Exploring protein sequence space for mechanistic studies and functional improvements

Zhai, Lili

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Exploring protein sequence space for mechanistic studies and functional improvements

Ph.D. Thesis

Lili Zhai

Department of Biotechnology and Biomedicine (DTU Bioengineering)

Technical University of Denmark

February 2023

Contents

Preface

The sequence space of proteins is vast due to the 20 different amino acids that make up these biopolymers. Even small proteins have an astronomical number of possible amino acid sequences. Moreover, a single amino acid change can lead to significant changes in protein properties and functions. Therefore, the exploration of protein sequence space is essential for two key reasons. Firstly, targeted changes in amino acid sequence can be used to improve the function of a given protein. Secondly, conducting experiments on how a process that the protein can undergo is affected by a specific amino acid change leads to improved mechanistic understanding. Exploring protein sequence space is crucial for various fields, including drug discovery, protein engineering, and basic research, as it offers opportunities to design proteins with desirable characteristics, investigate the relationship between protein sequence, structure, and function, and develop a deeper understanding of protein behavior and mechanisms. In this study, we performed protein sequence space explorations with two purposes: functional improvement and mechanistic study, using two distinct proteins, namely KOD DNA polymerase and α-synuclein. Therefore, this PhD project encompasses two topics: 1) DNA polymerase optimization by semi-rational sequence space exploration; and 2) Towards mechanistic studies of α-synuclein assembly by mRNA display.

Due to the global pandemic and its rapid spread, my PhD studies at DTU were seriously postponed. As a result, I had to delay my arrival at DTU until the third year of the program, leading to a significant delay in exploring protein sequence space in α-synuclein assembly using mRNA display. After consulting with my primary supervisor, Alexander K. Buell, and co-supervisor, Qingqing Xie at BGI, we decided to introduce an alternate sub-topic related to the directed evolution of DNA polymerase at BGI. This work on KOD DNA polymerase was carried out at BGI, and the work on α-synuclein was carried out at DTU during a 6-month stay in 2022.

This PhD project started in November 2019 as a collaboration between the Department of Biotechnology and Biomedicine (DTU Bioengineering), Technical University of Denmark, and BGI-Shenzhen. The work presented in this PhD thesis was carried out under the supervision of Professor Büll (main supervisor), Professor Laustsen-Kiel (co-supervisor), and Doctor Xie (co-supervisor). Projects included in this thesis were supported by DTU Bioengineering and Shenzhen Engineering Laboratory for Molecular Enzymology (No. DRC-SZ [2018]958), the National Natural Science Foundation of China, China Grant No. 21505134, and the Novo Nordisk Foundation (NNFSA170028392).

Publications

The list of manuscripts:

1 Lili Zhai, Zi Wang, Liu Fen, Chongjun Xu, Jingjing Wang, Hongyan Han, Qingqing Xie, Wenwei Zhang, Yue Zheng, Alexander K. Buell, Yuliang Dong. Semi-rational evolution of a recombinant DNA polymerase for catalytic activity on modified nucleotide incorporation. Submitted to BioRxiv (2023).

2 Ruyin Cao, Lili Zhai (co-first author), Qingqing Xie, Zi Wang, Yue Zheng, Wenwei Zhang, Alexander K. Buell, Xun Xu, Yuliang Dong, Chongjun Xu, Wenping Lyu. Rational evolution of a recombinant DNA polymerase for efficient incorporation of unnatural nucleotides by dual site boosting. Submitted to BioRxiv (2022).

3 Lili Zhai, Soumik Ray, Antonin Kunka, Alexander K. Buell. Exploratory experiments to probe the interactions between α-Syn and nucleic acids. To be submitted (2023).

The list of patents:

1. Lili Zhai, Qingqing Xie, et al. Recombinant KOD DNA polymerase

   Patent No: WO2022082482A1

2. Lili Zhai, Qingqing Xie, et al. Thermostable b-family DNA polymerase mutant and application thereof

   Patent No: WO2022247055A1

Acknowledgment

First, I would like to express my appreciation to my principal supervisor, Alexander K. Buell, co-supervisor, Professor Andreas Hougaard Laustsen-Kiel, and co-supervisor, Doctor Qingqing Xie, for offering me guidance and support during my PhD program. Great thanks to the BGI-DTU collaborative PhD program for offering me an opportunity to be a PhD student at DTU. Academically, I have obtained a lot of experience and knowledge about protein sequence space effect on mechanistic studies and functional improvements during the mentoring and discussion about my projects. More importantly, I have learned a lot from Alexander's rigorous attitude and creative thinking in scientific study, which I believe will influence me a lot in my future career.

I also give special thanks to my managers at BGI Research Institute, Yue Zheng, Yuliang Dong, and Wenwei Zhang, who keep providing resources to me for building the lab and doing experiments and are always encouraging and considerate.

I would like to thank my BGI and MGI colleagues Ruyin Cao, Zi Wang, Hui Zhang, Fen Liu, Jingjing Wang, and other members for working together with me during the program. We had beautiful times together. In addition, many thanks to Yue Zheng, and Qingqing Xie for providing constructive suggestions for my thesis.

I would also like to thank my DTU colleagues Soumik Ray, Rasmus Krogh Norrild, Seyedazad Farzadfard, Antonin Kunka, Suk Kyu Ko, Jacob Aunstrup Larsen, Louise Kjær Klausen, and Shuangyang Wang for helping me to conduct experiments and other parts of the PhD project.

Finally, I would like to thank my parents and friends for their encouragement. I deepest gratitude to my husband, Man Zhang, for his patience and full support during my thesis writing. Besides that, his optimistic attitude has influenced me, which has played a positive role in my persistence in completing my PhD project and has influenced my future life and study.

List of abbreviations

KOD pol: KOD DNA polymerase

PT-2Cy5: The primer and template duplex modified with dye Cy5 at both 5' terminal

Modified dATP: 3'-O-azidomethyl-dATP with dye Cy3 labeled

WT: Wild type

FRET: Fluorescence resonance energy transfer

ML: Machine learning

CSR: Compartmentalized Self-Replication

MM-GBSA: Molecular Mechanics Generalized Born Surface Area

DNBSEQ^{TM}: DNA NanoBalls technology

CoolMPS^{TM}: Massively Parallel Sequencing Chemistry

NGS: Next-generation Sequencing

α-Syn: α-synuclein

LLPS: Liquid-liquid phase separation

ssDNA: Single-strand DNA

dsDNA: Double strands DNA

R_h: Hydrodynamic radius

FIDA: Flow-Induced Dispersion Analysis

ThT: Thioflavin-T

ANS: 8-anilinonapthalene-1-sulfonic acid

MD: Molecular dynamics

SASAs: Solvent-accessible surface areas

FRAP: Fluorescence recovery after photobleaching

IDP: Intrinsically disordered proteins

ROI: Regions of interest

Abstract in English

The exploration of the vast protein sequence space is critical for our understanding of the mechanisms underlying the living world. Despite the significant challenges posed by the comprehensive exploration of this sequence space, it remains an essential approach to gain insights into novel functional proteins and advance our understanding of their mechanisms. This study aims to explore protein sequence space using different proteins, KOD DNA polymerase and α-synuclein, for functional improvements and mechanistic studies. To achieve this goal, we have divided our research into two distinct sub-projects that will be introduced in detail below.

In the first sub-project, primarily, we employed a semi-rational strategy to engineer a natural DNA polymerase (KOD pol) from *Thermococcus kodakaraensis* for improving catalytic efficiency for modified nucleotides, from which we gain insight into the link between sequence modifications and protein function. We designed hundreds of mutants, covering the finger, palm, thumb, and N-terminal domains, based on the analysis of the crystal structure of KOD pol. To efficiently screen libraries of mutants, we developed a high-throughput screening method based on FRET technology in a microplate reader. Through single-point and combinatorial mutations, we obtained a best-performing variant carrying eleven mutation sites. The best-performing variant demonstrated satisfactory performance on both the BGISEQ-500 and MGISEQ-2000 platforms. Subsequently, we developed a machine learning (ML) based virtual screening method trained on the kinetic data of hundreds of KOD variants and designed a virtual library with dual-site mutations based on the best-performing variant identified during the semi-rational screening process. We experimentally screened about 22% of the variants within this virtual library. The experimental results indicated that over 80% of the mutants exhibited higher catalytic efficiency compared to the parent strain. Among them, the best-performing variant KH showed an improvement in kinetic efficiency by more than an order of magnitude. We also analyzed the mechanism of the conformational changes in the finger domain at positions 485 and 451. These findings on the interactions of sequence, structure, and function of KOD pol, expand our so far limited knowledge of B-family DNA polymerases in their interactions with modified substrates and further our understanding of their ability to accept a wide range of modified nucleotides. The enzyme evolution technology and strategy in this study also have implications for guiding the design of other DNA polymerases for a broad range of biotechnological applications.

In the second sub-project, we explored the use of mRNA display for the study of α-synuclein

aggregation, which is related to Parkinson's disease (PD). The aim was to leverage the extremely high throughput nature of mRNA display in such a context for the first time. The idea behind this approach is to study the aggregation and phase separation of a very large number of sequence variants at the same time by incubating an mRNA-displayed sequence library of α-synuclein with an excess of aggregating/phase separating WT protein, followed by physical separation of the aggregates/droplets from the soluble protein and sequencing of the different pools to determine the fraction of the library associated with each pool. An important aspect determining the feasibility of this approach is the type and strength of the interactions between the α-synuclein and mRNA. Therefore, as a preliminary study of the mRNA-display approach, we first investigated the interactions between α-synuclein and DNA. We chose DNA as a proxy for RNA because it is easier to handle and because of the existing literature on α-synuclein-DNA interactions with which we wanted to compare our results.

We focused on the interactions between α-Syn and different types of DNA, including ssDNA with different lengths, dsDNA, and genomic DNA. The binding affinities of the monomeric α-Syn and DNA under the given conditions were measured using flow-induced dispersion analysis (FIDA). The approximate K_d values were calculated and compared, these results suggested that the binding affinity was weak and generally in the tens of micromolar range. In addition, liquid-liquid phase separation of α-Syn was carried out in the presence and absence of ssDNA and dsDNA to explore to what extent the interactions between DNA and α-Syn would influence this process. Interestingly, complex coacervates of α-Syn with both ssDNA and dsDNA formed, and ssDNA/dsDNA-PEG formed suspended drops in a larger α-Syn condensate. In addition, α-Syn can exhibit a liquid-to-solid phase transition due to aggregation even in the complex coacervates, whereas both ssDNA and dsDNA remained completely liquid-like even after 48h. These results could provide insights into the pathogenesis of neurodegenerative diseases associated with α-Syn and may have implications for the development of new therapeutic strategies. Additionally, we made the first steps towards the mRNA display method using wild-type α-synuclein, such as a demonstration of successful ligation of puromycin to mRNA. The next steps would be to generate an appropriate mRNA library of α-synuclein variants and perform cell-free expression to generate the mRNA-displayed library followed by screening and identifying these variants with high specificity and affinity for well-defined soluble or aggregated states. This exploration of the protein sequence space of α-synuclein has the potential to provide mechanistic insights into its role in neurodegenerative disorders at an unprecedented scale.

Abstract in Danish

Undersøgelse af det enorme proteinsekvensrum er afgørende for vores forståelse af de mekanismer, der ligger til grund for den levende verden. På trods af de betydelige udfordringer, som omfattende undersøgelse af sekvensrummet udgør, er det fortsat en essentiel tilgang for at få indsigt i nye funktionelle proteiner og fremme vores forståelse af deres mekanismer. Dette studie har til formål at udforske proteinsekvensrum ved hjælp af forskellige proteiner, KOD DNA-polymerase og α-synuclein, til funktionelle forbedringer og mekanistiske undersøgelser. For at opnå dette har vi opdelt vores forskning i to adskilte delprojekter, som vil blive introduceret i detaljer nedenfor.

I det første delprojekt brugte vi primært en semi-rationel strategi til at konstruere en naturlig DNA-polymerase (KOD pol) fra *Thermococcus Kodakaraensis* til optimering af katalytisk effektivitet for modificerede nukleotider, hvorfra vi får indsigt i sammenhængen mellem sekvensmodifikationer og protein funktion. Vi designede hundredvis af mutanter, der dækkede finger-, håndflade-, tommelfinger- og N-terminal domænerne, baseret på analysen af krystalstrukturen af KOD pol. For effektivt at screene biblioteker af mutanter udviklede vi en high-throughput screeningsmetode baseret på FRET-teknologi i en mikropladelæser. Gennem enkeltpunkts- og kombinatoriske mutationer opnåede vi en bedst ydende variant med elleve mutationssteder. Den bedst ydende variant viste tilfredsstillende effektivitet på både BGISEQ-500- og MGISEQ-2000-platformene. Efterfølgende udviklede vi en maskinlæring (ML) baseret virtuel screeningsmetode trænet på de kinetiske data fra hundredvis af KOD-varianter og designede et virtuelt bibliotek med dual-site mutationer baseret på den bedst ydende variant identificeret gennem den semi-rationelle screeningproces. Vi screenede eksperimentelt omkring 22% af varianterne i dette virtuelle bibliotek. De eksperimentelle resultater viste, at over 80% af mutanterne udviste højere katalytisk effektivitet sammenlignet med forældrestammen. Blandt dem viste den bedst ydende variant KH en forbedring i kinetisk effektivitet med mere end en faktor 10. Vi analyserede også mekanismen for de konformationelle ændringer i fingerdomænet ved position 485 og 451. Disse fund om interaktionerne af sekvens, struktur og funktion af KOD pol udvider vores hidtil begrænsede viden om B-familiens DNA-polymerasers interaktioner med modificerede substrater og yderligere vores forståelse af deres evne til at acceptere en bred vifte af modificerede nukleotider. Enzymudviklingsteknologien og -strategien i dette studie har også implikationer for at vejlede design af andre DNA-polymeraser til en bred vifte af bioteknologiske anvendelser.

I det andet delprojekt undersøgte vi brugen af mRNA-display til undersøgelse af α-synuclein-

aggregering, som er relateret til Parkinsons sygdom (PD). Målet var at udnytte mRNA-displayets ekstremt høje gennemløbskarakter i en sådan sammenhæng for første gang. Ideen bag denne tilgang er at studere aggregeringen og faseseparationen af et meget stort antal sekvensvarianter på samme tid ved at inkubere et mRNA-tagget sekvensbibliotek af α-synuclein med et overskud af aggregerende/faseseparerende WT-protein, efterfulgt af fysisk adskillelse af aggregaterne/dråberne fra det opløselige protein og sekvensering af de forskellige faser for at bestemme den fraktion af biblioteket, der er forbundet med hver fase. Et vigtigt aspekt, der bestemmer gennemførligheden af denne fremgangsmåde, er typen og styrken af interaktionerne mellem a-synuclein og mRNA. Derfor, som en foreløbig undersøgelse af mRNA-display-tilgangen, undersøgte vi først sammenspillet mellem α-synuclein og DNA. Vi valgte DNA som proxy for RNA, fordi det er lettere at håndtere og på grund af den eksisterende litteratur om α-synuclein-DNA-interaktioner, som vi ønskede at sammenligne vores resultater med.

Vi fokuserede på interaktionerne mellem α-Syn og forskellige typer DNA, herunder ssDNA med forskellige længder, dsDNA og genomisk DNA. Bindingsaffiniteterne af monomerisk a-Syn og DNA under de givne betingelser blev målt ved anvendelse af flow-induceret dispersionsanalyse (FIDA). De omtrentlige Kd-værdier blev beregnet og sammenlignet, disse resultater antydede, at bindingsaffiniteten var svag og generelt omkring ti mikromolær. Derudover blev væske-væskefaseseparation af α-Syn udført med og uden ssDNA og dsDNA for at undersøge, i hvilket omfang interaktionerne mellem DNA og α-Syn ville påvirke denne proces. Interessant nok opstod komplekse coacervater af α-Syn med både ssDNA og dsDNA, og ssDNA/dsDNA-PEG dannede dråber suspenderet større α-Syn kondensater. Derudover kan α-Syn gennemgå en væske-til-fast faseovergang på grund af aggregering selv i de komplekse coacervater, hvorimod både ssDNA og dsDNA forblev fuldstændig væskelignende selv efter 48 timer. Disse resultater kunne give indsigt i patogenesen af neurodegenerative sygdomme forbundet med α-Syn og kan have implikationer for udviklingen af nye terapeutiske strategier. Derudover tog vi de første skridt hen imod mRNA-display metoden ved hjælp af vildtype-α-synuclein, ved at demonstrere af vellykket ligering af puromycin til mRNA. De næste trin ville være at generere et passende mRNA-bibliotek af a-synuclein-varianter og udføre cellefri ekspression for at generere det mRNA-display bibliotek efterfulgt af screening og identifikation af disse varianter med høj specificitet samt affinitet for veldefinerede opløselige eller aggregerede tilstande. Denne undersøgelse af proteinsekvensrummet af α-synuclein har potentialet til at give mekanistisk indsigt i dets rolle i neurodegenerative lidelser i et hidtil uset omfang.

Project objectives

In this study, we generally aim to explore protein sequence space for functional improvements and mechanistic studies using two different proteins serving different purposes: KOD DNA polymerase, a member of the archaeal B-family DNA polymerase, and α-synuclein, which is implicated in neurodegenerative diseases like Parkinson's disease (PD). To achieve this goal, we have divided our research into two distinct sub-projects, which will be introduced in detail below.

1. DNA polymerase optimization by semi-rational sequence space exploration

Engineered B-family DNA polymerases are a driving force in numerous biotechnological applications, particularly in Next-Generation Sequencing (NGS). However, relatively little work has been reported on the engineering and modification of B family DNA polymerases to enable the efficient incorporation of cleavable fluorescent nucleotide reversible terminators in NGS applications. The objective of this sub-project is to engineer KOD DNA polymerase (KOD pol) to increase its catalytic efficiency for incorporating dye-labeled reversible terminators employed in NGS. To achieve this goal, we systematically designed and evaluated hundreds of KOD variants using both semi-rational and rational approaches developed in this study. This study provides a comprehensive understanding of the relationship between the sequence, structure, and function of KOD pol. The specific aims were:

1) To establish a high-throughput enzyme screening method for evaluating the capability of KOD mutants to incorporate modified nucleotides into DNA strands.

2) To design and screen KOD variants with mutations by semi-rational design in the active pocket and substrate-binding domain of KOD pol.

3) To obtain potential mutants that perform well on the BGI and MGI sequencing platforms.

4) To develop a machine learning (ML) based virtual screening method to reduce experimental efforts.

5) To provide valuable insights into the functional dynamics of KOD polymerases and have the potential to advance protein engineering of other B-family DNA polymerases.

2. Mechanistic studies of α-synuclein assembly by mRNA display

Recent research studies have extensively documented a variety of naturally occurring mutations in α-synuclein (α-Syn) that significantly influence the probability of an individual that carries the mutation contracting Parkinson's disease (PD). These α-Syn variants exhibit distinct characteristics compared to the wild-type α-Syn, including changes in aggregation propensity and binding properties. Therefore, there is a growing interest in identifying and characterizing more α-synuclein sequence variants related to PD. Currently employed methods of screening sequence space and variant characterization are very low throughput and limited to several dozens of variants.

In this thesis, we innovatively intend to use the mRNA display technique to explore its high throughput for the simultaneous characterization of many sequence variants with respect to their specificities and affinity for well-defined soluble or aggregated states formed by the wild-type α-Syn. Due to the known binding properties of α-Syn and nucleic acids, investigating the binding interactions between α-Syn and various types of nucleic acids is necessary, in order to be able to assess the potential effects of these binding interactions on the characterization analysis of mRNA-displayed sequence library of α-synuclein. Therefore, this sub-topic first aims to investigate the interaction between α-synuclein and nucleic acids, and secondly, we aim to develop the mRNA display technique to assemble α-Syn to explore and identify α-Syn variants related to PD. Specifically, the aims are as follows:

1) To characterize the interactions between monomeric and polymeric forms of α-Syn and various types of DNA, including single-stranded DNA (ssDNA) of different lengths, double-stranded DNA (dsDNA), and genomic DNA.

2) To conduct a comprehensive and systematic study of the interactions between α-synuclein and specific nucleic acids in liquid-liquid phase separation (LLPS).

3) To establish the mRNA display method using wild-type α-synuclein and then generate sequence libraries with the WT sequence as a starting point.

4) To screen and identify mRNA-displayed sequence library of α-synuclein and then analyze these sequenced variants with high soluble or aggregates properties.

Thesis outline

This thesis consists of an introduction, Chapter 1, Chapter 2, and a conclusion and future prospects. Chapter 1 and Chapter 2 constitute two sub-topics that are individually explored for two different proteins. The introduction provides an overview of the exploration of protein sequence space, its developmental history, and the related protein engineering technology. Chapter 1 and Chapter 2 highlight two proteins related to this study, namely KOD DNA polymerase (KOD pol) and α-synuclein (α-Syn), respectively, with different study purposes.

Chapter 1 comprises an introduction, two manuscripts, and a summary. The introduction provides an overview of KOD pol, DNA sequencing platforms of BGI, and the application of DNA polymerase in NGS. Manuscript 1 focuses on the directed evolution of KOD pol for dye-labeled reversible terminators used in DNA sequencing applications. This study describes a semi-rational design strategy to screen and obtain a KOD variant with eleven-point mutations, which performed exceptionally well in BGISEQ-500 and MGISEQ-2000 platforms compared to wild-type KOD pol. This study highlights the potential of semi-rational screening as a powerful tool for obtaining enzymes with desired properties and advancing their applications in the DNA sequencing field. Manuscript 2 is a continuation of the first manuscript. We focus on the development of rational design methods to further improve the incorporation efficiency of KOD variants. The study describes the establishment of machine learning models to predict mutation sites in the finger domain and thumb domain of KOD pol, leading to the continuous improvement of its performance. Additionally, the study explores the underlying molecular mechanisms behind the improved activity observed in the KOD variants obtained through rational design. This research provides valuable insights into the rational design of enzymes and demonstrates the potential of combining machine learning with experimental methods for enzyme engineering.

Chapter 2 contains an introduction, one manuscript draft, a small additional section, and a summary. The introduction provides an overview of α-Syn related to Parkinson's disease, studies of α-Syn and variants, and α-Syn phase separation. Manuscript 3 aims to explore the binding interaction between α-Syn and nucleic acids. Firstly, the binding affinities of the monomeric α-Syn and DNA under the given conditions were measured using flow-induced dispersion analysis (FIDA). Then, we conducted experiments to investigate the effect of DNA addition on the fibrillization process of α-Syn. Additionally, we evaluated the effect of DNA on phase separation of α-Syn. The results presented two distinct types of condensate droplets that result from the interaction between α-synuclein and two different types of DNA. This research

provides valuable insights into the interaction between α-synuclein and nucleic acids, which may have implications for understanding the pathogenesis of neurodegenerative diseases such as Parkinson's disease, and which will help us to access the effects of the interactions on our development of the mRNA display-based method. The last small section of this chapter presents the first steps towards our aim to use mRNA display technology to study the assembly of α-synuclein. We successfully optimized and established the initial steps of the mRNA display method using wild-type α-synuclein. Unfortunately, due to the outbreak of the epidemic and time constraints, this work has not yet been completed by the graduation deadline. The host laboratory at DTU will continue this research in the future, and the workflow of future studies is described.

Finally, the results of this study are concluded, and prospects are provided.

1 Introduction

Overview of protein sequence space exploration

Proteins are important components of all cells and tissues in the human body and they play a role in all vital body functions[1]. Proteins are polymers composed of amino acids that are linked together in a specific order to form a polypeptide chain. These amino acid building blocks can then combine in various ways to form proteins with hundreds of amino acid residues. The diversity of protein structures is a result of the sequence of amino acids in the protein molecule and the resulting three-dimensional structure[2]. Proteins typically have four levels of structure: primary, secondary, tertiary, and quaternary, and the structure of the protein determines its physical properties and functions[3]. The primary structure of a protein, which is its amino acid sequence, plays a major role in determining the three-dimensional and quaternary structure of the protein. Proteins play an essential role in the biomedical, molecular diagnostics, and molecular biology fields[4]. The investigation and comprehension of protein sequences, structures, and functions constitute a major driving force in the field of molecular biology[4].

Protein sequence space exploration entails the generation, analysis, and comparison of a vast array of protein sequences with the objective of enhancing mechanistic understanding and improving protein function. In the past decades, studies on protein sequence space exploration have grown exponentially, offering new perspectives for the study of protein functional diversity and mechanism[5]. The size of protein sequence space is astronomically vast. For example, the size of the sequence space for a protein with 100 amino acids could be estimated to be 20^{100}(approx. 10^{130}) using the typical approach[6], which involves the combination of the 20 commonly occurring amino acids. Even a single amino acid mutation can lead to substantial alterations in a protein's structure, function, or properties[5]. Consequently, researchers frequently employ targeted modifications to amino acid sequences in order to investigate functional enhancements of proteins. The exploration and development of protein sequence space typically rely on a diverse range of biotechnological methods, with protein engineering being a particularly important approach[6].

Protein engineering involves modifying or designing proteins to improve their properties or create novel functions. Protein engineering allows researchers to manipulate and explore the sequence space of proteins to optimize their stability, activity, specificity, or other desired characteristics[7]. Protein engineering encompasses a wide range of techniques, including computational modeling, and high-throughput protein screening approaches and directed

evolution[7]. Directed evolution is a powerful method employed to enhance the efficiency and capabilities of enzymes by introducing genetic variations. This process typically involves three main steps including starting with the introduction of mutations to protein genes and resulting in a library of variants[8], protein variants undergoing expression and separation processes, and the screening and selection for protein variants with improved properties or functions[9]. The strategies of directed evolution can be broadly categorized into three approaches (shown in Figure 1): irrational design, semi-rational design, and rational design[10].

The irrational design uses genetic engineering technology to gain diversity and generate multiple variants[11]. The most common methods generating genetic diversity libraries include DNA shuffling, error-prone PCR, random mutation, and chemical mutagenesis. These methods often require high-throughput screening platforms, such as fluorescence-activated cell sorting (FACS), nutrient deprivation selection, or tolerance to substrate toxicity, to select for higher enzymatic activity and fluorescence chromogenic selection[12,13,14]. Nishikawa et al. reported the construction of a gene screening system by combining giant unilamellar liposomes and a fluorescence-activated cell sorter (FACS) and achieved more than tenfold enrichment of the gene with higher β-glucuronidase activity[15]. However, the main limitations of the irrational design approach are the large library size and the difficulty in fully exploring the protein sequence.

The rational design approach relies on computer technology to simulate the evolutionary trajectory of natural proteins[16]. By using virtual mutations and screening, potentially useful mutants can be accurately predicted. Based on bioinformatics, a series of algorithms and programs are developed and used to predict the effects of mutations at specific sites on enzyme stability, folding, binding affinity, and enzyme activity[17]. Siegel et al. reported that they successfully developed a de novo enzyme that could catalyze bimolecular Diels-Alder reactions with high stereoselectivity and substrate specificity, by using computational design and experimental characterization of enzymes[18]. Although some successful studies have used the rational design approach[19], it faces challenges due to a lack of comprehensive understanding of the relationship between enzyme sequence, structure, and function, as well as the heavy computational demands and limited accuracy.

The semi-rational design approach uses a combination of the understanding of the physicochemical properties, protein structure, and structure-activity relationships, along with information on catalytic mechanisms, to create a library of reasonable mutants through techniques such as site-directed mutagenesis, saturation mutation, and combination mutations

of key residues in the active pocket or binding region[79,80]. The use of computer-aided technology is also incorporated[20]. This approach combines elements of both rational and irrational design, addressing the limitations of each and reducing technical requirements[19]. By constructing a relatively small library, and using high-throughput screening methods, target mutants can be rapidly obtained. The most common screening methods include fluorescence-activated droplets or cell sorting by using a microfluidic device[21], CSR (Compartmentalized Self-Replication)[22], RT-CSR (Reverse Transcription-Compartmentalized Self-Replication)[23], or microplate reader-based screening methods using Förster Resonance Energy Transfer (FRET)[24]. Scientists attempted to choose the best versions within the catalyst domain of Sh1B DNA polymerases that exhibited exceptional DNA production and high sensitivity, through the use of CSR selection[22]. The methods used to screen enzymes can be chosen based on their function and properties. DNA polymerases for sequencing are typically screened using microfluidic technology and evolutionary selection of potential mutants. Fluorescence-based microplate reader technology can also be used to screen for dominant mutants. For other types of molecular tool enzymes, such as PCR DNA polymerase, DNA ligase, or reverse transcriptase, CSR, RT-CSR, and microdroplets can be used to enrich and screen for dominant mutants[25,26,27]. Semi-rational design is one of the most widely applied and successful methods for enzyme screening.

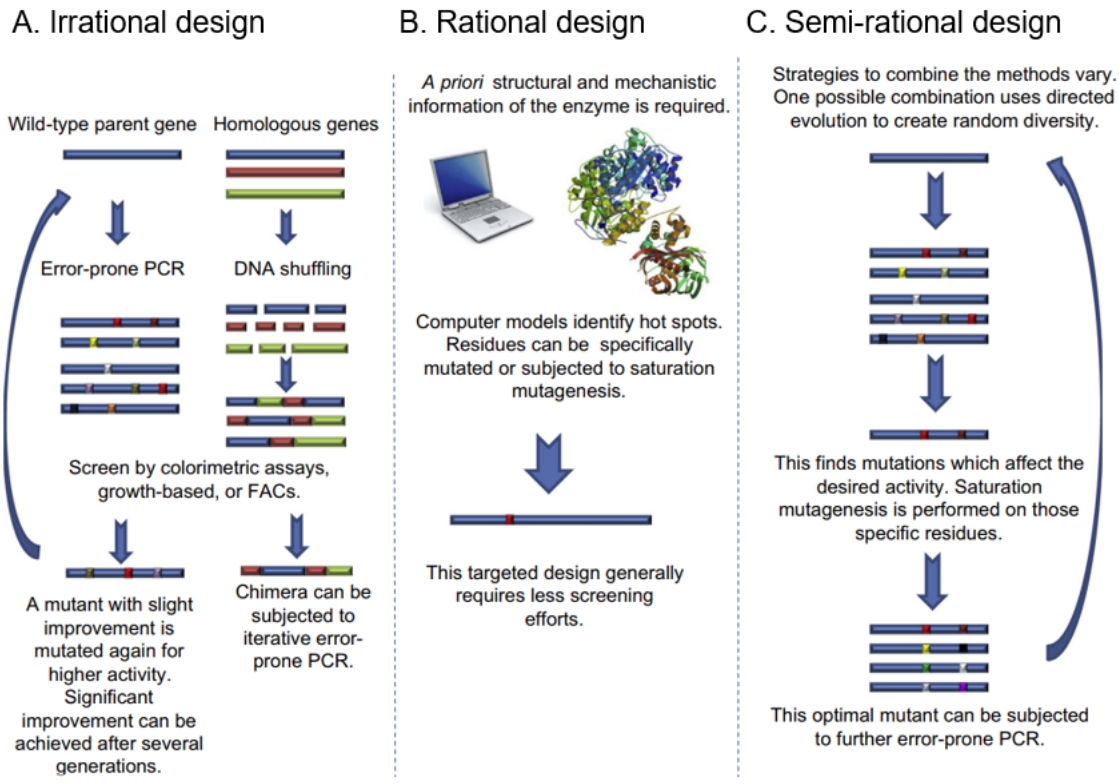


Figure 1. Overview of the main strategies of protein-directed evolution[28]. A. Irrational design simulates natural evolutionary processes in an accelerated way. Usually, diversified mutation library generation and high-throughput screening methods are employed in irrational design to accumulate beneficial mutations by iterating rounds of screening. B. Rational design relies on protein sequence and structural knowledge, which is not only used for modifying proteins but also for creating new proteins. A markable advantage of the rational design is that it is not limited by the availability of high-throughput screening methods. C. Semi-rational design is the combination of irrational design and rational design and is used widely and commonly by researchers. Semi-rational design targets specific residues located at the specific domain of an enzyme, in which the library size and the onerous screening task are reduced. The powerful tool can achieve diverse and efficient mutant libraries by identifying efficient mutations through structure-function analysis and predicting the effects of structural changes on enzyme activity.

In addition, display technology plays a crucial role in the exploration of protein sequence space in protein engineering. It allows for the presentation of expressed polypeptides on the surface of ribosomes, viruses, or cells in the form of fusion proteins while preserving their individual spatial structures and biological activities[29,30,31,32,33]. This technology is used to study the properties, recognition, and interactions of polypeptides or proteins by screening for specific functional structures. Types of display technology include phage display, ribosome display, mRNA display, and yeast surface display. Phage display technology, also known as phage display, was first developed by Smith 30 years ago[34,35], and this technique allows for the selection of peptides and proteins with desired functions and properties from molecular libraries through genetic engineering[32]. Ribosome display technology is a cell-free system that associates nascent proteins or polypeptides with their mRNA through ribosomes to form mRNA-ribosome-protein trimer complexes[32]. Ribosome display is widely used for antibody and protein library selection, in vitro enzyme-directed evolution, enzyme screening, etc[36]. Yeast surface display technology is a eukaryotic fusion protein expression system, in which a specific vector containing an exogenous gene is inserted into yeast cells[37]. It is widely used in protein evolution, specific ligand screening, novel drug discovery, monoclonal antibody election, optimization of immuno-biomolecular catalysis[38].

mRNA display technology is a powerful *in vitro* screening method for peptides and proteins[39]. This technique involves covalently linking mRNA to its encoded polypeptide or protein using a puromycin linkage, resulting in an mRNA-protein fusion. Compared to ribosome display technology, mRNA display has several advantages such as a smaller linking intermediate that does not interfere with subsequent screening and identification, and a more stable screening system. These fusions that bind stably are then reverse transcribed to cDNA and amplified using

a polymerase chain reaction. This technology can form a large-scale polypeptide or protein library (approx. 10^{13} sequence variants) without the limitations of *in vitro* cell expression system. It is mainly used in the screening of peptides and proteins with specific biological functions, the discovery of novel peptides and protein aptamers for various target molecules, the exploration of the sequence space of binding proteins, and the elucidation of protein interaction mechanisms. Anthony D. Keefe and Jack W. Szostak generated random mutant libraries with a sequence containing 80 contiguous amino acids using mRNA display technology, and then screened and enriched functional proteins that can bind ATP[40]. They obtained four novel ATP-binding proteins that were unrelated to each other or anything found in current biological protein databases[40]. Overall, mRNA display technology has a broad range of applications due to its conceptual simplicity and significant potential for development.

Exploring the vast space of protein sequences provides invaluable opportunities to design proteins with desired properties, investigate the intricate connections between protein sequence, structure, and function, and gain a deeper understanding of protein behavior and mechanisms[4,5,6].

2 Chapter I DNA polymerase optimization by semi-rational sequence space exploration

2.1 Introduction

2.1.1 Introduction of KOD DNA Polymerase

DNA polymerases are classified into evolutionary families based on their amino acid sequences and conserved motifs[41]. Currently, seven families of DNA polymerases have been described: A, B, C, D, X, Y, and RT[42]. The most widely studied polymerases belong to the A and B families, and KOD DNA polymerase is a typical representative of the B-family DNA polymerases[43]. KOD DNA polymerase (KOD pol) containing 744-amino acid is known for its high replication processivity, fidelity, and elongation rate[44,45]. Since the discovery of KOD pol, numerous scientists have conducted research on its sequence, functionality, and applications[13,15]. The sequence and crystal structure of KOD pol has been studied, revealing several conserved domains and residues that are critical for its function[46]. The detailed crystal structure (PDB: 1WNS)[46] of KOD pol has been determined and consists of five domains: exonuclease domain (131-326), palm domain (369-449, 500-587), thumb domain (591-774), finger domain (450-499), as well as the N-terminal (1-130, 327-368). The exonuclease domain of KOD pol has a beta-hairpin structure that helps to maintain the polymerase stably bound to the DNA template strand during 3'-5' exonuclease activity[47]. The beta-hairpin in KOD pol is made up of 12 and 13 amino acids, located distant from the DNA strand and not interacting with it. The exonuclease activity of KOD pol is often inactivated when used with chemically modified nucleotides to prevent cleavage of the modified nucleotides[48]. The N-terminal domain of KOD pol is thought to interact with the last resolved nucleotide of the DNA template strand. The active pocket of KOD pol, formed by the palm domain, incoming dATP, and closed finger domains, includes catalytic and DNA ligand residues. The distance between the amino acids of the KOD pol active pocket and the incoming dNTP is between 2.8-6.6 Å, and stacks with nascent base pairs form active complexes[49]. The palm domain of KOD pol has two long and one short alpha-helix and a six-stranded beta-sheet, structurally similar to the palm domain of 9°N DNA polymerase[50]. The finger subdomain of KOD pol changes conformation during DNA synthesis. The finger domain of KOD pol has two alpha-helices, similar in structure to the finger domain of DNA polymerase delta. The thumb domain, in an "open" conformation, is responsible for binding double-stranded DNA[51]. In addition, the binary complex structure of KOD pol and the DNA strand (PDB: 4K8Z)[52], as well as the ternary complex structure (PDB: 5OMF) of KOD pol with DNA and incoming dATP have been determined by X-ray

crystallography[49]. In general, wild-type (WT) KOD pol often demonstrates limitations or suboptimal performance in practical applications; however, these shortcomings can be mitigated through protein engineering[53].

The published research on KOD DNA polymerase focuses on the engineering of KOD pol using various methods to evaluate the effect of mutations at specific sites on functional impairment or enhancement[54,55]. The mutations characterized in those studies have led to the generation of functional diversity in KOD pol and have the potential for extensive application[56]. Here, we summarize the reported KOD variants and their effects on KOD pol in Table 1. Currently, research on KOD mainly focuses on improving its efficiency in the polymerase chain reaction (PCR), in synthesizing artificial genetic polymers such as TNA (threose nucleic acid), modified nucleotides, and developing new functions such as reverse transcriptase activity.

Table 1. Summarized variants information and their influences on reported KOD pol.

Variant	Influence	Reference
D141A/E143A	3'-5' exonuclease deficient	Kropp et al.[57]
H147E	30% of the 3'-5'exonuclease activity	Kuroita et al.[48]
H147K	276% of the 3'-5'exonuclease activity	Kuroita et al.[48]
H142Q	24% of the 3'-5'exonuclease activity	Kuroita et al.[48]
G245I/D	5-methylcytosine-sensitive activity increase	Huber et al.[58]
A485R/E664I	α-(L)-thyreophorans nucleic acid (TNA) triphosphates catalysis activity increase	Chim et al.[59]
M247R/L381R	PCR processivity ability improvement	Elshawadfy et al.[60]
L381R/K502R	PCR processivity ability improvement	Elshawadfy et al.[60]
M137L/K466R	RT activity	Pinheiro et al.[61]
D141A/D143E /A485L	Functionalized nucleotides with amphiphilic side chains incorporation	Hoshino et al.[47]
A485R/E143A	Incorporation fidelity improvement of tNTP incorporation	Horhota et al.[62]
L489Q/N491S	Improve the ability of polymerase towards threose nucleic acid synthesis	Nikoomanzar et al.[55]
A485R/N491S	Improve the ability of polymerase towards threose nucleic acid synthesis	Nikoomanzar et al.[55]
A485R/E664I	Improve the ability of polymerase towards threose nucleic acid synthesis	Nikoomanzar et al.[55]
W504R/Y505N	Potential increasing the TNA polymerase activity	Nikoomanzar et al.[55]
R476G/Y481F	Potential improvement TNA polymerase activity	Nikoomanzar et al.[55]

606G/T723A	polymerase activity improvement toward TNA synthesis	Nikoomanzar et al.[55]
P179S/L650R	Incorporation LNA nucleotides additionally modified at position 3' of the sugar moiety	Sabat. *et al*[63]
E664K/ G711V	Increase the binding affinity of polymerase to RNA/DNA heteroduplexes	Ellefson et al.[64]
E735K	Increase the binding affinity of polymerase to RNA/DNA heteroduplexes	Ellefson et al.[64]

Below, we present three representative reported works on the directed evolution of KOD pol. It is noteworthy that further engineering modifications of KOD pol are of substantial significance for exploring its potential in various biotechnological fields.

The first study employed a protein fusion method for improving the PCR performance of KOD DNA polymerase[60], shown in Figure 2. Elshawadfy *et al.* reported that the hybrid enzyme Pfu-TkodT from Pfu-pol *(Pyrococcus furiosus)* and Tkod-Pol *(Thermococcus kodakarensis)* showed faster elongation rates, improving PCR performance and processability[60]. The authors also claimed that Pfu-Tkod contained a "forked point" located in the junction of the exonuclease and polymerase channels of Tkod-Pol, leading to an improvement in PCR performance. In addition, Pfu-Tkod contains the thumb domain responsible for the double-stranded DNA binding of Tkod-Pol, again leading to improved performance in PCR. The authors also evaluated single, double, and triple arginine/thumb swap mutants towards PCR performance.

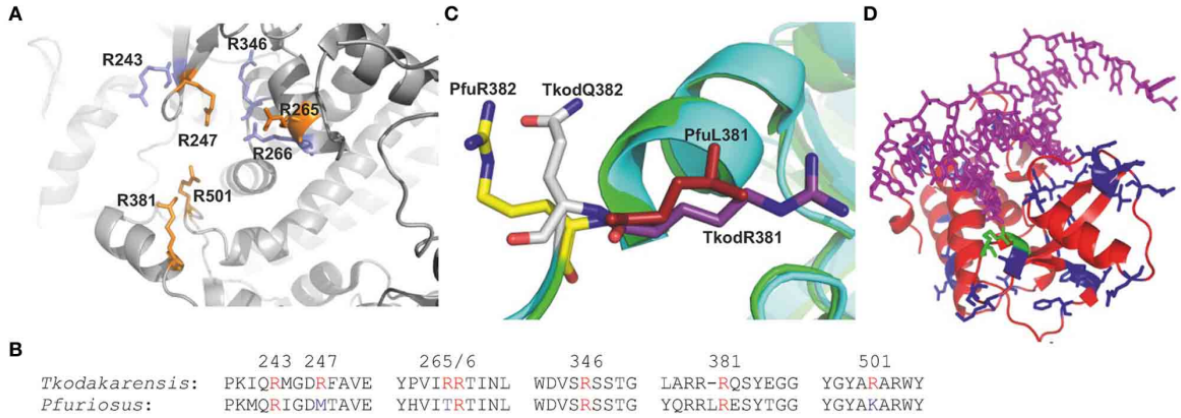


Figure 2. The graph illustrates the "Forked-point" arginine and the amino acid sequence arrangement of their neighboring locations in the family-B DNA polymerase[60]. It also compares the thumb domain of Tkod-Pol (PDB ID: 4K8Z) and Pfu-Pol (PDB ID: 2JGU) and shows the sequence alignment between Tkod-Pol and Pfu-Pol.

The second work presented the engineering of KOD DNA polymerase by the RT-CSR strategy

towards reverse transcriptase activity. Ellefson *et al*. developed a reverse transcription compartmentalized self-replication (RT-CSR) method which was a modified CSR method for the evolution of KOD pol with improved reverse transcriptase activity[64] and showed the workflow of RT- CSR in Figure 3. The authors started with WT KOD pol transcribing less than 5 RNA bases to generate the mutation library by using error-prone PCR and DNA shuffling methods. Then, they conducted eighteen rounds of RT-CSR with the stepwise addition of RNA bases into primers. The best variant B11 containing thirty-seven mutations was successfully achieved that featured significantly improved RT activity by being able to reverse transcribe at least 500 base pairs. Subsequently, the authors synthesized and tested several KOD variants obtained by modeling designs of several polymerases with favorable RT mutations. The most effective RTX variant, which contained fewer than half the mutations found in B11, was able to perform single-enzyme RT-PCR and had reverse transcription capabilities of more than 5 kb. Therefore, the authors successfully engineered a KOD DNA polymerase variant by a combination of directed evolution and rational design strategies.

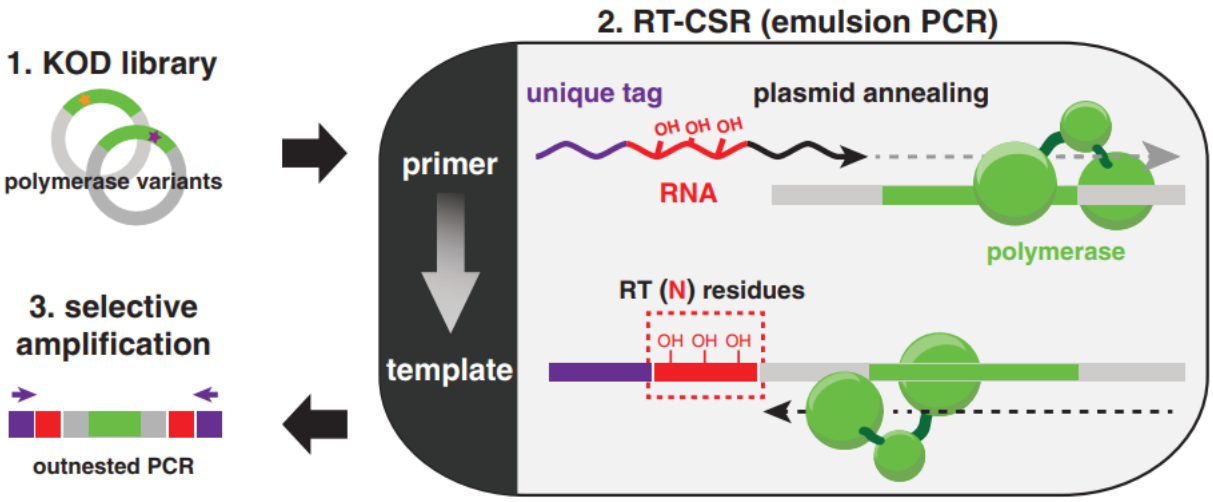


Figure 3. Framework for the directed evolution of a KOD pol by reverse transcription compartmentalizes self-replication (RT-CSR) [64]. Libraries of KOD pol mutants are generated, and expressed in Escherichia coli, and compartmentalized *in vitro*. Primers flanking the KOD DNA polymerase coding region allow for self-replication during emulsion PCR, which is designed with variable RNA bases separating the plasmid annealing portion from the unique recovery tag. Through the iterative rounds of RT-CSR, the number of RNA bases is gradually increased.

The third example described the successful application of a fluorescence-activated cell sorting (FACS) device or fluorescent-activated droplet sorting (FADS) instrument for mutant screening during KOD DNA polymerase evolution[55] (shown in Figure 4). Nikoomanzar *et al.* developed a high-throughput microfluidic approach which was described as droplet-based optical polymerase sorting (DrOPS) to explore the relationships between polymerase sequence space and function[55]. Starting from deep mutational scanning of the 48 amino acid sites in the finger subdomain of the KOD pol, they employed the DrOPS method to separate and select functional variants. The enrichment profile supplied an impartial view of each mutant's ability for synthesizing threose nucleic acids used as a model non-natural genetic polymer. The authors found that the FADS experiment showed higher levels of enrichment than the FACS experiment. After identifying highly enriched mutations, the authors selected thirteen single-point mutation sites for separate functional analysis and designed combinations of mutations based on the analysis result, leading to the discovery of a double mutant polymerase with ten-fold greater activity than the previous best TNA polymerase.

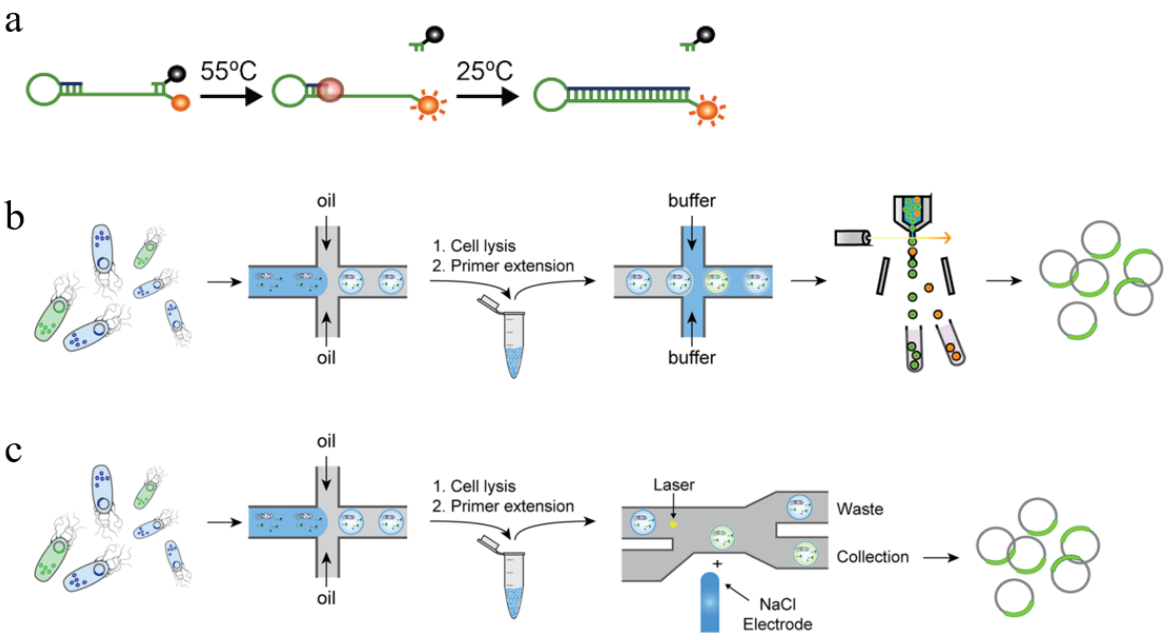


Figure 4. Overview of screening strategy of the DrOPS method[55]. (a) A schematic diagram of the fluorescence detector system used to select beneficial variants with polymerase activity. The detector is made up of a spontaneously initiated hairpin template (green) containing a downstream fluorophore (orange) and a DNA quencher (black) probe annealed to the 5' end of the template. The quencher will be dissociated from the hairpin template when the temperature is raised to 55 °C, allowing for the detection of polymerase activity. The KOD variants with

optimal polymerase activity cause the quencher to remain detached and lead to the production of a fluorescent signal that can be detected at room temperature. If there is no extension by the polymerase, the quencher will reattach to the template once cooled to room temperature, resulting in the absence of a fluorescence signal. (b) Schematic diagram of droplet-based optical polymerase sorting using double emulsion droplets with conventional FACS sorting. (c) Schematic diagram of droplet-based optical polymerase sorting using single emulsion droplets with FADS sorting.

2.1.2 General introduction of sequencing platform of BGI Genomics

BGI Genomics is an advanced technology organization that particularly specializes in next-generation sequencing (NGS)[65]. BGI Genomics has proprietary sequencers and high-performance bioinformatics analysis pipelines[66]. The BGISEQ series of domestically produced gene sequencers utilize advanced combinatorial Probe Anchor Synthesis (cPAS) and improved DNA nanoballs (DNBs) core sequencing technologies, which anchor DNA molecules and fluorescent probes on the DNBs[67]. The sequencing technology based on DNA nanoball (DNB) is collectively referred to as DNBSEQ™ technology (Figure 5A). The sequencing reagent kits typically contain preparation reagents for DNA nanoballs and general sequencing reagents for combinatorial probe anchor synthesis, which includes DNA polymerase as a critical component. During sequencing, DNA polymerase catalyzes the incorporation of fluorescently labeled reversible terminators (probes) into the sequencing template (Figure 5B). The fluorescent groups are then excited by a laser, and the resulting emission intensities are collected. These emissions are subsequently processed into digital signals, which are used to obtain DNA sequence information. Additionally, MGI Genomics has developed CoolMPS™ chemistry technology[68], which introduces cold nucleotides (modified nucleotides with reversible block groups) and four fluorescent-labeled antibodies to identify the incorporated base (Figure 5C). CoolMPS™ sequencing retains the high accuracy, low repeat sequence rate, and low label jumping rate[68].

In response to evolving market demands, BGI Genomics continuously updates its sequencing instruments and reagent series since unveiling its desktop sequencing system, BGISEQ-500, in 2015[65]. For instance, the newly launched DNBSEQ-T7 is a flexible, fast, and ultra-high-throughput gene sequencer capable of producing up to 6 TB of data per sequencing data per day[69]. This ongoing updating of sequencing instruments poses persistent challenges in terms of upgrading the DNA polymerase in the core constituents of sequencing reagents. Additionally, the limitations imposed by existing international patents of enzymes used in DNA sequencing

represent a significant challenge for BGI Genomics in its efforts to bring new sequencing instruments and technologies to the global market. In order to overcome these limitations, BGI Genomics needs to develop proprietary DNA polymerases with independent intellectual property rights. Furthermore, continuous optimization and functional improvement of DNA polymerases are required to enhance their catalytic efficiency for modified nucleotides and specificity to meet the diverse requirements of various sequencing instruments.

