Maskless Array Photolithography Synthesis Simulator

Anna Novin and Adi Tzhori

September 2023

1 Introduction

DNA synthesis plays a pivotal role in modern molecular biology, enabling scientists to engineer and manipulate DNA for various applications, including the cutting-edge field of information storage in DNA. At the heart of molecular biology lies the ability to synthesize custom DNA sequences with precision. These sequences serve as the genetic code that defines the traits of living organisms.

The remarkable data storage potential of DNA has garnered significant attention in recent years. DNA molecules can store vast amounts of digital information in their chemical structure, offering a durable and compact storage medium that holds promise for long-term data archiving. However, unlocking this potential requires a deep understanding of DNA synthesis processes.

While the field of DNA synthesis has witnessed tremendous advancements, it is essential to acknowledge the cost constraints associated with experimental approaches. Traditional DNA synthesis techniques, including Maskless Array Synthesis (MAS) and others, often demand substantial financial investments. These costs encompass equipment, reagents, operational expenses, and ongoing maintenance, making them a barrier for many research projects.

To address the cost challenges and make research in information storage in DNA more accessible and sustainable, simulation emerges as an attractive alternative. Simulators provide a cost-effective means of studying DNA synthesis processes, allowing researchers to model and experiment with various scenarios without the need for expensive physical equipment and materials.

In this project, we have created a simulator tailored specifically for Maskless Array Synthesis (MAS). This simulator is intended to facilitate the study of MAS processes in the context of DNA storage research while alleviating the financial constraints associated with physical experimentation. Grounded in the latest research insights and findings, including those detailed in the study titled "Chemical and photochemical error rates in light-directed synthesis of complex DNA libraries" [1], our MAS simulator represents a methodical effort to address the cost limitations. It provides a practical, cost-effective, and scalable solution for researchers engaged in the investigation of information storage in DNA through the MAS technique.

2 Maskless Array Photolithography Synthesis

Maskless Array Synthesis (MAS) is a highly precise method of DNA synthesis that relies on a combination of chemistry and optics. At its core are thousands to millions of synthesis sites, each equipped with the necessary chemical reagents, including nucleotide building blocks (A, C, G, T), activators, and deprotecting agents.

The DNA synthesis process in MAS primarily hinges on controlled chemical coupling reactions. These reactions take place between the nucleotide building blocks and the growing DNA strands. To ensure precision, the nucleotide building blocks carry protective groups that temporarily shield their reactive sites.

What sets MAS apart is its use of optical control. Optical effects, including diffraction and scattering, play a pivotal role in directing the synthesis process. Photolithography is employed to precisely direct light onto specific synthesis sites. The coordinated use of mirrors, mirror size, and separation between mirrors ensures that light is accurately directed. This precise optical control allows for the selective activation of synthesis sites, enabling the synthesis of custom DNA sequences.

A crucial element in MAS is the use of 5'-photosensitive protective groups. These groups temporarily shield the reactive sites on the nucleotide building blocks, preventing unintended reactions during the coupling process. However, during photolithography, exposure to light selectively removes these protective groups, allowing the desired chemical coupling to occur and facilitating the extension of the DNA sequence.

Capping reactions are employed after each nucleotide addition to minimize undesired side reactions. These reactions help ensure the accuracy and fidelity of the synthesized DNA strands.

In summary, MAS relies on the intricate interplay of chemistry and optics to enable the precise creation of custom DNA sequences. Its high precision, adaptability, and flexibility make it a valuable tool in various scientific fields, including molecular biology, genomics, proteomics, and DNA-based information storage. Understanding the scientific principles behind MAS is essential for optimizing its processes effectively and managing associated costs.

