEG-DNA [10] and VVC-DNA frameworks [11] are also classified as conventional methods.

With the advent of machine learning methods, particularly those that use deep learning based on artificial neural networks, image compression has evolved into LIC. A variety of LIC methods are emerging as competitors to conventional methods. Notable LIC methods include variational autoencoder (VAE) [12] that achieves a high compression ratio while effectively maintaining reconstruction fidelity, vector-quantized generative adversarial network (VQGAN) [13] that allows high-fidelity reconstruction of images through adversarial training, multilevel image compression (MLIC) [14] that effectively extracts multiscale features through hierarchical structures, and HybridFlow [15] that integrates VQGAN and MLIC.

Currently, image compression for DNA storage has also evolved into LIC, particularly through the application of deep learning methods based on artificial neural networks. For example, Strebel et al. [16] explored the integration of deep learning methods into DNA storage systems, aiming to enhance the efficiency and precision of encoding and decoding processes. However, their study revealed that these advanced methods were vulnerable to synthesis errors, which could significantly impact the reliability of the system. Zheng et al. [17] proposed an image encoding scheme for DNA storage, named

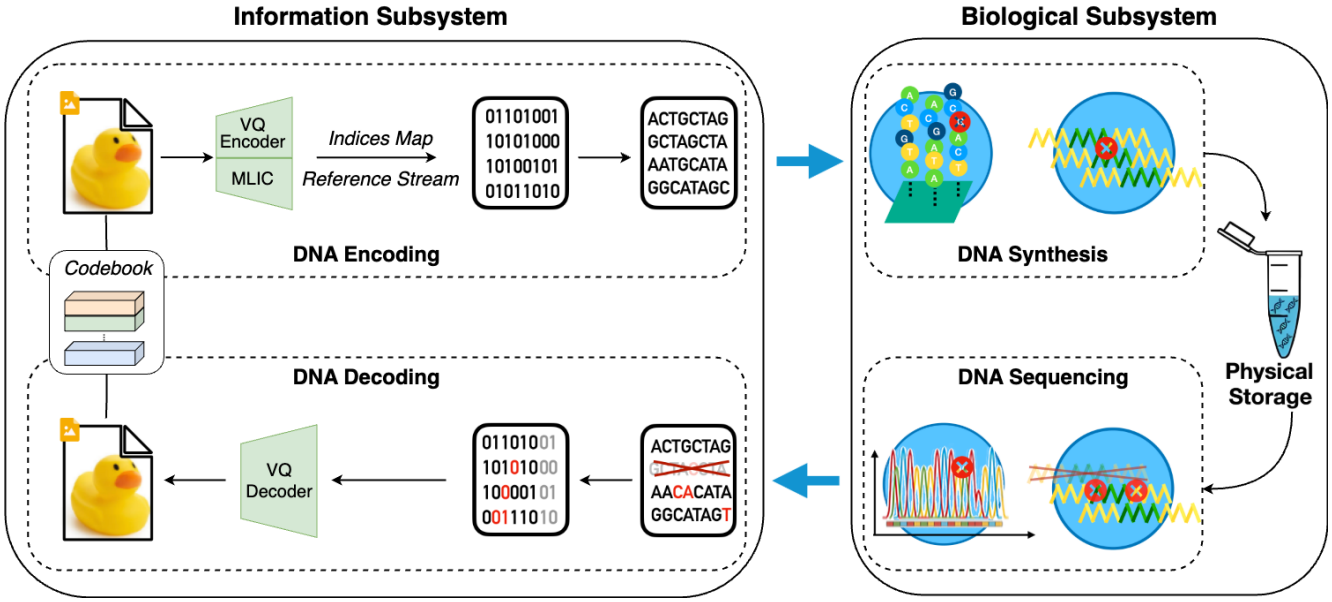

Fig. 1: Overview of HybridFlow-DNA

DNA-QLC, which integrates a quantized ResNet variational autoencoder (QRes-VAE) with a Levenshtein code (LC) for error correction. Wu et al. [18] proposed the DJSCC- DNA framework, which integrates deep learning methods into DNA storage to improve error tolerance and optimize biological constraints, while Pic et al. [19] explored techniques to improve the efficiency and reliability of DNA encoding. These previous studies demonstrated that image compression for DNA storage must take into account at least virtual information capacity, reconstruction fidelity, and error-tolerance capacity. Previous work has focused on improving the decoding process, such as employing residual CNN-based techniques to improve the robustness of decoding [20]. However, neither study fully addressed the challenge of achieving virtual information capacity while effectively maintaining reconstruction fidelity end-to-end.