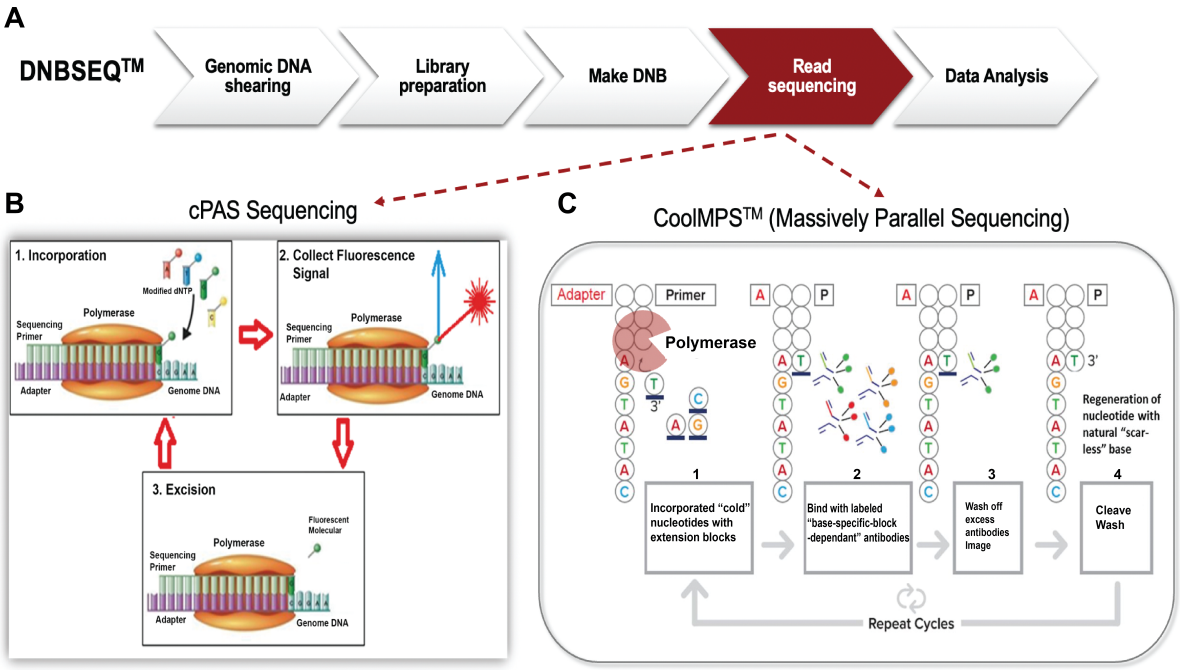


Figure 5. Schematic diagrams of the DNBSEQ™ sequencing technologies. A) The sequencing workflow is based on DNBSEQ™ technology, which includes five steps: DNA fragmentation, library construction, DNB generation, sequencing readout, and data analysis. B) This figure illustrates the principle of the cPAS sequencing technology. In Step 1, after the binding of sequencing primers and genomic DNA adapters, DNA polymerase extends the primer and incorporates fluorescently labeled reversible terminators (probes) into the template strand. In Step 2, excess modified substrates are washed away, followed by imaging and signal collection. In Step 3, the fluorescent and blocking groups of the modified substrates are cleaved to produce natural scar-less nucleotides for the next sequencing cycle. C) This figure depicts the principle of CoolMPS™ technology. In Step 1, DNA polymerase adds cold nucleotides to the primer strand. In Step 2, labeled "base-specific-block-dependent" antibodies are introduced to rapidly bind to these incorporated cold nucleotides. After washing out excess antibodies in Step 3, imaging and signal collection are performed. In Step 4, the blocking groups of the modified substrates are cleaved to generate natural scar-less nucleotides for the subsequent sequencing cycles.

2.1.3 Introduction of the application of DNA polymerase in NGS

Archaeal B family DNA polymerases and dye-labeled reversible terminators play a crucial role in driving next-generation sequencing (NGS) technologies[70,71]. Cleavable fluorescent nucleotide reversible terminators have been recently proposed for DNA extension in NGS[72]. The subsequent extension reaction was performed by employing substrate-specific DNA polymerase, which catalyzed the incorporation of modified nucleotides into the cleaved DNA products containing the cleavage scars[70]. Challenges in NGS with cleavable fluorescent nucleotide reversible terminators involve further improving the DNA polymerase to efficiently recognize the modified nucleotides[70]. B family DNA polymerases are widely engineered to be applied for catalyzing the incorporation of chemically modified nucleotides, because of their potential for tunability of the active site and acceptance of a broad range of substrates[73,52,57]. The engineering of B family DNA polymerases is essential to enhance their tolerance towards dye-labeled reversible terminators and improve their catalytic efficiency, enabling them to function as enzyme products in NGS applications[74]. However, relatively little work has been reported on the engineering and modification of B family DNA polymerases to enable the efficient incorporation of dye-labeled reversible terminators in NGS applications[75]. Additionally, the engineering of B family DNA polymerases to incorporate dye-labeled reversible terminators is hindered by a lack of comprehensive understanding regarding the structural determinants that govern polymerization efficiency[76].

KOD DNA polymerase (KOD pol) serves as a prototypical representative of B family DNA polymerases[77]. Engineered KOD DNA polymerase plays a significant role in various biotechnological applications, including PCR[78], TNA synthesis[62], and incorporation of base or sugar-modified nucleoside triphosphates[79]. Some KOD variants, specifically KOD variant (exo-), KOD variant (exo-, 485L), and KOD variant (P179S, L650R), have demonstrated tolerance and polymerization capabilities for certain unnatural nucleotides, such as 7-deaza-modified adenosine triphosphate and 5-substituted pyrimidine nucleoside triphosphates (dNamTPs) [71,55,47]. However, few studies have been reported on KOD variants that demonstrate efficient polymerization of dye-labeled reversible terminators for NGS applications[80]. In previous studies, a 9°N DNA polymerase variant (exo-, A485L/Y409V) has been reported to incorporate 3'-O-azidomethyl-terminators[70]. The crystal structures of 9°N DNA polymerase and KOD pol have been compared and are shown in Figure 6, revealing a common structural organization consisting of five domains: the exonuclease domain, palm domain, thumb domain,

finger domain, and N-terminal domain. In addition, KOD DNA polymerase and 9°N DNA polymerase share high sequence similarity (91%), and both belong to the B family of polymerases[81]. Thus, KOD pol has the potential to be engineered to efficiently incorporate dye-labeled reversible terminators, thereby expanding its applications in NGS.

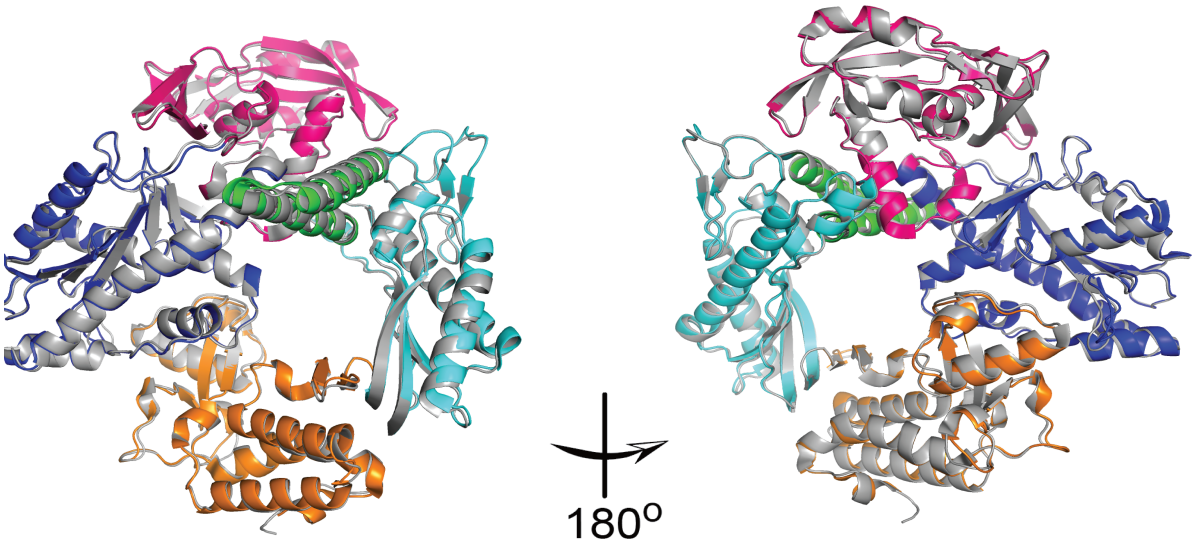


Figure 6. The comparison of the crystal structures of KOD DNA polymerase (PDB ID: 4K8Z) and 9°N DNA polymerase (PDB ID: 4K8X). Both crystal structures are depicted in a cartoon format. In the 9°N DNA polymerase structure, the entire protein is colored gray. In the KOD DNA polymerase structure, the exonuclease domain is represented in blue, the palm domain is colored cyan, the thumb domain is depicted in orange, the finger domain is shown in green, and the N-terminal domain is depicted in hot pink.

2.2 Part I Manuscript 1 - Semi-rational evolution of a recombinant DNA polymerase for catalytic activity on modified nucleotide incorporation

This manuscript describes the engineering of wild-type KOD DNA polymerase using a semi-rational design strategy, enabling it to efficiently incorporate dye-labeled reversible terminators commonly used in next-generation sequencing (NGS) applications. To achieve this objective, we initially established a high-throughput enzyme screening method based on microplate screening and FRET fluorescence technology to evaluate the capability of KOD variants to incorporate modified dATP into DNA strands. Subsequently, through structural analysis of the KOD polymerase and computer-assisted site prediction, we constructed several mutant libraries. A total of four rounds of mutant library screening were performed, with each round of mutant screening based on the best mutants from the previous round. In the final round of screening, we identified a KOD mutant, Mut_E10, with 11 mutation sites, and conducted functional tests on two different sequencing platforms. The experimental results demonstrate that this variant exhibits satisfactory sequencing performance on both platforms. We further analyzed and discussed the potential polymerization mechanism responsible for the functional enhancement of the KOD variant and outlined future studies. Additionally, we have applied for a patent (WO2022082482A1) to protect the effective mutation sites of KOD pol discovered in this study. This manuscript has been prepared for publication in the near future.

Semi-rational evolution of a recombinant DNA polymerase for modified nucleotide incorporation efficiency

Contributions: Lili Zhai developed the methodology, performed the experiments for KOD DNA polymerase library construction, mutant screening, and protein purification, and analyzed the data. Wang Zi performed the MD simulation. Jingjing Wang, Hongyan Han, and Fen Liu validated the performance of the best variant on various sequencing platforms.

Lili Zhai[1,2], Zi Wang[1], Liu Fen[4], Chongjun Xu[3], Jingjing Wang[3], Hongyan Han[1], Qingqing Xie[1,*], Wenwei Zhang[1*], Yue Zheng[1,*], Alexander K. Buell[2,*], Yuliang Dong[1,*]

1. BGI-Shenzhen, Beishan Industrial Zone, Shenzhen 518083, China

2. Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs. Lyngby 2800, Denmark

3. MGI, BGI-Shenzhen, Shenzhen 518083, China

4. MGI, BGI-Wuhan, Wuhan 430030, China

*Corresponding author

Abstract

Engineering improved B-family DNA polymerases to incorporate 3'-O-modified nucleotide reversible terminators is limited by an insufficient understanding of the structural determinants that define polymerization efficiency. To explore the key mechanism for unnatural nucleotide incorporation, we engineered a B family DNA polymerase from *Thermococcus Kodakaraenis* (KOD pol) by using semi-rational design strategies. We first scanned the active pocket of KOD pol through site-directed saturation mutagenesis and combinatorial mutations and identified a variant Mut_C2 containing five mutation sites (D141A/E143A, L408I/Y409A, A485E) using a high throughput microwell-based screening method. Mut_C2 demonstrated high catalytic efficiency in incorporating 3'-O-azidomethyl-dATP labeled with the Cy3 dye, whereas the wild-type KOD pol failed to incorporate it. Computational simulations were then conducted towards the DNA binding region of KOD pol to predict additional mutations with enhanced catalytic activity, which were subsequently experimentally verified. By a stepwise combinatorial mutagenesis approach, we obtained an eleven-mutation variant, named Mut_E10 by introducing additional mutations on the background of the Mut_C2 variant. Mut_E10, which carried six specific mutations (S383T, Y384F, V389I, V589H, T676K, and V680M) within the DNA-binding region, demonstrated over 20-fold improvement in kinetic efficiency as compared to Mut_C2. In addition, Mut_E10 demonstrated satisfactory performance in two different sequencing platforms (BGISEQ500 and MGISEQ2000), indicating its potential for commercialization. Our study demonstrates that an effective enhancement in its catalytic efficiency towards modified nucleotides can be achieved efficiently through combinatorial mutagenesis of residues in the active site and DNA binding region of DNA polymerase. These findings contribute to a comprehensive understanding of the mechanisms that underlie the incorporation of modified nucleotides by DNA polymerase. The beneficial mutation sites, as well as the nucleotide incorporation mechanism identified in this study, can provide valuable guidance for the engineering of other B-family DNA polymerases.

Key Words

Semi-rational evolution, KOD DNA polymerases, Enzyme engineering, FRET technology,

Molecular dynamics simulation, DNA sequencing, Modified nucleotides

Introduction

DNA polymerases discovered in nature play important roles in molecular biology, synthetic biology, and molecular diagnostics[1,2]. Wilde-type (WT) DNA polymerases mostly fail or perform poorly in practical applications[2]. However, these shortcomings can be overcome by enzyme engineering[3]. The application of DNA polymerase engineering to improve the performance of natural enzymes has been widely reported[1,4]. Engineered DNA polymerases are the work horses of numerous biotechnological applications such as in the fields of DNA synthesis[4], molecular diagnostics[5], and DNA sequencing[3,6,7].

Many DNA sequencing methods rely on engineered DNA polymerases that efficiently incorporate chemically modified nucleotides[8,9] such as the Sanger sequencing method, which relies on dideoxynucleoside triphosphates (ddNTPs)[7], and the sequencing-by-synthesis (SBS) method, which uses dye-labeled reversible terminators in an iterative loop[6]. SBS relies on the identification of each base as the DNA strand is extended by the cleavable fluorescent nucleotide reversible terminators that temporarily pause the DNA synthesis for sequence determination[10]. Challenges in using SBS with cleavable fluorescent nucleotide reversible terminators involve further improving the DNA polymerase that efficiently recognizes and incorporates the modified nucleotides[9]. Archaeal B-family DNA polymerases are widely engineered to be applied in catalyzing the addition of chemically modified nucleotide substrates in SBS sequencing, because of their potential for tunability of the active site and acceptance of broad substrate[11,12,13]. The engineering of B-family DNA polymerases relies on the development of various enzyme evolution techniques[14].

Enzyme evolution techniques usually include computational design[15,16], semi-rational design[17], and large-scale screening by directed evolution[18]. Semi-rational design is one of the most widely applied approaches, which enables targeted exploration of functional amino acid sites by selecting and mutating specific residues, with reducing library size and screening costs compared to random mutagenesis[14]. Semi-rational design is typically complemented by various screening methods to evaluate and select protein variants with desired characteristics[17]. Enzyme screening strategies can include DNA shuffling[19], CSR (Compartmentalized Self-Replication)[20], microfluidic screening technology[21], microplate reader screening technology[22], etc. Most screening techniques will utilize fluorescence as a readout because of its high sensitivity, time efficiency, the potential for high throughput, low cost, and other advantages[23]. Förster resonance

energy transfer (FRET) is a mechanism describing energy transfer between two fluorescent molecules[24], which is widely exploited in various protein mechanistic and engineering studies, such as studies of protein folding kinetics[25], exploring the spatial relationships between protein and substrates[26], and protein-protein interactions[26]. In particular, previous studies have reported the applicability of FRET technology for the screening of enzymatic activity[27,28].

KOD DNA polymerase (KOD pol) originates from *Thermococcus kodakarensis* which is an extremophilic archaeon, an important representative of B-family DNA polymerases[29,30]. Engineered KOD pol is the key component of numerous biotechnological applications such as multiple types of PCR[31], TNA synthesis[32], and incorporation of base or sugar-modified nucleoside triphosphates[33,34,29]. Some KOD variants have been reported to tolerate and polymerize certain unnatural nucleotides[35]. For example, the KOD variant (exo-) can tolerate 7-deaza-modified adenosine triphosphate[36] or microenvironment-sensitive fluorescent nucleotide probes[34]. KOD variant (exo-, A485L) can incorporate 5-substituted pyrimidine nucleoside triphosphates (dNamTPs)[37]. KOD variant (P179S, L650R) can tolerate LNA nucleotides additionally modified at position 3' of the sugar moiety[38]. However, significantly less work has been reported on the engineering and modification of B family DNA polymerases to enable the efficient incorporation of cleavable fluorescent nucleotide reversible terminators in Next-Generation Sequencing (NGS) applications. A previous study reported that the 9°N DNA polymerase variant (exo-, A485L/Y409V) can incorporate nucleotide reversible terminators, such as 3'-O-azidomethyl-dNTPs[9]. KOD DNA polymerase and the 9°N DNA polymerase share high sequence similarity (91%), and both belong to the B family of polymerases[39]. Thus, KOD polymerase has the potential to be engineered to efficiently incorporate reversible terminators, thereby expanding its applications in the field of DNA sequencing[39]. In Engineering KOD DNA polymerases to incorporate modified nucleotides, there are generally two main directions of sequence modification: mutations in the active site to enable discrimination and incorporation of specific nucleotides[40], and mutations in the DNA strand binding region to enhance catalytic efficiency by stabilizing the interactions between the polymerase, the DNA template, and the incoming nucleotide[41].

In this study, we established a FRET-based enzyme screening platform to screen mutants for their improved abilities to incorporate 3'-O-azidomethyl-dATP labeled with the Cy3 dye into a primer-templated and Cy5-labeled DNA using the FRET emission signal of Cy5 as an indicator. Initially, we scanned amino acids located in the active site pocket of KOD pol. Subsequently, we computationally predicted relevant residues to be replaced within the DNA binding region

of KOD pol. By employing enzyme kinetic screening of multiple mutant libraries, we successfully obtained a KOD variant that exhibited satisfactory performance on both BGI and MGI sequencing platforms. Through dissecting the specific residues within the active site and DNA binding region, this study provides insights into the key mechanisms underlying the incorporation of modified nucleotides by KOD pol. The effectiveness of semi-rational design strategies and the identified beneficial mutation sites in this study serve as valuable guidance for the engineering of other B-family DNA polymerases. In addition, the significance of our research lies in its implications for the engineering of DNA polymerases and their applications in NGS.

Results

Mutant screening strategy establishment

The comprehensive screening strategy developed for KOD pol involved multiple intricate steps including structural analysis, MD simulations, protein expression, mutant screening, mutant protein purification, and tests in sequencing applications, as illustrated in Figure 1A. The process began with a structural analysis of KOD pol to provide valuable insights into potential sites for mutagenesis. We performed saturation and combinatorial mutagenesis on the selected sites to construct multiple libraries. We employed MD simulations to predict more promising sites to do stepwise combinatorial mutagenesis. The high throughput expression of KOD pol variants was carried out in a 96-deep-well plate format in order to obtain sufficient quantities of protein for screening. The expressed proteins were subsequently subjected to semi-purification followed by quantification to ensure the reliability and comparability of the subsequent mutant screening experiments (Supplementary Figure 1). Then, the protein variants were screened for enzyme activity and kinetic performance in a 384-well plate using a microplate reader. After identifying variants with favorable kinetic characteristics, we performed rigorous purification using chromatography with three prepacked columns to eliminate potential interference from impurities on sequencing quality. Subsequently, the promising and highly purified proteins were tested for their sequencing performance on the BGISEQ-500 platform for SE50 testing and MGISEQ-2000 platform for PE100 testing.

In the screening method, the modified dATP, 3'-O-azidomethyl-dATP labeled with the Cy3 dye, shown in Figure 1B, was employed. Additionally, we employed a primer-template complex in which the single oligo strand was labeled with a fluorescent Cy5 dye at its 5' end (P/T-2Cy5), as illustrated in Figure 1C. Upon successful incorporation of 3'-O-azidomethyl-dATP labeled

with the Cy3 dye into P/T-2Cy5 by DNA polymerase, excitation at 530 nm of Cy3 results in fluorescence resonance energy transfer to the nearby Cy5 molecule, resulting in the emission at 676 nm by Cy5, as illustrated in Figure 1C. This increase in Cy5 FRET emission signal can be conveniently monitored inside a microplate reader, allowing for the measurement of the rate of modified nucleotide incorporation into the DNA strand. By comparing the incorporation rates (RFUs/min, with RFUs representing the increased Cy5 FRET emission signal), we were able to determine the catalytic efficiency of KOD variants for incorporating modified dATP.

Experimental validation for the screening method was performed successfully by comparing two KOD variants with distinguishing enzymatic activities, and WT KOD pol, shown in Supplementary Figure 1C. The developed screening method was effective in distinguishing between the different levels of enzymatic activities of the variants.

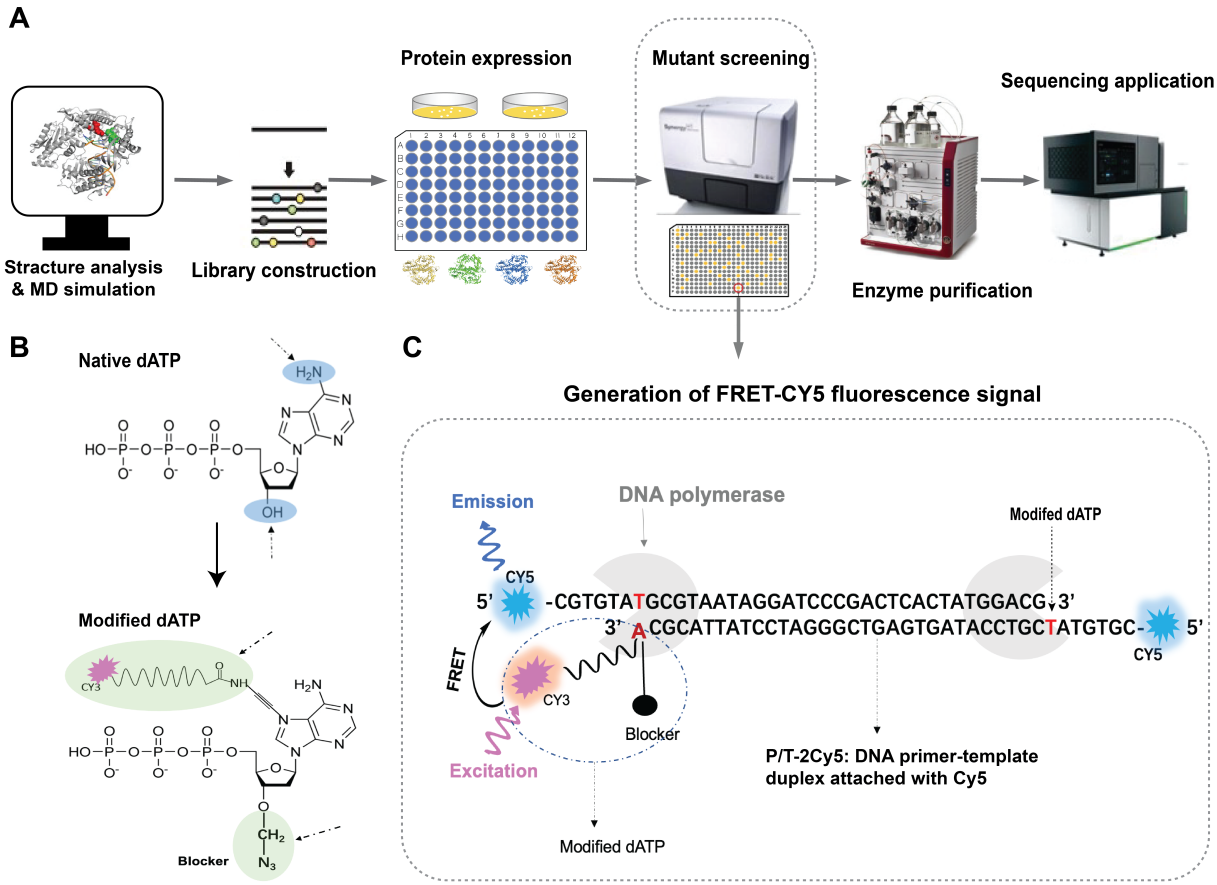


Figure 1. Schematic of KOD pol engineering experiments. A) Overview of the semi-rational evolution process involving structural analysis of KOD pol, MD simulation, protein expression, variant screening, variant purification, and sequencing application tests. Multiple libraries were constructed by site-direct mutagenesis, combinatorial mutagenesis, and stepwise combinatorial mutagenesis. The screening of the catalytic efficiency of the mutants was performed by

monitoring FRET Cy5 fluorescence as an indicator. B) The molecular structures of 3'-O-azidomethyl-deoxyadenosine triphosphate with dye Cy3 labeled (modified dATP) as well as native dATP. C) The process of the generation of the FRET Cy5 fluorescence emission signal. The FRET Cy5 fluorescence emission signal can be detected at 676 nm upon excitation of the incorporated Cy3 dye-labeled modified dATP at 530 nm. If the modified dATP is not properly incorporated into the P/T-2Cy5, no FRET Cy5 fluorescence emission signal is detected. DNA polymerase is shown in cartoon format and colored grey. Dye Cy5 and Cy3 are shown in cartoon format and colored blue and pink, respectively. The primer-template duplex is labeled with Cy5 dye at both 5' ends, which enhances signal intensity.

Selection of first-round mutation sites and mutant library construction

First, we analyzed the amino acids located in the active pocket of KOD pol. The crystal structures of KOD pol, including the open (PDB ID: 4K8Z)[42] and the closed (PDB ID:5OMF)[39] states, are superimposed and shown in Figure 2 A. A significant movement of the finger domain from the open state to the closed state can be observed, resulting in the formation of the active pocket by the palm and closed finger subdomains. Residues located in the active pocket of KOD pol directly impact nucleotide incorporation efficiency by interfacing with incoming nucleotides. In the active pocket, residues L408, Y409, P410, and A485 (Figure 2B) were reported to associate with the enhanced unnatural nucleotide incorporation efficiency[43]. In addition, in the engineering of B-family polymerases to incorporate modified nucleotides, exonuclease activity is often inactivated to prevent the excision of the incorporated modified nucleotides[44]. A previous study reported that an exonuclease-deficient variant (D141A and E143A) with A485L mutation derived from *Thermococcus kodakaraensis* (KOD) DNA polymerase, was prepared to incorporate unnatural nucleotides into a primer-template DNA[45]. Based on these findings, we selected residues L408, Y409, P410, and A485, as well as mutation sites D141A and E143A with exonuclease deficiency, as key sites for the first-round mutagenesis. Despite the fact that these positions have been previously reported to be important for the objectives of our study, there is still a lack of in-depth investigation into the effects of their combination and interactions with other related sites. Such an in-depth investigation will be valuable for the engineering of B-family DNA polymerases to improve the incorporation efficiency for unnatural nucleotides.

We performed alanine substitutions at residues L408, Y409, P410, and A485. Introducing alanine substitutions at critical positions in proteins represents a potent strategy for identifying the residues that play crucial roles in protein function, stability, and structure[46]. Alanine substitution allows for the evaluation of the individual contributions of each amino acid

sidechain to the overall functionality of the protein[46,47]. In this study, alanine substitution in the active pocket of KOD pol could create favorable conditions for the incorporation of modified dATP carrying sterically demanding groups, such as a flexible linker and fluorescence dye. Taking inspiration from this approach, we designed a variant of KOD DNA polymerase called Mut_1 (D141A/E143A, L408A/Y409A/P410A), as the parent variant in this round of mutagenesis. Notably, position 485 in the wild-type KOD pol is already an A, so no further mutation was introduced at this site in Mut_1. Next, we constructed the first-round library based on Mut_1, which involved saturation mutagenesis at positions 408, 409, 410, and 485. The mutant library was constructed (Figure 2C) by performing saturation mutagenesis at position 408 with Y409A/P410A, saturation mutagenesis at position 409 with L408/A/P410A, saturation mutagenesis at position 410 with L408A/Y409A, and saturation mutagenesis at position 485 in the case of the mutant with L408A/Y409A/P410A. These libraries were subjected to enzyme activity screening to investigate the impact of specific residue positions on the catalytic efficiency, which laid the groundwork for subsequent combinatorial mutagenesis.

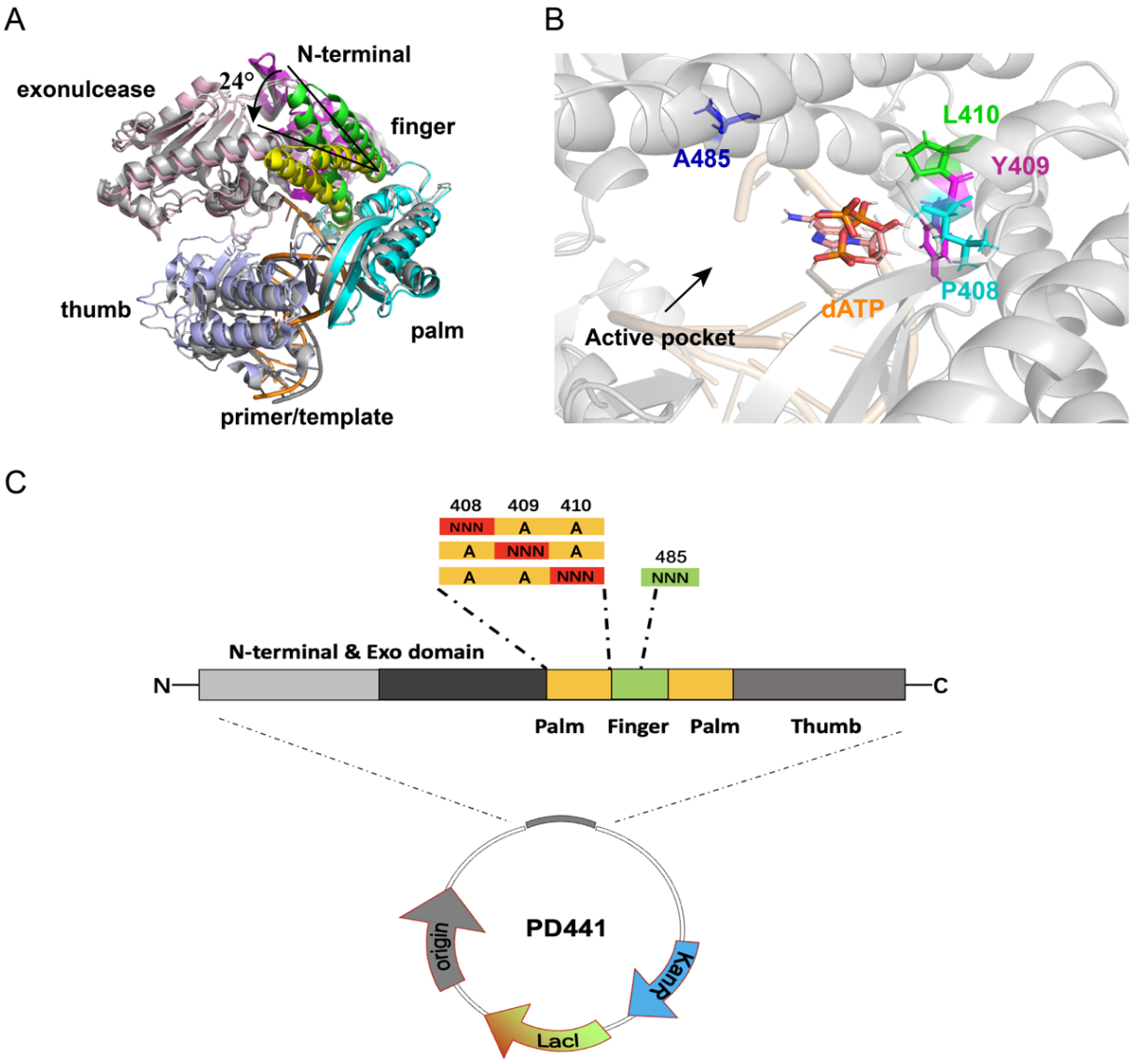


Figure 2. Illustration of structures of KOD pol and demonstration of library construction. A) Depiction of the superimposed structures of a ternary KOD complex (close state, PDB ID: 5MOF) and a binary KOD complex (open state, PDB ID: 4K8Z), which are shown in cartoon format. The finger domain of the ternary complex is shown in yellow and closed by an inward movement of approximately 24˚, in comparison to the finger domain of the binary complex shown in green. The other domains of the KOD ternary complex are colored in gray, the exonuclease of the binary complex in light pink, the N-terminal of the binary complex in pink, the thumb domain of the binary complex in light blue, the palm domain of the binary complex in cyan, and the primer-template of the KOD ternary and binary complex are colored in gray and orange, respectively. B) Illustration of the active pocket of KOD pol with position A485 colored in blue, L408 colored in cyan, Y409 colored in pink, and P410 colored in green. All residues are shown as sticks. The dATP is shown as sticks and depicted in orange elements, and the rest of the protein is shown in cartoon format and colored in gray. C) Illustration of the first-

round constructed library, which includes saturation mutations at positions 408, 409, 410, and 485. The plasmid is PD441 which is a high copy number vector for the expression of E. coli flagellin from an IPTG-inducible T5 promoter.

Mutant screening of the first-round mutagenesis and combinatorial (the second-round) mutagenesis

The screening results of the first-round library are presented in Figure 3. The data were shown as relative reaction velocity, V_{rel} as a measure for the enzyme activity of KOD mutants. V_{rel} of KOD variants was calculated with equation 1:

$$V_{rel} = (V|_{Mut})/(V|_{Mut\_1}) \quad \text{(eq. 1)}$$

$V|_{Mut\_1}$ represents the incorporation rate V (RFUs/min) of Mut_1, while $V|_{Mut}$ represents the incorporation rate V (RFUs/min) of the given variant. The parent variant Mut_1 had a V_{rel} value of 1, as shown in the top center in Figure 3A. When the V_{rel} value is 0, it signifies that the variant under study is unable to incorporate modified dATP or exhibit any measurable enzymatic activity.

At position 408, eight mutations (A/I/Y/M/P/V/S/C) were identified that showed catalytic activity for modified dATP incorporation, shown in Figure 3A (left). L408I is depicted in the top left of Figure 3A in the active site structure. Only two mutations at position 409, A and G, showed enzymatic activity, with the V_{rel} of mutation of the natural Y to A (Figure 3A, center) being higher than that to G. In our test, smaller amino acids such as alanine and glycine displayed the strongest positive effects at this position, suggesting it is a potential role as a steric gatekeeper in KOD pol. Residues that function as steric gatekeepers in the DNA polymerase family are typically highly conserved, and this phenomenon is also evident in other B-family DNA polymerases. For example, in 9 °N DNA polymerase, which belongs to the family B DNA polymerases, the conserved steric gate residue is Y409[48,49]. At position 410, nine mutations (I/C/S/V/P/L/M/G/A) exhibited an increased value of V_{rel} (Figure 3A right), including the original residue of the WT enzyme, P410 (Figure 3A above right), which exhibited the best performance in incorporating modified dATP, on the background of Mut_1. Most variants at position 485 exhibited an increased V_{rel} compared to Mut_1, shown in Figure 3B. Among these, six mutants, including F/I/Y/L/K/E, displayed values of V_{rel} more than three times higher than Mut_1, with E exhibiting the highest enzymatic activity. Residues L/E/A at position 485 are shown in the active site structure in the top of Figure 3B. A previous study highlighted that A485L is an essential mutation for nucleotide analog discrimination[50]. Interestingly, we found

that 485E performed satisfactorily in modified dATP incorporation, possibly due to the additional mutations included in motif A (L408A/Y409A/P410A) in our study compared to previously reported polymerase variants[45]. In addition, we observed relatively stable inter-subdomain contacts between A485E and Q332 (as illustrated in Figure 3B, top center). The contacts of those two residues may provide support for conformational changes in the finger domain, allowing for a rapid transition from a closed state to an open state during the polymerization process.

By individually evaluating saturation mutations at positions 408, 409, 410, and 485, some variants with catalytic activity were obtained. These findings suggest that mutations in specific residues play a crucial role in determining the catalytic activity of KOD pol. Further investigation is necessary to explore whether there exists any potential synergy for the incorporation efficiency of modified nucleotides between those variants that were found to have catalytic activity.

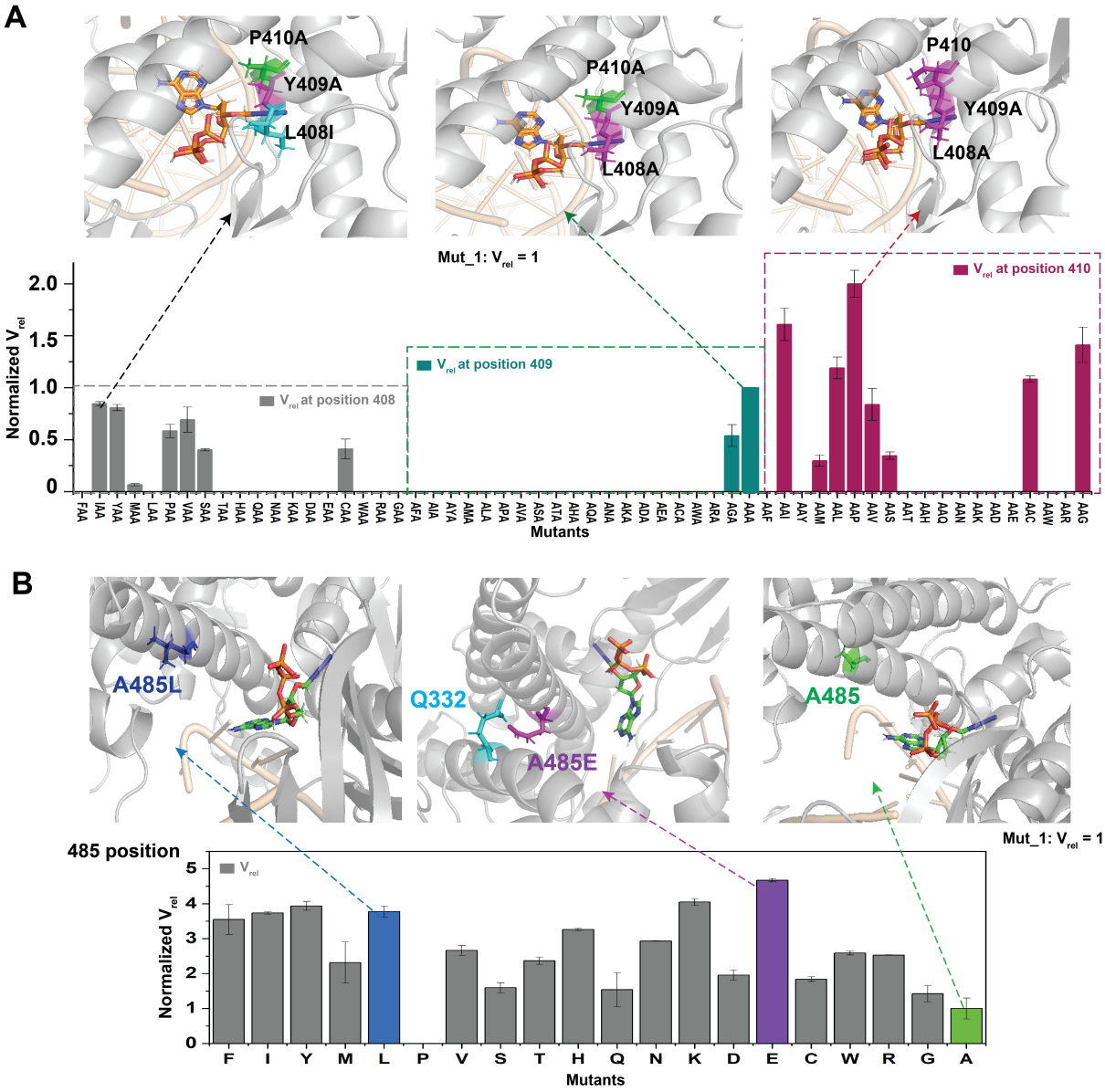


Figure 3. Summary of the screening results of saturation mutagenesis at positions 408, 409, 410, and 485 of KOD pol. A) The screening results of amino acid substitutions at positions 408, 409, and 410, are depicted as V_{rel} values. The bar charts show the screening results of saturation mutagenesis in positions 408 (gray), 409 (dark green), and 410 (red). The upper panels show locations of L408I (cyan), Y409A (pink), and P410A (green) amino acids as well as 3'-O-azidomethyl-dATP (orange) in the crystal structure of KOD DNA polymerase (PDB ID: 5MOF). The 3'-O-azidomethyl-dATP replaces the original dATP and was placed using homology modeling and the mutagenesis was performed by PyMOL. L408I (pink), Y409A (pink), P410A (green), and 3'-O-azidomethyl-dATP (orange) are shown as sticks. The rest of the protein structure is shown in cartoon format and colored grey, and the DNA strand is shown as cartoon and colored in wheat. B) The screening results of saturation mutagenesis at position

485, as V_{rel} like in A). The blue bar represents the A485L substitution, with the corresponding structure shown above. The purple bar represents the A485E substitution, with the corresponding structure shown above and an observed interaction with Q332 (cyan). The green bar represents the original A485 residue and its corresponding structure. The data were analyzed and plotted with OriginPro.

Based on the screening results of the previous round of saturation mutagenesis, mutants with increased V_{rel} values at positions 408/409/410/485, were selected for the second round of combinatorial mutant library construction, shown in Figure 4A. The screening results of thirty combined mutants from the combinatorial mutant library are displayed in Figure 4B. Eight mutants showed a V_{rel} increase of more than 8-fold compared to Mut_1 (Figure 4C), among which Mut_C2 (D141A/E143A, L408I/Y409A, A485E) exhibited the highest increase in V_{rel}, which was over 16-fold compared to Mut_1. In addition, most of those combinatorial mutants exhibited a higher catalytic efficiency compared to single-point mutants.

In order to further assess the catalytic efficiency of Mut_C2, tailored kinetic assays were developed for P/T-2Cy5 and 3'-O-azidomethyl-dATP labeled with the Cy3 dye (modified dATP), separately. These kinetic assays were conducted with varying concentrations of either P/T-2Cy5 or modified dATP ranging from 0 to 6 μM while maintaining the other one at a constant concentration of 2 μM. Fundamental kinetic parameters k_{cat} and K_m were obtained from nonlinear regression of the Michaelis-Menten equation. The kinetic results of Mut_C2 are shown in Figures 4E and 4F. The ratio of k_{cat}/K_m for the variation of P/T-2Cy5 was 1263.3, and the ratio of k_{cat}/K_m for the variation of modified dATP was 887.2. Notably, Mut_C2 exhibited similar performance in both types of substrate variations, indicating an improved efficiency in incorporating modified dATP into the DNA strand. The kinetic data of WT KOD pol shows only a flat line (Figures 4E and 4F), indicating that it failed to incorporate modified dATP, as expected. Mut_C2 was selected as the parent sequence for the next round of mutagenesis.

Following the evaluation of residues situated in the catalytic active center, our next step is to conduct a further assessment with a focus on the critical residues present in the DNA strand binding region of KOD pol. This involves investigating and analyzing the specific amino acid residues that are essential for binding to DNA and interacting with the polymerase enzyme.

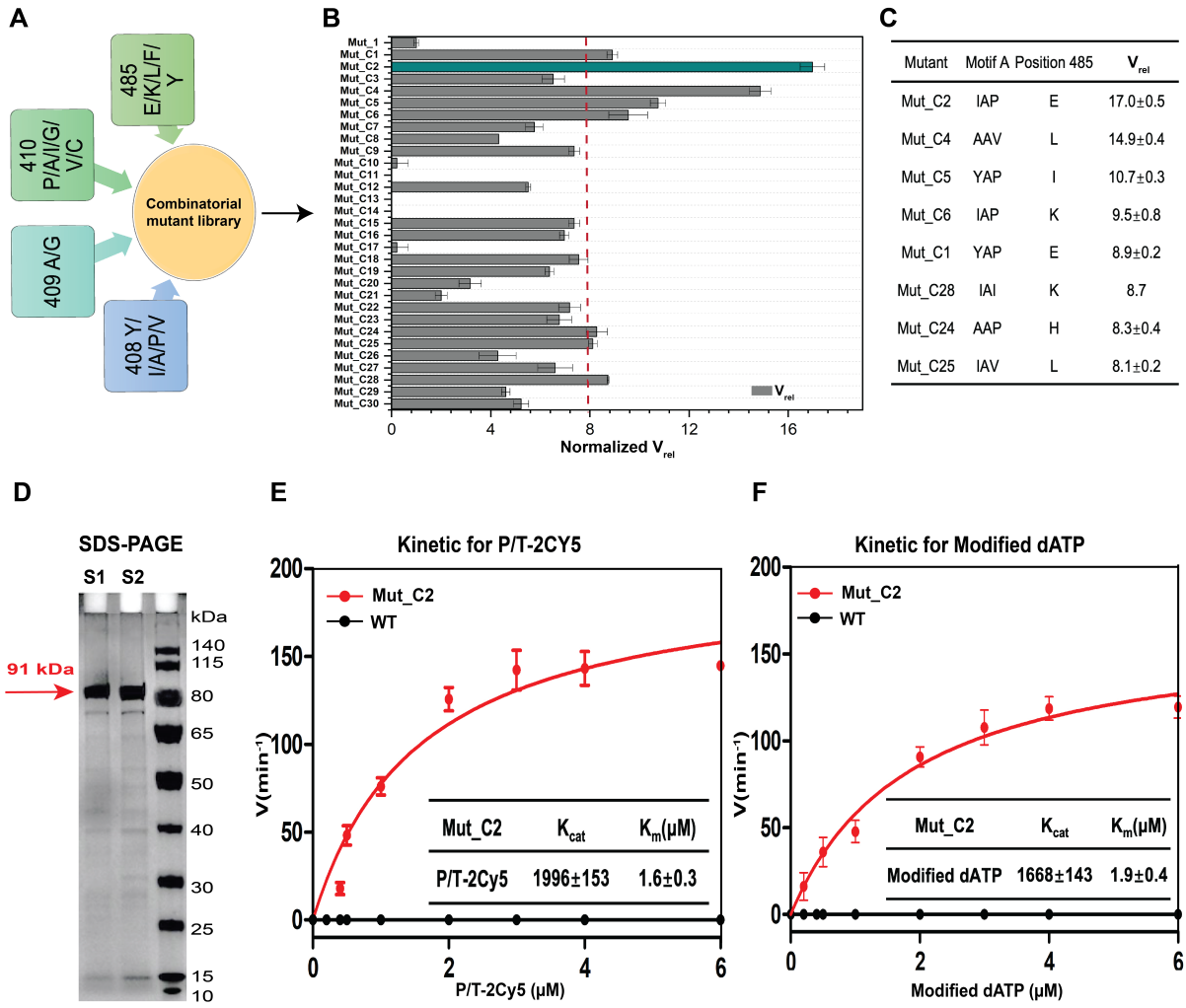


Figure 4. Overview of the screening results of combined mutagenesis at positions 408, 409, 410, and 485 of KOD pol. A) Schematic diagram of the combinatorial mutant library construction. The library was constructed by site-directed mutagenesis, combinatorial mutagenesis, and degenerate codon mutagenesis. Amino acid substitutions selected at position 408 include Y/I/A/P/N, at position 409 are A/G, at position 410 are P/A/I/G/V/C, and at position 485 are E/K/L/F/Y. B) displays the results of screening thirty combined mutants from the library, measuring their V_{rel} values which are shown in a bar graph. Mut_C2 (in dark green) showed the highest increase in V_{rel}, over 16-fold compared to Mut_1. The other mutants are colored gray. The data were analyzed and plotted using OriginPro. C) The mutants resulted in a more than 8-fold increase in V_{rel} compared to Mut_1. D) SDS-PAGE analysis (12% gel) with WT KOD pol represented by S1 and KOD variant Mut_C2 represented by S2, respectively. The amount of protein loaded was around 1.0 µg, and the purification method involved lysing cells with lysozyme at 37 °C for 10 min and centrifugation after heating at 80°C for 30 min. The gel was run at 120V voltage and 1-2 h running time at room temperature. E) and F) The kinetic test

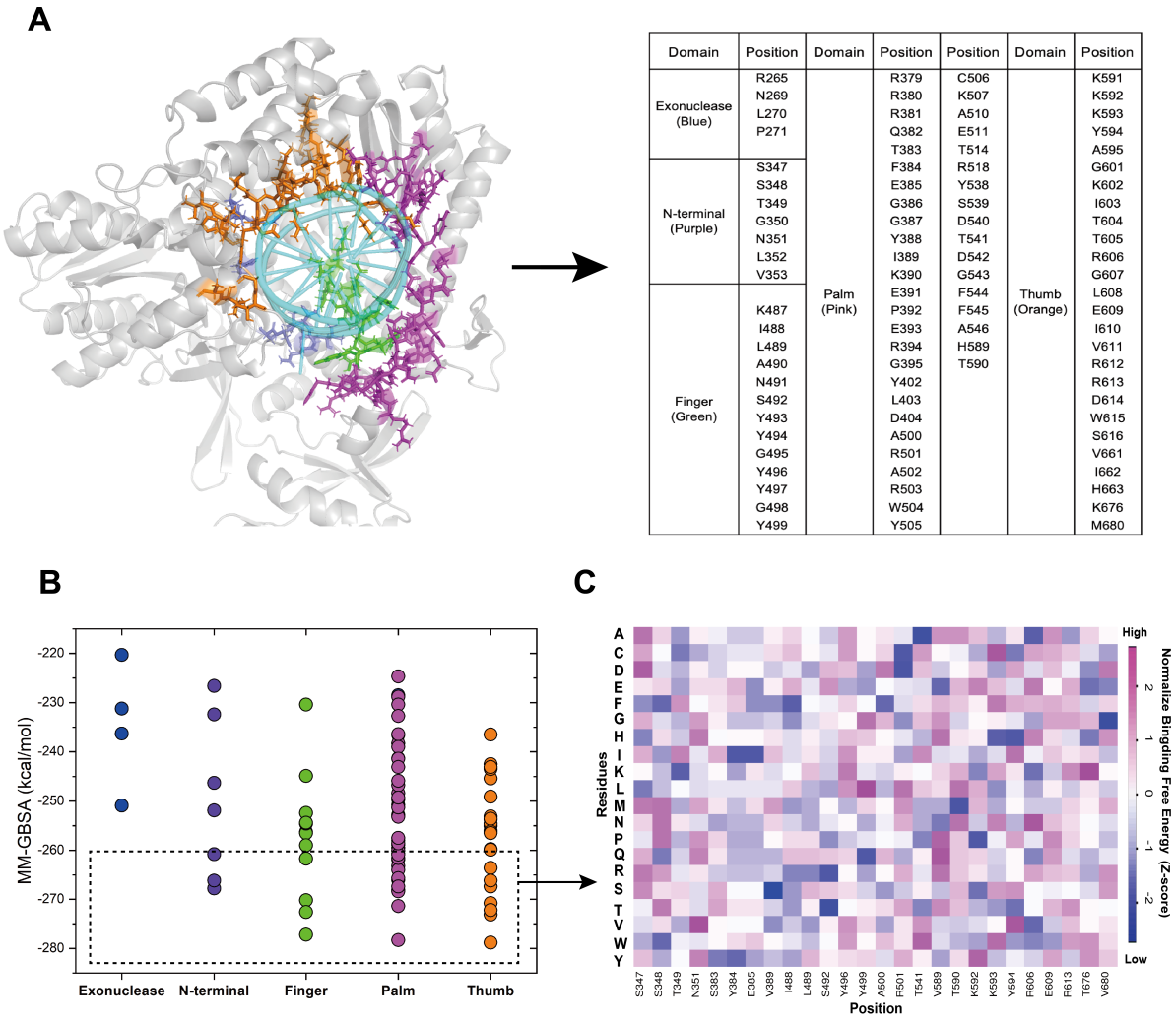
results of Mut_C2 and WT KOD pol towards variations of P/T-2Cy5 (E) and modified dATP (F) under the otherwise same conditions. The experimental data were fitted non-linearly using the Michaelis-Menten equation with GraphPad Prism 5, and the values of k_{cat} and K_m are presented in the table insert the graph. Each kinetic experiment was performed in triplicate.