[1] https://doi.org/10.1093/nar/gkab505

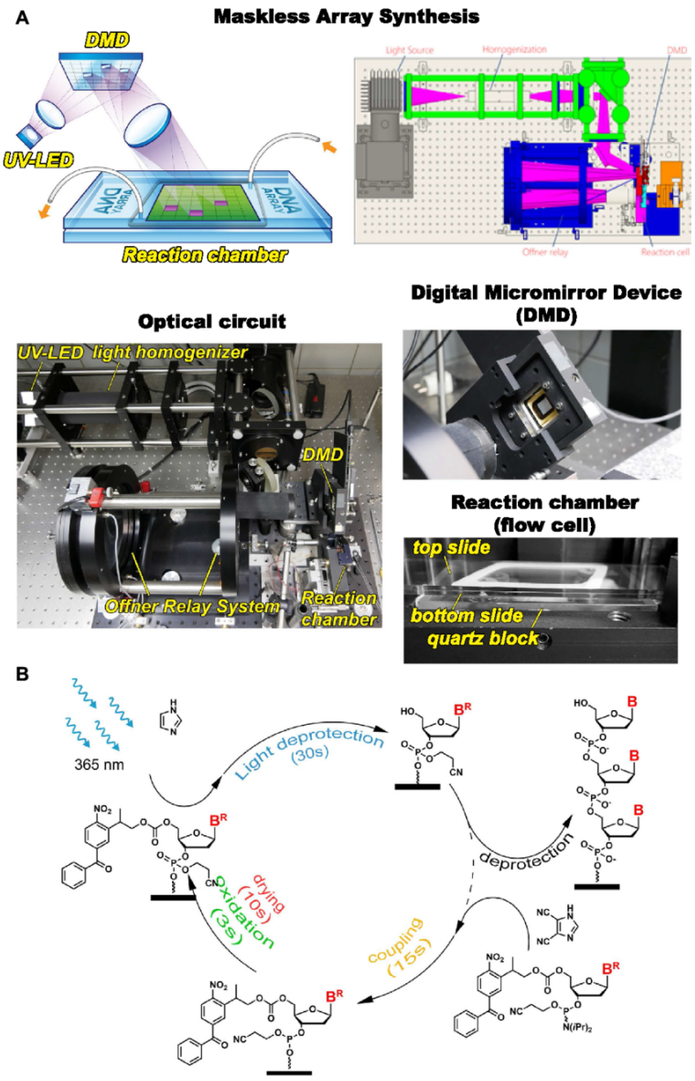

Figure 1: Schematic overview of Maskless Array Synthesis of DNA Libraries

3 Maskless Array Photolithography Synthesis Simulator

We conducted an in-depth analysis of Maskless Array Synthesis (MAS) and the error rates detailed in the referenced article, utilizing these insights to develop a simulator tailored to MAS operations. This simulator enables users to manipulate various critical parameters, simulating both device settings, library configurations and error rates.

Within the simulator, various types of errors in DNA synthesis can be simulated, including deletions, insertions, and substitutions. Deletion errors involve the omission of one or more nucleotide bases, resulting in gaps or missing information. Insertion errors, on the other hand, entail the addition of extra nucleotide bases, altering the sequence's length. Substitution errors encompass the replacement of one nucleotide base with another, potentially leading to genetic code and protein structure changes.

It's worth noting that the simulator is designed in a modular fashion, allowing for seamless integration of additional data, such as extra parameters or settings, as they become available.

3.0.1 Simulator User Interface

The simulator's user interface is depicted as follows:

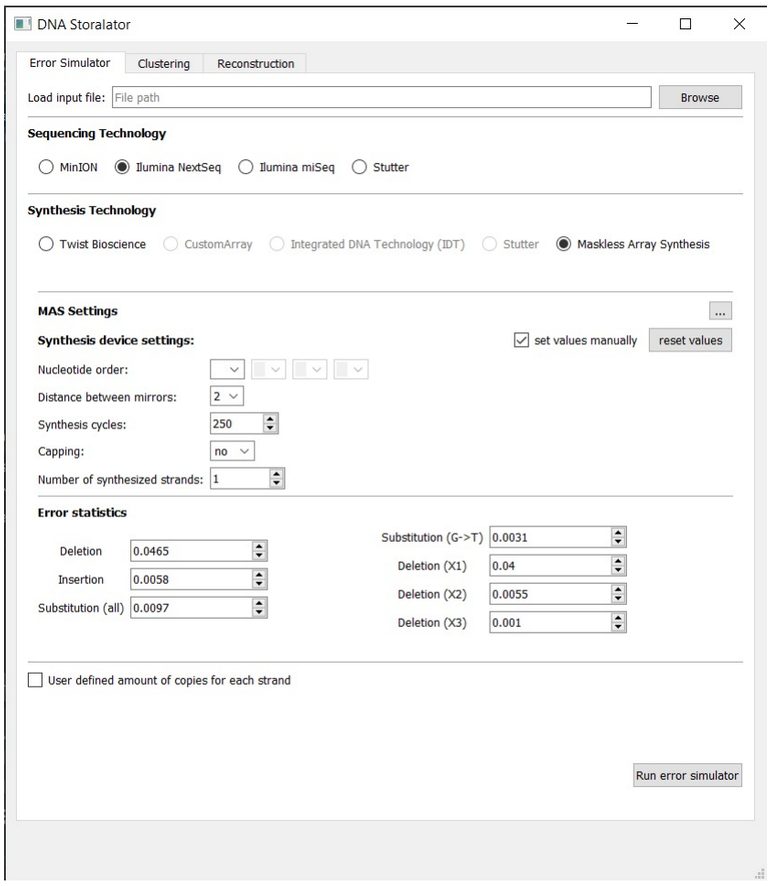

Figure 2: User Interface of Maskless Array Photolithography Synthesis Simulator

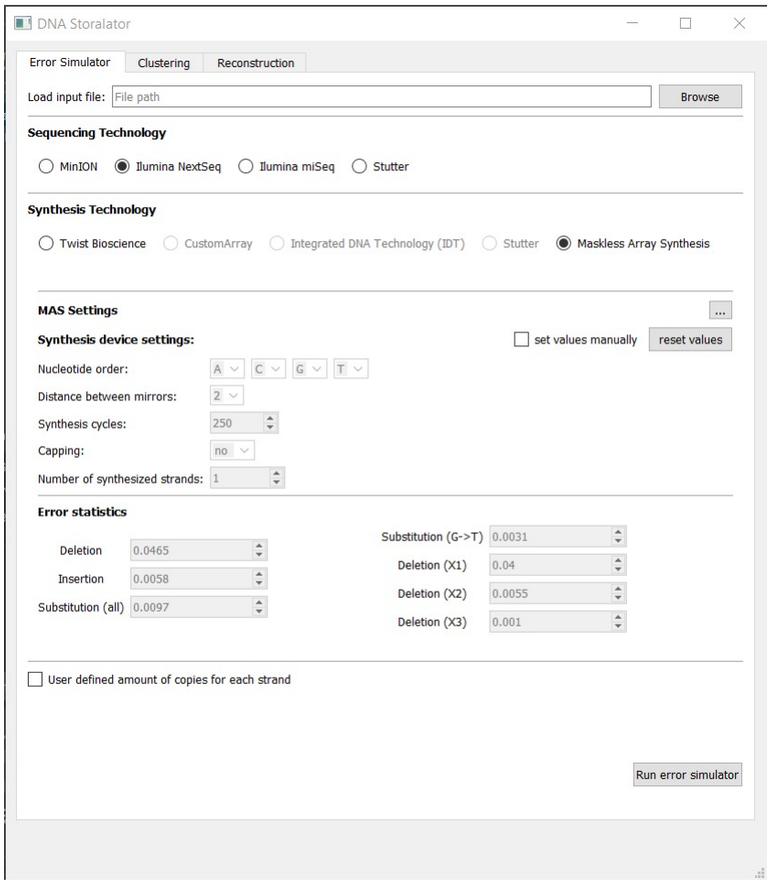

Figure 3: User Interface of Maskless Array Photolithography Synthesis Simulator

3.0.2 Simulator Configuration Options

The following configuration options can be precisely controlled and adjusted within our MAS simulator.

- Device and Library Configuration Options
  - Nucleotide order (default: A, C, G, T): Users can customize the order in which nucleotides are synthesized, allowing for versatile experimentation.
  - Distance Between Mirrors (default: 2): Users can customize the distance between mirrors.
  - Synthesis Cycles (default: 250): Users have the flexibility to modify the number of synthesis cycles, simulating variations in the synthesis process.
  - Capping (default: no): Users can opt to include or exclude capping reactions.

- Error Rates Configuration Options
  - Average Total Error Rate (default: 0.063)
  - Deletion (default: 0.0465)
  - Insertion (default: 0.0058)
  - Substitution (all) (default: 0.0097)
  - Substitution ($G \rightarrow T$) (default: 0.0031)
  - Deletion (x1) (default: 0.04)
  - Deletion (x2) (default: 0.0055)
  - Deletion (x3) (default: 0.001)

  The default error rates settings are set according to the device and library settings.

3.0.3 Simulator Code Flow

The simulator's code flow can be described as follows: it tracks two key variables — idx, representing the current position in the target strand, and `synthesis_step`, signifying the ongoing synthesis step. Simultaneously, an empty string named `synthesized_strand` serves as the repository for the progressively generated strand.

