D-CPCE: Collision-free DNA Payload Encoding with Uncompromised Logical Capacity

Xingyu Wang[1], Yunfei Gu[1,2,3], Chentao Wu[1,2,3,*], Jiabin Wang[4,*], Fei Wang[5], Jie Li[1,2,3], Guangtao Xue[1,2,3], Minyi Guo[1], Xuefei Chen[6], Ziwei Zhao[7], Zhijie Huang[8], Chunhai Fan[5,9]

[1]*School of Computer Science, Shanghai Jiao Tong University, Shanghai, China*

[2]*State Key Laboratory of Digital Finance (In Preparation), Shanghai Jiao Tong University, Shanghai, China*

[3]*Shanghai Key Lab. of Web3 Trusted Data Circulation and Governance, Shanghai Jiao Tong University, Shanghai, China*

[4]*Jiaxing Key Laboratory of Biosemiconductors, Xiangfu Laboratory (A), Jiashan, China*

[5]*State Key Laboratory of Synergistic Chem-Bio Synthesis, School of Chemistry and Chemical Engineering, New Cornerstone Science Laboratory, Frontiers Science Center for Transformative Molecules, Zhang Jiang Institute for Advanced Study and National Center for Translational Medicine, Shanghai Jiao Tong University, Shanghai, China*

[6]*Shanghai Computer Society, Shanghai, China*

[7] *Shanghai Artificial Intelligence Research Institute Co., Ltd.,, Shanghai, China*

[8]*School of Computer Science, Northwestern Polytechnical University, Xi'an, China*

[9]*Institute of Molecular Medicine, Shanghai Key Laboratory for Nucleic Acids Chemistry and Nanomedicine, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China*

[*]Corresponding Authors: {wuct, wjb0221}@sjtu.edu.cn

***Abstract*—Random access in DNA data storage relies on spatial multiplexing, utilizing finite libraries of orthogonal primers to biochemically address data pools. However, existing encoding architectures are fundamentally "primer-blind," frequently synthesizing payload sequences that cross-hybridize with the addressing library. When processing gigabyte-scale datasets, these primer-payload collisions scale catastrophically, invalidating up to 100% of the primer library and causing severe logical capacity collapse. To realize true mass-scale DNA storage, we propose the DNA Collision-Free Payload Constraint Encoding (D-CPCE). D-CPCE fundamentally inverts the conventional encoding paradigm by shifting the global constraint-resolution burden entirely to the payload subspace. By orchestrating adaptive Linear Feedback Shift Register (LFSR) scrambling and sequence escalation protocols, and base-3 rotational ciphers, D-CPCE forces dynamic payloads to cleanly route around the static primer library. Comprehensive evaluations across diverse, high-entropy datasets demonstrate that D-CPCE is the first architecture to achieve a 0% primer collision rate on the evaluated workloads against the evaluated primer library at the evaluated configuration. By perfectly preserving 100% of the logical primer capacity while maintaining a competitive overall encode rate of 1.18 bits/NT and a payload encode rate of 1.53 bits/NT, D-CPCE provides a vital architectural foundation for highly scalable, mass-storage DNA archives.**

***Index Terms*—DNA Storage, Next-Generation Storage, DNA Encoding Scheme, Primer-Payload Collision**

I. Introduction

Synthetic DNA has emerged as a disruptive medium for archival data storage due to its unparalleled volumetric information density and extreme physical durability. To transition DNA from a sequential cold-storage archive into a functional, mass-scale data system, random access is essential. This capability is achieved through spatial multiplexing, where unique, orthogonal DNA sequences (primers) are appended to encoded payloads to act as biochemical file addresses for Polymerase Chain Reaction (PCR) retrieval. This biological mechanism functions analogously to Content-Addressable Storage (CAS) in traditional IT systems, where the global primer library serves as the unique key space used to route and retrieve specific data blocks. However, as further detailed in Section II, designing this massive library of mutually orthogonal primers is severely restricted by rigid thermodynamic constraints, including narrow melting temperatures (T_m) and strict Hamming distance thresholds [18]. Consequently, a validated global primer library is a highly precious, finite resource that dictates the absolute logical capacity of the entire storage system. A critical vulnerability arises during the encoding phase: if a synthesized payload accidentally generates a subsequence that cross-hybridizes with *any* primer in the global library, a primer-payload collision occurs, causing severe non-specific binding and data dropout during retrieval.

