

Massively Parallel Homogeneous Amplification of Chip-Scale DNA for DNA Information Storage (MPHAC-DIS)

Corresponding Author: Professor Ping Song

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This paper brings the concept of designing PCR primers with uniform Gibbs free energy changes to enhance the homogeneity in amplification, therefore guaranteeing unbiased information retrieval in DNA data storage. This “FE-design concept” is important and timely as prediction of biochemical reaction efficiency with considerations from the energy perspective is becoming a trend. Also, I appreciate the authors’ thorough characterization and detailed analysis of the constructed DNA data storage system. Overall, this work is well presented and is of certain scientific value. However, there are some points I found not clear or requiring more supportive evidence. Please see below:

1. Page 6 of SI. In the paragraph about GC content on PCR efficiency. The authors discussed that “high GC content (>65%) can affect the efficiency of PCR due to the tendency of these sequences to fold into complex secondary structures”. However, in their simulation results shown in Figure S2-1b, it seems a high GC content had zero impact on PCR reaction as the efficiencies all reached 100% when GC content was higher than 50%. In addition, “GC contents constrained within 45% to 55%” in this simulation experiment should only contain the case when GC content was exactly 50%. However, it is confusing to me that in Figure S2-1c, the distribution of simulated efficiency for “with constraint” group was not completely consistent with what was displayed in Figure S2-1b when GC ratio was 50%.
2. In “primer generation” of the “Methods” section, the secondary structure screening solely relied on MFE structure. However, in some cases the MFE frequency may be quite low for a given sequence, which means only a small fraction is predicted to be in MFE structure. It would be good if the authors could also evaluate the impact of MFE frequency (even if the MFE structure is fully unfolded) on PCR efficiency.
3. Page 6 of SI, Section of “primer length”. The authors stated “Longer primers can decrease the likelihood of encountering low amplification efficiency primers”. Is it supported by experiments? Please include experimental results or any references.
4. I recommend the authors to use “Gibbs free energy change” instead of “Gibbs free energy” to represent ΔG throughout the article.
5. Please reconcile the parameters of free energy change filter throughout the article. In Page 9 of SI it is $-10.5 \text{ kcal mol}^{-1} \sim -12.5 \text{ kcal mol}^{-1}$. However, in the Methods section (Line 309) it is $-11.5 \text{ kcal mol}^{-1} \sim -12.5 \text{ kcal mol}^{-1}$ and in the results section (Line 86) it is $-10 \text{ kcal mol}^{-1} \sim -12.5 \text{ kcal mol}^{-1}$.
6. In Line 97, the authors mentioned “various primer sets could be mixed as needed to realize random access to individual or multiple files”. Multiplex PCR requires more careful design of primer sets to ensure the efficiency. In Figure S2-7 and Figure 3b, the authors have evaluated any possible cross-hybridization between primers from different sets and the variation of primer specificity with free energy change. However, in their primer design algorithm, the specificity check is missing, and I suppose the authors screen the specificity entirely by controlling free energy changes, which makes the specificity check less persuasive.
7. The authors may also need to evaluate the specificity of primers to ensure that no primers will target an unintended template, and in their primer design algorithm, an additional screening step may need to be carried out.
8. Was the simulated sequencing experiment described in Lines 121-126 based on the simulated amplification yield? I suggest the authors to specify the detailed procedure of the simulated sequencing experiment in the Methods section.
9. In the Methods section, Line 337. How was the amplification efficiency evaluated for a non-target? Has the association and dissociation rate changed compared to on-target amplification? The authors may need to specify this in the same paragraph.
10. Please specify the meaning of dashed lines in Figure S5-4. Are these indicating the threshold PCR cycles?
11. In Figure S5-4 and Figure S5-5, I am still not clear whether the threshold PCR cycle correlates more with the file size ratio or the highest initial concentration of a file. Besides, the authors discussed the situation of amplifying two files with varying

file size ratios, how about multiple files, as in reality many files with different sizes may need to be amplified simultaneously?

12. Figure S5-6. Since the amplified files will be subjected to sequencing to read out the data anyway, why not perform the adapter PCR as the first step to cover the “first PCR”? If so, to avoid bias in sequencing reads or even loss of data, the authors may also need to include sequencing adapters in the multiple screening steps of their PCR primer design, especially the dimer formation, secondary structures and specificity. Also, please correct “adaptor” to “adapter” in the figure.

13. In Figure S5-7. Is the threshold read count fixed or flexible depending on actual experimental conditions?

14. In Figure 4a, given that the last five files are in text format, how would the decoded files also be represented in pictures? Also, how could the sequencing coverage be represented in a 2D rectangle as the picture files did?

15. Figure S8-2. The authors analyzed the effect of GC content on sequencing errors. This metric represented the GC content of the whole sequence of a strand. Did the authors also evaluate GC-rich regions (narrow down to 10-20 nt) of a sequence in particular to see whether sequencing errors are more prone to these regions? Compared to using the overall GC content, analysis with local GC content may be more helpful to demonstrate the effect of GC content on sequencing errors.

16. SI section 8, “Effect of sequence secondary structure”. The examples the authors used to evaluate the effect of secondary structures all contained repetitive sequences, which may cause difficulties in algorithmic sequence assembly and produce sequencing errors. This makes the described evaluation not adequate to identify whether the sequencing errors originated from secondary structures or repetitive sequences. On the other hand, the secondary structures in the provided examples were actually not strong, given the base pairing probabilities of the hybridized regions not significantly high (around 50%). The authors may need to provide more examples excluding repetitive sequences but solely containing stable secondary structures to further evaluate the effect of secondary structure on sequencing errors. Lastly, for the subtitle “b) Secondary structure between payload region and primers”, please use “primer binding region” in place of “primers” to avoid misunderstanding.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

The study presents a promising approach to primer selection for DNA amplification that could be instrumental in large-scale information storage. Among the methods for DNA synthesis, chip-based synthesis is particularly suited for large-scale storage as it enables the simultaneous synthesis of a wide variety of oligonucleotides. However, this approach yields low initial concentrations of each oligonucleotide type, necessitating the use of PCR amplification for decoding. The authors highlight that conventional primer design strategies, which rely on fixed-length primers, result in a wide range of primer-template hybridization energies, leading to uneven amplification of different DNA sequences. To address this issue, the authors propose a fixed-energy primer design approach that avoids fixed primer lengths and instead optimizes hybridization energy for uniform amplification.

As the authors state, previous studies have already considered the energy of primer hybridization or related parameters, such as melting temperature (e.g., doi:10.1038/nbt.4079). However, this study provides a comprehensive comparison of the fixed-length versus fixed-energy approaches, emphasizing and explaining the necessity for such design. Through theoretical modeling, the authors determine the optimal hybridization energies that balance both high specificity and hybridization efficiency. These findings are substantiated by experimental data, demonstrating that fixed-energy primers enable uniform amplification during PCR and achieve high accuracy in encoding and decoding various types of data. Despite the comprehensive design of the study, and promise of the fixed-energy primer design for DNA storage systems, several concerns need to be addressed:

1. My main concern of the proposed fixed-energy approach is that it uses not factual energy but predicted energy based on a theoretical model, which is far from ideal when it comes to experiments. This way any number of other theoretical metrics can be devised with varying results and the used approach does not seem as the optimum. For example, using a neural network to design primers based on considerable experimental data may give much better results. While I greatly appreciate the thoroughness of the study that involved a huge amount of work, I am not sure that the manuscript clearly defines why the study conceptually presents a significant inventive step over the previous works.

2. The results of the fixed-energy approach should be directly compared with those obtained using existing methods to highlight the significance of the research. For instance, the study by Lee Organick et al. (Nat. Biotechnol. 2018) employs similar filters for primer design based on melting temperature rather than hybridization energy, and also results in accurate data decoding.

3. The study does not present any statistical analyses to evaluate the significance of the results, which should be included for quantitative comparisons of different groups.

4. To prove that homogenous amplification is independent of initial template concentrations, PCR experiments should be conducted using various initial concentrations of the plasmid.

5. Even a library of 16,095 primers seems insufficient for storing significant amounts of data. This limitation, along with challenges related to inaccurate oligo synthesis, suggests that DNA information storage is still more of a concept than a viable storage system. Therefore, the authors should discuss future plans and strategies to address these challenges.

To improve comprehension of the material, several adjustments are recommended:

5. Key information, such as the encoding-decoding strategy, should be incorporated into the main text to ensure clarity for readers.

6. Terms like “fold-80 base penalty” and “Ct-value” should be defined to make the content accessible to a broader audience.

7. The term “primer” is inconsistently used to refer to both a segment of a template and the primer itself. These should be clearly differentiated as reverse complementary sequences to avoid confusion.

8. P.6 “Given the typically low initial concentration of the template during actual PCR, we simulated the hybridization yield of

primers when $A \gg B$." The designations "A" and "B" are not clearly identified in the text.

9. P.5 "Based on the theoretical model, we explored amplification efficiency and specificity under different primer free energies." Theoretical model should be supplemented with a brief description or at least a reference to the supplementary information.

10. The description of Fig. 5e seems to have mixed-up labels (decoding images and restored images)

Several points also require further clarification:

11. Table S6-2 lists various libraries without describing the differences between similar ones (e.g., lib1 vs. lib7; lib6 vs. lib8).

12. The detailed description of the image recovery algorithm is missing and should be included to provide a complete understanding of the methodology used.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

Summary and recommendation: This manuscript introduces MPHAC-DIS, a novel strategy for DNA-based information storage that utilizes simulation-guided primer design to achieve unbiased and efficient amplification. The authors demonstrate remarkable results, showcasing simultaneous access to diverse multimedia files with high accuracy even at very low sequencing depths. The study is well-presented and comprehensive, making it accessible to a broad audience, including those interested in DNA storage and DNA design. I recommend minor revisions before accepting this paper in Nature Communications.

Specific Changes/Suggestions for Revisions:

- For the experimental validation of the strategy: a detailed error rate analysis and long-term stability, particularly at low sequencing depths, is valuable to build a better understanding of the types and sources of errors will aid in developing error-correction strategies. The authors have those sections buried in SI, they should consider incorporating those discussions from the Supplementary Information into the main manuscript, or within the methodology section.
 - Additional storage condition testing: While the current study evaluates buffer variations, it'll be important to expand testing or discussion to explore the system's performance under different temperatures, assess the impact of varying humidity levels on DNA stability, evaluate the effects of UV light exposure, particularly for long-term storage scenarios, and the potential of dry storage conditions, as explored in this example <https://www.nature.com/articles/s41467-019-09517-y>. Note that I am not affiliated with this study, but it's worth considering how extreme storage conditions could impact the DNA storage system's performance and data integrity. Evaluating/discussing its resilience under such conditions is essential for real-world applications.
 - Benchmarking/Comparative analysis: A significant drawback of the paper is the comparison with only one prior method (fixed 20nt primers), which the authors acknowledge isn't the only approach used in the past. Can the authors expand the analysis to include a wider range of existing primer design strategies, including those that consider factors like T_m (melting temperature). This would provide a more comprehensive understanding of its advantages and limitations in terms of storage density, retrieval accuracy, cost, and scalability (providing a comparative table like Table S1 including all these variables would help).
 - Results Figure 3 and 4: They are very comprehensive and clear but consider adding the coefficient of variance (CV) to plot (d) and the curve in section (i), it would help quantify the variance between the two methods. Similarly, consider adding them to the histograms in Figure 4e and other relevant figures.
- Also, could the " $A \gg B$ " and " $A=B$ " cases in section (h) be experimentally validated and compared to the theoretical predictions?
- The Supplementary Information file is extensive and, while comprehensive, feels like a separate paper. It contains multiple discussion points that could be integrated into the main manuscript, as noted above. The reference to 5 figures (S2-6) on line 94 is an example of how 5 figures are lumped into one reference point and could be confusing and overwhelming. This large amount of data in the SI could be streamlined for clarity, ensuring each figure conveys a distinct idea. In many instances, the text refers to sections within the SI instead of specific figures. This is unclear for the reader and contributes to the perception of the SI as a separate publication rather than supporting material.
 - Expand the discussion on the limitations of the method (where does it break?) and potential areas for future improvement.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I appreciate the authors for their careful revision. They have addressed most of my concerns and I believe this work is ready for publication.

Additional minor suggestions:

1. For the original Question 14 regarding the creation of 2D rectangle for text files, please specify the methods in the main text, preferably in the Methods section.
2. There is a small mistake in Figure S2-6 of the revised version, where the top tick label on the Y-axis should probably be

1000000.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

The authors have made substantial improvements to the manuscript, effectively addressing the key concerns raised during the initial review. The revised version is now clearer and more cohesive, with enhanced explanations and a more streamlined presentation of the methodology. Importantly, the authors have provided more context for their findings, which makes the manuscript more compelling and positions it as a significant contribution to the field.

However, one aspect still requires further clarification. While an additional check for primer specificity has been introduced to ensure orthogonality between primers, the possible occurrence of partially complementary sequences within the oligonucleotide outside the primer-binding region remains a concern. The authors should discuss how it might impact the uniformity of amplification or propose strategies to mitigate this issue. Once this discussion is addressed, the manuscript can be recommended for publication.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

The authors have carefully considered and addressed all the points raised. The manuscript has undergone significant changes, with improvements in clarity, organization, and scientific rigor. Overall, I think the revisions made the article suitable for publication with a great impact on the field and a compelling case for the use of MPHAC-DIS in DNA-based information storage. I recommend this manuscript for publication without further changes.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain

permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

REVIEWER COMMENTS

Response to reviewers' comments:

We thank the reviewers for their careful reading of our manuscript and highly appreciate their constructive remarks and suggestions. We have considered their feedback to improve the manuscript and included a detailed point-by-point response to their comments below. All changes to the text are highlighted in red in the revised manuscript.