Computational screening

The identification of pivotal residues implicated in DNA binding can facilitate subsequent enhancements in DNA polymerase catalytic activity. To achieve this goal of identifying key residues involved in the DNA strand binding region of KOD pol, we selected 93 specific amino acid residues, most of them located within a 4 Å distance from the DNA strand in the crystal structure of KOD pol (PDB ID: 5MOF) (Figure 5A). These residues were chosen based on their potential to interact with the DNA through hydrogen bonds, salt bridges, and other types of interactions critical for DNA replication. The selected residues are located across critical domains of KOD pol, including thumb, finger, and palm regions, which play an essential role in DNA replication (Figure 5A). Previous studies reported that the binding process between DNA polymerase and the DNA strand with the correct dNTP pairing is slower than the chemical incorporation process[51]. Raper *et al.* proposed that the nucleotide binding and incorporation must be much faster than the binding equilibration of a polymerase and DNA (E + DNA ⇌ E·DNA)[52], indicating the binding of DNA polymerase and DNA strands may be a rate-limiting step in the whole process. Therefore, mutations at these selected positions will affect the binding between KOD pol and DNA strands, either increasing or decreasing the binding rate, which could in turn affect the overall efficiency of KOD pol for polymerizing unnatural nucleotides. Mutagenesis in the DNA strand binding region provides valuable insights for a comprehensive understanding of the mechanism underlying the polymerization of modified nucleotides by KOD pol.

Virtual saturation mutagenesis was performed for the selected 93 residues using molecular dynamics simulation and the binding energy with dsDNA was predicted for each mutant. The workflow of simulation and calculation involving a method called MM-GBSA is presented in Supplementary Figure 2 and Supplementary Figure 3. We developed an automatic processing pipeline in Python to perform this virtual screening, including mutant pretreatment, molecular dynamics simulation, and binding energy calculation (Supplementary Figure 2). The average binding energy value of saturation mutations of the 93 residues, obtained from the MM-GBSA method, is plotted in Figure 5B and listed in Supplementary Table 1. All these residues with an average computed binding energy < -260 kcal/mol, which was lower than that of WT KOD pol,

were selected for further investigation (Figure 5B, dash box). Some interesting positions previously identified by Kropp *et al*[10], including residues 349, 383, 389, 541, 592, and 606, were also included. The resulting next-round library consisted of 26 distinct mutation sites, and their binding energy was visualized as a heat map in Figure 5C. Variants with more negative computational free energy values are expected to exhibit stronger binding affinities to dsDNA. Finally, we selected two variants at each site with the most negative binding free energies for further experimental screening.



Domain	Position	Domain	Position	Position	Domain	Position
Exonuclease (Blue)	R265		R379	C506		K591
	N269		R380	K507		K592
	L270		R381	A510		K593
	P271		Q382	E511		Y594
			T383	T514		A595
N-terminal (Purple)	S347		F384	R518		G601
	S348		E385	Y538		K602
	T349		G386	S539		I603
	G350		G387	D540		T604
	N351		Y388	T541		T605
	L352		I389	D542		R606
	V353		K390	G543		G607
Finger (Green)	K487	Palm (Pink)	E391	F544	Thumb (Orange)	L608
	I488		P392	F545		E609
	L489		E393	A546		I610
	A490		R394	H589		V611
	N491		G395	T590		R612
	S492		Y402			R613
	Y493		L403			D614
	Y494		D404			W615
	G495		A500			S616
	Y496		R501			V661
	Y497		A502			I662
	G498		R503			H663
	Y499		W504			K676
			Y505			M680

Figure 5. Results of computational screening of residues involved in polymerase binding to dsDNA. A) The location of 93 specific amino acid residues, most of them located within a 4 Å distance from the DNA strand in the crystal structure of KOD DNA polymerase (PDB ID: 5MOF). The corresponding residues are shown as sticks and color-coded on the structure: blue for the Exo domain, purple for N-term, green for the finger, pink for the palm, and orange for the thumb subdomain. The rest of the protein structure is shown in cartoon format and colored grey. The DNA strand is shown in cartoon format and colored cyan. The right panel displays a

table presenting all residues that were selected for MD simulation. B）The average binding free energy value of the 93 residues calculated by MM-GBSA, with each region color-coded accordingly. The residues within the black dashed box were predicted to exhibit stronger binding affinity. C) Heatmap of the binding free energy of 26 distinct mutation sites selected from panel B (dashed box).

Mutant screening of the third-round mutagenesis

The third-round library consisting of 52 mutants containing the 26 critical amino acids identified in virtual screening was constructed by site-direct mutagenesis. Mut_C2 (D141A/E143A, L408I/Y409A, A485E) which exhibited the best performance in the second-round screening, was used as the parent or starting variant for constructing the library in this round. The enzymatic activity of these mutants was quantified using equation 2, which is similar to equation 1 except that Mut_C2 serves as the reference. The V_{rel} of Mut_C2 was therefore 1.

$$V_{rel} = (V|_{Mut})/(V|_{Mut\_2}) \quad \text{(eq. 2)}$$

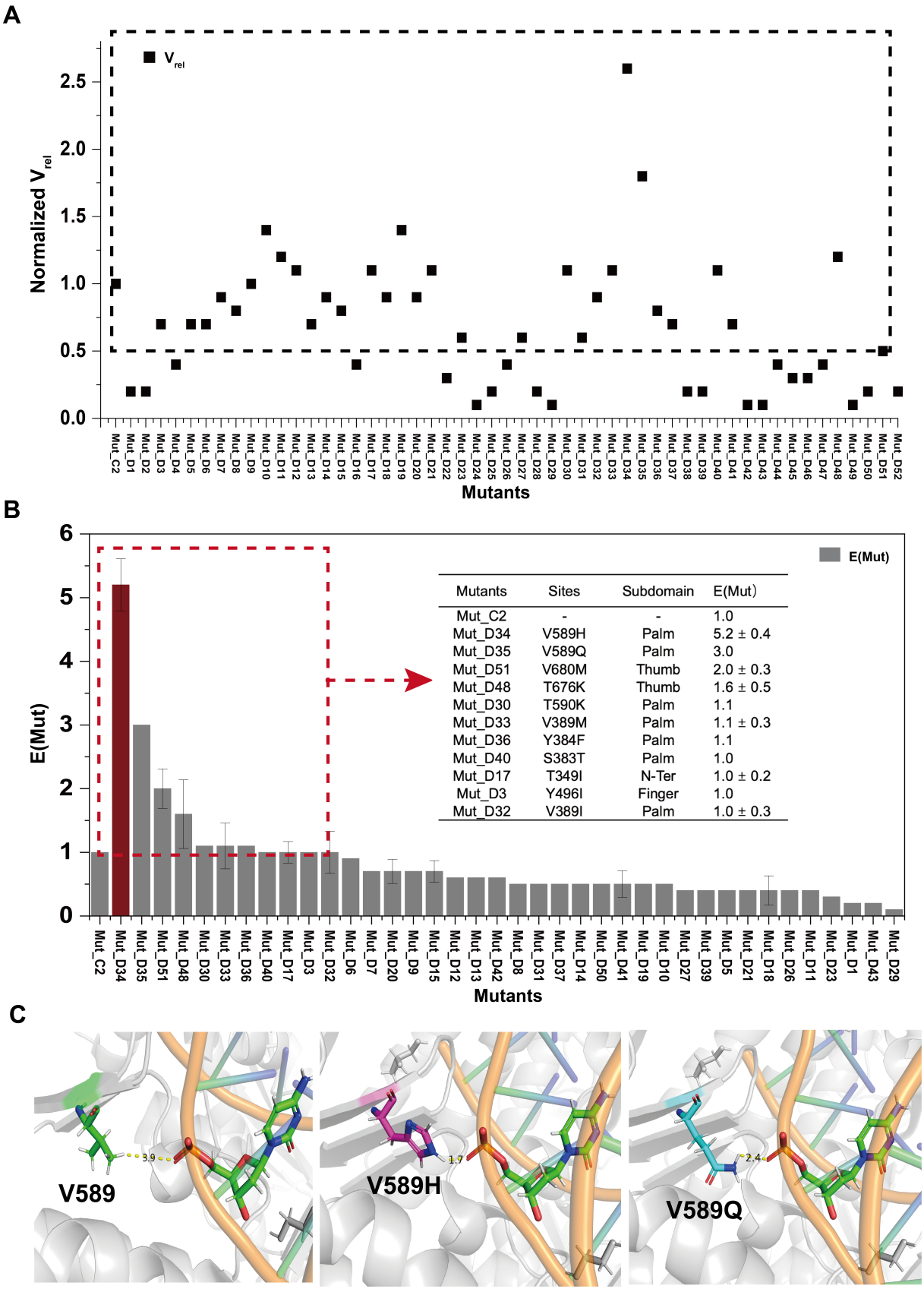
The kinetic performance of KOD variants is described by the Michaelis-Menten parameters k_{cat} and K_m. The relative kinetic performance of those mutants is characterized by the ratio of $k_{cat}/K_m|_{Mut}$ relative to that of Mut_C2. $k_{cat}/K_m|_{Mut}$ and $k_{cat}/K_m|_{Mut\_C2}$ were measured and calculated in the same way, with varying P/T-2Cy5 ranging from 0 to 6 μM while maintaining the modified dATP at a constant concentration of 2 μM. We define the relative kinetic performance of the mutant variants as E(Mut), as shown in Equation 3. The subscript after the vertical bar "|" indicates the mutant used for measurement. The Mut_C2 variant served as the reference, with a catalytic efficiency (E(Mut)) value of 1.

$$E(Mut) = (k_{cat}/K_m|_{Mut})/(k_{cat}/K_m|_{Mut\_C2}) \quad \text{(eq. 3)}$$

The enzyme activity of 52 variants was screened under the same experimental conditions as described above, and their V_{rel} values were calculated and compared (Figure 6 A). 14 variants displayed a V_{rel} value greater than that of Mut_C2, with one variant, Mut_D34 (V589H), exhibiting an over 2-fold greater V_{rel} to Mut_C2. We selected 39 mutant variants, with a V_{rel} value higher than 0.5, to perform further kinetic screening measurements. Although some mutant variants did not show significantly higher than Mut_C2, we still selected them because they were located close to the DNA strand in the structure of KOD DNA polymerase.

E(Mut) of these variants was determined and presented in Figure 6 B. Among this set, 11 variants displayed more favorable E(Mut) values than Mut_C2, listed in the table embedded in Figure 6 B. Mut_D34, which contains the V589H mutation, outperformed the other variants

with the highest observed E(Mut) value of 5.2. Mut_D35, carrying the V589Q mutation with no charged side chain, showed improved kinetic activity (E(Mut) = 3.0) compared to Mut_C2 (V589). The observed increase in catalytic activity in Mut_D34 and Mut_D35 can be attributed to the presence of V589H and V589Q mutations, which are located in the palm domain of KOD pol. The distance of atom contacts between residues V, H, and Q at position 589 and the DNA strand in the crystal structure of KOD pol (PDB ID: 5MOF) is shown in Figure 6C, which are 3.9 Å, 1.7 Å, and 2.4 Å, respectively. A closer distance implies a higher possibility of hydrogen bonds forming, which can subsequently impact the binding between DNA and the protein. In addition, we also speculated that the positively charged side chain of histidine (H) at position 589 increased the binding affinity between KOD pol and the DNA strand, facilitating the incorporation process by virtue of strong and favorable electrostatic interaction with the double-stranded DNA. This has been supported by previous studies[39]. For example, Kropp *et al.* reported that the structure of KOD pol has a long crack with an electro-positive potential extending from the DNA binding region at the thumb domain upwards along the β-hairpin and the palm domain to the N-terminal domain[11]. The increased positive charge introduced in the DNA binding region of the polymerase results in enhanced affinity towards the template DNA, leading to a more stable binding[53]. Mut_D51, which carried the V680M mutation, and Mut_48 carrying the T676K mutation, demonstrated 2-fold and 1.6-fold improvements in E(Mut) compared to Mut_C2, respectively. Among the investigated mutation sites, five were located in the palm domain (V589H/Q, T590K, V389I, S383T, Y384F), two in the thumb domain (T676K, V680M), one in the finger domain (Y496I), and one in the N-terminal region (T349I). Among the mutation sites investigated in this study, residues 383/384/389 were located in the loop of the KOD pol structure, while residues 589/590/676/680 contributed to the binding between KOD pol and the DNA strand. The screening results showed that those mutated residues either directly or indirectly influence the binding of KOD pol and the DNA strand, leading to an increase in the incorporation efficiency during polymerization. Therefore, we performed further combinations with these well-performing mutants to explore possible interactions and synergies between the identified sites. To accomplish this, we generated stepwise combination variants based on Mut_D34 exhibited the best performance in this round of screening.



Mutants	Sites	Subdomain	E(Mut)
Mut_C2	-	-	1.0
Mut_D34	V589H	Palm	5.2 ± 0.4
Mut_D35	V589Q	Palm	3.0
Mut_D51	V680M	Thumb	2.0 ± 0.3
Mut_D48	T676K	Thumb	1.6 ± 0.5
Mut_D30	T590K	Palm	1.1
Mut_D33	V389M	Palm	1.1 ± 0.3
Mut_D36	Y384F	Palm	1.1
Mut_D40	S383T	Palm	1.0
Mut_D17	T349I	N-Ter	1.0 ± 0.2
Mut_D3	Y496I	Finger	1.0
Mut_D32	V389I	Palm	1.0 ± 0.3

Figure 6. Screening results of third-round variants constructed based on Mut_C2. A) The

enzyme activity screening results of these variants. The V_{rel} (black square) of these variants was calculated according to equation 2 and plotted using Origin 2021. The variants with a $V_{rel} > 0.5$ are included inside the dashed box. B) The relative kinetic screening results of those variants with a $V_{rel} > 0.5$ are shown in the dashed box in Panel A. The kinetic parameters of the Michaelis-Menten equation, such as k_{cat} and K_m, were calculated by analyzing the data using GraphPad Prism 5 software. E(Mut) of these mutants was calculated according to equation 3 and plotted using Origin 2021. Mut_D34 corresponds to the red column, while the other variants are shown as grey bars. B) The site information, location, and kinetic results of variants with E(Mut) > 1. C) The location of residue 589 with V/H/Q placed in the crystal structure of KOD DNA polymerase (PDB ID: 5MOF). The DNA strand is depicted in a cartoon format, with the backbone-colored orange, while one of the deoxyribonucleotides is shown as sticks and colored by element. The residues at position 589 are displayed as sticks and colored green for V (in the left panel), magenta for H (in the middle panel), and cyan for Q (in the right panel). Atom contacts identified between residues (V, H, Q) at position 589 and the DNA strand are illustrated as an example. The values of V_{rel} and E(Mut) were summarized in Supplementary Table 2. The distances predicted by the PyMOL software are represented by dashed yellow lines.

Stepwise combination of effective mutations

Based on the last-round screening results, the best-performing variant Mut_D34 was selected as the parent for a fourth round of combinational mutagenesis to further improve the catalytic activity. To comprehensively assess the catalytic performance of stepwise combined variants, we conducted kinetic assays towards both substrates of P/T-2Cy5 and modified dATP. These assays allowed us to evaluate the kinetic performance of each new variant individually with respect to the DNA strand and modified dATP. The relative kinetic performance was evaluated using Equation 4, which is similar to Equation 3. In this round of screening, the Mut_D34 variant served as a reference, with a catalytic efficiency (E(Mut)) value of 1.

$$E(Mut) = (k_{cat}/K_m|_{Mut})/(k_{cat}/K_m|_{Mut\_D34}) \quad \text{(eq. 4)}$$

To distinguish the relative kinetic performance of two substrates, we used E(Mut)/P/T-2Cy5 to represent the kinetic measurement towards varying P/T-2Cy5. Similarly, E(Mut)/modified dATP represents the kinetic measurement towards varying modified dATP.

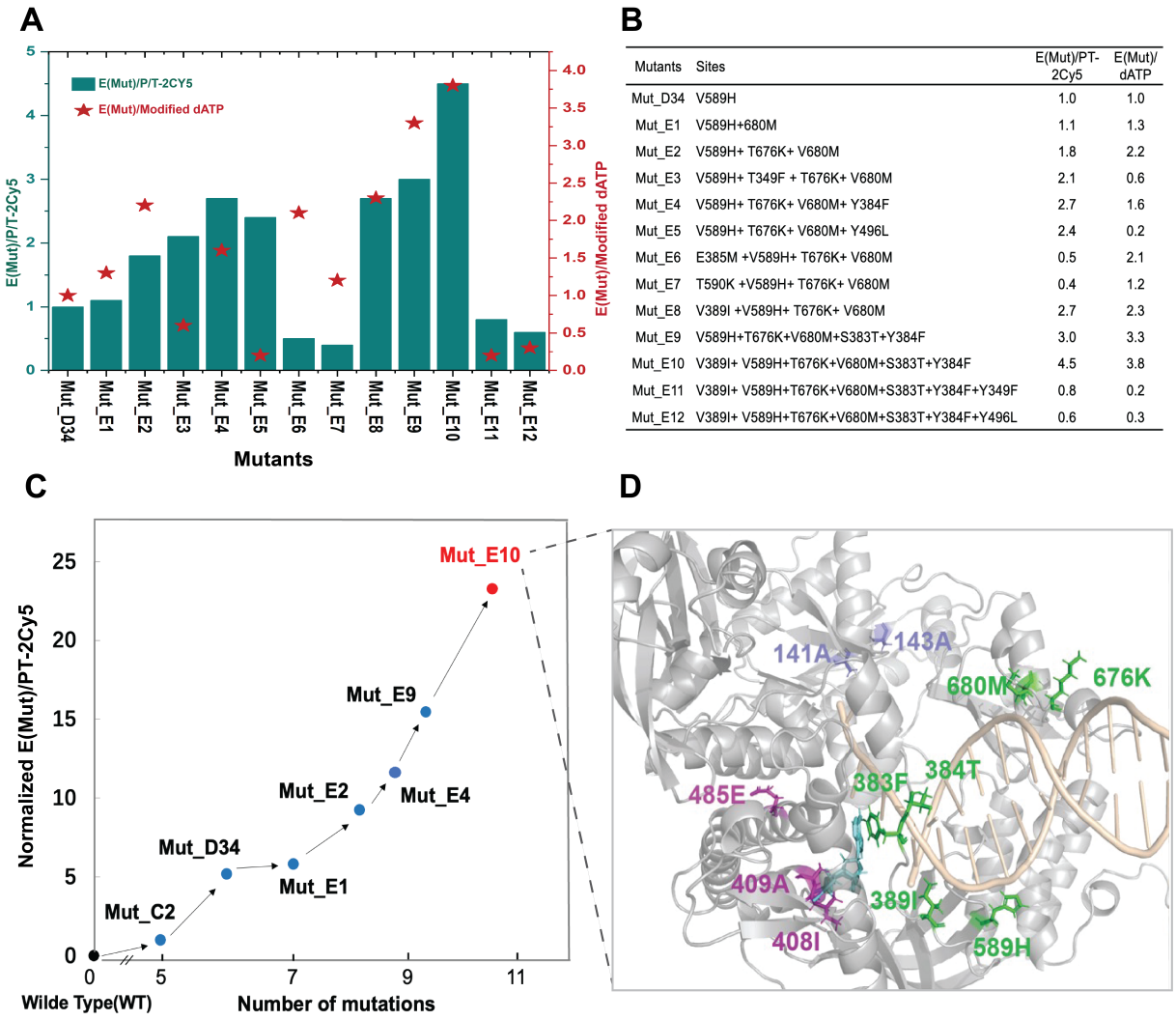
The screening results for E(Mut)/P/T-2Cy5 of combined variants are shown in Figures 7A and 7B. We combined V680M (Mut_D51) with the mutations of variant Mut_D34, generating the combination variant Mut_E1 which displayed slight improvement in E(Mut), confirming that the V680M mutation stacked on Mut_D34 improved catalytic performance. We then added

T676K (Mut_D48) to Mut_E1, which introduced a positive charge located on the thumb domain. The results demonstrated that these three mutation sites displayed positive interactions for improving polymerase catalytic efficiency. The resulting improved variant was designated as Mut_E2. Further single site mutations were introduced to Mut_E2, including T590K, V389H, Y496L, T349F, E385M, and Y384F. Four mutants carrying specific residues, namely Mut_E3 (Y349F), Mut_E4 (Y384F), Mut_E5 (Y496L), and Mut_E8(V389I), displayed increased E(Mut) compared to Mut_E2. We then generated a two-point and three-point superposition of S383T, Y384F, and V389I, leading to the variants Mut_E9 and Mut_E10 with further improved kinetic performance. The E(Mut) value of Mut_E9, which carries the mutations S383T and Y384F, showed a 3-fold increase compared to Mut_E1 and approximately a 1.6-fold increase compared to Mut_E2. The E(Mut) value of Mut_E10 carrying S383T, Y384F, and V389I, showed a 4.5-fold increase compared to Mut_E1. However, attempts to further combine Mut_E10 with Y349F and Y496L did not result in an improved variant.

E(Mut)/modified dATP values of these mutants were calculated and are also shown in Figures 7A and 7B. Interestingly, we found that some variants behaved inconsistently in kinetic experiments against P/T-2Cy5 and modified dATP. For example, Mut_E3 and Mut_E5 exhibited a two-fold increase compared to Mut_D34 in the P/T-2Cy5 kinetic assay but displayed poorer kinetic performance in the modified dATP. Except for some variants that exhibit inconsistent performance with P/T-2Cy5 and modified dATP, most exhibit consistent behavior. Notably, Mut_E10 exhibited the highest performance for variations of both substrates, indicating a high level of efficiency in incorporating modified dATP into the DNA strand. Overall, our data show that introducing combinatorial mutations at specific positions within the DNA-binding region of KOD pol significantly affects its efficiency in incorporating modified nucleotides during the polymerization process.

The whole semi-rational evolution process leading to the identification of Mut_E10 is depicted in Figures 7C and 7D. With two-round libraries screening, we obtained a five-mutation variant Mut_C2 (D141A/E143A, L408I/Y409A, A485E). Mut_C2 demonstrated a higher catalytic efficiency in incorporating modified dATP compared to wild-type KOD pol. We then performed virtual screening and selected several mutation sites located in the DNA binding region for further investigation. By combining the V589H mutation with the mutations present in Mut_C2, we obtained a variant called Mut_D34. Mut_D34 exhibited a 5.2-fold improvement in E(Mut)/PT-2Cy5 compared to Mut_C2. Based on Mut_D34, we conducted stepwise combinatorial mutation screening and ultimately obtained a best-performance variant Mut_E10.

Mut_E10 exhibited an exceptional catalytic efficiency, with a relative increase of over 23-fold in E(Mut)/PT-2Cy5 compared to Mut_C2. This indicates a significant improvement in the polymerase's ability to incorporate modified dATP. Our results highlight the importance of combinatorial mutations of multiple amino acids in improving the incorporation efficiency of KOD pol for modified nucleotides.



Mutants	Sites	E(Mut)/PT-2Cy5	E(Mut)/dATP
Mut_D34	V589H	1.0	1.0
Mut_E1	V589H+680M	1.1	1.3
Mut_E2	V589H+ T676K+ V680M	1.8	2.2
Mut_E3	V589H+ T349F + T676K+ V680M	2.1	0.6
Mut_E4	V589H+ T676K+ V680M+ Y384F	2.7	1.6
Mut_E5	V589H+ T676K+ V680M+ Y496L	2.4	0.2
Mut_E6	E385M +V589H+ T676K+ V680M	0.5	2.1
Mut_E7	T590K +V589H+ T676K+ V680M	0.4	1.2
Mut_E8	V389I +V589H+ T676K+ V680M	2.7	2.3
Mut_E9	V589H+T676K+V680M+S383T+Y384F	3.0	3.3
Mut_E10	V389I+ V589H+T676K+V680M+S383T+Y384F	4.5	3.8
Mut_E11	V389I+ V589H+T676K+V680M+S383T+Y384F+Y349F	0.8	0.2
Mut_E12	V389I+ V589H+T676K+V680M+S383T+Y384F+Y496L	0.6	0.3

Figure 7. Screening results of stepwise combinational mutagenesis. Mut_D34 was used as the parental variant in this round, and additional mutation sites were introduced using site-directed mutagenesis. A) Relative kinetic screening results for variants, which include their kinetic performance towards P/T-2CY5 (E(Mut)/P/T-2Cy5, displayed as green bars), as well as their performance towards modified dATP (E(Mut)/modified dATP, displayed as red stars). The kinetic assays were conducted for 1-2 hours at 40°C with varying concentrations of P/T-2CY3 or modified dATP, using excitation at 530 nm and detection at 676 nm. The data were analyzed and plotted with OriginPro. B) The mutation site information and the corresponding dynamic results of combined variants. C) Illustration of the evolution process of Mut_E10 (red spots),

which comprises 11 mutation sites compared to the wild-type KOD pol. During its semi-rational evolution process, key mutants (black spots) include Mut_C2, Mut_D34, Mut_E1, Mut_E2, and Mut_E9. D) The locations of eleven mutation sites from the variant Mut_E10 are shown in the crystal structure of KOD DNA polymerase (PDB ID: 5MOF). These residues are shown as sticks, the rest of the protein structure is shown in cartoon format and colored grey. The sites located in the exonuclease region, including D141A and E143A, are colored light blue. The sites associated with the catalytic activity center, including L408I, Y409A, and A485E, are colored light magenta. The sites related to DNA binding, including S383T, Y384F, V389I, V589H, T676K, and V680M, are colored green. The DNA strand is shown in cartoon format and colored in light wheat. The 3'-O-azidomethyl-dATP is shown in sticks and colored cyan.

Sequencing performance validation of KOD variants in NGS

In order to examine whether the observed improvement in the polymerization of Mut_E10 brought significant values in kinetic performance in sequencing applications, we tested the performance of Mut_E10 in different NGS platforms. Since the launch of the Human Genome Project, next-generation sequencing (NGS) platforms have had a very significant impact on biological research[54]. We first compared the sequencing performance of Mut E10 to that of WT and Mut_C2 on the BGISEQ-500 platform with SE50 sequencing, which can assess the relative improvements achieved through the use of the engineered variants. Furthermore, we conducted a deeper sequencing evaluation of Mut_E10 on the MGISEQ-2000 platform, which involved PE100 sequencing. We performed rigorous purification for WT, Mut_C2, and Mut_E10. The protein purity of each sample was confirmed to exceed 95% through SDS-PAGE analysis, shown in Supplementary Figure 4.

The sequencing results of WT KOD pol, Mut_C2, and Mut_E10 on the BGISEQ-500 platform are presented in Figure 8. The BGISEQ-500 platform featuring cPAS and DNA Nanoballs (DNB™) technology is a widely used NGS platform[55]. The library used for sequencing was *E. coli*, the number of sequencing cycles was SE50, and the temperature was 55°C. We compared the performance of the three different enzymes, WT KOD pol, Mut_C2, and Mut_E10, by replacing the original enzyme in the supplied sequencing kit using the BGISEQ-500 sequencing platform. Notably, the WT KOD pol group did not show any bright spots on the chip (Figure 8A), indicating that no fluorescently labeled probes were extended. In contrast, the Mut_C2 group presented bright spots, indicating the successful extension of modified probes, with the bright spots representing the incorporated fluorescently labeled probes in the DNB event, as shown in Figure 8A. Then, the sequencing results of Mut_E10 and Mut_C2 were analyzed and compared. The quality distribution of the sequencing reads, including Q30, ESR, and mapping

rate, represents the overall sequencing quality of the enzyme-replaced platforms, shown in Figures 8B and 8D. The results revealed that Mut_E10 exhibited superior performance compared to Mut_C2 which has an apparent decrease after 20 cycles. A higher Q value indicates a lower probability (P) of base misidentification[56]. Mut_E10 had a higher Q30 (> 93%) throughout the entire cycle, making it highly suitable for NGS applications. The data revealed that Mut_E10 exhibited superior performance in terms of lag% and runon% compared to Mut_C2 (Figure 8C). These parameters reflect the incorporating ability of Mut_E10 as the sequencing enzyme in the NGS application. Based on these results, we conclude that Mut_E10 is a promising candidate for the BGISEQ-500 sequencing platform. Subsequently, we investigated the performance of Mut_E10 on another sequencing platform, CoolMPS™ sequencing.

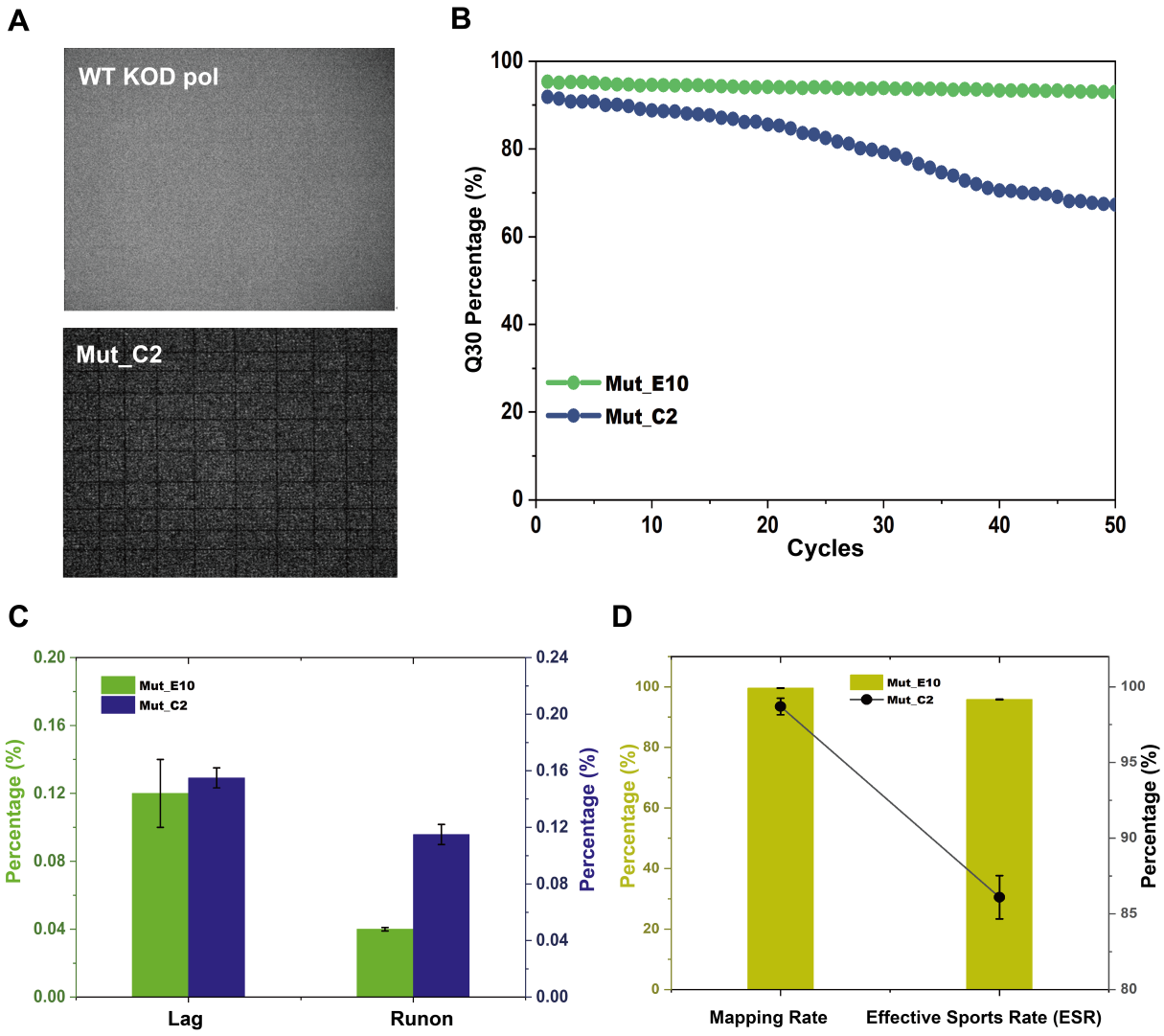


Figure 8. Sequencing results of WT KOD pol, Mut_C2, and Mut_E10 on the BGISEQ-500 platform. A) Representative images in the sequencing process. During the sequencing process, DNA polymerases polymerize modified nucleotides and then remove excess fluorescence followed by imaging of the chip. The WT KOD pol group showed no bright spots on the chip, indicating no fluorescently labeled probes were extended. In contrast, the Mut_C2 group exhibited bright spots, indicating the successful extension of modified probes. The area of a small square in the chip shown in the diagram is 0.882*0.882 mm. B) Quality portion distribution on reads (Q30) for Mut_C2 (blue line) and Mut_E10 (green line). Mut_E10 had a stable curve and performed better in each sequencing cycle in comparison to Mut_C2. C) Lag and Runon comparison for Mut_C2 (blue) and Mut_E10 (green) in sequencing results. Mut_E10 had lower values, indicating better extension efficiency. D) Mapping rate and ESR comparison between Mut_C2 (black squares) and Mut_E10 (olive). Mut_E10 had a higher mapping rate and ESR, suggesting better sequencing quality.

Furthermore, we applied Mut_E10 on the MGISEQ-2000 platform featuring CoolMPS™ technology for PE100 sequencing. CoolMPS™ is a novel massively parallel sequencing chemistry, which was developed by the Drmanac group at MGI[57]. The sequencing experiments were performed in duplicates on a single chip in Lane1 and Lane2. The library used for sequencing was *E. coli*, the temperature was 55 °C, and the sequencing process was carried out following the published protocols[58]. The significant sequencing results were listed in a concise table in Figure 9A. Both the first and second strands exhibited high mapping rates exceeding 99%. The ESR value was reported to be above 78%, and the average error rate was less than 0.16, further indicating the accuracy and quality of the sequencing data. During the base-calling of sequencing, the Q30 score remained consistently higher than 95% for the first-strand sequencing, and a gradual decrease in Q30 was observed for the second-strand sequencing (Figure 9 B), which is considered acceptable for PE100 testing. The average lag% and runon% values for both strands were comparatively low and are depicted in Figures 9C and 9D. In addition, Korostin reported that Illumina HiSeq 2500 and MGISEQ-2000 have similar performance characteristics after comparing both whole-genome sequencings (PE150) in terms of sequencing quality, number of errors, and performance[58]. Our sequencing data also exhibited comparable performance to that of HiSeq 2500, including high mapping rates (>99%) and Q30 values (>97%), with Q30 being around 96%. The sequencing data obtained is satisfactory for the MGISEQ-2000 sequencing platform, according to the reported range of NGS analysis generated[52,56]. Taken together, these results demonstrate that Mut_E10 is suitable for practical applications in the MGISEQ-2000 platform.

Finally, considering that the enzyme in the original sequencing reagent has been directly

replaced, it is believed that there is still potential for enhancing the sequencing quality of Mut_E10 on both sequencing platforms. One potential avenue for improvement lies in optimizing the sequencing buffer to further enhance the performance and accuracy of the sequencing process[59]. Consequently, the sequencing quality achieved with Mut_E10 on both the BGISEQ-500 and MGISEQ-2000 platforms fell within the acceptable range comparable to the "gold standard" set by NGS analysis generated using the Illumina platform[58].

A

MGISEQ-2000	Mut_E10	
Category	Value_1	Value_2
Reference	E.coli	E.coli
Cycle Number	210	210
ESR (%)	78.74	81.23
Q30 (%)	95.97	96.42
Total Mapping Rate (%)	99.98	99.98
Error Rate 1 (%)	0.16	0.12
Error Rate 2 (%)	0.16	0.12
Split Rate (%)	99.24	99.23
Runon 1 (%)	0.04	0.04
Runon 2 (%)	0.12	0.08
Lag 1 (%)	0.06	0.06
Lag 2 (%)	0.13	0.14

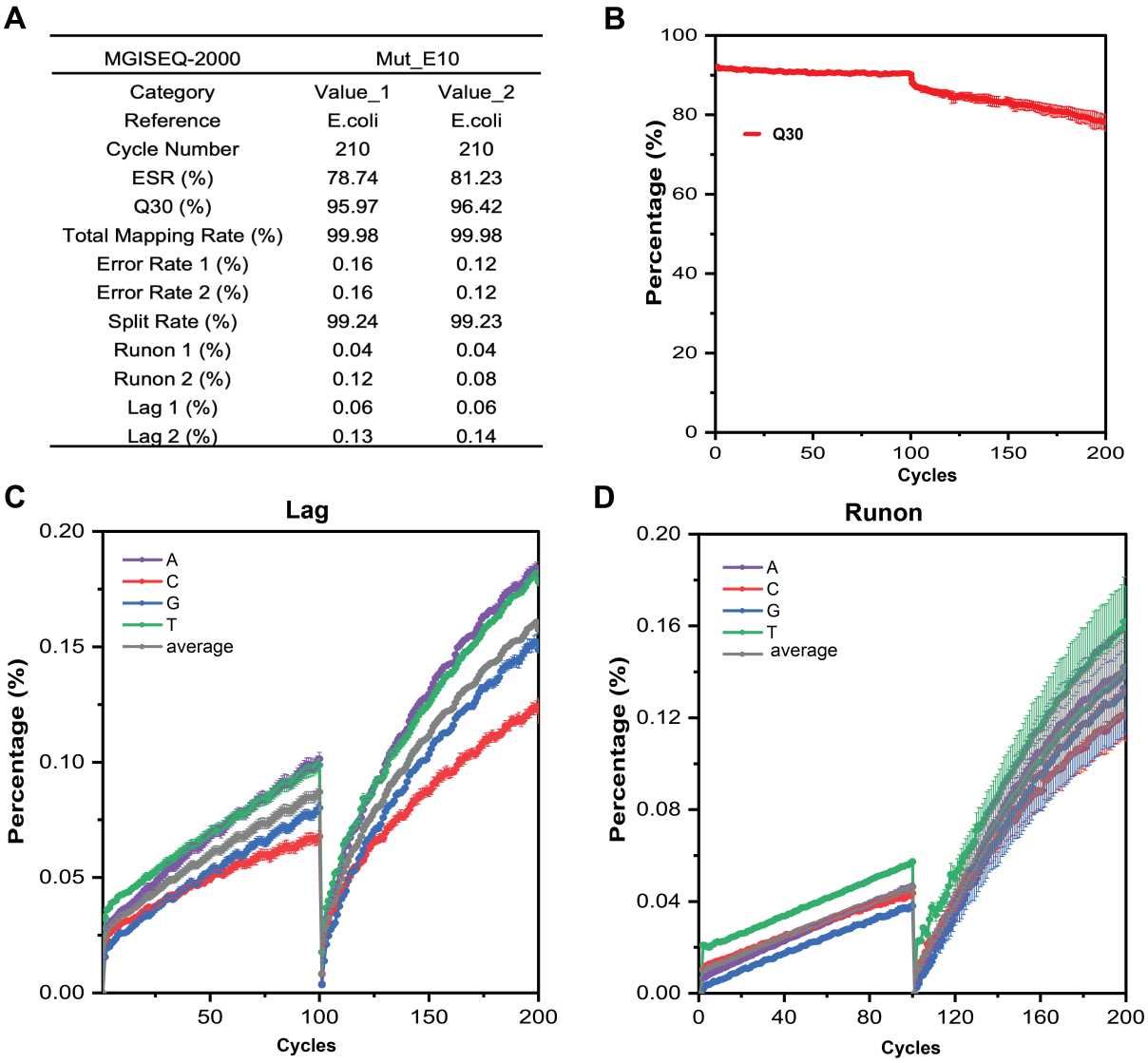


Figure 9. Sequencing results of Mut_E10 on the MGISEQ-2000 platform. A) Overview of the significant sequencing information and major sequencing quality metrics or data, with PE100 sequencing. Note: Error Rate: base error rate; Split Rate: barcode split rate. Value_1 and Value_2 are duplicates of sequencing data obtained from Lane 1 and Lane 2, respectively, on the same chip during sequencing. B) The average Q30 value of Value_1 and Value_2 during sequencing in panel A), which serves as an indicator of base calling accuracy. C) and D) show

the average lag and runon percentages of Value_1 and Value_2 during sequencing in panel A), respectively. These values serve as indicators of the completeness of the polymerization reaction that occurred within the DNB.

Discussion

Mut_E10 displayed satisfactory performance on two types of sequencing platforms including BGISEQ-500 and MGISEQ-2000, as indicated by our sequencing data. These results suggest that Mut_E10 possesses the potential for commercialization and is adaptable to different sequencing platforms. To our best knowledge, the cPAS technology of the BGISEQ-500 platform utilized 3'-O-blocked reversible terminators labeled with four different fluorescent dyes (four colors)[9,60]. On the other hand, the CoolMPS™ technology of the MGISEQ-2000 platform achieved sequencing using monophosphate nucleotides with a 3'-O-azidomethyl blocking group and an NHS linker on phosphate[61,62], which can rapidly bind (within 30 seconds) with nucleobase-specific fluorescently labeled antibodies[57]. Previous studies reported DNA polymerases can incorporate dNMPs at apurinic/apyrimidinic (AP) sites or similar damaged sites but with generally low efficiency and a propensity for error insertion[63,64]. Remarkably, Mut_E10 exhibits high efficiency in incorporating both chemically modified dNTPs and dNMPs. The incorporation efficiency of unnatural nucleotides depends on the linker used to anchor the modification to the nucleobase, especially with sterically demanding groups[33]. Therefore, we speculate that the steric size requirements for the incorporation of chemically modified dNTPs and dNMPs by DNA polymerases may be similar in this study. However, the mechanisms responsible for the incorporation of both unnatural substrates in these two sequencing platforms, are yet to be fully understood. Our analysis focuses on the potential causes or mechanisms by which these identified mutations improve the polymerization efficiency of Mut_E10. Therefore, we conducted further analysis on these identified mutations. Except for the two sites (D141A/E143A) in the exonuclease region, the remaining nine mutation sites can be categorized into three distinct categories.

The first category of mutational sites includes those that affect the binding of Mut_E10 polymerase with the DNA strand. These sites are S383T, Y384F, V389I, V589H, T676K, and V680M, as shown in Figure 7D. MM-GBSA calculations indicated that modifications at these sites enhanced the ability of Mut_E10 to bind the DNA strand, accelerating DNA strand capture. Notably, residues 589H and 676K possess positive charges and are expected to exhibit higher binding affinities for negatively charged DNA when the polymerase binds to the DNA strand. Residues 383T, 384F, and 389I are located near the DNA strand and nucleotides that form the

back of the active pocket likely contribute positively to the formation of a stable active pocket. Furthermore, residues 383T, 384F, 389I, and 676K are situated on the coil structure. To avoid disrupting the original coil structure, we screened for mutants that were similar to the original residues. Interestingly, experimental results revealed that if the properties of the residue side chains at these points changed significantly, the calculated binding energy might be stronger, but the experimental test results did not provide any indications for an increased binding affinity (data not shown). These findings indicate that a stable coil or loop structure of the DNA polymerase plays an important role in modified nucleotide polymerization activity. Excessive and disruptive changes to these residues on the coil structure are thus not recommended.

The second category of sites includes those where mutations altered the dNTP catalytic pockets, such as L408I and Y409A, as shown in Figure 7D. The modified dATP structure (with an azido methyl terminal group at the 3' end, a long linker, and Cy3 fluorescent group) is depicted in Figure 1B. When the modified dATP with a blocking group enters the catalytic pocket, its blocking group would collide with the surrounding L408/Y409/P410 residue side chains, as shown in Figure 7D. The mutations in motif A (L408I/Y409A/P410) of Mut_E10 have smaller side chains, thus providing more space to accommodate the incoming modified dATP with blocking group, making Mut_E10 more active for modified dATP incorporation. Interestingly, Mut_1 with motif A (L408A/Y409A/P410A) should have even more space in this area for modified dATP, however, the catalytic activity of this variant was not optimal. There may be other critical factors that should be considered affecting the catalytic efficiency of DNA polymerase besides the space in motif A, such as metal ions. For example, metal ions can affect the incorporation efficiency of DNA polymerase for dNMPs or dNTPs[65]. Further study could consider the interaction between metal ions and crucial residues. Overall, our findings suggest that when engineering DNA polymerase to incorporate a modified substrate with a larger group, using small side chain residues in the catalytic pockets is recommended.

In the third site category, A485E mutation affects the conformational change of the finger domain of KOD pol, as shown in Figure 7D. The binary complex structure of KOD pol that bound only to the DNA strand was compared with the ternary complex structure of KOD pol that bound to DNA complex and dATP, as depicted in Figure 2A. Previous reports have suggested that the movement of the finger domain may be advantageous for the fidelity of KOD DNA polymerase[66]. In Mut_E10, the A485 residue on the rotating alpha helix of the finger domain was substituted with A485E, and this resulted in the formation of a salt bridge (as illustrated in Figure 3B, top center) with Q332 on the opposite helix of the exonuclease domain,

thereby preventing the rotation of the finger domain. Evidently, this mutation had a detrimental effect on the catalytic activity of KOD pol when natural nucleotides were used. However, in the case of modified nucleotides employed in NGS, which possess long linkers and fluorescent groups, the locking of the finger domain has the potential to enhance the stability of the modified substrate access channel. This improvement can lead to an increased entry and exit rate of modified nucleotides from the catalytic pocket. The experimental outcomes indicated that this mutation significantly improved the enzyme's catalytic efficiency.

In addition, the semi-rational evolution strategy employed for enhancing the efficiency of modified nucleotide incorporation by KOD pol exhibited satisfactory efficiency compared to some reported methods[66,67]. Kennedy *et al*. conducted a steady-state kinetic analysis to examine the DNA synthesis process by DNA polymerase using natural or modified nucleotides[67]. The authors employed denaturing polyacrylamide gel electrophoresis to distinguish and assess the extension of individually modified nucleotides in an exogenous primer. In contrast, our approach utilizes fluorescence as an indicator for kinetic assays, offering advantages such as higher throughput, increased sensitivity, and a more streamlined analysis process. Nikoomanzar *et al*. developed a microfluidic-based deep mutational scanning method to evolve a replicative DNA polymerase (KOD pol) from *Thermococcus kodakarensis* for TNA synthesis[66]. They identified positive sites in the finger subdomain of KOD pol through a single round of sorting from 912 mutants and subsequently combined them step-by-step to obtain a double mutant variant. In contrast, our screening strategies achieved higher screening efficiency, resulting in Mut_E10 containing nine mutation sites (excluding 141A/143A), with fewer experimentally validated mutants (not exceeding 150). Nevertheless, our screening method using a microplate reader does not possess a comparable level of throughput as their approach, which employed a microfluidic-based droplet screening strategy.

Therefore, the screening method for future studies could consider employing microfluidic-based droplet screening combined with FRET fluorescence as an indicator, as this approach has the potential to overcome the limitations in screening throughput. For instance, we tested a small fraction of the variants from the computational library and combinatory library, and future studies could employ higher-throughput screening methods to assess more variants that may have been missed during the evolutionary process. Furthermore, further exploration of residues for continually improving catalytic efficiency of KOD pol can be conducted around the amino acid residues neighboring these key positions, such as motif B including residues 484/485/486, and residues around 389/589/676/680.

Conclusion

To conclude, this study presents a comprehensive workflow for engineering WT KOD DNA polymerase to incorporate unnatural nucleotides suitable for NGS. The workflow can be summarized into five stages: construction of the screening strategy, identification of residues in the active pocket for improvement, identification of residues located in the DNA strand binding region for improvement, step-wised combination, and characterization of variants in a specific application. The variant Mut_E10, which had eleven mutation sites, was found to enable successful compatibility between the two sequencing platforms, namely BGISEQ500 and BGISEQ2000. These beneficial mutation sites identified in this study could act as inspiration for the engineering and improvement of other archaeal B-family DNA polymerases. In addition, the successful engineering of KOD DNA polymerase with improved catalytic efficiency for unnatural nucleotides proves the great potential of the enzyme engineering strategy constructed, which offers a new solution of polymerase engineering to fulfill versatile applications.

Methods and Materials

Reagents and instruments

The primer and template labeled at the 5'-terminus by Cy5 dye were ordered from Invitrogen. The primer-template (P/T-2Cy5) was prepared by mixing the primer and template at a 1:1 molar ratio and annealing at 80 °C for 10 min and then cooling down to room temperature in a water bath. The sequence information of P/T - 2Cy5 is shown in Figure 1C. The primers designed with mutations were ordered from Genscript Biotech. 3'-O-azidomethyl-dATP-Cy3 (modified dATP) was supplied by the BGI synthetic chemistry group, and the structure is shown in Figure 1B. Potassium phosphate, $MgSO_4$, KCl, $(NH4)_2SO_4$, Tris, NaCl, HCl, EDTA, NaOH, imidazole, glycerol, lysozyme, LB broth, kanamycin, PMSF, and IPTG were purchased from Thermo Fisher. 12% SDS-PAGE gel and gel running buffer were purchased from Bio-Rad. Affinity chromatography columns (HisTrap HP 5ml column) and ion exchange columns (HiTrap Q FF 5mL and HiTrap SP HP 5ml column) were purchased from GE Healthcare. The ÄKTA Pure Protein Purification system was purchased from GE Healthcare. The microplate reader was purchased from Bio Tek.

Mutation library construction

The codon-optimized gene encoding this exonuclease-deficient KOD pol (named Mut_1) was

synthesized by Genscript Biotech and then cloned into vector pD441 with kanamycin resistance. The corresponding expressed protein features a 6xHis tag at the N-terminus. The first mutant library consisted of single-site saturation mutagenesis at four positions: 408, 409, 410, and 485. The saturation mutagenesis at positions 408, 409, 410, and 485 was performed based on the background of triple alanine substitutions at the other three positions. The site-directed mutagenesis kits were purchased from Thermo Fisher. The PCR reaction conditions were as follows: one cycle at 98 °C for 10 s; 20 cycles at 98 °C for 10 s,72 °C for 2.5 min, followed by elongation at 72 °C for 5 min and hold at 4 °C. The PCR products were visualized by 1.2% agarose gel electrophoresis and ultraviolet light. Then, 1 μL of DpnI (NEB) was added to the PCR mix and incubated for 2 h at 37 °C. 5μL of the digested PCR product was transformed into the 50 μL *E. coli* DH5α competent cells (Tiangen). Recombinant plasmids were prepared by using the QIAprep Spin Miniprep Kit (QIAGEN) and confirmed by sequencing analysis. 1 μL of each correct plasmid was transformed into 50 μL *E. coli* BL21(DE3) competent cell (Tiangen) for protein expression.

Protein expression and purification

1 mL of LB medium in a 96-deep-well plate was used for small-scale expression, and 1 L of LB medium in a conical flask was used for large-scale expression. Cells grown in LB medium containing 50 μg/mL kanamycin were induced at OD600 nm = 0.6 - 0.8 by the addition of 0.5 mM IPTG. Protein production was carried out at 25 °C, 220 rpm overnight.
Semi-purification was performed using cell pellets collected from 1 mL culture in a 96-deep-well plate. The cell pellets were resuspended in Buffer 1 (20 mM Tris-HCl, 10 mM KCl, 10 mM $(NH4)_2SO_4$, 0.1% Triton, 4 mM $MgSO_4$, 1.25 mM PMSF, and 1 mg/mL lysozyme, pH 7.6) with a ratio of 0.04 g cell pellets per mL buffer. The resuspended cells were incubated at 37 °C for 10 minutes, followed by thermal denaturation at 80 °C for 30 minutes. The supernatant containing proteins was collected after centrifugation at 12,000 X g for 20 minutes. The protein concentration was measured and estimated by using the Bradford protein assay[68].