The heart of the simulation lies in the synthesis step. When the current nucleotide (nt) in the target strand matches the expected nucleotide based on the current synthesis step, the simulator introduces randomness. It generates a random number r between 0 and 1 and compares it to predefined deletion error rates (dx1, dx2, dx3). Depending on r, it may skip one, two, or three nucleotides or, if no deletion occurs, the current nucleotide is added correctly to the `synthesized_strand`.

Conversely, when nt doesn't match the expected nucleotide, the simulator calculates the total error rate as the sum of insertion and substitution error rates. It then generates a random number r between 0 and 1 and evaluates whether it's less than the total error rate. If not, no error occurs, and no nucleotide is added to the `synthesized_strand`. If r is less than the total error rate, the simulator chooses between insertion and substitution errors based on their weighted probabilities. For insertion errors, it adds the expected nucleotide for the current synthesis step to the `synthesized_strand`. For substitution errors, it adds the expected nucleotide and increments idx by 1.

The simulator cycles through the synthesis order at each step, ensuring that the correct order of nucleotides is followed. After processing the entire target strand, the `synthesized_strand` is appended to the result.

For a visual representation of this code flow, please refer to the figure presented on the next page.

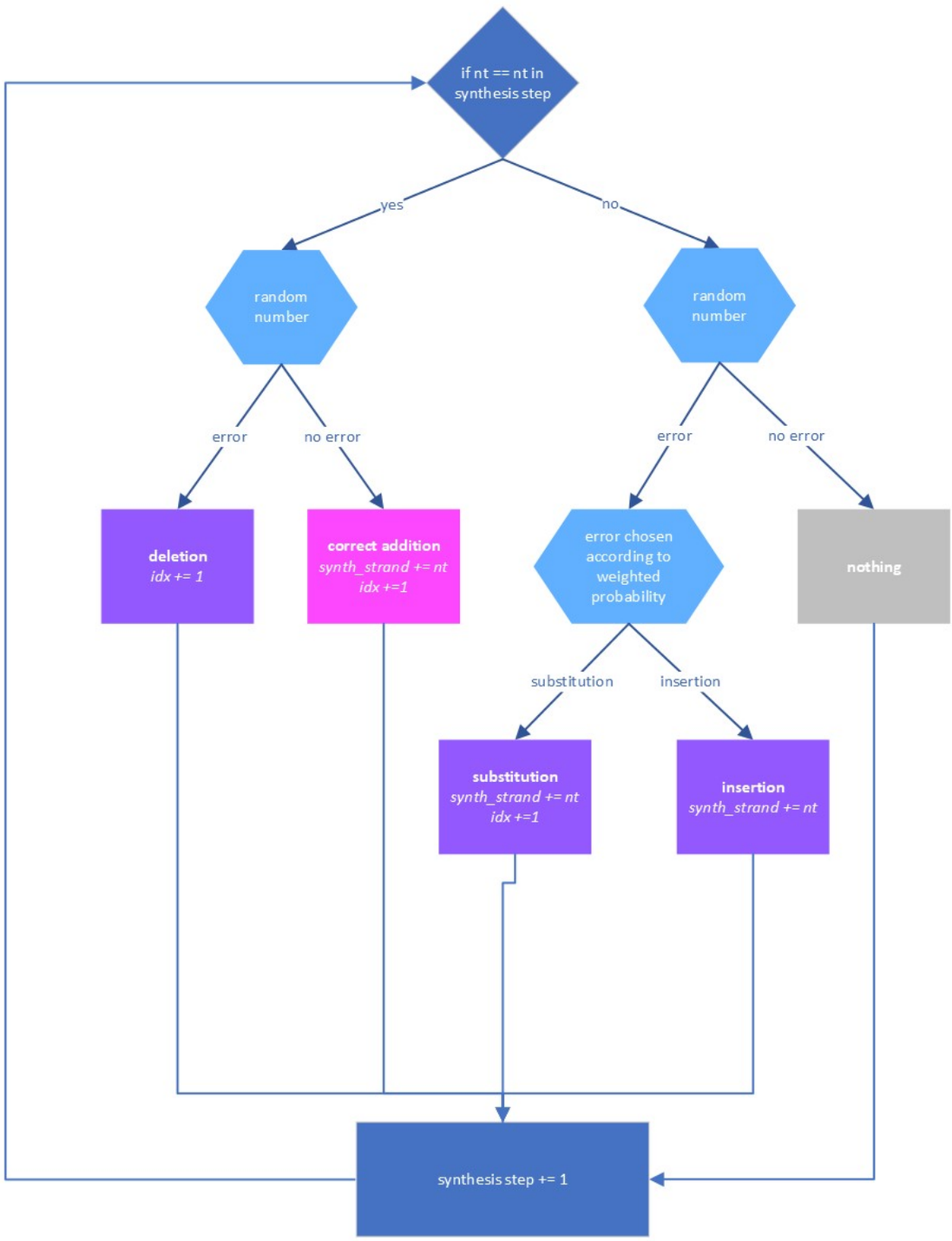

Figure 4: Code Flow of Maskless Array Photolithography Synthesis Simulator

4 Empirical Evaluation

4.1 Methodology

Our evaluation of code accuracy employed a methodical analysis of error rates, entailing a meticulous comparison between input strands and their corresponding output counterparts. The reported results represent the average outcomes derived from the synthesis of 1000 output strands for each unique input strand.

In the initial phase, we scrutinized errors in isolation, examining each error type independently. Subsequently, we undertook an analysis that considered the cumulative effects of all errors. This comprehensive testing methodology ensures a rigorous and scientifically valid assessment of the code's performance and reliability.

4.2 Systematic Error Analysis: Evaluating Individual Errors

In the following tests, we conducted a thorough analysis of individual error rates. This involved a focused evaluation of a single error type, with all other potential errors being assigned a probability of 0.

The error rates we assessed encompassed a range of values, [0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.2], and pertained to deletion, insertion, and substitution errors.