Existing DNA storage architectures universally fail to resolve this scaling bottleneck. The conventional paradigm is inherently "primer-blind," optimizing payloads strictly for localized constraints while ignoring global collisions. Specifically, these local optimizations involve avoiding homopolymer runs and balancing GC-content (the ratio of G and C bases) to maintain thermodynamic stability during retrieval [2], [17]. When collisions inevitably occur, current systems handle the failure by permanently discarding the affected primer. While this capacity loss is negligible for small experimental archives, it is catastrophic at scale. Our evaluations reveal that when processing large, high-entropy datasets, even state-of-the-art codecs—including those designed with heuristic collision-awareness—suffer massive capacity collapse, burning between 58% and 100% of the usable primer library.

To fully exploit the high-density benefits of DNA storage without artificially capping its logical capacity, we must funda-

mentally invert the encoding paradigm: rather than discarding precious, inflexible primers to accommodate static payloads, the encoding architecture must force the payload to dynamically adapt and route around the static primer library.

In this paper, we present the DNA Collision-Free Payload Constraint Encoding (D-CPCE), an architecture that actively unifies localized biochemical constraint resolution with global addressing preservation. By orchestrating an adaptive Linear Feedback Shift Register (LFSR) scrambling mask, dynamic sequence escalation, and base-3 rotational ciphers, D-CPCE perfectly resolves the collision crisis.

Our specific contributions are as follows:

- We expose the critical scaling limitations of existing DNA codecs, demonstrating experimentally that heuristic, payload-first encoding inevitably causes catastrophic logical capacity loss in mass-storage environments.
- We propose an encoding architecture that shifts the constraint-resolution burden entirely to the payload subspace via pseudo-random masking and bit-shift escalation.
- We are the first to demonstrate an encoding architecture that fully eliminates primer-payload collisions. D-CPCE preserves 100% of the logical primer library, successfully maintaining absolute capacity at scale.
- We demonstrate that absolute biochemical compliance does not require abandoning data density. D-CPCE achieves 100% logical capacity retention while maintaining a moderate and competitive overall encode rate of roughly 1.18 bits/NT.

II. Background and Motivation

A. Scaling Bottleneck

To fully harness the unprecedented volumetric density of DNA data storage, systems must support random access—the ability to selectively retrieve specific files from a vast, pooled synthetic DNA archive without sequencing the entire data lake. This is achieved through Polymerase Chain Reaction (PCR) spatial multiplexing, where unique orthogonal primer sequences are appended to the ends of encoded DNA payloads to act as file-specific biochemical addresses.

Designing these addressing primers is an extraordinarily rigid biochemical challenge. Unlike payload sequences, orthogonal primer libraries must adhere to strict thermodynamic constraints, including narrow melting temperature (T_m) bounds, minimal secondary structures (hairpins), mutually high Hamming distances, balanced GC content and avoidance of long homopolymers to prevent cross-hybridization. Due to these severe limitations, a pool of $30,000$ mutually orthogonal primers represents a highly precious, finite system resource that dictates the absolute logical capacity of the storage archive.

B. Primer-Payload Collision

The conventional DNA encoding paradigm operates sequentially: binary data is encoded into a DNA payload, and a pre-designed primer is subsequently attached. However, if the generated payload accidentally contains a subsequence that is identical or complementary to *any* primer in the global library, a primer-payload collision occurs. During PCR retrieval, these collisions cause non-specific binding, resulting in massive data dropout and readout failure.

To prevent retrieval failure in current architectures, if a payload collides with a primer, that primer is permanently invalidated and discarded. If a DNA archive only stores a handful of files requiring a small primer library, this capacity loss is negligible. However, as we scale toward mass-storage regimes requiring tens of thousands of primers, the statistical probability of a gigabyte-scale payload colliding with the primer library approaches 100%. Under the conventional paradigm, attempting to store massive datasets results in the catastrophic degradation of the system's logical capacity, forcing the system to throw away thousands of hard-to-design primers.

To guarantee 100% logical capacity and make mass-scale DNA storage viable, we must fundamentally invert this paradigm. Instead of throwing away precious, inflexible primers when collisions occur, we must shift the constraint-resolution focus entirely to the *payload side*, dynamically altering the payload sequences to route around the static primer library.

C. Limitations of Existing Works

Recent advancements in DNA data storage have established robust architectures for random access [19], [20], composite-letter density [21], channel characterization [22], and constraint-integrated coding with forward error correction [23], [24]. Despite these advances, the vast majority of existing DNA storage codecs (e.g., DNA Fountain [5], Grass [4], Blawat [3]) are completely "primer-blind." They focus strictly on localized constraints like GC-content and homopolymers, completely ignoring global primer collisions. Consequently, when deployed against large addressing libraries, they suffer near-total capacity collapse.

To the best of our knowledge, Collision Aware Code (CAC) [7] is the only existing literature that explicitly identifies and attempts to mitigate primer-payload collisions. However, CAC relies on localized, greedy triplet-mapping heuristics. To quantify the effectiveness of current literature, we benchmarked 8 prominent encoding methods on a standard image (`mona_lisa.jpg`) against a pre-generated library of 30,000 primers.