The objective of this study is to develop an image compression framework for DNA storage, with a focus on maximizing virtual information capacity, minimizing loss of fidelity, and improving error tolerance. Consequently, we proposed a novel framework, named HybridFlow-DNA, based on a previously proposed HybridFlow [15]. HybridFlow-DNA integrates HybridFlow with the adaptive dynamic DNA fountain encoding scheme (ADF-DNA). In the present study, ADF-DNA was proposed, modified from dynamic DNA fountain encoding proposed in our previous study [11]. Ultimately, we demonstrated that HybridFlow-DNA can achieve a high virtual information capacity while effectively maintaining reconstruction fidelity, as evidenced by simulation results.

II. Proposed Method

A. HybridFlow Architecture

HybridFlow-DNA includes VQGAN, MLIC, and ADF-DNA (Fig. 1). In the present study, the ADF-DNA module was integrated into HybridFlow-DNA. However, the use of alternative DNA encoding modules does not affect the overall performance of HybridFlow-DNA. In the present study, HybridFlow-DNA was proposed, inheriting HybridFlow [15], which is a codebook-based compression method. In HybridFlow-DNA, the VQGAN/MLIC encoder generates two categories of information (indices map and reference information) for each image. VQGAN extracts critical latent features of an input image and quantizes them into an indices map using a visual codebook. The visual codebook is crucial for image reconstruction, serving as a shared prior knowledge between the encoder and decoder. Meanwhile, MLIC captures spatial and contextual dependencies to provide reference information, thereby facilitating the decoding process. It is important to note that the VQGAN/MLIC encoder and decoder operate outside the DNA encoding and decoding in the HybridFlow-DNA workflow (Fig.1).

HybridFlow's masked image modeling allows only a portion of the indices map to be transmitted, and the missing indices are inferred through a mask predictor, effectively reducing the bitrate while maintaining reconstruction quality. This feature makes HybridFlow particularly suitable for the error-prone DNA storage environment, as the dual- stream design offers complementary benefits: ensuring high performance and helping in masked region recovery.

B. DNA Storage of an Image

In both HybridFlow and HybridFlow-DNA, the visual codebook is crucial for image reconstruction, serving as a shared prior knowledge between the encoder and decoder. For image transmission, it is assumed that the visual codebook is synchronized between the encoder and the decoder. For DNA storage, the visual codebook needs to be stored once for all images. Given that the visual codebook is highly sensitive to

errors, it needs to be duplicated and stored using multiple methods to ensure reliability. For each image, the map of the indices and the reference information are processed by deep learning methods that tolerate a certain degree of errors, thereby enabling them to be stored with the image data in a compressed file. Therefore, visual codebooks, indices maps, and reference information (Eq. (1)) must be stored using two different strategies.

$$\text{Image}_n = \{C, d_n(C), y_n(C)\}, \quad C \in \mathbb{R}^{T \times L} \tag{1}$$

In **Eq. (1)**, C represents a visual codebook, where T is the total number of words in the visual codebook and L is the dimension of a word; d represents the indices map of an image and y represents the reference information, where n denotes the number of images.

