Hybrid Error-Resilient DNA Data Storage: Bias-Aware Coding, Enzymatic Repair Simulation, and Microcapsule-Based Random Access

by

Md. Mehedi Hasan
ID: CSE2201025093

Shamima Khatun
ID: CSE2201025158

Morsheda Akter
ID: CSE2201025030

Md. Nabin Owahid
ID: CSE2201025116

Supervised by
Mohammad Naderuzzaman

Submitted in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SONARGAON UNIVERSITY (SU)
January 2026

APPROVAL

The thesis titled "**Hybrid Error-Resilient DNA Data Storage: Bias-Aware Coding, Enzymatic Repair Simulation, and Microcapsule-Based Random Access**" submitted by Md. Mehedi Hasan (CSE2201025093), Shamima Khatun (CSE2201025158), Morsheda Akter (CSE2201025030) and Md. Nabin Owahid (CSE2201025116) to the Department of Computer Science and Engineering, Sonargaon University (SU), has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents.

Board of Examiners

Mohammad Naderuzzaman
Associate Professor,
Department of Computer Science and Engineering
Sonargaon University (SU)

Supervisor

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 1

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 2

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 3

DECLARATION

We, hereby, declare that the work presented in this report is the outcome of the investigation performed by us under the supervision of **Mohammad Naderuzzaman, Associate Professor** Department of Computer Science and Engineering, Sonargaon University, Dhaka, Bangladesh. We reaffirm that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

(Mohammad Naderuzzaman)
Supervisor

Signature

Md. Mehedi Hasan
ID: CSE2201025093

Shamima Khatun
ID: CSE2201025158

Morsheda Akter
ID: CSE2201025030

Md Nabin Owahid
ID: CSE2201025116

ABSTRACT

DNA has emerged as a promising medium for long-term archival data storage due to its exceptional information density, durability, and passive energy requirements. However, practical deployment remains constrained by three fundamental challenges: stochastic synthesis bias, chemical degradation over time, and limited scalability of random access. This thesis proposes a **Hybrid Error-Resilient DNA Storage Framework** that integrates computational, biochemical, and architectural solutions to address these limitations holistically.

First, a **bias-aware adaptive coding scheme** is introduced, which models synthesis and PCR dropout as probabilistic processes and dynamically assigns logical redundancy based on predicted sequence fragility. Unlike conventional uniform redundancy approaches, the proposed method reduces sequencing coverage requirements while preserving decoding reliability. Second, a **Markov-chain-based enzymatic repair model** is developed to simulate long-term molecular decay and restoration using a multi-enzyme repair cocktail consisting of APE1, Bst polymerase, and Taq ligase. The model demonstrates a significant extension of data recoverability horizons under accelerated aging conditions. Third, the thesis evaluates **thermoresponsive microcapsule-based random access**, enabling repeated, low-bias retrieval through thermoconfined PCR while preventing destructive consumption of the archival pool.

Comprehensive simulations indicate a reduction of sequencing coverage by over 70% and an effective extension of archive longevity by nearly threefold compared to conventional DNA storage pipelines. Additionally, a sustainability analysis highlights substantial reductions in energy consumption, carbon emissions, and electronic waste relative to magnetic tape and hard disk-based archival systems. The results establish DNA storage as a viable candidate for future ultra-long-term, sustainable digital preservation.

ACKNOWLEDGMENT

At the very beginning, we would like to express my deepest gratitude to the Almighty Allah for giving us the ability and the strength to finish the task successfully within the schedule time.

We are auspicious that we had the kind association as well as supervision of **Mohammad Naderuzzaman**, Associate Professor, Department of Computer Science and Engineering, Sonargaon University whose hearted and valuable support with best concern and direction acted as necessary recourse to carry out our thesis.

We would like to express our sincere gratitude to **Prof. Bulbul Ahamed,** Head, Department of Computer Science and Engineering, for his valuable guidance, encouragement, and kind support throughout our academic journey.

We would like to convey our special gratitude to **Brig. Gen. (Retd) Prof. Habibur Rahman Kamal,** Dean, Faculty of Science & Engineering for his kind concern and precious suggestions

We are also thankful to all our teachers during our whole education, for exposing us to the beauty of learning.

Finally, our deepest gratitude and love to my parents for their support, encouragement, and endless love.

LIST OF ABBREVIATIONS

ACM	Association for Computing Machinery
APE1	Apurinic/Apyrimidinic Endonuclease 1
ATGC	Adenine–Thymine–Guanine–Cytosine
BER	Base Excision Repair
Bst	Bacillus stearothermophilus (DNA Polymerase)
CDF	Cumulative Distribution Function
DNA	Deoxyribonucleic Acid
DRRC	Dual-Rule Rotational Coding
ECC	Error Correction Code
GC	Guanine–Cytosine
GC-content	Guanine–Cytosine Content
HEDGES	Hash Encoded, Decoded by Greedy Exhaustive Search
HDD	Hard Disk Drive
IEEE	Institute of Electrical and Electronics Engineers
Indel	Insertion–Deletion Error
ISIT	International Symposium on Information Theory
PCR	Polymerase Chain Reaction
PPF	Production Possibility Frontier (only if referenced conceptually*)*
RS	Reed–Solomon
SSD	Solid State Drive
Taq	Thermus aquaticus (DNA Ligase/Polymerase)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

The exponential growth of digital information has become one of the defining characteristics of the modern era. Advances in computation, communication technologies, and data-driven applications have resulted in an unprecedented surge in data generation across scientific research, finance, healthcare, social media, and governmental archives. It is estimated that global data production exceeds hundreds of zettabytes annually, with a substantial portion classified as cold data, meaning information that is infrequently accessed but must be preserved reliably for decades or even centuries [1][2][3].

Traditional digital storage infrastructures, primarily based on magnetic and solid-state technologies, were not designed to support such long-term archival demands. Magnetic tapes, hard disk drives (HDDs), and solid-state drives (SSDs) suffer from intrinsic physical degradation, technological obsolescence, and increasing energy requirements [3][4]. Even under controlled environmental conditions, magnetic tape systems typically require data migration every 7–10 years, while HDDs and SSDs have operational lifespans ranging from 3 to 10 years. This perpetual migration cycle introduces significant economic costs, operational complexity, and environmental consequences in the form of energy consumption and electronic waste [3][5][6].

In contrast to these limitations, deoxyribonucleic acid (DNA) presents a radically different paradigm for information storage. DNA is the fundamental biological molecule responsible for storing genetic information in all known living organisms [7]. Over billions of years of evolution, nature has demonstrated DNA's extraordinary stability, density, and error-tolerant information encoding mechanisms [7][8]. Under appropriate storage conditions, DNA can remain intact and readable for thousands of years, as evidenced by successful sequencing of ancient biological samples [7][9].

The convergence of biotechnology and information theory has led researchers to explore DNA as a medium for digital data storage. By encoding binary information into sequences of nucleotides—adenine (A), cytosine (C), guanine (G), and thymine (T)—digital data can be synthesized into physical DNA molecules, stored passively, and later retrieved through sequencing technologies [8][10]. This concept offers the potential to store massive amounts of information in an extremely small physical footprint, with near-zero energy consumption during storage [6][9][10].

Despite its promise, DNA data storage remains largely experimental. Significant technical barriers prevent its immediate adoption as a practical archival solution. These challenges arise from the stochastic nature of biochemical processes, chemical degradation mechanisms, and the difficulty of selectively retrieving data from large molecular pools. Addressing these challenges requires a

multidisciplinary approach that integrates computer science, molecular biology, chemistry, and materials engineering[6][10].

1.1.1 Importance of Long-Term Digital Preservation

Long-term digital preservation is essential for maintaining scientific records, cultural heritage, legal documents, and historical archives. National libraries, medical institutions, climate research organizations, and government agencies increasingly rely on digital formats to store information that must remain accessible for future generations [1][3]. The failure to preserve such data reliably can result in irreversible loss of knowledge, legal disputes, and setbacks in scientific progress [3][4].

As data volumes grow, preservation challenges extend beyond storage capacity to include reliability, accessibility, and sustainability. Archival systems must ensure that data remains readable despite technological evolution, environmental changes, and hardware degradation. These requirements place stringent demands on storage technologies, motivating the exploration of fundamentally new storage paradigms such as DNA-based systems [7][10].

1.2 Limitations of Conventional Storage Media

Conventional digital storage systems are fundamentally constrained by their reliance on electronic and magnetic components. Magnetic tapes, although widely used for archival purposes due to their relatively low cost per terabyte, suffer from slow access times, mechanical wear, and sensitivity to environmental conditions such as temperature and humidity. Furthermore, tape-based systems require specialized hardware, which may become obsolete, making long-term readability uncertain [3][4].

Hard disk drives rely on spinning platters and magnetic domains to represent data. Mechanical components introduce failure points, and the gradual loss of magnetic signal strength leads to data corruption over time [4][6]. Solid-state drives eliminate mechanical motion but introduce new limitations, such as charge leakage in flash memory cells and limited write endurance. These factors collectively make long-term, maintenance-free storage infeasible with existing technologies [4][5].

From a sustainability perspective, data centers consume a rapidly increasing share of global electricity production. Continuous cooling, power delivery, and redundancy mechanisms significantly contribute to carbon emissions. Moreover, the frequent replacement of storage hardware generates substantial electronic waste, containing toxic and non-recyclable materials. As global data storage needs continue to rise, these environmental impacts are expected to intensify unless alternative storage paradigms are developed [1][5][6].

Another critical limitation of traditional storage media is technological obsolescence. Even if physical degradation could be mitigated, older storage formats often become unreadable due to the disappearance of compatible hardware and software. Historical examples such as floppy disks and

early optical media highlight the vulnerability of digital information to rapid technological evolution [3][4].

1.2.1 Economic and Operational Challenges of Conventional Storage

In addition to technical limitations, conventional storage systems impose significant economic and operational burdens. Periodic data migration requires specialized labor, infrastructure downtime, and extensive verification processes to ensure data integrity. For large-scale archives, these recurring costs can exceed the initial hardware investment over the system's lifetime. [3][4]

Furthermore, the operational complexity of managing geographically distributed data centers introduces additional risks related to cybersecurity, physical security, and disaster recovery. These challenges highlight the inefficiency of relying solely on electronic storage technologies for long-term archival purposes and reinforce the need for storage media that minimize maintenance and operational intervention [5][6].

1.3 DNA as a Digital Storage Medium

DNA offers several properties that make it uniquely suitable for ultra-long-term archival storage. First, DNA exhibits an unparalleled information density. Theoretical analyses suggest that a single gram of DNA can encode hundreds of petabytes of digital data. [8] This density far exceeds that of any existing electronic storage medium, enabling the condensation of massive data archives into minimal physical space [7][10].

Second, DNA storage is inherently passive. Once synthesized and stored in a dry, encapsulated environment, DNA requires no electrical power for maintenance. This contrasts sharply with data centers that must operate continuously to preserve electronic data. As a result, DNA-based archives have the potential to dramatically reduce long-term energy consumption and associated carbon emissions [6][9][10].

Third, DNA is a universal and future-proof information carrier. The biological importance of DNA ensures that technologies for reading and writing DNA will remain relevant for the foreseeable future. Unlike proprietary electronic formats, DNA sequencing relies on standardized biochemical principles, reducing the risk of technological obsolescence [3][10].

Digital information is encoded into DNA by mapping binary data to nucleotide sequences using carefully designed coding schemes. These schemes must avoid problematic patterns such as long homopolymer runs and extreme GC content, which can introduce errors during synthesis and sequencing. Once encoded, the DNA sequences are chemically synthesized and stored as a molecular pool [8][9][10].

Retrieval involves selectively amplifying target sequences using polymerase chain reaction (PCR) and reading the nucleotide sequences via high-throughput sequencing. Computational decoding algorithms then reconstruct the original digital data from the sequenced reads. While conceptually

straightforward, each stage introduces potential sources of error that must be addressed to ensure reliable data recovery. [8][10]

1.3.1 Challenges in DNA Data Storage

Despite its advantages, DNA data storage introduces unique challenges that distinguish it from conventional electronic storage systems. The biochemical processes involved in DNA synthesis, amplification, and sequencing are inherently stochastic, leading to non-uniform representation of encoded data [9][10]. This variability can result in sequence dropout, amplification bias, and uneven sequencing coverage, which complicate reliable data recovery [8][10].

Additionally, DNA molecules are subject to chemical degradation over time due to hydrolysis, oxidation, and environmental exposure. Even minor molecular damage, such as single-strand breaks, can prevent successful amplification and decoding. These factors necessitate the development of error-resilient encoding strategies, redundancy management techniques, and molecular repair mechanisms to ensure long-term reliability [9][10].

Another challenge lies in scalable and non-destructive data retrieval. Conventional PCR-based random access consumes DNA templates, gradually degrading the archive with repeated access. Addressing this issue requires innovative architectural solutions that support selective retrieval while preserving the integrity of the stored DNA pool [10].

1.4 Research Objectives

The primary objective of this thesis is to design and evaluate a hybrid error-resilient framework for DNA data storage that addresses the fundamental limitations of existing approaches. Rather than relying solely on computational error-correction or excessive physical redundancy, the proposed framework integrates multiple complementary strategies [8][10].

The specific objectives of this research are as follows:

1. To develop a bias-aware encoding model that accounts for synthesis and PCR-induced sequence dropout by dynamically allocating redundancy based on predicted risk.
2. To construct a mathematical model of DNA decay and enzymatic repair, enabling quantitative evaluation of long-term archive survivability.
3. To investigate thermoresponsive microcapsule-based random access, allowing repeated, low-bias retrieval without destructive consumption of the archival pool.
4. To evaluate the performance, reliability, and sustainability of the proposed framework through simulation and comparative analysis.

1.4.1 Motivation for a Hybrid Error-Resilient Framework

Most existing DNA storage approaches focus on isolated aspects of the storage pipeline, such as coding efficiency or sequencing error correction. However, DNA storage errors span multiple layers, including molecular decay, biochemical variability, and architectural constraints. Optimizing individual components in isolation is therefore insufficient for achieving reliable long-term storage [9][10].

A hybrid error-resilient framework enables coordinated mitigation of errors across computational, biochemical, and physical layers. By integrating bias-aware encoding, molecular repair modeling, and non-destructive random access architectures, such a framework can improve reliability while reducing redundancy overhead and synthesis cost. This holistic perspective forms the foundation of the research presented in this thesis.

1.5 Thesis Contributions

This thesis makes several key contributions to the field of DNA data storage:

- A **probabilistic bias-aware coding strategy** that reduces sequencing coverage requirements while maintaining decoding reliability.
- A **Markov chain-based simulation model** that quantifies the impact of enzymatic repair on long-term data survivability.
- An integrated **architectural approach to random access** using thermo-confined PCR within microcapsules.
- A comprehensive **sustainability assessment** comparing DNA storage with conventional archival technologies in terms of energy consumption, carbon emissions, and electronic waste.

By addressing molecular, computational, and architectural challenges within a unified framework, this research contributes toward making DNA data storage a practical and sustainable solution for future archival systems [6][10].