Responses to Reviewer #1:

This paper bring the concept of designing PCR primers with uniform Gibbs free energy changes to enhance the homogeneity in amplification, therefore guaranteeing unbiased information retrieval in DNA data storage. This “FE-design concept” is important and timely as prediction of biochemical reaction efficiency with considerations from the energy perspective is becoming a trend. Also, I appreciate the authors' thorough characterization and detailed analysis of the constructed DNA data storage system. Overall, this work is well presented and is of certain scientific value. However, there are some points I found not clear or requiring more supportive evidence. Please see below:

Response: Thank you for your positive feedback on our work. We appreciate your thorough review and constructive suggestions. In the following point-by-point responses, the response content is in regular font size, while the article content is in a slightly smaller font size. The original text is in **blue**, and the modified text is highlighted in **red**.

1. Page 6 of SI. In the paragraph about GC content on PCR efficiency. 1) The authors discussed that “high GC content (>65%) can affect the efficiency of PCR due to the tendency of these sequences to fold into complex secondary structures”. However, in their simulation results shown in Figure S2-1b, it seems a high GC content had zero impact on PCR reaction as the efficiencies all reached 100% when GC content was higher than 50%. 2) In addition, “GC contents constrained within 45% to 55%” in this simulation experiment should only contain the case when GC content was exactly 50%. However, it is confusing to me that in Figure S2-1c, the distribution of simulated efficiency for “with constraint” group was not completely consistent with what was displayed in Figure S2-1b when GC ratio was 50%.

Response: Thank you for your comments. 1) The statement on Page 6 of SI about high GC content affecting PCR efficiency refers to the potential reduction in polymerase extension efficiency and the increased likelihood of complex secondary structure formation in sequences with high GC content and the impact can be mitigated by adjusting the buffer, such as using denaturing agents or commercial high-GC amplification buffers. To ensure the quantifiability in our model, we focused solely on the thermodynamics and kinetics of nucleic acid hybridization. As a result, the efficiency in this model is primarily determined by free energy. For sequences of fixed length, a higher GC content results in more negative free energy, which explains why our results show nearly 100% efficiency for high GC content sequences. In addition, we did not account for the effects of high GC content at this stage, but we incorporated a GC content filtering step in the final primer generation procedure.

2) In Figure S2-1b, we interpolated the data in 10% increments, randomly generating 100 sets of primers for each GC content level. However, in Figure S2-1c, we randomly generated a new set of 1,000 primers. The “with constraint” group includes sequences with GC contents of 45% (9 G/C, 11 A/T), 50% (10 G/C, 10 A/T), and 55% (11 G/C, 9 A/T), which explains the differences between Figure S2-1b and Figure S2-1c. To avoid confusion, we have revised the original text, and here is the revised version:

GC content. The GC content represents the proportion of guanine and cytosine bases in the primer sequence. DNA templates with high GC content (>65%) can affect PCR efficiency due to the reduced polymerase extension efficiency and the increased likelihood of forming complex secondary structures, which can hinder amplification (while we did not account for this factor in our simulations, we controlled the GC content in further steps to mitigate these issues). Moreover, the GC content also influences the distribution of primer binding free energy. For instance, at a fixed primer length of 20 nt, the free energy distributions vary significantly at different GC contents, ranging from approximately 0 kcal/mol to -24 kcal/mol (Fig. S2-1a). Correspondingly, these diverse free energy distributions result in distinct PCR efficiencies (Fig. S2-1b). By constraining the GC content within the range of 45% to 55%, the amplification efficiencies of the primers are substantially improved in comparison to when they are unconstrained, with a majority of primers exhibiting high efficiency levels around 100% (Fig. S2-1c).

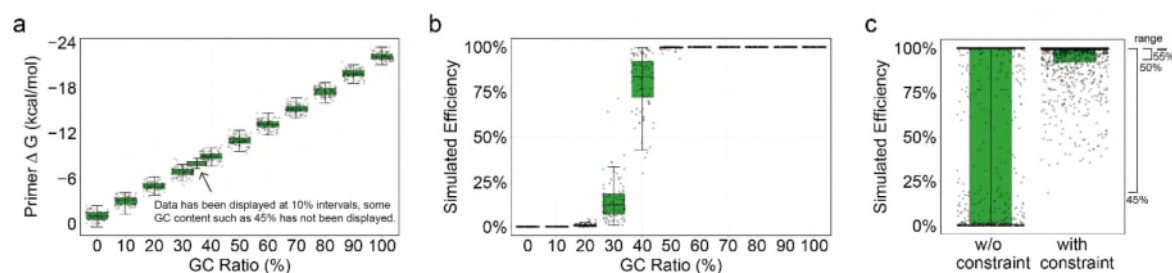


Figure S2-1. Free energy and efficiency distribution of 20-nt primers at different GC contents. (a) Distribution of free energy for primers with a length of 20 nt at various GC contents. 100 sequences were randomly generated for calculation at each GC content. Calculations were conducted at a temperature of 63°C and a Na^+ salt concentration of 0.22 M. (b) Simulated amplification efficiencies of primers at different GC contents, with 100 sequences generated for each GC content in 10% increments. (c) Comparison of the distribution of amplification efficiency between primers with GC contents constrained to 45%-55% (45%, 50%, and 55%) versus primers with unconstrained GC contents. In this panel, 1,000 primers were regenerated for each condition.

2. In “primer generation” of the “Methods” section, the secondary structure screening solely relied on MFE structure. However, in some cases the MFE frequency may be quite low for a given sequence, which means only a small fraction is predicted to be in MFE structure. It would be good if the authors could also evaluate the impact of MFE frequency (even if the MFE structure is fully unfolded) on PCR efficiency.

Response: Thank you for your professional suggestion. It is indeed true that for some sequences, the MFE structure represents only a small fraction. Incorporating MFE frequency into the screening process could potentially increase the number of sequences passing the filter (since sequences with secondary structures in MFE may mostly remain in a single-stranded state). However, in our actual filtering, given that primers are typically short, we observed that the current secondary structure filter does not eliminate a significant number of sequences (only a 4% reduction, from 16,791 to 16,095). Additionally, the MFE frequency of most of the filtered sequences was above 0.5 (Fig. R1). Finally, considering the MFE frequency would require additional computational resources, which would slow down the process. Weighing these factors, we finally chose to use the current secondary structure filter, which preserves the majority of sequences while maintaining an efficient runtime.

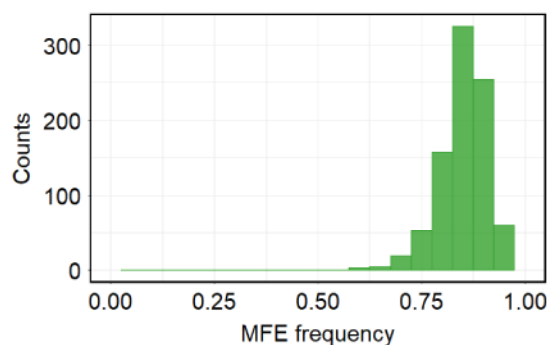


Figure R1. Distribution of MFE frequency of the filtered sequences.

3. Page 6 of SI, Section of “primer length”. The authors stated “Longer primers can decrease the likelihood of encountering low amplification efficiency primers”. Is it supported by experiments? Please include experimental results or any references.

Response: We apologize for the lack of clarity in the original text. The intention behind our statement was to emphasize that longer primers can reduce the likelihood of low amplification efficiency due to free energy change (ΔG). When primers are lengthened, the ΔG becomes more negative, improving the binding stability and thus enhancing amplification efficiency. However, once ΔG reaches a certain threshold, the amplification yield approaches saturation (close to 100%). Beyond this point, further increasing the primer length may actually increase the risk of primer dimers (Fig. S2-4) and secondary structure formation (Fig. S2-5), which in turn could decrease amplification efficiency¹.

We have revised the original text, and the updated version is provided below:

Primer length is a key indicator of primer design, and primers that are too short may result in inadequate efficiency. While 20 nt is a commonly used length, our previous results indicate that even with constrained GC content, certain primers still exhibit poor amplification efficiency (Fig. S2-1c). Longer primers can make ΔG more negative and provide stronger binding. Figure S2-2 illustrates the distribution of free energy for various primer lengths. Among them, the distribution of energy will be more concentrated when there is a GC content constraint. However, excessive length increases the likelihood of primer dimer formation (discussed later), and this is one of the primary reasons why we do not recommend fixed-length primer design.

Reference:

[1]. Hendling, M. and I. Barisic, In-silico Design of DNA Oligonucleotides: Challenges and Approaches. Comput. Struct. Biotechnol. J., 2019. 17: p. 1056-1065.

4. I recommend the authors to use “Gibbs free energy change” instead of “Gibbs free energy” to represent ΔG throughout the article.

Response: Thank you for your suggestion. We have changed the full name of ΔG from “Gibbs free energy” to “Gibbs free energy change”. The updated version is provided:

The binding free energies of programmable probes have been widely studied to achieve the binding specificity and sensitivity. Here, we introduce binding energy controlled massively parallel homogeneous amplification of chip-scale DNA (MPHAC), which ensures highly specific primer hybridization with homogeneous standard Gibbs free energy change (ΔG°), enabling high-throughput, bias-free and selective amplification of target sequences by PCR.

5. Please reconcile the parameters of free energy change filter throughout the article. In Page 9 of SI it is -10.5 kcal mol⁻¹ ~ -12.5 kcal mol⁻¹. However, in the Methods section (Line 309) it is -11.5 kcal mol⁻¹ ~ -12.5 kcal mol⁻¹ and in the results section (Line 86) it is -10 kcal mol⁻¹ ~ -12.5 kcal mol⁻¹.

Response: Thank you for identifying this inconsistency. We have standardized the free energy change filter across the manuscript to ensure consistency. The correct range for the free energy change is now set to -10.5 kcal mol⁻¹ ~ -12.5 kcal mol⁻¹. Here is the updated text:

Page 9 of SI: Combining the above explorations, we set up six primer filtering steps. These include: primer free energy (-10.5 kcal mol⁻¹ ~ -12.5 kcal mol⁻¹), GC content (45% ~ 55%), homopolymer (no AAAA, TTTT, CCC and GGG), dimer (max badness less than 1), and secondary structure (MFE free energy = 0 kcal mol⁻¹).

Methods section: Subsequently, the standard free energy of each primer was calculated using the nearest neighbor model and primers with free energy in the range of -10.5 kcal mol⁻¹ to -12.5 kcal mol⁻¹ were retained.

Results section: Of these, only 8.3% passed the ΔG° filter (-10.5 kcal mol⁻¹ to -12.5 kcal mol⁻¹), and a scant 2% conformed to the GC content criteria.

6. In Line 97, the authors mentioned “various primer sets could be mixed as needed to realize random access to individual or multiple files”. Multiplex PCR requires more careful design of primer sets to ensure the efficiency. In Figure S2-7 and Figure 3b, the authors have evaluated any possible cross-hybridization between primers from different sets and the variation of primer specificity with free energy change. However, in their primer design algorithm, the specificity check is missing, and I suppose the authors screen the specificity entirely by controlling free energy changes, which makes the specificity check less persuasive.

Response: Thank you for your professional suggestion. An additional specificity check is indeed necessary to ensure orthogonality between primers. In response, we have incorporated a BLAST (Basic Local Alignment Search Tool) step in the primer design process to ensure specificity (Fig. 1b). This step aligns primer sequences with themselves and checks for mismatches and potential off-targets, ensuring precise amplification during PCR. After applying the BLAST filter, we retained 11,335 primers (Fig. S2-6). We have updated the text and images accordingly. The revised content is provided below:

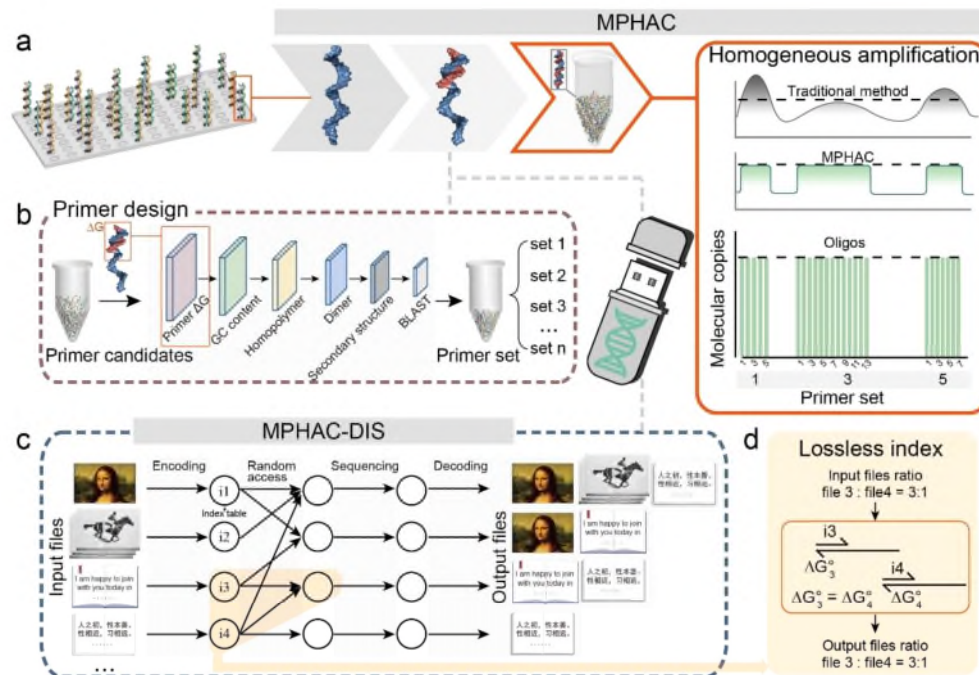


Fig. 1 | Schematic workflow of MPHAC and MPHAC-DIS. **a**, Low concentration single-stranded DNA molecules synthesized on chip were amplified with MPHAC to obtain uniform high-concentration double-stranded DNA molecules. **b**, Primer design and filter workflow. **c**, Workflow of multimodal data storage with random access using MPHAC-DIS. **d**, Lossless index for DNA information storage to maintain files ratio consistency before and after storage.