Similarly to traditional Fountain-DNA [5], ADF-DNA (Fig. 2(a)) begins with pre-processing tailored to different data types, followed by the common steps of packaging, Luby transformation, and adaptive DNA mapping ((Fig. 2(c)). One important feature of ADF-DNA is that it decomposes a codebook into quantized values Q and residuals R (Eq. (2) and Algorithm 1). This decomposition improves the information capacity of ADF-DNA. Quantized values are restricted within 8-bit precision, enabling compact encoding, while residuals typically concentrate near zero with a sparse distribution, facilitating efficient DNA encoding.

$$Q \leftarrow \text{Quantize}(C, 8 \text{ bits}), R \leftarrow C - Q \tag{2}$$

C. Error-tolerance capacity

HybridFlow-DNA exhibits a certain degree of error tolerance capacity, which inherits from HybridFlow [15]. HybridFlow's masked image modeling allows only a portion of the indices map to be transmitted, and the missing indices are inferred through a mask predictor. This feature makes HybridFlow particularly suitable for the error-prone DNA storage environment, as its dual-stream design offers complementary benefits: ensuring high performance and helping in masked region recovery. Furthermore, ADF-DNA is inherently tolerant of errors [5].

III. Experimental Methodology and Analysis

A. Experimental Design

Two standard image datasets (Kodak and CLIC) were used to train the models. Each data set was used independently for testing the models to facilitate comparisons among the different methods. The performance of each method or framework was evaluated in terms of information capacity and image reconstruction fidelity using two metrics, which are virtual information capacity (bits/nt) and the perceptual image patch similarity (LPIPS), respectively. In the present study, LPIPS was calculated between the compressed image and its original image pairwise using VGG, instead of the more commonly used peak signal-to-noise ratio (PSNR), as PSNR often fails to capture the nuances of human visual perception, particularly in the context of LIC methods [21], [22]. Recent LIC studies

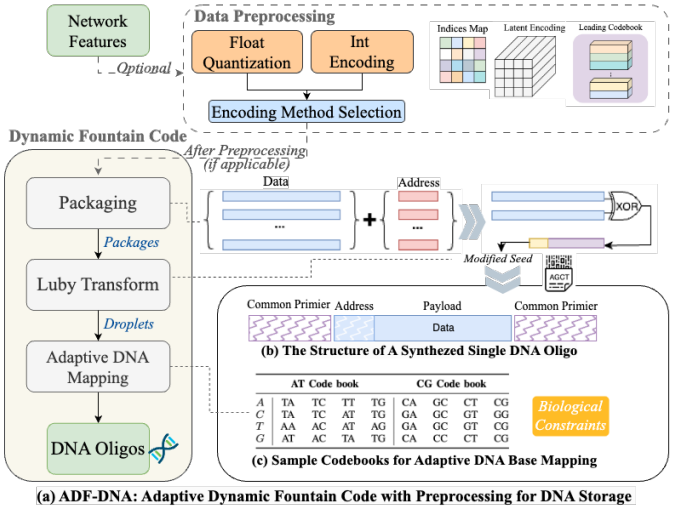

Fig. 2: Workflow of ADF-DNA
(a) ADF-DNA for DNA Storage. (b) A synthesized DNA sequence consists of three fundamental regions: the common primer, address, and data payload regions. (c) Codebooks for adaptive DNA base mapping are designed to adapt flexibly to target DNA sequence length and biological constraints, providing high efficiency with lower computational complexity.

Algorithm 1 Residual Encoding with ADF-DNA for Codebook C

Require: Codebook C
Ensure: DNA sequences for storage
Quantization:
$Q \leftarrow \text{Quantize}(C, 8 \text{ bits})$ {8-bit quantization}
$R \leftarrow C - Q$ {Calculate residuals}
ADF-DNA Encoding:
$Q_{enc}, R_{enc} \leftarrow \text{Encode}(Q, R, 1.83)$ {With redundancy}
DNA Mapping:
$Q_{dna}, R_{dna} \leftarrow \text{MapToDNA}(Q_{enc}, R_{enc})$
Decoding and Reconstruction:
$Q', R' \leftarrow \text{Decode}(\text{SequenceDNA}())$
$C' \leftarrow Q' + R'$ {Reconstruct codebook}
return C' =0

have further explored deep generative models and GAN-based architectures for low-bitrate image coding, achieving high-quality reconstructions at extreme compression rates.