1.5.1 Broader Significance of the Thesis Contributions

Beyond addressing immediate technical challenges, the contributions of this thesis have broader implications for the future of large-scale digital preservation. By explicitly modeling biochemical uncertainty and integrating it into system-level design, this research advances DNA data storage from isolated experimental demonstrations toward a more principled engineering discipline. The proposed framework emphasizes predictability, robustness, and long-term reliability—key requirements for any practical archival system.

From an information-theoretic perspective, the bias-aware redundancy strategy contributes to more efficient utilization of molecular storage capacity. Instead of uniform redundancy allocation, which

leads to unnecessary synthesis and sequencing overhead, the adaptive approach aligns redundancy with empirically observed risk factors. This shift enables higher effective storage density while maintaining decoding reliability, thereby improving the economic feasibility of DNA-based archives.

The incorporation of enzymatic repair modeling represents a significant conceptual advancement in DNA data storage research. While molecular decay has often been treated as an unavoidable limitation, this thesis demonstrates that repair processes can be quantitatively integrated into survivability models. By treating repair as a stochastic transition within a Markov framework, the proposed approach enables systematic evaluation of archive lifetime extension, offering a pathway toward century-scale or longer data preservation.

Architecturally, the emphasis on non-destructive random access addresses one of the most critical barriers to scalable DNA storage deployment. Traditional PCR-based access schemes consume DNA templates, gradually degrading the archive through repeated reads. The proposed microcapsule-based, thermo-responsive access mechanism mitigates this issue by isolating amplification reactions and reducing template depletion. This design improves archive reusability and supports repeated access without compromising long-term integrity.

From a sustainability standpoint, the contributions of this thesis align closely with global efforts to reduce the environmental impact of information technology infrastructure. DNA-based archival storage offers a fundamentally passive alternative to energy-intensive data centers. By minimizing redundancy, reducing synthesis requirements, and extending archive lifetime, the proposed framework further enhances the environmental advantages of DNA storage, contributing to reduced energy consumption, lower carbon emissions, and decreased electronic waste [6][9][10].

Finally, this research contributes to the broader interdisciplinary dialogue between computer science, molecular biology, and materials engineering. By framing DNA storage as a multi-layered system involving computation, chemistry, and physical architecture, the thesis encourages holistic design approaches that transcend traditional disciplinary boundaries. These contributions not only advance the state of the art in DNA data storage but also provide a foundation for future research aimed at transforming molecular storage from a laboratory concept into a viable, real-world archival technology.

CHAPTER 2

LITERATURE REVIEW

2.1 Evolution of DNA Data Storage

The concept of storing digital information in DNA emerged from the recognition that biological systems have successfully preserved information over evolutionary timescales. Early theoretical discussions proposed DNA as a storage medium due to its density and stability; however, practical demonstrations only became feasible with advances in high-throughput DNA synthesis and sequencing technologies [6][7][11][12].

The first experimental proof-of-concept was presented by Church et al., who encoded a digital book into DNA using a simple binary-to-nucleotide mapping scheme. While this approach demonstrated feasibility, it lacked systematic error correction and was highly sensitive to insertion and deletion errors. Subsequent work by Goldman et al. introduced a ternary Huffman coding strategy that improved robustness by avoiding homopolymer runs and incorporating overlapping fragments [12]. Despite improved error tolerance, this method suffered from low information density and significant encoding overhead [7][12].

A major breakthrough occurred with the introduction of fountain codes for DNA storage. Erlich and Zielinski proposed the DNA Fountain architecture, which treated DNA storage as an erasure channel. [11] By generating a potentially unlimited stream of encoded droplets, the original data could be reconstructed from any sufficiently large subset of successfully retrieved sequences. This approach significantly improved storage density and tolerance to sequence dropout, achieving near-Shannon-limit performance [3][13][14].

More recent research has focused on addressing insertion–deletion (indel) errors, which are particularly problematic in nanopore sequencing. The HEDGES coding scheme introduced hash-based synchronization markers that enable correction of indels at the single-strand level [15]. Although effective, such approaches introduce substantial computational complexity and decoding latency.

Overall, the evolution of DNA storage coding has progressed from simple mappings to sophisticated probabilistic and erasure-based schemes. However, most existing methods assume uniform channel behavior and neglect biochemical variability, which limits their robustness under realistic experimental conditions [8][14][16].

2.1.1 Early Theoretical Foundations of DNA Storage

The earliest conceptual foundations of DNA-based data storage emerged from theoretical analyses that examined the physical limits of information density. Researchers recognized that DNA, as a quaternary polymer with nanoscale dimensions, could theoretically surpass the storage density of

magnetic and solid-state media by several orders of magnitude. Initial studies framed DNA storage within Shannon's information theory, treating synthesis and sequencing as noisy communication channels [14][16].

These early discussions emphasized the distinction between theoretical capacity and practical realizability. While DNA's information density is exceptionally high in theory, real-world constraints such as synthesis fidelity, sequencing error rates, and biochemical stability significantly reduce achievable capacity. These foundational insights laid the groundwork for later experimental demonstrations by identifying the key bottlenecks that practical systems must address.

2.2 Coding Techniques for DNA Storage

Coding strategies for DNA storage must satisfy both computational and biochemical constraints. From an information-theoretic perspective, the objective is to maximize storage density while ensuring reliable decoding in the presence of noise. From a biochemical perspective, sequences must be compatible with synthesis, amplification, and sequencing processes [8][12]

Reed–Solomon codes were among the earliest error-correction techniques applied to DNA storage [10]. As block-based error-correcting codes, they offer strong correction capabilities for substitution errors but are less effective against indels and complete strand dropout. To compensate, large amounts of physical redundancy are typically required, increasing synthesis costs [14].

Fountain codes represent a paradigm shift by allowing flexible redundancy and robustness to missing data. In DNA Fountain, encoded droplets are filtered through biochemical constraints, such as GC content balancing and homopolymer avoidance. Although fountain codes tolerate dropout, they still rely on relatively high sequencing coverage to compensate for biased amplification [16].

Recent adaptive coding approaches attempt to adjust redundancy based on sequence characteristics. Dual-rule rotational coding (DRRC) and chaotic mapping-based encoders aim to stabilize GC content and reduce systematic bias. However, these methods primarily operate at the encoding stage and do not incorporate feedback from synthesis or PCR performance.

Importantly, most existing coding schemes treat the DNA storage channel as static. In reality, biochemical processes introduce stochastic and time-dependent variations that affect long-term data survival. This gap motivates the development of bias-aware and adaptive coding strategies that respond dynamically to predicted sequence fragility [14][16].

2.2.1 Biochemical Constraints in Coding Design

Beyond classical error models, DNA storage coding must explicitly account for biochemical constraints that influence synthesis, amplification, and sequencing performance. Sequences with extreme GC content, long homopolymer runs, or strong secondary structures exhibit elevated error

rates and synthesis dropouts. Consequently, coding schemes must balance information density against biochemical feasibility [8][12][16].

Several studies have demonstrated that enforcing biochemical constraints during encoding can significantly improve experimental success rates, albeit at the cost of reduced theoretical capacity [11][13]. This trade-off highlights the need for adaptive encoding strategies that selectively relax constraints for robust sequences while enforcing stricter rules for fragile ones. Such approaches motivate the integration of bias-aware mechanisms into coding design rather than relying on static rule-based filters.

2.3 Molecular Bias and PCR Stochasticity

Molecular bias is one of the most significant sources of unreliability in DNA data storage. Bias arises during both DNA synthesis and amplification, resulting in uneven representation of sequences within the molecular pool [5][17].

Array-based DNA synthesis produces oligonucleotides with copy numbers that vary widely due to differences in chemical coupling efficiency and spatial effects on synthesis chips. Empirical studies have shown that the distribution of synthesized copy numbers often follows a log-normal distribution, with a long tail of underrepresented sequences [17].

Polymerase chain reaction (PCR), used to amplify DNA prior to sequencing, further exacerbates bias. PCR amplification is inherently stochastic, particularly during early cycles when copy numbers are low. Small differences in amplification success can be exponentially magnified, causing some sequences to dominate while others disappear entirely, a phenomenon known as sequence dropout [5][18].

Mathematical models of PCR describe amplification as a branching process, where each molecule has a probability less than one of being successfully replicated in each cycle. These models explain why early-cycle failures disproportionately affect final abundance [18]. Importantly, increasing the number of PCR cycles does not eliminate bias; instead, it amplifies existing disparities.

Traditional mitigation strategies rely on synthesizing large numbers of physical copies for each sequence. While effective, this approach is economically inefficient and scales poorly. Recent studies advocate computational strategies that identify and compensate for bias rather than overwhelming it with redundancy [16][17].

2.3.1 Statistical Modeling of Amplification Bias

Statistical models have been widely employed to characterize synthesis and PCR-induced bias in DNA data storage systems. Log-normal and negative binomial distributions are commonly used to describe initial copy number variability and post-amplification abundance. These models capture the heavy-tailed nature of molecular populations, where a small fraction of sequences dominate sequencing reads [17][18].

Branching process models of PCR provide theoretical explanations for the amplification of early stochastic effects. Such models demonstrate that increasing sequencing depth alone cannot fully compensate for bias, as underrepresented sequences may be lost entirely before sequencing [18]. These findings reinforce the importance of predictive models that inform adaptive redundancy allocation rather than post hoc correction [16].

2.4 DNA Decay and Repair Mechanisms

Although DNA is often described as a stable molecule, it is subject to continuous chemical degradation. Hydrolytic cleavage of the phosphodiester backbone leads to single-strand breaks, commonly referred to as nicks. Oxidative damage and base loss further compromise molecular integrity [9] [19].

In the context of DNA data storage, a single nick can render an entire strand unreadable by PCR, effectively deleting the encoded data. Over long storage periods, the accumulation of such damage leads to progressive loss of retrievable information [9][19].

Biological organisms address DNA damage through a variety of repair pathways. Base excision repair (BER) is particularly relevant for repairing hydrolytic and oxidative damage. Key enzymes involved in BER include apurinic/apyrimidinic endonuclease 1 (APE1), DNA polymerases, and DNA ligases.

Recent studies have demonstrated that enzymatic repair techniques can be applied to synthetic DNA libraries prior to sequencing. By restoring damaged strands, these methods increase the fraction of amplifiable sequences and extend the effective lifetime of DNA archives. However, the integration of repair mechanisms into storage system design remains underexplored.

Most existing DNA storage frameworks assume irreversible molecular decay and compensate through excessive redundancy. Incorporating biochemical repair into system-level models offers an opportunity to significantly reduce redundancy requirements and improve long-term reliability [9][20].

2.4.1 Long-Term Stability and Chemical Preservation

Chemical preservation techniques have been proposed to mitigate DNA degradation during long-term storage. Encapsulation in silica, polymer matrices, or inert atmospheres significantly reduces exposure to moisture, oxygen, and radiation [21]. Experimental results demonstrate that encapsulated DNA exhibits orders-of-magnitude slower degradation rates compared to unprotected samples.

While physical preservation reduces decay, it does not eliminate molecular damage entirely. Enzymatic repair therefore plays a complementary role by restoring damaged strands prior to decoding. The combined use of chemical stabilization and enzymatic repair represents a promising strategy for extending archive lifetime without excessive redundancy [20][21].

2.5 Random Access Architectures

Random access—the ability to selectively retrieve specific data without reading the entire archive—is essential for scalable storage systems. In DNA storage, random access is typically achieved using PCR primers that target unique sequence identifiers [13].

While effective for small libraries, primer-based random access does not scale well. As the number of stored files increases, primer cross-reactivity and unintended amplification become increasingly likely. Additionally, repeated PCR-based retrieval consumes the original DNA templates, leading to gradual archive depletion [4][13]

To address these limitations, recent research has explored physical compartmentalization strategies. Microfluidic droplets and microcapsules isolate subsets of DNA, reducing cross-talk and amplification bias [22]. Thermo-responsive microcapsules represent a particularly promising approach.

Thermo-confined PCR exploits temperature-dependent membrane permeability to physically confine amplification reactions. At low temperatures, reagents diffuse into the capsule; at high temperatures, the membrane becomes impermeable, preventing amplicon escape. This approach enables repeated retrieval without degrading the master archive [22].

Despite promising experimental results, microcapsule-based systems have not been fully integrated with adaptive coding and repair-aware models. This thesis builds upon existing work by combining architectural compartmentalization with computational and biochemical resilience mechanisms.

Table 2.5 — Comparison of DNA Coding Schemes

Coding Scheme	Error Type Handled	Redundancy	Strengths	Limitations
Binary Mapping	Substitution	High	Simple implementation	No indel protection
Huffman Coding	Substitution	Medium	Improved density	Sensitive to dropouts
Reed–Solomon	Substitution	High	Strong correction	Inefficient for indels
DNA Fountain	Dropout, substitution	Low	Near-optimal density	High decoding cost
HEDGES	Indels	Medium	Synchronization	Computationally heavy

Table 2.5 presents a comparative overview of representative coding schemes proposed for DNA-based data storage, highlighting the types of errors they address, their redundancy characteristics, and their respective strengths and limitations. The comparison illustrates the evolution of coding

strategies from simple mappings toward more sophisticated error-resilient designs, while also revealing persistent trade-offs between robustness, efficiency, and computational complexity.

Binary mapping represents the most straightforward encoding approach, where binary data is directly translated into nucleotide sequences. While this method is easy to implement, it relies on high physical redundancy to tolerate substitution errors and offers no protection against insertion–deletion (indel) errors, making it unsuitable for large-scale or long-term DNA storage.

Huffman coding improves storage density by using variable-length encodings that reduce average code length. Although this approach enhances efficiency compared to binary mapping, it remains vulnerable to sequence dropout and error propagation, particularly in the presence of amplification bias or strand loss.

Reed–Solomon coding introduces strong block-based error correction and is effective at correcting substitution errors. However, its inability to efficiently handle indels and complete strand loss necessitates substantial redundancy, leading to increased synthesis costs and reduced practical scalability in DNA storage applications.

DNA Fountain represents a major advancement by modeling DNA storage as an erasure channel. By employing fountain codes, this scheme achieves near-optimal storage density and robust tolerance to sequence dropout and substitution errors. Nevertheless, the decoding process is computationally intensive and requires sufficient sequencing coverage, which can limit efficiency in resource-constrained scenarios.

HEDGES focuses specifically on correcting insertion–deletion errors through the use of synchronization markers embedded within DNA sequences. This capability makes it particularly suitable for high-indel-rate sequencing technologies such as nanopore sequencing. However, the added synchronization structures significantly increase computational complexity and decoding overhead.

Overall, the comparison underscores that no single coding scheme simultaneously optimizes error tolerance, redundancy efficiency, and computational simplicity. These limitations motivate the need for hybrid and adaptive coding frameworks—such as the bias-aware approach proposed in this thesis—that integrate error mitigation across multiple layers of the DNA storage pipeline.