As shown in Table I, there is a highly non-linear relationship between sequence collisions (the percentage of encoded payloads containing at least one forbidden primer substring) and valid capacity (the percentage of the 30,000-primer library that survives). Because a single poorly-constrained payload sequence can cross-hybridize with multiple primers, and vice versa, a massive 98.4% sequence collision rate may only invalidate roughly 25% of the primer library. Nonetheless, our results confirm that CAC fails to fundamentally solve the collision problem at scale. While CAC maintains roughly

TABLE I
PRIMER CAPACITY LOSS ON MONA_LISA.JPG (30,000 PRIMER LIBRARY)

Method	Sequence Collisions (%)	Usable Primers	Valid Capacity (%)
Church [1]	98.590	14215	47.383
Goldman [2]	93.339	26333	87.777
Blawat [3]	98.114	19331	64.437
Grass [4]	96.079	18877	62.923
DNA Fountain [5]	96.873	18094	60.313
YYC [6]	97.072	18429	61.430
CAC [7]	98.411	22729	75.763

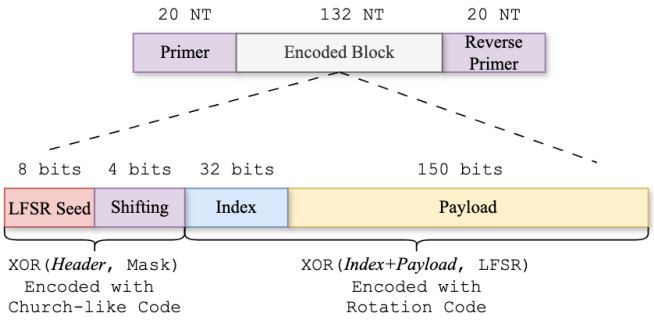

Fig. 1. Oligo Structure of D-CPCE with encoded block breakdown.

75.7% valid capacity on this small image payload, our mass-scale evaluations (Section IV) reveal that its valid capacity collapses to just 42.3% on gigabyte-scale datasets. Currently, no literature fully maintains 100% logical capacity for mass-scale archiving.

D. Motivation and Design Goals

In their foundational analysis, the authors of CAC [7] observed that rotating ciphers (such as the base-3 implementation by Goldman *et al.*) naturally exhibit the lowest baseline primer-payload collision rates among existing architectures due to their inherent homopolymer suppression and continuous state transitions. We validate this observation and strategically select the rotation code architecture as the foundational layer for our proposed codec.

Our objective is to design an encoding method that achieves two simultaneous goals:

1) **Maximum Logical Capacity:** Achieve strict 100% collision avoidance to perfectly preserve the 30,000-primer library, enabling true mass-scale random access.
2) **Maximum Encode Rate:** While CAC's collision-aware heuristic restricts its payload rate to a highly inefficient 1.0 bit/nucleotide (NT), the theoretical information limit for strict homopolymer-free encoding is $\log_2(3) \approx 1.58$ bits/NT. This limit arises because completely avoiding homopolymers restricts each subsequent sequence position to only 3 valid base choices out of the 4 available nucleotides. Our goal is to push the overall physical encode rate as close to this theoretical ceiling as possible while strictly enforcing all biochemical constraints.

III. SYSTEM DESIGN: THE D-CPCE ARCHITECTURE

Unlike traditional computer storage architectures where data blocks reside at fixed, linearly addressable physical memory locations, DNA data storage relies on biochemical spatial multiplexing. Data retrieval is executed via Polymerase Chain Reaction (PCR), where specific primer pairs act as the retrieval addresses for entire data pools (e.g., individual files). Because a single spatial primer pair typically corresponds to thousands or millions of unique oligonucleotides (short, synthetic DNA strands) , the encoded sequences must be meticulously structured to ensure strict global ordering while actively evading cross-hybridization with the primer library itself. Designing a robust primer library is biochemically complex, constrained by thermodynamic stability and the prevention of secondary structures; for a comprehensive discussion on these challenges, we direct readers to Organick et al. [18]. We construct our primer pool following this established methodology. Furthermore, to actively evade cross-hybridization, we enforce a strict 12-mer exact match constraint. Unlike the relaxed thresholds (e.g., 14-mer or 16-mer), this 12-mer criterion serves as a rigorous standard that provides a stricter bound against primer-payload collisions. Figure 1 illustrates the oligo structure of D-CPCE.