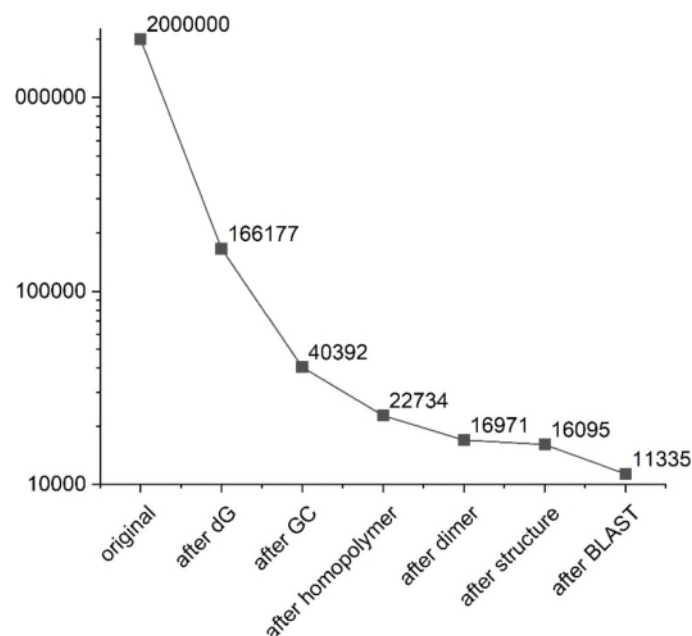


Figure S2-6. Number of primers remaining after each round of filtering steps. Here, 2,000,000 primers were randomly generated to be used as the initial primer pool and 11,335 primers were obtained after 6 rounds of filtering.

Methods

Primer generation

First, 2,000,000 primers of 15 to 30 nt were randomly generated. Subsequently, the standard free energy of each primer was calculated using the nearest neighbor model and primers with free energy in the range of $-10.5 \text{ kcal mol}^{-1}$ to $-12.5 \text{ kcal mol}^{-1}$ were retained. The GC content check (45%-55%) and homopolymer check (AAAA, TTTT, CCC, GGG) of the primers were then performed. For retained primers, badness (a metric describing the probability of dimer formation between two primers) between all primers was calculated. Dimer check was performed using max badness, and only primers with max badness less than 1 were retained. Furthermore, the minimum free energy (MFE) structure of each primer was analyzed by NUPACK. Primers with free energy equal to 0 kcal mol^{-1} passed the secondary structure check and were included in the final primer library. Finally, crosstalk between primers and their reverse complementary sequences was verified using BLAST to ensure that the number of complementary bases (including shift complementarity) was less than or equal to 16.

7. The authors may also need to evaluate the specificity of primers to ensure that no primers will target an unintended template, and in their primer design algorithm, an additional screening step may need to be carried out.

Response: Thank you for your suggestion. We would like to emphasize that we have included a BLAST step at the final stage of our primer design process, which effectively evaluates primer specificity. This step compares the primer sequences against themselves and the primer binding regions (reverse complement sequences), allowing us to screen for specificity between primers and verify that each primer targets its intended template without unintended interactions.

8. Was the simulated sequencing experiment described in Lines 121-126 based on the simulated amplification yield? I suggest the authors to specify the detailed procedure of the simulated sequencing experiment in the Methods section.

Response: Thank you for your helpful suggestion. Yes, the simulated sequencing experiment described in Lines 121-126 was indeed based on the simulated amplification yield. We have added further details to clarify the procedure in the Methods section. Below is the updated text with the additional information:

100,000-plex primer simulation

Firstly, 100,000 primers were generated for each of the FL-design and FE-design, respectively. The ΔG° of each sequence was then calculated to determine the free energy distributions for both designs. Subsequently, the amplification efficiency of each primer was calculated based on the theoretical model, and the coverage of each primer was derived using the following equation:

$$Coverage_i = \frac{Depth \cdot n \cdot Efficiency_i^{Cycle}}{\sum_{i=1}^n Efficiency_i^{Cycle}} \quad (1)$$

Here, n represents the total number of primers, which in our case is 100,000. Depth refers to the sequencing depth, set to $50\times$, and Cycle denotes the number of PCR amplification cycles, set to 20.

9. In the Methods section, Line 337. How was the amplification efficiency evaluated for a non-target? Has the association and dissociation rate changed compared to on-target amplification? The authors may need to specify this in the same paragraph.

Response: Thank you for your suggestion. For non-targets, we assume that they differ from the target by n bases, which results in a change in ΔG ($\Delta\Delta G$). This difference impacts the binding affinity of the non-target sequence to the primer. While the binding rate constants for nucleic acid hybridization (spanning tens of bases) are usually fixed (e.g. $10^{-6} \text{ M}^{-1} \text{ s}^{-1}$)², following the relationship $\Delta G = k_{hyb}/k_{dis}$, non-specific targets exhibit a binding rate comparable to that of the specific target but dissociate more rapidly due to their weaker binding energy. We simulated various scenarios with different $\Delta\Delta G$ values to observe their effect on amplification efficiency. Across these simulations, we consistently found the same optimal thermodynamic parameters (see Fig. S3-5). We have clarified this in the Methods section to provide a more explicit explanation of how non-target amplification was evaluated. Here is the updated version:

$Efficiency_{target}$ signifies the amplification efficiency of the matching target sequence, and $Efficiency_{non-target}$ indicates the amplification efficiency of the non-target. Here, non-target denotes a non-specific template with multiple base differences compared with target, whose ΔG with the primer is more positive than that of target with primer, and the difference between the two is $\Delta\Delta G$, which is fixed.

Reference:

[2]. Simmel, F.C., B. Yurke and H.R. Singh, Principles and Applications of Nucleic Acid Strand Displacement Reactions. Chem. Rev., 2019. 119(10): p. 6326-6369.

10. Please specify the meaning of dashed lines in Figure S5-4. Are these indicating the threshold PCR cycles?

Response: The dashed lines in Figure S5-4 do indeed represent the threshold PCR cycles. We have updated the figure and its caption to include a more detailed explanation. Here is the revised version:

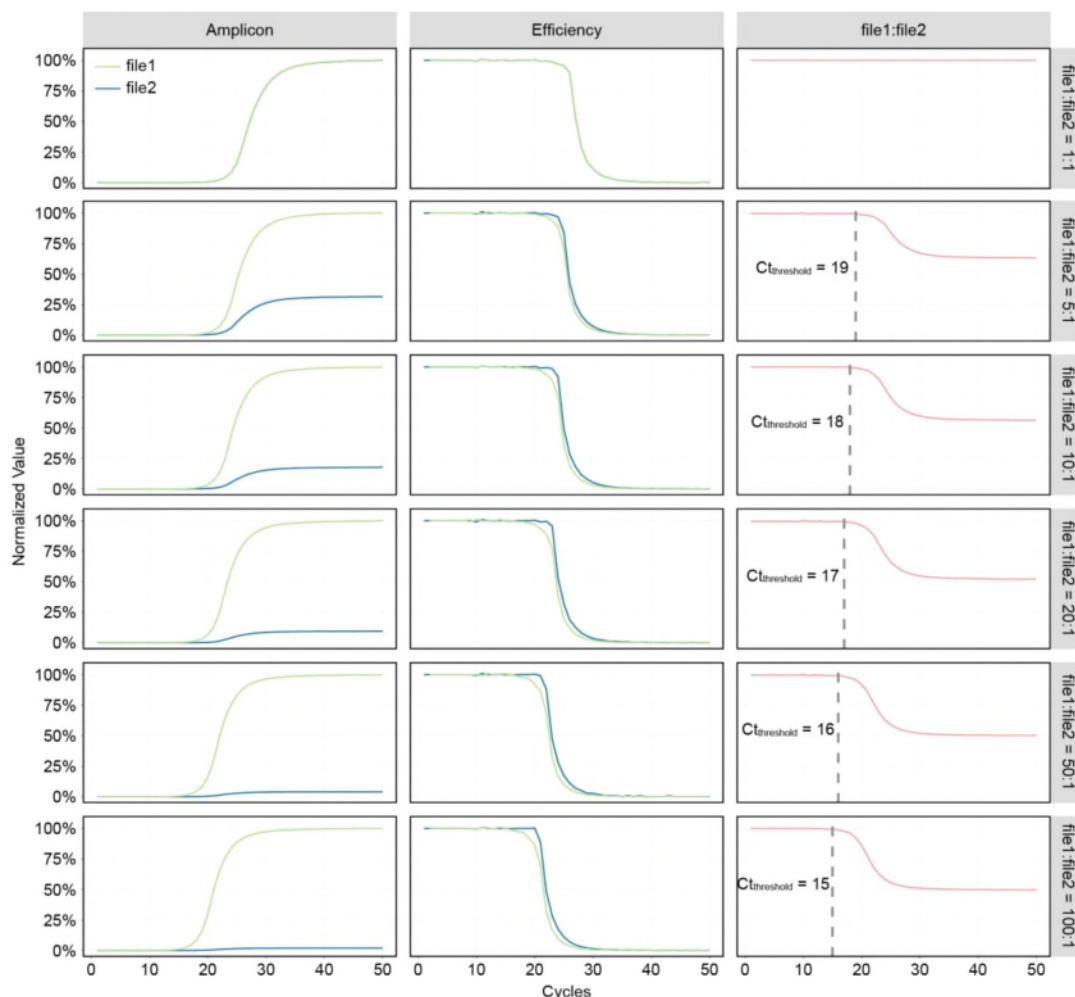


Figure S5-4. Effect of File Size on Amplification. The amplification curves (left), amplification efficiencies (middle), and molecule number ratios (right) of the two files are shown while the size ratios of the two files varied from 1:1 to 100:1. The values have been normalized in the analysis. The dashed line indicates the threshold PCR cycles that start to introduce bias.

11. In Figure S5-4 and Figure S5-5, I am still not clear whether the threshold PCR cycle correlates more with the file size ratio or the highest initial concentration of a file. Besides, the authors discussed the situation of amplifying two files with varying file size ratios, how about multiple files, as in reality many files with different sizes may need to be amplified simultaneously?

Response: Thank you for your feedback, and we apologize for that the original description in the article was not sufficiently precise. When amplifying multiple files, we introduced a key parameter in our model that considers the interaction between the polymerase and the double-stranded template (dsDNA). According to the literature, polymerase has a binding capacity to dsDNA. When polymerase binds to excess dsDNA, it is unable to extend the primer-template complex, which reduces its overall activity³.

Our modeling results demonstrate that when polymerase activity begins to be impacted by the binding of dsDNA (at the threshold PCR cycle), there is a slight difference in amplification efficiency depending on the template concentrations. This difference can introduce amplification bias. Moreover, The threshold cycle is primarily related to the initial total concentration of the templates to be amplified. Only in cases where there is a significant concentration disparity between two templates, the higher concentration template tends to dictate the critical cycle threshold. Therefore, for large-scale file amplification, it is crucial to calculate the total initial concentration of all the templates to be amplified. This allows for a more accurate estimate of the optimal number of PCR cycles. In order to make the above points clearer, we have revised the original text, and the following is the revised version:

Here, we introduced a key parameter in our model: the interaction between the polymerase and the double-stranded template (dsDNA). When polymerase binds to excess dsDNA, it is unable to extend the primer-template complex, which reduces its overall activity. As shown in Figure S5-4, the file amplification curve and amplification efficiency present as opposite sigmoid curves. Notably, as the amplification curve enters the exponential period, the amplification efficiency of the file begins to decline as well. It is observed that the amplification efficiency of the two files remains approximately equivalent for various file size ratios. However, upon closer examination, a slight delay in the amplification efficiency of file2 is noticeable when the amplification efficiency of file1 begins to decrease. This difference in amplification efficiency due to the delay introduces a bias, resulting in a change in the size ratio of the two files.

Notably, the point at which the ratio changes correlates with the size ratio of the files. This is because threshold cycle is actually related to the overall size of the file to be accessed (the sequences needed to be amplified). Larger files correspond to higher initial concentrations, leading to an earlier onset of the exponential period (when generated dsDNA can significantly affect polymerase activity) and, consequently, an earlier introduction of bias. As a result, it is essential to ensure that the number of amplification cycles remains below the threshold during library preparation. Figure S5-5 displays the thresholds at various initial concentrations, revealing a linear correlation between the thresholds and $\log_{10}$ (concentration). When performing large-scale file random access, the total concentration of the file to be accessed can be calculated and a suitable PCR cycle can be selected based on this curve.

Reference:

[3]. Kainz, P., The PCR plateau phase – towards an understanding of its limitations. *Biochimica. et. biophysica. acta*, 2000. 1494(1): p. 23-27.

12. Figure S5-6. Since the amplified files will be subjected to sequencing to read out the data anyway, why not perform the adapter PCR as the first step to cover the “first PCR”? If so, to avoid bias in sequencing reads or even loss of data, the authors may also need to include sequencing adapters in the multiple screening steps of their PCR primer design, especially the dimer formation, secondary structures and specificity. Also, please correct “adaptor” to “adapter” in the figure.

Response: Thank you for your insightful question. While using adapter PCR as the first step could indeed simplify the process and offer some advantages, adapter primers are typically quite long (50-70 nucleotides). The increased length of primers enhances their binding strength, which also increases their tolerance for mismatches. This heightened tolerance can lead to a higher likelihood of inter-primer hybridization, resulting in the formation of primer dimers^{1,4}. Primer dimers, in turn, can interfere with the amplification of the target template. To mitigate this risk, we opted to use conventional primers for the initial pre-amplification. Afterward, we added universal sequences to both ends of the target sequences through three additional cycles of adapter PCR. This approach minimizes the potential for primer-dimer formation during the critical early stages of amplification.

Finally, we have also corrected the spelling from “adaptor” to “adapter” throughout the article. Below is the revised original text:

Next, 15 μ L purified amplicon, 2.5 μ L adapter FP-mix, 2.5 μ L adapter RP-mix, 0.5 μ L polymerase, 1 μ L dNTP (10 mM), 10 μ L 5 $\times$ HF buffer and 18.5 μ L ddH₂O were mixed to total volume of 50 μ L. The thermocycling protocol included an initial 98 °C denaturation for 2 minutes, followed by 3 cycles of 98 °C for 20 seconds, 63 °C for 30 seconds and 72 °C for 30 seconds. Similarly, the amplicon products were then purified from the reaction mixture. This step added the adapter sequences to the ends of the amplicon.

Reference:

[1]. Hendling, M. and I. Barisic, In-silico Design of DNA Oligonucleotides: Challenges and Approaches. *Comput. Struct. Biotechnol. J.*, 2019. 17: p. 1056-1065.