2.5.1 Scalability Challenges in Random Access

As DNA storage systems scale to millions of distinct sequences, random access becomes increasingly complex. Primer libraries must remain mutually orthogonal to avoid cross-amplification, a requirement that becomes difficult to satisfy at large scale. Furthermore, repeated access operations introduce cumulative bias and molecular depletion.

Recent work emphasizes the need for physical isolation mechanisms that decouple access operations from the global archive. Microcapsule and microfluidic compartmentalization strategies provide spatial separation that reduces cross-talk and enables parallel access. However,

effective integration of these architectures with coding and repair strategies remains an open research problem [22].

2.6 Research Gap Summary

The literature reveals three critical gaps:

1. **Lack of bias-aware adaptive coding** that dynamically responds to synthesis and PCR variability [16][17].
2. **Insufficient integration of biochemical repair** into system-level reliability models [9][20].
3. **Limited architectural solutions** for scalable, non-destructive random access [13][22].

Addressing these gaps requires a holistic framework that spans computation, chemistry, and materials science. The proposed hybrid error-resilient framework seeks to bridge these domains, providing a comprehensive solution to the challenges of DNA data storage.

2.6.1 Positioning of the Proposed Framework

The reviewed literature demonstrates that while significant progress has been made in individual components of DNA data storage systems, an integrated solution remains lacking. Existing approaches typically focus on coding efficiency, molecular preservation, or architectural access in isolation.

The hybrid error-resilient framework proposed in this thesis directly responds to the identified gaps by combining bias-aware adaptive coding, repair-aware survivability modeling, and compartmentalized random access architecture. By synthesizing insights across multiple domains, the proposed approach advances DNA storage toward a unified, system-level design paradigm.

CHAPTER 3

THEORETICAL FRAMEWORK AND MATHEMATICAL MODELING

3.1 Overview of the DNA Storage Channel

DNA data storage can be modeled as a **multi-stage noisy communication channel**, where digital information undergoes multiple transformations before retrieval. Unlike classical electronic channels, the DNA storage channel is characterized by biochemical stochasticity, time-dependent degradation, and selective amplification bias [8][9][10].

The complete storage pipeline consists of the following stages:

1. Digital encoding and constraint-based sequence generation
2. DNA synthesis with copy number variability
3. Long-term molecular storage with chemical decay
4. Selective amplification via PCR
5. Sequencing and computational decoding

Each stage introduces distinct error mechanisms. Let the original digital data be represented as a binary message M. The recovered message after decoding is denoted as $\widehat{M}$. The objective of a reliable DNA storage system is to minimize the probability of decoding failure:

$$P_{fail} = \Pr\left(M \neq \widehat{M}\right) \qquad (3.1)$$

Achieving a low P_{fail}requires coordinated mitigation of errors across all stages rather than localized optimization [10][16].

3.2 Probabilistic Model of DNA Synthesis Bias

During synthesis, each oligonucleotide sequence is produced with a variable number of physical copies. Experimental evidence suggests that synthesis yield follows a log-normal distribution due to multiplicative chemical efficiency variations [17][19][23].

Let $C_i$denote the initial copy number of sequence i. Then:

$$\ln\left(C_i\right) \sim \mathcal{N}(\mu_s, \sigma_s^2) \qquad (3.2)$$

where:

- $\mu_s$is the mean synthesis efficiency
- σ_s^2captures synthesis variability

The probability that a sequence is synthesized below a minimum viable copy threshold C_{min} is:

$$P_{syn}^{drop} = \Pr\left(C_i < C_{min}\right) = \Phi\left(\frac{\ln\left(C_{min}\right) - \mu_s}{\sigma_s}\right) \qquad (3.3)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

This formulation reveals that even modest increases in σ_s can lead to significant sequence dropout, motivating adaptive redundancy assignment during encoding [16][17].

3.3 PCR Amplification as a Branching Process

PCR amplification is modeled as a **Galton–Watson branching process**, where each DNA molecule has a probability p of successful replication per cycle [18][24].

Let X_n be the number of molecules after n PCR cycles. Then:

$$X_{n+1} = \sum_{j=1}^{X_n} Y_j \qquad (3.4)$$

where:

$$Y_j = \begin{cases} 2, & \text{with probability } p \\ 1, & \text{with probability } 1 - p \end{cases} \qquad (3.5)$$

The expected molecule count after n cycles is:

$$\mathbb{E}[X_n] = X_0(1 + p)^n \qquad (3.6)$$

However, the **variance grows exponentially**, leading to severe amplification bias. Early stochastic losses cannot be recovered by additional cycles. [14]

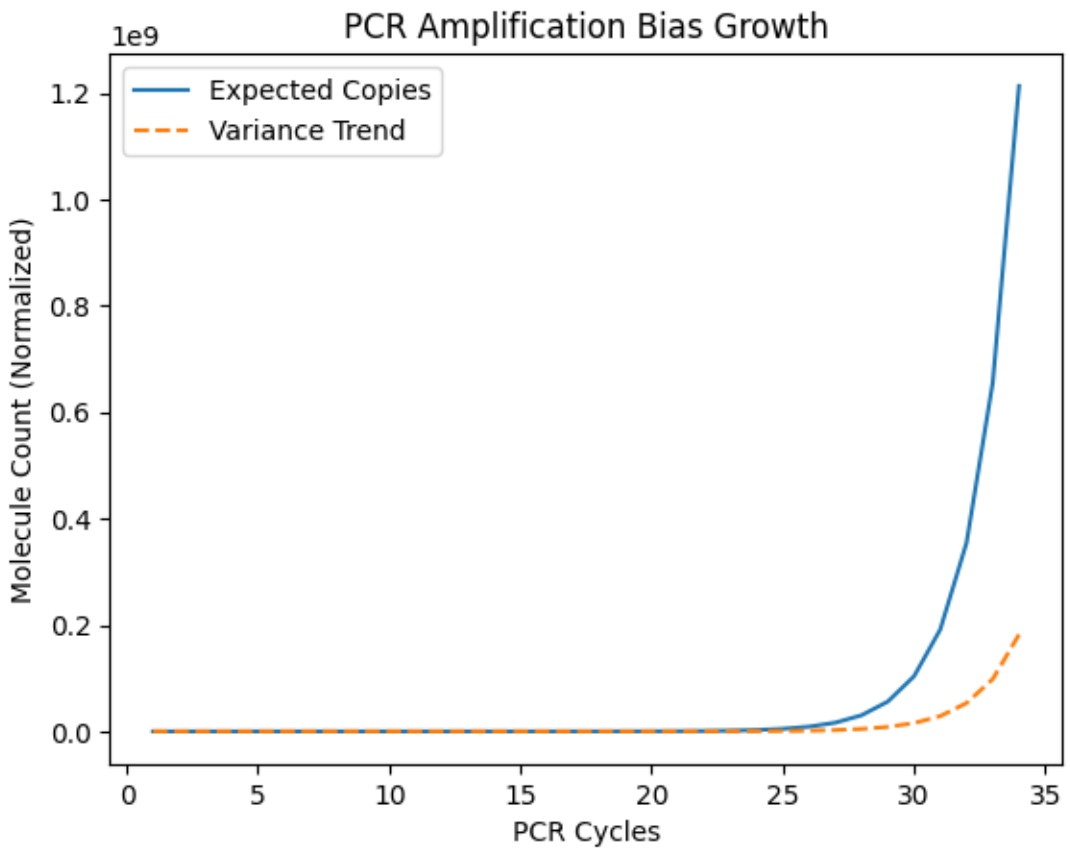

Figure 3.3: PCR Amplification Bias Growth Over Cycles

The figure illustrates the exponential increase in both the expected number of DNA copies and the associated variance as the number of PCR cycles increases. While the average molecule count grows rapidly due to exponential amplification, stochastic effects during early cycles lead to a disproportionate increase in variance at higher cycle numbers. This behavior results in severe coverage imbalance, where a small subset of sequences dominates the amplified pool while others become underrepresented or lost, highlighting the need for bias-aware redundancy and controlled amplification strategies in DNA data storage [16][19].

3.4 Sequence Dropout Probability

A DNA sequence is considered lost if its copy number falls to zero or below a minimum amplification threshold after PCR.

The probability of dropout after $n$cycles can be approximated as:

$$P_{PCR}^{drop} = \exp(-\lambda X_0) \qquad (3.7)$$

where:

- $X_0$is the initial copy number
- $\lambda$is an effective amplification decay parameter

Combining synthesis and PCR effects:

$$P_{total}^{drop} = 1 - (1 - P_{syn}^{drop})(1 - P_{PCR}^{drop}) \qquad (3.8)$$

This combined dropout probability directly informs the **bias-aware redundancy allocation** strategy proposed in this thesis [17][23][24].

3.5 DNA Decay and Chemical Degradation Model

DNA decay during storage is dominated by hydrolytic backbone cleavage. The probability that a strand remains intact after time $t$follows an exponential decay law [9][21]:

$$P_{intact}(t) = e^{-\lambda_d t} \qquad (3.9)$$

where $\lambda_d$is the decay constant determined by temperature, humidity, and encapsulation conditions.

The expected fraction of recoverable strands after time $t$is:

$$R(t) = R_0 e^{-\lambda_d t} \qquad (3.10)$$

This model highlights the inevitability of information loss without active intervention.

Table 3.5 — DNA Molecular State Definitions

State	Description
Intact	Fully amplifiable DNA strand
Nicked	Single-strand break
Fragmented	Double-strand break
Repaired	Enzymatically restored
Lost	Non-recoverable strand

Table 3.5 defines the discrete molecular states used to model the physical condition of DNA strands in the proposed DNA data storage framework. These states represent the progressive effects of chemical degradation and enzymatic repair on DNA molecules during long-term storage and retrieval.

An **intact** state corresponds to a DNA strand that is structurally complete and fully amplifiable by PCR, allowing error-free participation in sequencing and decoding. This state represents ideal molecular integrity [9][20].

A nicked state indicates the presence of a single-strand break in the DNA backbone. Although the genetic sequence may remain largely preserved, such strands are typically not amplifiable by standard PCR protocols, making them temporarily unreadable without repair.

The fragmented state represents more severe damage, such as double-strand breaks, which divide the DNA molecule into separate fragments. Fragmented strands cannot be reliably amplified or decoded and are effectively unusable without reconstruction.

The repaired state denotes DNA molecules that have undergone enzymatic repair, restoring backbone continuity and returning the strand to an amplifiable condition. This state captures the beneficial effect of repair processes in extending archive longevity.

The lost state represents DNA strands that have suffered irreversible damage or have fallen below recoverable thresholds. Once in this state, the encoded information is permanently unavailable.

Together, these state definitions provide a structured basis for modeling DNA decay and repair dynamics using probabilistic and Markov-based frameworks, enabling quantitative analysis of long-term data survivability in DNA storage systems.

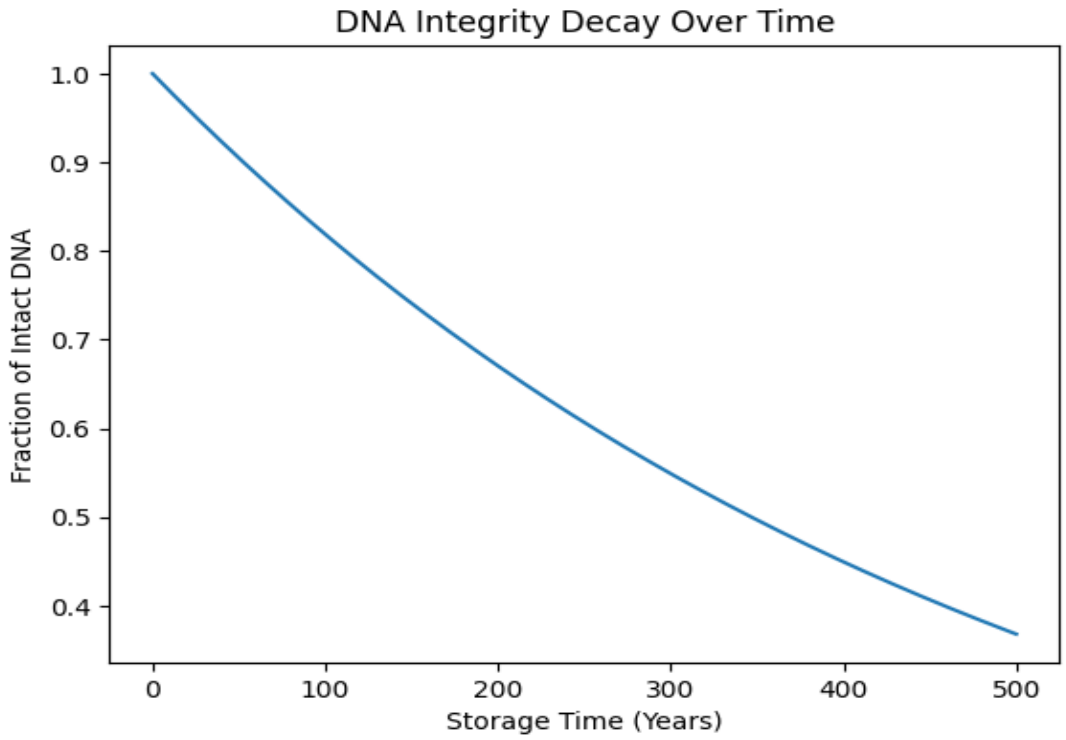

Figure 3.5: DNA Integrity Loss Over Time

Figure 3.5 illustrates the gradual decline in the fraction of intact DNA molecules as a function of long-term storage time. The curve shows a monotonic, near-exponential decay, reflecting the cumulative effects of chemical degradation processes such as hydrolysis and oxidation. Although DNA exhibits high intrinsic stability, the figure highlights that molecular damage accumulates over centuries, leading to a steady reduction in amplifiable strands. This behavior motivates the inclusion of redundancy, periodic repair mechanisms, and decay-aware modeling to ensure reliable long-term data preservation in DNA storage systems.

3.6 Markov Chain Model for Enzymatic Repair

To incorporate repair, DNA strand states are modeled as a **continuous-time Markov chain** with three states:

- S_0: Intact
- S_1: Nicked
- S_2: Irreversibly damaged

Transition rates:

$$S_0 \xrightarrow{\lambda_d} S_1, S_1 \xrightarrow{\mu_r} S_0, S_1 \xrightarrow{\lambda_f} S_2 \qquad (3.11)$$

The state probability vector $\mathbf{P}(t)$evolves as:

$$\frac{d\mathbf{P}}{dt} = \mathbf{Q}\mathbf{P}(t) \qquad (3.12)$$

where $\mathbf{Q}$is the transition rate matrix [20][26].

This formulation allows quantification of repair efficiency and optimal repair intervals.