We rigorously purified mutant proteins for use on NGS platforms and the purification methods as the following process. For protein purification, cell pellets collected from 1 L cultivation were resuspended in Buffer 2 (500 mM NaCl, 5% glycerol, 20 mM imidazole, 1.25 mM PMSF, 50 mM potassium phosphate, pH 7.4) and then disrupted with a high-pressure homogenizer AH-1500 purchased from ATS engineering company. After a 30-minute thermal denaturation at 80 °C, the centrifugation at 12,000 rpm 4 °C (Beckman Avanti J-26) for 30 mins was

performed to remove cell debris. After centrifugation, the supernatant was filtered through a 0.22 µm membrane filter. The subsequent purification process was performed using ÄKTA pure 25 (GE Healthcare) with Nickel affinity chromatography (5 mL HisTrap HP column, GE Healthcare), anion ion exchange chromatography (5 mL HiTrap Q FF column, GE Healthcare), and cation ion exchange chromatography (5 mL HiTrap SP HP column, GE Healthcare), sequentially. For the affinity purification process, the supernatant was loaded onto the pre-equilibrated column with a flow rate of 2.5 mL/min so that the retention time will be 2 min. A wash step using 50 mL Buffer 2 with a flow rate of 5mL/min was added before elution. The elution procedure was performed with a linear gradient of Buffer 3 (500 mM NaCl, 5% glycerol, 500 mM imidazole, 50 mM potassium phosphate, pH 7.4) from 0 to 100% in 10 CV with a flow rate of 5 mL/min. Fractions with 280 nm UV signal above 100 mAu were collected. The collected samples from Ni purification were diluted 6-fold with Buffer 4 (5% glycerol, 25 mM potassium phosphate, pH 6.6) to decrease the concentration of NaCl to 500 mM. The diluted sample was loaded onto the pre-equilibrated HiTrap Q FF 5 mL column (GE Healthcare) at a flow rate of 2.5 mL/min. The flowthrough sample was collected and loaded onto the pre-equilibrated HiTrap SP HP 5 mL column (GE Healthcare) with a flow rate of 2.5 mL/min. The HiTrap SP HP 5 mL column was pre-equilibrated with Buffer 5 (50 mM NaCl, 5% Glycerol, 50 mM Potassium Phosphate, pH 7.4). The elution procedure was performed with a linear gradient of Buffer 6 (1 M NaCl, 5% Glycerol, 50 mM Potassium Phosphate, pH 6.6) from 0 to 60% in 10 CV. The first peak eluted was collected. Finally, the collected sample was dialyzed with Buffer 7 (20 mM Tris, 200 mM KCl, 0.2 mM EDTA, 5% glycerol, pH 7.4) at 4 °C overnight and then stored at -20 °C with 50% glycerol. Protein concentration was determined by measuring the absorbance at 280 nm using a microplate reader (Bio Tek) and calculated using the extinction coefficient of 1.393 M^{-1} cm^{-1} predicted by the ExPASy server. The purity of the protein was analyzed using 12% SDS-PAGE.

Enzyme activity screening of KOD variants

The polymerization reaction mixture contained 0.1 µM P/T-2Cy5, 0.25 µM modified dATP, 2.5 µM native dC/G/TTP each, 2 µg BSA, 1X KOD Screening Buffer (20 mM Tris-HCl, 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 0.1% Triton, 4 mM $MgSO_4$, pH 8.8) and 0.5 µg KOD protein. The reaction was initiated by adding each semi-purified KOD mutant protein. The enzyme activity assay was performed in 384-well (black clear-bottom) plates (Corning®) at 40 °C for 1-2 hours. The FRET Cy5 fluorescence signal was measured as relative fluorescence units (RFUs) in

appropriate time intervals by exciting at 530 nm and detecting the emission at 676 nm. Measurements were performed with a gain setting at 60 and shaking for 10 seconds before measurement in a microplate reader (Bio Tek). The enzyme activity (reaction rate V) was defined as the maximum slope of the FRET Cy5 signal, expressed as RFUs/min. The data were analyzed and plotted using OriginPro.

Enzyme kinetic assays

The reaction mixture contained 10 µM native dC/G/TTP each, 2 µg BSA, 1X KOD Screening Buffer (20mM Tris-HCl, 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 0.1% Triton, 4 mM $MgSO_4$, pH 8.8) and 0.5 µg KOD protein. The kinetic assays were carried out with varying concentrations of P/T-2Cy5 or modified dATP ranging from 0 to 6 µM, while the concentration of modified dATP or P/T-2Cy5 remained constant at 2 µM. The reaction was initiated by adding KOD mutant protein. The assay was also performed in 384-well (black clear-bottom) plates (Corning®) at 40 °C for 1-2 hours. The FRET Cy5 fluorescence signal was measured by exciting at 530 nm and detecting the emission at 676 nm. Measurements were performed with a gain setting at 60 and shaking for 10 seconds before measurement in a microplate reader (Bio Tek). The maximum slope (reaction rate V) of the FRET Cy5 signal for each concentration of P/T-2Cy5 was calculated and used for determining the kinetic parameters (k_{cat} and K_m). The kinetic data were fitted non-linearly using the Michaelis-Menten equation with GraphPad Prism 5 to obtain k_{cat} and K_m.

MM-GBSA screening method

The complex structure of KOD pol and dsDNA adopts the 5OMF PDB structure. Additional small molecules in both structures were removed, and the systems were energy minimized in 80 mM sodium chloride aqueous solution, followed by NVT ensemble MD relaxation of 50 ns[69]. We used the Modeler 9.19 software to carry out saturation mutagenesis at 93 residues on the polymerase and then processed all 1860 mutants uniformly. Due to this large number of calculations, we designed an automatic processing workflow in Python. In an 80 mM sodium chloride aqueous solution, five rounds of energy minimization were first carried out for the structure generated by homologous modeling. In the five rounds of simulations, the binding coefficient to heavy atoms of protein and DNA was gradually weakened from 80 kcal / (mol · Å) to 0 kcal / (mol · Å). After that, NVT ensemble MD was carried out at 350 K, and finally, NPT ensemble MD of 10-25 ns was carried out at 310 K. We sampled the trajectory of the system

that reached equilibrium in the last part and calculated the binding energy by the MM-GBSA method[70].

Sequencing in the NGS platform

BGISEQ-500 is a desktop platform developed for DNA or RNA sequencing by BGI. It utilizes DNA NanoBalls (DNBs) technology (DNBSEQ™). The dNTPs used in DNBSEQ™ technology were modified with four different fluorescent dyes in the base and a reversible blocking group in its 3'-side. To evaluate the sequencing performance of Mut_E10 in the BGISEQ-500 platform, an *E. coli* library prepared by following the instructions described in the manufacturer's use manual of MGIEasy PCR-Free DNA Library Prep Kit (PN: 1000013452, MGI-tech) was used to make DNB following BGISEQ-500 protocol[60]. The DNBs were then loaded into the patterned nanoarrays. The SE50 sequencing kit was supplied by the BGI sequencing group. WT KOD pol, Mut_C2, and Mut_E10 were used to replace the original enzyme in the sequencing kit and performed sequencing tests in the BGISEQ-500 sequence platform employing the SE50 mode. Base calling and mapping were performed as previously described[55].

MGISEQ-2000 is a high-throughput platform developed by MGI. The sequencing performance of Mut_E10 was also tested in the DNBSEQ-2000 platform employing CoolMPS™ technology. The cold nucleotides used in CoolMPS™ technology are monophosphate nucleotides with a 3'-O-azidomethyl blocking group. CoolMPS™ PE100 High-throughput Sequencing kit (PN: 1000016935) was purchased from MGI-tech. The same E. coli library used for BGISEQ-500 was the same for MGISEQ-2000. The sequencing process followed the instructions described in the MGISEQ-2000 High-throughput Sequencing Set User Manual[57]. Mut_E10 was used to replace the original enzyme in the CoolMPS™ PE100 sequencing kit. Mut_E10 was used to sequence the flow cells loaded with DNBs prepared with the E. coli library in different runs. Base calling and mapping steps were performed as reported previously[58].

Sequence information

>Seq_1(WT KOD DNA pol)

MILDTDYITEDGKPVIRIFKKENGEFKIEYDRTFEPYFYALLKDDSAIEEVKKITAERHG
TVVTVKRVEKVQKKFLGRPVEVWKLYFTHPQDVPAIRDKIREHPAVIDIYEYDIPFAK
RYLIDKGLVPMEGDEELKMLAFAIATLYHEGEEFAEGPILMISYADEEGARVITWKN
VDLPYVDVVSTEREMIKRFLRVVKEKDPDVLITYNGDNFDFAYLKKRCEKLGINFAL
GRDGSEPKIQRMGDRFAVEVKGRIHFDLYPVIRRTINLPTYTLEAVYEAVFGQPKEKV

YAEEITTAWETGENLERVARYSMEDAKVTYELGKEFLPMEAQLSRLIGQSLWDVSRS
STGNLVEWFLLRKAYERNELAPNKPDEKELARRRQSYEGGYVKEPERGLWENIVYL
DFRSLYPSIIITHNVSPDTLNREGCKEYDVAPQVGHRFCKDFPGFIPSLLGDLLEERQKI
KKKMKATIDPIERKLLDYRQRAIKILANSYYGYYGYARARWYCKECAESVTAWGRE
YITMTIKEIEEKYGFKVIYSDTDGFFATIPGADAETVKKKAMEFLKYINAKLPGALELE
YEGFYKRGFFVTKKKYAVIDEEGKITTRGLEIVRRDWSEIAKETQARVLEALLKDGD
VEKAVRIVKEVTEKLSKYEVPPEKLVIHEQITRDLKDYKATGPHVAVAKRLAARGVK
IRPGTVISYIVLKGSGRIGDRAIPFDEFDPTKHKYDAEYYIENQVLPAVERILRAFGYRK
EDLRYQKTRQVGLSAWLKPKGT*

Acknowledgment

This study was supported by Shenzhen Engineering Laboratory Molecular Enzymology (Fa Gaishen [2018] No. 958). W.L. Thanks to the National Natural Science Foundation of China, China Grant No. 21505134. Alexander K. Buell thanks the Novo Nordisk Foundation for support (NNFSA170028392). Thanks to China National GeneBank DataBase & BGI's Sequencing Platform and MGI's CoolMPS™ platform.

Author Contributions

Conceived and designed the experiments: Lili Zhai, Yue Zheng. Performed the experiments: Lili Zhai. Performed the simulation: Zi Wang. Performed the application test: Liu Fen, Jingjing Wang, Hongyan Han. Analyzed the data: Lili Zhai. Contributed reagents/materials/analysis tools: Qingqing Xie, Yue Zheng, Wenwei Zhang, Yuliang Dong. Wrote the paper and SI: Lili Zhai, Zi Wang. Reviewed and revised the paper: Lili Zhai, Yue Zheng, Qingqing Xie, Chongjun Xu, and Alexander Kai Buell.

Supplementary Information

Experimental validation

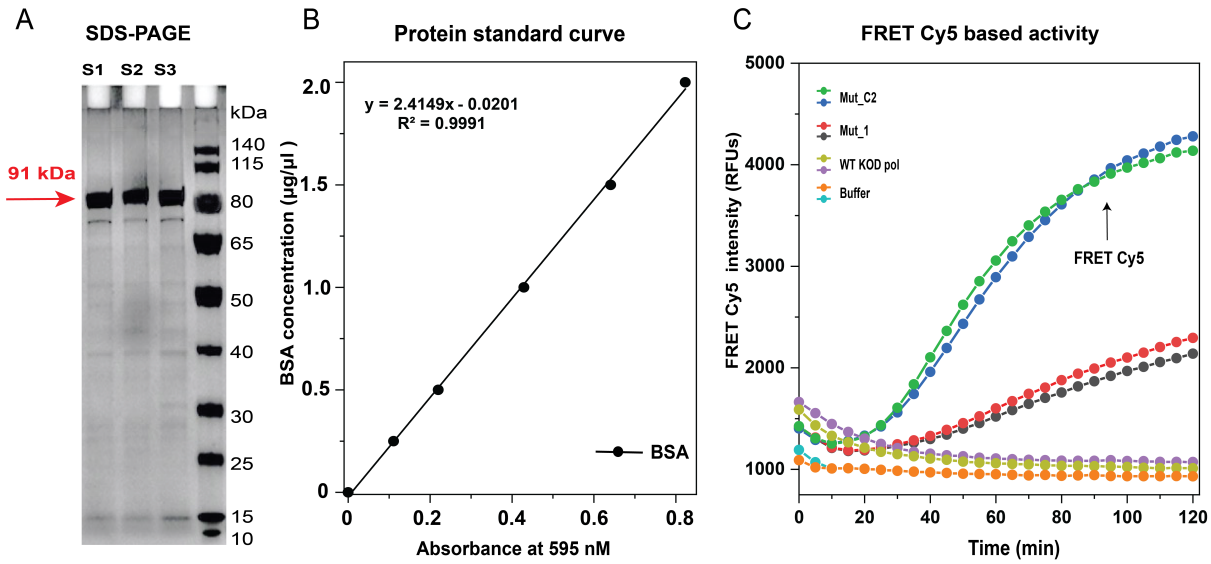
We conducted a validation of our screening method, using the identified mutants as an exemplar. In order to validate our screening method, we utilized wild-type KOD pol and two mutants (Mut_1 and Mut_C2) with different enzymatic activity from the first-round screening as examples selected from the initial screening due to their distinct catalytic efficiencies toward the modified dATP. In this approach, we performed a quantitative assessment of the enzyme activity of the KOD variants, which was crucial for evaluating their catalytic performance.

After semi-purification, mutant proteins were analyzed by SDS-PAGE gel electrophoresis, and the estimated protein purity was approximately 80% based on ImageJ analysis of the gels, as shown in Supplementary Figure 1A. In order to quantify the mutant protein, we established a standard curve using bovine serum albumin (BSA) as a standard and measured the absorbance at 595 nm through the Bradford assay[38,39]. The BSA concentration range was between 0 to 2.0 μg/μL, as shown in Supplementary Figure 1B. Through linear regression analysis of the standard curve, we were able to consistently quantify the KOD mutant protein. After calculating the total protein concentration (Conc. Total) using the Bradford method, the final target protein concentration was determined by multiplying the Conc. Total by 80%. This approach enabled us to determine the protein concentration of KOD mutants obtained in the previous step and ensure their purity for subsequent analysis and characterization.

We calculate the maximum slope of the increased fluorescence signal (FRET Cy5) as a surrogate for the enzyme activity of KOD mutants. To assess the FRET Cy5-based enzymatic activity of the KOD mutants, we added 0.5 μg of each mutant protein to the reaction system, with the reaction buffer serving as the negative control (no enzyme added). As shown in Supplementary Figure 1C, a steady starting was observed before the rise of the FRET Cy5 fluorescence signal, which could be attributed to a temperature increase during the reaction. WT KOD DNA polymerase failed to generate a FRET Cy5 fluorescence signal, indicating WT KOD pol cannot catalyze modified dATP into the DNA strand. In contrast, Mut_C2 exhibited a higher catalytic efficiency compared to Mut_1, generating more FRET Cy5 fluorescence signals within the same time frame. This observation suggests that Mut_C2 is capable of efficiently incorporating modified dATP into the DNA strand, making it a promising candidate for further characterization and development. Overall, our experimental screening approach was effective in distinguishing different levels of enzyme activities of the mutants.

In addition, fluorescently labeled substrates used in our experiments were at risk of fluorescence quenching during storage and usage, which raises concerns about the reliability and

reproducibility of data obtained from experiments utilizing fluorescent labeling, particularly in situations where extended storage or multiple freeze-thaw cycles are involved. To get rid of the influence of fluorescence quenching, we preferred using the direct readout of the fluorescent signal values in each batch test, not the absolute substrate concentrations. Positive control was employed in each experiment as a reference for all the other tested mutants. By using this compromise solution, the efficiencies of mutants tested not in the parallel experiments can be compared. During every round of the evolution process, the parent mutant was utilized as a positive control.



Supplementary Figure 1. Experimental validation of the reliability of the screening methods. A） This figure displays the SDS-PAGE analysis (12% gel) with WT KOD pol represented by S1, KOD variant Mut_1 represented by S3, and KOD variant Mut_C2 represented by S2. The volume of samples loaded was around 1 μg, and the purification method involved lysing cells with lysozyme at 37 °C for 10 min and centrifugation after heating at 80°C for 30 min. The gel was run at 120V voltage and 1-2 h running time at room temperature. B) The protein standard curve was generated using bovine serum albumin (BSA) as the protein standard. Absorbance was measured at 595nm using a microplate reader (BioTek) at room temperature after the mixture 15 minutes of incubation in the dark. C) FRET Cy5-based enzyme activities of wild-type KOD pol (olive and purple) and KOD variants Mut_1 (red and black), Mut_C2 (green and blue), and buffer (orange and cyan) are presented. In these experiments, 0.5 μg protein was used for measurements. The reactions were performed at 40°C for 2 hours with excitation at 530 nm and detection at 676 nm using a microplate reader (BioTek). The data were directly output using the Gene5 software of the microplate reader. The curve was plotted using OriginPro. Each experiment was performed by duplicates under the same conditions.

Computational simulation

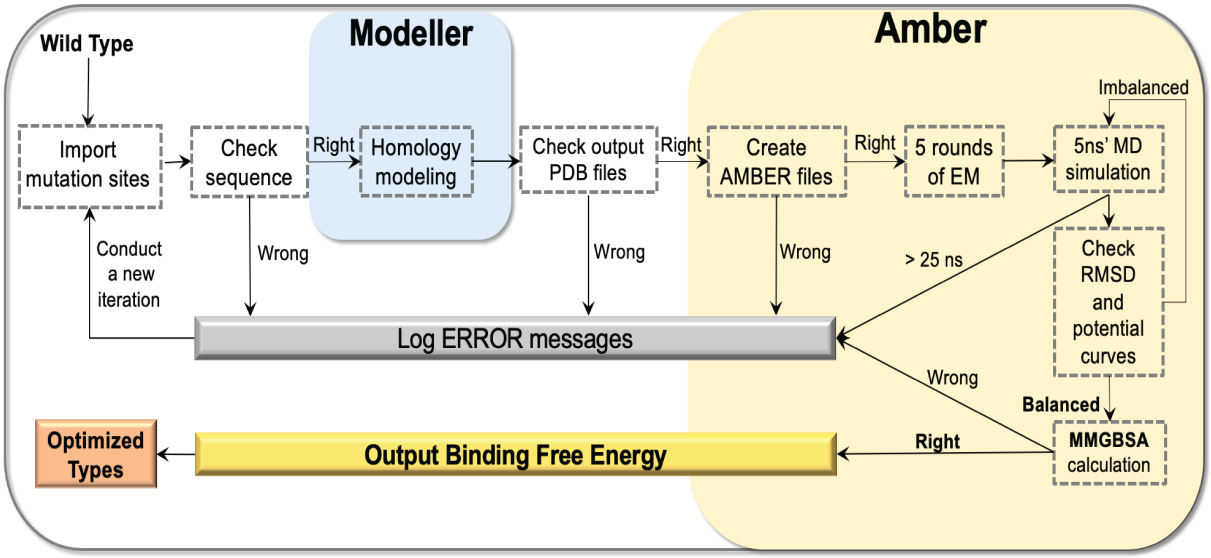
We then calculated the changes in binding free energy between the mutants and dsDNA using the MM-GBSA method. The automation process of the MM-GBSA calculation and MD simulation is shown in Supplementary Figure 2 and Supplementary Figure 3. Due to the high standard deviation of MM-GBSA calculation, during data analysis, we first calculate the average value of the calculation results of 20 mutants at a single site and then select some mutants for further experimental testing to determine their lower average value. We expressed and tested several enzyme mutants with an average binding energy of < 260 kcal/mol and some interesting positions. These successfully constructed and expressed mutants were subjected to measurement of enzymatic activity and polymerization kinetics towards the DNA primer-template, and the results are shown in Supplementary Table 2.

Supplementary Table 1. Average binding free energy (Kcal/mol) of 93 mutation sites in DNA-binding region.

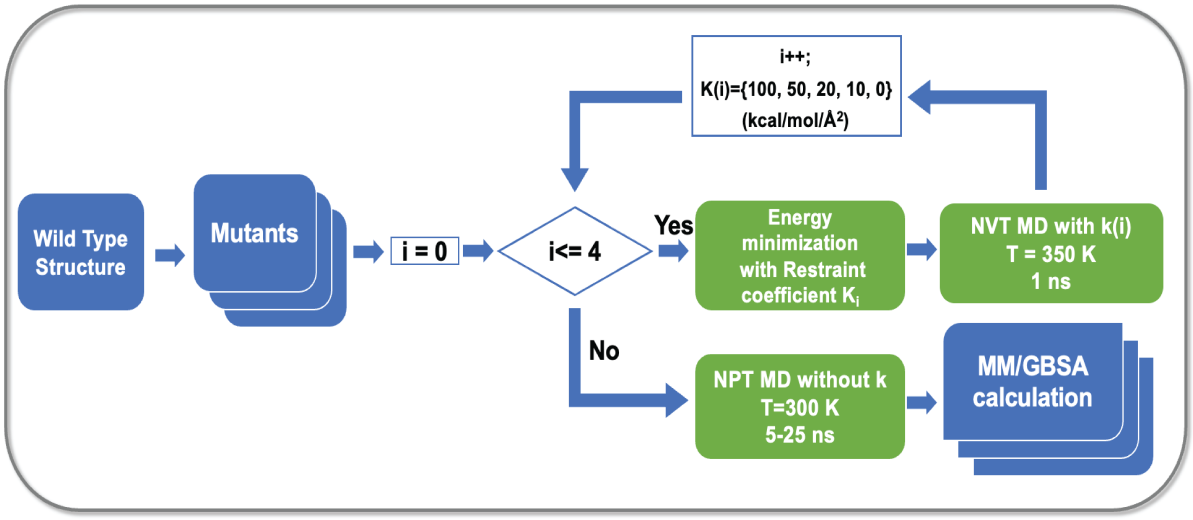
Mutant	Avg. [1] (Kcal/mol)	Mutant	Avg. [1] (Kcal/mol)	Mutant	Avg. [1] (Kcal/mol)	Mutant	Avg. [1] (Kcal/mol)
R265	-236.3	P392	-228.8	W504	-264	G601	-256.3
N269	-231.2	E393	-259.5	Y505	-260	K602	-255.1
L270	-250.9	R394	-258.9	C506	-249.6	I603	-254.6
P271	-220.3	G395	-263.2	K507	-230.4	T604	-242.5
S347	-272.9	Y402	-260.1	A510	-251.3	T605	-270.8
S348	-252.3	L403	-241.4	E511	-243.1	R606	-242.0
T349	-235.1	D404	-258.8	T514	-232.7	G607	-259.8
G350	-260.8	K487	-256.3	R518	-246.8	L608	-256.2
N351	-271.6	I488	-231.9	Y538	-261.6	E609	-273.1
L352	-251.9	L489	-244.9	S539	-266.1	I610	-272.2
V353	-246.3	A490	-277.2	D540	-260.9	V611	-243.5
R379	-278.3	N491	-270.2	T541	-257.5	R612	-253.6
R380	-252.8	S492	-239.6	D542	-268.1	R613	-267.4
R381	-236.6	Y493	-256.4	G543	-258.6	D614	-263.6
Q382	-238.3	Y494	-256.5	F544	-257.5	W615	-259.9
T383	-255.7	G495	-254.6	F545	-268.4	S616	-278.8
F384	-228.5	Y496	-261.7	A546	-245.9	V661	-243.1

E385	-228.5	Y497	-252.4	H589	-265.7	I662	-249.1
G386	-271.4	G498	-259.0	T590	-267.4	H663	-256.2
G387	-241.3	Y499	-254.4	K591	-243.2	K676	-256.2
Y388	-260.7	A500	-250.2	K592	-245.4	M680	-253.1
I389	-238.3	R501	-249.0	K593	-255.3		
K390	-250.8	A502	-248.4	Y594	-256.0		
E391	-258	R503	-253.2	A595	-253.1		

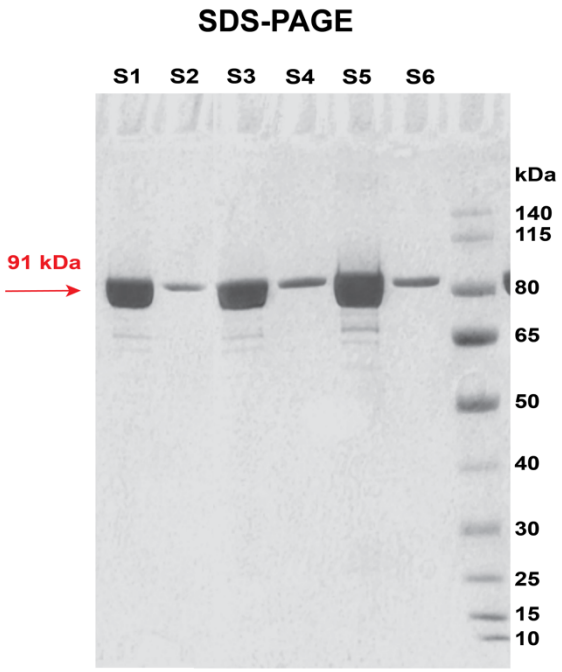
(1): The average binding free energy (Kcal/mol) for each site is calculated. The binding energy of each variant was measured by the MM- GBSA screening (details in the **Materials and Methods** section).



Supplementary Figure 2. Automation process of MM-GBSA calculation. EM signifies Energy Minimization; MD signifies Molecular Dynamics; RMSD signifies the Root-Mean-Square Deviation.



Supplementary Figure 3. Mutants' pretreatment and MD simulation process. i was the cycle number of energy minimization; k_i was the restraint coefficient; T was temperature.



Supplementary Figure 4. The SDS-PAGE analysis (12% gel) for rigorously purified proteins with WT KOD pol represented by S1 (5 µg) and S2 (0.25 µg), KOD variant Mut_C2 represented by S2(5 µg) and S4 (0.25 µg), and KOD variant Mut_E10 represented by S5 (5 µg) and S6 (0.25 µg). We employed a three-step purification process involving Ni affinity chromatography and anion-cation exchange chromatography using ÄKTA. The estimated protein purity was approximately 95% based on ImageJ analysis of the gels.

Supplementary Table 2. Kinetic data towards the P/T-2Cy5 of combined mutants.

Mutants	Sites	Subdomain	V_{rel} (1)	E(Mut)/P/T-2Cy5 (2)
Mut_C2	-	-	1.0	1.0
Mut_D34	V589H	Palm	2.6 ± 0.21	5.2 ± 0.41
Mut_D35	V589Q	Palm	1.8	3.0
Mut_D51	V680M	Thumb	0.5 ± 0.07	2.0 ± 0.31
Mut_D48	T676K	Thumb	1.2 ± 0.07	1.6 ± 0.54
Mut_D30	T590K	Palm	1.1 ± 0.35	1.1
Mut_D33	V389M	Palm	1.1 ± 0.21	1.1 ± 0.36
Mut_D36	Y384F	Palm	0.8± 0.14	1.1
Mut_D40	S383T	Palm	1.1	1.0
Mut_D17	T349I	N-Ter	1.1 ± 0.07	1.0 ± 0.17
Mut_D3	Y496I	Finger	0.7 ± 0.21	1.0
Mut_D32	V389I	Palm	0.9 ± 0.07	1.0 ± 0.33
Mut_D6	I488Q	Finger	0.7 ± 0.11	0.9
Mut_D7	L489Y	Finger	0.9 ± 0.14	0.7
Mut_D20	E385M	Palm	0.9 ± 0.14	0.7 ± 0.19
Mut_D9	I488L	Finger	1	0.7
Mut_D15	S348T	N-Ter	0.8± 0.14	0.7 ± 0.17
Mut_D12	S347I	N-Ter	1.1 ± 0.21	0.6
Mut_D13	S347M	N-Ter	0.7 ± 0.21	0.6
Mut_D42	E609Y	Thumb	0.1	0.6 ±
Mut_D8	L489K	Finger	0.8	0.5 ±
Mut_D31	T590L	Palm	0.6 ± 0.14	0.5
Mut_D37	Y384W	Palm	0.7± 0.14	0.5
Mut_D14	S348M	N-Ter	0.9 ± 0.21	0.5
Mut_D50	V680D	Thumb	0.2	0.5
Mut_D41	S383I	Palm	0.7	0.5 ± 0.21
Mut_D19	A500G	Palm	1.4 ± 0.14	0.5
Mut_D10	N351H	N-Ter	1.4 ± 0.21	0.55 ± 0.04
Mut_D27	R501M	Palm	0.6	0.4
Mut_D39	Y594V	Palm	0.2	0.4
Mut_D5	Y499L	Finger	0.7 ± 0.07	0.4
Mut_D21	E385W	Palm	1.1 ± 0.14	0.4

Mut_D18	A500D	Palm	0.9 ± 0.07	0.4 ± 0.23
Mut_D26	R501L	Palm	0.4	0.4
Mut_D11	N351Y	N-Ter	1.2 ± 0.07	0.4
Mut_D23	K592Y	Palm	0.6	0.3
Mut_D1	S492E	Finger	0.2	0.2
Mut_D43	E609G	Thumb	0.1	0.2
Mut_D29	T541W	Palm	0.1	0.1
Mut_D38	Y594I	Palm	0.2	0
Mut_D2	S492Y	Finger	0.2	NR
Mut_D4	Y496L	Finger	0.4	NR
Mut_D16	T349F	N-Ter	0.4	NR
Mut_D22	K592L	Palm	0.3	NR
Mut_D24	K593F	Palm	0.1	NR
Mut_D25	K593W	Palm	0.2	NR
Mut_D28	T541I	Palm	0.2	NR
Mut_D44	R606H	Thumb	0.4	NR
Mut_D45	R606N	Thumb	0.3	NR
Mut_D46	R613G	Thumb	0.3	NR
Mut_D47	R613T	Thumb	0.4	NR
Mut_D49	T676Y	Thumb	0.1	NR
Mut_D52	Y499G	Finger	0.2	NR

(1): The relative enzyme activity (V_{rel}) of each KOD variant was calculated with equation 2 (eq. 2). Enzyme activity was measured by enzyme activity screening method (details in **Materials and Methods** section).
(2): The relative kinetic performance (E(Mut)/ P/T-2Cy5) of each KOD variant was calculated with equation 3 (eq. 3). The kinetic was measured by enzyme kinetic assay (details in **Materials and Methods** section).
The data presented represent the mean and standard deviation values obtained from duplicated independent experiments. 0: No increased fluorescence signal is observed. NR: Measurements are not performed and reported.

Reference

1. Zhang, L., Kang, M., Xu, J. & Huang, Y. Archaeal DNA polymerases in biotechnology. *Appl. Microbiol. Biotechnol.* **99**, 6585–6597 (2015).
2. Terpe, K. Overview of thermostable DNA polymerases for classical PCR applications: from molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* **97**, 10243–10254 (2013).
3. Pinheiro, V. B. Engineering-driven biological insights into DNA polymerase mechanism. *Curr. Opin. Biotechnol.* **60**, 9–16 (2019).

4. Aschenbrenner, J. & Marx, A. DNA polymerases and biotechnological applications. *Curr. Opin. Biotechnol.* **48**, 187–195 (2017).
5. Wang, Y., Shi, Y., Hellinga, H. W. & Beese, L. S. Thermally controlled intein splicing of engineered DNA polymerases provides a robust and generalizable solution for accurate and sensitive molecular diagnostics. *Nucleic Acids Res.* gkad368 (2023) doi:10.1093/nar/gkad368.
6. Leconte, A. M. *et al.* Directed Evolution of DNA Polymerases for Next-Generation Sequencing. *Angew. Chem.* **122**, 6057–6060 (2010).
7. Chen, C.-Y. DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. *Front. Microbiol.* **5**, (2014).
8. Huber, C., von Watzdorf, J. & Marx, A. 5-methylcytosine-sensitive variants of *Thermococcus kodakaraensis* DNA polymerase. *Nucleic Acids Res.* gkw812 (2016) doi:10.1093/nar/gkw812.
9. Guo, J. *et al.* Four-color DNA sequencing with 3′- *O* -modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc. Natl. Acad. Sci.* **105**, 9145–9150 (2008).
10. Gardner, A. F. *et al.* Therminator DNA Polymerase: Modified Nucleotides and Unnatural Substrates. *Front. Mol. Biosci.* **6**, 28 (2019).
11. Rashid, N. & Aslam, M. An overview of 25 years of research on Thermococcus kodakarensis, a genetically versatile model organism for archaeal research. *Folia Microbiol. (Praha)* **65**, 67–78 (2020).
12. Wynne, S. A., Pinheiro, V. B., Holliger, P. & Leslie, A. G. W. Structures of an Apo and a Binary Complex of an Evolved Archeal B Family DNA Polymerase Capable of Synthesising Highly Cy-Dye Labelled DNA. *PLoS ONE* **8**, e70892 (2013).
13. Kropp, H. M., Diederichs, K. & Marx, A. The Structure of an Archaeal B-Family DNA Polymerase in Complex with a Chemically Modified Nucleotide. *Angew. Chem. Int. Ed.* **58**, 5457–5461 (2019).
14. Eriksen, D. T., Lian, J. & Zhao, H. Protein design for pathway engineering. *J. Struct. Biol.* **185**, 234–242 (2014).
15. Tobin, M. B., Gustafsson, C. & Huisman, G. W. Directed evolution: the ‘rational’ basis for ‘irrational’ design. *Curr. Opin. Struct. Biol.* **10**, 421–427 (2000).
16. Siegel, J. B. *et al.* Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **329**, 309–313 (2010).
17. Ge, Q., Jing, Z., Ping, Z., Jibin, S. & Zhoutong, S. Recent advances in directed evolution. 9 (2018).
18. Coco, W. M. *et al.* DNA shuffling method for generating highly recombined genes and evolved enzymes. *Nat. Biotechnol.* **19**, 354–359 (2001).
19. DNA shuffling method for generating highly recombined genes and evolved enzymes.pdf.
20. Tubeleviciute, A. & Skirgaila, R. Compartmentalized self-replication (CSR) selection of Thermococcus litoralis Sh1B DNA polymerase for diminished uracil binding. *Protein Eng. Des. Sel.* **23**, 589–597 (2010).
21. Nikoomanzar, A., Vallejo, D. & Chaput, J. C. Elucidating the Determinants of Polymerase Specificity by Microfluidic-Based Deep Mutational Scanning. *ACS Synth. Biol.* **8**, 1421–1429 (2019).
22. Zeng, W., Guo, L., Xu, S., Chen, J. & Zhou, J. High-Throughput Screening Technology in Industrial Biotechnology. *Trends Biotechnol.* **38**, 888–906 (2020).
23. Zeng et al. - 2020 - High-Throughput Screening Technology in Industrial.pdf.
24. Zhang, C. & Johnson, L. W. Microfluidic Control of Fluorescence Resonance Energy Transfer: Breaking

the FRET Limit*. *Angew Chem* 4 (2007).
25. Wu, P. G. & Brand, L. Resonance Energy Transfer: Methods and Applications. *Anal. Biochem.* **218**, 1–13 (1994).
26. Chan, F. K.-M. *et al.* Fluorescence resonance energy transfer analysis of cell surface receptor interactions and signaling using spectral variants of the green fluorescent protein. *Cytometry* **44**, 361–368 (2001).
27. Stryer, L., Thomas, D. D. & Meares, C. F. Diffusion-Enhanced Fluorescence Energy Transfer. *Annu. Rev. Biophys. Bioeng.* **11**, 203–222 (1982).
28. Chen, Y. *et al.* Screening strategy of TMPRSS2 inhibitors by FRET-based enzymatic activity for TMPRSS2-based cancer and COVID-19 treatment. *Am. J. Cancer Res.* **11**, 827–836 (2021).
29. Kuwahara, M. *et al.* Study on Suitability of KOD DNA Polymerase for Enzymatic Production of Artificial Nucleic Acids Using Base/Sugar Modified Nucleoside Triphosphates. *Molecules* **15**, 8229–8240 (2010).
30. Atomi, H., Fukui, T., Kanai, T., Morikawa, M. & Imanaka, T. Description of *Thermococcus kodakaraensis* sp. nov., a well studied hyperthermophilic archaeon previously reported as *Pyrococcus* sp. KOD1. *Archaea* **1**, 263–267 (2004).
31. Sawai, H. *et al.* Expansion of structural and functional diversities of DNA using new 5-substituted deoxyuridine derivatives by PCR with superthermophilic KOD Dash DNA polymeraseElectronic supplementary information (ESI) available: sequencing of the PCR products (108mer DNA) from substrates 1 and 2. See http://www.rsc.org/suppdata/cc/b1/b107838k/. *Chem. Commun.* 2604–2605 (2001) doi:10.1039/b107838k.
32. Horhota, A. *et al.* Kinetic Analysis of an Efficient DNA-Dependent TNA Polymerase. *J. Am. Chem. Soc.* **127**, 7427–7434 (2005).
33. Hottin, A. & Marx, A. Structural Insights into the Processing of Nucleobase-Modified Nucleotides by DNA Polymerases. *Acc. Chem. Res.* **49**, 418–427 (2016).
34. Ghosh, P. *et al.* Microenvironment-Sensitive Fluorescent Nucleotide Probes from Benzofuran, Benzothiophene, and Selenophene as Substrates for DNA Polymerases. *J. Am. Chem. Soc.* **144**, 10556–10569 (2022).
35. Aschenbrenner, J. & Marx, A. DNA polymerases and biotechnological applications. *Curr. Opin. Biotechnol.* **48**, 187–195 (2017).
36. Hottin, A. & Marx, A. Structural Insights into the Processing of Nucleobase-Modified Nucleotides by DNA Polymerases. *Acc. Chem. Res.* **49**, 418–427 (2016).
37. Hoshino, H. *et al.* Consecutive incorporation of functionalized nucleotides with amphiphilic side chains by novel KOD polymerase mutant. *Bioorg. Med. Chem. Lett.* **26**, 530–533 (2016).
38. Sabat, N. *et al.* Towards the controlled enzymatic synthesis of LNA containing oligonucleotides. *Front. Chem.* **11**, 1161462 (2023).
39. Bergen, K., Betz, K., Welte, W., Diederichs, K. & Marx, A. Structures of KOD and 9°N DNA Polymerases Complexed with Primer Template Duplex. *ChemBioChem* **14**, 1058–1062 (2013).
40. Kennedy, E. M., Hergott, C., Dewhurst, S. & Kim, B. The Mechanistic Architecture of Thermostable *Pyrococcus furiosus* Family B DNA Polymerase Motif A and Its Interaction with the dNTP Substrate.

Biochemistry **48**, 11161–11168 (2009).
41. Pinheiro, V. B. Engineering-driven biological insights into DNA polymerase mechanism. *Curr. Opin. Biotechnol.* **60**, 9–16 (2019).
42. Hashimoto, H. *et al.* Crystal structure of DNA polymerase from hyperthermophilic archaeon Pyrococcus kodakaraensis KOD111Edited by R. Huber. *J. Mol. Biol.* **306**, 469–477 (2001).
43. Fujiwara, S., Takagi, M. & Imanaka, T. Archaeon Pyrococcus kodakaraensis KOD1: application and evolution. in *Biotechnology Annual Review* vol. 4 259–284 (Elsevier, 1998).
44. Kropp, H. M., Diederichs, K. & Marx, A. The Structure of an Archaeal B-Family DNA Polymerase in Complex with a Chemically Modified Nucleotide. *Angew. Chem. Int. Ed.* **58**, 5457–5461 (2019).
45. Hoshino, H. *et al.* Consecutive incorporation of functionalized nucleotides with amphiphilic side chains by novel KOD polymerase mutant. *Bioorg. Med. Chem. Lett.* **26**, 530–533 (2016).
46. Moreira, I. S., Fernandes, P. A. & Ramos, M. J. Computational alanine scanning mutagenesis—An improved methodological approach. *J. Comput. Chem.* **28**, 644–654 (2007).
47. A site-directed mutagenesis method particularly useful for creating otherwise difficult-to-make mutants and alanine scanning | Elsevier Enhanced Reader. https://reader.elsevier.com/reader/sd/pii/S0003269711006270?token=97A2E4FAD118E49845706173377D968EA9EC324A72DE205AC4FC81C27225030A8C57E8D1A328DB0E3AB8943940AA2ED0&originRegion=eu-west-1&originCreation=20220413125914 doi:10.1016/j.ab.2011.09.019.
48. Brown, J. A. & Suo, Z. Unlocking the Sugar "Steric Gate" of DNA Polymerases. *Biochemistry* **50**, 1135–1142 (2011).
49. Gardner, A. Determinants of nucleotide sugar recognition in an archaeon DNA polymerase. *Nucleic Acids Res.* **27**, 2545–2553 (1999).
50. Gardner, A. F. *et al.* Therminator DNA Polymerase: Modified Nucleotides and Unnatural Substrates. *Front. Mol. Biosci.* **6**, 28 (2019).
51. O'Flaherty, D. K. & Guengerich, F. P. Steady-State Kinetic Analysis of DNA Polymerase Single-Nucleotide Incorporation Products. *Curr. Protoc. Nucleic Acid Chem.* **59**, (2014).
52. Raper, A. T., Reed, A. J. & Suo, Z. Kinetic Mechanism of DNA Polymerases: Contributions of Conformational Dynamics and a Third Divalent Metal Ion. *Chem. Rev.* **118**, 6000–6025 (2018).
53. Blasco, M. A., Méndez, J., Lázaro, J. M., Blanco, L. & Salas, M. Primer Terminus Stabilization at the φ29 DNA Polymerase Active Site. *J. Biol. Chem.* **270**, 2735–2740 (1995).
54. Leconte, A. M. *et al.* Directed Evolution of DNA Polymerases for Next-Generation Sequencing. *Angew. Chem.* **122**, 6057–6060 (2010).
55. Huang, J. *et al.* A reference human genome dataset of the BGISEQ-500 platform. *GigaScience* **6**, (2017).
56. Lang, J. *et al.* Evaluation of the MGISEQ-2000 Sequencing Platform for Illumina Target Capture Sequencing Libraries. *Front. Genet.* **12**, 730519 (2021).
57. Drmanac, S. *et al. CoolMPS ™ : Advanced massively parallel sequencing using antibodies specific to each natural nucleobase*. http://biorxiv.org/lookup/doi/10.1101/2020.02.19.953307 (2020) doi:10.1101/2020.02.19.953307.
58. Korostin, D. *et al.* Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing. *PLOS ONE* **15**, e0230301 (2020).

59. Tiacci, L. Buffer allocation vs. sequencing optimization: which of the two is most effective to improve the efficiency of assembly lines? *IFAC-Pap.* **55**, 452–457 (2022).
60. Huang, J. *et al. BGISEQ-500 Sequencing v1*. https://www.protocols.io/view/bgiseq-500-sequencing-pq7dmzn (2018) doi:10.17504/protocols.io.pq7dmzn.
61. Zavgorodny, S. *et al.* 1-alkylthioalkylation of nucleoside hydroxyl functions and its synthetic applications: a new versatile method in nucleoside chemistry. *Tetrahedron Lett.* **32**, 7593–7596 (1991).
62. Drmanac, R. et al. Stepwise sequencing by non-labeled reversible terminators or natural nucleotides. (Google Patents, 2018).
63. Yudkina, A. V. & Zharkov, D. O. Miscoding and DNA Polymerase Stalling by Methoxyamine-Adducted Abasic Sites. *Chem. Res. Toxicol.* **35**, 303–314 (2022).
64. Endutkin, A. V., Yudkina, A. V., Zharkov, T. D., Kim, D. V. & Zharkov, D. O. Recognition of a Clickable Abasic Site Analog by DNA Polymerases and DNA Repair Enzymes. *Int. J. Mol. Sci.* **23**, 13353 (2022).
65. Tabor, S. & Richardson, C. C. Effect of manganese ions on the incorporation of dideoxynucleotides by bacteriophage T7 DNA polymerase and Escherichia coli DNA polymerase I. *Proc. Natl. Acad. Sci.* **86**, 4076–4080 (1989).
66. Nikoomanzar, A., Vallejo, D. & Chaput, J. C. Elucidating the Determinants of Polymerase Specificity by Microfluidic-Based Deep Mutational Scanning. *ACS Synth. Biol.* **8**, 1421–1429 (2019).
67. Kennedy, E. M., Hergott, C., Dewhurst, S. & Kim, B. The Mechanistic Architecture of Thermostable *Pyrococcus furiosus* Family B DNA Polymerase Motif A and Its Interaction with the dNTP Substrate. *Biochemistry* **48**, 11161–11168 (2009).
68. Kielkopf, C. L., Bauer, W. & Urbatsch, I. L. Bradford Assay for Determining Protein Concentration. *Cold Spring Harb. Protoc.* **2020**, pdb.prot102269 (2020).
69. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **10**, 449–461 (2015).
70. Poli, G., Granchi, C., Rizzolio, F. & Tuccinardi, T. Application of MM-PBSA Methods in Virtual Screening. *Molecules* **25**, 1971 (2020).

2.3 Part II Manuscript 2 - Rational evolution of a recombinant DNA polymerase for efficient incorporation of unnatural nucleotides by dual site boosting

This manuscript is a continuation of the studies presented in Manuscript 1, with a primary focus on further enhancing the catalytic efficiency of the KOD variant Mut_E10 (renamed as RF) in incorporating modified nucleotides and developing rational design methods to reduce experimental costs. Patent protection has been applied for Mut_E10; hence it was renamed RF (reference). In this study, we established a machine-learning model to predict specific site mutations in the thumb and finger domains of KOD DNA polymerase. We performed virtual screening on a designed virtual library containing 200 variants with double-point mutations and subsequently selected variants that ranked in the top 22% for experimental validation. The experimental results demonstrated that over 80% of these enzyme variants achieved higher catalytic efficiency than the parent RF, indicating that our rational design method is more efficient than traditional mutagenesis methods. Finally, we performed an analysis of the functional dynamics and potential mechanisms responsible for enhancing the catalytic efficiency of the optimal KOD variant. Subsequently, we applied for a patent (WO2022247055A1) to protect the effective mutation sites discovered in this manuscript. This manuscript has been prepared for publication in the near future.

Rational evolution of a recombinant DNA polymerase for efficient incorporation of unnatural nucleotides by dual-site boosting

Contributions: Ruyin Cao developed the machine learning methodology, performed the mutants' simulation, and analyzed the data. Lili Zhai developed the experimental methodology, performed the experiments for KOD DNA polymerase's library construction, mutant screening, and purification on the bench, and analyzed the data.

Ruyin Cao[#, *,1], Lili Zhai[*,1], Qingqing Xie[1], Zi Wang[1], Yue Zheng[1], Wenwei Zhang[1], Alexander K. Buell[2], Xun Xu[1], Yuliang Dong[#,1], Chongjun Xu[#,3], Wenping Lu[#,4]

1. BGI-Shenzhen, Beishan Industrial Zone, Shenzhen 518083, China

2. Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs. Lyngby 2800,

Denmark

3. Complete Genomics Inc., 2904 Orchard Pkwy, San Jose, CA, 95134, USA

4. Department of Physics, RWTH Aachen University, Aachen 52062, Germany

∗ These authors contributed equally to this work

Corresponding author

Abstract

Machine learning-based modeling is increasingly used to assist in function-oriented enzyme engineering, which typically relies on a predefined sequence space. However, defining the determinant amino acid positions for constructing a combinatorial mutational library remains a challenge in protein engineering. In this work, we employed dual-site design and the fine-tuning of functional dynamics to engineer and analyze a recombinant KOD DNA polymerase, leading to significant improvement in the efficient incorporation of 3'-O-azidomethyl-dATP labeled with the fluorescent dye Cy3. This includes identifying key sites/regions outside of the known active center, machine learning-assisted mutant screening, and understanding the underlying mechanism of kinetics boosting. By utilizing hundreds of training data points and analyzing dozens of machine-learning predicted variants, we have discovered that the catalytic efficiency of the highly engineered KOD variant can be further enhanced by at least one order of magnitude through specific mutations at two sites on the finger subdomain: A485I and S451L. Contrary to position 485, which is known to dominate the local conformation of B-family DNA polymerases, position 451 is a new active site discovered in our approach. The synergy of A485I and S451L on kinetics boosting is driven by an enhanced dynamical-cooperative regulation mechanism of the state distribution of 'open/closed' conformations and the dynamics of the nucleotide of template DNA for substrate pairing. These findings contribute to a deeper understanding of functional dynamics engineering in KOD pol and have implications for the design of other B-family DNA polymerases for the efficient incorporation of modified nucleotides.

Key Words

Machine learning, Molecular dynamics simulation, Next generation sequencing, DNA Polymerase engineering, Rational evolution, Dual-site mutation, Regulation mechanism

Introduction

Machine learning (ML) models have been proposed to assist traditional protein engineering protocols, in efficiently identifying improved protein variants by reducing the experimental burden of testing many variants[1]. However, ML approaches alone do not offer any mechanistic insights as to how amino acid (a.a.) sequence positions of an enzyme individually contribute to its comprehensive function[2-4]. Defining the determinant a.a. positions upon which the combinatorial space for a specific functional goal is constructed is still a challenge in protein science[3,4]. Practically, this aspect is also the key to minimizing both the time and economic cost of training data collection. A common strategy is to consider potential active sites, such as conserved a.a. positions or those involved in the binding pocket/interface[5]. These approaches might be beneficial in the case of proteins targeting small ligands and those with evolutionarily conserved functional residues[6]. A common strategy is to consider potential active sites, such as conserved a.a. positions or those involved in the binding pocket/interface. These approaches might be beneficial in the case of proteins targeting small ligands and those with evolutionarily conserved functional residues[7,8].

An important example of this challenging category is the B Family of DNA polymerase, the most widespread polymerase in all domains of life exhibiting multiple functional states during nucleotide ligation[9,10,11]. In modern biotech and biomedical industries, such as DNA sequencing and next-generation therapeutics, the function-oriented modifications of B-family DNA polymerases are highly sought after[12,13]. A B-family DNA polymerase (KOD) from *Thermococcus kodakaraensis*, has been highly engineered to incorporate dye-labeled reversible terminators in our previous studies. In our previous work, a highly engineered KOD variant carrying eleven mutation sites demonstrated satisfactory performance in next-generation sequencing platforms such as BGI Genomics[14] or MGI Tech[15]. Multiple residue combinations discovered in our previous studies demonstrate synergistic effects that contribute to improved catalytic efficiency for the incorporation of dye-labeled reversible terminators. However, utilizing traditional mutagenesis methods[16], such as random mutagenesis[17] and site-direct mutagenesis[16], to identify effective synergy between sites or the underlying mechanisms in an evolutionary landscape would be insufficient and challenging[18,19]. Therefore, the utilization of machine learning in protein engineering for mutation site prediction and structure prediction has garnered increasing attention from researchers, as it holds promise for enhancing protein functionality and facilitating mechanistic studies[20]. Machine learning-assisted screening methods can analyze large datasets and identify patterns within protein sequences and structures,

enabling the prediction of optimal combinations of sites for generating protein variants with desired properties[20].

In this study, we aimed to enhance the catalytic performance of a highly engineered KOD variant (obtained in our previous study) for incorporating dye-labeled reversible terminators. In order to achieve this, a medium-sized dataset was collected to construct our machine learning (ML) models. A virtual library was designed and screened using the machine learning models, followed by the experimental screening of the top 22% of variants. The best predicted KOD variant exhibited catalytic performance improvements of up to 18-fold relative to the parent KOD variant while requiring only around 12% of the experimental costs compared to the training data. The analysis of functional dynamics, as well as these beneficial mutation sites identified in this study, can provide valuable insights for incorporating modified nucleotides of KOD pol. Additionally, the developed efficient rational design approach in this study, which aids in experimental screening to improve the catalytic efficiency of KOD pol, can also be considered valuable guidance for the engineering of other B-family DNA polymerases.