[4]. Song, P., et al., Selective multiplexed enrichment for the detection and quantitation of low-fraction DNA variants via low-depth sequencing. *Nat. Biomed. Eng.*, 2021. 5(7): p. 690-701.

13. In Figure S5-7. Is the threshold read count fixed or flexible depending on actual experimental conditions?

Response: Based on previous experience⁵, we typically use a threshold read count of 5, which has been shown to ensure accurate decoding under normal conditions. This threshold helps filter out low-confidence reads while maintaining high decoding accuracy. However, in experiments with very low sequencing depths, achieving a read count

greater than this threshold can be challenging, even when the reads are correct. In such cases, lowering the threshold can sometimes improve decoding results, as it allows more reads to be considered, despite the lower coverage.

Reference:

[5]. Kim, J., et al., Metastable hybridization-based DNA information storage to allow rapid and permanent erasure. Nat. Commun., 2020. 11(1).

14. In Figure 4a, given that the last five files are in text format, how would the decoded files also be represented in pictures? Also, how could the sequencing coverage be represented in a 2D rectangle as the picture files did?

Response: Thank you for your question. Indeed, the last five files are text files, and the images provided in the figure serve primarily as a visual representation to illustrate the type of files being stored. The actual decoding results can be better seen through the corresponding sequencing coverage diagrams.

Regarding the representation of sequencing coverage in a 2D rectangle (as with the image files), here is the approach: During the encoding process, the information is split and stored across multiple sequences. After sequencing, we calculate the coverage of each sequence by aligning the reads. This results in an $N \times 1$ vector, where N represents the number of sequences. To visualize this coverage as a rectangular image, we transform the $N \times 1$ vector into an $M \times K$ matrix (where M and K are chosen to suit the visualization). A heatmap is then generated based on the values in this matrix, where each point in the heatmap corresponds to the coverage of a particular sequence. This arrangement allows for intuitive visualization, where the sequences are organized into a grid-like shape to convey the coverage data clearly.

15. Figure S8-2. The authors analyzed the effect of GC content on sequencing errors. This metric represented the GC content of the whole sequence of a strand. Did the authors also evaluate GC-rich regions (narrow down to 10-20 nt) of a sequence in particular to see whether sequencing errors are more prone to these regions? Compared to using the overall GC content, analysis with local GC content may be more helpful to demonstrate the effect of GC content on sequencing errors.

Response: Thank you for your valuable suggestion. In response, we split the sequence into 20 nt segments (domains) for localized error rate analysis. Since sequencing error frequency tends to increase near the 3' end of reads (Fig. S8-14), we ensured fair comparison by analyzing error rates of domains with varying GC content at the same positions. We selected three domains at different positions and categorized them into two groups based on GC content: $\geq 50\%$ and $< 50\%$. The results show no significant difference between the groups with different GC contents (Fig. S8-3). However, it is important to note that the coding method we used inherently maintains the GC content of the payload region between 40% and 60%, meaning there are no GC-rich regions with very high GC content. This may explain why we did not observe a significant difference between the $\geq 50\%$ and $< 50\%$ groups.

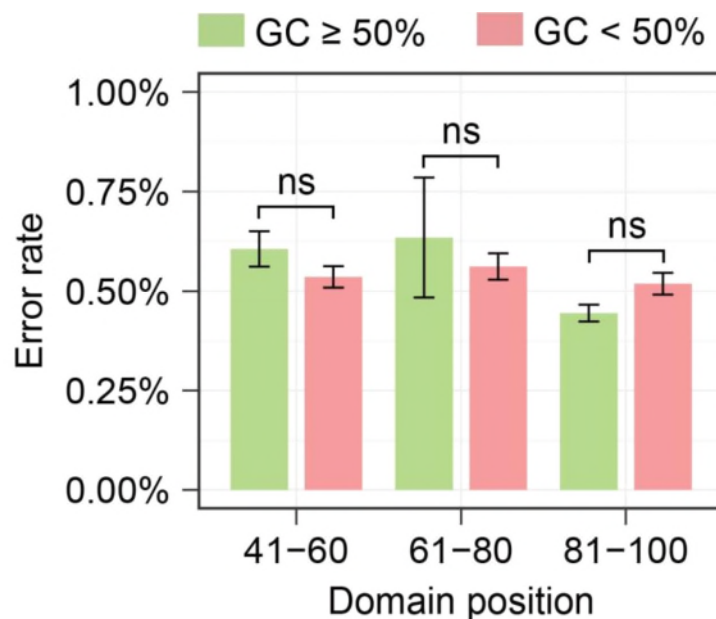


Figure S8-3. Comparison of error rates in domains with varying GC content. In this context, domains refer to 20nt short motifs. We selected three domains at different positions and classified them into two groups based on their GC content: $\geq 50\%$ and $< 50\%$. The error rate was then calculated for each group. The results indicated no significant difference in error rates between the two groups.

16. SI section 8, “Effect of sequence secondary structure”. The examples the authors used to evaluate the effect of secondary structures all contained repetitive sequences, which may cause difficulties in algorithmic sequence assembly and produce sequencing errors. This makes the described evaluation not adequate to identify whether the sequencing errors originated from secondary structures or repetitive sequences. On the other hand, the secondary structures in the provided examples were actually not strong, given the base pairing probabilities of the hybridized regions not significantly high (around 50%). The authors may need to provide more examples excluding repetitive sequences but solely containing stable secondary structures to further evaluate the effect of secondary structure on sequencing errors. Lastly, for the subtitle “b) Secondary structure between payload region and primers”, please use “primer binding region” in place of “primers” to avoid misunderstanding.

Response: Thank you for your suggestion. In response, we have added some secondary structure diagrams of non-repetitive sequences with distinct secondary structures to the supplementary material for further evaluation. Here is the updated version:

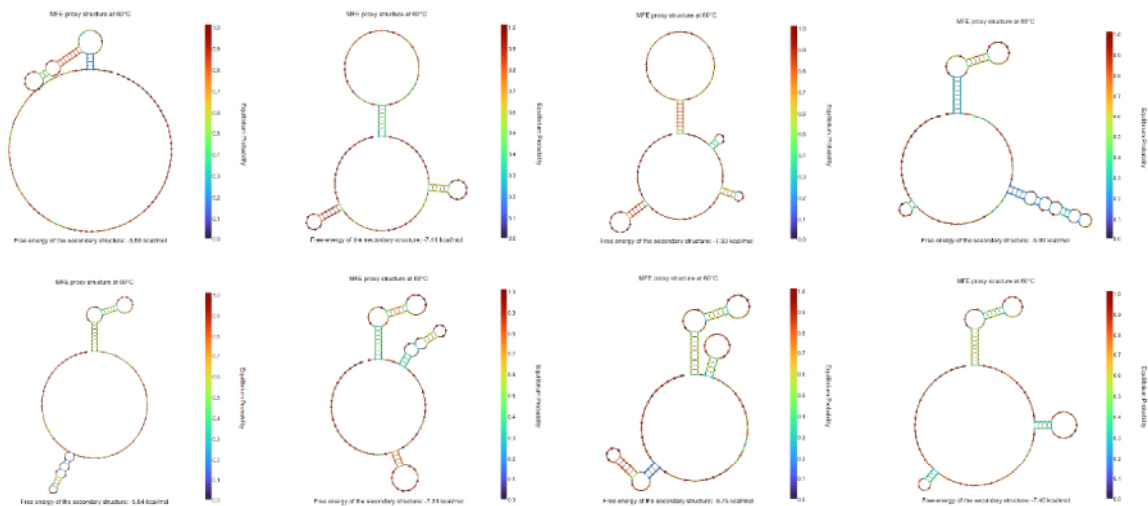


Figure S8-12. Secondary structure of non-repetitive error-prone sequences. The sequences shown here are those that have failed to decode in multiple experiments.

Furthermore, we have corrected the terminology “primer” and “primer binding region” throughout the manuscript. The corrected text is provided below:

b) Secondary structure between payload region and **primer binding region**. Having observed the influence of self-secondary structure of the payload region on the decoding accuracy, we will now investigate whether the secondary structure between the payload region and the **primer binding region** also plays a role.

Responses to Reviewer #2

The study presents a promising approach to primer selection for DNA amplification that could be instrumental in large-scale information storage. Among the methods for DNA synthesis, chip-based synthesis is particularly suited for large-scale storage as it enables the simultaneous synthesis of a wide variety of oligonucleotides. However, this approach yields low initial concentrations of each oligonucleotide type, necessitating the use of PCR amplification for decoding. The authors highlight that conventional primer design strategies, which rely on fixed-length primers, result in a wide range of primer-template hybridization energies, leading to uneven amplification of different DNA sequences. To address this issue, the authors propose a fixed-energy primer design approach that avoids fixed primer lengths and instead optimizes hybridization energy for uniform amplification.

As the authors state, previous studies have already considered the energy of primer hybridization or related parameters, such as melting temperature (e.g., doi:10.1038/nbt.4079). However, this study provides a comprehensive comparison of the fixed-length versus fixed-energy approaches, emphasizing and explaining the necessity for such design. Through theoretical modeling, the authors determine the optimal hybridization energies that balance both high specificity and hybridization efficiency. These findings are substantiated by experimental data, demonstrating that fixed-energy primers enable uniform amplification during PCR and achieve high accuracy in encoding and decoding various types of data.

Response: Thank you for your encouraging feedback on our work. We appreciate your recognition of the potential impact of our fixed-energy primer design approach. Regarding the comparison with melting temperature (T_m) strategies, we acknowledge that studies like Lee Organick et al. (Nat. Biotechnol. 2018) have utilized T_m-based filters effectively. However, our results indicate significant differences in amplification uniformity, which we elaborate on in our response to question 2. Additionally, we have expanded our comparative analysis to include a new table (Table S3-1) detailing various primer design strategies. This addition highlights the advantages of our fixed-energy approach in terms of primer capacity and amplification homogeneity. Thank you once again for your valuable insights.

Table S3-1 Simulated performance comparison between two designs

	FL+Tm ^a	FE ^b
Primer capacity ^c	2.53×10 ¹¹	1.67×10 ¹⁴
Efficiency range	78.2%-100%	>99%
Correct ratio ^d	87.08%	100%
Fold-80 base penalty	1.28	1

^a FL+Tm is defined here as a fixed length of 20 nt, with the T_m value constrained between 55°C and 60°C. This standard is derived from the work of Lee Organick et al¹.

^b FE refers to the limitation of the Gibbs free energy change, set between -10.5 kcal mol⁻¹ and -12.5 kcal mol⁻¹.

^c Primer capacity is defined as the maximum number of primers that can be generated. Initially, we generate 10,000 random primers and calculate the percentage that passes the filter, which is then used to derive the primer capacity.

^d Correct ratio indicates the percentage of successfully decoded files. For this analysis, we have established a sequencing depth of 50× and 30× coverage as the threshold for successful decoding.

Reference:

[1]. Organick, L., et al., Random access in large-scale DNA data storage. Nat. Biotechnol., 2018. 36(3): p. 242-248.

Below are our point-by-point responses, with the response content in regular font size and the article content in a slightly smaller font size. Original content is highlighted in **blue**, and modifications are marked in **red**.

Despite the comprehensive design of the study, and promise of the fixed-energy primer design for DNA storage systems, several concerns need to be addressed:

1. 1) My main concern of the proposed fixed-energy approach is that it uses not factual energy but predicted energy based on a theoretical model, which is far from ideal when it comes to experiments. 2) This way any number of other theoretical metrics can be devised with varying results and the used approach does not seem as the optimum. For example, using a neural network to design primers based on considerable experimental data may give much better results. While I greatly appreciate the thoroughness of the study that involved a huge amount of work, 3) I am not sure that the manuscript clearly defines why the study conceptually presents a significant inventive step over the previous works.

Response: Thank you for your insightful comment. 1) For the first point, we ensure that our predicted values align with real-world applications through both prior experience^{2,3} and experimental validation. The thermodynamic calculation methods we employed have been previously validated (Table R1)^{4,5}, further supporting the reliability of our approach. Moreover, our experimental results demonstrate that the predictions are consistent with experimental outcomes (Fig. S4-6), validating the reliability of our simulation in practical scenarios.

Table R1. The previously validated nearest-neighbor thermodynamic parameters for DNA Watson-Crick pairs.

Propagation sequence	ΔH° (kcal mol ⁻¹)	ΔS° (e.u.)	ΔG_{37}° (kcal mol ⁻¹)
AA/TT	-7.6	-21.3	-1.00
AT/TA	-7.2	-20.4	-0.88
TA/AT	-7.2	-21.3	-0.58
CA/GT	-8.5	-22.7	-1.45
GT/CA	-8.4	-22.4	-1.44
CT/GA	-7.8	-21.0	-1.28
GA/CT	-8.2	-22.2	-1.30
CG/GC	-10.6	-27.2	-2.17
GC/CG	-9.8	-24.4	-2.24
GG/CC	-8.0	-19.9	-1.84
Initiation	+0.2	-5.7	+1.96
Terminal AT penalty	+2.2	+6.9	+0.05
Symmetry correction	0.0	-1.4	+0.43

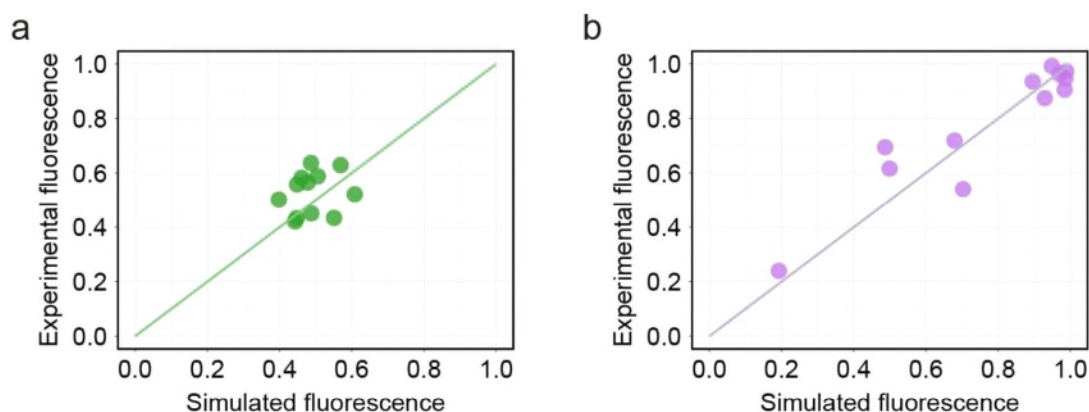


Figure S4-6. Endpoints Comparison of two groups. (a) FE-primer group, and (b) FL-primer group.