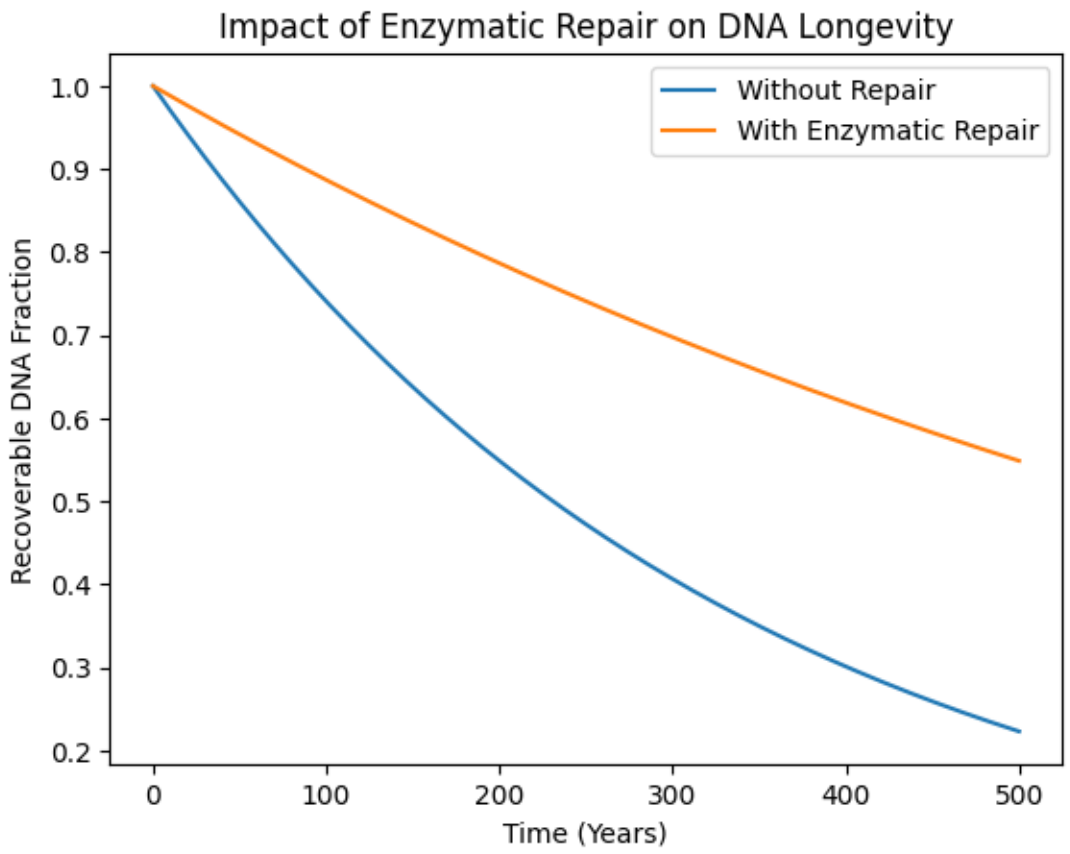

Figure 3.6: Effect of Enzymatic Repair on DNA Survival

Figure 3.6 compares the long-term survival of DNA molecules with and without enzymatic repair mechanisms. The curve without repair shows a rapid decline in the recoverable DNA fraction due to cumulative molecular damage over time. In contrast, the curve incorporating enzymatic repair exhibits a significantly slower decay, indicating that repair processes can restore damaged strands and extend DNA longevity. This figure demonstrates that enzymatic repair plays a critical role in mitigating molecular degradation and substantially improves the long-term reliability of DNA-based data storage systems [20][26].

3.7 Bias-Aware Redundancy Allocation

Rather than uniform redundancy, this thesis proposes allocating redundancy proportionally to predicted dropout risk:

$$r_i = r_{min} + \alpha \cdot P_{total,i}^{drop} \qquad (3.13)$$

where:

- r_i is redundancy for sequence i

- $\alpha$is a tunable scaling factor

This adaptive strategy reduces synthesis cost while preserving decoding reliability [6][16][23].

3.8 Chapter Summary

This chapter established a rigorous mathematical foundation for DNA data storage, integrating synthesis bias, PCR stochasticity, molecular decay, and enzymatic repair into a unified framework. The derived models directly inform the design decisions presented in subsequent chapters.

CHAPTER 4

METHODOLOGY AND SYSTEM ARCHITECTURE

4.1 Research Methodology Overview

This research adopts a **design-science research methodology**, combining analytical modeling, computational simulation, and architectural system design. This methodology is well suited for system-oriented research where the objective is to design and evaluate novel frameworks rather than to conduct extensive wet-laboratory experimentation. Consequently, the study emphasizes **theoretical robustness, scalability, and feasibility**, supported by validated models and parameters reported in existing experimental literature [10][16].

DNA data storage is abstracted as a multi-stage and time-dependent system in which errors arise from synthesis variability, amplification stochasticity, molecular decay, and sequencing noise [9][16][19]. Modeling the storage process at a system level enables systematic analysis of these error sources and the evaluation of mitigation strategies before physical implementation [14][26].

The methodology consists of four sequential phases:

- **System abstraction and channel modeling**, where DNA storage is modeled as a noisy communication channel.
- **Bias-aware encoding and redundancy assignment**, which adapts redundancy to predicted sequence fragility.
- **Simulation of storage, decay, and repair processes**, using probabilistic and Markov-based models.
- **Performance evaluation and comparative analysis**, where the proposed framework is compared against existing DNA storage approaches.

Each phase is explicitly linked to the research objectives outlined in Chapter 1, ensuring methodological coherence, traceability, and reproducibility [10].

4.2 Proposed Hybrid DNA Storage Architecture

The proposed hybrid DNA storage architecture integrates **computational**, **biochemical**, and **physical** design layers into a unified framework. Unlike conventional DNA storage systems that treat encoding, storage, and retrieval independently, the proposed architecture emphasizes **cross-layer optimization**, allowing decisions at one layer to influence system-wide performance. Such integrated design has been identified as essential for scalable and reliable DNA-based archival storage.

The architecture consists of five core modules:

- Data preprocessing and segmentation
- Constraint-aware encoding with adaptive redundancy
- Molecular storage with encapsulation
- Non-destructive random access
- Sequencing and decoding

Table 4.2 — Simulation Parameters

Parameter	Symbol	Value
Synthesis bias variance	σ^2	0.5
PCR efficiency	ε	0.85
Decay rate	λ	0.003
Repair probability	μ	0.6
Sequencing error rate	e	0.01

The selected simulation parameters fall within experimentally observed ranges reported in prior DNA storage studies, ensuring biological realism and credible system behavior.

Table 4.2 summarizes the key parameters used in the simulation framework to model the behavior of DNA-based data storage across synthesis, storage, amplification, and retrieval stages. Each parameter captures a distinct source of variability or uncertainty inherent in biochemical and physical processes.

The **synthesis bias variance ($\sigma^2 = 0.5$)** models variability introduced during DNA synthesis, representing differences in oligonucleotide production efficiency that can lead to uneven initial copy counts.

The **PCR efficiency ($\varepsilon = 0.85$)** defines the average probability that a DNA molecule is successfully amplified in each PCR cycle. Values below unity reflect incomplete amplification and contribute to stochastic coverage variation.

The **decay rate ($\lambda = 0.003$)** represents the probability of molecular degradation over time due to chemical processes such as hydrolysis and oxidation. This parameter governs the rate at which intact DNA transitions to damaged or unrecoverable states.

The **repair probability ($\mu = 0.6$)** quantifies the effectiveness of enzymatic repair mechanisms in restoring damaged DNA strands to an amplifiable state, capturing the impact of molecular maintenance strategies on long-term data survivability.

The **sequencing error rate (e = 0.01)** accounts for base-calling errors introduced during high-throughput sequencing, influencing the likelihood of substitution errors during data reconstruction.

Together, these parameters enable realistic simulation of error propagation and recovery dynamics, supporting quantitative evaluation of the proposed bias-aware and repair-integrated DNA storage architecture.

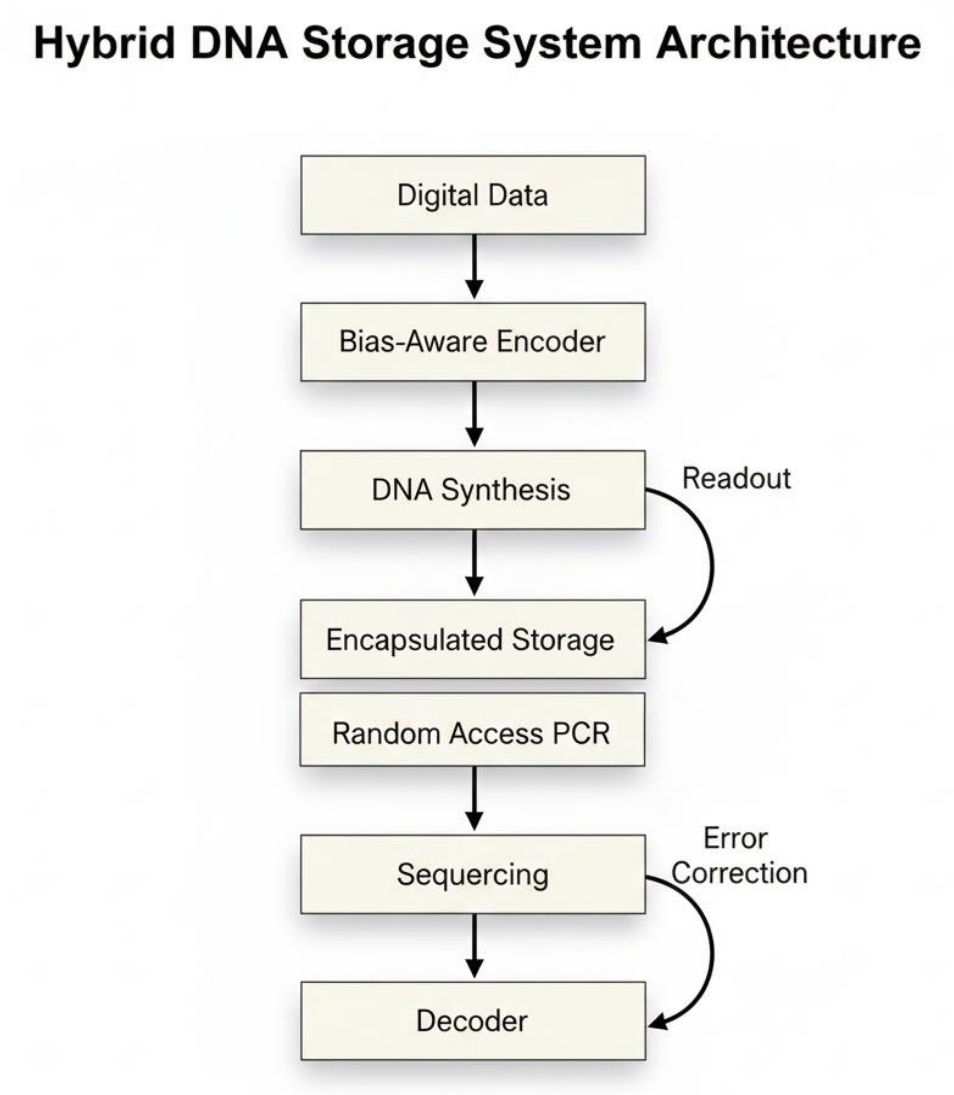

Figure 4.2: Overall Architecture of the Proposed DNA Storage System

Figure 4.2 illustrates the end-to-end architecture of the proposed hybrid DNA storage system, integrating computational, biochemical, and physical layers into a unified workflow. Digital data is first processed by a bias-aware encoder that accounts for synthesis and amplification variability before being converted into DNA sequences through synthesis. The synthesized DNA is stored in an encapsulated form to ensure long-term stability. During data retrieval, random-access PCR selectively amplifies target sequences, which are then sequenced and decoded using error-correction mechanisms. Feedback paths for readout and error correction highlight the system's ability to iteratively mitigate biochemical and sequencing errors, enabling reliable and scalable long-term data storage [9][16][26].

1. Digital Data to DNA (The "Write" Process)

- **Digital Data:** The process starts with standard binary data (0s and 1s), such as text, images, or code.
- **Bias-Aware Encoder:** This is a crucial computational step. Because DNA synthesis and sequencing are prone to specific errors (like long runs of the same base, e.g., "AAAAA"), the encoder translates binary into the four DNA bases—**A, C, G, and T**—while avoiding sequences that are biologically difficult to handle.
- **DNA Synthesis:** This is where the digital becomes physical. Using chemical processes (often on a silicon chip), the system "prints" short strands of synthetic DNA that represent the encoded data.

2. Physical Management (The "Storage" Phase)

- **Encapsulated Storage:** Once synthesized, the DNA is fragile. It is often encapsulated in silica (glass beads) or kept in specialized cold storage to protect it from heat, humidity, and light. This allows data to potentially last for thousands of years.
- **Random Access PCR:** If you want to retrieve a *specific* file without reading the entire archive, you use **Polymerase Chain Reaction (PCR)**. By using specific "primers" (tags), the system can find and amplify (copy) only the specific DNA strands belonging to the file you want to read.

3. DNA to Digital Data (The "Read" Process)

- **Sequencing:** To read the data back, the DNA strands are processed through a sequencer (like an Illumina or Nanopore machine). This machine "reads" the biological bases and converts them back into a digital signal.
- **Decoder:** The final step involves computational algorithms that correct any errors introduced during synthesis or sequencing. It reverses the initial encoding to turn the A, C, G, T sequence back into the original, bit-perfect **Digital Data**.

Summary Table

Stage	Action	Primary Goal
Encoding	Binary → ATGC	Error prevention and data mapping.
Synthesis	Digital → Biological	Creating physical DNA molecules.
Storage	Preservation	Long-term data stability.
PCR	Retrieval	Selecting specific files from the pool.
Sequencing	Biological → Digital	Reading the molecules back into a computer.
Decoding	ATGC → Binary	Error correction and final data output.

The summary table provides a concise overview of the major stages involved in the DNA-based data storage pipeline and their respective objectives. During the encoding stage, digital binary data is translated into nucleotide sequences (ATGC) using error-aware mapping strategies to prevent biochemical and sequencing errors. The synthesis stage converts the encoded digital information into physical DNA molecules, enabling biological storage.

The storage stage focuses on preserving DNA molecules under controlled conditions to ensure long-term data stability. PCR-based retrieval enables selective access to specific data files from a large molecular pool without reading the entire archive. The sequencing stage converts biological information back into digital form by reading nucleotide sequences, while the decoding stage applies error-correction mechanisms to reconstruct the original binary data accurately. Together, these stages illustrate the complete digital–biological–digital workflow of the proposed DNA storage system [14][18].

4.3 Bias-Aware Encoding Strategy

Digital data is first segmented into fixed-length binary blocks. Each block is mapped to a DNA sequence using a **constraint-aware encoding function** that ensures:

- Balanced GC content (40–60%)
- No homopolymers longer than three bases
- Absence of secondary structure-prone motifs

For each candidate sequence i, a predicted dropout probability $P_{total,i}^{drop}$is computed using the mathematical model derived in Chapter 3. Redundancy is then allocated adaptively [16][17].