It's important to note that each reported result signifies the average error rate obtained from synthesizing 1000 strands for a single input strand. This rigorous approach ensures a comprehensive examination of error rates across a broad spectrum of scenarios.

Test Results for Deletion Error Rates

In the table below, we exclusively evaluated deletion error rates, aligning the set error rate values against their corresponding measured counterparts, along with their respective standard deviations. This focused analysis provides a clear insight into the accuracy of the deletion error rates.

Test Results for Deletion Error Rate		
Set Error Rate	Measured Error Rate	Standard Deviation
0.01	0.0094	0.0004
0.02	0.0200	0
0.03	0.0296	0.0002
0.04	0.0399	0
0.05	0.0500	0.0001
0.06	0.0592	0.0005
0.07	0.0690	0.0007
0.08	0.0784	0.0011
0.09	0.0887	0.0009
0.1	0.1013	0.0009
0.11	0.1085	0.001
0.12	0.1216	0.0012
0.13	0.1311	0.0008
0.14	0.1406	0.0005
0.15	0.1494	0.0004
0.16	0.1622	0.0016
0.17	0.1686	0.001
0.18	0.1806	0.0005
0.19	0.1912	0.0009
0.20	0.1980	0.0014

Test Results for Insertion Error Rates

In the table provided below, we conducted a focused assessment solely on insertion error rates. This involved aligning the preset error rate values with their corresponding measured values, and also accounting for their respective standard deviation

Test Results for Insertion Error Rate		
Set Error Rate	Measured Error Rate	Standard Deviation
0.01	0.0092	0.0005
0.02	0.0191	0.0006
0.03	0.0277	0.0016
0.04	0.0373	0.0018
0.05	0.0462	0.0026
0.06	0.0560	0.0028
0.07	0.0642	0.0041
0.08	0.0750	0.0035
0.09	0.0848	0.0036
0.1	0.0913	0.0061
0.11	0.1037	0.0044
0.12	0.1125	0.0053
0.13	0.1236	0.0045
0.14	0.1296	0.0073
0.15	0.1406	0.0066
0.16	0.1494	0.0075
0.17	0.1579	0.0085
0.18	0.1682	0.0083
0.19	0.1806	0.0066
0.20	0.1898	0.0071

Test Results for Substitution Error Rates

In the following table, our examination was concentrated solely on substitution error rates. We aligned the predetermined error rate values with their corresponding measured values and accounted for their respective standard deviations.

Test Results for Substitution Error Rate		
Set Error Rate	Measured Error Rate	Standard Deviation
0.01	0.0093	0.0005
0.02	0.0183	0.0012
0.03	0.0284	0.0011
0.04	0.0385	0.001
0.05	0.0472	0.0019
0.06	0.0583	0.0012
0.07	0.0676	0.0017
0.08	0.0786	0.001
0.09	0.0867	0.0023
0.1	0.0989	0.0007
0.11	0.1074	0.0018
0.12	0.1182	0.0012
0.13	0.1309	0.0006
0.14	0.1393	0.0004
0.15	0.1482	0.0012
0.16	0.1582	0.0012
0.17	0.1685	0.001
0.18	0.1798	0.0001
0.19	0.1915	0.0011
0.20	0.2002	0.0002

Conclusions for Tests for Individual Errors

The tables above reveal a consistent trend: the measured error rates closely mirror the set error rates across all conducted tests for deletion, insertion and substitution error rates. This agreement is further substantiated by the calculated standard deviations for each set of results.