Results and discussion

Medium-sized preliminary screening

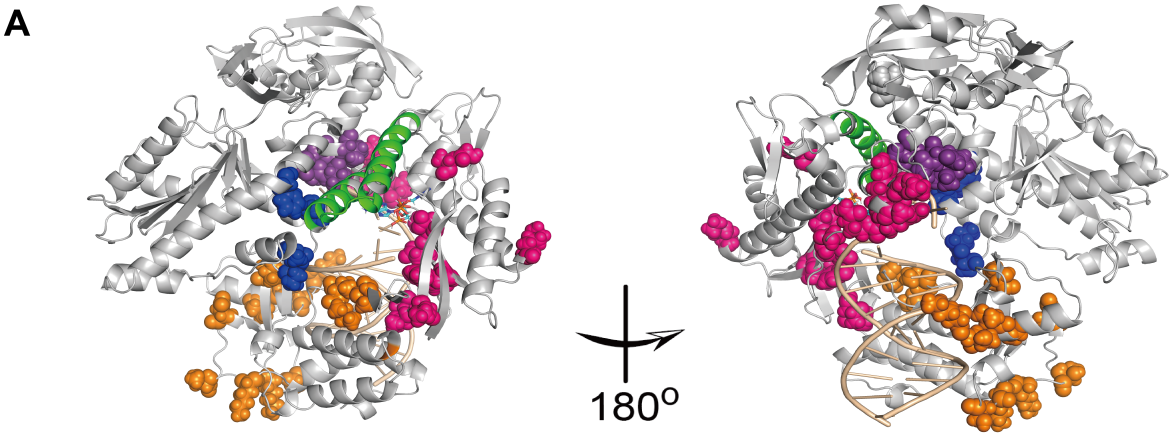
A medium-sized preliminary library consisting of 349 KOD mutants covering over 70 sites was constructed by site-directed mutagenesis[21], random mutagenesis[22], and multi-codon scanning mutagenesis[4]. Over 70 sites were strategically selected within the finger, thumb, palm, N-terminal, and exonuclease subdomains of KOD pol, as illustrated in Figure 1A. These sites were selected based on the identified mutation sites from our previous study and nearby positions, as well as considering some positions that have been previously suggested to potentially affect catalytic activity[23,24]. This library was constructed based on RF, a highly engineered KOD variant referenced in a public patent (WO2022082482A1), carrying eleven mutation sites. In our previous study, RF demonstrated a high catalytic efficiency for efficiently incorporating dye-labeled reversible terminators into the DNA strand. To screen the medium-sized library, the kinetic performance of those variants was evaluated. This involved assessing their ability to incorporate one cycle of 3'-O-azidomethyl-dATP labeled with the fluorescent dye Cy3 (modified dATP), into a primer-template complex. The primer-template complex was labeled with the fluorescent Cy5 dye at its 5' ends (P/T-2Cy5). All expressed variant proteins were subjected to crude purification followed by quantification to ensure the reliability and comparability of the subsequent mutant kinetic screening experiments. The methodology of the

kinetic measurements employed in this screening process is described in the method section and is equivalent to the approach in our previous study. The kinetic performance of KOD variants is described by the Michaelis-Menten parameters k_{cat} and K_m. The relative kinetic performance of those mutants is characterized by the ratio of $k_{cat}/K_m|_{Mut}$ relative to that of the KOD variant RF. $k_{cat}/K_m|_{Mut}$ and $k_{cat}/K_m|_{RF}$ were measured and calculated in the same way, with varying P/T-2Cy5 ranging from 0 to 6 µM while maintaining the modified dATP at a constant concentration of 2 µM. In addition, in the previous experiment, we performed simultaneous measurements of kinetic towards P/T-2Cy5 and dATP. The results showed that both varying P/T-2Cy5 and dATP were effective for screening mutant variants. We define the relative kinetic performance of the mutated variants as E(Mut), as shown in Equation 1. The subscript after the vertical bar "|" indicates the variant used for measurement. The RF variant served as the reference, with a catalytic efficiency (E(Mut)) value of 1.

$$E(Mut) = (k_{cat}/K_m|_{Mut})/(k_{cat}/K_m|_{RF}) \quad \text{(eq. 1)}$$

The results of the relative kinetic screening of 349 KOD variants are displayed in Figure 1B. The dataset consists of data for 149 single-point mutants and 200 variants with multiple combined mutants. The E(Mut) of the single-point mutants (green dashed box in Figure 1B) was generally lower than that of variants with multiple mutations (red dashed box in Figure 1B). The average catalytic efficiency (E(Mut)) of single-point mutants was approximately 0.70, while in contrast, the average E(Mut) of variants with multiple point mutations was around 1.4. Increased E(Mut) of the latter can be attributed to the synergistic effects of advantageous mutations at different sites. However, the average E(Mut) of all these variants is around 1, implying that the preliminary engineering resembles a random walk in the fitness landscape of E(Mut). One possible reason for the challenges in improving of E(mut) in this manner can be attributed to the fact that the 349 tested mutants represent only a very small subset of the relevant combinatorial space that encompasses hundreds of amino acid sites. Also, the possible reason could be that the E(mut) landscape of KOD is essentially epistatic enriched with low-fitness “functional holes”[25]. These initial results highlighted the difficulty in engineering KOD pol toward significantly enhanced incorporation kinetics for modified dATP. The variant KI (A485I) (Figure 1B red half-solid circular) demonstrated a 2.9-fold improvement in the value of E(Mut) compared to RF (A485E) (Figure 1B green half-solid circle). The best-performing variant of our set (Figure 1B yellow-solid circle), carrying two mutations (K676S, M680I), exhibited approximately a 9-fold increase in E(Mut) compared to RF. Overall very few moderately improved mutants (E(mut) > 4) were discovered.

The relative kinetic data of 349 variants were collected and analyzed to generate a medium-sized dataset for constructing our machine-learning model. To evaluate the efficiency of our machine learning-based predictions for the generation of improved polymerase variants, we chose RF as the starting variant for machine learning-based screening.



Exonuclease (Blue)	267 270 271 325	Finger (Green)	451 452 453 454	Finger (Green)	476 477 478 479	Palm (Pink)	574 584 589 590
N-terminal (Purple)	347 348 349 350 351 353 367		455 457 458 459 461 462 463		480 481 482 484 485 486 488	Thumb (Orange)	605 609 610 630 674 676 680
Palm (Pink)	375 379 380 381 382 385 386 387		465 466 467 469 470 471 472 474 475	Palm (Pink)	490 491 493 496 497 531 539 542 545		682 698 705 707 718 723 726

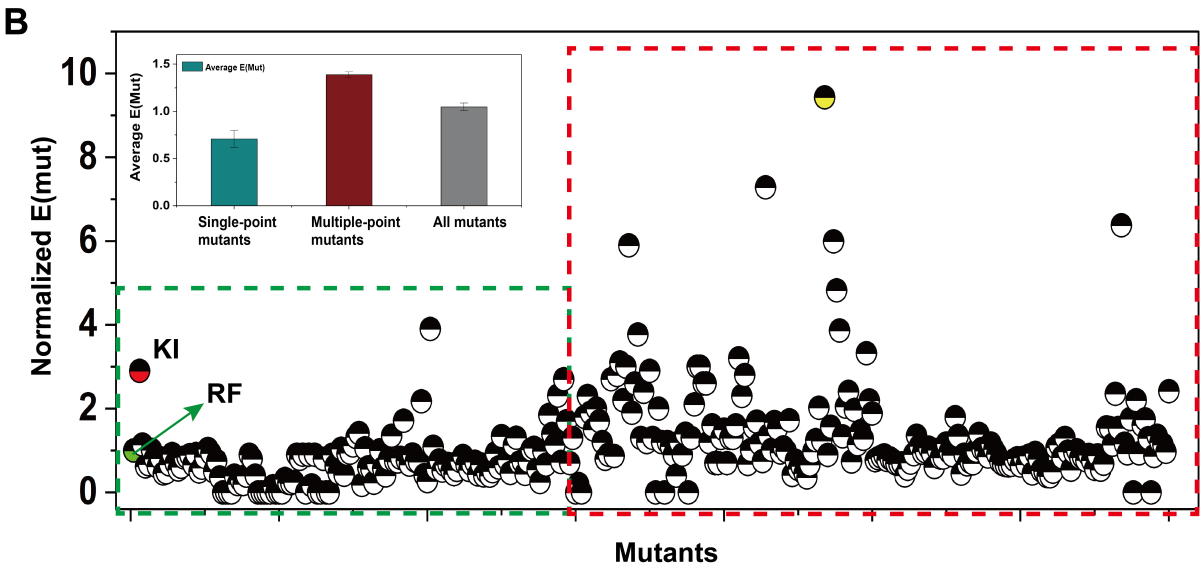


Figure 1. The kinetic screening results of the medium-sized preliminary library. A) The figure illustrates the locations of 70 positions within the crystal structure of KOD DNA polymerase (PDB ID: 5MOF). The selected sites within the finger domain are depicted in a green cartoon format, while the remaining selected sites are represented as spheres. To aid visualization, the

corresponding positions are color-coded on the structure: the exonuclease domain is highlighted in blue, the N-terminal domain in purple, the palm domain in pink, and the thumb regions in orange. The rest of the protein and the DNA strands are depicted in cartoon format. The protein is colored grey, while the DNA strand is colored wheat. Detailed information on the 70 positions is provided in the accompanying table. B) The figure displays the relative kinetic results (normalized E(Mut)) of the 349 tested KOD variants. The kinetic data were fitted to the Michaelis-Menten equation to determine k_{cat} and K_m. E(Mut) of these mutants was calculated using equation 1. The parent variant, RF, is depicted as a green half-solid circle, while the mutant KI is represented in a red half-solid circle. The best-performing variant is shown as a yellow half-solid circle, while the remaining variants are displayed as black half-solid circles. The average E(mut) values for single-point mutants (green), multiple-point mutants (red), and all mutants (gray) subsets are represented using bars and shown in the upper dash box. All E(Mut) values of these variants are summarized in Data Set of appendix A.

Structural insights into the functional dynamics of KOD pol

In order to construct our machine learning model, we initially performed an analysis of a single incorporation process of dye-labeled reversible terminators by a DNA polymerase. This process consists of significant conformational changes in the DNA polymerase, which can be observed in five distinct states[26] (Figure 2A). Firstly, the DNA duplex composed of one primer and one template is bound within a groove jointly formed by the palm and thumb[27]. The incoming nucleotide is added to the 3'-OH ends of the primer at the catalytic cavity, which is jointly formed by the finger, palm, and thumb. States I to V for the incorporation of modified nucleotides are the same as for natural nucleotide incorporation[19-21]. DNA polymerase translocation along the DNA duplex (state V→I, Figure 2A) is blocked by dye-labeled reversible terminators. Also, DNA polymerase translocation along the DNA duplex (state V→ I) is not considered here as our kinetic assay only quantifies the incorporation of a single dye-labeled reversible terminator, and therefore no further nucleotide can attach. Thus, in the catalytic cycle of DNA polymerization, the conformational changes of polymerase could be the rate-limiting steps of single-nucleotide incorporation[28-29].

In addition, KOD pol adopts a disk-shaped structure comprising five subdomains[23], i.e. the N-terminal (N-term.), exonuclease (Exo.), palm, finger, and thumb (Figure 2B). In addition to the apo state, KOD pol complex states I (binary)[30], II (ternary with open finger)[31], and III (ternary with closed finger)[32] have been captured by X-ray crystallography. The atomic fluctuations of the exonuclease, finger, palm, and particularly the thumb domains are all notably suppressed in their complexed states, as observed from the left shift of the normalized B-factor distribution

of C_α atoms (I-III, Figure 2C).

Subsequently, we conducted a comprehensive assessment of the structural dynamics of the KOD variant RF. To assess the structural dynamics of RF, we performed microsecond-scale all-atom molecular dynamics (MD) simulations, focusing on its apo state and binary state I. In the apo state, the finger and thumb of RF demonstrated the greatest flexibility (Figure 2D), with the tip of the finger subdomain swinging between the palm and exonuclease subdomains, inducing the closed and open conformations of the enzyme[33,34]. The closed-to-open conformational transition of the fingertip coincided with the outward movement of the thumb region, resulting in a looser binding pocket of the RF for the DNA stands. In line with the B'-factor analysis (Figure 2C), the overall flexibility of the enzyme is greatly reduced in the binary state (Figure 2D), with the dynamics of the thumb domain being the most significantly affected. The structural stabilization of the RF variant is mainly observed in the thumb region, which is the major binding interface of the DNA strand. Although the finger subdomain is still the most flexible region in the binary state, it has a higher probability (~55% in the apo state vs. ~82% in the binary state, Figure 2E) of adapting to a closed conformation (the fingertip towards the palm region). The biased conformational distribution of 'open' and 'closed' states might limit the kinetics of state transitions of I→II and IV→V in the catalytic cycle (Figures 2A and 2B).

Overall, we suspect that the finger and thumb subdomains are the most "active" regions upon the binding of nucleotides during the functioning of the highly engineered RF. However, the complete sequences of the finger subdomain (covering 52 a.a. sites) and thumb subdomain (covering 183 a.a. sites) still represent a great challenge for direct wet-lab kinetics testing. Thus, we focused on specific positions located in the finger subdomain and the thumb subdomain for the virtual library screening.

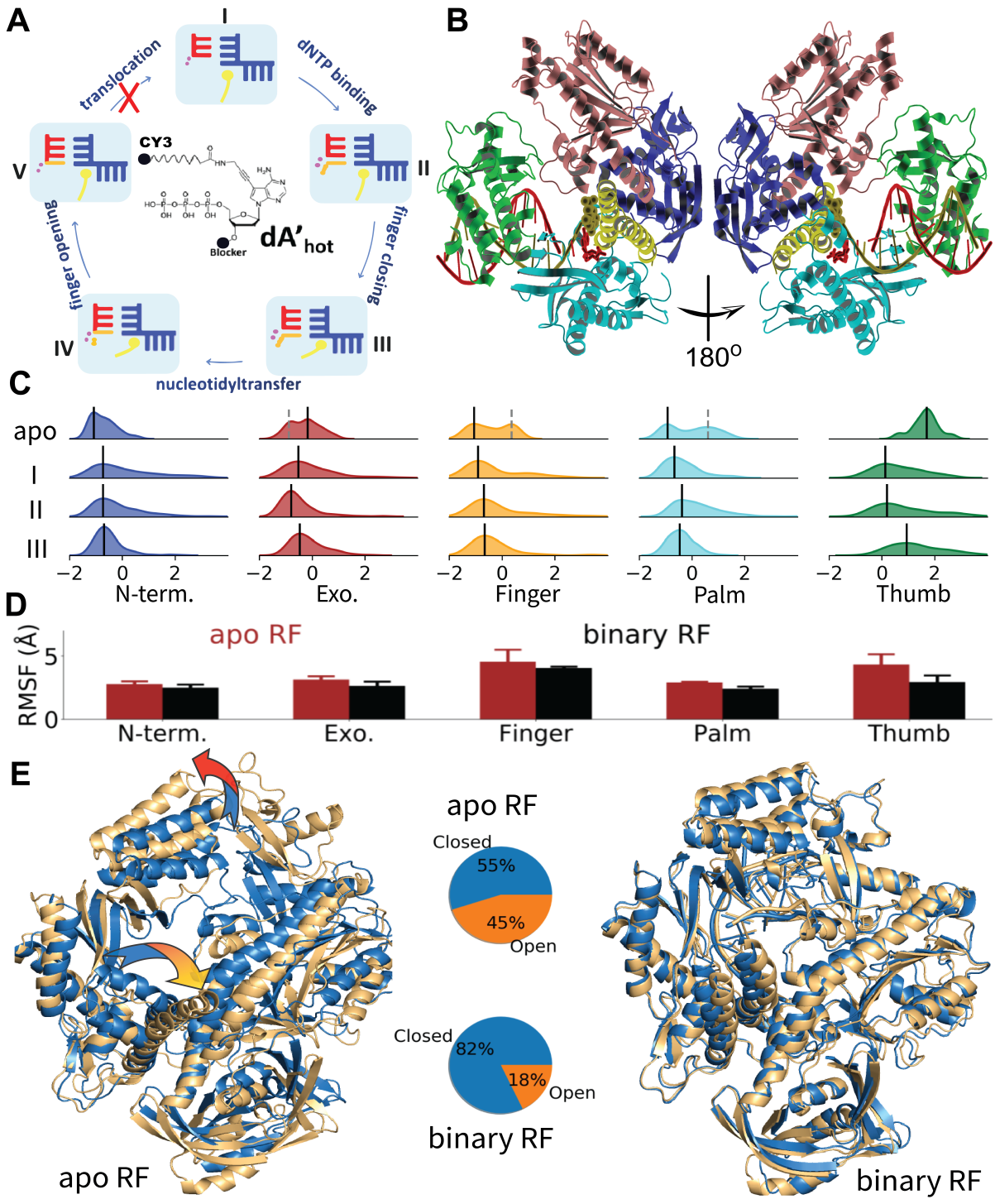


Figure 2. Conformational dynamics of KOD pols. A) Kinetic scheme of an incorporation cycle. Starting from a DNA-bound binary state (I) with an open finger (yellow) and unoccupied active site, one nucleotide enters the active site (II) and the finger domain closes (III) to establish tight binding with the incoming nucleotide; the nucleotide is then added to the 3'-end of the primer strand and pyrophosphate is released from the active site (IV). Note that the translocation of

DNA polymerase is hampered by a 3'-terminated nucleotide (see insert, 3'-O-blocked-dATP labeled with the fluorescent dye Cy3). B) The subdomains of KOD in the 3-dimensional structure are rendered in different colors: N-terminal (blue), exonuclease active site (salmon), finger (yellow), palm (cyan), and thumb (green). C) Structural fluctuations of each subdomain of KOD across the apo state (1WNS), binary state (I, 4K8Z), finger-open ternary state (II, 5VU7), and finger-close ternary state (III, 5OMF), characterized by the density distribution of Cα atoms' normalized B-factor. D) The root mean-square-fluctuation (RMSF) statistics of each functional subdomain of RF in apo and binary states. E) The most weighted (blue) and the second-most weighted (gold) conformations as representative structures of the RF variant sampled in its apo and binary states, respectively. Two arrows mark the major local conformational transitions between the two representatives in the apo form. The probability of open and closed states was estimated by the weightings of the representatives in clustering analyses.

Machine learning model building and virtual library design

Based on those structural insights (as illustrated in Figure 2), our machine learning (ML) model construction can be divided into three main stages (highlighted by the red dashed box in Figure 3A). The three stages include data collection and analysis, representation of mutant features, and implementation of the model. A total of 349 variants from the medium-size library were characterized, and the sequence changes with respect to RF and the kinetic assay-based E(mut) values were collected for model construction. Subsequently, we performed feature engineering on the 349 variants by considering their sequence variations relative to RF and leveraging available crystallographic templates[30,31,32]. In one ML model, each variant was represented as a vector encoding sequence-based and structure-based features. The sequence-based features referred to the characteristics derived from the primary sequence of each KOD variant. These features include commonly used protein descriptors such as the isoelectric point (PI), flexibility, charge, and other commonly used characteristics, and so on[35]. Additionally, three features were calculated to quantify the sequence similarities of the whole protein, finger subdomain, and thumb subdomain between KOD variants and RF. A score derived from PROVEAN was also calculated for each variant to quantify the functional effects of site mutations. The structure-based features are obtained through structural modeling and analysis of the KOD variants. Structural features included solvent-accessible surface areas (SASAs) of hydrophobic residues in the whole protein, finger subdomain, and thumb subdomain, and the backbone B-factor averaged across the protein and subdomains. Moreover, the binding free energy between the variant and the DNA double strands is considered a structural feature. These features serve as

inputs for the machine learning models to predict the E(mut) values and identify variants with enhanced properties compared to RF. The machine learning models were implemented using the Extremely Randomized Trees algorithm within the Scikit-learn library[36,37]. The medium-sized dataset was copied into three child datasets for ML model building, by using the Y-stratified train_test_split function with different random state settings. For each child dataset, the training-testing dataset was used to build an extra tree regressor[38] (models Etr 1-3, Figure 3 C) with five-fold cross-validation. The three regressors respectively reached a coefficient of determination R^2 score of 0.61, 0.57, and 0.60 on the corresponding testing samples, shown in Table S1. We implemented the permutation importance method to quantify each feature's contribution to the three regressors. For each model, the importance of each feature was calculated on their respective testing datasets, shown in Figure 3D. In all three models, features "thumb_blosum80" and "finger_blosum80", which quantified the sequence similarity between the mutant and KOD_RF in thumb subdomain (a.a. 652-729) and finger subdomain (a.a. 447-507) respectively, were shown as the most significant contributing factors. In addition, the accuracy of the three models using three different regression algorithms was calculated and compared, resulting in accuracy values of 75%, 71%, and 71% respectively, as presented in Table S1. The best model selected from each child dataset was then aggregated to output a mean E(Mut) value for each variant, shown in Figure 3C.

The ML models were then employed to rationally identify undiscovered mutations on the finger subdomain and thumb domain towards higher E(mut) at lower experimental throughputs. A virtual library screening was performed, and the overview is shown in Figure 3A (gray dash box). Firstly, hidden mutation-kinetics patterns within the tested pool of KOD mutants were successfully learned by three extra tree regressor models with 24 calculated sequence/structure features. The virtual library considered 9 a.a. sites of KOD pol in total, shown in Figure 3B, with 6 positions located on the finger domain (a.a. 451, 453, 480, 484, 485, 486) and 3 thumb-located positions which are in close contact with DNA template in binary complex (a.a. 665, 666, 667). In the medium-sized library screening, single mutations at residues 451, 453, 480, and particularly 485 were found to have an enhancing effect on E(Mut). Structural neighbors of residue 485, such as 484 and 486, were also taken into consideration. Notably, the selection of the three thumb-located sites (amino acids 665, 666, 667) was based on their specific identification as the sites of interaction with the primer strand[39,40]. The residue at position 485 has been reported to play a crucial role in controlling the specificity of DNA polymerase for incorporating unnatural nucleotides[41]. Therefore, we chose this site as a fixed position in the design of dual-site mutants, which would enable us to evaluate the effectiveness of our mutant

design. Each mutant in the virtual library contains two mutations in comparison to RF, with one located at residue 485 and the other at one of the remaining eight sites. Furthermore, considering the distinct roles of the finger and thumb residues in different functional stages of the incorporation cycle, simultaneous amino acid substitutions within these two subdomains may lead to cumulative effects on the overall kinetics of KOD pol. The virtual library comprises 200 mutants. For each of those mutants, calculated features were fed into three regressors to output a mean predicted E(Mut) from the three prediction models. The predicted E(Mut) of those mutants ranged from 0.7 to 2.2, as shown in Figure 3E.

Overall, by leveraging machine learning algorithms, we successfully constructed a virtual library, enabling us to efficiently screen and identify potential mutants. Subsequently, we selected 45 variants from the virtual library for experimental validation based on their predicted E(Mut) values exceeding the threshold of 1.65 (indicated by the red dash box in Figure 3E), enabling us to reduce experimental costs.

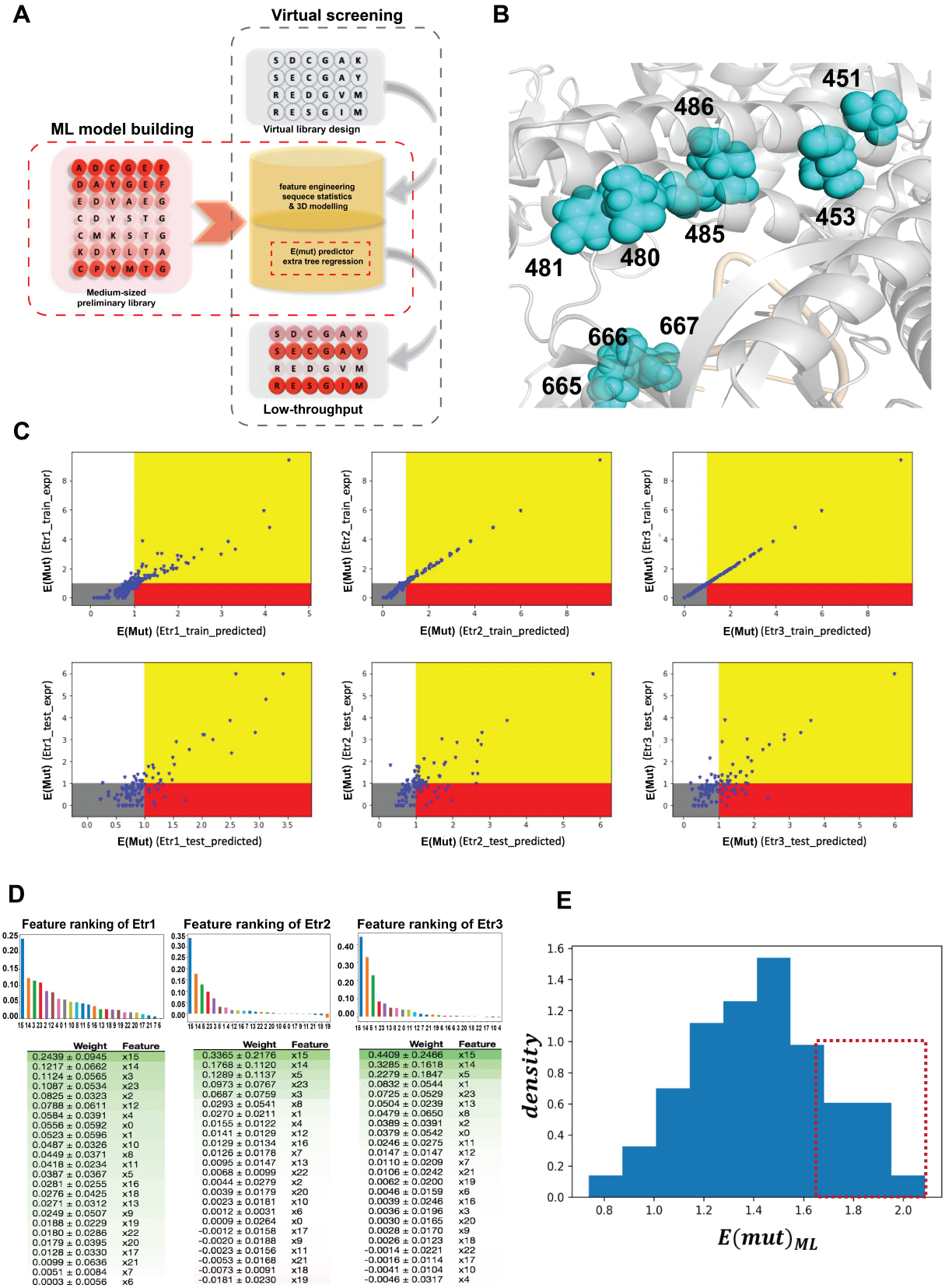


Figure 3. Results of machine learning (ML) model building and virtual library screening. A) Scheme of

ML model building and virtual screening workflow. For ML model construction, sequence-based and structure-derived features, along with tested E(Mut) compared to RF, were utilized to build an E(Mut) predictor for the library (red dash box). The virtual library was designed and screened, and the developed predictor was employed to rank a library of KOD mutants with designed mutations (gray dashed box). B) The location of these positions selected for virtual library construction is depicted on the structure of KOD pol (PDB ID: 5MOF), corresponding to amino acid residues 451, 453, 480, 481, 485, 486, 665, 666, and 667. The residues are depicted as spheres and colored in cyan, while the remaining protein is shown in a cartoon format and colored in grey. C) The extra tree regression modeling of E(Mut) for KOD variants is presented. Each plot displays the predicted and experimentally tested E(Mut) values of the samples on the X-axis and Y-axis, respectively. The upper and lower plots in each column depict the model performances of the built extra tree regressors (Etr 1-3) on the training and testing sets. The evaluation metrics, including the R^2 score and accuracy rate for both the training and test samples, are provided in Table S1. D) The feature ranking of the three extra tree regressors (Etr 1-3) is shown, with 24 features indexed from x0 to x23. These features correspond to various characteristics such as 'PI', 'Flexibility', 'Charge', 'Aliphaticness', 'TotalEntropy', 'Instability', 'aromaticity', 'gravy', 'Mutabilityscale', 'Volumescale', 'Polarizabilityscale', 'jascale', 'ASAInTripeptidescale', 'wholeseq_blosum80', 'finger_blosum80', 'thumb_blosum80', 'w_hsasa', 'f_hsasa', 't_hsasa', 'w_bf', 'f_bf', 't_bf', 'md12_pd_deltaG', 'provean_score'. E) The distribution of predicted E(Mut) values is displayed for variants carrying designed double mutations located on the finger domain and the thumb domain. Variants in the red dash box were selected for experimental validation.

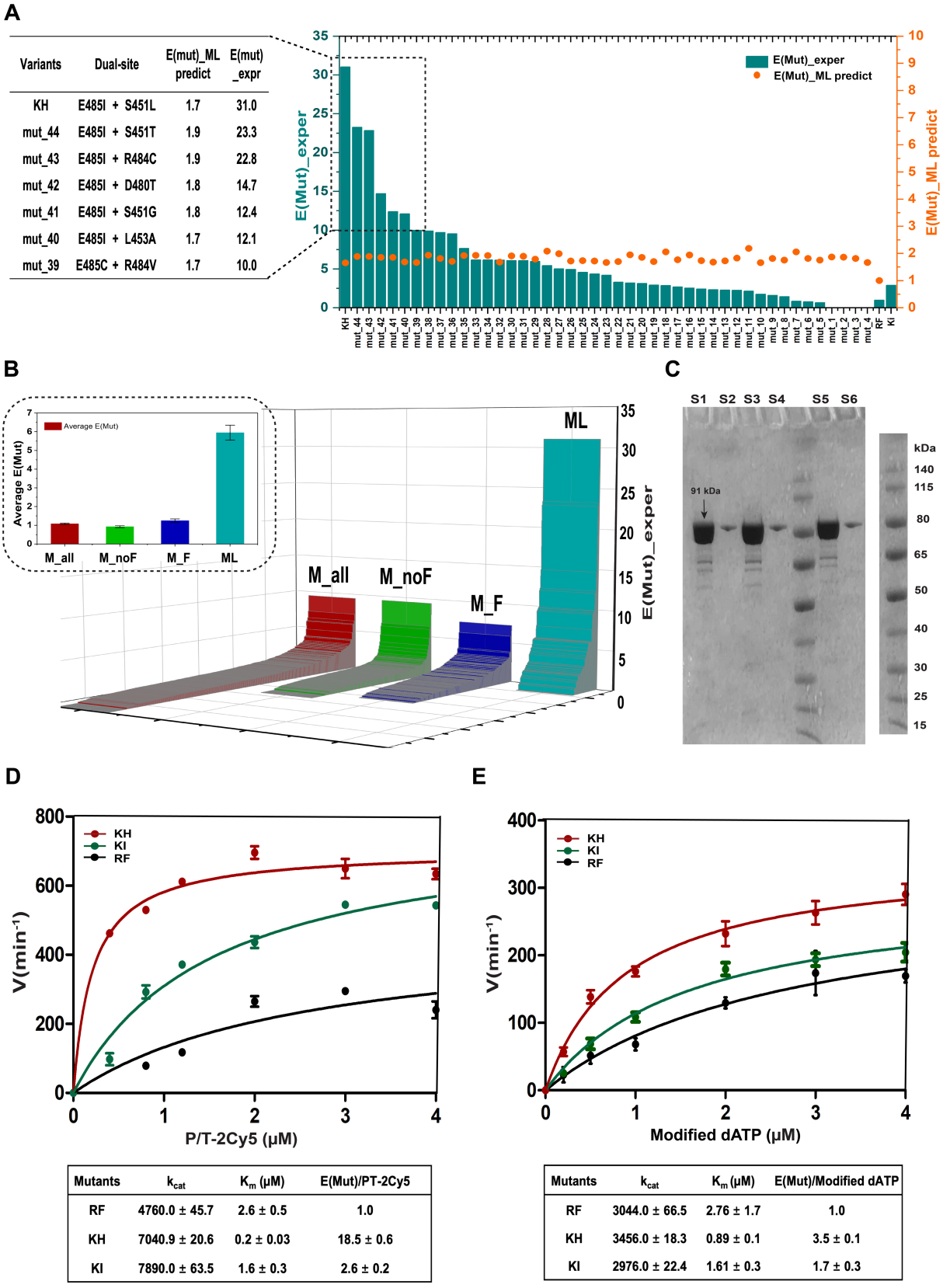
Machine learning-assisted mutant screening

We constructed a library of those 45 variants by site-directed mutagenesis and subsequently assessed their relative kinetic performance using the same experimental procedures as for the initial library, analyzed according to equation (eq.1). E(Mut) of those mutants was calculated and shown in Figure 4A. Despite the ML-predicted E(Mut) values of the 45 mutants ranging from 1.65 to 2.2, the experimental data revealed considerable variations in their catalytic efficiencies for incorporating modified dATP, ranging from 0 to 31 (Figure 4 A). The disparity between predictions and experimental results may potentially stem from the limited size of the training database. In other studies, the number of data points used for machine learning training reach up to 19,000[42]. Employing a larger dataset of experimentally characterized variants usually implies more accurate prediction outcomes[43]. The most promising mutant, named KH, contained two mutations, A485I and S451L, which resulted in a 31-fold increase in E(Mut) compared to RF. Interestingly, KI (A485I) from the medium-sized library revealed a 2.9-fold increase in catalytic efficiency. Seven mutated variants showed a significant increase in kinetic E(mut) values, exceeding ten-fold, as shown in the dashed box in Figure 4A. Interestingly, all

seven mutants shared a common feature, as the dual mutation sites were located in the finger domain of the KOD pol. Therefore, we distinguished the variants with mutations located in the finger region from the initial medium-sized library and compared them with the ML-predicted variants. The results of this comparison are presented in Figure 4B. The average E(Mut) of all variants from the medium-sized library was around 1 (noted as M_all, shown in the upper insert dashed box in Figure 4B). Variants with finger-located mutations (noted as M_F, represented in the upper insert dashed box in Figure 4B) exhibited a discernible improvement in average E(Mut) values, around 1.2. In comparison, other mutants exclusive to the finger region (noted as M_noF, shown in the upper insert dashed box in Figure 4B) had a lower average E(Mut) value of around 0.9. These findings suggest that mutations occurring at specific sites within the finger domain exhibit a higher efficiency in enhancing the incorporation of modified nucleotides. The average E(Mut) of ML-predicted variants increased more than 5-fold compared to the non-ML-predicted variants, as shown in the upper inset of Figure 4B. The results indicate that applying the machine learning strategy to design mutations in the finger subdomain significantly enhances the efficiency of the screening process. In addition, these results suggest that the rationally designed dual-site mutations on the finger subdomain are more effective in enhancing the DNA polymerase's kinetic performance compared to traditional mutagenesis.

To confirm whether the observed enhancement in catalytic activity is influenced by potential impurities in the semi-purified mutant proteins, we performed rigorous purification using chromatography with three prepacked columns to purify KH, KI, and RF variants. The protein purity of each sample was confirmed to exceed 95% through SDS-PAGE analysis, shown in Figure 4C. Subsequently, we performed tailored kinetic assays developed for P/T-2Cy5 and 3'-O-azidomethyl-dATP labeled with the Cy3 dye (modified dATP), separately. The kinetic assays were conducted with varying concentrations of either P/T-2Cy5 or modified dATP ranging from 0 to 6 μM while maintaining the other one at a constant concentration of 2 μM. To distinguish between the kinetic performance of two substrates, we used E(Mut)/P/T-2Cy5 to represent the kinetic measurement towards varying P/T-2Cy5 and E(Mut)/modified dATP to represent the kinetic measurement towards varying modified dATP. The kinetic results of KH, KI, and RF are shown in Figures 4D and 4E. The improvement in the E(Mut)/P/T-2Cy5 value of KI compared to RF was approximately 2.6-fold, which is consistent with the previous testing results (2.9-fold). Despite the potential for activity loss during the purification process, KH still exhibited approximately 18-fold rate acceleration, as demonstrated by the kinetic curves shown in the right panel of Figure 4D. Both KH and KI variants showed an increase in the

E(Mut)/modified dATP value compared to RF, although the increase may not be as significant as observed with the E(Mut)/P/T-2Cy5. The results demonstrated a significant improvement in the catalytic efficiency of the KH variant in incorporating the modified dATP compared to RF. This suggests that KH may exhibit more satisfactory performance in sequencing applications relative to RF. The mechanism underlying the enhancement in kinetics observed in KOD variant KH was subsequently analyzed in comparison to KI and RF.



Variants	Dual-site	E(mut)_ML predict	E(mut) _expr
KH	E485I + S451L	1.7	31.0
mut_44	E485I + S451T	1.9	23.3
mut_43	E485I + R484C	1.9	22.8
mut_42	E485I + D480T	1.8	14.7
mut_41	E485I + S451G	1.8	12.4
mut_40	E485I + L453A	1.7	12.1
mut_39	E485C + R484V	1.7	10.0

Mutants	k_{cat}	K_m (μM)	E(Mut)/PT-2Cy5
RF	4760.0 ± 45.7	2.6 ± 0.5	1.0
KH	7040.9 ± 20.6	0.2 ± 0.03	18.5 ± 0.6
KI	7890.0 ± 63.5	1.6 ± 0.3	2.6 ± 0.2

Mutants	k_{cat}	K_m (μM)	E(Mut)/Modified dATP
RF	3044.0 ± 66.5	2.76 ± 1.7	1.0
KH	3456.0 ± 18.3	0.89 ± 0.1	3.5 ± 0.1
KI	2976.0 ± 22.4	1.61 ± 0.3	1.7 ± 0.3

Figure 4. The screening results of machine learning aided KOD variants. A) The figure

illustrates a comparison between the experimentally determined E(Mut) values (shown by green bars) and the ML-predicted E(Mut) values (represented by orange dots) of the mutant variants. Notably, the black dashed box highlights the mutants that exhibit a ten-fold improvement in kinetic performance. B) Statistical distributions of tested E(Mut) values of KOD variants from the medium-sized library. The red bars (M_all) correspond to all tested mutants, the green bars (M_noF) represent mutations on non-finger regions, and the blue bars (M_F) indicate mutations on the finger region. The E(Mut) values of KOD variants from the virtual library were noted as ML, shown in bars and colored cyan. The average E(mut) values for the M_all, M_noF, M_F, and ML subsets are represented using bars of the corresponding colors, respectively. The average E(Mut) is shown in the upper dash box. C) The SDS-PAGE analysis (12% gel) for rigorously purified proteins with RF represented by S1 (5 µg) and S2 (0.25 µg), KI represented by S2(5 µg) and S4 (0.25 µg), and KH represented by S5 (5 µg) and S6 (0.25 µg). The estimated protein purity was approximately 95% based on ImageJ analysis of the gels. D) and E) The kinetic results of RF, KI, and KH towards P/T-2Cy5 (D) and modified dATP (E) under the same conditions. The experimental data were fitted non-linearly using the Michaelis-Menten equation, and the values of k_{cat} and K_m are presented in the table bottom of the graph. Each kinetic experiment was performed in duplicate.

The mechanism underlying kinetics boosting of KOD polymerase.

Single mutations at positions 485 or 451 in the KOD variant RF resulted in an enhancement in catalytic efficiency, with improvements of about 2-fold, respectively (Figure 5C and Figure 1B). The single mutations achieved a lower level of improvement compared to the combined mutations introduced to the KH (A485I and S451L), which showed an impressive over 18-fold improvement in catalytic efficiency. In other words, the synergy between S451L and A485L mutations contributes to a comprehensive kinetic boosting of KOD pol. However, the two sites are evolutionarily uncorrelated (Figure 5A) and their respective degrees of conservation are divergent[44] (position 485 is highly conserved to alanine whereas position 451 is non-conserved, Figure 5B). Consistent with their weak evolutionary correlation, the two sites have no close contact in 3D space (Cß-Cß distance: 21.6 Å). We then explored the synergy of the two sites on the structural dynamics of KH via all-atom molecular dynamics (MD) simulations. As in the simulations of the RF variant (Figure 2), three independent MD simulations for the KH in both the apo state and binary state were performed (6 µs in total), in order to capture the synergy of the two sites on the structural dynamics of KH. For comparison purposes, MD simulations were also conducted for a single-site mutant with the A485I mutation only, denoted as KI (Figure 5C). In the apo state, the bending angles of the two helices (Helix@485 and Helix@451) in the finger subdomain were found to be indistinguishable on average among the KH, RF, and KI

variants (Figure 5D). The fingertips of the three systems are free to adapt to the "open" and/or "closed" conformations regardless of the two sites' a.a. contents (Figure 5E and Figure 2E). Interestingly, after the binding to the DNA duplex, the mutation A485L induces a significant conformational change of the finger subdomain (Figure 5D). After the binding of the DNA duplex, the helix@485 in both the KH and KI is less bent towards the palm region than in their RF counterpart (reduced from 20º to 5º on average, Figure 5D). The bending angles of helix@451 of the KH and KI are also slightly decreased. Overall, the finger subdomain of the two mutants is more likely to avoid the closed conformation than that of the parent RF in the binary complex state (Figure 5F), resulting in a larger space for the binding of modified nucleotides in the functional states I-III (Figure 2A). The increased available space is advantageous for accommodating dye-labeled reversible terminators, thereby improving the kinetic performance. However, the local conformational shift of the finger domain is indistinguishable between the KH bearing the double mutations and the KI only with a mutation on a.a. 485 (Figure 5D). Therefore, the local conformational shift of the finger mainly relies on the mutation at a.a. 485, and this shift becomes significant after the binding of the DNA duplex.

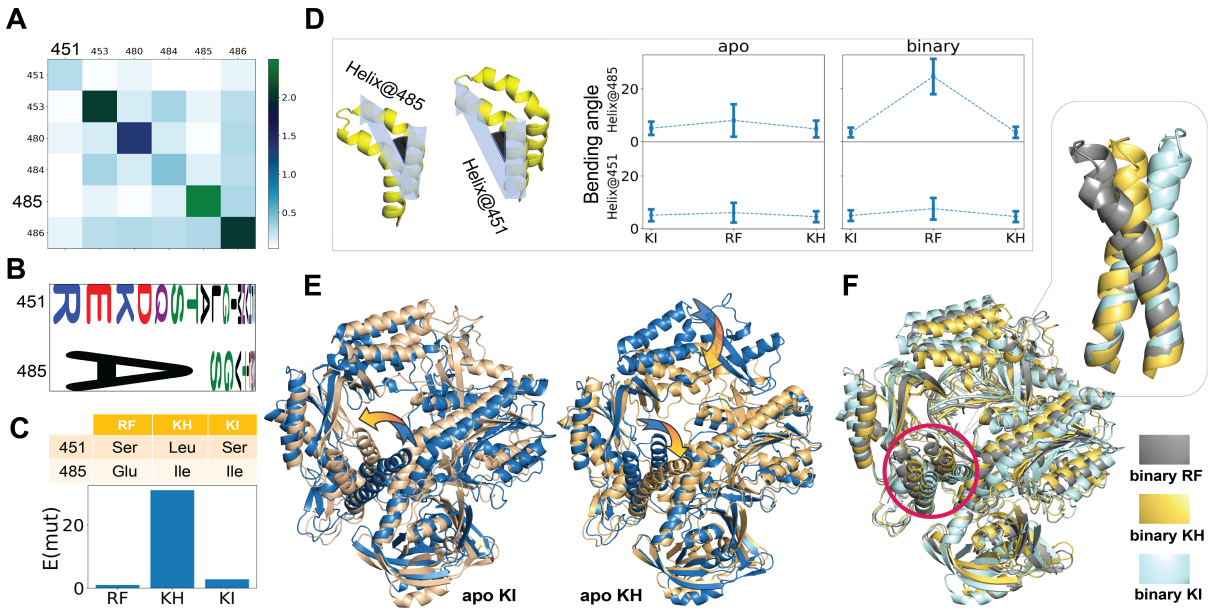


Figure 5. Comparison of sequence and structure of the parent KOD RF and ML predicted KOD mutants. A) The evolutionary covariance score matrix of KOD on a.a. 451, 453, 480, 481, 485, 486. A higher score means a stronger evolutionary correlation between pairwise positions. The correlation between a.a. 451 and a.a. 485 is around 0.04. B) Probability-based sequence logo plot of KOD drawn by logo maker, of a.a. 451 and 485. C) The type of mutations on the a.a. 451 and 485 sites of the RF, KH, and KI (upper panel); the measured E(Mut) of the three variants (lower panel). D) The bending of the finger subdomain of the RF, KI, and KH in MD simulations are characterized by the bending angle of the two

finger helices (Helix@485 and Helix@451). Note the inserts represent the definition of the two angles (black area). E) The most weighted (blue) and the second-most weighted (orange) representatives of the KI and KH sampled in apo state, respectively. The arrows highlight the local conformational transitions between the two representative structures. F) Superposition of the most weighted representative of the RF, KH, and KI in a binary state. The enlarged plot highlights the conformational difference in their finger subdomains.

There is no direct interaction between sites 451 and 485, and the nucleotides of the DNA duplex, as the two sites are located at the back of the DNA strand binding groove (Figure 6A). The communication (if any) between the two sites and the DNA duplex must be mediated by other subdomains. Indeed, several closely contacted (closest heavy atom distance less than 0.5 nm on average) residues of a.a. 451 and 485 are detected in the N-term. (a.a. 364), palm (a.a. 446), and Exo. (a.a. 329, 332, 333, and 336) subdomains (Figures 6B and 6C). Interestingly, the KH and KI share the same mutation on a.a. 485 (Ile), while there are two close contacts (485-329 and 485-332) missing on the KI (Figure 6B). On the other hand, the two close contacts of a.a. 451 (451-364 and 451-446) are both indistinguishable from their average distances for all the three KOD pols (Figure 6B), although the KH is different from KI and RF on the a.a. 451 (Figure 6C). The distance distribution (lower panels in Figure 6C) of these contacts shows that the KH, which features dual-site mutations, forms the most stable inter-subdomain contacts: all distance distributions are single peaked within 5 Å. For the single-site mutated KI and the parent RF, dual-peak distributions are observed for one (451-364) and three (451-364, 485-332, and 485-329) inter-subdomain contacts, respectively. These results indicate that the effect of the mutations is more consistent with a change in the dynamics of the residue-to-residue contacts, rather than with changes in the average structures.

Therefore, we characterized the pairwise dynamical correlation of all residues and/or nucleotides by the normalized covariance of their center-of-mass and then determined the dynamically cooperative groups by maximizing a net correlation (equation 2, see method) of the whole complex[45,46]. We found that the dynamics of the DNA duplex are mainly cooperative to the palm and thumb subdomain, while the T-DNA is partially cooperative to the N-term., Exo. and finger (Figure 6D). The situation is similar for the single-site mutated KI, in which the T-DNA is still partially cooperative with Exo. In both RF and KI, the T-DNA strand is split into two cooperative groups at the 5'-terminal (A (768)) (Figure S1). Importantly, the core of the binding pocket (T (769), right panel of Figure 6A), on which only a correct incoming nucleotide should be base-paired (right panel, Figure 6A), is a covalent bond to the uncooperative A (768) of T-DNA (Figure S1). From a thermal stability point of view, an

extended correlated dynamical network around the reaction center could enhance the catalysis of an enzyme[47]. Indeed, for the best-performed KH, the entire P-DNA and T-DNA strands are clustered in a single cooperative group (Figure 6D), indicating an improved dynamic cooperation of A (768) with the whole DNA duplex structure.

Finally, we quantified the correlation between different mutations and the structural dynamics of the DNA duplex, by estimating the sum of maximum flow ($\sum MaxF$)[48] stemming from a.a. 451 and 485 to each nucleotide of the DNA duplex (see methods section), mediated by all possible residues (Figure 6E). The higher $\sum MaxF$, the larger the capacity of information (in the form of cooperative conformational dynamics) spread from the two sites to the DNA duplex (and vice versa). Results show that the $\sum MaxF$ is well correlated to the E(Mut) values for the RF, KH, and KI (red dashed line in the upper right panel of Figure 6E). The best-performed KH has the highest communication capacity of conformational dynamics between the dual-site mutations and the DNA strand. Indeed, an explicit communication pathway (colored as red lines, Figure 6F) between the two mutation sites 485 and 451 and the T (769) for substrate-pairing in the core of the binding pocket is observed in the KH only (Figure S1).

Taken together, we conclude here that the observed enhancement of incorporation kinetics generated by the dual-site mutations (S451L and A485I) of the best-performed KH is a result of a change in conformational dynamics within the local enzyme subdomain and concerted dynamical motion of the bound DNA duplex. Namely, A485I regulates the local conformation of the finger subdomain, which dominates the processes of "capture" and "release" of a single-nucleotide; while S451L modulates the internal dynamical cooperation of the DNA duplex, which extends the correlation network around the core of the binding pocket. In particular, position 451 itself is a non-conserved site and the effect of mutations on the average conformation of the finger subdomain is insignificant. These properties imply that site 451 could be easily ignored by traditional active-site prediction and test protocols. Furthermore, we discovered that tuning the DNA duplex dynamics, especially the nucleotide for substrate base pairing, via the specific combination of amino acids 485 and 451, could accelerate the incorporation efficiency of the substrate with a half-closed binding pocket, implying an unexplored regulation mechanism of the polymerase functionality.

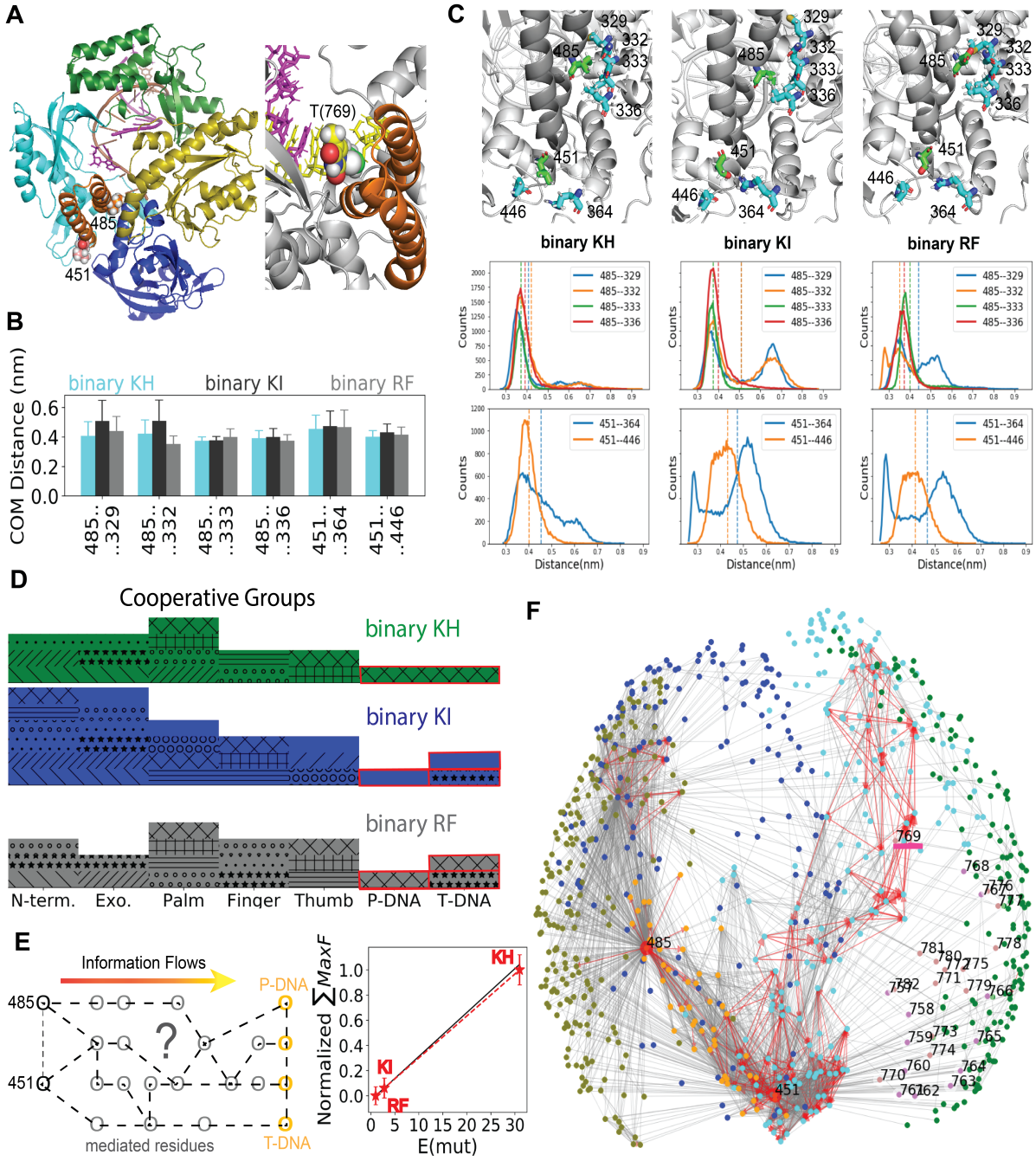


Figure 6. Analysis of the conformational dynamics of the KH, KI, and RF variants in their binary states. A) The relative positions of the two mutation sites (451 and 485) and the DNA duplex in the 3-D structure. The protein is colored according to its functional domains. The nucleotide T (769) for substrate pairing is highlighted as vdW spheres in the right panel. B) and C) Inter-subdomain close contacts of a.a. 451 and 485 of the three systems. The average distances and their distributions are reported in B) and the lower panel of C), respectively. The vertical dashed lines marked the average distances, respectively. Note the representative conformation of the close contacted residues is shown as an illustration for each system. D) The

dynamically cooperative groups assigned to each subdomain and the DNA strand of the KH, KI, and RF, separately. Different fill patterns represent different cooperative groups of each system, and two subdomains sharing the same pattern means their conformational dynamics are cooperative in the same binary complex. E) Left panel: a schematic diagram of the information flow from the two mutation sites (a.a. 451 and 485) to the DNA duplex mediated by unknown residues; Right panel: the correlation between the measured E(Mut) and the sum of maximum flows of correlated movements from the two sites to the DNA duplex. Note the black solid line represents an ideal correlation estimated by the trend of RF and KI. F) The core of the binding pocket, T (769) is nested by the superimposed pathways of all maximal flows stemming from a.a. 451 and 485 (labeled) to the DNA duplex (labeled) as observed in binary KH. Here each vertex represents a residue or nucleotide, and each edge represents the mean flow between two vertices. The edges with weightings greater than 0.5 are highlighted in red, and between 0.5 and 0.1 are rendered in grey, otherwise are not displayed for clarity. Note the relative position of the vertices is assigned according to their overall pairwise correlation.