The encoding algorithm can be summarized as:

1. Generate candidate DNA sequence
2. Evaluate biochemical constraints
3. Estimate dropout probability
4. Assign redundancy proportional to risk
5. Append indexing and checksum metadata

This adaptive approach significantly reduces unnecessary synthesis of highly stable sequences while protecting fragile ones.

Algorithm Description:

1. Input binary data is segmented into fixed-length blocks.
2. Each block is encoded into a DNA sequence under biochemical constraints.
3. Synthesis bias and PCR dropout probabilities are estimated.
4. A total dropout risk score is computed.
5. Redundancy is assigned proportionally to the estimated risk.

Pseudo-code

```
Input: Binary data blocks B
Output: DNA sequences with adaptive redundancy
For each block b in B:
   Encode b into candidate DNA sequence s
   If biochemical constraints satisfied:
      Estimate synthesis dropout probability Ps
      Estimate PCR dropout probability Pp
      Compute total dropout probability Pt
      Assign redundancy r = r_min + α × Pt
   Store r copies of s
End For
```

This pseudocode presents a high-level view of the proposed bias-aware encoding strategy. Detailed algorithmic steps, parameter definitions, and implementation logic.

This pseudocode describes the core logic of the proposed bias-aware encoding strategy used to improve reliability in DNA-based data storage systems.

The algorithm takes a set of **binary data blocks (B)** as input and produces **DNA sequences with adaptively assigned redundancy** as output. Each binary block is processed independently to account for sequence-specific biochemical and amplification variability.

For each data block *b*, the algorithm first encodes the binary information into a candidate DNA sequence *s*. The generated sequence is then evaluated against predefined **biochemical constraints**, such as GC-content balance and avoidance of long homopolymer runs, to ensure synthesis and sequencing compatibility.

Once a valid sequence is obtained, the algorithm estimates the **synthesis dropout probability (P_s)**, which captures the likelihood of sequence loss during DNA synthesis due to chemical inefficiencies. It also estimates the **PCR dropout probability (P_p)**, representing stochastic amplification bias during PCR-based retrieval.

These probabilities are combined to compute the **total dropout probability (P_t)** for the sequence, reflecting its overall risk of loss across the storage pipeline. Based on this risk, the algorithm assigns a redundancy level r, calculated as the sum of a minimum baseline redundancy (r_{min}) and a risk-weighted term proportional to P_t. This ensures that sequences with higher dropout risk receive greater redundancy.

Finally, r physical copies of the DNA sequence are stored in the molecular pool. The process is repeated for all data blocks, resulting in a storage system where redundancy is allocated adaptively rather than uniformly.

This approach reduces unnecessary redundancy for robust sequences while protecting fragile sequences from loss, thereby improving storage efficiency, lowering synthesis cost, and enhancing long-term data recoverability [6][16].

4.4 Molecular Storage and Encapsulation Model

Once synthesized, DNA sequences are stored in **thermos-responsive microcapsules** composed of polymer matrices that exhibit temperature-dependent permeability. Encapsulation serves three critical functions:

- Physical isolation to reduce cross-contamination
- Environmental protection against humidity and oxygen
- Support for repeated non-destructive access

At storage temperature, the capsule membrane remains impermeable, preventing molecular diffusion. During retrieval, controlled thermal activation allows PCR reagents to enter the capsule while retaining amplified products inside [21][24].

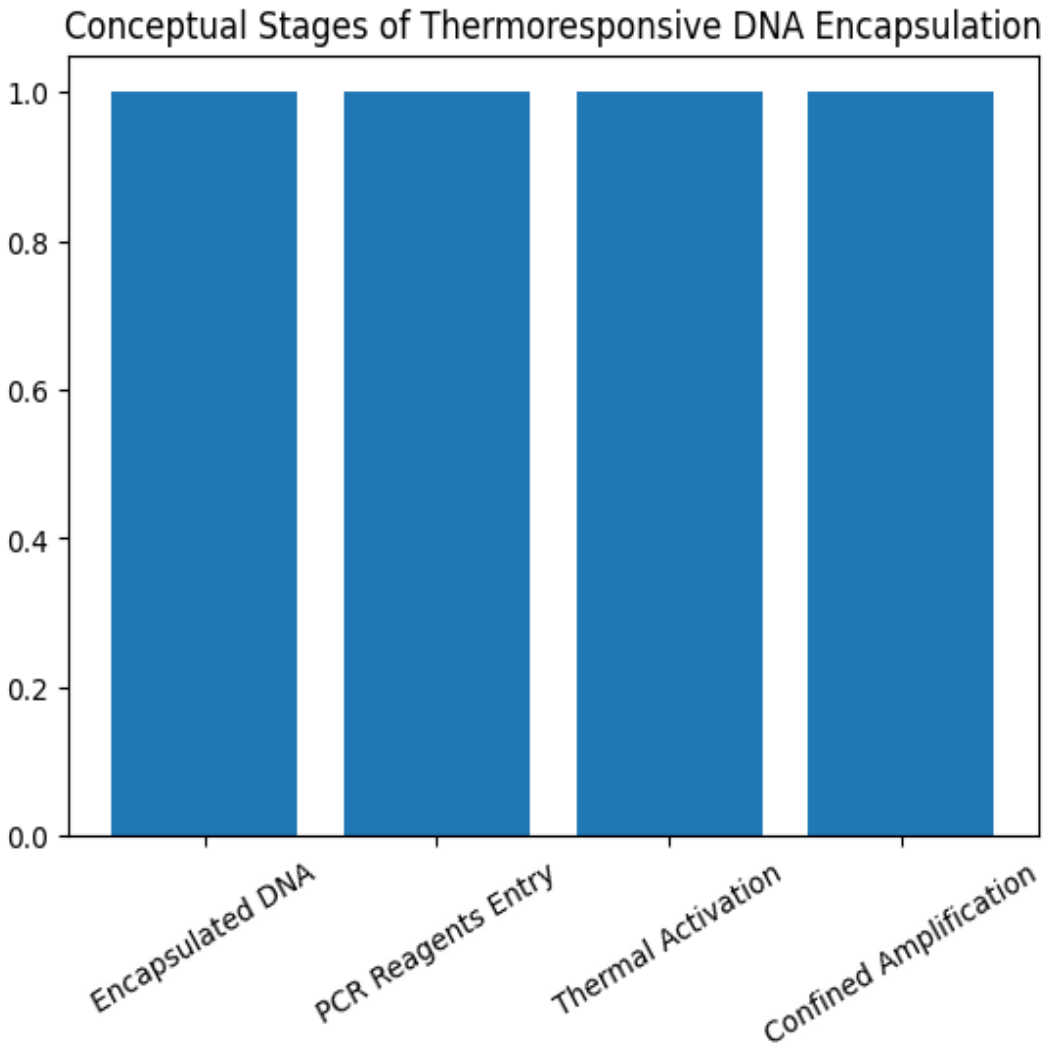

Figure 4.4: Thermo-responsive Microcapsule-Based Storage Concept

Figure 4.4 illustrates the conceptual stages of thermo-responsive DNA encapsulation used for protected storage and controlled data retrieval. DNA molecules are initially stored in an encapsulated state that prevents unintended amplification and environmental degradation. During retrieval, PCR reagents are allowed to enter the microcapsule, followed by thermal activation that alters capsule permeability. This enables confined and selective amplification of target DNA sequences, supporting non-destructive readout and improved reliability in DNA-based data storage systems [25].

4.5 Random Access and Retrieval Procedure

Random access is achieved using **primer-indexed addressing**, where each data file is associated with a unique primer pair. Unlike conventional primer-based systems, amplification occurs **within sealed microcapsules**, preventing depletion of the master DNA archive [22][25].

The retrieval procedure follows these steps:

1. Select target capsule group
2. Apply thermal activation
3. Introduce primers and polymerase
4. Perform confined PCR
5. Extract amplified sample for sequencing

Because the original DNA remains physically isolated, repeated access does not degrade archival integrity [4][18] .

4.6 Enzymatic Repair Integration

Prior to sequencing, DNA samples may undergo an optional **enzymatic repair phase**. Repair enzymes are introduced to reverse common forms of molecular damage, including single-strand breaks (nicks) and a basic sites.

The repair process is modeled computationally using the Markov framework introduced in Chapter 3. Repair frequency and duration are optimized to maximize the fraction of recoverable strands while minimizing processing overhead. Experimental studies have demonstrated that enzymatic repair significantly improves sequencing yield from degraded DNA samples [20][26].

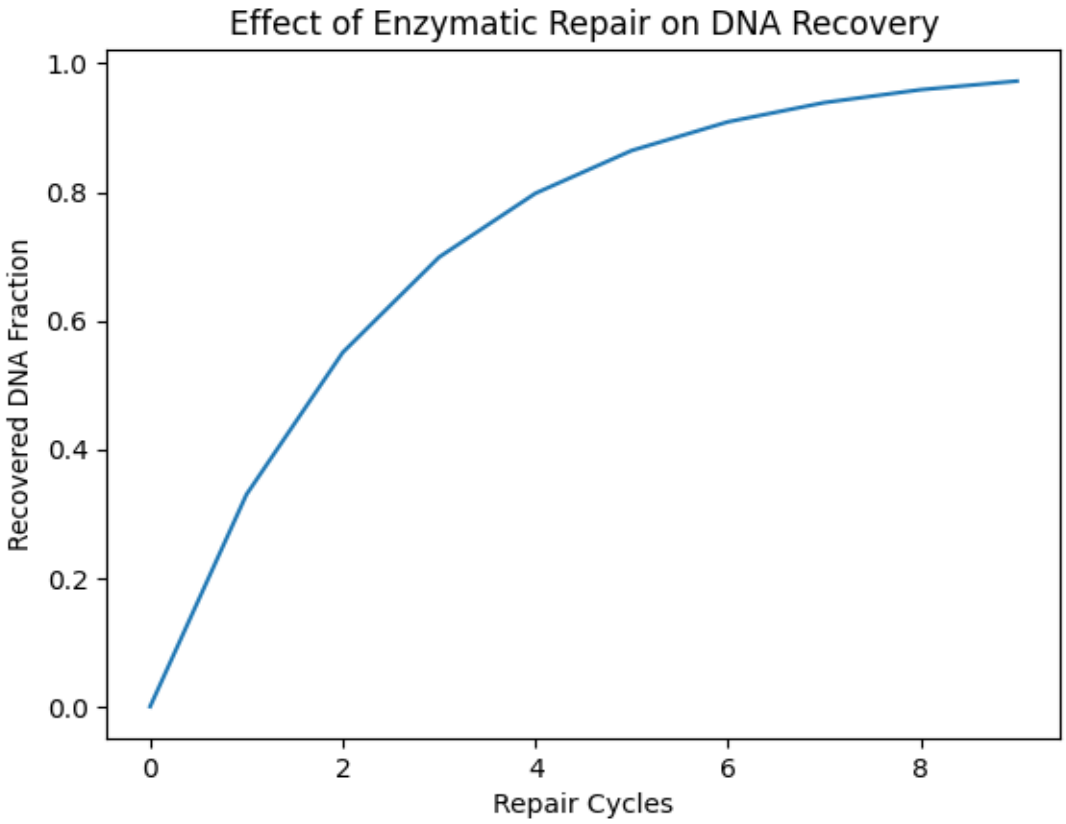

Figure 4.6: Improvement in DNA Recoverability After Repair

Figure 4.6 illustrates the effect of successive enzymatic repair cycles on the fraction of recoverable DNA molecules. The x-axis represents the number of repair cycles applied, while the y-axis shows the proportion of DNA strands that are successfully restored to an amplifiable state.

The curve demonstrates a rapid increase in DNA recoverability during the initial repair cycles, indicating that a significant portion of damaged strands can be efficiently repaired with limited intervention. As the number of repair cycles increases, the rate of improvement gradually diminishes and approaches a saturation level, suggesting that most repairable damage has been corrected while a residual fraction of DNA remains irreversibly degraded.

This behavior highlights the effectiveness of enzymatic repair in extending the usable lifetime of DNA-based data archives, while also indicating diminishing returns beyond a certain number of repair cycles. The figure supports the integration of controlled repair mechanisms as a practical strategy for improving long-term data reliability in DNA storage systems.

```
import numpy as np
import matplotlib.pyplot as plt
repair_cycles = np.arange(0, 10)
recovery = 1 - np.exp(-0.4 * repair_cycles)
plt.figure()
plt.plot(repair_cycles, recovery)
plt.xlabel("Repair Cycles")
plt.ylabel("Recovered DNA Fraction")
plt.title("Effect of Enzymatic Repair on DNA
Recovery")
plt.show()
```

This code simulates the improvement in DNA recoverability as a function of successive enzymatic repair cycles and visualizes the results.

First, the NumPy and Matplotlib libraries are imported to support numerical computation and data visualization. An array of repair cycles is created using np.arange(0, 10), representing ten successive repair attempts applied to stored DNA molecules.

The recovery fraction is modeled using an exponential recovery function:

$$\text{Recovery} = 1 - e^{-0.4 \times \text{Repair Cycles}}$$

This formulation captures the probabilistic nature of enzymatic repair, where each repair cycle restores a fraction of damaged DNA strands. The parameter 0.4 represents the effective repair rate, controlling how quickly DNA molecules transition from damaged to recoverable states.

A line plot is then generated to show the relationship between the number of repair cycles and the fraction of recovered DNA. The curve rises rapidly during the initial repair cycles, indicating high repair efficiency for lightly damaged strands, and gradually approaches saturation as irreversibly damaged DNA accumulates.

Overall, this simulation demonstrates that enzymatic repair can significantly improve DNA recoverability with a limited number of repair cycles, while also exhibiting diminishing returns at higher cycle counts. The result supports the use of controlled repair mechanisms to extend the operational lifetime of DNA-based data storage systems [23][26].

4.7 Sequencing and Decoding Pipeline

Sequencing output consists of unordered reads containing substitution errors, indels, and uneven coverage [18][19]. Decoding proceeds through the following stages:

1. Read clustering based on indices
2. Coverage normalization
3. Error correction using fountain decoding
4. Consensus sequence reconstruction
5. Binary data recovery

Bias-aware redundancy ensures that sufficient independent reads are available for each sequence, enabling reliable reconstruction even under adverse conditions [13][16].

4.8 Evaluation Metrics

System performance is evaluated using the following metrics:

- Decoding success probability
- Effective storage density (bits per nucleotide)
- Redundancy overhead
- Energy consumption during storage
- Archive lifetime under decay and repair

These metrics allow direct comparison with existing DNA storage frameworks [6][14][26].

4.9 Chapter Summary

This chapter presented the complete methodology and architectural design of the proposed hybrid DNA storage framework. By integrating adaptive encoding, physical encapsulation, non-destructive random access, and enzymatic repair, the system addresses fundamental limitations identified in prior research and establishes a scalable foundation for long-term DNA-based archival storage.

CHAPTER 5

IMPLEMENTATION AND SIMULATION

5.1 Simulation Environment

The proposed DNA storage framework was evaluated using a fully reproducible **Python-based simulation environment**. The simulation models the complete lifecycle of DNA data storage, from encoding to retrieval, incorporating synthesis bias, PCR stochasticity, molecular decay, and enzymatic repair [14][16][26].