A. Payload Scrambling and Logical Block Assembly

The encoding pipeline begins at the raw binary level. To disrupt naturally occurring low-entropy regions or structural repetitions in the source data (such as extended sequences of null bytes), the raw payload bits are XOR-scrambled using a pseudo-random sequence generated by a Linear Feedback Shift Register (LFSR). This LFSR is initialized with a variable k-bit seed, establishing our primary mathematical search space for constraint resolution.

Once the payload is scrambled, we must establish internal sequence ordering. Because target molecules float freely in a biochemical solution, a single PCR primer pair extracts an unordered "soup" of millions of oligos; thus, the system relies on an embedded binary index to reconstruct the file post-sequencing. These index bits are appended to the front of the scrambled payload bits. We define this combined continuous binary sequence (Index + Payload) as the *logical block*.

To minimize the baseline probability of biochemical collisions before active filtering begins, the entire logical block is translated into a DNA sequence using a base-3 rotation cipher (e.g., the Goldman codec). This rotational translation intrinsically suppresses homopolymers and maintains a high baseline payload density.

B. The Constraint Engine: Retries and Escalation

Once the logical block is translated into DNA, it is passed into the constraint verification engine. Here, the encoded sequence undergoes a unified dual-verification process. First, it is evaluated for strict biochemical compliance—specifically ensuring the global GC content remains safely within the

Algorithm 1 D-CPCE Encoding and Decoding Pseudocode

Require: Index $index$, Payload Bits $payload$, Collision Hash Table C, Max Seed S_{max}, Max Shift K_{max}, Primer Pair P_{pair}
Ensure: State S, Valid oligo O_{valid}, Seed s, Shift k

{— ENCODING PIPELINE —}
$D \leftarrow concatenate(index, payload)$
for $k = 0$ **to** $K_{max} - 1$ **do**
  for $s = 0$ **to** $S_{max} - 1$ **do**
    // Generate deterministic LFSR mask
    $M_s \leftarrow$ LFSR_Generate(s)
    // Apply shift and XOR mask
    $logical\_chunk \leftarrow Encode\_Rotation((D \lll k) \oplus M_s)$
    $header \leftarrow Generate\_Header\_With\_Church(s, k, i)$
    $O_{valid} \leftarrow Pack(P_{pair}, header, logical\_chunk)$
    // Validation
    if $0.40 \leq$ GC_Ratio$(O_{valid}) \leq 0.60$ **then**
      if Check_Collision$(O_{valid}, C) ==$ False **then**
        return SUCCESS, $oligo, s, k$
      end if
    end if
  end for
end for
return FAILURE, $\varnothing, S_{max}, K_{max}$

{— DECODING PIPELINE —}
Require: Oligo $oligo$
Ensure: Decoded Payload $payload$, Decoded Index $index$
// Reconstruct metadata header and logical chunk from oligo
$(header, chunk) \leftarrow Separate(oligo)$
$(s, k) \leftarrow Decode\_Church(header)$
$decoded\_chunk \leftarrow Decode\_Rotation(chunk)$
// Regenerate exact LFSR mask
$M_s \leftarrow$ LFSR_Generate(s)
// Reverse masking and shifting
$decoded\_chunk \leftarrow (decoded\_chunk \oplus M_s) \ggg k$
$(index, payload) \leftarrow Separate(decoded\_chunk)$
return $payload, index$

viable 40%–60% bounds (homopolymer runs are intrinsically prevented by the rotational cipher, requiring no active filtering). Second, the sequence is scanned against the global primer library to detect any forbidden 12-mer primer-payload collisions.

Our architecture utilizes a deterministic retry mechanism to resolve any failure in either verification step. If the sequence violates the GC thresholds *or* triggers a primer collision, the system discards the encoded block, increments the LFSR seed, generates a new pseudo-random XOR mask, and repeats the logical block assembly and rotation encoding.

In extremely rare, highly adversarial entropy scenarios, the system may exhaust all 2^k available LFSR seeds without finding a valid, collision-free sequence. To solve this without expanding the LFSR footprint (which would waste valuable nucleotide space), we introduce a dynamic *escalation* protocol. Upon seed exhaustion, the engine shifts the raw bits of the entire logical block by exactly one bit position. This single-bit shift radically alters the downstream base-3 rotational translation, thereby generating a completely orthogonal set of mathematical states. The LFSR search is then restarted on this shifted block, virtually guaranteeing absolute constraint resolution without halting the encoding pipeline.

C. Robust Header Encoding and Global Assembly

To ensure the decoder can accurately reverse the scrambling and escalation process, the system must permanently store the specific LFSR seed and the escalation shift offset used to successfully encode the block. We pack these parameters into a compact binary *header*, which is placed directly in front of the encoded logical block.