2) Regarding the second point, while neural network-based primer design can indeed yield useful results, it often requires large training datasets. Neural networks demand vast amounts of data for learning; otherwise, the model may generate biased features based on the training set, leading to overfitting and reduced generalizability. As more data becomes available, accuracy may improve, but currently, it is not feasible to generate such large-scale PCR data. Additionally, due to the "black box" nature of neural networks, it is challenging to ensure that all primers satisfy strict design criteria, such as primer dimer screening. In contrast, our method is based on experimentally verified features and well-established PCR principles, providing an intuitive reflection of PCR efficiency. Our approach, grounded in well-established theoretical principles, offers strong interpretability and ensures that all primers strictly adhere to the design criteria.

3) Lastly, unlike conventional genomic PCR, primer design in DNA storage does not depend on fixed template sequences and is subject to biological limitations like DNA synthesis and sequencing. To our knowledge, there is no high-throughput, bias-free PCR amplification model for DNA storage. To address this, we constructed a PCR amplification model based on the thermodynamics and kinetics of primer hybridization, further combined with enzyme extension, leading to optimal thermodynamic parameters. Notably, compared to traditional fixed-length designs, our fixed-energy approach allows for the generation of a significantly larger number of primers. Through strict free energy control and filters for GC content, homopolymer, sequence secondary structure, etc., we achieved bias-free amplification. We hope this provides a clearer understanding of how our work represents an advancement over previous methodologies.

Reference:

- [2]. Song, P., et al., Selective multiplexed enrichment for the detection and quantitation of low-fraction DNA variants via low-depth sequencing. *Nat. Biomed. Eng.*, 2021. 5(7): p. 690-701.
- [3]. Xie, N.G., et al., Designing highly multiplex PCR primer sets with Simulated Annealing Design using Dimer Likelihood Estimation (SADDLE). *Nat. Commun.*, 2022. 13(1).
- [4]. SantaLucia, J. and D. Hicks, The Thermodynamics of DNA Structural Motifs. *Annual Review of Biophysics and Biomolecular Structure*, 2004. 33(1): p. 415-440.
- [5]. Wang, C., J.H. Bae and D.Y. Zhang, Native characterization of nucleic acid motif thermodynamics via non-covalent catalysis. *Nat. Commun.*, 2016. 7(1).

2. The results of the fixed-energy approach should be directly compared with those obtained using existing methods to highlight the significance of the research. For instance, the study by Lee Organick et al. (*Nat. Biotechnol.* 2018) employs similar filters for primer design based on melting temperature rather than hybridization energy, and also results in accurate data decoding.

Response: Thank you for your suggestion. The work by Lee Organick et al., published in *Nature Biotechnology* in 2018, represents a significant milestone in the field of DNA storage, successfully encoding over 200 MB of data. In their approach, they also employed a primer design strategy based on melting temperature filters, which, like our method, enabled accurate data decoding. However, our study reveals a key difference: the homogeneity of sequencing results in their method differs by two orders of magnitude (Fig. R1). While it can still achieve accurate decoding, this generally requires higher sequencing coverage due to its inhomogeneity. In contrast, our method exhibits more uniform sequencing results (Fig. 4b, top), allowing for accurate decoding even at lower sequencing depths, which underscores the advantages of our hybridization energy-based primer design approach.

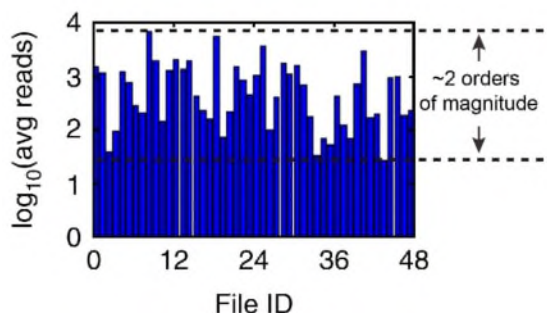


Figure R1. Sequencing homogeneity results for primers designed based on melting temperature. This result is from Figure 2a of the article by Lee Organick et al. published in *Nature Biotechnology* in 2018.

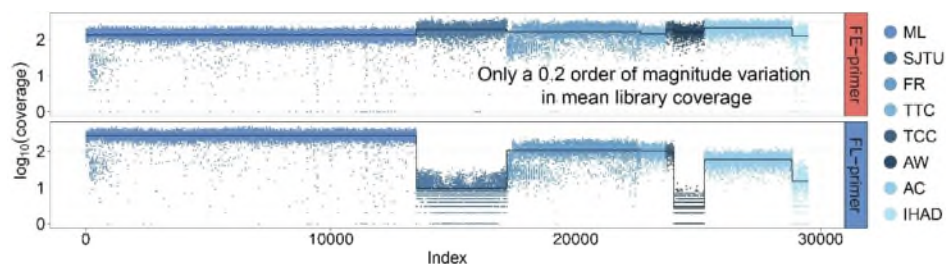


Figure 4b. Coverage comparison between FE-primers and FL-primers.

3. The study does not present any statistical analyses to evaluate the significance of the results, which should be included for quantitative comparisons of different groups.

Response: Thank you for your valuable suggestion. We have now incorporated a statistical significance analysis in Figure 4c to highlight the differences between the two groups. The updated version is provided below:

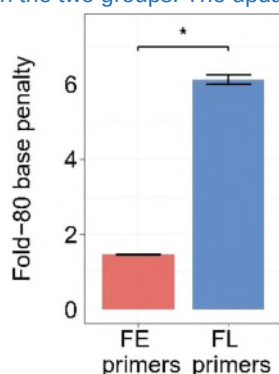


Fig. 4c, Fold-80 base penalty of FE-primers and FL-primers. Fold-80 base penalty is a metric used to assess coverage uniformity in sequencing experiments. It indicates how much more sequencing is needed for 80% of the targets to reach the mean coverage. The closer the Fold-80 base penalty is to 1, the better the coverage uniformity. Data are presented as mean $\pm$ standard error, with $n = 2$ for each group. Statistical significance was determined using an independent two-sample t-test. Asterisks denote levels of statistical significance (* denotes $p < 0.05$).

4. To prove that homogenous amplification is independent of initial template concentrations, PCR experiments should be conducted using various initial concentrations of the plasmid.

Response: Thank you for your suggestion. We conducted PCR experiments using plasmid templates of 600,000, 60,000, and 6,000 copies/rxn. The results indicate that while lower template concentrations affect the Ct values, overall, the FE primers exhibit very uniform amplification, whereas the FL primers show some degree of variability (Fig. S4-3 and Fig. S4-4).

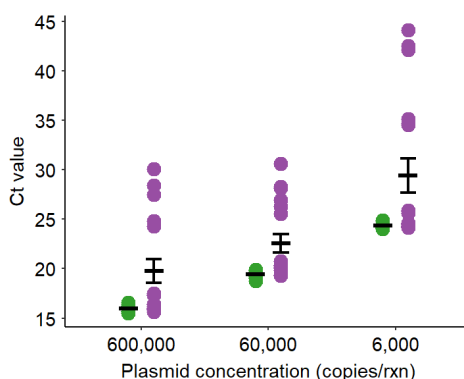


Figure S4-3. Amplification results of two sets of primers at different initial concentrations of plasmids. Each group comprised six primer pairs, with each primer pair repeated three times.

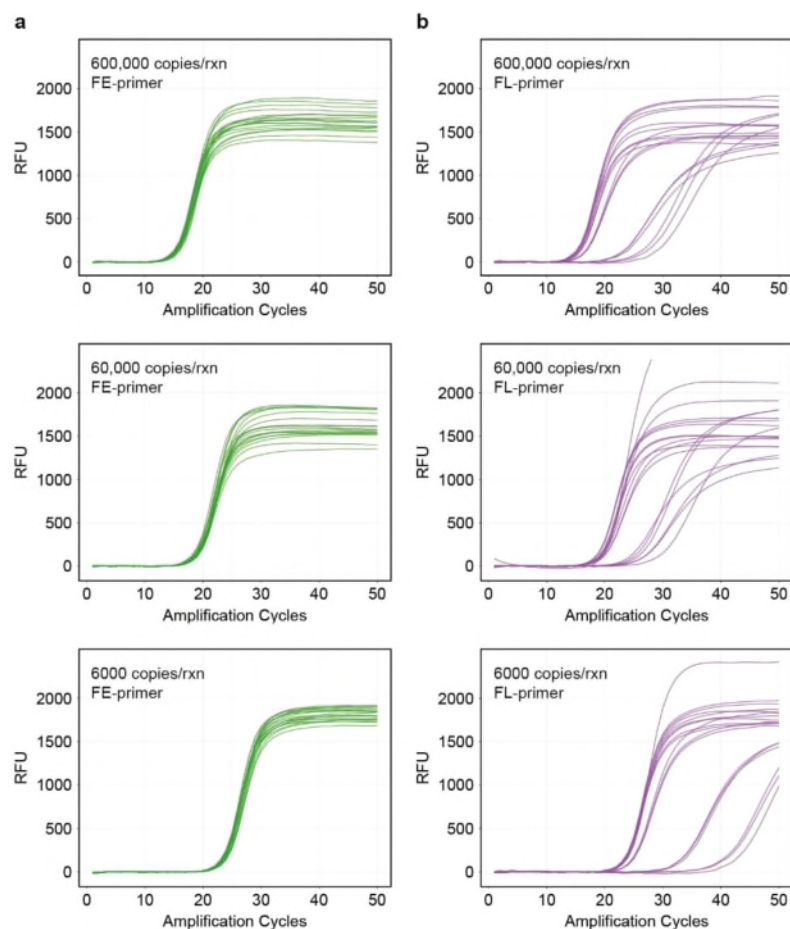


Figure S4-4. PCR amplification curves of two groups at different initial concentrations of plasmids. (a) FE-primer group, and (b) FL-primer group.

5. Even a library of 16,095 primers seems insufficient for storing significant amounts of data. This limitation, along with challenges related to inaccurate oligo synthesis, suggests that DNA information storage is still more of a concept than a viable storage system. Therefore, the authors should discuss future plans and strategies to address these challenges.

Response: Thanks to your suggestion. While we generated tens of thousands of primers, which is higher compared to other works, it is indeed still insufficient for large-scale DNA storage. We have now expanded the discussion section to the limitations of the current approach and the potential strategies for overcoming these obstacles in future research. The updated text is provided below:

While the MPHAC-DIS method is effective for DNA information storage, it does have certain limitations. Firstly, although we optimized and generated a library comprising tens of thousands of primer pairs through our primer design workflow, practical wet-lab experiments face challenges in handling such a large number simultaneously, which can lead to non-specific amplification and primer dimer formation. Beyond algorithmic solutions, primer dimers and longer non-specific amplicons can be reduced using size selection with magnetic beads and enzymatic cleavage. Secondly, even tens of thousands of primer pairs are insufficient for large-scale DNA storage. Future advancements will require faster algorithms and methods capable of generating primers with higher capacity. Additionally, techniques such as isothermal amplification can be employed to improve access accuracy during random access, and physical addressing remains a viable option. Finally, biases introduced during the DNA synthesis stage can affect the reliability of the storage system. To mitigate this, we can introduce error correction codes and incorporate partial redundancy to ensure decoding accuracy. Moreover, there is a need to develop more economical and higher-quality DNA synthesis techniques, such as enzymatic synthesis.

To improve comprehension of the material, several adjustments are recommended:

6. Key information, such as the encoding-decoding strategy, should be incorporated into the main text to ensure clarity for readers.

Response: Thank you for your suggestion. We have incorporated the information of the encoding-decoding strategy into the main text and provided references to the supplementary information. Below is our revised text:

To further validate the effectiveness of MPHAC, we implemented it in DNA information storage (MPHAC-DIS) (Supplementary Section 5). Since all files are fundamentally binary data, during the DNA storage encoding phase, we converted 1 byte (8 bits) of binary information into 5 nucleotides based on a predefined transcoding table (Fig. S5-2). Primer binding regions were then appended to both ends of the information sequence to enable random access (Fig. S5-3). In MPHAC-DIS, 35,406 DNA oligos were synthesized to achieve the concurrent encoding and storage of multimodal data, encompassing texts, images, and videos. During the decoding process, the most plausible sequences were selected for information retrieval using frequency statistics (Fig. S5-7). Supplementary Section 6 showed the detailed file information and access combinations.

7. Terms like "fold-80 base penalty" and "Ct-value" should be defined to make the content accessible to a broader audience.

Response: Thank you for your suggestion. We have now added definitions for the terms "fold-80 base penalty" and "Ct-value" in the captions to ensure the content is more accessible to a broader audience. Here is the updated version:

fold-80 base penalty: Fold-80 base penalty is a metric used to assess coverage uniformity in sequencing experiments. It indicates how much more sequencing is needed for 80% of the targets to reach the mean coverage. The closer the Fold-80 base penalty is to 1, the better the coverage uniformity. (Fig. 4c caption)

Ct value: Here, Ct value (Cycle threshold value) represents the cycle number at which the fluorescence signal exceeds a defined threshold, indicating the presence of the target. Theoretically, identical concentrations of the target should result in the same Ct value. (Fig. 3i caption)

8. The term "primer" is inconsistently used to refer to both a segment of a template and the primer itself. These should be clearly differentiated as reverse complementary sequences to avoid confusion.

Response: Thank you for your suggestion. We have now distinguished between the terminology "primer" and "primer binding region" throughout the manuscript to clearly differentiate between the primer itself and the corresponding segment of the template. We have updated the terminology consistently across the entire text.

9. P.6 "Given the typically low initial concentration of the template during actual PCR, we simulated the hybridization yield of primers when $A \gg B$." The designations "A" and "B" are not clearly identified in the text.