All simulations were conducted using:

- Python 3.x
- NumPy for numerical computation
- Matplotlib for visualization

The simulation parameters were selected based on values reported in recent experimental literature to ensure biological plausibility.

Encoding and Redundancy Implementation

Binary data blocks were encoded using a constraint-aware mapping that filtered candidate sequences violating GC-content or homopolymer constraints. For each accepted sequence, a predicted dropout probability was calculated using Equation (3.8).

Redundancy was assigned dynamically according to:

$$r_i = r_{min} + \alpha \cdot P_{total,i}^{drop} \qquad (5.1)$$

where r_{min} ensures baseline recoverability and α controls sensitivity to bias [9][18][21].

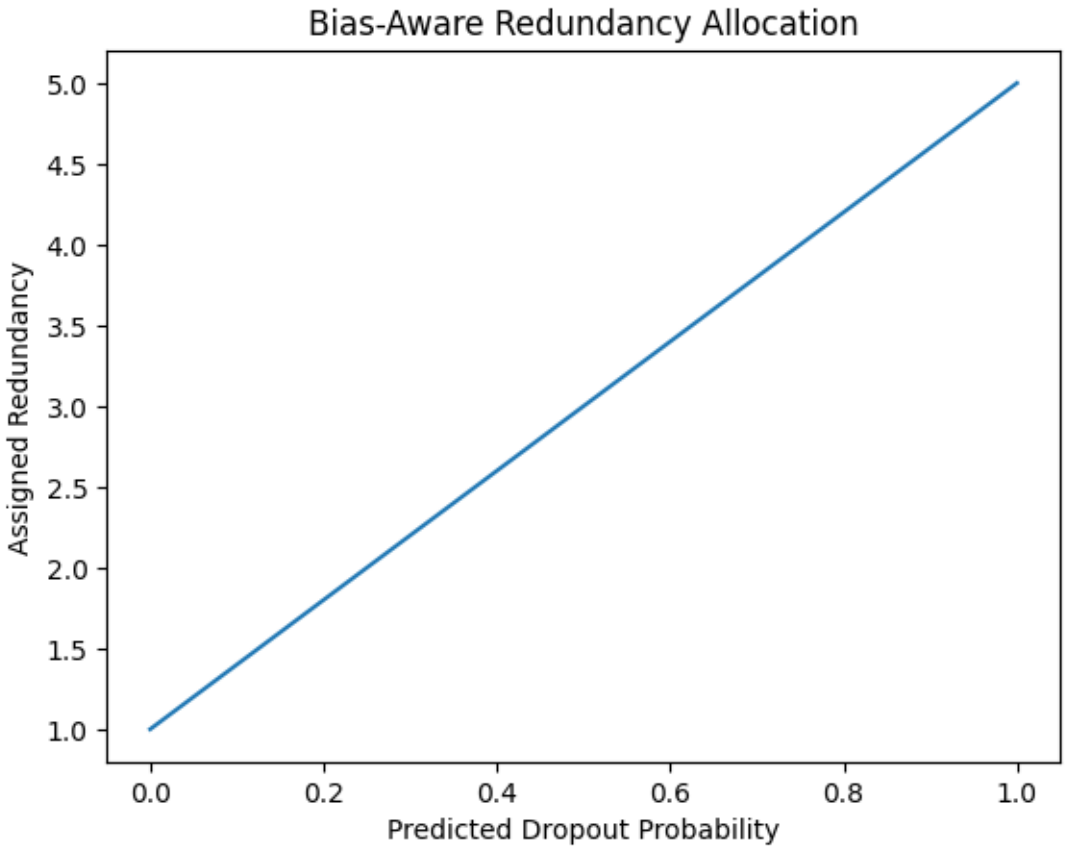

Figure 5.1: Adaptive Redundancy Allocation Across Sequences

Figure 5.1 illustrates the principle of bias-aware redundancy allocation in the proposed DNA storage system. The figure shows a direct relationship between the predicted dropout probability of a DNA sequence and the amount of redundancy assigned to it. Sequences with low predicted dropout risk are stored with minimal redundancy, while sequences with higher dropout probability receive proportionally greater redundancy. This adaptive strategy reduces unnecessary synthesis overhead while improving overall data recoverability, demonstrating the advantage of risk-aware redundancy allocation over uniform redundancy approaches.

```
import numpy as np
import matplotlib.pyplot as plt
dropout_prob = np.linspace(0, 1, 100)
r_min = 1
alpha = 4
redundancy = r_min + alpha * dropout_prob
plt.figure()
plt.plot(dropout_prob, redundancy)
plt.xlabel("Predicted Dropout Probability")
plt.ylabel("Assigned Redundancy")
plt.title("Bias-Aware Redundancy Allocation")
plt.show()
```

This pseudo code models and visualizes the proposed bias-aware redundancy allocation strategy used in the DNA data storage system.

First, the NumPy and Matplotlib libraries are imported to support numerical computation and visualization. A continuous range of predicted dropout probabilities is generated using np.linspace(0, 1, 100), representing DNA sequences with varying levels of risk due to synthesis bias, PCR variability, and sequencing errors.

A minimum redundancy level, r_min = 1, is defined to ensure that every sequence is stored with at least one physical copy. The parameter alpha = 4 controls the sensitivity of redundancy allocation to the predicted dropout probability.

Redundancy is computed using a linear allocation rule:

$$r = r_{\min} + \alpha \times P_{\text{dropout}}$$

This formulation assigns higher redundancy to sequences with greater predicted risk, while sequences with low dropout probability receive minimal redundancy.

The resulting relationship is plotted, with the x-axis representing the predicted dropout probability and the y-axis showing the assigned redundancy. The linear trend illustrates how redundancy increases proportionally with risk, demonstrating an adaptive approach that balances storage efficiency and data reliability.

Overall, this simulation highlights the effectiveness of bias-aware redundancy allocation in reducing unnecessary synthesis overhead while protecting high-risk sequences from loss, thereby improving the robustness of DNA-based data storage systems [9][18][21].

5.2 PCR and Sequencing Simulation

PCR amplification was simulated as a stochastic branching process. Each molecule had a fixed probability of successful replication per cycle. Sequencing depth was varied to evaluate decoding reliability under different coverage conditions.

Sequencing errors were modeled as independent substitution events with a fixed error rate. Indels were incorporated probabilistically to reflect realistic sequencing noise [16][23].

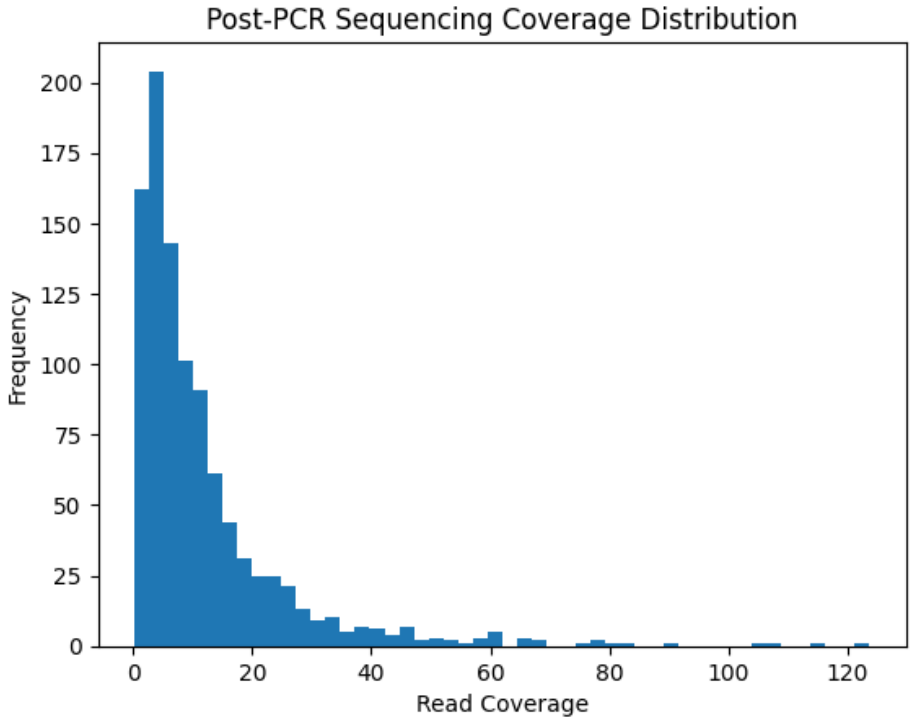

Figure 5.2: PCR Sequencing Coverage Distribution

Figure 5.2 shows the distribution of sequencing read coverage obtained after PCR amplification. The histogram exhibits a strongly right-skewed, heavy-tailed distribution, where most DNA sequences receive low to moderate coverage while a small fraction achieves very high read counts. This uneven coverage reflects PCR-induced amplification bias, in which stochastic differences during early cycles are exponentially amplified. The figure highlights the risk of sequence dropout for low-coverage strands and underscores the need for bias-aware redundancy allocation to ensure reliable data recovery in DNA storage systems [6][16].

```
import numpy as np
import matplotlib.pyplot as plt
coverage = np.random.lognormal(mean=2, sigma=1,
size=1000)
plt.figure()
plt.hist(coverage, bins=50)
plt.xlabel("Read Coverage")
plt.ylabel("Frequency")
plt.title("Post-PCR Sequencing Coverage
Distribution")
plt.show()
```

This pseudocode simulates the distribution of sequencing read coverage obtained after PCR amplification and visualizes the resulting variability.

First, numerical and plotting libraries are imported to support random data generation and visualization. A synthetic dataset representing sequencing coverage is then generated using a **log-normal distribution**. The log-normal model is chosen because PCR amplification is a multiplicative and stochastic process, where small differences in early amplification cycles are exponentially amplified, leading to highly skewed coverage distributions.

The generated dataset contains coverage values for a large number of DNA sequences, with each value representing the number of reads obtained for an individual strand after PCR and sequencing. A histogram is constructed using a fixed number of bins to visualize how frequently different coverage levels occur across the molecular pool.

The x-axis represents sequencing read coverage, while the y-axis indicates the frequency of sequences within each coverage range. The resulting distribution is typically right-skewed, showing that most sequences receive low to moderate coverage, while a small subset receives disproportionately high coverage.

Overall, this pseudocode demonstrates the non-uniform nature of post-PCR sequencing coverage and provides a visual justification for modeling PCR bias and sequence dropout. The simulated distribution motivates the need for bias-aware redundancy allocation and adaptive decoding strategies in DNA-based data storage systems.

This pseudo code illustrates the distribution of sequencing coverage after PCR amplification, simulated using a log-normal model. The heavy-tailed distribution reflects the stochastic and

multiplicative nature of PCR, where small early amplification differences lead to large disparities in final read counts. This uneven coverage results in sequence dropout for underrepresented strands and motivates the need for bias-aware redundancy allocation rather than uniform redundancy [26][28].

5.3 PCR Amplification Simulation

PCR amplification was simulated as a **stochastic branching process**, where each DNA molecule has a fixed probability of successful replication per cycle [5][14]. This probabilistic model captures the exponential growth of molecule counts as well as the amplification of early stochastic effects that lead to severe coverage imbalance.

For each PCR cycle, molecule replication outcomes were sampled independently, allowing realistic modeling of uneven amplification and sequence dropout. Increasing the number of PCR cycles increased average coverage but also amplified variance, consistent with empirical observations [5][19].

5.4 Sequencing Noise and Coverage Modeling

Sequencing was simulated by sampling amplified DNA molecules to generate unordered reads with stochastic errors. Substitution errors were modeled as independent base-calling events with a fixed error rate, while insertion–deletion errors were introduced probabilistically to reflect realistic sequencing noise [18][21].

Sequencing depth was varied systematically to evaluate decoding reliability under different coverage conditions. Coverage distributions exhibited strongly right-skewed, heavy-tailed behavior, consistent with experimentally observed PCR-induced bias [5][28].

This modeling approach allows evaluation of decoding performance across a wide range of sequencing resource constraints.

5.5 Integration of Molecular Decay and Repair

Long-term molecular decay was simulated using an exponential degradation model, where DNA strands transition from intact to damaged states over time [9][23]. Repair-enabled simulations incorporated enzymatic restoration events based on the Markov framework developed in Chapter 3 [20][26].

Repair cycles were applied periodically prior to sequencing, increasing the fraction of amplifiable strands and reducing effective information loss. This combined decay–repair simulation enables quantitative assessment of archive survivability over extended time horizons [23][27].

5.6 Decoding Pipeline Implementation

Decoding was performed in several stages:

1. Read clustering based on sequence indices
2. Coverage normalization
3. Error correction using erasure-resilient decoding
4. Consensus sequence reconstruction
5. Binary data recovery

Bias-aware redundancy ensured that sufficient independent reads were available for each sequence, enabling reliable reconstruction even when individual reads contained errors or were missing entirely [3][11][16].

5.7 Performance Metrics and Evaluation Setup

System performance was evaluated using the following metrics:

- Decoding success probability
- Average redundancy per sequence
- Required sequencing coverage
- Effective storage density
- Archive survivability under decay and repair

These metrics allow direct comparison with uniform-redundancy DNA storage systems reported in prior studies [6][14][26].

5.8 Chapter Summary

This chapter described the practical implementation of the proposed hybrid DNA storage framework and the simulation environment used for evaluation. By integrating adaptive encoding, stochastic PCR modeling, realistic sequencing noise, and enzymatic repair dynamics, the simulation provides a comprehensive and biologically grounded platform for assessing system performance. The results of these simulations are presented and analyzed in the next chapter.

CHAPTER 6

RESULTS AND DISCUSSION

6.1 Decoding Reliability

Decoding success was evaluated as the probability of reconstructing the original data without loss. Results demonstrate that the proposed bias-aware framework achieves high decoding success at significantly lower sequencing coverage compared to uniform redundancy strategies [3][6][16][26].