Conclusion and future perspectives

In summary, our study presents a workflow for the rational design of double-site mutants based on a highly engineered KOD variant RF, with the goal of enhancing the enzyme's efficiency in the incorporation of modified nucleotides. The workflow can be summarized into three stages: a medium-sized preliminary library construction and screening, the machine learning (ML) model building and virtual library screening, and the experimental kinetic characterization of the top 22% of the variants predicted by the ML algorithm. We successfully enhanced the catalytic efficiency of KOD variants towards unnatural nucleotides by more than one order of magnitude. In comparison to the parent variant RF, the KOD variant KH exhibits highly enhanced kinetic performance, indicating its potential as a more favorable candidate for future sequencing applications. Therefore, in future studies, it would be worthwhile to further assess the KOD variant KH on NGS sequencing platforms to evaluate its improvement in catalytic efficiency and to assess its potential commercial value. Moreover, it should be noted that the database utilized for machine learning training in this study was only a medium-sized database. In the future, high-throughput screening methods could be developed to screen more variants for the training of machine learning algorithms. Acquiring a larger and more diverse sequence dataset for training purposes is crucial to improving the accuracy and reliability of predictions in machine learning applications. Furthermore, future rational design of potential sites that may possess synergistic functionality can be attempted in more mutation sites of the finger domain and thumb domain, which might be missed in this study.

Consequently, we propose that the rational integration of detailed insight into the functional dynamics and knowledge-based machine learning models, is a powerful strategy for efficiently determining key sites and accelerating the engineering of complex enzymes, towards a targeted biological function. We believe that both the strategy and the findings presented in this work will enable advanced protein engineering of other B family DNA polymerases or other nucleotide-involving types of protein machinery in the future.

Materials and Methods

Reagents and instruments

DNA templates (5' CGTGTATGCGTAATAGGATCCCGACTCACTATGGACG 3') and primers (5' CGTGTATCGTCCATAGTGAGTCGGGATCCTATTACGC 3') labeled at the 5'-terminus by Cy5 dye were synthesized by Invitrogen. The Cy5 labeled DNA/primer complex product (P/T-2Cy5) was obtained by the following steps: 200 µL of labeled primer (100 µM) was annealed to 200 µL of the labeled DNA template (100 µM) by heating to 80 °C for 10 minutes in a thermomixer and then slowly cooling down to room temperature by turning off the power. Mutant primers were purchased from GeneScrip, and dNTPs, Pfu DNA polymerase, and Dpn1 were bought from NEB. 3'-O-azidomethyl-dATP-Cy3 (modified dATP) was supplied by the BGI synthetic chemistry group. Sequences of parent variant RF and all mutants were integrated into a commercial vector pD441-pelB from DNA2.0. LB broth, kanamycin, and IPTG were bought from Thermo Fisher. The ÄKTA Pure Protein Purification system and columns were purchased from GE Healthcare. The Spark Control microplate reader was purchased from TECAN. PCR reactions performed in a Bio-Rad Life ECO™ Gradient Thermal Cycler were carried out in a cycler 96 system with visualization of DNA for motion via SYBR safe (Invitrogen).

Library construction and protein expression

Site-directed mutagenesis strategy was performed for library construction. The PCR reaction components consisted of 10X Pfu Buffer with MgSO4 (2.5 µL), 10mM dNTPs mix (0.5 µL), DNA template (75 ng), Pfu DNA polymerase (1U), nuclease-free water to 25uL. Thermal cycling conditions were 95 °C for 3 minutes denaturation step, followed by 20 cycles at 95 °C for 30 seconds, 53 °C for 30 seconds, 72 °C for 7 minutes, and final extension at 72 °C for 10 minutes. Then, 1 µL of DpnI was added to each PCR tube to digest template DNA at 37 °C for 1 hour. 5 µL of PCR product of each mutant was transformed into the E. coli DH5α competent

cell. Preparations of recombinant plasmids were conducted using the QI Aprep Spin Miniprep Kit from QIAGEN, and then these plasmids were sequenced by BGI (Genome Sequencing Company) Group. Next, those recombinant plasmids with the expected sequencing results were transformed in E. coli BL21 (DE3), and then one colony was selected for cultivating and protein expression.

Mutant protein expression and purification

1 mL of LB medium in a 96-deep-well plate was used for small-scale expression, and 1L of LB medium in a conical flask was used for large-scale expression. Cells grown in LB medium containing 50 μg/ml kanamycin were induced at OD600 nm = 0.6 - 0.8 by the addition of 0.5 mM IPTG. Protein production was carried out at 25 °C, 220 rpm overnight.

Crude purification was performed using cell pellets collected from 1 mL culture in a 96-deep-well plate. The cell pellet was collected and resuspended in lysis buffer 1 (20 mM Tris-HCl, 10 mM KCl, 10 mM $(NH4)_2SO4$, 0.1% Triton, 4 mM MgSO4, 1.25 mM PMSF, and 1 mg/ml lysozyme, pH 7.6) with a ratio of 0.04 g cell pellet per mL buffer. The resuspended cells were incubated at 37 °C for 10 minutes, followed by thermal denaturation at 80 °C for 30 minutes. After centrifugation at 12,000 X g for 20 minutes, the supernatant was collected, and the crude enzyme concentration was determined. The approximate degree of purity of these KOD variants was checked via SDS-PAGE and the estimated protein purity was approximately 80% based on ImageJ analysis of the gels. The protein concentration was measured and estimated by using the Bradford protein assay[49]. After calculating the total protein concentration (Conc. Total) using the Bradford method, the final target protein concentration was determined by multiplying the Conc. Total by 80%. The crudely purified protein variants were directly used for the kinetic screening.

Purified polymerases of RF, KI, and KH were obtained by the following methods. The cell pellet was suspended in lysis buffer 2 (500 mM NaCl, 5% glycerol, 20 mM imidazole, 1.25 mM PMSF, 50 mM potassium phosphate, pH 7.4) and then crushed with a homogenizer AH-1500 purchased from ATS engineering company. After a 30-minute thermal denaturation at 80 °C, centrifugation at 12,000 rpm 4 ℃ (Beckman Avanti J-26) for 30 min was performed to remove cell debris. Next, the supernatant was filtered with a 0.22 μm pore-size membrane. Then, the subsequent purification process was performed using ÄKTA pure 25 (GE Healthcare) with pre-equilibrated Nickel affinity chromatography (5 mL HisTrap HP column, GE Healthcare). The filtrate was loaded with a flow rate of 2.5 mL/min so that the retention time

will be 2 min and then a wash step was performed with 50 mL buffer 2. The elution procedure was performed with a linear gradient of Buffer 3 (500 mM NaCl, 5% glycerol, 500 mM imidazole, 50 mM potassium phosphate, pH 7.4) from 0 to 100% in 10 CV with a flow rate of 5 mL/min. Fractions with 280 nm UV signal above 100 mAu were collected. The collected elute was diluted 6-fold with Buffer 4 (5% glycerol, 25 mM potassium phosphate, pH 6.6) and then loaded onto a pre-equilibrated HiTrap Q HP column (5 mL). The flow-through was collected. The collected flow-through was uploaded in a pre-equilibrated HiTrap SP HP column (5ml), followed by a wash step with 50mL Buffer 5 (50 mM NaCl, 5% Glycerol, 50 mM Potassium Phosphate, pH 7.4). The elution was carried out with a linear gradient of Buffer 6 (1 M NaCl, 5% Glycerol, 50 mM Potassium Phosphate, pH 6.6) from 0 to 60% in 10 CV. The first peak was collected. The eluate was dialyzed against Buffer 7 (20 mM Tris 200 mM KCl, 0.2 mM EDTA, 5% glycerol, pH 7.4) for 18 hours, and then stored in 50% glycerol in a final concentration of 1mg/mL. Protein concentration was determined by measuring the absorbance at 280 nm using a microplate reader (Bio Tek) and calculated using an extinction coefficient of 1.393 M^{-1} cm^{-1} predicted by the ExPASy server. The purity of the protein was analyzed using 12% SDS-PAGE. Purified enzymes were stored at -80°C until being used.

Kinetic measurements of KOD variants

The kinetic assays were conducted with varying concentrations of either P/T-2Cy5 or 3'-O-azidomethyl-dATP labeled with the Cy3 (modified dATP) ranging from 0 to 6 µM while maintaining the other reagent at a constant concentration of 2 µM. Upon successful incorporation of 3'-O-azidomethyl-dATP labeled with the Cy3 dye into P/T-2Cy5 by DNA polymerase, excitation at 530 nm of Cy3 results in fluorescence resonance energy transfer to the nearby Cy5 molecule, resulting in the emission at 676 nm by Cy5. The 1X reaction buffer used for kinetic assays comprises the following components: 0.1 mg/mL BSA, 20 mM Tris-HCl, 10 mM KCl, 10 mM $(NH4)_2SO4$, 0.1% Triton, 4 mM MgSO4, with a pH of 8.8. The premixed solution consisting of 1X reaction buffer, P/T-2Cy5, and modified dATP in 384-well (black clear-bottom) plates (Corning®) was prepared on ice. In the final step of the reaction system, 0.5 µg DNA polymerase was added to the premixed solution and then the kinetic measurement was performed using a Spark Control plate reader (TECAN) at a temperature of 40 ºC for 1 - 2 hours. The gain setting for the measurements was set to 60. Before each measurement, the reaction mixture was shaken for 10 seconds to ensure proper mixing. During the measurement, fluorescence resonance energy transfer (Cy5 FRET) signals were monitored.

After the measurement, the maximum slope (reaction rate V) of the Cy5 FRET signal for each concentration of P/T-2Cy5 or modified dATP was calculated and used for determining the kinetic parameters (k_{cat} and K_m). k_{cat} and K_m were obtained from the Michaelis-Menten equation[50] through non-linearly fitting by GraphPad Prism 5. RF was used as the positive control and its k_{cat}/K_m value was defined as 1, and the E(Mut) values of all KOD mutants were calculated according to eq.1. The same experimental methods for the quantification of the enzyme kinetics were used for both crudely purified and fully purified KOD variants.

Machine learning-assisted screening

The machine learning (ML) models built in this study were aimed at identifying KOD variants with enhanced E(Mut) values compared to RF. Model building was based on characterized E(Mut) of a medium-sized pool of KOD variants all seeded from the RF variant. For each kinetically characterized variant, its mutation contents with respect to RF and assay-based E(Mut) value of mutant proteins were collected for model construction. Each variant was represented as a vector encoding sequence-based and structure-based features in one ML model. The primary sequence of each variant was inputted into a ProtFeat Python[51] class to generate common protein descriptors, including PI point, flexibility, charge, aliphaticness, total entropy, instability, aromaticity, gravy, mutability scale, volume scale, polarizability scale, and ja a scale as an in tripeptide scale. Apart from including features directly generated from ProtFeat, three features were also calculated to quantify the sequence similarities of whole protein, finger subdomain, and thumb subdomain between the variant and RF. Besides, a score derived from PROVEAN[52] was calculated for each variant to quantify the functional effect of site mutations. Structural features of each variant include solvent-accessible surface areas (SASAs) of hydrophobic residues located respectively in whole protein, finger and thumb subdomains, backbone B-factor averaged across the whole protein, finger and thumb subdomain as well as the binding free energy between the variant and the DNA double strands. Structural models of apo state and binary complex state for each variant were built with modeler v9.23[53]. Structural features were calculated based on conformations sampled from MD trajectories performed with AMBER18 suite of programs.

The machine learning models built here were implemented with the Extremely Randomized Trees algorithm within Scikit-learn[54]. Extremely Randomized tree is one ensemble learning technique that aggregates results from M decision trees to output a mean result for regression problems or a majority vote for classification problems. Each Decision Tree in the Extra Trees

Forest is built from the original training input, without random bootstrapping as used by Random Forest. For each decision tree, a random subset of K features is drawn, and selected features are split into random cut points and the best feature chosen for each node is chosen based on the total reduction of some mathematical criteria (typically the Gini Index). The rationale behind the Extra Trees method is that the explicit randomization of the cut point and features combined with ensemble averaging should be able to reduce variance.

The dataset used for ML model building has 349 samples in total and was copied into 3 child datasets using Y-stratified train_test_split function (ratio between training samples and testing samples is 7:3) with different random states. For each child dataset, the training subset and testing subset based on Y-stratified splitting were respectively used for regressor building and performance evaluation, and hyperparameter tuning was implemented using 5-fold GridSearch CV. The best model selected from each child dataset was then aggregated to output a mean E(Mut) value for each variant (Figure 3C). A restricted library carrying two mutations on the finger subdomain (positions 451, 453, 480, 484,485,486) with respect to RF was constructed based on results from preliminary experimental testing and structural insights into KOD functionality. We only included 159 double mutants in this site pool for feature engineering, as structural feature calculation (described above) is computationally intensive. Each variant was assigned a mean predicted E(Mut) ML score by the three ML models (Figure 3C) and about 45 variants (Figure 3E, and Table S2) according to score ranking were then expressed by site-directed mutagenesis and kinetically characterized.

Extensive MD simulations

For the extensive MD simulations[55] aimed at providing support for the mechanistic study of the improved enzymatic efficiency of the polymerase variants, structural models of RF, KI, and KH in apo and binary complex states were constructed with modeler v9.23 based on a crystal structure template. Protein, DNA duplex, and water atoms were respectively treated with ff14SB, OL15, and TIP3P force field parameters. Each model was solvated in an octahedral water box, with Cl^- or Na^+ ions added to neutralize the system. Each solvated system was minimized using gradually decreasing restraint on heavy backbone atoms, heated up to 313K in the NVT ensemble, and equilibrated for 1 ns in the NPT ensemble. For each of RF, KI, and KH, 3 independent 1 - μs MD simulations of both the apo state and the binary state (18 μs in total) were performed using the AMBER18 suite of programs, with the CUDA implementation for GPUs. 30,000 conformations were collected from the three independent MD simulations of

each system.

Representative conformation selection and dynamical cooperative groups analysis.

A uniform parameter-less and unsupervised clustering strategy was employed to perform representative conformation selection and dynamical cooperative group analysis, by which the net information of the whole system *C(L)* is maximized after the assignment of clustering labels L = [Li, ...Lk] [46,45], as following according to equation 2:

$$C(L) = \sum_{i=1}^{R} P_{i,L_i} + \text{penalty term} \qquad \text{(eq. 2)}$$

Here, P_i, L_i is a measure of pairwise structure similarity or dynamical correlation between a non-representative data point *i* and a representative data point L_i. The penalty

term equals -∞ if L_i = k and L_k = k are not satisfied at the same time, which means that only the net information between the (only one) representative data point and the nonrepresentative data points within each cluster is considered. Maximizing the target function *C(L)* ensures that the information of the whole system is preserved as much as possible after the clustering. The parameter-less property of this strategy reduces the possibility of over-interpreting data in our analysis.

In the representative structure clustering, the P_i, L_i is the cosine similarity of Cα atoms of each pair of collected conformations. In the dynamical cooperative group analysis, the P_i, L_i is the normalized covariance of each pair of center-of-mass of a residue and/or a nucleotide across time. Before any analysis, the rotational and translational degrees of freedom were removed by least squares fitting the backbone atoms to a common reference structure of each system. To generate a reasonable number of clusters for explainable analysis, the clustering procedure can be applied to the clustered groups. Bootstrapping of the conformations of each ensemble was employed for error analysis. Python packages mdtraj[56] and scikit-learn[37] were employed to perform the trajectory analysis, covariance and similarity calculation, and the maximizing of *C(L)*.

Maximum flow analysis

The upper limit of information spreading capability between two isolated particles is proportional to the normalized covariance of their center of mass across time. For multiple particles, their communication paths can be characterized with an undirected graph, in which

each pair of particles is connected by an edge with an assigned capability if their normalized covariance is not zero. Here, the center-of-mass of each residue and/or nucleotide is represented as a particle in the graph. The communication of the two mutation sites 451 and 485 with the DNA duplex can be factorized as the superposition of the maximum (or minimum cost) flows from either 451 or 485 to each nucleotide of the DNA duplex in the graph, by assuming that different dynamics signals spreading in the graph are independent. This assumption is made based on a prerequisite that not all the nucleotides must be involved in the dynamic regulation of the T (769) in the binding pocket. To calculate these maximum flows, we employed the push–relabel maximum flow algorithm implemented in Network[57].

Acknowledgment

This study was supported by Shenzhen Engineering Laboratory Molecular Enzymology (Fa Gaishen [2018] No. 958). W.L. Thanks to the National Natural Science Foundation of China, China Grant No. 21505134. Alexander K. Buell thanks the Novo Nordisk Foundation for support (NNFSA170028392). Thanks to China National GeneBank DataBase & BGI's Sequencing Platform and MGI's CoolMPS™ platform.

Author Contributions

Performed the virtual screening: Ruyin Cao. Performed the experiments: Lili Zhai and Qingqing Xie. Contributed reagents/materials/analysis tools: Yue Zheng, Wenwei Zhang, Yuliang Dong. Wrote the paper and SI: Ruyin Cao, Lili Zhai, and Wenping Lu. Reviewed and revised the paper: Lili Zhai, Ruyin Cao, Chongjun Xu, Alexander Kai Buell, and Wenping Lu.

Supplementary Information

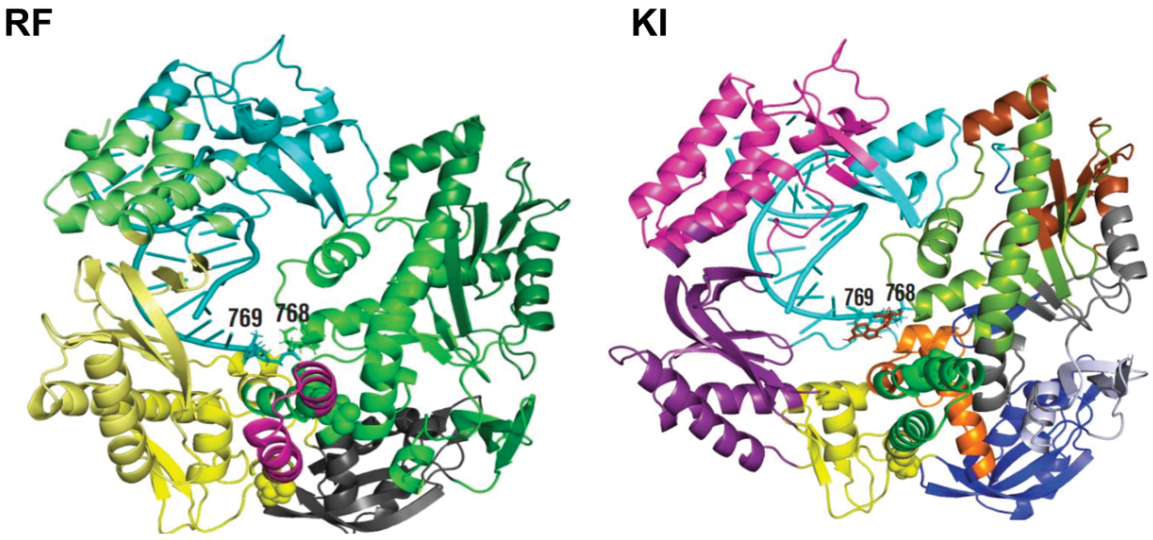


Figure S1. The dynamical cooperative groups (rendered in different colors) of binary RF (left) and KI (right) show that the nucleotide 768 of template DNA is uncooperative to other nucleotides including the core of the binding pocket (769). The two mutation sites 451 and 485 are highlighted as sphere models.

Table S1. Performances of Extra tree regressor (Etr 1-3), Random Forest regressor (RF), and XGBoost regressor (XGB) on the KOD_train dataset. The dataset was divided into three different training/test pairs, and each dataset pair was used to train a regressor with Extra tree, Random Forest, and XGBoost algorithms implemented within Scikit-Learn. Evaluation metrics were shown, including R^2 score and accuracy on training and testing subsets under each data split.

	$R^2_{train\_Etr}$	$R^2_{test\_Etr}$	$R^2_{train\_RF}$	$R^2_{test\_RF}$	$R^2_{train\_XGB}$	$R^2_{test\_XGB}$
train_test_split pair 1	0.74	0.61	0.82	0.49	0.98	0.39
train_test_split pair 2	0.99	0.57	0.90	0.69	0.98	0.36
train_test_split pair 3	1.00	0.60	0.89	0.56	0.98	0.56
	Acc_{train_Etr}	Acc_{test_Etr}	Acc_{train_RF}	Acc_{test_RF}	Acc_{train_XGB}	Acc_{test_XGB}
train_test_split pair 1	0.87	0.75	0.86	0.66	0.96	0.64
train_test_split pair 2	0.97	0.71	0.90	0.50	0.94	0.63
train_test_split pair 3	1.00	0.71	0.56	0.70	0.92	0.70

Table S2. The screening results of 45 KOD variants in virtual screening and experiment screening. For each tested variant, the mutation contents are relative to reference the mutant RF.

E(Mut)_ML predict was predicted using ML-model. E(Mut)_expr was characterized with the kinetic assay.

Variants	Dual site	E(Mut)_ML predict [1]	E(Mut)_expr [2]
RF	A485E	1	1
KI	A485I	/	2.9
KH	A485I + S451L	1.65	31.05
mut_44	A485I + S451T	1.89	23.28
mut_43	A485I + R484C	1.89	22.84
mut_42	A485I + D480T	1.85	14.72
mut_41	A485I + S451G	1.85	12.38
mut_40	A485I + L453A	1.68	12.11
mut_39	A485C + R484V	1.67	10.00
mut_38	A485I + S451N	1.94	9.89
mut_37	A485I + S451H	1.81	9.70
mut_36	A485C + R484I	1.71	9.54
mut_35	A485I + D480S	1.92	7.65
mut_33	A485I + D480Q	1.93	6.18
mut_34	A485I + D480M	1.93	6.18
mut_32	A485I + S451K	1.68	6.15
mut_30	A485I + L453N	1.91	6.09
mut_31	A485I + D480R	1.90	6.09
mut_29	A485I + S451M	1.79	5.98
mut_28	A485I + D480W	2.09	5.44
mut_27	A485I + D480K	1.99	5.03
mut_26	A485I + S451R	1.72	4.96
mut_25	A485I + Q665R	1.73	4.57
mut_24	A485I + Q665S	1.72	4.37
mut_23	A485I + D480E	1.67	4.19
mut_22	A485I + S451A	1.70	3.31
mut_21	A485I + I666P	1.95	3.20
mut_20	A485I + I666T	1.85	3.14
mut_19	A485I + L453T	1.70	2.92
mut_18	A485I + D480C	2.06	2.89
mut_17	A485I + D480G	1.77	2.70
mut_16	A485I + D480P	1.94	2.55
mut_15	A485I + D480F	1.73	2.42
mut_14	A485I + D480H	1.68	2.34
mut_13	A485I + S451P	1.73	2.29
mut_12	A485I + Q665K	1.83	2.27
mut_11	A485I + I666A	2.18	2.14
mut_10	A485I + R484K	1.66	1.74
mut_9	A485I + Q665C	1.81	1.60
mut_8	A485F + I486K	1.75	1.44

mut_7	A485I	+	I666N	2.06	0.86
mut_6	A485I	+	R484D	1.81	0.79
mut_5	A485I	+	I666K	1.76	0.66
mut_1	A485I	+	I666W	1.87	0
mut_2	A485I	+	I666H	1.87	0
mut_3	A485I	+	I666R	1.82	0
mut_4	A485C	+	I486R	1.67	0

(1): The kinetic performance of each KOD variant was predicted by using ML-model. The kinetic was measured by the machine learning-assisted screening method (details in the **Materials and Methods** section).
(2): The kinetic performance of each KOD variant was calculated with equation (eq.1). The kinetic was measured by enzyme kinetic assay (details in the **Materials and Methods** section). 0: No increased fluorescence signal is observed.
The data presented represent the mean values obtained from independent experiments.

References

1. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
2. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci.* **116**, 8852–8858 (2019).
3. Pierce, N. A. & Winfree, E. Protein Design is NP-hard. *Protein Eng. Des. Sel.* **15**, 779–782 (2002).
4. Liu, J. & Cropp, T. A. A method for multi-codon scanning mutagenesis of proteins based on asymmetric transposons. *Protein Eng. Des. Sel.* **25**, 67–72 (2012).
5. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
6. Miller, M., Bromberg, Y. & Swint-Kruse, L. Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Sci. Rep.* **7**, 41329 (2017).
7. Sequeiros-Borja, C. E., Surpeta, B. & Brezovsky, J. Recent advances in user-friendly computational tools to engineer protein function. *Brief. Bioinform.* **22**, bbaa150 (2021).
8. Rapp, L. R. *et al.* Substrate Anchoring and Flexibility Reduction in CYP153A $_{M.aq}$ Leads to Highly Improved Efficiency toward Octanoic Acid. *ACS Catal.* **11**, 3182–3189 (2021).
9. Choi, W. S. *et al.* How a B family DNA polymerase has been evolved to copy RNA. *Proc. Natl. Acad. Sci.* **117**, 21274–21280 (2020).
10. Elshawadfy, A. M. *et al.* DNA polymerase hybrids derived from the family-B enzymes of Pyrococcus furiosus and Thermococcus kodakarensis: improving performance in the polymerase chain reaction. *Front. Microbiol.* **5**, (2014).
11. Kropp, H. M., Diederichs, K. & Marx, A. The Structure of an Archaeal B-Family DNA Polymerase in Complex with a Chemically Modified Nucleotide. *Angew. Chem. Int. Ed.* **58**, 5457–5461 (2019).
12. Aschenbrenner, J. & Marx, A. DNA polymerases and biotechnological applications. *Curr. Opin.*

Biotechnol. **48**, 187–195 (2017).
13. Leconte, A. M. *et al.* Directed Evolution of DNA Polymerases for Next-Generation Sequencing. *Angew. Chem.* **122**, 6057–6060 (2010).
14. Drmanac, S. *et al. CoolMPS ™ : Advanced massively parallel sequencing using antibodies specific to each natural nucleobase*. http://biorxiv.org/lookup/doi/10.1101/2020.02.19.953307 (2020) doi:10.1101/2020.02.19.953307.
15. Huang, J. *et al.* A reference human genome dataset of the BGISEQ-500 sequencer. *GigaScience* **6**, (2017).
16. Eriksen, D. T., Lian, J. & Zhao, H. Protein design for pathway engineering. *J. Struct. Biol.* **185**, 234–242 (2014).
17. McCullum, E. O., Williams, B. A. R., Zhang, J. & Chaput, J. C. Random Mutagenesis by Error-Prone PCR. in *In Vitro Mutagenesis Protocols* (ed. Braman, J.) vol. 634 103–109 (Humana Press, 2010).
18. DeBalsi, K. L., Longley, M. J., Hoff, K. E. & Copeland, W. C. Synergistic Effects of the in cis T251I and P587L Mitochondrial DNA Polymerase γ Disease Mutations. *J. Biol. Chem.* **292**, 4198–4209 (2017).
19. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci.* **116**, 8852–8858 (2019).
20. Mazurenko, S., Prokop, Z. & Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal.* **10**, 1210–1223 (2020).
21. A site-directed mutagenesis method particularly useful for creating otherwise difficult-to-make mutants and alanine scanning. *Anal. Biochem.* **420**, 163–170 (2012).
22. Stemmer, W. P. C. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370**, 389–391 (1994).
23. Bergen, K., Betz, K., Welte, W., Diederichs, K. & Marx, A. Structures of KOD and 9°N DNA Polymerases Complexed with Primer Template Duplex. *ChemBioChem* **14**, 1058–1062 (2013).
24. Nikoomanzar, A., Vallejo, D. & Chaput, J. C. Elucidating the Determinants of Polymerase Specificity by Microfluidic-Based Deep Mutational Scanning. *ACS Synth. Biol.* **8**, 1421–1429 (2019).
25. Wittmann, B. J., Yue, Y. & Arnold, F. H. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst.* **12**, 1026-1045.e7 (2021).
26. Wu, J. *et al.* 3′- *O* -modified nucleotides as reversible terminators for pyrosequencing. *Proc. Natl. Acad. Sci.* **104**, 16462–16467 (2007).
27. Johnson, K. A. The kinetic and chemical mechanism of high-fidelity DNA polymerases. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **1804**, 1041–1048 (2010).
28. Guo, J. *et al.* Four-color DNA sequencing with 3′- *O* -modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc. Natl. Acad. Sci.* **105**, 9145–9150 (2008).
29. Chim, N., Shi, C., Sau, S. P., Nikoomanzar, A. & Chaput, J. C. Structural basis for TNA synthesis by an engineered TNA polymerase. *Nat. Commun.* **8**, 1810 (2017).
30. Hashimoto, H. *et al.* Crystal structure of DNA polymerase from hyperthermophilic archaeon Pyrococcus kodakaraensis KOD111Edited by R. Huber. *J. Mol. Biol.* **306**, 469–477 (2001).
31. Bergen, K., Betz, K., Welte, W., Diederichs, K. & Marx, A. Structures of KOD and 9°N DNA Polymerases Complexed with Primer Template Duplex. *ChemBioChem* **14**, 1058–1062 (2013).
32. Kropp, H. M., Betz, K., Wirth, J., Diederichs, K. & Marx, A. Crystal structures of ternary complexes of

archaeal B-family DNA polymerases. *PLOS ONE* **12**, e0188005 (2017).
33. Nikoomanzar, A., Vallejo, D., Yik, E. J. & Chaput, J. C. Programmed Allelic Mutagenesis of a DNA Polymerase with Single Amino Acid Resolution. *ACS Synth. Biol.* **9**, 1873–1881 (2020).
34. Nikoomanzar, A., Vallejo, D. & Chaput, J. C. Elucidating the Determinants of Polymerase Specificity by Microfluidic-Based Deep Mutational Scanning. *ACS Synth. Biol.* **8**, 1421–1429 (2019).
35. Ong, S. A., Lin, H. H., Chen, Y. Z., Li, Z. R. & Cao, Z. Efficacy of different protein descriptors in predicting protein functional families. *BMC Bioinformatics* **8**, 300 (2007).
36. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
37. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. Preprint at http://arxiv.org/abs/1201.0490 (2018).
38. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
39. Kropp, H. M., Betz, K., Wirth, J., Diederichs, K. & Marx, A. Crystal structures of ternary complexes of archaeal B-family DNA polymerases. *PLOS ONE* **12**, e0188005 (2017).
40. Wynne, S. A., Pinheiro, V. B., Holliger, P. & Leslie, A. G. W. Structures of an Apo and a Binary Complex of an Evolved Archeal B Family DNA Polymerase Capable of Synthesising Highly Cy-Dye Labelled DNA. *PLoS ONE* **8**, e70892 (2013).
41. Gardner, A. F. *et al.* Therminator DNA Polymerase: Modified Nucleotides and Unnatural Substrates. *Front. Mol. Biosci.* **6**, 28 (2019).
42. Lu, H. *et al.* Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **604**, 662–667 (2022).
43. Whelan, S. & Goldman, N. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Mol. Biol. Evol.* **18**, 691–699 (2001).
44. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
45. (Lv), W. L., Arnesano, F., Carloni, P., Natile, G. & Rossetti, G. Effect of *in vivo* post-translational modifications of the HMGB1 protein upon binding to platinated DNA: a molecular simulation study. *Nucleic Acids Res.* **46**, 11687–11697 (2018).
46. Frey, B. J. & Dueck, D. Clustering by Passing Messages Between Data Points. *Science* **315**, 972–976 (2007).
47. Bunzel, H. A. *et al.* Evolution of dynamical networks enhances catalysis in a designer enzyme. *Nat. Chem.* **13**, 1017–1022 (2021).
48. Goldberg, A. V. & Tarjan, R. E. A new approach to the maximum-flow problem. *J. ACM* **35**, 921–940 (1988).
49. Kielkopf, C. L., Bauer, W. & Urbatsch, I. L. Bradford Assay for Determining Protein Concentration. *Cold Spring Harb. Protoc.* **2020**, pdb.prot102269 (2020).
50. Raper, A. T., Reed, A. J. & Suo, Z. Kinetic Mechanism of DNA Polymerases: Contributions of Conformational Dynamics and a Third Divalent Metal Ion. *Chem. Rev.* **118**, 6000–6025 (2018).
51. Ofer, D. & Linial, M. ProFET: Feature engineering captures high-level protein functions. *Bioinformatics* **31**, 3429–3436 (2015).
52. Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid

substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).
53. Šali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
54. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. Preprint at http://arxiv.org/abs/1201.0490 (2018).
55. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **10**, 449–461 (2015).
56. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
57. Yazulla, S. & Schmidt, J. Radioautographic localization of 125I alpha-bungarotoxin binding sites in the retinas of goldfish and turtle. *Vision Res.* **16**, 878–880 (1976).

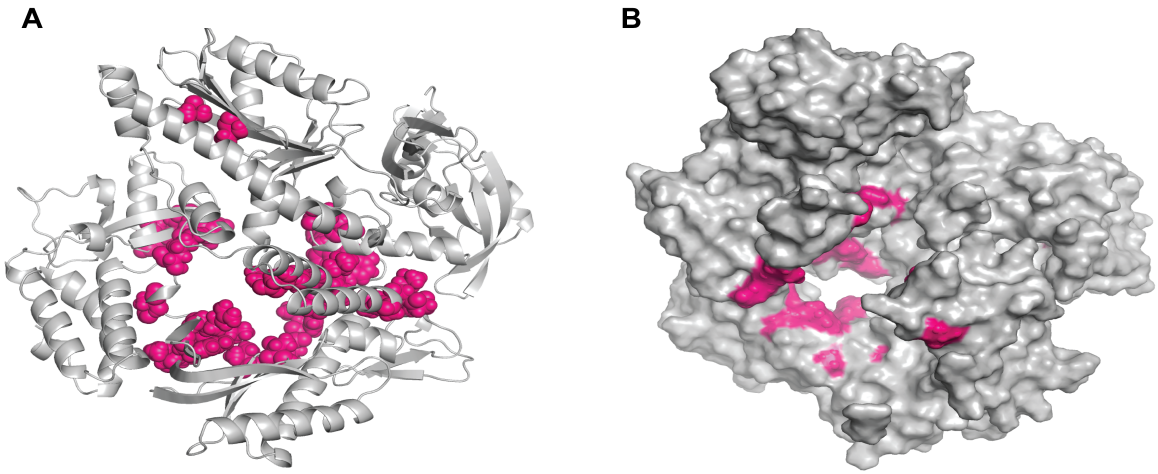
2.4 Summary of KOD DNA Polymerase optimisation studies

In this study, we employed two protein engineering strategies, namely semi-rational and rational design, to engineer a wild-type KOD DNA polymerase (KOD pol) in order to enhance its efficiency in incorporating dye-labeled reversible terminators that are commonly utilized in next-generation sequencing (NGS) applications. We developed a high throughput screening approach based on FRET (Förster Resonance Energy Transfer) technology to identify KOD mutants with enhanced capabilities in incorporating modified nucleotides into the fluorescence-labeled DNA strand. This mutant screening method is employed in both semi-rational and rational screening.

Initially, we employed a semi-rational evolution approach to enhance the catalytic activity of KOD pol. The workflow can be divided into five distinct stages including mutation site selection and sites predicted by molecular dynamics (MD) simulation, library construction, variant screening, stepwise combination, and variant performance testing in the NGS application. The MD method was employed to predict mutation sites in the DNA binding region of KOD pol aimed at improving catalytic efficiency by assessing the binding affinity between KOD pol and the DNA strand, employing the MM-PGSA method. In addition, we identified more than thirty specific sites located within the active pocket and the DNA binding region of KOD pol. A variant named Mut_E10, carrying eleven mutations, was discovered to exhibit successful compatibility with the two DNA sequencing platforms, indicating its significant potential for commercialization as an enzyme. We have applied for patent protection for Mut_E10 (renamed as RF) and these beneficial mutation sites identified in Mut_E10 exhibited the synergy for the incorporation efficiency of modified nucleotides. Our further studies were performed based on the variant RF to continuously enhance its catalytic efficiency towards modified nucleotides through the implementation of rational design strategies. We presented a comprehensive workflow for the rational design of double-site mutants based on the KOD variant RF. The whole workflow of the rational design included three stages: a medium-sized preliminary library construction and screening, the machine learning (ML) model building and virtual library screening, and the experimental validation with kinetic characterization for these top ML-assisted mutants. Through screening ML-assisted mutants, we successfully identified a double mutant KH that exhibited a noteworthy 18-fold improvement in catalytic kinetics compared to its parent enzyme RF. KH carries twelve mutation sites, out of which two sites （A485I and S451L）were identified through rational screening, shown in the following Figures 1D and 1E. The remaining eleven sites, including the A485E mutation, were identified

through semi-rational design in the parent mutant RF. In comparison to the parent variant RF, the KOD variant KH exhibits highly enhanced kinetic performance for modified nucleotides, indicating its potential as a more favorable candidate for future sequencing applications. Future studies would be worthwhile to assess the KOD variant KH on NGS sequencing platforms to evaluate its improvement in catalytic efficiency and to assess its potential commercial value.

Consequently, a total of 37 sites of KOD pol were identified during the whole evaluation process to assess catalytic efficiency for the incorporation of dye-labeled 3'-O-azidomethyl-dATP. Most of these sites were located in both the active pocket and the DNA binding region, as depicted in Figures 1A and 1B. The relevant information regarding the mutated residues of these 37 sites was summarized and presented in the table shown in Figure 1C. Specifically, 32 mutation sites were identified through the semi-rational screening approach, while 9 mutation sites were identified through the rational screening approach. During the entire evaluation process, we evaluated hundreds of KOD variants. Notably, those mutants displaying high catalytic activity consistently possessed multiple combinatorial mutations, often in different structural domains of the protein. This observation strongly suggests the significant role of synergistic effects from multi-site combinations in protein engineering. Therefore, further exploration of residues to continuously enhance catalytic efficiency can be conducted around the neighboring amino acid residues of these identified key positions. Furthermore, in future studies, the rational design approach developed in this study can be utilized to identify potential multiple sites in other regions of KOD pol, specifically the palm domain and thumb domain, with a focus on positions near the active site and the DNA strand, where synergistic functionality may exist.



C

N-terminal	Finger	Palm	Thumb
S347I/M	S451/L/T/G/H/N/K/M/R/A/P	T349I/F	V589H/Q
S348M/T	L453A/N/T	N351H/Y	T590K/L
	D480T/S/Q/M/R/W/K/E/C/G/P/F/H	S383T/I	K592Y/L
	R484C/V/I/K/D	Y384F/W	K593F/W
	A485R/N/D/C/E/Q/G/H/I/L/K/M/F/P/S/T/W/Y/V	E385M/W	Y594V/I
	I486K/R	V389M/I	R606H/N
	I488Q/L	L408A/R/N/D/C/E/Q/G/H/I/L/K/M/F/P/S/T/W/Y/V	E609Y/G
	L489Y/K	Y409A/R/N/D/C/E/Q/G/H/I/L/K/M/F/P/S/T/W/V	R613G/T
	S492E/Y	P410A/R/N/D/C/E/Q/G/H/I/L/K/M/F/S/T/W/Y/V	Q665R/S/K/C
	Y496I/L	A500G/D	I666P/T/A/N/K/W/H/R
	Y499G/L	R501M/L	T676K/Y
		T541I/W	V680M/D

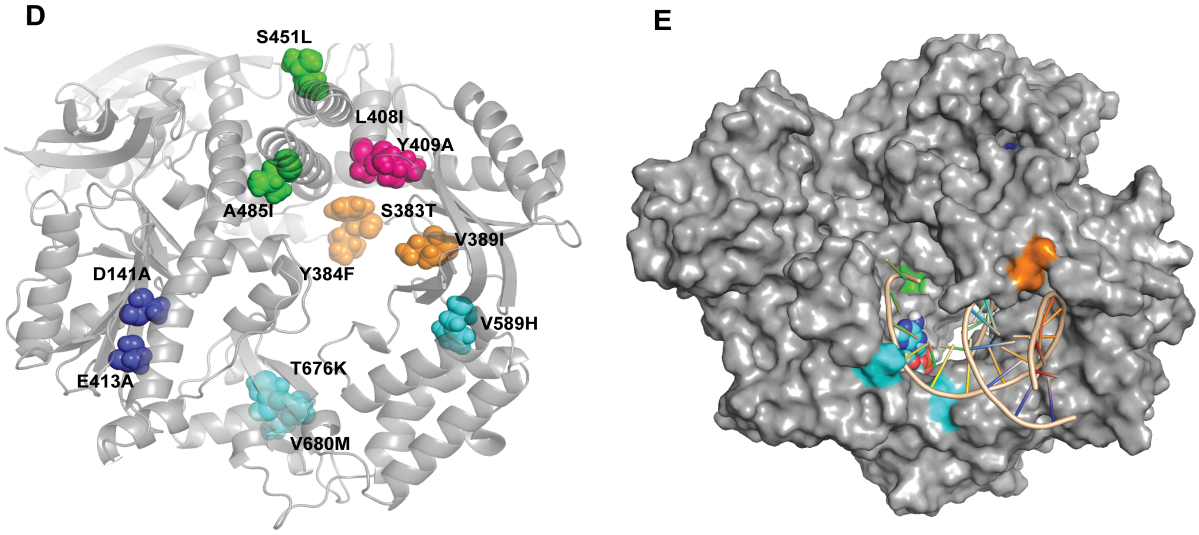


Figure 1. Summary of screened mutation sites of KOD pol. A) The location of 37 sites in the crystal structure of KOD DNA polymerase (PDB ID: 5MOF) is shown during the whole evaluation process. Residues of these mutation sites are represented as spheres, colored in hot

pink. The remaining protein structure is depicted in cartoon format and colored gray. B) Surface view of mutation sites in KOD DNA polymerase structure from A), using the same color scheme. C) The table summarizes the information regarding the residues of the 37 mutation sites. D) The locations of twelve mutation sites from the best-performing variant KH are indicated. These residues are represented as spheres, while the rest of the protein structure is displayed in cartoon format, colored gray. Specifically, D141A and E143A are colored blue, L408I and Y409A are colored hot pink, A485I and S451L are colored green, S383T, Y384F, and V389I are colored orange, and V589H, T676K, and V680M are colored cyan. E) A surface view of the KOD variant KH structure from Figure 1C is provided, using the same color scheme. Additionally, the 3'-O-azidomethyl-dATP is represented as big spheres, colored in elements. The DNA strands are displayed in cartoon format, colored wheat.

In summary, the advantage of our semi-rational screening strategy lies in its integration of experimental screening and computational methods, leading to improvements in targeted characteristics without requiring an in-depth understanding of protein structure and function. The successful engineering of KOD pol with improved catalytic efficiency for modified nucleotides proves the great potential of the semi-rational screening strategy we have developed. In addition, compared to traditional mutagenesis methods, rational screening requires only approximately 12% of the experimental costs while achieving comparable results in terms of training data utilization. In our rational screening, the Extremely Randomized Trees algorithm and Scikit-learn used have complementary features, such as lower bias and higher accuracy. Despite the challenges faced in machine learning-based enzyme design, the advancement in experimental and computational technologies is expected to make prediction methods based on machine learning more reliable. The disadvantages of the screening method can be summarized as the screening throughput is relatively low compared to microfluidic screening techniques, which hinders the ability to perform high-throughput screening of larger libraries within a limited time. Therefore, for future studies, the screening method could consider employing microfluidic-based droplet screening combined with FRET fluorescence as an indicator to overcome limitations in screening throughput. This study significantly expands our comprehension of the interplay between sequence, structure, and functional enhancement in KOD pol. In addition, this work contributes to a deeper understanding of the structural determinants that discriminate modified nucleotides with unique large group specificity in KOD pol. The findings highlight the effectiveness of both semi-rational and rational design strategies, as well as the identification of beneficial mutation sites, serving as valuable guidance for the engineering of other B family DNA polymerases.

3 Chapter II Mechanistic studies of α-synuclein assembly by mRNA display

3.1 Introduction of α-synuclein

α-synuclein (α-Syn) is the primary component of Lewy bodies, which are pathological markers found in all synucleinopathies, including Parkinson's disease (PD), Parkinson's disease with dementia (PDD), and dementia with Lewy bodies (DLB)[82].α-Syn is a naturally disordered and unfolded presynaptic neuronal protein that can form harmful oligomers and amyloid fibrils when it aggregates[83]. α-Syn is a protein consisting of 140 amino acids that is primarily found in presynaptic nerve terminals[84] (Figure 1A). The primary structure of human α-Syn is composed of three distinct regions[83] (Figures 1A and 1B): the amphipathic N-terminal region (residues 1-60), a hydrophobic aggregation-prone 'non-amyloid-β component (NAC, residues 61-95), and a flexible acidic C-terminal region (residues, 96-160) containing a random coil. In the natural state, the N-terminus of the protein is positively charged and contains main mutation sites related to Parkinson's disease[84] (Figure 1A). Although the NAC region plays a significant role in α-Syn aggregation, the majority of familial mutations occur at the N terminus, which illustrates its importance in α-Syn misfolding and aggregation[85]. The C-terminal domain plays a regulatory role in the aggregation of alpha-synuclein[85]. Generally, α-synuclein exists as a monomer under physiological conditions[83]. However, the misfolding and abnormal aggregation of α-synuclein result in the formation of insoluble aggregates and protofibrils (Figure 1D), which are toxic to cells[86]. These insoluble aggregates and protofibrils interact with other molecules such as ubiquitin, neurofilaments, and lipid membranes, leading to the formation of Lewy bodies which are a key pathological feature of Parkinson's disease (PD)[87]. Generally, Parkinson's disease is known to progress through several distinct stages, as shown in Figure 1C. The initial phase of PD and PDD share a similar α-Syn pathology that is influenced by several key factors, including genetic risk factors, aging, and exposome[85]. As aggregation of α-synuclein intensifies over time, as shown in Figure 1C, individuals with PD may progress to develop dementia[85]. The aggregation of α-synuclein can potentially spread to adjacent cells and interact with a diverse range of biological proteins, such as amyloid β and tau, leading to massive neuronal cell death and disease[86]. Therefore, the underlying mechanism of α-Syn aggregation has been crucial in understanding the pathogenesis of Parkinson's disease (PD)[88,89].

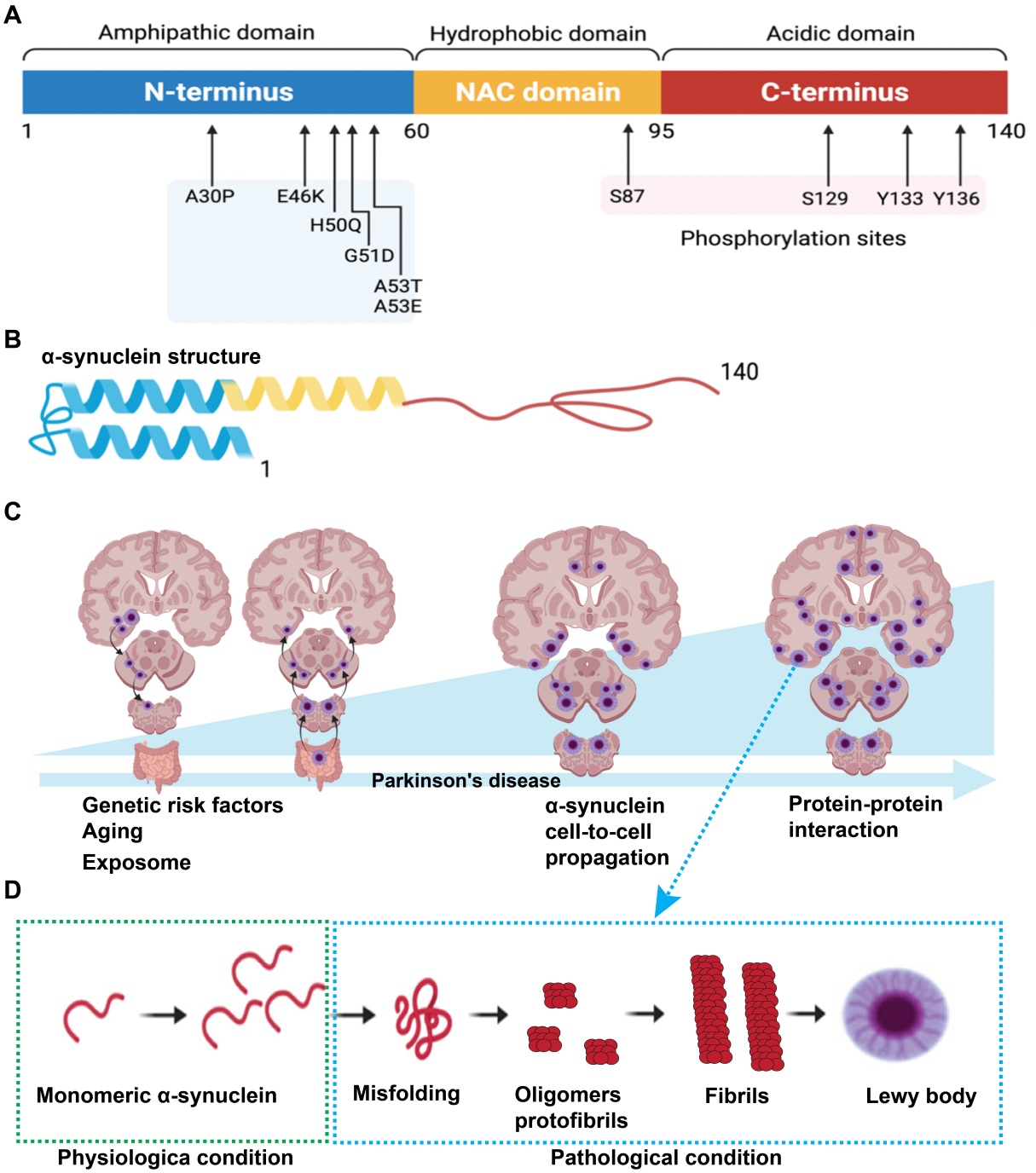


Figure 1. α-synuclein aggregation and Parkinson's disease (PD) pathological[84]. A) The structure of the α-synuclein monomer is depicted, consisting of three distinct regions including the N-terminus (blue), NAC region (yellow), and C-terminus (red). These regions correspond to three distinct domains, including the amphipathic domain (blue) associating main alpha-synuclein gene mutations associated with Parkinson's disease, the hydrophobic domain (yellow) responsible for promoting aggregation, the acidic domain (red) containing the main phosphorylation sites. B) The tertiary structure of the α-synuclein monomer is shown, with each

region colored accordingly. C) The different stages of Parkinson's disease are depicted. The early stage of PD starts with some main factors including genetic risk factors, aging, and exposome. As the global burden of α-synuclein pathology progressively increases over time, as illustrated by the blue arrow and the brain figures, individuals with PD may experience the development of dementia. α-synuclein aggregation can spread from one cell to another, and it can interact with various biological proteins as well as pathological proteins like amyloid βprotein and tau. D) Comparison of α-synuclein in physiological and pathological conditions is presented. In physiological condition: α-synuclein exists as a monomer and is structurally disordered. In pathological conditions: α-synuclein undergoes a series of aggregations, starting with misfolding, leading to the formation of oligomers and protofibrils. These intermediates then progress further to fibrillation, forming insoluble aggregates. In this pathological state, α-synuclein can exhibit toxicity through various mechanisms associated with the pathology of the disease.