Response: Thank you for pointing this out. In the context of our discussion, A refers to the primer, and B denotes the reverse complementary sequence of the primer. We have now added a schematic representation in supplementary materials (Fig. S4-2) and added these definitions to the figure caption for clarity. Here is the updated version:

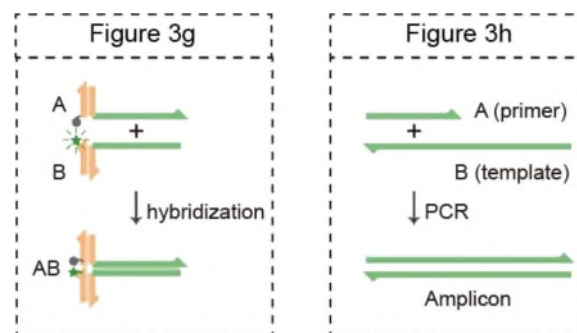


Figure S4-2. Schematic representation of the sequence represented by A and B in Figure 3.

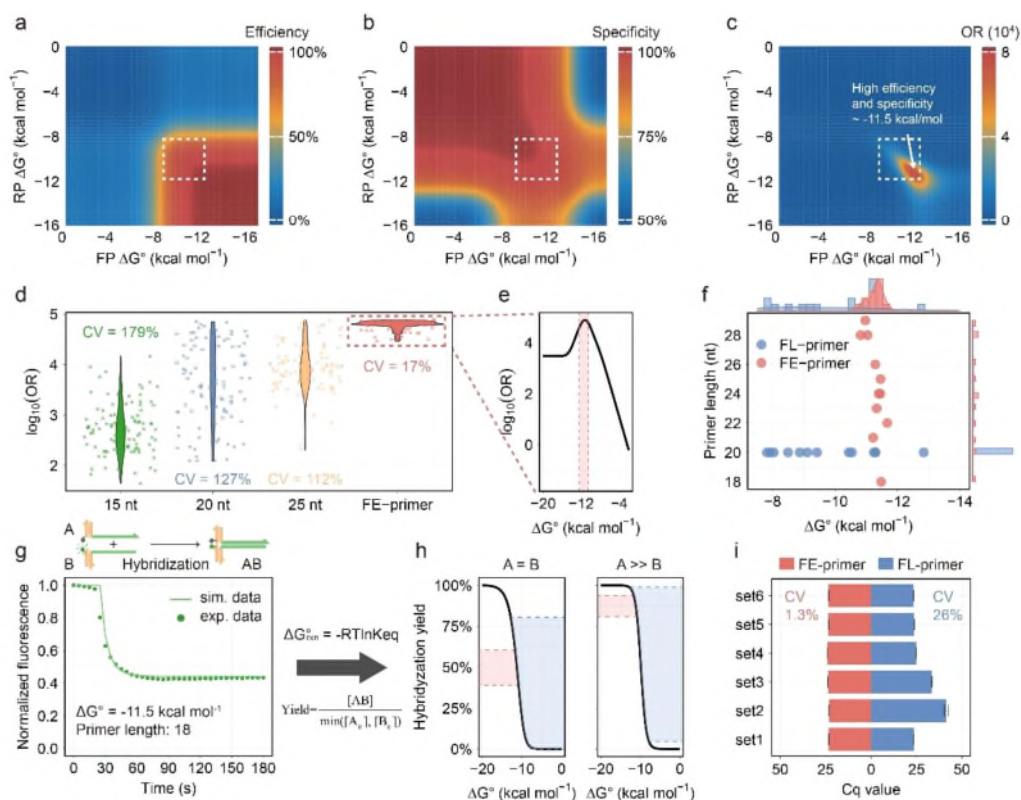


Fig. 3 | Parameters optimization of MPHAC. **a-c**, Simulated results of amplification efficiency (**a**), specificity (**b**) and odds ratio (OR) (**c**) under different primer free energies based on the theoretical model. Details of the model construction are provided in Supplementary Section 3. The white dashed box represents the free energy range for 80% of FL-primers in 100,000-plex primer pairs. Amplification efficiency positively correlates with free energy, while specificity exhibits the opposite trend. We considered the amplification efficiency and specificity together by OR, and the results suggest that the primer free energy can be optimized around $-11.5 \text{ kcal mol}^{-1}$. **d, e**, OR values under different length FL-primers and FE-primers (**d**) and different primer free energies (**e**). Here, the coefficient of variation (CV) is a measure of the relative dispersion of a data set, expressed as the ratio of the standard deviation to the mean. **f**, Primer length and free energy distribution. Six sets of FE-primers and six sets of FL-primers were designed based on plasmid sequences. **g**, Monitoring primer hybridization fluorescence kinetics using X-probe. Experimental data are represented by dots, and simulated data is depicted by lines. **h**, Simulated hybridization yields for primers when $A=B$ (left) and $A \gg B$ (right). Here, **A** refers to the primer, and **B** denotes the reverse complementary sequence of the primer (template). A detailed schematic representation can be seen in Fig. S4-2. The range of hybridization yields for FE-primers is considerably narrower (pink) than that of FL-primers (blue). **i**, PCR results for FE- and FL-primers. Error bars represent the results of three parallel experiments. Here, Ct value (Cycle threshold value) represents the cycle number at which the fluorescence signal exceeds a defined threshold, indicating the presence of the target. Theoretically, identical concentrations of the target should result in the same Ct value.

10. P.5 “Based on the theoretical model, we explored amplification efficiency and specificity under different primer free energies.” Theoretical model should be supplemented with a brief description or at least a reference to the supplementary information.

Response: We have added a brief description of the theoretical model along with a reference to the supplementary information for further details. Here is the updated text:

Based on the theoretical model that primarily considers primer hybridization and extension (Supplementary Section 3), we explored amplification efficiency and specificity under different primer free energies.

11. The description of Fig. 5e seems to have mixed-up labels (decoding images and restored images) Several points also require further clarification:

Response: Thank you for pointing out the mislabeling issue. We have corrected the mixed-up labels in the description of Fig. 5e. Below is the revised version:

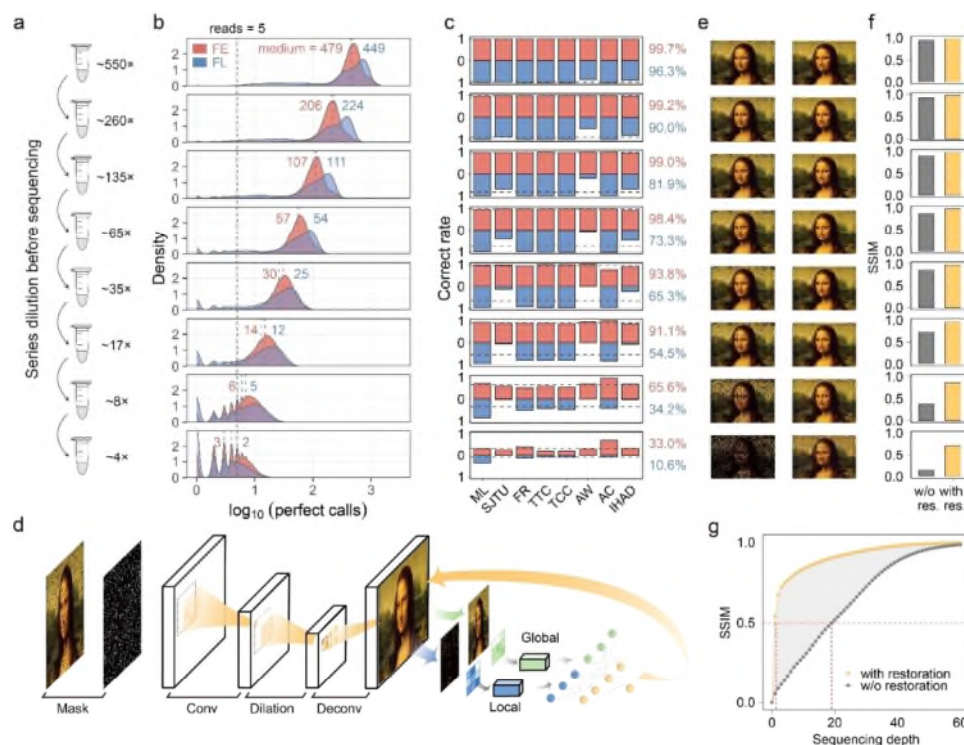


Fig. 5 | Study of MPHAC-DIS using ultra-low coverage sequencing. **a**, The amplified libraries were serially diluted before sequencing, and then mixed together for sequencing. Finally, libraries with different sequencing depths were obtained by demultiplexing the sequencing results. **b**, Density distribution of the exact matching sequences. The red and blue dashed lines indicate the median number of perfect calls under the FE-design and the FL-design, respectively. **c**, Decoding accuracy for each library file. The red and blue dashed lines indicate the average decoding correct rate for the FE-design versus the FL-design, respectively. **d**, Image restoration strategy employing a convolutional neural network. **e**, Decoding results of ML image without restoration (left) and the results with restoration (right). **f**, Comparison of the corresponding structural similarity index measure (SSIM, a metric used to measure the similarity between two images) of ML image before (w/o res.) and after restoration (with res.). The SSIM values indicate the similarity of each image to the original image. The result reveals that the image restoration technique has a more pronounced impact at lower sequencing depths. **g**, Fitting results for the quality of decoded images with and without image restoration at lower sequencing depths. Using SSIM > 0.5 as the threshold, a sequencing depth of 19× is necessary before image restoration. However, with image restoration, a sequencing depth of only 1× is needed, showcasing the efficacy of image restoration technology in further reducing sequencing costs.

12. Table S6-2 lists various libraries without describing the differences between similar ones (e.g., lib1 vs. lib7; lib6 vs. lib8).

Response: Thank you for your comment. The differences between lib1 vs. lib7 and lib6 vs. lib8 represent two sets of parallel replicated experiments. These experiments were conducted to verify reproducibility and consistency under identical conditions. We have now added a description in the original paper, and the updated explanation is as follows:

We conducted experiments as shown in Figure 4, creating nine libraries that utilized various primer combinations for random access to specific files. These experiments showcased the effective orthogonality of the DNA storage system. Details of the random access for each library are documented in Table S6-2. In this context, lib1 vs. lib7 and lib6 vs. lib8 are two sets of parallel replicated experiments performed under identical conditions to assess the reproducibility and consistency of the results.

13. The detailed description of the image recovery algorithm is missing and should be included to provide a complete understanding of the methodology used.

Response: Thank you for your valuable suggestion. In response, we have added a detailed description of the image recovery algorithm in Section S10 of the Supplementary Information. This section provides a comprehensive explanation of our two-stage image inpainting process, including the use of a Convolutional Neural Network (CNN) for initial filling and a contextual attention mechanism for refining results. We hope this addition provides a clearer understanding of our methodology.

Section S10 Image Restoration Model

The image restoration model proposes a two-stage image inpainting process aimed at effectively filling in missing regions within an image (Fig. S10-1). Stage 1 employs a Convolutional Neural Network (CNN) to perform an initial, rough filling of the missing areas. While the continuous CNN predicts the missing parts, it often produces a blurred output. Stage 2 enhances the coarse results through two distinct strategies. First, an analogous convolutional operation is applied to further refine the filled regions. Second, a contextual attention mechanism is integrated, which enables the network to leverage patterns and textures from other intact areas of the image. This mechanism captures contextual information from the preserved regions and applies it to fill the missing areas with more realistic details.

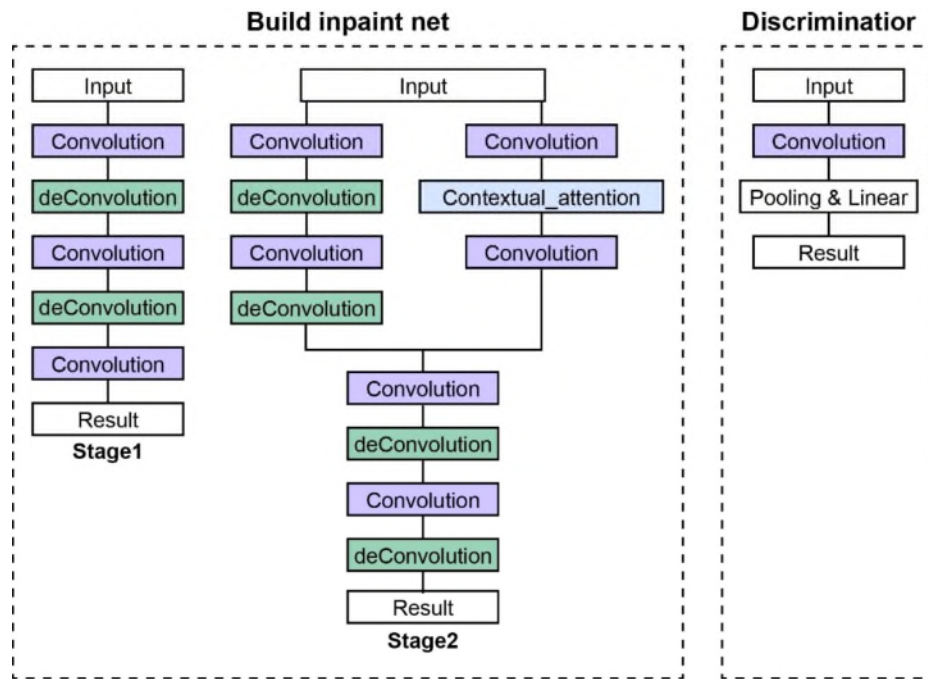


Figure S10-1. Framework of image restoration model.

The discriminator plays a pivotal role in assessing the authenticity of the generated images by differentiating between real and inpainted (generated) images. It is constructed using a sequence of convolutional layers specifically designed to downsample the input image while progressively extracting relevant features. Within the inpainting framework, the discriminator is responsible for classifying patches from both the real image (original content) and the generated image (filled regions).

Moreover, the model utilizes masks derived from DNA storage decoding results to enhance localization accuracy. Unlike conventional masks, which may be generated randomly or manually, the DNA decoding results yield a more informative mask. This advancement facilitates improved localization of areas requiring repair and identification of occluded regions. This design strives to incorporate as many accurate pixels as possible during the learning phase, thereby enabling the generator to more effectively learn the structures and features of unoccluded areas while minimizing unnecessary error propagation.