Because the successful decoding of the payload is entirely dependent on the integrity of this header, it requires the strictest possible biochemical stability. Therefore, instead of using the base-3 rotation cipher, the header is independently encoded using a rigid, Church-like substitution scheme yielding 1.0 bit/NT. In this scheme, binary `0` is mapped dynamically to $\{\mathrm{A}, \mathrm{C}\}$, and binary `1` is mapped to $\{\mathrm{T}, \mathrm{G}\}$. This specific mapping algorithm perfectly balances the GC content of the header at exactly 50% and mathematically prevents any homopolymer runs, creating a highly stable biochemical anchor for the oligo's metadata.

Crucially, the dual-verification constraint scan is not limited to the payload alone. The system executes its final global collision and GC check on the *entire assembled sequence* (the Church-encoded Header concatenated with the Rotation-encoded Logical Block). Only when this unified sequence is proven to be strictly GC-compliant and 100% orthogonal to the global primer pool does the pipeline proceed to the final step: appending the specifically assigned spatial primer pair to the 5' and 3' flanks of the validated sequence, rendering the oligonucleotide ready for biological synthesis.

IV. Evaluation

A. Experimental Setup

Datasets. We benchmarked our architecture against nine diverse, adversarial datasets representing realistic production workloads (Table II).

TABLE II
Summary of Evaluated Datasets

Category	Dataset Name	Type	#Files	Size
AI Models	ALBERT [8]	.safetensors	1	45.2 MB
	MobileNet [9]	.bin	1	13.6 MB
	Huawei-Noah [10]	.bin	1	59.8 MB
Images	Caltech-101 [11]	.jpg	9,144	14.0 MB
Videos	Blender [12]	.mkv	1	1.1 GB
	Derf Video [13]	.y4m	1	130.5 MB
Texts & Binaries	Linux FS [14]	.c, .h	2,085	47.0 MB
	Shackleton-s [15]	. (text)	18,687	39.28 MB
	Hetzner [16]	.bin	1	10.0 GB
Total			**29922**	**11.4 GB**

Evaluation Metrics. We assess performance using five primary metrics:

- **Payload Rate (bits/NT):** Logical bits divided strictly by payload nucleotides.

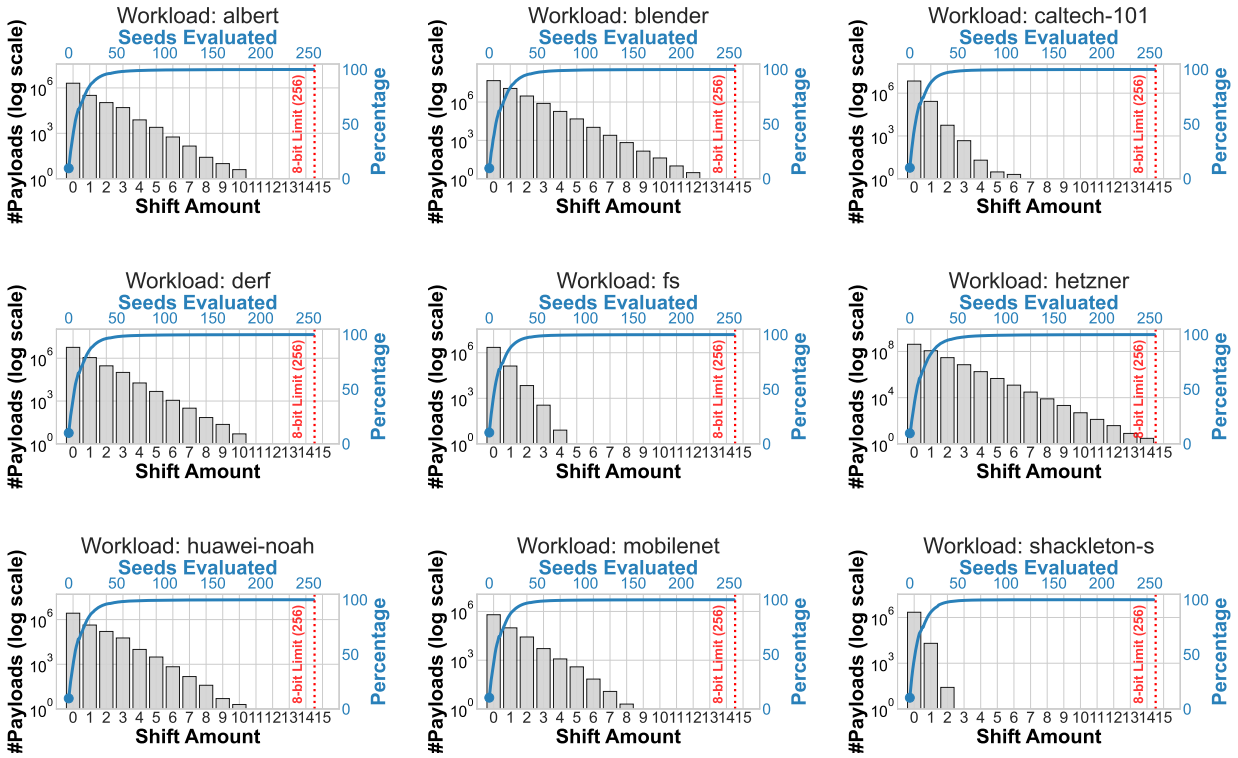

Fig. 2. Constraint resolution and execution breakdown for the $\langle 8, 16, 150 \rangle$ configuration, demonstrating 100% primer retention and strict GC compliance within an 8-bit search space.