Previous studies have extensively studied the mechanisms underlying the aggregation of α-synuclein, such as protein aggregation and structure, synaptic function and neurotransmitter release, functional regulation and interactions, binding properties, and the association of α-synuclein with Parkinson's disease[90]. Reports have shown that the aggregation of α-Syn involved multiple processes to form amyloid fibers, such as primary nucleation, secondary nucleation, fragmentation, and elongation contribute to the growth kinetics of α-synuclein aggregates[91]. In addition, the aggregation process of α-synuclein to form amyloid fiber is characterized by three main distinct phases (Figure 2A): the lag phase, the elongation phase, and the plateau phase[89]. The aggregation growth conditions, such as inhibitors, buffer composition, salt, temperature, etc., significantly affect the formation of α-Syn fibrils[89,92]. Moreover, mutations located in the N-terminus region of α-Syn associated with the early onset of familial PD are also known to modulate its aggregation[93]. These mutations include A30P, E46K, H50Q, G51D, A53T, and A53E, as shown in Figure 1A, and have been identified in individuals with early-onset familial PD[84,94].

In addition, recent studies have revealed that in neurodegenerative diseases, certain neurotoxic proteins, such as α-synuclein, tau protein, and TDP-43, have been found to undergo liquid-liquid phase separation (LLPS) within cells[95]. LLPS plays a critical role in the assembly and organization of complex biomolecular systems[96]. In recent years, LLPS of α-synuclein has been recognized as an alternative nucleation pathway that leads to the formation of dynamic supramolecular assemblies during the early stages of aggregation[97]. The process of α-Syn aggregation in phase separation is associated with three main distinct phases (Figure 2C), including liquid-liquid phase separation, liquid-to-solid phase transition of α-Syn droplets, and

hydrogel formation[98]. The stronger self-association between protein molecules triggers the irreversible transition of these assemblies from liquid to solid, forming amyloid-like hydrogels that may serve as reservoirs, encapsulating toxic low-molecular-weight intermediates and fibrils[97]. In addition, the process of LLPS in α-synuclein has emerged as an alternative non-canonical aggregation pathway, where a crowded microenvironment with high local concentrations promotes the primary nucleation process of aggregation[98]. Mukherjee et al., through a comparative analysis of the conventional misfolding pathway and the non-canonical aggregation pathway, as shown in Figure 2B, provided new insights into the mechanisms of α-synuclein LLPS and amyloid formation[98]. Studies have suggested that abnormal LLPS is a crucial mechanism underlying the aberrant accumulation of α-synuclein and other proteins in neurodegenerative diseases such as PD, and its interaction with iron metabolism disorders is a key driver of ferroptosis in PD[99]. Moreover, current research has revealed that familial mutations of α-Syn can significantly interfere with or regulate its liquid-liquid phase separation (LLPS) dynamics[100]. Specifically, the E46K and A53T mutations have been observed to stimulate α-Syn LLPS, which subsequently leads to its aggregation and precipitation[99]. Therefore, the exploration of the sequence space of α-synuclein has revealed more variants with a propensity for either solution or aggregation in vitro, as well as studying their aggregation or binding properties can contribute to a better understanding of the relationship between α-synuclein and neurodegenerative diseases[101].

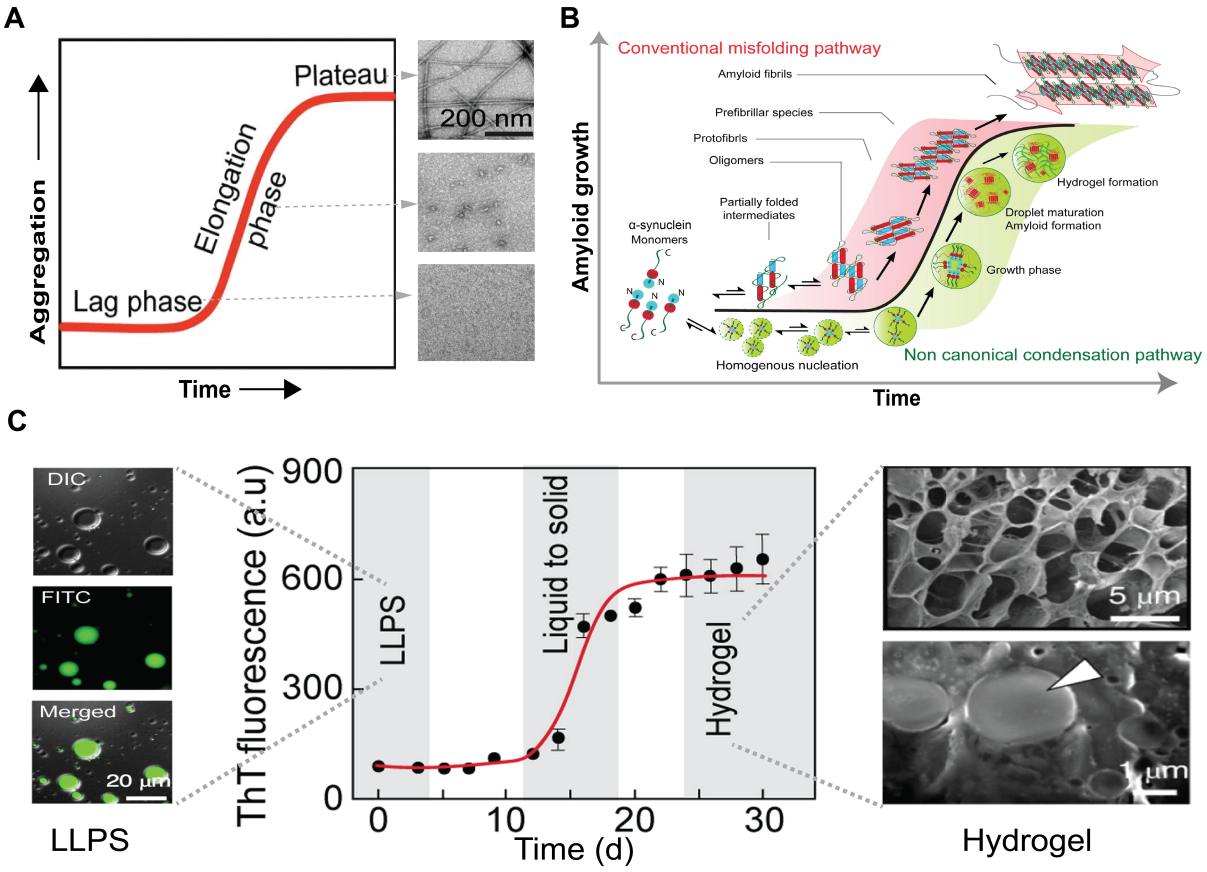


Figure 2. The mechanisms of α-synuclein LLPS and amyloid formation[98]. Scientists have investigated the mechanisms of α-synuclein liquid-liquid phase separation (LLPS) and amyloid formation using various techniques, including time-dependent ThT fluorescence assay, transmission electron microscopy, fluorescence microscopy, and others. The pathways of α-synuclein aggregation can be categorized into the conventional misfolding pathway and the non-fibrillar coacervation pathway. A) The conventional misfolding pathway is depicted, including the lag phase, the elongation phase, and the plateau phase. B) This figure illustrates the comparative analysis of the conventional misfolding pathway and the non-canonical aggregation pathway. C) The non-canonical aggregation pathway is shown, involving liquid-liquid phase separation, followed by the liquid-to-solid phase transition of α-Syn droplets and subsequent hydrogel formation.

3.2 Part I Manuscript 3 - Exploratory experiments to probe the interactions between α-Syn and nucleic acids

This section presents a preliminary study conducted using the mRNA-display approach to explore the sequence space of α-Syn. The objective is to provide some insights into the potential impact of these interactions on the process of α-Syn assembly by mRNA display. We evaluated the binding interaction between α-Syn and various types of nucleic acids, including single-stranded DNA (ssDNA) of different lengths, double-stranded DNA (dsDNA), and genomic DNA, through three parts. We first evaluated the binding affinity of α-Syn for different types of DNA. The results indicate that there is a binding interaction between α-Syn and DNA, and this binding interaction is weak. Next, we evaluated the influence of DNA on the α-Syn fibrillization process. We found that DNA molecules displayed a weak inhibitory effect on the α-Syn fibrillization process. In follow-up studies, we primarily monitored the effect of single and double-stranded DNA on the liquid-liquid phase separation of α-Syn. The results presented the complex coacervation of α-Syn with both ssDNA and dsDNA and ssDNA/dsDNA-PEG formed suspended drops in a larger α-Syn condensate. α-Syn can exhibit a liquid-to-solid phase transition due to aggregation even in the complex coacervates. Consequently, these findings can provide a valuable basis for speculating the effects of the mRNA tail on the subsequent process of α-Syn assembly by mRNA display. In addition, this study can provide valuable insights and implications into the interaction mechanism between nucleic acids and α-synuclein, which is related to the pathogenesis and therapy of Parkinson's disease (PD).

Exploratory experiments to probe the interactions between α-Syn and nucleic acids

Contributions: Soumik Ray and Lili Zhai developed the experimental methodology. Lili Zhai performed the experiments for the binding affinity measurements, the ThT assay, and the phase separation assay on the bench, and analyzed the data. Ray Soumik repeated the phase separation experiments, replenished FRAP experiments, and analyzed the data. Antonin Kunka performed the ANS assay and analyzed the data.

Lili Zhai [1,2], Antonin Kunka[1], Soumik Ray[1], Alexander K. Buell[1]*

[1]Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs. Lyngby, 2800 Denmark

[2] BGI-Shenzhen, Beishan Industrial Zone, Shenzhen 518083, China

*Corresponding author: alebu@dtu.dk

Abstract

α-Synuclein (α-Syn) is a naturally disordered and unfolded presynaptic neuronal protein that can form filamentous aggregates, which are prominent and presumed to be significant etiological factors in Parkinson's disease (PD). Recent research reports indicate that α-synuclein demonstrates DNA-binding properties. However, our understanding of the effect of these interactions on aggregation and phase separation of α-synuclein is limited. In this study, we aimed at elucidating the extent of these effects. We quantified the binding interactions between α-Syn and different types of DNA using flow-induced dispersion analysis (FIDA). The results suggest that the binding affinity is weak and generally in the low tens of micromolar range. Then, we observed the influence of DNA on the amyloid fibril formation of α-Syn. The results demonstrated that DNA has weak inhibitory effects on α-Syn aggregation. Furthermore, we explored the influences of different DNA, including both ssDNA and dsDNA, on the phase separation of α-Syn. The results demonstrated that two distinct types of complex coacervates formed in the presence of ssDNA and dsDNA, respectively. In addition, α-Syn can exhibit a liquid-to-solid phase transition due to aggregation even in the complex coacervates, while both ssDNA and dsDNA remained completely liquid-like after 48h. The present results provide new insight into the interaction mechanism of nucleic acids and α-synuclein related to the pathogenesis of PD.

Introduction

Parkinson's disease (PD) is the second most prevalent progressive neurodegenerative disorder after Alzheimer's disease (AD) with an estimated incidence of about 1 - 4% that is expected to increase as a result of an increasing life expectancy of the population[1,2]. Together with PD, dementia with Lewy bodies (DLB), and multiple system atrophy (MSA) and others, these disorders form a large, heterogeneous group of diseases collectively termed synucleinopathies whose hallmark is the pathological aggregation and intracellular deposition of α-synuclein (α-Syn)[3]. The physiological function of α-Syn includes vesicle trafficking, modulation of dopamine release[4], protection against oxidative stress[5], and transcriptional regulation of specific genes through its binding properties to DNA[6]. Due to the binding properties of α-Syn and DNA, there is an increasing need for a detailed description of its interaction mechanisms

with DNA[6].

Direct interaction between α-Syn and DNA was observed in several model systems both *in vitro* and *in vivo*[7]. *In vivo*, evidence suggests that α-Syn binds unprotected histone-free chromatin DNA and can modulate transcription by binding to several promoters[8]. In vitro, single-stranded circular and supercoiled plasmid DNA were shown to induce different conformational changes in α-Syn, including partial folding and helix formation, and have opposing effects on α-Syn aggregation[9]. Conversely, several studies demonstrate structural changes of DNA induced by α-Syn binding, including stabilization and structural change of GC-rich oligonucleotides from B-DNA to altered B-DNA conformation[10,11], or extension and compaction of DNA under confined conditions (e.g., nanofluidic channel) depending on the degree of α-Syn C-terminal truncations[12]. In addition, Using electrophoretic mobility assays (EMSA), two groups recently reported binding of α-syn with 100 - 500 bp dsDNA in the micromolar concentration range of the protein with increasing affinity in a DNA length-dependent manner[8]. Surprisingly, Jos *et al*., show that the removal of the acidic C-terminus (α-Syn_1-103) results in a complete lack of binding to dsDNA fragments based on the EMSA[8], indicating that the molecular mechanism of the interaction might be more complex than simple charge complementarity between the positively charged N-terminal region and the negative DNA.

However, the comprehensive understanding of the α-Syn/DNA interactions is limited by there being few studies aiming to systematically investigate the dependence of type, length, and sequence of DNA on the binding affinity and aggregation of α-synuclein[8]. The possible reason for such a knowledge gap is arguably two-fold. First, biophysical studies use different model DNA systems that range from a few bp oligos to >100 kb gDNA which differ significantly in their physiochemical and colloidal properties[13]. Second, the effect of DNA on α-Syn aggregation to form amyloid fibrils is typically assessed only at conditions of initially homogeneous monomeric solutions that aggregate into fibrils via a heterogeneous nucleation mechanism involving suitable interfaces[14], whilst its effect on other α-Syn aggregation pathways remains unexplored. Generally, the process of α-Syn aggregation assays in a microplate reader can be effectively investigated by specific amyloidogenic fluorescent markers[15], such as Thioflavin T fluorescence (ThT) and the fluorescent dye 8-Anilinonaphthalene 1-Sulfonic Acid (ANS). However, DNA interacts with these specific amyloidogenic fluorescent markers[16], posing an additional challenge to exploring the impact of DNA on the amyloid fibril formation of α-Syn.

In addition, studies reported that α-Syn can undergo liquid-liquid phase separation(LLPS),

whereby a protein solution spontaneously separates into a protein-rich (dense) and a protein-depleted (dilute) phase[17]. LLPS is a widespread phenomenon associated with the aberrant protein aggregation observed in various neurodegenerative diseases[18]. Moreover, LLPS has been extensively studied to elucidate the fundamental mechanisms underlying phase separation[18] and to construct cell-like structures or bioreactors for investigating the pathological mechanisms and functions associated with neurodegenerative diseases[19]. The latest advancements in LLPS research have demonstrated that it is driven by multivalent protein-protein or protein-DNA interactions, and it plays a critical role in several essential cellular processes[20,21]. The phase behavior of proteins is influenced by various external environmental factors such as temperature, ionic strength, and pH[22]. Additionally, the presence of nucleic acids, including RNA and DNA, can serve as inducers or regulators for the phase separation[23]. Notably, some DNA binding proteins or cationic polypeptides, such as Ddx4 and RP3, undergo LLPS the driving force and mechanism of which is highly modulated by the presence of DNA[24,25].

In this study, we investigate the interaction between α-Syn and four types of DNA including short ssDNA, long ssDNA, dsDNA, and genomic DNA. We first quantified the binding affinity of α-Syn for these types of DNA. Then, we assessed the influence of DNA on the aggregation process of the α-Syn under conditions where the latter forms a homogeneous monomeric solution. The results demonstrated that the binding interaction of α-Syn and DNA is weak. Furthermore, we explored the influence of DNA on α-Syn phase separation. The results demonstrated that in the presence of α-Syn, two distinct types of coacervates formed with ssDNA and dsDNA, respectively. Moreover, whilst initial condensates of both components display liquid-like properties, α-Syn gradually solidifies whereas DNA droplets remain liquid-like. Our results provide new insights into the α-Syn/DNA system, not only crucial for understanding the progression of PD but also for providing information to support the development of novel therapies for neurodegenerative disease.

Results and Discussion

α-Synuclein and DNA interact with micromolar affinity

We primarily focused on investigating the binding affinity of α-Syn for four types of DNA : (i) 8 nt short single-stranded DNA, including ssDNA_$(AT)_4$, ssDNA_$(GC)_4$, and ssDNA_$(GCAT)_2$; (ii) 60 nt long single-stranded DNA, ssDNA_$(AT)_{30}$, ssDNA_$(GC)_{30}$, and ssDNA_$(GCAT)_{15}$; (iii) the corresponding double-stranded counterparts, dsDNA_$(GCAT)_{15}$ and genomic DNA; and (iv) a non-repetitive 30 nt-long single-stranded DNA, $ssDNA_{30}$, which was previously used

in our study on Ddx4 LLPS[24], covering different sequences and lengths. We employed the changes in hydrodynamic radius (R_h) upon the binding interaction measured using flow-induced dispersion (FIDA) analyses. In this approach, we employed both Alexa Fluor 488 labeled α-Syn or DNA molecules as indicators. These indicators enabled us to monitor changes in the R_h of the complex formed by α-Syn and DNA when titrated with either DNA or α-Syn, thereby indicating binding between α-Syn and DNA. We analyzed the data and fitted binding curves using three different binding models (1:1, 1:2, and 1:3) to obtain the K_d values. This experimental design allowed us to investigate the binding interactions between α-Syn and DNA and analyze the effects of varying DNA or α-Syn concentrations on the complex formation.

First, we conducted titration experiments using a constant concentration of α-Syn by employing different DNA concentrations. In the absence of DNA, the R_h (hydrodynamic radius) of the α-Syn monomer was approximately 2 nm (Figure 1A). In the presence of DNA, the R_h of the complex value increased (Figure 1A), indicating that α-Syn interacts with DNA molecules. The results demonstrated that most of the binding curves do not saturate in the concentration range that we were able to explore and thus we only obtained approximate K_d values for equilibrium dissociation constants by fitting the data with three different models (1:1, 1:2, and 1:3). The fitting results of nine DNA sequences using three models, along with the corresponding R^2 values, are summarized in Table 1. The binding isotherms of α-Syn for nine DNA sequences, fitted with a 1:1 model, are displayed in Figure 1A.

Next, we conducted titration experiments using a constant concentration of DNA by employing different α-Syn concentrations. There were three DNA sequences involved, namely ssDNA_$(GCAT)_{15}$, dsDNA_$(GCAT)_{15}$, and $ssDNA_{30}$. Each type of DNA was tested in three independent experiments to ensure the rigor of the data obtained. The data from three sets of independent experiments were fitted using the 1:1 binding model (Figure 1B). The binding affinities were determined as follows: ssDNA_$(GCAT)_{15}$ with a K_d of 12.1 ± 4.8 μM, dsDNA_$(GCAT)_{15}$ with a K_d of 27.1 ± 3.0 μM, and $ssDNA_{30}$ with a K_d of 78.8 ± 7.9 μM, as shown in the upper left corner of Figure 1B. The results demonstrated that the binding between the three types of DNA and α-Syn was not saturated, similar to the results described above, suggesting that the determined K_d values can be considered approximate equilibrium dissociation constants. Interestingly, we observed a decrease of R_h at high α-Syn concentrations (>100 μM) in titrations of ssDNA_$(GCAT)_{15}$ and one replicate of $ssDNA_{30}$ following the initial increase of complex size. This shape of the binding curve is reminiscent of the so-called hook effect (also called the prozone effect), a phenomenon that can sometimes be detected in

immunoassays involving multivalent binding partners[26]. We speculate that the occurrence of such a phenomenon in the present study could be caused by intramolecular interactions of α-Syn molecules, its binding to two or more distinct states of the DNA (e.g., hairpins), conformational changes of α-Syn induced by its binding to the repetitive ssDNA[27], or ssDNA compaction.

Overall, the results demonstrated a significant increase in R_h value when titrated with either DNA or α-Syn, indicating the formation of larger complexes due to the binding interactions between DNA and α-Syn. In a previous study, the binding of α-syn with 100 - 500 bp dsDNA was examined by an electrophoretic mobility assay (EMSA) as a function of DNA length[8]. In the binding model employed in their study, there are three potential binding sites for α-Syn on DNA. In our study, the highest R_h value observed in the binding of α-Syn and DNA ranged from 4 to 6 nm (Figures 1A and 1B). Notably, the R_h value for the individual molecules of α-Syn monomer, ssDNA_$(GCAT)_{15}$, dsDNA_$(GCAT)_{15}$, and $ssDNA_{30}$ is approximately 2.0 nm (Figure 1B), 2.8 nm, 2.5 nm, and 2.4 nm (Figure 1B), respectively. Therefore, the results suggested that the binding model of α-Syn and dsDNA_(GCAT)15 could possibly be 1:1 in our study. In addition, all K_d values range in the micromolar range, as shown in Table 1 and Figure B. Comparatively, in other studies, the binding affinity (K_d value) of α-Syn for DNA aptamers has been reported to be in the nanomolar range[28,29]. The lowest approximate K_d values were found to be in the low tens of micromolar range in our study, which is several orders of magnitude higher than that reported in other studies[28,29]. Therefore, we concluded that the binding interactions between DNA and α-Syn are relatively weak for the DNA sequences investigated in this study.

In addition, a previous report has suggested that different incubation times may have a significant impact on the magnitude of the measured K_d value[30]. In order to observe if different incubation times had an impact on the binding affinity, we performed time-dependent titration experiments. The results show that the incubation time had almost no effect on the binding isotherms in our experiments (Figure 1C). Furthermore, we found that the R_h value tended to increase in the presence of 200 mM NaCl (Figure 1D), which might be attributed to the influence of different salt concentrations on the structural features of α-Syn molecules. A previous study has proposed that the terminal domains of α-Syn can directly interact with Na^+ and Cl^- ions, thereby affecting the structural characteristics of monomeric α-Syn[31]. These findings suggest that the presence of NaCl may induce conformational changes in α-Syn, resulting in an increase in the R_h value. Therefore, in subsequent experiments, we further

investigated the phase separation of α-Syn under conditions with and without 200 mM salt.

Consequently, our experiments clearly demonstrated a weak binding interaction between α-Syn and four types of DNA, and the binding model might be 1:1. Although K_d values obtained from the fitting process may deviate slightly from the true affinities due to factors like the exclusion of the hook effect and incomplete saturation, the overall order of affinities between different DNA types can still be estimated with a high level of confidence. Specifically, the relative differences in the K_d values can provide valuable information regarding the dependence of the binding strengths on sequence features of the DNA. Additionally, it was observed that the K_d values of ssDNA_$(GCAT)_{15}$ and its double-stranded counterpart dsDNA_$(GCAT)_{15}$ were consistently the lowest in our data set when titrated with either DNA or α-Syn. Therefore, these two types of DNA were selected to further investigate the influence of DNA on α-Syn phase separation. In parallel, we explored the influence of different types of DNA that featured the lowest K_d values on the aggregation of α-Syn, despite the challenges associated with investigating the kinetic aggregation process in the presence of DNA.

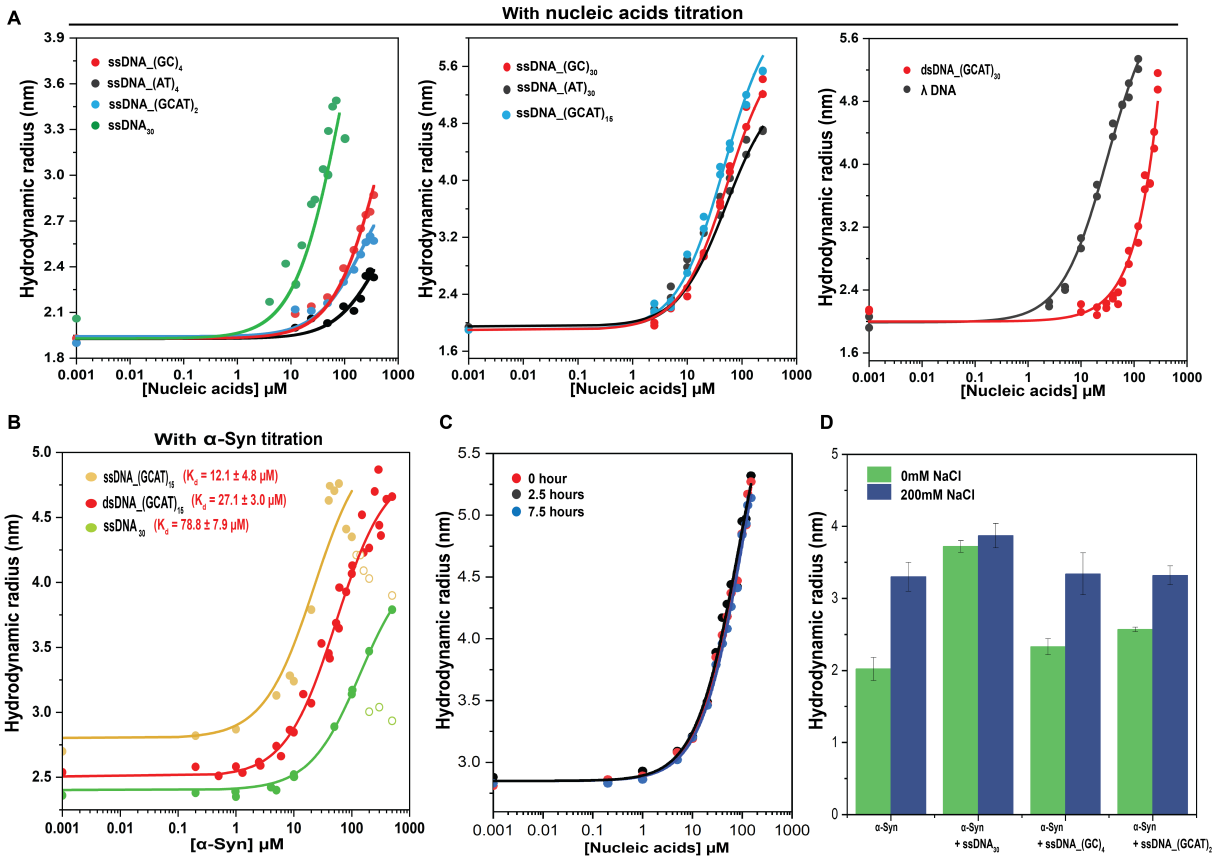


Figure 1. Experimental determination of the binding affinity of α-Syn for different types of DNA. A) Titration experiments using a constant concentration of α-Syn that is titrated with

different types of DNA (2.5, 5, 10, 20, 40, 60, 80, 100, 120, 240 and 300 µM). The measurements involved nine DNA sequences. Left panel: Binding isotherms of α-Syn and ssDNA_$(AT)_4$ (black), ssDNA_$(GC)_4$ (red), ssDNA30 (green), and ssDNA_$(GCAT)_2$ (blue) are presented. Middle panel: Binding isotherms of α-Syn and ssDNA_$(AT)_{30}$ (black), ssDNA_$(GC)_{30}$ (red), and ssDNA_$(GCAT)_{15}$ (blue) are shown. Right panel: Binding isotherms of α-Syn and dsDNA_$(GCAT)_{15}$ (red) and lambda DNA (black) are presented. The fits to the 1 to 1 binding model are shown as the corresponding-colored curves. K_d values are obtained by the fits shown in Table 1. In all experiments, 20 or 30 µM α-Syn spiked with 50 nM Alexa Fluor 488 labeled α-Syn was used. All reactions were conducted in 10 mM phosphate buffer at room temperature. Each experiment was performed with duplicate measurements. B) Titration experiments using a constant concentration of DNA and titrating it with different α-Syn concentrations (0, 0.5, 1, 5, 10, 20, 30, 40, 60, 80, 100, 150, 250, 400, and 500 µM). The measurements involved three DNA sequences: ssDNA_$(GCAT)_{15}$ (orange), dsDNA_$(GCAT)_{15}$ (red), and $ssDNA_{30}$ (green). The fits to the 1 to 1 binding model are shown as the corresponding-colored curves. K_d values obtained from the fits are shown in the upper left corner of the graph. The data points excluded from the fitting (hook effect) are shown as open symbols and colored in the corresponding colors. In all experiments, 3 µM ssDNA or 1 µM dsDNA spiked with 30 nM Alexa Fluor 488-labeled DNA molecules were used. All reactions were conducted in 10 mM phosphate buffer at room temperature. Each experiment was performed with duplicate or triplicate measurements. C) Time-dependent titration experiments were performed by measuring titration curves after different incubation times at room temperature, including 0 hours (red), 2.5 hours (black), and 7.5 hours (blue), of the reaction mixture before the beginning of the measurement. The fits to the 1 to 1 binding model are shown as the corresponding-colored curves. In all cases, 30 µM α-Syn was used and titrated with different DNA concentrations (2.5, 5, 10, 20, 40, 60, and 100 µM) in 10 mM phosphate buffer. D) Changes of R_h of these complexes including α-Syn, α-Syn with $ssDNA_{30}$, α-Syn with ssDNA_$(GC)_4$, α-Syn with ssDNA_$(GCAT)_2$ in the presence of 200 mM NaCl (blue), or absence of salt (green), are shown.

Table 1. The binding affinities according to the analysis of the data with different binding models (1:1, 1:2, and 1:3) of α-Syn on four types of DNA were determined using the FIDA software.

Initial α-Syn Con. (µM)	DNA types	Model 1:1 K_d (µM) [(1)]	R^2	Model 1:2 K_d (µM) [(2)]	R^2	Model 1:3 K_d (µM) [(3)]	R^2
30	ssDNA_ $(AT)_4$	333.0 ± 37.0	0.967	136.7 ± 27.7	0.901	130.5 ± 49	0.871
30	ssDNA_$(GC)_4$	143.0 ± 28.0	0.992	82.1 ± 13.7	0.961	84.7 ± 30.0	0.950
30	ssDNA_$(GCAT)_2$	207.9 ± 21.0	0.995	106.6 ± 17.2	0.977	103.9 ± 33.0	0.943
20	$ssDNA_{30}$	33.4 ± 6.0	0.991	19.9 ± 8.5	0.971	19.4 ± 13.0	0.938

30	ssDNA_$(AT)_{30}$	12.4	± 1.9	0.980	18.8	± 8.8	0.932	20.3 ± 16.0	0.736
30	ssDNA_ $(GC)_{30}$	16.9	± 1.1	0.997	19.5	± 9.2	0.976	21.8 ± 17.0	0.912
30	ssDNA_$(GCAT)_{15}$	13.8	± 0.9	0.997	14.1	± 9.3	0.969	17.3 ± 14.0	0.921
30	dsDNA_$(GCAT)_{15}$	11.5	± 0.1	0.997	11.3	± 4.7	0.982	12.2 ± 9.0	0.967
30	λ DNA	201	± 6.6	0.96	144.1	± 15.9	0.696	123.2 ± 34.0	0.735

(1): K_d value of each DNA sequence was obtained by the 1:1 binding model with equation 3 (n = 1).

(2): K_d value of each DNA sequence was obtained by the 1:2 binding model with equation 3 (n = 2).

(3): K_d value of each DNA sequence was obtained by the 1:3 binding model with equation 3 (n = 3).

The binding affinity was measured by the binding affinity measurements with FIDA (details in the **Materials and Methods** section).

The data presented represent the mean values obtained from duplicated or triplicated independent experiments.

Table 2. Different types of DNA information.

oligo Types	Sequence (5'-3')	length
ssDNA_$(GC)_4$	GCGCGCGC	8nt
ssDNA_$(AT)_4$	ATATATAT	8nt
ssDNA_$(GCAT)_2$	GCATGCAT	8nt
$ssDNA_{30}$	TTTTTCCTAGAGAGTAGAGCCTGCTTCGTGG	31nt
ssDNA_$(GC)_{30}$	GCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCG CGCGCGCGCGCGCGCGCGC	60nt
ssDNA_$(AT)_{30}$	ATATATATATATATATATATATATATATATATATATAT ATATATATATATATATATAT	60nt
ssDNA_$(GCAT)_{15}$	GCATGCATGCATGCATGCATGCATGCATGCATGCA TGCATGCATGCATGCATGCAT	60nt
dsDNA_$(GCAT)_{15}$	GCATGCATGCATGCATGCATGCATGCATGCATGCA TGCATGCATGCATGCATGCAT	60bp
$ssDNA_{30}$ (1)	Alexa Fluor 488- TTTTTCCTAGAGAGTAGAGCCTGCTTCGTGG	31nt
ssDNA_$(GCAT)_{15}$ (2)	Alexa Fluor 488- GCATGCATGCATGCATGCATGCATGCATGCATGCA TGCATGCATGCATGCATGCAT	60nt
ssDNA_$(GCAT)_{15}$ (3)	Alexa Fluor 555- GCATGCATGCATGCATGCATGCATGCATGCATGCA TGCATGCATGCATGCATGCAT	60bp

(1): ssDNA30 was synthesized with the 5' end labeled Alexa Fluor 488.

(2): ssDNA_(GCAT)15 was synthesized with the 5' end labeled Alexa Fluor 488.

(3): ssDNA_(GCAT)15 was synthesized with the 5' end labeled Alexa Fluor 555.

The fluorescence-labeled dsDNA_(GCAT)15 was obtained by annealing one fluorescence-labeled ssDNA_(GCAT)15 with one non-fluorescence unlabeled complementary sequence.

Effects of different types of DNA on α-Syn fibrillization

Next, we performed a characterization to investigate how DNA influences the fibrillization kinetics of α-Syn. It is known that DNA interacts with specific amyloidogenic fluorescent markers, which poses a significant challenge in investigating the kinetics of the aggregation process. Therefore, we performed the experiments using two fluorescence dyes, namely 8-anilinonapthalene-1-sulfonic acid (ANS) and Thioflavin T (ThT), and involved five DNA sequences, namely ssDNA_$(AT)_{30}$, ssDNA_$(GC)_{30}$, ssDNA_$(GCAT)_{15}$, and $ssDNA_{30}$, along with one double-stranded DNA, dsDNA_$(GCAT)_{15}$.

In order to compare any potential interference from the binding of DNA molecules and fluorescent markers, we first performed binding assays between DNA and two fluorescence dyes (ANS and Thioflavin T). α-Syn is known to be able to form distinct fibril morphologies even in a single reaction mixture from initially homogeneous monomeric samples that show variable sensitivity to ThT and potentially other fluorescent dyes[32]. In addition, previous studies have shown that ANS is a suitable probe for α-Syn aggregation that yields very similar data as when ThT is used, in the absence of DNA[33]. We carried out fluorescence assays using a concentration of 50 μM ThT[16] and 100 μM ANS[34], respectively, with titrated $ssDNA_{30}$. The resulting fluorescence curve in the case of ThT exhibited a binding isotherm that reached saturation at approximately 50 μM, with half of the maximum amplitude observed at 3 μM of ssDNA (Figure 2A). Similar to ThT, an increase in fluorescence intensity was observed in the presence of DNA, albeit the effect was much weaker, and the amplitude change was lower, exhibiting linear dependence on DNA concentration (Figure 2A). These findings suggest a stronger interaction between ThT and DNA compared to ANS. In addition, the binding mode and physiochemical properties of ANS are completely different from ThT which is positively charged and binds specifically to the cross-beta sheet structures, typical for amyloid fibrils[35]. Therefore, we also evaluated the time course of ThT and ANS fluorescence intensity using different mixing schemes of DNA, dye, and pre-formed α-Syn fibrils, shown in Figure 2B. We conducted experiments involving 40 μM pre-formed α-Syn fibrils (equivalent to monomer concentration) combined with either 50 μM ThT or 100 μM ANS. These experiments were performed in the presence of buffer alone or in the presence of 30 μM $ssDNA_{30}$. The results presented that the ThT signal in the sample containing DNA was lower compared to the sample with fibrils alone (Figure 2B). However, the signal from ANS was consistent in both the samples with and without DNA in the fibril samples (Figure 2B). Therefore, we speculate that the lack of ThT signal during aggregation of α-Syn in the presence of DNA is caused predominantly by

the depletion of free ThT from solution by the DNA, preventing its binding to the fibrils. Conversely to ThT, the ANS does not seem to be depleted by the DNA to a similar extent, which is expected due to the different physiochemical properties of the two dyes.

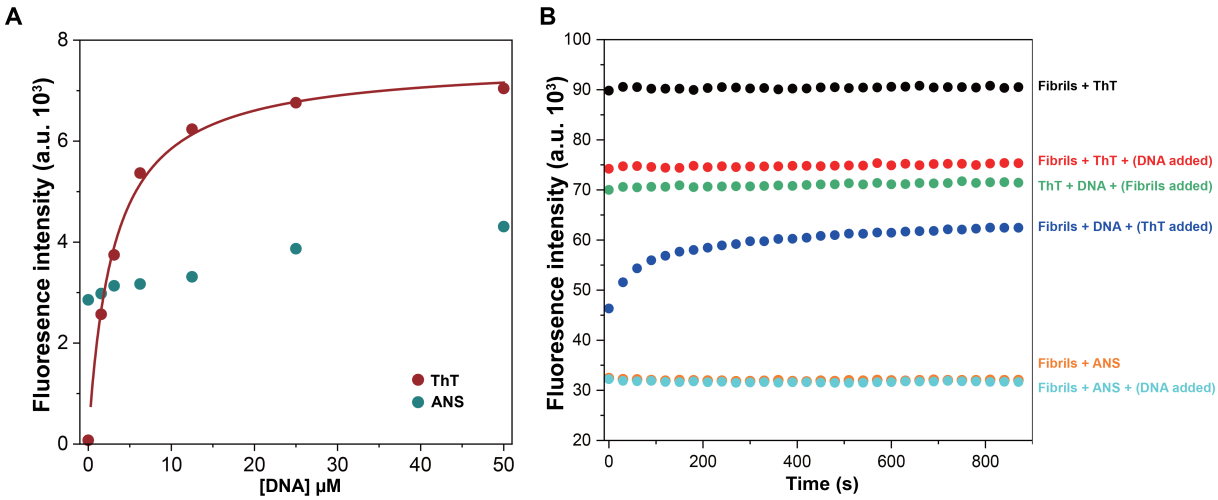


Figure 2. DNA interferes with the fluorescence quantum yield of aggregation-sensitive dyes including 8-anilinonapthalene-1-sulfonic acid (ANS) and Thioflavin T (ThT). A) Fluorescence intensity of 50 µM ThT (red dots) and 100 µM ANS (green dots) as a function of $ssDNA_{30}$ concentration. The measurement was performed at room temperature. B) Time course of ThT and ANS fluorescence intensity upon different mixing schemes of $ssDNA_{30}$, dye, and pre-formed fibrils as indicated on the right side of the graph with different corresponding colors. The measurement was performed at 37 °C.

Subsequently, we investigated the fibrillization process of 100 µM α-Syn monomer using ThT and ANS fluorescence assays in the presence and absence of DNA, and the results are shown in Figure 3. The aggregation of α-Syn to form fibrils is relatively slow at physiological pH due to the presence of a large energy barrier during primary nucleation[36]. Aggregation kinetics are therefore typically measured under shaking conditions in the presence of a glass bead to provide an additional surface promoting heterogeneous nucleation, and to enhance the fibril fragmentation rate[37]. Therefore, in both the ThT and ANS experiments, we employed shaking in the presence of a glass bead to accelerate the aggregation rate of α-Syn monomers in physiological pH. In the absence of DNA, the lag time for α-Syn fibrillization in the ThT experiment was around 20 hours (Figures 3A and 3C), and the lag time in the ANS experiment was around 40 hours (Figures 3B and 3D). In addition, the time courses of the increase in ThT fluorescence intensity were described by empirical sigmoidal curves, which consisted of an initial lag phase, followed by a rapid growth phase, and finally reached a plateau phase[38], consistent with our results (the red-colored binding curve in Figure 3A).

In the ThT experiment, the influence of three different DNA concentrations (30 μM, 10 μM, and 2 μM) on α-Syn aggregation was investigated, as illustrated in Figure 3A. In the presence of 30 μM and 10 μM DNA, the samples demonstrated a pronounced inhibition of fluorescence signal, and the lag time was significantly delayed, as shown in the left panel and middle panel of Figure 3A. Such inhibition observed is surprising because the binding affinity between the DNA and the α-Syn monomer is low. Under the experimental conditions employed, 70 - 90% of the initial monomer concentration should theoretically remain unbound and available for the aggregation reaction based on the K_d values. Therefore, we suspected that the observed inhibition might be attributed to the strong binding interaction of DNA and ThT, as shown in Figure 2A. Additionally, in the presence of 2 μM DNA, the α-Syn aggregation process can be observed by an increase in the fluorescence signal, as depicted in the right panel of Figure 3A. The increased ThT fluorescence signal is indicative of fibril formation. In the parallel ANS experiment, the influence of DNA concentrations of 30 μM and 10 μM on the α-Syn fibrillization process was performed, as depicted in Figure 3B. The results demonstrated that the α-Syn fibrillization process remained similar for both DNA concentrations (Figure 3B), and there was no obvious systematic trend observed in the aggregation lag time with increasing concentrations of DNA (Figure 3D). The results indicate that the influence of DNA on the α-Syn fibrillization process was weak. In addition, we compared the lag time and fluorescence amplitudes of α-Syn fibrillization kinetics at the end of the reaction in the ThT assay (Figure 3C) and the ANS assay (Figure 3D), both in the absence and presence of DNA. In the presence of DNA, both ThT amplitudes and ANS amplitudes of the α-Syn fibrillization process were similarly decreased (Figures 3C and 3D). Moreover, the lag time of the α-Syn fibrillization process exhibited a slight trend prolonged in both the ThT experiment (the longest lag time: about 53 hours) and the ANS experiment (the longest lag time: about 70 hours) in the presence of DNA, compared to the lag time (ThT: 20 hours and ANS: 40 hours, respectively) in absence of DNA. Overall, we can conclude that DNA molecules displayed a weak inhibitory effect on the α-Syn fibrillization process, i.e., prolonged lag time and decreased fluorescence amplitudes. This is consistent with other studies that α-Syn aggregation could be inhibited by DNA aptamers selected from a random oligonucleotide pool[28,29]. These previous results were interpreted as DNA aptamers increasing the lag time of α-Syn aggregation by stabilizing its off-pathway oligomers. However, it should be noted here that the sequences that we investigated in these kinetic experiments were selected from the very small pool listed in Table 1.

This inhibitory action is consistent with the weak, but clearly identified binding interaction between DNA and α-Syn monomers. It is likely that DNA can also interact with other species

along the aggregation pathway, e.g., oligomers and fibrils. Next, we set out to investigate the effects that DNA, in the form of ssDNA_$(GCAT)_{15}$ and dsDNA_$(GCAT)_{15}$, has on the phase separation and subsequent amyloid formation under quiescent conditions.

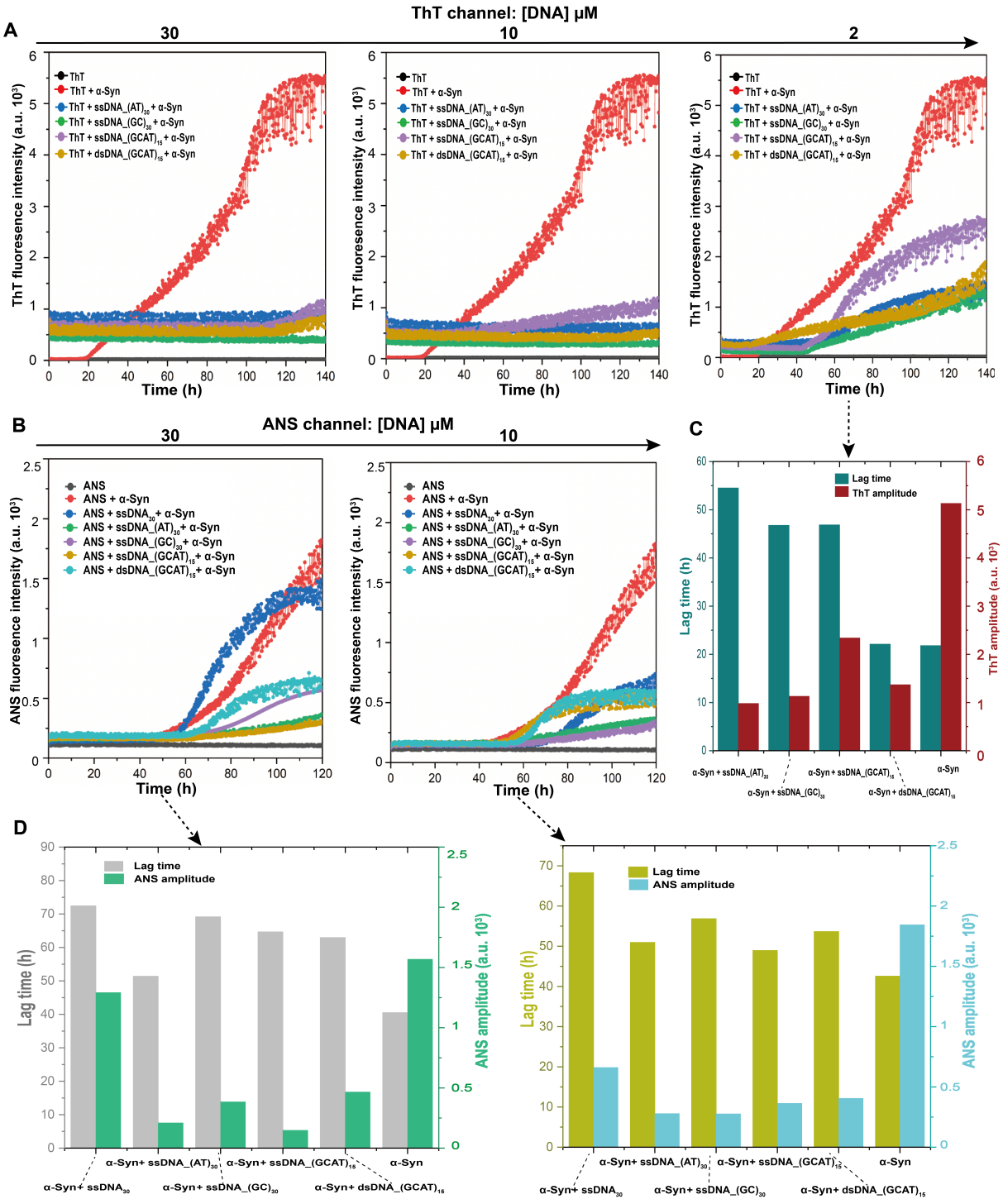


Figure 3. The kinetics of fibril formation of α-Syn monomer both in the absence and presence of DNA was monitored by Thioflavin T (ThT) and 8-Anilinonaphthalene-1-sulfonic acid

(ANS), respectively. A) Left panel: The ThT fluorescence traces for α-Syn fibrillization in the absence of DNA (red), in the presence of four DNA sequences: ssDNA_$(AT)_{30}$ (blue), ssDNA_$(GC)_{30}$ (green), ssDNA_$(GCAT)_{15}$ (purple), and dsDNA_$(GCAT)_{15}$ (yellow), respectively. In these experiments, 100 μM α-Syn monomer, 30 μM DNA and 50 μM ThT were employed in 10 mM phosphate buffer and the aggregation process was monitored at 37 °C for 140 hours with shaking in the presence of one glass bead. The ThT fluorescence traces for the buffer is colored black. Middle panel: The DNA concentration is reduced to 10 μM, while all other measurement conditions remain the same as in the left panel. In the right panel, the DNA concentration is further reduced to 2 μM, while all other measurement conditions remain the same as those in the left and middle panels. B) Left panel: The ANS fluorescence traces for α-Syn fibrillization in the absence of DNA (red), in the presence of five DNA sequences: ssDNA$_{30}$ (blue), ssDNA_$(AT)_{30}$ (green), ssDNA_$(GC)_{30}$ (purple), ssDNA_$(GCAT)_{15}$ (yellow), and dsDNA_$(GCAT)_{15}$ (cyan) are shown. In these experiments, 100 μM α-Syn monomer and 30 μM DNA were employed in 10 mM phosphate buffer and the aggregation process was monitored by 100 μM ANS, at 37 °C for 120 hours with shaking in the presence of one glass bead. The ANS fluorescence traces for the buffer is colored black. In the right panel, the DNA concentration is reduced to 10 μM, while all other measurement conditions remain the same as in the left panel. C) The lag time (green) and the ThT amplitude (red) at the end of the aggregation reactions in the absence and presence of 2 μM DNA are presented as bars. D) In the left panel: The lag time (grey) and the ANS amplitude (green) at the end of the aggregation reactions in the absence and presence of 30 μM DNA are depicted as bars. In the right panel: The lag time (olive) and the ANS amplitude (cyan) at the end of the aggregation reactions in the absence and presence of 10 μM DNA are depicted as bars. The values represent the amplitudes derived from one individual experiment.

DNA and α-synuclein undergo phase separation in the presence of salt and PEG

In order to effectively investigate the phase separation of DNA and α-synuclein *in vitro*, we conducted individual phase separation experiments with α-synuclein, ssDNA_$(GCAT)_{15}$, and dsDNA_$(GCAT)_{15}$, using polyethylene glycol as a crowder in each case. We found that at sodium phosphate buffer pH 7.4 in the presence of 20% w/v PEG-8000, 150 μM α-Syn can undergo phase separation at high salt concentrations, which is consistent with previous findings by us and others[39,40,41,42], shown in the top of Figure 4. Next, we assessed the phase separation regimes of ssDNA_(GCAT)15 and dsDNA_(GCAT)15, including DNA, PEG, and NaCl concentrations. In the absence of salt, only a few small droplets were formed by the 100 μM DNA compared to that in the presence of salt, as shown in Figure 4 Middle. In the presence of 150 mM salt, in the case of both ssDNA_$(GCAT)_{15}$ and dsDNA _$(GCAT)_{15}$, phase separation

was observed, as shown in Figure 4, middle. This is consistent with the findings of other studies, where it has been observed that DNA undergoes phase separation in the presence of hydrophilic polymers such as PEG and salts[43].

Subsequently, we conducted a systematic screening of the phase separation of 150 μM α-synuclein in the presence of varying concentrations of ssDNA and dsDNA, as well as different concentrations of PEG-8000, as depicted in the bottom of Figure 4. As a negatively charged polymer, DNA most commonly undergoes LLPS in response to positively charged proteins or peptides[23]. Previous studies have shown that DNA strongly modulates the phase separation of cationic polypeptides and proteins enriched with positively charged residues, such as histones[44,45,46]. In the absence of such complementary charged macromolecules, high molecular crowder concentrations are needed to provide a sufficient driving force for phase separation. At the highest concentration of PEG used (20% w/v), micron-sized droplets were formed regardless of the DNA concentration and DNA types (Figure 4 Bottom). When 10% PEG was used, the phase separation behavior differed between ssDNA and dsDNA. Droplets formed in the presence of ssDNA_$(GCAT)_{15}$ at concentrations above 20 μM, whereas for dsDNA_$(GCAT)_{15}$, droplets were only observed at concentrations above 100 μM. No phase separation was observed in the presence of 5% PEG for both ssDNA and dsDNA cases.

Consequently, the results indicate that ssDNA_$(GCAT)_{15}$ modulates phase separation of α-Syn more strongly than its double-stranded counterpart, probably due to their different persistence lengths, structures and/or charge densities. To further investigate the differential modulation of the phase separation of α-Syn by ssDNA and dsDNA, we performed phase separation experiments using fluorescently labeled α-Syn, ssDNA_$(GCAT)_{15}$, and dsDNA_$(GCAT)_{15}$.