Responses to Reviewer #3

Summary and recommendation: This manuscript introduces MPHAC-DIS, a novel strategy for DNA-based information storage that utilizes simulation-guided primer design to achieve unbiased and efficient amplification. The authors demonstrate remarkable results, showcasing simultaneous access to diverse multimedia files with high accuracy even at very low sequencing depths. The study is well-presented and comprehensive, making it accessible to a broad audience, including those interested in DNA storage and DNA design. I recommend minor revisions before accepting this paper in Nature Communications.

Response: Thank you for your positive feedback on our study. We appreciate your recognition of the significance of our work in the context of DNA data storage. Your suggestions have been effective in improving our work. Below are our point-by-point responses to your comments, with the response content in regular font size and the article content in a smaller font size, where original text is in **blue** and modified content is highlighted in **red**:

Specific Changes/Suggestions for Revisions:

- For the experimental validation of the strategy: a detailed error rate analysis and long-term stability, particularly at low sequencing depths, is valuable to build a better understanding of the types and sources of errors will aid in developing error-correction strategies. The authors have those sections buried in SI, they should consider incorporating those discussions from the Supplementary Information into the main manuscript, or within the methodology section.

Response: Thank you for your suggestion. We have now incorporated the discussion on error analysis and sources of errors into the results section of the main manuscript. Below is the updated text:

We also conducted a detailed error analysis of MPHAC-DIS (Supplementary Section 8). In the sequencing results, approximately 80% of the reads were completely correct, while errors existed in 20% of the reads (Fig. S7-2). Among the erroneous sequences, some appeared in multiple experiments (Fig. S8-1). Additionally, within the error sequences, certain sequences shared the same payloads (Fig. S8-5 to S8-8), and most of these payloads had relatively high GC content (>55%, Fig. S8-2). These sequences often formed secondary structures either within the payloads themselves or with the primer binding regions (Fig. S8-5 and S8-8 to S8-10). Further analysis of the relationship between errors and base positions revealed that the likelihood of errors increased toward the 3' end of the sequencing reads (Fig. S8-13 and S8-14). This may be related to the higher error rates associated with DNA synthesis and sequencing at longer lengths. Here, it's noted that sequencing reads are reverse complements of the synthesized sequences, errors introduced at the 5' end of the synthesized DNA appear at the 3' end of the sequencing reads. When analyzing the types of errors, we found that single-base substitutions were the most common, with G>T substitutions being the most frequent (Fig. 4f). Our results are consistent with Erlich's findings, and quantifying errors helps in understanding their sources as well as correcting them^{12,33}.

- Additional storage condition testing: While the current study evaluates buffer variations, it'll be important to expand testing or discussion to explore the system's performance under different temperatures, assess the impact of varying humidity levels on DNA stability, evaluate the effects of UV light exposure, particularly for long-term storage scenarios, and the potential of dry storage conditions, as explored in this example <https://www.nature.com/articles/s41467-019-09517-y>. Note that I am not affiliated with this study, but it's worth considering how extreme storage conditions could impact the DNA storage system's performance and data integrity. Evaluating/discussing its resilience under such conditions is essential for real-world applications.

Response: Thank you for your suggestion. In response, we have added a section in the results discussing the impact of environmental factors—such as temperature, humidity, and UV light exposure—on the long-term stability of DNA data storage. We have also explored potential strategies to enhance the system's resilience under extreme storage conditions. We hope this addition provides a more comprehensive understanding of the system's durability in real-world applications. Below is the updated text:

Environmental factors are also critical to the long-term stability of DNA data storage systems. Temperature fluctuations, for instance, can accelerate DNA degradation, with higher temperatures exacerbating hydrolytic and oxidative damage to DNA strands. Humidity also plays a significant role in DNA integrity, as excessive moisture introduces water molecules that promote hydrolysis, ultimately leading to data loss. Furthermore, extended exposure to UV light can trigger thymine dimer formation and other photochemical reactions, further compromising the stored DNA. Studies have demonstrated that dry storage, particularly when using encapsulation techniques, can effectively mitigate these risks by creating a stable, inert environment that shields DNA from external factors. Additionally, for DNA pools where some damage has occurred, data restoration techniques offer another solution. The combination of preservation strategies with restoration methods helps to ensure a more robust DNA storage system.

• **Benchmarking/Comparative analysis:** A significant drawback of the paper is the comparison with only one prior method (fixed 20nt primers), which the authors acknowledge isn't the only approach used in the past. Can the authors expand the analysis to include a wider range of existing primer design strategies, including those that consider factors like Tm (melting temperature). This would provide a more comprehensive understanding of its advantages and limitations in terms of storage density, retrieval accuracy, cost, and scalability (providing a comparative table like Table S1 including all these variables would help).

Response: Thank you for your insightful comment. In response, we have expanded our analysis to include a comparative evaluation of various primer design strategies, specifically addressing both fixed length with constrained melting temperatures and our fixed energy approach. We included a new table (Table S3-1) that details the simulated performance of these designs, demonstrating our method's superiority in terms of primer capacity and amplification homogeneity. The results indicate that designs based on fixed energy yield a significantly higher primer capacity and maintain consistent efficiency and correctness. We hope this addition provides a more comprehensive understanding of the advantages and limitations of our approach.

Table S3-1 Simulated performance comparison between two designs		
	FL+Tm ^a	FE ^b
Primer capacity ^c	2.53×10 ¹¹	1.67×10 ¹⁴
Efficiency range	78.2%-100%	>99%
Correct ratio ^d	87.08%	100%
Fold-80 base penalty	1.28	1

^a FL+Tm is defined here as a fixed length of 20 nt, with the Tm value constrained between 55°C and 60°C. This standard is derived from the work of Lee Organick et al¹.

^b FE refers to the limitation of the Gibbs free energy change, set between -10.5 kcal mol⁻¹ and -12.5 kcal mol⁻¹.

^c Primer capacity is defined as the maximum number of primers that can be generated. Initially, we generate 10,000 random primers and calculate the percentage that passes the filter, which is then used to derive the primer capacity.

^d Correct ratio indicates the percentage of successfully decoded files. For this analysis, we have established a sequencing depth of 50× and 30× coverage as the threshold for successful decoding.

Reference:

[1]. Organick, L., et al., Random access in large-scale DNA data storage. Nat. Biotechnol., 2018. 36(3): p. 242-248.

• Results Figure 3 and 4: They are very comprehensive and clear but consider adding the coefficient of variance (CV) to plot (d) and the curve in section (i), it would help quantify the variance between the two methods. Similarly, consider adding them to the histograms in Figure 4e and other relevant figures. Also, could the "A>>B" and "A=B" cases in section (h) be experimentally validated and compared to the theoretical predictions?

Response: Thank you for your valuable suggestion. In response to your first point, we have added the CV values to the relevant figure to help quantify the variance between the two methods. The updated version is provided below:

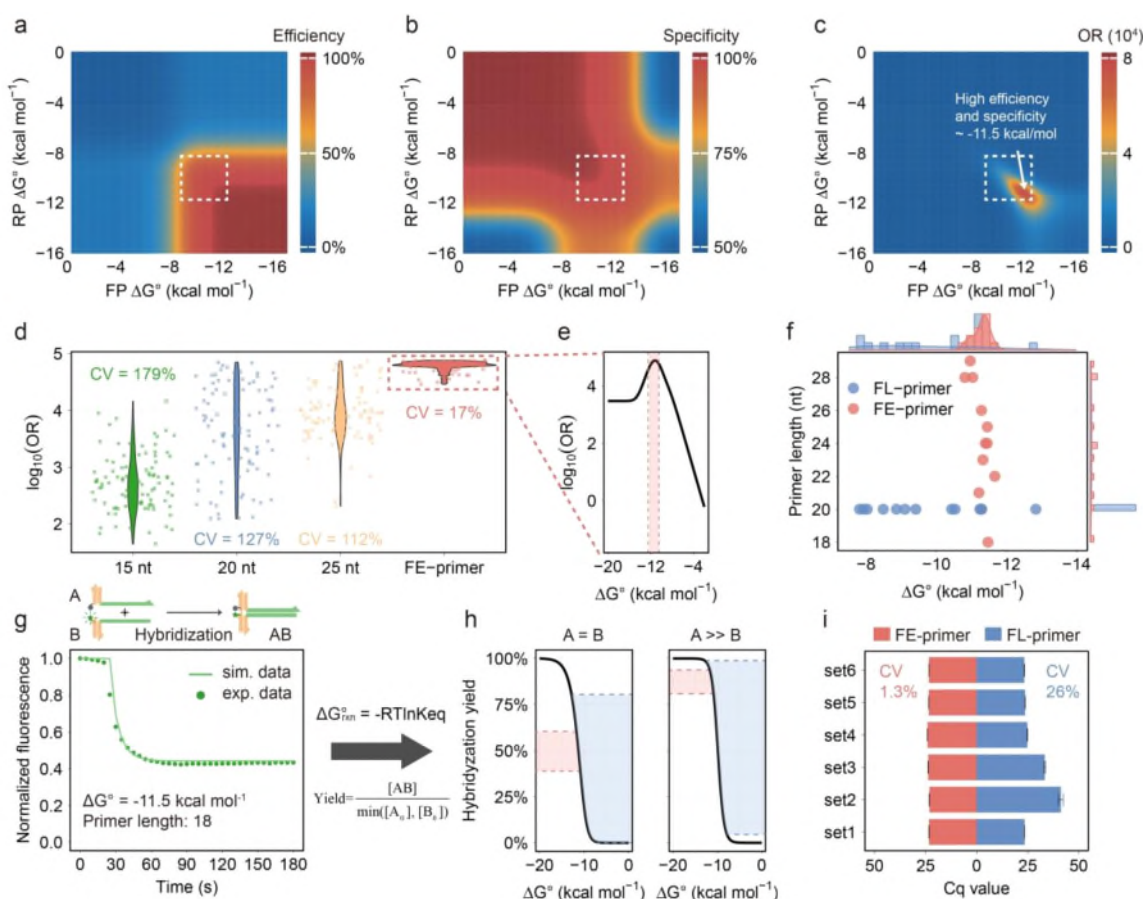


Fig. 3 | Parameters optimization of MPHAC. **a-c**, Simulated results of amplification efficiency (**a**), specificity (**b**) and odds ratio (OR) (**c**) under different primer free energies based on the theoretical model. Details of the model construction are provided in Supplementary Section 3. The white dashed box represents the free energy range for 80% of FL-primers in 100,000-plex primer pairs. Amplification efficiency positively correlates with free energy, while specificity exhibits the opposite trend. We considered the amplification efficiency and specificity together by OR, and the results suggest that the primer free energy can be optimized around -11.5 kcal mol⁻¹. **d, e**, OR values under different length FL-primers and FE-primers (**d**) and different primer free energies (**e**). Here, the coefficient of variation (CV) is a measure of the relative dispersion of a data set, expressed as the ratio of the standard deviation to the mean. **f**, Primer length and free energy distribution. Six sets of FE-primers and six sets of FL-primers were designed based on plasmid sequences. **g**, Monitoring primer hybridization fluorescence kinetics using X-probe. Experimental data are represented by dots, and simulated data is depicted by lines. **h**, Simulated hybridization yields for primers when A=B (left) and A>>B (right). Here, A refers to the primer, and B denotes the reverse complementary sequence of the primer (template). The range of hybridization yields for FE-primers is considerably narrower (pink) than that of FL-primers (blue). **i**, PCR results for FE- and FL-primers. Error bars represent the results of three parallel experiments. Here, Ct value (Cycle threshold value) represents the cycle number at which the fluorescence signal exceeds a defined threshold, indicating the presence of the target. Theoretically, identical concentrations of the target should result in the same Ct value.