- **Overall Rate (bits/NT):** Total physical footprint efficiency, including primer and algorithmic overhead.
- **Sequence Collisions (%):** Percentage of oligos containing ≥ 1 forbidden 12-mer primer match.
- **Usable Primers:** Count of surviving primers from the global 30,000-primer dictionary.
- **Logical Capacity (%):** Percentage of viable primer addresses retained.

System Configuration. Evaluation was executed on a dual-socket 64-core ARM64 server (2.60 GHz) with 64 MB L3 cache, running Ubuntu Linux (Kernel 5.4.0).

B. Constraint Resolution and Density Analysis

Baseline Implementation and Parameters. We re-implemented all comparison codecs in C++. To ensure fairness, we fixed the routing index at 32 bits and calibrated payload lengths to yield ∼100-nucleotide (NT) physical footprints (Church/CAC/YYC: 100 bits; Grass: 128; Goldman: 150; Blawat: 160; Fountain: 200). We enforced strict constraints (max homopolymer 3, GC 40%–60%) where supported. Notably, to prevent infinite stalling on highly entropic data, YYC bypassed GC constraints to achieve its theoretical 2.0 bits/NT rate.

Primer Degradation vs. Sequence Collisions. Table III reveals a nonlinear relationship between sequence collisions and usable primers. While most methods exhibit $> 91\%$ collision rates, surviving capacities vary drastically (0% to 86%). This occurs because a single poorly-constrained sequence can invalidate dozens of orthogonal primers. Therefore, absolute sequence collision percentage is a secondary symptom; primer retention is the fundamental metric dictating random-access capacity.

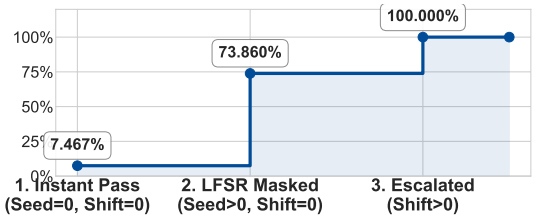

Fig. 3. CDF of instant pass, LFSR masking and shift escalations.

TABLE III
EVALUATION OF ENCODING METHODS: DENSITY AND PRIMER COMPATIBILITY

Method	Usable Primers	Capacity (%)	Collisions (%)	Payload (bits/NT)	Overall (bits/NT)
Church [1]	2276	7.587	98.141	1.000	0.767
Goldman [2]	25863	86.210	93.333	1.574	1.173
Grass [4]	80	0.267	91.068	1.778	1.231
Blawat [3]	4358	14.527	97.498	1.600	1.200
CAC [7]	12690	42.300	98.711	1.000	0.767
DNA Fountain [5]	0	0.000	97.033	1.906	1.487
YYC [6]	0	0.000	97.040	2.000	1.487
D-CPCE (Ours)	**30000**	**100.000**	**0.000**	**1.528**	**1.181**

High-density algorithms like DNA Fountain and YYC suffer

$> 97\%$ collision rates, collapsing usable capacity to 0%. Legacy codes preserve $< 15\%$ capacity, while Goldman retains 86.2%. Notably, the collision-aware CAC codec preserves only 42.3% capacity at scale, confirming that greedy, primer-blind heuristics fail against adversarial entropy. Conversely, D-CPCE explicitly isolates the payload subspace via our $O(1)$ LFSR engine. Though incurring a moderate density penalty (1.181 bits/NT), D-CPCE achieves 0% sequence collisions and preserves 100% logical capacity, proving deterministic avoidance is vital for scaling spatial multiplexing.

C. Engine Dynamics and Resolution Effort

To understand exactly how D-CPCE achieves perfect capacity retention, we analyzed the internal workload of the constraint engine. Figure 3 shows the cumulative distribution of the three resolution stages: (a) *instant pass*, representing oligos that directly satisfy biochemical constraints without any primer-payload collisions; (b) *LFSR masking*, representing oligos that successfully avoid collisions through seed retries; and (c) *shift escalations*, representing adversarial oligos that resort to bit-shifting to eliminate collisions. Across all tested workloads, 7.5% of sequences naturally satisfy constraints (Instant Pass), the LFSR masking layer advances the cumulative curve to 73.8%, and structural shift escalations resolve the remaining adversarial sequences to reach 100%. Therefore, of the 92.5% of sequences that initially collide, 71.7% are efficiently mitigated by the $O(1)$ LFSR masking layer, while the remaining 28.3% are successfully handled by structural shift escalations.