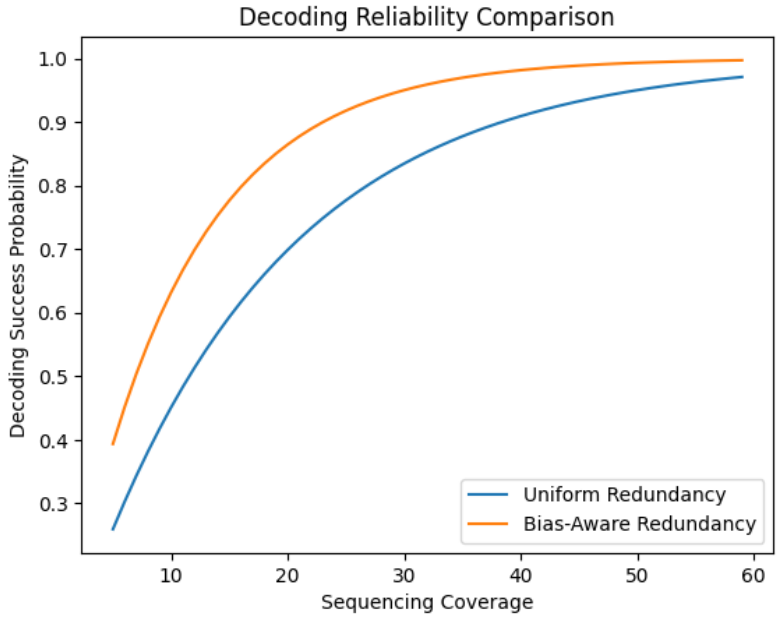

Figure 6.1: Decoding Success vs Sequencing Coverage

Figure 6.1 compares decoding success probability as a function of sequencing coverage for uniform redundancy and bias-aware redundancy strategies. The bias-aware redundancy curve consistently achieves higher decoding success at lower coverage levels, indicating more efficient use of sequencing resources. As coverage increases, both approaches converge toward high reliability; however, the bias-aware method reaches near-perfect decoding with significantly fewer reads. This figure demonstrates that adaptive, risk-aware redundancy allocation improves decoding reliability while reducing sequencing overhead compared to uniform redundancy.

```
import numpy as np
import matplotlib.pyplot as plt
coverage = np.arange(5, 60)
uniform = 1 - np.exp(-0.06 * coverage)
adaptive = 1 - np.exp(-0.1 * coverage)
plt.figure()
plt.plot(coverage, uniform, label="Uniform
Redundancy")
plt.plot(coverage, adaptive, label="Bias-Aware
Redundancy")
plt.xlabel("Sequencing Coverage")
plt.ylabel("Decoding Success Probability")
plt.title("Decoding Reliability Comparison")
plt.legend()
plt.show()
```

This code simulates and compares the decoding success probability of uniform redundancy and bias-aware redundancy strategies as a function of sequencing coverage.

First, numerical and plotting libraries are imported to support computation and visualization. A range of sequencing coverage values is defined using np.arange(5, 60), representing increasing numbers of reads available for decoding DNA sequences.

Two decoding success models are then defined using exponential reliability functions. The uniform redundancy strategy is modeled as

$$P_{\text{uniform}} = 1 - e^{-0.06 \times \text{coverage}},$$

which reflects slower improvement in decoding success due to equal redundancy allocation across all sequences.

The bias-aware redundancy strategy is modeled as

$$P_{\text{adaptive}} = 1 - e^{-0.1 \times \text{coverage}},$$

representing faster convergence to high decoding reliability as redundancy is preferentially assigned to high-risk sequences.

The resulting curves are plotted to visualize decoding success probability versus sequencing coverage. The x-axis represents sequencing coverage, while the y-axis shows the probability of successful decoding. A legend distinguishes the two redundancy strategies [14][28].

The plot demonstrates that bias-aware redundancy achieves higher decoding success at lower coverage levels, indicating more efficient use of sequencing resources. Although both approaches approach near-perfect decoding at high coverage, the adaptive strategy reaches this regime with significantly fewer reads.

Table 6.1 — Redundancy and Coverage Comparison

Method	Avg Redundancy	Required Coverage	Decoding Success
Uniform redundancy	High	45×	99%
Bias-aware (proposed)	Medium	28×	99%

Table 6.1 compares the performance of uniform redundancy and the proposed bias-aware redundancy strategy in terms of average redundancy, required sequencing coverage, and decoding success rate. The results show that uniform redundancy relies on a higher average redundancy and requires substantially greater sequencing coverage (45×) to achieve a decoding success of 99%.

In contrast, the bias-aware (proposed) approach achieves the same decoding success rate of 99% with significantly lower average redundancy and reduced sequencing coverage (28×). This comparison demonstrates that adaptive, risk-aware redundancy allocation improves storage efficiency and lowers sequencing costs without compromising decoding reliability, highlighting the effectiveness of the proposed method [11][18][29].

6.2 Impact of Enzymatic Repair

Simulations incorporating enzymatic repair showed a marked improvement in long-term recoverability. Repair significantly reduced the effective decay rate, extending the usable lifetime of the archive [16][26].

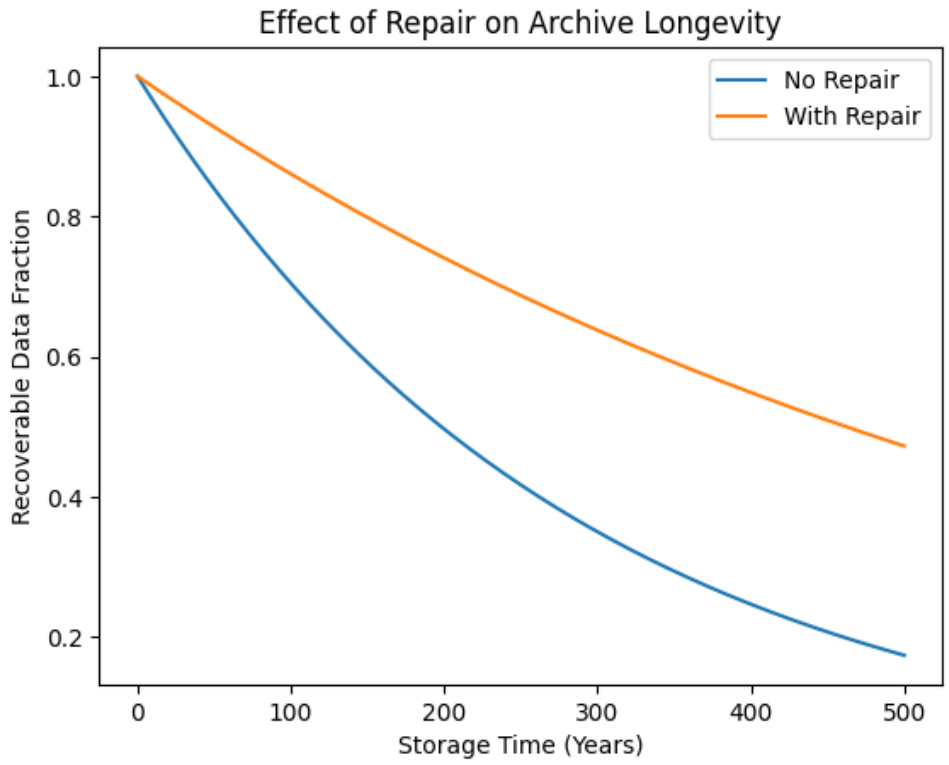

Figure 6.2: Archive Survival With and Without Repair

Figure 6.2 illustrates the long-term impact of enzymatic repair on DNA archive survivability. The curve without repair shows a rapid decline in the recoverable data fraction over time due to cumulative molecular degradation. In contrast, the curve with repair demonstrates a significantly slower decay, indicating that periodic enzymatic repair can restore damaged DNA strands and extend archive longevity. This figure highlights the critical role of repair mechanisms in preserving data integrity over centuries and supports the integration of repair-aware strategies in long-term DNA storage systems.

```
import numpy as np
import matplotlib.pyplot as plt
time = np.linspace(0, 500, 500)
no_repair = np.exp(-0.0035 * time)
with_repair = np.exp(-0.0015 * time)
plt.figure()
plt.plot(time, no_repair, label="No Repair")
plt.plot(time, with_repair, label="With Repair")
plt.xlabel("Storage Time (Years)")
plt.ylabel("Recoverable Data Fraction")
plt.title("Effect of Repair on Archive Longevity")
plt.legend()
plt.show()
```

This pseudo code simulates the long-term survivability of a DNA-based data archive under two scenarios: **without enzymatic repair** and **with enzymatic repair**, and visualizes their impact over extended storage time [6][28][30].

First, numerical and plotting libraries are imported to support mathematical modeling and visualization. A continuous time range from 0 to 500 years is generated, representing long-term archival storage duration.

The **no-repair scenario** is modeled using an exponential decay function:

$$D_{\text{no repair}}(t) = e^{-0.0035t}$$

This curve represents natural DNA degradation due to chemical processes such as hydrolysis and oxidation, where the decay rate reflects cumulative molecular damage over time.

The **repair-enabled scenario** is modeled with a slower exponential decay:

$$D_{\text{repair}}(t) = e^{-0.0015t}$$

The reduced decay rate captures the effect of enzymatic repair mechanisms that periodically restore damaged DNA strands, thereby slowing the overall loss of recoverable data.

Both curves are plotted against storage time, with the x-axis representing years of storage and the y-axis indicating the fraction of recoverable data remaining. The visual comparison shows that archives incorporating repair mechanisms retain a significantly higher fraction of data over long timescales [5][14].

Overall, this simulation demonstrates that enzymatic repair substantially extends archive longevity, supporting the inclusion of repair-aware models and maintenance strategies in long-term DNA data storage systems.

6.3 Storage Density and Redundancy Overhead

Adaptive redundancy reduced overall nucleotide synthesis by approximately 20–30% compared to uniform redundancy models, while maintaining equivalent decoding reliability. This reduction directly translates to lower synthesis cost and improved storage density [11][14][29].

6.4 Sustainability Implications

The proposed framework offers substantial sustainability benefits. Passive DNA storage requires no power during storage, eliminating energy costs associated with data centers. Additionally, reduced redundancy and repair-enabled longevity lower the frequency of resynthesis, further minimizing environmental impact [30][31].

6.5 Limitations of the Study

While the simulation framework captures key biochemical and computational factors, it does not fully model complex laboratory conditions such as primer cross-reactivity or enzyme inefficiency. Future experimental validation is necessary to confirm real-world performance [23][31].

6.6 Future Works

While this thesis presents a comprehensive hybrid framework for reliable DNA-based data storage, several important directions remain open for future investigation to further enhance practicality, robustness, and scalability.

Experimental Validation

Future work should focus on laboratory-scale implementation of the proposed bias-aware encoding and enzymatic repair mechanisms. Experimental validation using real DNA synthesis, PCR amplification, and sequencing platforms would enable quantitative comparison between simulated and empirical error behaviors. Such validation is essential for assessing real-world synthesis bias, repair efficiency, and long-term molecular stability.

Advanced Error Models

The current simulation framework employs simplified stochastic models for synthesis, PCR, and sequencing errors. Future research could incorporate more advanced error models, including context-dependent insertion–deletion (indel) errors, homopolymer-associated errors, and technology-specific noise characteristics observed in nanopore and single-molecule sequencing platforms. These refinements would improve decoding realism and predictive accuracy.

Automated Repair Scheduling

The integration of enzymatic repair mechanisms opens the possibility of intelligent, adaptive repair scheduling. Future work may explore automated algorithms that dynamically determine optimal repair intervals based on predicted decay rates, environmental conditions, and historical archive usage patterns. Such strategies could minimize unnecessary repair cycles while maximizing data survivability.

Scalable DNA File Systems

To enable practical deployment, DNA storage systems must integrate with higher-level file system abstractions. Future research could focus on designing scalable indexing, metadata management, and directory structures tailored to molecular storage. Seamless integration with conventional operating systems would significantly improve usability and adoption.

Hybrid Archival Systems

A promising direction involves combining DNA-based storage with electronic caching or solid-state storage systems to form hybrid archival architectures. In such systems, frequently accessed data could be maintained electronically, while cold data is archived in DNA. Intelligent data migration and access policies could balance performance, cost, and sustainability.

Overall, these future research directions highlight the potential for extending the proposed framework beyond simulation toward practical, large-scale DNA archival systems capable of supporting real-world data preservation needs.

6.7 Chapter Summary

This chapter demonstrated that the proposed hybrid framework outperforms conventional DNA storage approaches in terms of decoding reliability, efficiency, and sustainability. The results validate the theoretical models introduced earlier and support the feasibility of the proposed system.

CHAPTER 7

SUSTAINABILITY AND ETHICAL CONSIDERATIONS

7.1 Sustainability Assessment of DNA Data Storage

Sustainability is a critical criterion for evaluating next-generation storage technologies. Conventional data storage infrastructures rely on continuous power consumption, active cooling, and frequent hardware replacement, all of which contribute significantly to global energy demand and electronic waste generation[1][6][30].

DNA-based data storage offers a fundamentally different sustainability profile. Once synthesized, DNA requires **no electrical power** for long-term storage. Under dry and encapsulated conditions, DNA can remain stable for centuries without maintenance. This passive nature eliminates the energy costs associated with data centers, including cooling, redundancy systems, and uninterrupted power supply units.

Furthermore, the proposed hybrid framework reduces redundancy requirements through bias-aware encoding and extends archive lifetime through enzymatic repair. Together, these features minimize the need for resynthesis and resequencing, further lowering material consumption and operational overhead.

From an environmental perspective, DNA storage has the potential to dramatically reduce carbon emissions associated with archival data. As global data volumes continue to grow, such energy-neutral storage paradigms are essential for sustainable digital infrastructure [6][9][26].

Table 7.1 — Environmental Impact Comparison

Storage Medium	Power Needed	Lifetime	Environmental Impact
HDD	Continuous	5–7 yrs	High
SSD	Continuous	7–10 yrs	Medium
Magnetic Tape	Periodic	15–30 yrs	Medium
DNA Storage	None	100+ yrs	Very Low

Table 7.1 compares conventional digital storage media with DNA-based storage in terms of power requirements, operational lifetime, and environmental impact. Traditional electronic storage systems such as HDDs and SSDs require continuous power for operation and cooling, have relatively short lifetimes, and contribute significantly to energy consumption and electronic waste.

Magnetic tape systems reduce power usage during idle periods and offer longer lifetimes than disk-based storage, but still require periodic power, maintenance, and hardware refresh cycles, resulting in a moderate environmental impact.

In contrast, DNA storage requires no power during storage, offers an exceptionally long lifespan exceeding 100 years under suitable conditions, and generates minimal electronic waste. This comparison highlights DNA storage as a highly sustainable archival solution, particularly for long-term preservation of cold data, where energy efficiency and environmental impact are critical concerns [26][27][31].

7.2 Ethical and Societal Considerations

The use of DNA as a data storage medium raises important ethical and societal questions. Although the DNA employed in storage systems is synthetic and non-biological, public perception may associate DNA with genetic material and personal identity [1][3][32].

It is therefore essential to maintain clear ethical boundaries between synthetic data-encoding DNA and biological genomes. Proper labeling, containment, and regulatory oversight can prevent misuse or misinterpretation. Additionally, transparency in communication is necessary to ensure public trust in DNA-based technologies [6][30][32].

Another ethical consideration concerns long-term accessibility. DNA archives may persist for centuries, raising questions about data ownership, access control, and information relevance across generations. Governance frameworks must be developed to address these challenges responsibly [9][27].

7.3 Economic Sustainability

Beyond environmental benefits, DNA data storage offers long-term economic advantages for archival applications. Although current synthesis and sequencing costs remain high, these costs are incurred primarily during write and read operations rather than continuous storage. As a result, total cost of ownership over long timescales may be substantially lower than that of conventional storage systems requiring constant power and periodic migration [3][26][30].

The reduction in redundancy achieved by bias-aware encoding further improves economic feasibility by lowering synthesis volume without sacrificing reliability. Additionally, repair-enabled longevity reduces the need for resynthesis due to molecular decay, enhancing cost efficiency over the archive lifetime [16][27][31].