4.3 Comprehensive Error Analysis: Assessing Multiple Errors

In the following test, we conducted a comprehensive analysis of multiple error rates simultaneously. We established identical error rates for deletion, insertion, and substitution, selecting values from the range [0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.2].

Each reported result represents the average error rate derived from synthesizing 1000 strands for a single input strand. This meticulous approach ensures a thorough examination of combined error rates across a wide range of scenarios, providing robust and comprehensive insights into our analysis.

Test Results for Multiple Errors

In the subsequent table, we provide a detailed account of our comparison between the predefined error rates and the measured error rates. This examination encompasses error rates set and measured concurrently for deletion, insertion, and substitution. Additionally, we include the respective standard deviations for a comprehensive understanding of the variability in our findings.

Test Results for Combined Error Rates								
Set Error Rates			Measured Error Rates			Standard Deviation		
del	in	sub	del	in	sub	del	in	sub
0.01	0.01	0.01	0.0100	0.0090	0.0094	0	0.0006	0.0004
0.02	0.02	0.02	0.0193	0.0191	0.0182	0.0004	0.0006	0.0013
0.03	0.03	0.03	0.0288	0.0283	0.0287	0.0008	0.0012	0.0009
0.04	0.04	0.04	0.0382	0.0377	0.0380	0.0012	0.0016	0.0014
0.05	0.05	0.05	0.0474	0.0484	0.0476	0.0018	0.0011	0.0016
0.06	0.06	0.06	0.0563	0.0582	0.0589	0.0026	0.0013	0.0008
0.07	0.07	0.07	0.0658	0.0686	0.0680	0.0029	0.0009	0.0014
0.08	0.08	0.08	0.0728	0.0784	0.0769	0.005	0.0011	0.0011
0.09	0.09	0.09	0.0816	0.0885	0.0873	0.0059	0.001	0.0019
0.1	0.1	0.1	0.0885	0.0982	0.0966	0.0081	0.0012	0.0023
0.11	0.11	0.11	0.0973	0.1087	0.1082	0.009	0.0009	0.0013
0.12	0.12	0.12	0.1050	0.1178	0.1187	0.0106	0.0106	0.0009
0.13	0.13	0.13	0.1110	0.1256	0.1298	0.0134	0.0031	0.0001
0.14	0.14	0.14	0.1213	0.1372	0.1399	0.0132	0.002	0
0.15	0.15	0.15	0.1280	0.1482	0.1484	0.0155	0.0013	0.0011
0.16	0.16	0.16	0.1345	0.1610	0.1604	0.018	0.0007	0.0003
0.17	0.17	0.17	0.1418	0.1704	0.1704	0.0199	0.0003	0.0003
0.18	0.18	0.18	0.1492	0.1826	0.1792	0.0218	0.0019	0.0005
0.19	0.19	0.19	0.1539	0.1877	0.1903	0.0255	0.0016	0.0002
0.20	0.20	0.20	0.1607	0.2025	0.1987	0.0277	0.0018	0.0009

Conclusions for Tests for Multiple Errors

The table above illustrates that the measured combined error rates for deletion, insertion, and substitution closely align with the predefined combined error rates across all tested combinations. This alignment is further reinforced by the calculated standard deviations for each set of results, indicating the consistency and reliability of our findings.

5 Conclusions

In this project, we introduced a simulator tailored for the Maskless Array Synthesis (MAS) method. This simulator serves as a rapid and cost-effective alternative to conducting actual MAS syntheses, making it an invaluable tool for advancing research in DNA storage. By mitigating the financial constraints linked to physical experimentation, it opens new avenues for in-depth studies of MAS processes. We conducted rigorous testing of the simulator's accuracy by analyzing individual and combined error rates, comparing the results to predefined values. The outcomes of these tests revealed a remarkably close alignment between the predefined and measured values, affirming the consistency and reliability of our simulator. As a versatile tool for exploring the complexities of MAS, our simulator holds promise in advancing our understanding of DNA synthesis, error rates, and their implications.