In order to simulate DNA storage, biological constraints such as GC-content control and homopolymer avoidance were imposed on DNA encoding using ADF-DNA. This approach ensures that the simulation more closely approximates the conditions encountered in real-world DNA storage applications. However, it should be noted that errors that may arise during the DNA storage process were not artificially introduced, as the quantitative evaluation of the error-tolerance capacity was beyond the scope of the present study.

TABLE I: VIRTUAL INFORMATION CAPACITY COMPARISON

Method	Virtual information density (bits/nt)	Error correction strategy	Application Scenario
Goldman's Method [4]	1.0	Logical Redundancy	General digital data storage
Fountain-DNA [5]	1.9	Logical Redundancy	General digital data encoding
JPEG-DNA [10]	4.2	No Extra Design	General image storage
VVC-DNA [11]	120	Extra Designed Strategies	General images storage with high complexity
DNA-QLC [17]	2.9	Levenshtein Code	General image coding
DJSCC-DNA [18]	8	Hybrid Strategies	General images but limited by the size and content
HybridFlow-DNA	1757	Masking Module	Extreme low-bitrate for high visual quality required

B. Results and Analysis

1) Virtual Information Capacity Comparison: The compression ratio of HybridFlow reached the highest level of 960 on a comparative large dataset (data not shown), and ADF-DNA's Shannon information capacity can achieve 1.83 bits/nt, lower than 2 bits/nt, due to biological constraints. HybridFlow-DNA outperformed six other methods (Goldman's method, Fountain-DNA, JPEG-DNA, VVC-DNA, DNA-QLC, and DJSCC-DNA) (Table I), achieving the highest virtual information capacity of 1757 bits / nt (960×1.83). It is important to note that the compression ratio was calculated by dividing the size of the original image file by that of the compressed file, which includes the indices maps and reference information, but not the visual codebooks.

2) Reconstruction Fidelity Comparison: The comparison was performed between HybridFlow-DNA and two other methods (VVC-DNA and JPEG-DNA) on reconstruction fidelity, as these are the only methods currently available for such a comparison. HybridFlow-DNA maintains a lower LPIPS score, compared to VVC-DNA and JPEG-DNA (Fig. 3), particularly in the region of low image storage capacity region that represents extremely low bitrates. Although HybridFlow-DNA exhibits a certain degree of error-tolerance capacity (Table I), this comparison did not take into account errors that may arise during the DNA storage process.

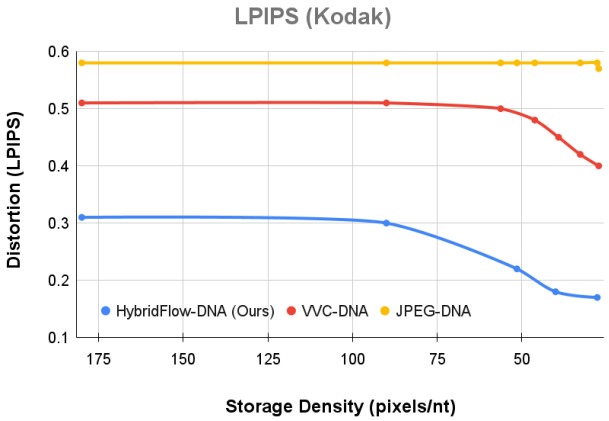

Fig. 3: Reconstruction Fidelity Comparison
The comparison was performed between HybridFlow-DNA and two other methods (VVC-DNA and JPEG-DNA) using the Kodak dataset. A lower LPIPS score indicates a higher fidelity of image reconstruction.

IV. Conclusion

In the present study, we proposed a novel image compression framework for DNA Storage, named HybridFlow-DNA. HybridFlow-DNA is established through the integration of VQGAN and MLIC with the ADF-DNA scheme. This integration takes advantage of the strengths of VQGAN in the high-fidelity reconstruction of images, the hierarchical feature extraction capabilities of MLIC, and the robustness of ADF-DNA. HybridFlow-DNA is designed to be flexible and adaptable, allowing for the integration of other DNA encoding modules. This flexibility is crucial for advancing information processing techniques for DNA storage, as it enables researchers to explore and incorporate novel methods that may further enhance the efficiency and reliability of DNA storage systems.