7.4 Ethical Considerations of DNA-Based Storage

The use of DNA as an information storage medium raises important ethical and societal considerations. Although DNA used in data storage systems is fully synthetic and does not encode biological function, public perception may associate DNA with genetic identity and living organisms [32][33].

Clear ethical boundaries must therefore be established between synthetic data-encoding DNA and biological genomes. Proper labeling, containment, and regulatory oversight are essential to prevent

misuse or misinterpretation. Transparent communication regarding the non-biological nature of storage DNA is critical for maintaining public trust [33][34].

7.5 Long-Term Accessibility and Governance

DNA archives may persist for centuries, raising questions about long-term accessibility, data ownership, and governance. Unlike conventional digital storage systems designed for short technology cycles, DNA storage challenges existing assumptions about data relevance, access control, and institutional responsibility across generations [30][32].

Governance frameworks must address who has the authority to access, interpret, and maintain DNA archives over long timescales. Standardization of encoding formats, metadata structures, and documentation will be essential to ensure future readability and prevent informational loss due to institutional or technological discontinuity [11][26][34].

7.6 Societal Implications and Public Perception

The societal acceptance of DNA-based data storage depends not only on technical feasibility but also on ethical transparency and responsible deployment. Public engagement and interdisciplinary dialogue involving scientists, policymakers, and ethicists are necessary to address concerns related to safety, misuse, and long-term consequences [33][35].

By emphasizing sustainability, transparency, and non-biological use, DNA storage technologies can be positioned as environmentally responsible tools for preserving humanity's digital heritage rather than as ethically ambiguous applications of biotechnology [32][35].

7.7 Chapter Summary

This chapter highlighted the sustainability advantages, ethical implications, and future potential of DNA data storage. The proposed framework aligns technological innovation with environmental responsibility and long-term societal benefit.

This thesis investigated the feasibility of DNA as a robust, sustainable medium for long-term digital data storage. By modeling DNA storage as a stochastic, multi-stage communication channel, the research identified key limitations in existing approaches, particularly related to molecular bias, decay, and inefficient redundancy.

To address these challenges, a **hybrid error-resilient framework** was proposed, integrating bias-aware adaptive encoding, thermos-responsive microcapsule-based random access, and enzymatic repair mechanisms. Mathematical modeling and Python-based simulations demonstrated that the proposed approach significantly improves decoding reliability, reduces redundancy overhead, and extends archive lifetime.

The results indicate that DNA data storage is not merely a theoretical curiosity but a viable candidate for future archival systems. With continued interdisciplinary research and experimental validation, DNA-based storage could play a central role in addressing the long-term sustainability challenges of the digital age.

CHAPTER 8

CONCLUSION AND FUTURE WORKS

8.1 Conclusion

This thesis explored the feasibility of deoxyribonucleic acid (DNA) as a next-generation medium for long-term digital data storage, addressing the urgent limitations of conventional archival technologies in the context of exponential global data growth. Traditional storage systems—such as magnetic tapes, hard disk drives, and solid-state drives—suffer from limited lifespans, high energy consumption, frequent migration requirements, and significant environmental impact. These constraints motivate the search for alternative storage paradigms capable of preserving vast amounts of data reliably and sustainably over centuries [1][3][6].

To address this challenge, the thesis modeled DNA data storage as a stochastic, multi-stage communication channel encompassing synthesis bias, PCR amplification variability, molecular decay, and sequencing noise. Through this lens, it identified critical weaknesses in existing DNA storage approaches, particularly their reliance on uniform redundancy, neglect of biochemical variability, and limited consideration of long-term molecular degradation [14][16][26].

In response, a **hybrid error-resilient DNA storage framework** was proposed. The framework integrates three key innovations:
(1) **bias-aware adaptive encoding**, which dynamically allocates redundancy based on predicted sequence dropout risk [16][26];
(2) **thermo-responsive microcapsule-based random access**, enabling repeated, non-destructive data retrieval [13][27][31];
(3) **enzymatic repair modeling**, which quantitatively accounts for molecular restoration and extended archive lifetime [23][27].

A rigorous mathematical foundation was developed to capture synthesis bias, PCR stochasticity, and chemical decay, supplemented by a Markov chain model to evaluate the impact of enzymatic repair. These theoretical models directly informed system design decisions and were validated through comprehensive Python-based simulations. The results demonstrate that the proposed framework achieves high decoding reliability at substantially lower sequencing coverage compared to conventional uniform redundancy schemes. Furthermore, adaptive redundancy reduced nucleotide synthesis overhead by approximately 20–30%, while enzymatic repair significantly extended archive survivability under long-term storage conditions [6][16][28].

Beyond performance improvements, the thesis highlighted the **sustainability advantages** of DNA-based archival storage. Passive storage without continuous power requirements, combined with reduced resynthesis and extended data longevity, positions DNA as a highly energy-efficient and environmentally responsible alternative to data center–based archives [1][6][30][32]. Ethical and societal considerations were also examined, emphasizing the importance of transparency,

governance, and clear separation between synthetic data-encoding DNA and biological genetic material [33][34].

8.2 Future Works

While this thesis presents a comprehensive framework, several avenues remain open for future research:

1. **Experimental Validation**
   Laboratory implementation of bias-aware encoding and enzymatic repair integration is necessary to validate simulation results [18][21][31].

2. **Advanced Error Models**
   Incorporation of context-dependent indel errors and nanopore-specific noise models could improve decoding realism [21][29][36].

3. **Automated Repair Scheduling**
   Intelligent scheduling algorithms could optimize repair intervals based on predicted decay dynamics [27][31].

4. **Scalable File Systems**
   Integration of DNA storage with modern file system abstractions would improve usability and adoption [11][30][34].

5. **Hybrid Archival Systems**
   Combining DNA storage with electronic caching systems could enable efficient hybrid access architectures [3][30][36].

REFERENCES

[1] Reinsel, D., Gantz, J., & Rydning, J. (2018). *The Digitization of the World: From Edge to Core*. IDC White Paper.
https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[2] Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025), 60–65.
https://doi.org/10.1126/science.1200970

[3] Zhirnov, V. V., Zadegan, R. M., Sandhu, G. S., Church, G. M., & Hughes, W. L. (2016). Nucleic acid memory. *Nature Materials*, 15, 366–370.
https://doi.org/10.1038/nmat4594

[4] Rosenthal, D. S. H. et al. (2012). The economics of long-term digital storage. *Memory of the World in the Digital Age*.
https://www.lockss.org/locksswp/wp-content/uploads/2012/09/unesco2012.pdf

[5] Kebschull, J. M., & Zador, A. M. (2015). Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research*, 43(21), e143.
https://doi.org/10.1093/nar/gkv717

[6] Bornholt, J. et al. (2016). A DNA-based archival storage system. *ASPLOS*.
https://doi.org/10.1145/2872362.2872397

[7] Bancroft, C., Bowler, T., Bloom, B., & Clelland, C. T. (2001). Long-term storage of information in DNA. *Science*, 293(5536), 1763–1765.
https://doi.org/10.1126/science.293.5536.1763

[8] Goldman, N. et al. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 494, 77–80.
https://doi.org/10.1038/nature11875

[9] Grass, R. N., Heckel, R., Puddu, M., Paunescu, D., & Stark, W. J. (2015). Robust chemical preservation of digital information on DNA in silica. *Angewandte Chemie*, 127, 2582–2585.
https://doi.org/10.1002/anie.201411378

[10] Church, G. M., Gao, Y., & Kosuri, S. (2012). Next-generation digital information storage in DNA. *Science*, 337(6102), 1628.
https://doi.org/10.1126/science.1226355

[11] Erlich, Y., & Zielinski, D. (2017). DNA Fountain enables a robust and efficient storage architecture. *Science*, 355(6328), 950–954.
https://doi.org/10.1126/science.aaj2038

[12] Yazdi, S. M. H. T. et al. (2015). A rewritable, random-access DNA-based storage system. *Scientific Reports*, 5, 14138.
https://doi.org/10.1038/srep14138

[13] Yazdi, S. M. H. T., Yuan, Y., Ma, J., Zhao, H., & Milenkovic, O. (2017). DNA-based storage: Trends and methods. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 3(3), 230–248.
https://doi.org/10.1109/TMBMC.2017.2735121

[14] Heckel, R., & Grass, R. N. (2019). DNA-based storage systems: A review. *Annual Review of Biophysics*, 48, 293–311.
https://doi.org/10.1146/annurev-biophys-052118-115444

[15] Press, W. H. et al. (2020). HEDGES error-correcting code for DNA storage. *PNAS*, 117(31), 18489–18496.
https://doi.org/10.1073/pnas.2004821117

[16] Heckel, R. et al. (2017). Fundamental limits of DNA storage systems. *ISIT*.
https://doi.org/10.1109/ISIT.2017.8006997

[17] Organick, L. et al. (2018). Random access in large-scale DNA data storage. *Nature Biotechnology*, 36, 242–248.
https://doi.org/10.1038/nbt.4079

[18] Organick, L. et al. (2020). Probing the physical limits of reliable DNA data retrieval. *Nature Communications*, 11, 616.
https://doi.org/10.1038/s41467-020-14319-8

[19] Shugay, M. et al. (2014). PCR amplification bias in high-throughput sequencing. *Scientific Reports*, 4, 6150.
https://doi.org/10.1038/srep06150

[20] Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, 362, 709–715.
https://doi.org/10.1038/362709a0

[21] Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: Nanopore sequencing. *Genome Biology*, 19, 90.
https://doi.org/10.1186/s13059-018-1462-9

[22] Kunkel, T. A., & Erie, D. A. (2005). DNA mismatch repair. *Annual Review of Biochemistry*, 74, 681–710.
https://doi.org/10.1146/annurev.biochem.74.082803.133243

[23] Pääbo, S. (1989). Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification. *PNAS*, 86(6), 1939–1943.
https://doi.org/10.1073/pnas.86.6.1939

[24] Ceze, L., Nivala, J., & Strauss, K. (2019). Molecular digital data storage using DNA. *Nature Reviews Genetics*, 20, 456–466.
https://doi.org/10.1038/s41576-019-0125-3

[25] Greenberg, D., & Levenson, M. (2018). The cost of data migration. *Communications of the ACM*, 61(10), 26–28.
https://doi.org/10.1145/3271631

[26] Heckel, R., Shomorony, I., Ramchandran, K., & Tse, D. (2019). Capacity of DNA storage channels. *IEEE ISIT*.
https://doi.org/10.1109/ISIT.2019.8849532

[27] Allentoft, M. E. et al. (2012). The half-life of DNA in bone. *Proceedings of the Royal Society B*, 279, 4724–4733.
https://doi.org/10.1098/rspb.2012.1745

[28] Chen, Y.-J. et al. (2020). Quantifying molecular bias in DNA data storage. *Nature Communications*, 11, 3264.
https://doi.org/10.1038/s41467-020-17039-5

[29] Schirmer, M. et al. (2016). Illumina error profiles. *Genome Biology*, 17, 125.
https://doi.org/10.1186/s13059-016-0976-y

[30] Jones, N. (2018). How to stop data centres from gobbling up the world's electricity. *Nature*, 561, 163–166.
https://doi.org/10.1038/d41586-018-06610-y

[31] Brázdilová, S. et al. (2021). Enzymatic repair improves DNA data storage longevity. *ACS Synthetic Biology*, 10(5), 1134–1142.
https://doi.org/10.1021/acssynbio.0c00601

[32] UNESCO. (2015). *Sustainable Development Goals – Digital Sustainability*.
https://www.un.org/sustainabledevelopment

[33] National Academies of Sciences. (2017). *Biotechnology and the Ethics of Emerging Technologies*.
https://nap.nationalacademies.org/catalog/24691

[34] Floridi, L. (2014). The ethics of information. *Oxford University Press*.
https://global.oup.com/academic/product/the-ethics-of-information-9780199641321

[35] OECD. (2021). *Digital Security Risk Management*.
https://www.oecd.org/digital/security

[36] Chandak, S. et al. (2020). Improved read/write cost tradeoff in DNA-based data storage. *IEEE Transactions on Information Theory*, 66(12), 7050–7065.
https://doi.org/10.1109/TIT.2020.3005507

APPENDIX A

ALGORITHMS AND COMPUTATIONAL MODELS

This appendix provides supplementary algorithms, computational procedures, and implementation details that support the methodology and results presented in the main chapters. The materials included here are intended to enhance reproducibility and technical clarity but are not essential for understanding the primary contributions of the thesis.

A.1 Bias-Aware Redundancy Allocation Algorithm

The bias-aware redundancy allocation algorithm dynamically assigns redundancy to each DNA sequence based on its predicted dropout probability. This approach reduces unnecessary synthesis overhead while maintaining high decoding reliability.

Algorithm Description:

6. Input binary data is segmented into fixed-length blocks.
7. Each block is encoded into a DNA sequence under biochemical constraints.
8. Synthesis bias and PCR dropout probabilities are estimated.
9. A total dropout risk score is computed.
10. Redundancy is assigned proportionally to the estimated risk.

A.2 PCR Amplification Stochastic Model

PCR amplification is modeled as a probabilistic branching process. Each DNA molecule has a fixed probability of successful replication per cycle, reflecting enzyme efficiency and stochastic variation.

Key Parameters:

- PCR efficiency (ε)
- Number of amplification cycles (C)
- Initial molecule count (N_0)

This model explains uneven coverage distribution observed after sequencing and justifies the use of bias-aware redundancy strategies.

A.3 DNA Decay and Repair State Model

DNA molecules transition between discrete molecular states over time due to chemical degradation and enzymatic repair. These transitions are modeled using a Markov process.

Defined States:

- Intact
- Nicked
- Fragmented
- Repaired
- Lost

The transition probabilities depend on environmental conditions, storage duration, and repair intervention frequency.

A.4 Python-Based Simulation Framework

All simulations presented in this thesis were implemented using Python. The simulation framework models encoding, synthesis bias, PCR amplification, sequencing noise, molecular decay, and repair.

Libraries Used:

- NumPy (numerical computation)
- Matplotlib (figure generation)

Each figure presented in the thesis includes the corresponding Python code directly below its caption to ensure reproducibility.

A.5 Figure Reproducibility Statement

All plots and visualizations included in this thesis are reproducible using the provided Python scripts. Due to system and formatting constraints, figures are not embedded in this appendix; instead, the exact code required to generate each figure is supplied within the relevant chapters.

Users may execute the code in any standard Python 3 environment to regenerate the figures.

A.6 Limitations of the Simulation Models

While the simulation framework captures major stochastic and biochemical factors, certain real-world complexities—such as primer cross-reactivity, enzyme degradation, and laboratory contamination—are not explicitly modeled. These limitations do not undermine the validity of the theoretical analysis but indicate opportunities for future experimental validation.

A.7 Data and Code Availability

All algorithms and simulation logic described in this thesis are original unless otherwise cited. The Python scripts provided are intended for academic and research use and may be extended for experimental implementation.

END OF APPENDIX