To further dissect this computational effort on a per-workload basis, Figure 2 visualizes the constraint resolution dynamics across all nine individual datasets. Because tracking both seed retries and shift escalations requires observing two distinct mechanisms, the figure employs a dual-axis representation:

- **Shift Distribution (Bar Charts):** Mapped to the bottom X-axis and the left Y-axis (log scale), the gray bars display the raw count of payloads resolved at each shift amount. The overwhelming trend across all workloads is a massive concentration at Shift 0, followed by a steep logarithmic decay.
- **Seed Resolution (Solid Curves):** Mapped to the top X-axis and the right Y-axis (linear scale), the black curves represent the Cumulative Distribution Function (CDF) of the LFSR seeds evaluated. The steep initial vertical ascent of these curves demonstrates that the vast majority of LFSR-mitigated collisions are resolved within the first few dozen seed permutations.

Crucially, the trend across all nine subplots proves the mathematical viability of the D-CPCE architecture. For every single dataset, the solid black CDF curve reaches 100% cumulative resolution well before hitting the $2^8 = 256$ seed boundary (marked by the vertical dashed line). This demonstrates that our compact 8-bit search space, when paired with the 4-bit escalation protocol, is fundamentally sufficient to guarantee 0% collisions without requiring larger, space-wasting metadata headers. The number of retries grows with the increase in file size, with the 10 GB hetzner dataset requiring the maximum number of shift amounts (0-14).

D. Parameter Sensitivity and Payload Scaling

To analyze internal dynamics, we evaluated varying $\langle \text{LFSR\_Bits}, \text{Index\_Bits}, \text{Payload\_Bits} \rangle$ configurations. Figure 2 details the optimal baseline $\langle 8, 16, 150 \rangle$ footprint (~ 100 physical NT payload), demonstrating 100% primer retention.

The Collision Wall at Extended Lengths. Scaling the payload to 300 bits ($\langle 8, 16, 300 \rangle$) exponentially increases the probability of primer matches, driving execution time up by $3.2\times$. Furthermore, the 8-bit LFSR (255 shifts) struggles here: while maintaining 100% collision reduction, 12 out of 330,592,566 sequences fail to converge within the $[0.4, 0.6]$ GC bounds.

Dynamic Escalation and Length Limitations. Escalating to a 10-bit LFSR ($\langle 10, 16, 300 \rangle$) restores strict GC bounds but incurs a $7\times$ slowdown (64.6 hours vs. 9.1 hours for the 10 GB dataset) versus the 8-bit baseline.

1) **Biochemical Length Limits:** The $\langle 8, 16, 150 \rangle$ oligo totals ~ 172 NT. Scaling to 300 bits pushes this to ~ 270 NT. Modern synthesis yields degrade and indel errors spike for oligos exceeding 200 NT [17], [18], rendering the 300-bit payload practically unviable.
2) **Payload Segmentation Reduces Collisions:** Longer sequences inherently possess fewer mathematically valid, collision-free states. Breaking large payloads into 150-bit chunks exponentially expands valid biochemical routing states, achieving 0% collisions via a faster $O(1)$ search space.

While this escalation reduces system throughput from 0.31 MB/s (baseline) to 0.04 MB/s, even this worst-case parallel throughput remains $\sim 40\times$ faster than state-of-the-art biological synthesis ($\sim$0.001 MB/s). Trading compute time for spatial efficiency thus represents a highly practical overhead to perfectly preserve valuable primer capacity.

E. Limitations and Future Work

D-CPCE assumes consensus algorithms are applied across the redundancy of PCR-amplified reads to resolve baseline noise prior to decoding. Single-base substitutions are strictly confined to local symbol boundaries because the rotation cipher evaluates only adjacent base transitions, and the independent, point-to-point LFSR XOR mask prevents any descrambling avalanche effect. However, uncorrected header errors or payload insertions/deletions (indels) can cause block erasures. Resolving indels requires specialized synchronization algorithms rather than traditional Forward Error Correction (FEC), and thus remains a challenge in DNA storage. Consequently, we leave the integration of outer-layer checksums and advanced erasure-recovery mechanisms to future work.