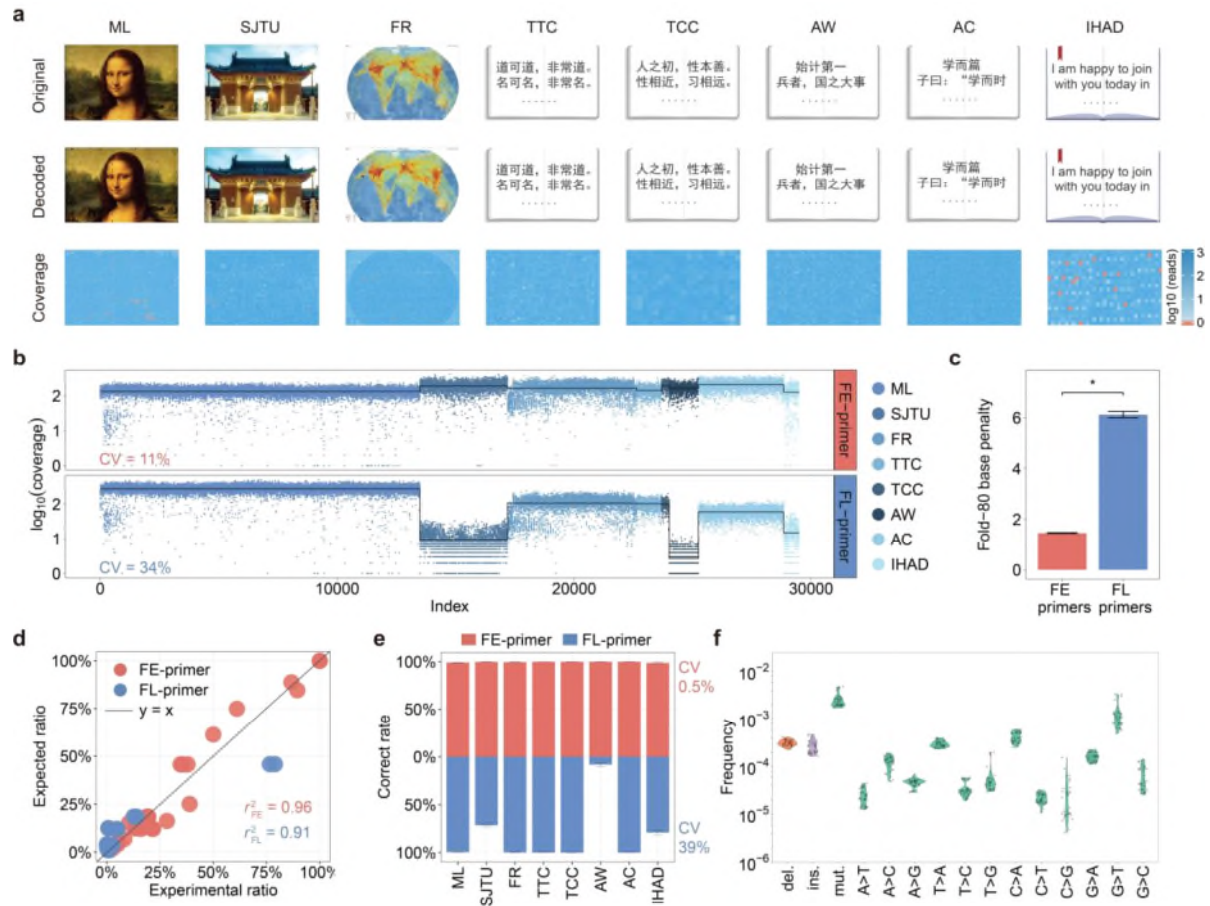


Fig. 4 | Homogeneous amplification and lossless index of MPHAC-DIS. We encoded files of multimodal data, including text, pictures, etc., based on the encoding principle (Supplementary Section 5). **a**, Simultaneous retrieval of 8 distinct files in the DNA pool using primer combinations. The top-to-bottom presentation includes original files, decoding results, and the distribution of sequence coverage, with red pixels indicating repeat reads lower than 5. Notably, the sequence coverage within each file is uniformly distributed, showcasing excellent intra-file homogeneity. **b**, Coverage comparison between FE-primers and FL-primers. The results illustrate a high degree of consistency and highlighting the outstanding inter-file amplification homogeneity of MPHAC-DIS. **c**, Fold-80 base penalty of FE-primers and FL-primers. Fold-80 base penalty is a metric used to assess coverage uniformity in sequencing experiments. It indicates how much more sequencing is needed for 80% of the targets to reach the mean coverage. The closer the Fold-80 base penalty is to 1, the better the coverage uniformity. Data are presented as mean $\pm$ standard error, with $n=2$ for each group. Statistical significance was determined using an independent two-sample t-test. Asterisks denote levels of statistical significance (* denotes $p < 0.05$). **d**, Correlation between the experimental ratios and the expected ratios of various files. Through multiple random-access experiments, we observed a linear relationship between the actual and expected ratio, providing additional confirmation of the inter-file amplification homogeneity of MPHAC-DIS. **e**, Decoding accuracy comparison between FE primers and FL primers. The correct rate for the majority of files exceeds 99%, underscoring the exceptional performance of MPHAC-DIS. **f**, Statistics on the frequency of sequence error types in MPHAC-DIS. Data points were obtained from the results of 30 libraries.

For the second point, we have experimentally validated the results for the case where $A = B$ (Fig. S4-6), and the experimental results are consistent with the simulation results, further confirming the credibility of the simulations. In the case of $A \gg B$, the initial concentration of template B is typically very low (e.g. fM level), making it difficult for the instrument to detect the signal, which is why we only present simulation results in this instance. However, the thermodynamic calculation methods we used have been previously validated^{1,2}, ensuring the accuracy of our calculated parameters. Additionally, there is also substantial prior work in this area^{3,4}, providing a solid theoretical foundation that supports our conclusions despite the lack of experimental validation for $A \gg B$.

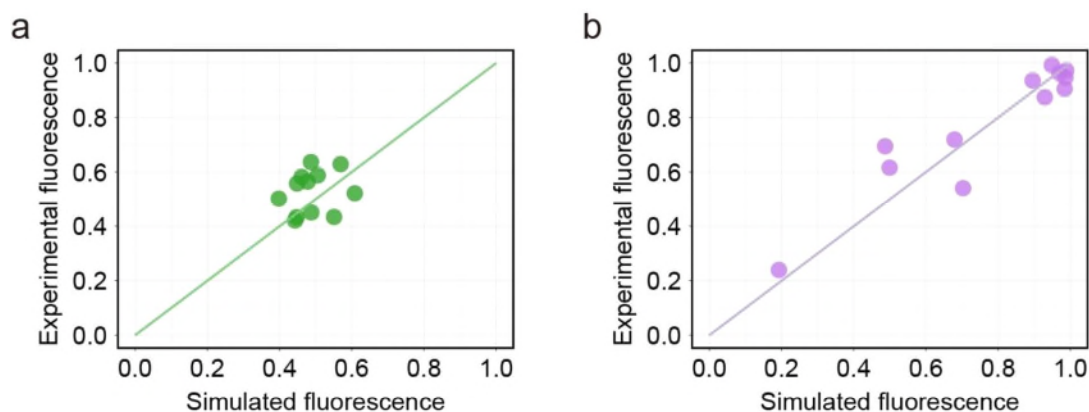


Figure S4-6. Endpoints Comparison of two groups. (a) FE-primer group, and (b) FL-primer group.

Reference:

- [1]. SantaLucia, J. and D. Hicks, The Thermodynamics of DNA Structural Motifs. Annual Review of Biophysics and Biomolecular Structure, 2004. 33(1): p. 415-440.
- [2]. Wang, C., J.H. Bae and D.Y. Zhang, Native characterization of nucleic acid motif thermodynamics via non-covalent catalysis. Nature Communications, 2016. 7(1).
- [3]. Zhang, D.Y., S.X. Chen and P. Yin, Optimizing the specificity of nucleic acid hybridization. Nature Chemistry, 2012. 4(3): p. 208-214.
- [4]. Wang, J.S. and D.Y. Zhang, Simulation-guided DNA probe design for consistently ultraspecific hybridization. Nature Chemistry, 2015. 7(7): p. 545-553.

- The Supplementary Information file is extensive and, while comprehensive, feels like a separate paper. It contains multiple discussion points that could be integrated into the main manuscript, as noted above. The reference to 5 figures (S2-6) on line 94 is an example of how 5 figures are lumped into one reference point and could be confusing and overwhelming. This large amount of data in the SI could be streamlined for clarity, ensuring each figure conveys a distinct idea. In many instances, the text refers to sections within the SI instead of specific figures. This is unclear for the reader and contributes to the perception of the SI as a separate publication rather than supporting material.

Response: Thank you for your valuable feedback. In response, we have updated the main manuscript to ensure that references point to specific figures in the Supplementary Information rather than referring to entire sections. We believe this adjustment will streamline the information and improve the overall clarity of the references. We would also like to clarify that the reference "S2-6" points specifically to the sixth figure in Section S2, rather than to five separate figures. We appreciate your suggestion and hope this revision enhances the readability and better integrates the Supplementary Information with the main text.

- Expand the discussion on the limitations of the method (where does it break?) and potential areas for future improvement.

Response: Thank you for your suggestion. We have expanded the discussion section to explore the limitations of our current approach and potential strategies for addressing these challenges in future research. Below is the updated text:

While the MPHAC-DIS method is effective for DNA information storage, it does have certain limitations. Firstly, although we provide a library of tens of thousands of primer pairs for random access, practical wet-lab experiments face challenges in handling such a large number simultaneously, which can lead to non-specific amplification and primer dimer formation. Beyond algorithmic solutions, primer dimers can be reduced using size selection with magnetic beads and enzymatic cleavage. Secondly, even tens of thousands of primer pairs are insufficient for large-scale DNA storage. Future advancements will require faster algorithms and methods capable of generating primers with higher capacity. Additionally, techniques such as isothermal amplification can be employed to improve

access accuracy during random access, and physical addressing remains a viable option. Finally, biases introduced during the DNA synthesis stage can affect the reliability of the storage system. To mitigate this, we can introduce error correction codes and incorporate partial redundancy to ensure decoding accuracy. Moreover, there is a need to develop more economical and higher-quality DNA synthesis techniques, such as enzymatic synthesis.

Response to reviewers' comments:

We sincerely thank the reviewers for their continued feedback. We appreciate their careful reading and valuable suggestions, which have helped us further improve the manuscript. We have addressed the remaining concerns, and detailed point-by-point responses are provided below. All changes are highlighted in **red**.

Responses to Reviewer #1:

I appreciate the authors for their careful revision. They have addressed most of my concerns and I believe this work is ready for publication.

Response: Thank you for your encouraging feedback. We are pleased that our revisions have successfully addressed your concerns, and we greatly appreciate your support for the publication of our work.

Additional minor suggestions:

1. For the original Question 14 regarding the creation of 2D rectangle for text files, please specify the methods in the main text, preferably in the Methods section.

Response: Thank you for your suggestion. In response, we have added the relevant details regarding the creation of 2D rectangles for text files in the Methods section of the main manuscript. We hope this provides a clearer understanding of the methodology used. Here is the updated version:

Visualization of Sequencing Coverage Results

To visually represent the sequencing coverage results, we employed the following approach. During the encoding phase, the information is split and stored across multiple sequences. After sequencing, the coverage of each sequence is calculated by aligning the reads, resulting in an $N \times 1$ vector, where N represents the number of sequences. To transform this vector into a two-dimensional format, it is rearranged into an $M \times K$ matrix, with M and K chosen to suit the visualization needs. For image files, this vector is rearranged to match the structure of the original image, allowing the coverage data to align visually with the image layout. Finally, a heatmap is generated from this matrix, with each point in the heatmap representing the coverage of a specific sequence, thereby providing a clear and intuitive visual representation.

2. There is a small mistake in Figure S2-6 of the revised version, where the top tick label on the Y-axis should probably be 1,000,000.

Response: Thank you for pointing out this oversight. We have corrected the Y-axis tick label in Figure S2-6, updating it to 1,000,000 to accurately reflect the data. We appreciate your attention to detail. Here is the revised version:

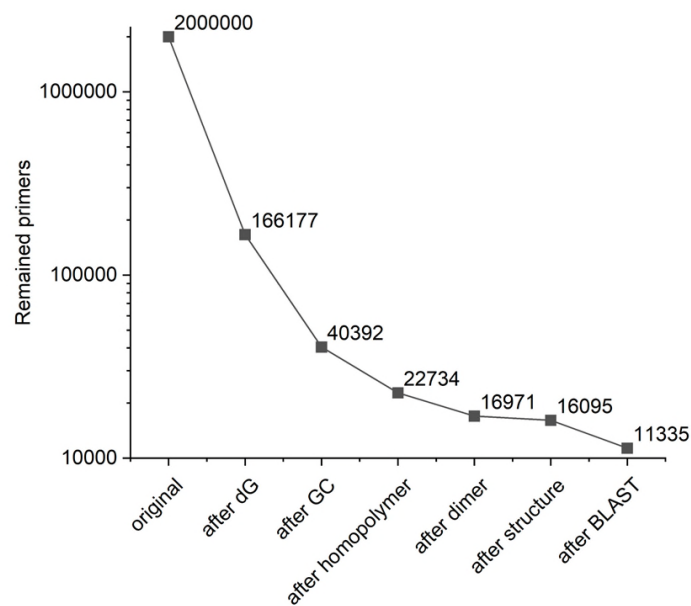


Figure S2-6. Number of primers remaining after each round of filtering steps. Here, 2,000,000 primers were randomly generated to be used as the initial primer pool and 11,335 primers were obtained after 6 rounds of filtering.

Responses to Reviewer #2:

The authors have made substantial improvements to the manuscript, effectively addressing the key concerns raised during the initial review. The revised version is now clearer and more cohesive, with enhanced explanations and a more streamlined presentation of the methodology. Importantly, the authors have provided more context for their findings, which makes the manuscript more compelling and positions it as a significant contribution to the field.

Response: Thank you for your positive feedback and recognition of the improvements we've made. We are pleased that the revised version has clarified our methodology and enhanced the overall cohesion of the manuscript.

However, one aspect still requires further clarification. While an additional check for primer specificity has been introduced to ensure orthogonality between primers, the possible occurrence of partially complementary sequences within the oligonucleotide outside the primer-binding region remains a concern. The authors should discuss how it might impact the uniformity of amplification or propose strategies to mitigate this issue. Once this discussion is addressed, the manuscript can be recommended for publication.

Response: Thank you for your insightful comment. We acknowledge the concern about partially complementary sequences outside the primer-binding region and their impact on amplification uniformity. In response, we have added a discussion in the manuscript addressing this issue. Here is the updated version:

While the MPHAC-DIS method is effective for DNA information storage, it does have certain limitations. Firstly, although we optimized and generated a library comprising tens of thousands of primer pairs through our primer design workflow, practical wet-lab experiments face challenges in handling such a large number simultaneously, which can lead to non-specific amplification and primer dimer formation. Beyond algorithmic solutions, primer dimers and longer non-specific amplicons can be reduced using size selection with magnetic beads and enzymatic cleavage. Secondly, even tens of thousands of primer pairs are insufficient for large-scale DNA storage. Future advancements will require faster algorithms and methods capable of generating primers with higher capacity. Additionally, techniques such as isothermal amplification can be employed to improve access accuracy during random access, and physical addressing remains a viable option. Thirdly, the potential occurrence of partially complementary sequences within the oligonucleotide outside the primer-binding region could also impact the uniformity of amplification. To address this, more flexible encoding strategies can be employed to avoid motifs related to primers and their complementary sequences within the payload. Finally, biases introduced during the DNA synthesis stage can affect the reliability of the storage system. To mitigate this, we can introduce error correction codes and incorporate partial redundancy to ensure decoding accuracy. Moreover, there is a need to develop more economical and higher-quality DNA synthesis techniques, such as enzymatic synthesis.

Responses to Reviewer #3:

The authors have carefully considered and addressed all the points raised. The manuscript has undergone significant changes, with improvements in clarity, organization, and scientific rigor. Overall, I think the revisions made the article suitable for publication with a great impact on the field and a compelling case for the use of MPHAC-DIS in DNA-based information storage. I recommend this manuscript for publication without further changes.

Response: Thank you very much for your encouraging feedback. We are pleased that the revisions have met your expectations and that you recognize the potential impact of our work. We greatly appreciate your recommendation for publication.