References

[1] D. Reinsel, J. Gantz, and J. Rydning, "The Digitization of the World from Edge to Core," *IDC WHITE PAPER*, vol. 2018, no. US44413318, 2018.

[2] G. M. Church, Y. Gao, and S. Kosuri, "Next-Generation Digital Information Storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.

[3] V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church, and W. L. Hughes, "Nucleic Acid Memory," *Nature Materials*, vol. 15, no. 4, pp. 366–370, 2016.

[4] 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.

[5] Y. Erlich and D. Zielinski, "DNA Fountain Enables a Robust and Efficient Storage Architecture," *Science*, vol. 355, no. 6328, pp. 950–954, 2017.

[6] C. Pan, S. H. T. Yazdi, S. K. Tabatabaei, A. G. Hernandez, C. Schroeder, and O. Milenkovic, "Image Processing in DNA," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8831–8835.

[7] B. Li, L. Ou, and D. Du, "IMG-DNA: Approximate DNA Storage for Images," in *Proceedings of the 14th ACM International Conference on Systems and Storage*, 2021, pp. 1–9.

[8] G. K. Wallace, "The JPEG Still Picture Compression Standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[9] B. Bross, J. Chen, S. Liu, J.-R. Ohm, B. Segall, T. Suzuki, Y.-K. Wang, R. Weerakkody, M. Winken, and Y. Yamamoto, "Overview of the Versatile Video Coding (VVC) Standard and Its Applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[10] M. Dimopoulou, E. G. San Antonio, and M. Antonini, "A JPEG-Based Image Coding Solution for Data Storage on DNA," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 786–790.

[11] 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)*. IEEE, 2023, pp. 1–5.
[12] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational Image Compression with a Scale Hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.
[13] P. Esser, R. Rombach, A. Blattmann, and B. Ommer, "Taming Transformers for High-Resolution Image Synthesis," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12 873–12 883, 2021.
[14] S. Zou, N. Wang, H. Xu, M. Lin, Z. Zhan, and W. Wang, "Multi-level Image Coding with Visual Codebook," *arXiv preprint arXiv:2211.06393*, 2022.
[15] L. Lu, Y. Xie, W. Jiang, W. Wang, X. Lin, and Y. Wang, "HybridFlow: Infusing Continuity into Masked Codebook for Extreme Low-Bitrate Image Compression," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 3010–3018.
[16] S. Strebel, N. Monnier, D. Lazzarotto, M. Testolina, and T. Ebrahimi, "Towards Learning-Based Image Compression for Storage on DNA Support," in *Applications of Digital Image Processing XLVI*, vol. 12674. SPIE, 2023, pp. 230–243.
[17] Y. Zheng, B. Cao, X. Zhang, S. Cui, B. Wang, and Q. Zhang, "DNA-QLC: An Efficient and Reliable Image Encoding Scheme for DNA Storage," *BMC GENOMICS*, vol. 25, no. 1, p. 266, 2024.
[18] W. Wu, L. Xiang, Q. Liu, and K. Yang, "Deep Joint Source-Channel Coding for DNA Image Storage: A Novel Approach with Enhanced Error Resilience and Biological Constraint Optimization," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2023.
[19] X. Pic and M. Antonini, "Image Storage on Synthetic DNA Using Autoencoders," *arXiv preprint arXiv:2203.09981*, 2022.
[20] C. Ruan, L. Yang, R. Han, S. Gao, H. Wu, and N. Ling, "Robust DNA Image Storage Decoding with Residual CNN," in *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2024, pp. 1–5.
[21] Y. Pei, Y. Liu, N. Ling, Y. Ren, and L. Liu, "An end-to-end deep generative network for low bitrate image coding," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2023, pp. 1–5.
[22] Y. Pei, Y. Liu, and N. Ling, "MobileViT-GAN: A Generative Model for Low Bitrate Image Coding," in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2023, pp. 1–5.