V. Conclusion

To transition DNA from a sequential archive into a scalable, random-access mass-storage system, the field must resolve the

primer-payload collision bottleneck. Our evaluations exposed a critical flaw in existing DNA codecs: payload-first, primer-blind encoding inevitably results in catastrophic addressing collisions, burning precious spatial multiplexing capacity as dataset sizes scale.

In this paper, we introduced D-CPCE, an encoding scheme that fundamentally solves this crisis by forcing the payload to dynamically adapt to the static addressing library. Through a unified constraint engine leveraging LFSR pseudo-random masking, base-3 rotational ciphering, and a robust 4-bit structural escalation protocol, D-CPCE systematically eliminates all biochemical collisions. Our evaluation proves that D-CPCE is the first DNA encoding architecture to achieve 0% primer-payload collisions, preserving 100% of logical capacity across adversarial, gigabyte-scale datasets. Furthermore, D-CPCE demonstrates that absolute biochemical compliance can be achieved without sacrificing practical data density, securing a competitive overall encode rate of 1.18 bits/NT and a payload rate of 1.528 bits/NT. Ultimately, this architecture guarantees collision-free spatial routing at scale, paving the way for the realization of ultra-dense, exabyte-scale DNA data centers.

VI. Acknowlegement

We sincerely thank our shepherd, Swaminathan Sundararaman, and the anonymous MSST reviewers for their constructive feedback and insightful comments. This work is partially sponsored by the National Key R&D Program of China (Grant No.2023YFB4502900), the National Natural Science Foundation of China (No.U25B2022, 62272394), the Explorers Program of Shanghai (Basic Research Funding No.25TS1410900), and the Shandong Provincial Natural Science Foundation (Project No.ZR2023LZH020).

References

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

[2] N. Goldman *et al.*, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, 2013.

[3] M. Blawat *et al.*, "Forward error correction for DNA data storage," *Procedia Computer Science*, vol. 80, pp. 1011–1022, 2016.

[4] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angewandte Chemie International Edition*, vol. 54, no. 8, pp. 2552–2555, 2015.

[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] P. Ping *et al.*, "Towards practical and robust DNA-based data archiving using the yin-yang codec system," *Nature Computational Science*, vol. 2, pp. 234–242, 2022.

[7] Y. Wei, B. Li, and D. H. C. Du, "An encoding scheme to enlarge practical DNA storage capacity by reducing primer-payload collisions," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*, pp. 317–331, 2024.

[8] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," *arXiv preprint arXiv:1909.11942*, 2019.

[9] A. G. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017.

[10] Huawei Technologies, "Huawei Noah's Ark Lab Open Source Datasets and Models," [Online]. Available: https://github.com/huawei-noah.

[11] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *Proc. 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 178–186, 2004.

[12] Blender Foundation, "Sintel 1080p Open Movie," [Online]. Available: https://download.blender.org/demo/movies/Sintel.2010.1080p.mkv.

[13] Xiph.org, "Derf's Video Test Media Collection: mad900_cif.y4m," [Online]. Available: https://media.xiph.org/video/derf/y4m/mad900_cif.y4m.

[14] L. Torvalds et al., "The Linux Kernel Source Tree, Version 6.19.11," [Online]. Available: https://www.kernel.org/.

[15] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research," in *European Conference on Machine Learning (ECML)*, Springer, pp. 217–226, 2004.

[16] Hetzner Online GmbH, "10GB Speed Test Binary File," [Online]. Available: https://ash-speed.hetzner.com/10GB.bin.

[17] S. Kosuri and G. M. Church, "Large-scale de novo DNA synthesis: technologies and applications," *Nature Methods*, vol. 11, no. 5, pp. 499–507, 2014.

[18] L. Organick et al., "Random access in large-scale DNA data storage," *Nature Biotechnology*, vol. 36, no. 3, pp. 242–248, 2018.

[19] J. Bornholt *et al.*, "A DNA-based archival storage system," in *Proceedings of the 21st International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pp. 637–649, 2016.

[20] S. M. H. Tabatabaei Yazdi *et al.*, "Portable and error-free DNA-based data storage," *Scientific Reports*, vol. 7, p. 5011, 2017.

[21] L. Anavy *et al.*, "Data storage in DNA with fewer synthesis cycles using composite letters," *Nature Biotechnology*, vol. 37, pp. 1229–1236, 2019.

[22] R. Heckel *et al.*, "Characterization of the DNA data storage channel," *Scientific Reports*, vol. 9, p. 9663, 2019.

[23] W. H. Press *et al.*, "HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints," *Proceedings of the National Academy of Sciences*, vol. 117, no. 42, pp. 25966–25972, 2020.

[24] M. Welzel *et al.*, "DNA-Aeon provides flexible arithmetic coding for constraint satisfaction and error correction in DNA storage," *Nature Communications*, vol. 14, p. 628, 2023.