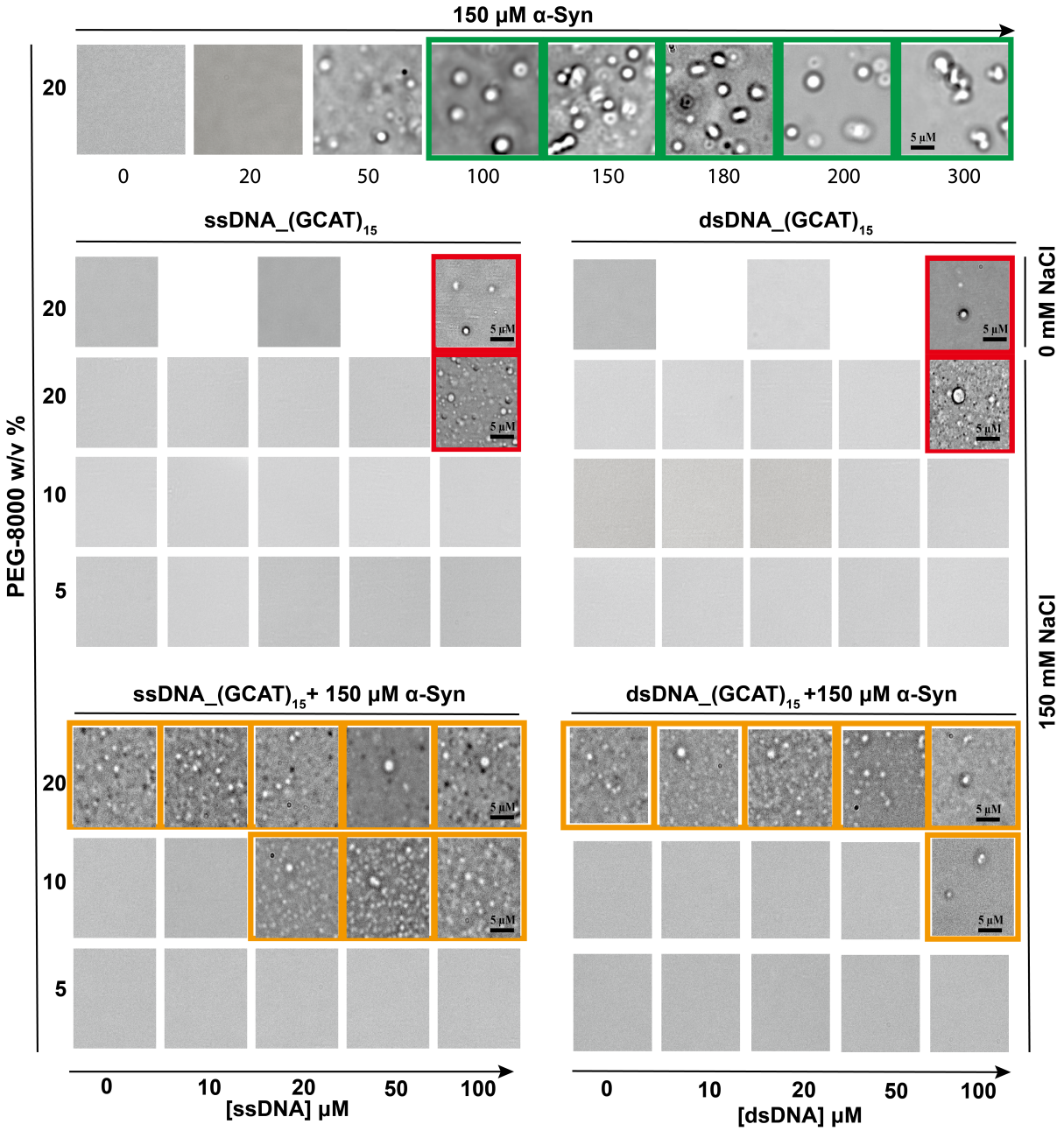


Figure 4. Microscopy analysis of α-synuclein, ssDNA, and dsDNA phase separation regimes. Top: Phase separation regime of 150 μM α-synuclein in the presence of 20% w/v PEG-8000 and varying concentrations of NaCl. Conditions, where droplets are observed, are highlighted by green boxes. Middle: Phase separation regimes of 0-100 μM ssDNA (left) and dsDNA (right) at varying concentrations of PEG-8000 and NaCl indicated on the left and right sides of the panel, respectively. Conditions, where droplets are observed, are highlighted by red boxes. Bottom: Phase separation regimes of 150 μM α-synuclein and 0-100 μM ssDNA (left) and dsDNA (right) in the presence of 150 mM NaCl and varying concentrations of PEG-8000

indicated on the left side of the panel. Conditions, where droplets are observed, are highlighted by orange boxes.

DNA and α-synuclein form two types of segregative coacervate systems

In order to visualize the multi-component condensates under a confocal fluorescence microscope, we performed the phase separation experiments with mixtures containing 100 nM of Alexa488-labeled α- SynA140C[47] and 100 nM of Alexa555-labeled ssDNA_$(GCAT)_{15}$ or dsDNA_$(GCAT)_{15}$. We used 130 µM α-Syn, 20% w/v PEG-8000, 200 mM NaCl in 10 mM sodium phosphate buffer, pH 7.4 to induce phase separation at room temperature, in the presence of 30 µM ssDNA_$(GCAT)_{15}$ and dsDNA_$(GCAT)_{15}$. In these experiments, the sequence of addition of the reaction mixture components was important since we wanted to establish α-Syn and ssDNA/dsDNA interactions prior to the induction of the phase separation. The reaction components were therefore added in the following order: (i) α-Syn (ii) NaCl (iii) ssDNA/dsDNA and finally, (iv) PEG-8000 (Figure 5 B). As a control, we performed α-synuclein phase separation experiments in the absence of DNA, as shown in Figure 5A. The results demonstrated that α-Syn underwent phase separation when 200 mM salt was present (right panel in Figure 5A), while no phase separation was observed in the absence of salt (left panel in Figure 5A). Intriguingly, we found complex coacervation of α-Syn with both ssDNA and dsDNA (Figure 5D). Two types of coacervates could be identified: (i) Type-I coacervates, where α-Syn was specifically localized on the periphery of the α-Syn/nucleic acid assemblies (left panel in Figure 5D), and (ii) Type-II coacervates, where the nucleic acid formed suspended drops inside a larger α-Syn condensate (right panel in Figure 5D). The average size of these multi-component condensates was substantially higher, with a broader size distribution than for those formed by only α-Syn in the presence of PEG (Figure 5E). Strikingly, no complete co-localization of α-Syn and DNA in the coacervate systems was observed irrespective of the type of DNA used. Therefore, it is likely that at this pseudo-equilibrium state, the complex coacervation between the two components is segregative, rather than associative. Phase separation of α-Syn is referred to as a pseudo-equilibrium state because the dilute phase concentration is not stable in time[47]. The major factor causing the progressive decrease in the dilute phase concentration of α-Syn is likely due to aggregation of the dilute phase monomers in contact with the droplets. Nevertheless, it was surprising to see highly negatively charged nucleic acids undergoing segregative condensation without α-Syn. One way of achieving this could be associative phase separation between ssDNA/dsDNA with PEG. To test this possibility, we carried out a parallel experiment with α-Syn and dsDNA in which we spiked the solution

with 10 µg/ml Alexa488 labeled mPEG-5000 and 100 nM of Alexa555 labeled dsDNA in the absence of labeled α-Syn (Figure 5F). Interestingly, we observed co-localization of the dsDNA and mPEG-5000 in type-II coacervates, indicating associative phase separation between nucleic acid and PEG molecules[48] (Figure 5F). In these experiments, we also noted that the ssDNA or dsDNA-PEG formed stable, suspended drops within a larger α-Syn condensate, but never the other way round (i.e., no large DNA condensate contained α-Syn droplets inside them).

Consequently, we observed the formation of two types of segregative coacervates: Type-I coacervates with ssDNA and Type-II coacervates with dsDNA. Next, we set out to investigate whether the presence of ssDNA or dsDNA had an impact on the liquid-to-solid phase transition of the dense phase α-Syn.

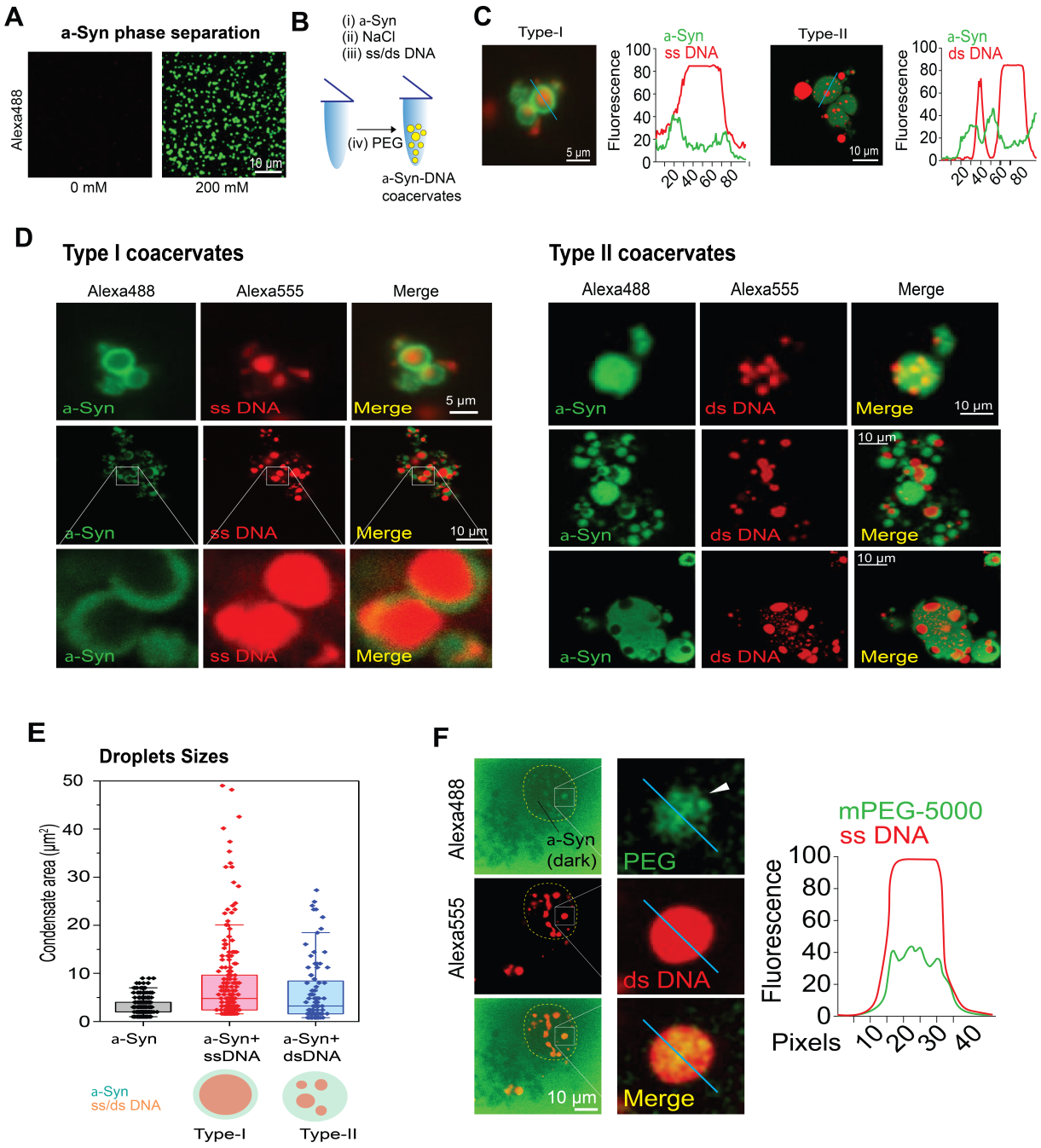


Figure 5. Complex coacervation of α-Syn and DNA in the presence of PEG. A) Phase separation of α-Syn in the presence of PEG 8000. In the absence of salt, no phase separation was observed, and green droplets indicate phase separation in the presence of 200 mM NaCl. B) Schematic illustration of the sequence of addition of α-Syn, NaCl, ssDNA/dsDNA, and PEG into the complex coacervation mix. C) Multi-channel fluorescence microscopy images and respective fluorescence intensity profiles of Alexa488 (α-Syn) and Alexa555 (DNA) in the complex coacervates (type-I and II) are shown. D) Type-I and type-II coacervates are indicated. Left panels: Alexa488 (green, α-Syn), Alexa555 (red, ssDNA), and merged channels showing

spatial partitioning of α-Syn and ssDNA into complex coacervates. Bottom panel: Enlarged image of the type-I coacervates. Right panels: Alexa488 (green, α-Syn), Alexa555 (red, dsDNA), and merged channels showing spatial partitioning of α-Syn and dsDNA into complex coacervates. Type-II coacervates are indicated. E) Size distribution profiles were obtained at similar times after the preparation of α-Syn phase-separated droplets in the absence (left) and in the presence of ssDNA or dsDNA (middle and right). (Lower panel) Schematic describing the spatial organization of α-Syn and DNA in type-I and type-II coacervates. F) PEG partitioning inside α-Syn-DNA coacervates is shown. Fluorescence channels: Alexa488 (green, mPEG-5000), Alexa555 (red, ssDNA), and merged channels showing spatial partitioning/co-localization of PEG and dsDNA into complex coacervates. The yellow dashed circle marks the boundary of a large α-Syn condensate in which the PEG-dsDNA droplets are suspended (Type-II). Right channel: The insets are enlarged for better visualization, and the fluorescence intensity profile of PEG and DNA are plotted across the diameter of one droplet.

Time evolution of the viscoelastic properties is different for individual components in the two types of coacervates

In order to examine the time evolution of the viscoelastic properties for the individual components in the Type-I and Type-II coacervates, we carried out the following procedure. Firstly, we incubated 20 µl of α-Syn, ssDNA/dsDNA, and PEG coacervate solutions in 1.5 ml Eppendorf tubes at room temperature. Subsequently, the incubated samples were examined using confocal fluorescence microscopy and subjected to fluorescence recovery after photobleaching (FRAP) experiments. In these experiments, we investigated the diffusion behavior of both α-Syn and nucleic acids within the complex coacervates. In the microscopy experiments, the type-I coacervates showed aggregated α-Syn morphology and clustering of many α-Syn droplets around them (left channel in Figure 6A). No difference in the spatial organization of the components could be identified in 48 h aged type-II coacervates, although clustering of smaller α-Syn droplets on the surface could be detected (right channel in Figure 6A). Interestingly, we observed a significant decrease in α-Syn fluorescence recovery (ca. 20% recovery with ssDNA and 50% recovery with dsDNA) after 48 h of incubation (Figure 6B). This observation indicates that even within the complex coacervates, α-Syn can undergo a liquid-to-solid phase transition due to its aggregation. Interestingly, both ssDNA and dsDNA remained completely liquid-like even after 48 h (Figure 6B). Subsequently, a multi-channel FRAP experiment was performed where we simultaneously bleached α-Syn and dsDNA condensates in a single, aged (48 h) coacervate. In this experiment, we observed notable fluorescence recovery only for dsDNA but not α-Syn (Figure 6D), in agreement with our previous observations. The liquid-like nature of DNA was further confirmed by detectable

fusion events between two DNA droplets inside an aged type-II coacervate after 48 h of incubation (Figure 6D, lower panel). In addition, previous reports have indicated that newly formed α-Syn phase-separated assemblies can be dissolved at around 60°C[49]. We performed experiments to investigate whether this behavior was influenced in the coacervates when DNA was present. To assess this, we performed temperature-dependent light scattering measurements at 636 nm wavelength on α-Syn droplets both in the absence and presence of ssDNA or dsDNA (Figure 6C). Our results showed no apparent differences in the dissolution behavior of the coacervates of α-Syn in the presence or absence of nucleic acids (Figure 6C).

Taken together, our experiments clearly demonstrate the solidification of a single component in a three-component coacervate system. This finding implies that the concentration enrichment of an aggregation-prone intrinsically disordered protein (IDP) by recruitment into any cellular condensate may trigger its aggregation without affecting the other components of the condensate, in particular in the case of segregative phase separation.

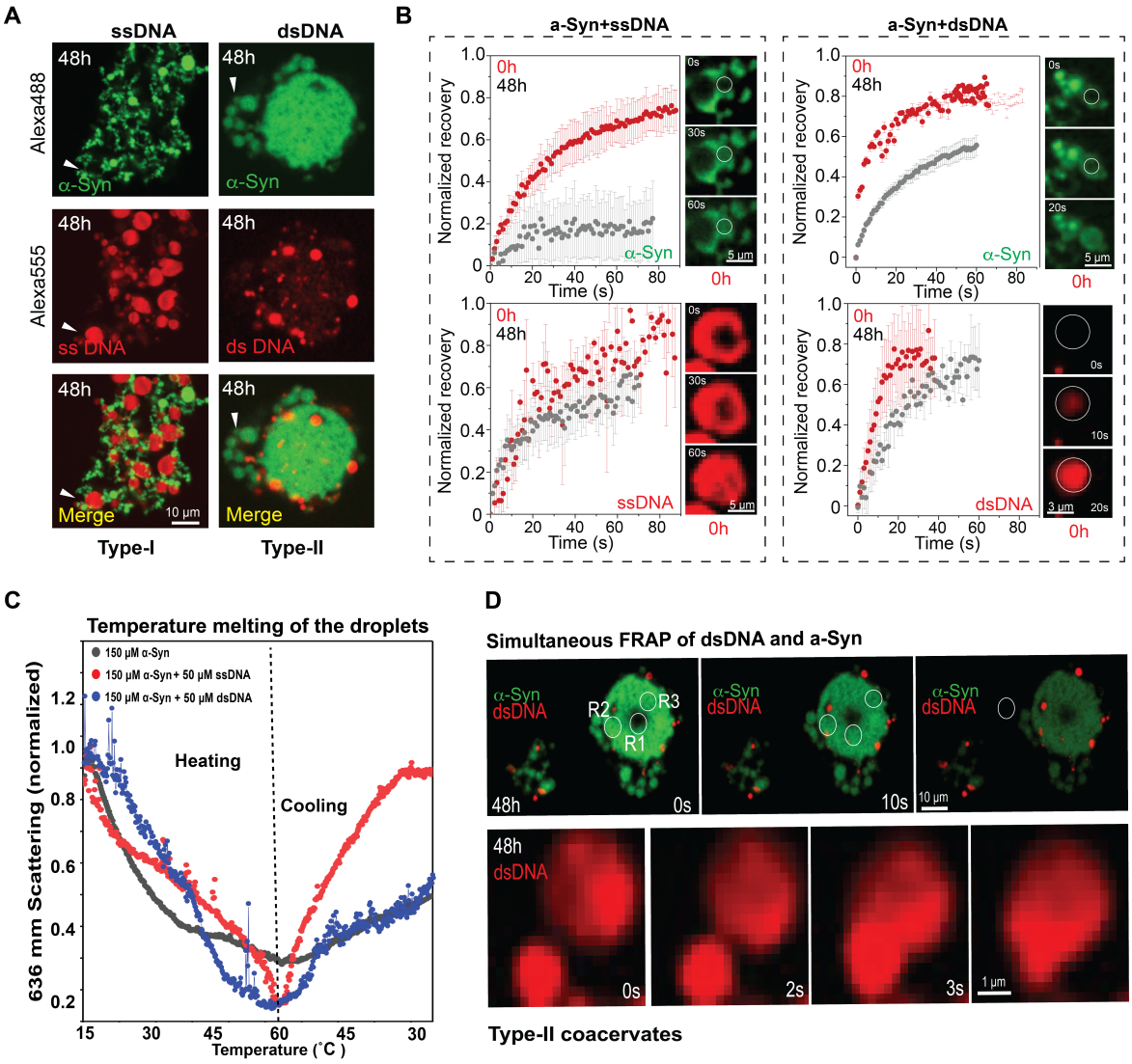


Figure 6. The aggregation of α-Syn in complex coacervates. A) Multi-channel fluorescence microscopy images showing localization of α-Syn and ssDNA or dsDNA in complex coacervates after 48 h of incubation at 25°C. White triangular pointers mark the presence of smaller α-Syn droplets clustered around the aged coacervates. B) (left) The upper panel shows the normalized fluorescence recovery of Alexa488 labelled 140C-α-Syn in a type-I complex coacervate at 0 h (just after formation) and after 48 h of incubation. Photobleaching experiments were performed using 100% laser power (488 nm) for 1s for a given ROI and the framerate of acquisition post-bleaching was kept at 1 frame per second for 80 - 90 s. The right panel shows representative snapshots of α-Syn in a type-I coacervate after photobleaching. The lower panel shows normalized fluorescence recovery of Alexa555 labeled ssDNA in a type-I complex coacervate at 0 h (immediately after formation) and after 48 h of incubation. Photobleaching was performed using 100% laser power (488 nm) for 4s for a given ROI and the framerate of acquisition post-bleaching was kept at 1 frame per second for 80-90s. The right panel shows

representative snapshots of ssDNA in a type-I coacervate after photobleaching. B) (right) The upper panel shows normalized fluorescence recovery of Alexa488 labelled 140C-α-Syn in a type-II complex coacervate. Photobleaching was performed using 100% laser power (488 nm) for 1s for a given ROI and the framerate of acquisition post-bleaching was kept at 1 frame per second for 60-80s. The right panel shows representative snapshots of α-Syn in a type-II coacervate after photobleaching. The lower panel shows normalized fluorescence recovery of Alexa555 labeled dsDNA in a type-II complex coacervate at 0 h (just after formation) and after 48 h of incubation at 25°C. Photobleaching was performed using 100% laser power (488 nm) for 4s for a given ROI and the framerate of acquisition post-bleaching was kept at 1 frame per second for 40-60s. The right panel shows representative snapshots of dsDNA in a type-II coacervate after photobleaching. In all experiments, the values represent the mean ± standard deviation (SD) for n = 3 independent experiments. C) Normalized 636 nm scattering intensity of α-Syn droplets in the absence and presence of ssDNA or dsDNA as a function of temperature. The temperature is shifted from 15 °C → 60 °C → 30 °C. The decrease of scattering intensity upon heating, followed by an increase of scattering intensity upon cooling indicates reversible LLPS behaviors of droplets and coacervates. D) (upper panel) Simultaneous FRAP of α-Syn and dsDNA in a type-II coacervate showing full recovery of dsDNA but partial/little recovery of α-Syn after 48 h of aging. The regions of interest (ROIs) are marked with white circles (a-c). All FRAP curves were corrected for the passive bleaching and background fluorescence of Alexa488 labeled α-SynA140C performed using well-established protocols[24]. The lower panel of the figure displays fluorescence microscopy images capturing a fusion event between two dsDNA droplets within a 48-hour type-II coacervate.

Conclusion and further perspective

In this study, we conducted a three-part investigation into the binding interaction between α-Syn and DNA, including the evaluation of binding affinities, the evaluation of the influence of DNA on the amyloid fibril formation of α-Syn, and the assessment of the effects of DNA on α-Syn phase separation. Initially, we evaluated the binding affinities of four types of DNA, which consisted of a total of nine different DNA sequences. The K_d values were calculated and compared using three fitting models (1:1, 1:2, and 1:3). The results demonstrated that α-Syn binds to all types of DNA, and the most likely binding model appeared to be 1:1, but probably depending on the molecular weight of the DNA employed. Additionally, the molecular weight of lambda DNA is significantly larger than that of α-Syn, therefore, the 1:1 binding model is possibly unsuitable for lambda DNA in this case. In previous studies, it was observed that in the presence of α-Syn, the lambda DNA molecule can be stretched, and the DNA became fully coated with α-Syn protein as visualized by atomic force microscopy (AFM)[50,51]. The authors found that the binding between α-Syn and lambda DNA is weak, because the DNA molecule

exhibited extension with increasing concentration of α-Syn, but the magnitude of the increase in DNA extension was relatively small. In addition, most K_d values were found to be in the tens of micromolar range, which is several orders of magnitude higher than the reported nanomolar range observed by others[28,29]. Based on the comparison with other studies, our findings indicated that the binding interaction between α-Syn monomer and nucleic acids is weak. Furthermore, the results demonstrated that the binding of α-Syn and different DNA sequences seems not to be sequence-specific due to the similar K_d values and R_h values of the complex formed by α-Syn and ssDNA_$(AT)_{30}$, ssDNA_$(GC)_{30}$, and ssDNA_$(GCAT)_{15}$, respectively. Consistently, a previous study reported that the electrostatic interactions of α-Syn and DNA may lead to forming non-specific complexes of α-Syn with DNA[52]. Next, we selected five DNA sequences that exhibited the highest affinities for α-Syn to evaluate the influence of DNA on the amyloid fibril formation of α-Syn. We examined the impact of both high and low DNA concentrations on the process of α-Syn aggregation by employing two chemically very distinct fluorescence dyes, namely ANS and ThT. At high DNA concentrations (tens of μM), almost no aggregation of α-Syn was observed in the ThT assay. However, at low DNA concentrations, we observed varying degrees of delay and reduction in the lag and plateau phases of α-Syn aggregation curves, respectively, in both the ThT and ANS experiments. Based on these observations, we concluded that DNA has weak inhibitory effects on the amyloid fibril formation of α-synuclein. Comparably, in the presence of Thioflavin T (ThT) as a tracer molecule for α-syn fibrillization, double-stranded DNA in the presence of α-synuclein enhances the kinetics of α-synuclein fibril formation and single-stranded circular DNA can delay the aggregation process by approximately 25 hours, as reported by Cherny et al.[53]. However, in our study, we did not observe any promoting effects of dsDNA on the fibrillization of α-synuclein in the ThT assay. The difference in results might be attributed to the molecular size of dsDNA used in Cherny et al.'s study (ranging from 2 kb to 3 kb), which is significantly larger than the 60 bp DNA used in our study. Interestingly, despite differences in type, molecular weight, and conformation, both the ssDNA used in our study and the circular single-stranded DNA used by Cherny et al. show inhibitory effects on the fibrillation of α-synuclein. Finally, we evaluated the influence of two DNA sequences, including ssDNA and its double-stranded counterpart dsDNA, on the phase separation of α-Syn. The results demonstrated that two distinct types of complex coacervates: Type-I coacervates with ssDNA and Type-II coacervates with dsDNA, were formed. The two coacervates displayed strikingly different changes in viscoelastic properties over time, with α-Syn droplets showing decreased FRAP values whilst the DNA-enriched ones remained virtually unchanged. This is arguably a consequence of the α-Syn and

DNA de-mixing in the dense phase and the lack of interactions between them, resulting in almost completely independent phase separation behavior. To the best of our knowledge, this study represents the first description of such DNA/α-Syn complex coacervation, which presents a highly intriguing and potentially biologically relevant system.

However, the underlying mechanisms of this binding interaction could be highly complex, unclear, and challenging to study due to the complex aggregation process of α-Syn and the involvement of various types of interactions between α-Syn monomers, oligomers fibrils, and DNA[52,53]. In addition, our study on the underlying mechanism of α-Syn and DNA interactions does have some limitations, such as the lack of clear saturation in the binding curves of α-Syn and DNA, and the potential interference of the binding of DNA with the fluorescent dyes used to monitor α-Syn aggregation. The exact mode of binding, as well as the stoichiometry of the interaction between DNA and α-Syn likely depends on the length of the DNA, but our binding data are not of sufficiently high quality to allow a more detailed analysis of the binding process. Future studies could continue to reveal the binding mechanism and related molecular interactions between DNA and α-Syn, further uncovering the binding sites and patterns between DNA and α-Syn. Furthermore, the use of customized amyloidogenic fluorescent markers that do not interact with DNA could be employed to monitor the aggregation process. Lastly, in our current study, we have focused on the influence of DNA on α-Syn phase separation, but other environmental factors such as temperature, ionic strength, and pH might also impact this interplay. Future research could consider the influence of these environmental factors on DNA/α-Syn complex coacervation and explore their relationships with phase behavior and disease relevance.

Overall, we can conclude that the binding interaction between α-Syn monomer and DNA molecules was weak, and the influence of DNA molecules on the α-Syn fibrillization process or α-Syn aggregates in phase separation was low. This study provides new molecular insight into understanding the binding interaction between α-Syn and DNA. This study expands our understanding of protein-nucleic acid interactions in the context of amyloid fibril formation and phase separation, scenarios that are crucially important for the understanding of neurodegeneration.

Materials and Methods

Reagents and chemicals

The pT7-7 plasmid carrying the WT human α-synuclein gene (Addgene plasmid #36046) was

used for recombinant expression of wild-type human α-synuclein (UniProt ID P37840), and the gene was sequenced. Sodium phosphate buffer, Tris, NaCl, HCl, NaOH, Benzonase, PEG-8000, DTT, LB broth, ampicillin, ANS, and ThT were all purchased from VWR (Denmark). Three types of short ssDNA by Muralidhar L.Hegde[54] with the sequence ssDNA_$(GC)_4$, ssDNA_$(AT)_4$, ssDNA_$(GCAT)_2$, and one medium by Nott et al.[55] with the sequence $ssDNA_{30}$ as well as the Alexa Fluor 488 labeled version, two types of long ssDNA: ssDNA_$(GC)_{30}$, ssDNA_$(GCAT)_{15}$ as well as the 5' end Alexa Fluor 555 labeled version and 5' end Alexa Fluor 488 labeled version, were synthesized with HPLC purification by TAG Copenhagen (Denmark), and the sequence information was exhibited in Table 2. The fluorescence-labeled dsDNA_(GCAT)15 was obtained by annealing one fluorescence-labeled ssDNA_(GCAT)15 with one non-fluorescence unlabeled complementary sequence synthesized with HPLC purification by TAG Copenhagen (Denmark). Lambda DNA was ordered from Merk (USA). Other relevant salts, buffer components, and materials in this study were purchased from Sigma (USA) and VWR (Denmark) and dissolved in Milli-Q water. Fluorescently labeled α-Synuclein-A140C-Alexa Fluor 488 was a kind gift from Prof. Celine Galvagnion. We bought BL21(DE3) capable E. coli cells from New England Biolabs (USA).

Wild type α-Synuclein

MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATV
AEKTKEQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEG
ILEDMPVDPDNEAYEMPSEEGYQDYEPEA

α-Syn expression and purification

A single colony of BL21(DE3) *E. coli* cells transformed by the plasmid encoding the WT human α-Syn gene was used to inoculate 10 mL LB media supplemented with ampicillin (Amp, 100 µg/mL final concentration). Following the overnight growth at 37 °C (ca. 12 hours), 1 mL of the starter culture was used to inoculate 1 L of fresh LB (100 µg/mL Amp) media. Protein expression was induced when the cell density of the 1L culture OD_{600} reached 0.8 using IPTG with 1 mM final concentration. After the 4 hours of growth at 37 °C under shaking conditions with 180 rpm, cells were harvested by centrifugation (4 °C, 8,000 X g, 15 min) and stored at - 20 °C. Next, cells were thawed and resuspended in 20 ml of 10 mM Tris-HCl, 1 mM EDTA, pH 8.0 with 1 mM protease inhibitor PMSF, and then sonicated on ice in 10 s intervals of 40% amplitude with 30 s pauses between each cycle for 2 minutes. Next, 1 µL of Benzoase (Merck) was added to remove nucleic acid, and the cells were centrifuged at 20,000 X g for 30 minutes

at 4 °C. Resulting cell-free extract was boiled for 20 minutes and centrifuged (20,000 X g, 20 min, 4 °C). α-Syn was precipitated by addition of saturated ammonium sulfate solution (4 mL of 4 M $(NH_4)_2SO_4$ per 1 ml of supernatant). The solution was incubated for 15 minutes on a stirring platform at 4 °C followed by centrifugation at 20,000 X g for 20 min at 4 °C. Pellet was resuspended in 7 ml of 25 mM Tris-HCl pH 7.7 supplemented with 1 mM DTT and dialyzed two times against the same buffer. α-Syn was followed purified on a HiTrap Q HP 5 ml column (Cytiva, USA), followed by size exclusion chromatography (SEC) using HiLoad 16/600 Superdex 200 pg. column (Cytiva, USA). The monomeric α-Syn eluted by 10 mM of sodium phosphate buffer (pH 7.4) and stored at -80 °C until use. Spectrophotometric analysis with the theoretical molar extinction coefficient (ε280 nm) of 5,960 M^{-1} cm^{-1} was employed in order to ascertain the protein concentration.

Binding affinity measurements with FIDA

The binding affinity of α-Syn and DNA was performed in two experiments by titration with DNA or α-Syn, respectively. Firstly, titration assays were done with 20 or 30μM α-Syn spiked with 20 - 50 nM Alexa Fluor 488 labeled α-Syn molecular by applying different DNA concentrations (2.5, 5, 10, 20, 40, 60, 80, 100, 120, 240 and 300 μM). Then, titration assays were done with 3 μM ssDNA or 1 μM dsDNA spiked with 30 nM Alexa Fluor 488 labeled DNA molecular by applying different α-Syn concentrations (0, 0.5, 1, 5, 10, 20, 30, 40, 60, 80, 100, 150, 250, 400, and 500 μM). Time-dependent titration assays were developed by adding different incubation times (0, 2.5, and 7.5 h) of the mixture before the beginning of the measurement at 25°C. In all cases, 30 μM α-Syn was used by applying different DNA concentrations (2.5, 5, 10, 20, 40, 60, and 100 μM). All these experiments were performed in 10 mM phosphate buffer at room temperature. The experiments of effects slat on R_h were performed including α-Syn, α-Syn with $ssDNA_{30}$, α-Syn with $ssDNA\_(GC)_4$, α-Syn with $ssDNA\_(GCAT)_2$. In these experiments, both in the absence and presence of 200 mM NaCl, 20 μM α-Syn spiked with 20 nM Alexa Fluor 488 labeled α-Syn molecular, 80 μM $ssDNA_{30}$, 350 μM $ssDNA\_(GC)_4$, 350 μM $ssDNA\_(GCAT)_2$, respectively.

The quantification of the binding affinities between various DNA constructs and α-Syn was carried out using a FIDA instrument equipped with the light-emitting-diode-induced fluorescence detection with excitation and emission wavelengths of 480 nm and > 510 nm, respectively (Fida Biosystems ApS, Denmark). Flow-Induced Dispersion Analysis (FIDA), a fixation-free technique often is used for describing and quantifying biomolecular interactions

and protein concentrations in their natural conditions. The standard capillary with an inner diameter of 75 µm, a total length of 100 cm, and a distance from the sample inlet to the detection window of 84 cm (Fida Biosystems ApS, Denmark) was used for all experiments.

The hydrodynamic radii (R_h) of α-Syn, various DNA constructs, and complexes of DNA and α-Syn were measured and determined in the 10 mM sodium phosphate buffer pH 7.4 in the present and absence of salt by using the Taylor dispersion analysis (TDA)[56,57]. A small volume (10 nl) of sample labeled with Alexa Fluor 488 was loaded in the capillary filled with the buffer. The methods are referenced from previous studies[58,59] and shown in Table 3. The dispersion profile of the sample was fitted using the built-in Gaussian function to determine the diffusivity (D) value. This diffusivity value was then recalculated to obtain the hydrodynamic radius (R_h) using the Stokes-Einstein equation[60]:

$$R_h = \frac{k_B T}{6\pi\eta D} \qquad \text{(eq. 1)}$$

where k_B is the Boltzmann constant, T is the absolute temperature and η the viscosity.

Table 3. Samples were analyzed by FIDA using the following methods.

Steps	Solution	Pressure	Time
NaOH wash	1 M NaOH	3500	60
H_2O wash	H_2O	3500	60
Buffer wash	10 mM phosphate buffer	1500	40
Sample injection	α-Syn/DNA mixture	75	20
Sample analysis	10 mM phosphate buffer	1500	75

The binding affinities of α-Syn and different DNA constructs were assessed from the changes of the R_h measured using the Flow-induced dispersion analysis (FIDA). For each binding experiment, samples containing fixed concentrations of labeled indicator (i.e., DNA-Alexa Fluor 488) and increasing concentrations of the unlabeled binding partner (analyte, i.e., α-Synuclein) were prepared. The pre-mixed samples were applied to the capillary and the apparent R_h of the indicator-analyte complex was determined as described in the previous section. The analyte binding results in increased diffusivity, broadening of the indicator's dispersion profile, and increased R_h. The plots of analyte concentration versus the apparent R_h corrected for the viscosity changes by the software were fitted to one of the three binding models (1:1, 1:2, and 1:3) available in the FIDA software that can be summarized by the following equations:

$$nA + I \overset{K_d}{\leftrightarrow} A_nI \quad \text{(eq. 2)}$$

$$R_h = (1 + [A]^n/K_d)/\left((1/R_I - 1/R_{IA}) + (1 + [A]^n/K_d)/R_{IA}\right) \quad \text{(eq. 3)}$$

where n is the number of analytes (A) molecules (n = 1,2 or 3) that bind to one molecule of the indicator I, R_h is the experimentally determined hydrodynamic radius, *n* is the stoichiometry of the binding, [A] - analyte concentration, K_d - dissociation constant, R_I - hydrodynamic radius of the indicator, and R_{IA} hydrodynamic radius of the analyte-indicator complex.

Liquid-liquid phase separation and correlative microscopy experiments

The phase separation of α-Syn was formed with over 100 µM of α-Syn in 10 mM phosphate buffer, 150 - 200 mM NaCl, 20% (w/v) PEG-8000, at room temperature. The samples were incubated at room temperature to investigate the droplets' formation and maturation. All the experiments were carried out under the same conditions.

Using an inverted confocal fluorescence microscope (LMI-005-Leica Microsystems Confocal Microscope SP8), all microscopy and fluorescence recovery after photobleaching (FRAP) experiments were carried out and recorded at X 60 oil immersion magnification. An excitation wavelength of 488 nm and an emission channel of 500 - 600 nm; an excitation wavelength of 555 nm and an emission channel of 520 - 620 nm, and the fluorescence merge channels were used for all labeled proteins/DNA used in our study. These images were obtained with a 16-bit depth in a resolution of 1024 X 1024. For FRAP, a bleaching radius of 1 - 2 µm was chosen depending on the a-Syn/nucleic acid coacervate size. The coacervates were photobleached with a 488 nm laser at 100% power for 1 - 4 s. The fluorescence recovery was recorded for 30 - 90 s (post-bleach) and subsequently corrected for the effect of passive bleaching and the background fluorescence following previously established protocols[40]. The fluorescence recovery (post-bleach) was normalized with respect to the pre-bleach fluorescence intensity for individual coacervates. All FRAP experiments were carried out at room temperature. The images were subsequently processed and analyzed using ImageJ (NIH, USA)

Analysis of the size distribution of phase droplets use microscopy (LMI-005-Confocal Microscope - SP8, Leica Microsystems, Germany). 5 µl of phase-separated samples including 150 µM α-Syn, 150 µM α-Syn with 50 µM ssDNA, and 150 µM α-Syn with 50 µM dsDNA were dropped onto a clean glass slide immediately after LLPS at room temperature. These droplets were captured using a 60 X oil immersion, a 488 nm excitation wavelength, and a 500 - 600 nm emission channel. The photos were captured with a 16-bit depth with 2048 X 2048

resolution. Using ImageJ, the size of n is over 100 droplets per sample was determined (NIH, USA).

Temperature effects of droplets were monitored using a multichannel spectrophotometer ProbeDrum equipped with temperature control (ProbationLabs, Lund, Sweden). Fluorescence emission spectra excited at 280 and 509 nm were monitored simultaneously with 636 nm absorption spectra as the temperature of a 60 µL sample placed in a high-precision Cell Quartz glass cuvette with a 3 X 3 mm optical path (Hellma Analytics, Germany) increased from 15 to 60 °C and decreased back to room temperature at constant 2 °C/min rate. Initial data analysis was performed using Origin 2021 (Origin Laboratories, USA).

Aggregation kinetics monitored by fluorescent markers

Kinetics of α-Syn aggregation in the presence of different types of DNA was monitored using two amyloid-sensitive dyes: Thioflavin T (ThT) and 8-anilinonapthalene-1-sulfonic acid (ANS). The experiment was carried out in 96-well half-area non-binding plates (Corning, USA) with 100 µM α-Syn, and 2 - 30 µM DNA in 10 mM sodium phosphate buffer pH 7.4. Samples were measured in duplicates of 100 µL per well. A glass bead (2 mm diameter, BioSpec) was added to each well, and plates were sealed with adhesive sealing sheets (Thermo Fisher Scientific, USA). The fluorescence was excited at 350 nm (ANS) and 440 nm (ThT) and monitored at 480 nm using FLUOstar Omega plate reader (BMG Labtech, Germany) for 120 or 140 hours at 37 °C in 10-minute intervals with continuous orbital shaking (300 rpm) between the readings.

Titration experiments for the binding interaction of DNA molecules and fluorescent markers were performed using a constant dye concentration (100 µM ANS and 50 µM Thioflavin T) by employing different $ssDNA_{30}$ concentrations (0, 1.5, 3.0, 6, 12.5, 25, and 50 µM). The measurements were investigated by FLUOstar Omega plate reader (BMG Labtech, Germany) at room temperature.

Time course experiments for ThT and ANS fluorescence intensity upon different mixing schemes of $ssDNA_{30}$, dye, and pre-formed fibrils were performed. We conducted experiments involving 40 µM pre-formed α-Syn fibrils (equivalent to monomer concentration) combined with either 50 µM ThT or 100 µM ANS. The measurements were investigated by FLUOstar Omega plate reader (BMG Labtech, Germany) at 37 °C for 15 min.

Acknowledgment

This study was supported by the Novo Nordisk Foundation (NNFSA170028392). This study was also supported by Shenzhen Engineering Laboratory Molecular Enzymology, the National Natural Science Foundation of China, China Grant No. 21505134.

Author Contributions

Performed the experiments and analyzed the data: Lili Zhai, Soumik Ray, and Antonin Kunka. Contributed reagents/materials/analysis tools: Alexander Kai Buell. Wrote the manuscript: Lili Zhai, Soumik Ray, and Antonin Kunka. Reviewed and revised the paper: Lili Zhai, Soumik Ray, Antonin Kunka, and Alexander K. Buell.

Reference

1. Lucking, C. B. & Brice, A. Alpha-synuclein and Parkinson's disease. 57, 15 (2000).
2. Jos, S. et al. Molecular insights into α-synuclein interaction with individual human core histones, linker histone, and dsDNA. Protein Sci. 30, 2121–2131 (2021).
3. Cascella, R. et al. The release of toxic oligomers from α-synuclein fibrils induces dysfunction in neuronal cells. Nat. Commun. 12, 1814 (2021).
4. Venda, L. L., Cragg, S. J., Buchman, V. L. & Wade-Martins, R. α-Synuclein and dopamine at the crossroads of Parkinson's disease. Trends Neurosci. 33, 559–568 (2010).
5. Li, S. & Le, W. Milestones of Parkinson's Disease Research: 200 Years of History and Beyond. Neurosci. Bull. 33, 598–602 (2017).
6. Jiang, K. et al. Alpha-Synuclein Modulates the Physical Properties of DNA. Chem. – Eur. J. 24, 15685–15690 (2018).
7. Sohrabi, T. et al. Common Mechanisms Underlying α-Synuclein-Induced Mitochondrial Dysfunction in Parkinson's Disease. J. Mol. Biol. 435, 167992 (2023).
8. Jos, S. et al. Molecular insights into α-synuclein interaction with individual human core histones, linker histone, and dsDNA. Protein Sci. 30, 2121–2131 (2021).
9. Abraham, J. N. & Nardin, C. Interaction of polymers with amyloidogenic peptides: Interaction of polymers with amyloidogenic peptides. Polym. Int. 67, 15–24 (2018).
10. Vasquez, V. et al. Chromatin-Bound Oxidized α-Synuclein Causes Strand Breaks in Neuronal Genomes in in vitro Models of Parkinson's Disease. J. Alzheimers Dis. 60, S133–S150 (2017).
11. Dent, S. E. et al. Phosphorylation of the aggregate-forming protein alpha-synuclein on serine-129 inhibits its DNA-bending properties. J. Biol. Chem. 298, 101552 (2022).
12. Jiang, K., Rocha, S., Kumar, R., Westerlund, F. & Wittung-Stafshede, P. C-terminal truncation of α -synuclein alters DNA structure from extension to compaction. Biochem. Biophys. Res. Commun. 568, 43–47 (2021).

13. Jones, N. A. et al. Polymer chemical structure is a key determinant of physicochemical and colloidal properties of polymer–DNA complexes for gene delivery. Biochim. Biophys. Acta BBA - Gene Struct. Expr. 1517, 1–18 (2000).
14. Debnath, K., Sarkar, A. K., Jana, N. R. & Jana, N. R. Inhibiting Protein Aggregation by Small Molecule-Based Colloidal Nanoparticles. Acc. Mater. Res. 3, 54–66 (2022).
15. Page, M. J. et al. Serum amyloid A binds to fibrin(ogen), promoting fibrin amyloid formation. Sci. Rep. 9, 3102 (2019).
16. Biancardi, A., Biver, T. & Mennucci, B. Fluorescent dyes in the context of DNA-binding: The case of Thioflavin T: BIANCARDI et al. Int. J. Quantum Chem. 117, e25349 (2017).
17. Sawner, A. S. et al. Modulating α-Synuclein Liquid–Liquid Phase Separation: Published as part of the Biochemistry virtual special issue "Protein Condensates". Biochemistry 60, 3676–3696 (2021).
18. Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. Science 357, eaaf4382 (2017).
19. Mukherjee, S. et al. Liquid-liquid Phase Separation of α-Synuclein: A New Mechanistic Insight for α-Synuclein Aggregation Associated with Parkinson's Disease Pathogenesis. J. Mol. Biol. 435, 167713 (2023).
20. Ray, S., Mason, T. O., Boyens-Thiele, L., Jahnke, N. & Buell, A. K. Mass photometric detection and quantification of nanoscale α-synuclein phase separation. http://biorxiv.org/lookup/doi/10.1101/2022.05.03.490467 (2022) doi:10.1101/2022.05.03.490467.
21. Ray, S. et al. α-Synuclein aggregation nucleates through liquid–liquid phase separation. Nat. Chem. 12, 705–716 (2020).
22. Hyman, A. A., Weber, C. A. & Jülicher, F. Liquid-Liquid Phase Separation in Biology. Annu. Rev. Cell Dev. Biol. 30, 39–58 (2014).
23. King, J. T. & Shakya, A. Phase separation of DNA: From past to present. Biophys. J. 120, 1139–1149 (2021).
24. Stender, E. G. P. et al. Capillary flow experiments for thermodynamic and kinetic characterization of protein liquid-liquid phase separation. Nat. Commun. 12, 7289 (2021).
25. Shakya, A. & King, J. T. DNA Local-Flexibility-Dependent Assembly of Phase-Separated Liquid Droplets. Biophys. J. 115, 1840–1847 (2018).
26. Rayan, G., Guet, J.-E., Taulier, N., Pincet, F. & Urbach, W. Recent Applications of Fluorescence Recovery after Photobleaching (FRAP) to Membrane Bio-Macromolecules. Sensors 10, 5927–5948 (2010).
27. Hegde, M., L. DNA induced folding/fibrillation of alpha-synuclein: new insights in Parkinson's disease. Front. Biosci. 15, 418 (2010).
28. Zheng, Y. et al. Novel DNA Aptamers for Parkinson's Disease Treatment Inhibit α-Synuclein Aggregation and Facilitate its Degradation. Mol. Ther. - Nucleic Acids 11, 228–242 (2018).
29. Tsukakoshi, K., Abe, K., Sode, K. & Ikebukuro, K. Selection of DNA Aptamers That Recognize α-Synuclein Oligomers Using a Competitive Screening Method. Anal. Chem. 84, 5542–5547 (2012).
30. Jarmoskaite, I., AlSadhan, I., Vaidyanathan, P. P. & Herschlag, D. How to measure and evaluate binding affinities. eLife 9, e57264 (2020).
31. Ramis, R., Ortega-Castro, J., Vilanova, B., Adrover, M. & Frau, J. Unraveling the NaCl Concentration Effect on the First Stages of α-Synuclein Aggregation. Biomacromolecules 21, 5200–5212 (2020).

32. Ziaunys, M., Sakalauskas, A., Mikalauskaite, K. & Smirnovas, V. Rapid restructurization of conformationally-distinct alpha-synuclein amyloid fibrils at an elevated temperature. PeerJ 10, e14137 (2022).
33. Younan, N. D. & Viles, J. H. A Comparison of Three Fluorophores for the Detection of Amyloid Fibers and Prefibrillar Oligomeric Assemblies. ThT (Thioflavin T); ANS (1-Anilinonaphthalene-8-sulfonic Acid); and bisANS (4,4′-Dianilino-1,1′-binaphthyl-5,5′-disulfonic Acid). Biochemistry 54, 4297–4306 (2015).
34. Matulis, D. & Lovrien, R. 1-Anilino-8-Naphthalene Sulfonate Anion-Protein Binding Depends Primarily on Ion Pair Formation. Biophys. J. 74, 422–429 (1998).
35. Sulatsky, M. I. et al. Effect of the fluorescent probes ThT and ANS on the mature amyloid fibrils. Prion 14, 67–75 (2020).
36. Buell, A. K. et al. Solution conditions determine the relative importance of nucleation and growth processes in α-synuclein aggregation. Proc. Natl. Acad. Sci. 111, 7671–7676 (2014).
37. Pronchik, J., He, X., Giurleo, J. T. & Talaga, D. S. In Vitro Formation of Amyloid from α-Synuclein Is Dominated by Reactions at Hydrophobic Interfaces. J. Am. Chem. Soc. 132, 9797–9803 (2010).
38. Mehra, S., Gadhe, L., Bera, R., Sawner, A. S. & Maji, S. K. Structural and Functional Insights into α-Synuclein Fibril Polymorphism. Biomolecules 11, 1419 (2021).
39. Sawner, A. S. et al. Modulating α-Synuclein Liquid–Liquid Phase Separation: Published as part of the Biochemistry virtual special issue “Protein Condensates”. Biochemistry 60, 3676–3696 (2021).
40. Stender, E. G. P. et al. Capillary flow experiments for thermodynamic and kinetic characterization of protein liquid-liquid phase separation. Nat. Commun. 12, 7289 (2021).
41. Sawner, A. S. et al. Modulating α-Synuclein Liquid–Liquid Phase Separation: Published as part of the Biochemistry virtual special issue “Protein Condensates”. Biochemistry 60, 3676–3696 (2021).
42. Siegert, A. et al. Interplay between tau and α-synuclein liquid–liquid phase separation. Protein Sci. 30, 1326–1336 (2021).
43. Biswas, N. et al. Phase separation in crowded micro-spheroids: DNA–PEG system. Chem. Phys. Lett. 539–540, 157–162 (2012).
44. Shakya, A. & King, J. T. DNA Local-Flexibility-Dependent Assembly of Phase-Separated Liquid Droplets. Biophys. J. 115, 1840–1847 (2018).
45. King, J. T. & Shakya, A. Phase separation of DNA: From past to present. Biophys. J. 120, 1139–1149 (2021).
46. Mimura, M. et al. Quadruplex Folding Promotes the Condensation of Linker Histones and DNAs via Liquid–Liquid Phase Separation. J. Am. Chem. Soc. 143, 9849–9857 (2021).
47. Stender, E. G. P. et al. Capillary flow experiments for thermodynamic and kinetic characterization of protein liquid-liquid phase separation. Nat. Commun. 12, 7289 (2021).
48. Biswas, N. et al. Phase separation in crowded micro-spheroids: DNA–PEG system. Chem. Phys. Lett. 539–540, 157–162 (2012).
49. Ray, S., Mason, T. O., Boyens-Thiele, L., Jahnke, N. & Buell, A. K. Mass photometric detection and quantification of nanoscale α-synuclein phase separation. http://biorxiv.org/lookup/doi/10.1101/2022.05.03.490467 (2022) doi:10.1101/2022.05.03.490467.
50. Jiang, K., Rocha, S., Kumar, R., Westerlund, F. & Wittung-Stafshede, P. C-terminal truncation of α -

synuclein alters DNA structure from extension to compaction. Biochem. Biophys. Res. Commun. 568, 43–47 (2021).
51. Jiang, K. et al. Alpha-Synuclein Modulates the Physical Properties of DNA. Chem. – Eur. J. 24, 15685–15690 (2018).
52. Hegde, M., L. DNA induced folding/fibrillation of alpha-synuclein: new insights in Parkinson's disease. Front. Biosci. 15, 418 (2010).
53. Cherny, D., Hoyer, W., Subramaniam, V. & Jovin, T. M. Double-stranded DNA Stimulates the Fibrillation of α-Synuclein in vitro and is Associated with the Mature Fibrils: An Electron Microscopy Study. J. Mol. Biol. 344, 929–938 (2004).
54. Hegde, M. L. & Rao, K. S. J. DNA induces folding in α-synuclein: Understanding the mechanism using chaperone property of osmolytes. Arch. Biochem. Biophys. 464, 57–69 (2007).
55. Nott, T. J. et al. Phase Transition of a Disordered Nuage Protein Generates Environmentally Responsive Membraneless Organelles. Mol. Cell 57, 936–947 (2015).
56. Pedersen, M. E., Østergaard, J. & Jensen, H. Flow-Induced Dispersion Analysis (FIDA) for Protein Quantification and Characterization. in Clinical Applications of Capillary Electrophoresis (ed. Phillips, T. M.) vol. 1972 109–123 (Springer New York, 2019).
57. Poulsen, N. N. et al. Flow induced dispersion analysis rapidly quantifies proteins in human plasma samples. The Analyst 140, 4365–4369 (2015).
58. Pedersen, M. E., Gad, S. I., Østergaard, J. & Jensen, H. Protein Characterization in 3D: Size, Folding, and Functional Assessment in a Unified Approach. Anal. Chem. 91, 4975–4979 (2019).
59. Restan, M. S., Pedersen, M. E., Jensen, H. & Pedersen-Bjergaard, S. Electromembrane Extraction of Unconjugated Fluorescein Isothiocyanate from Solutions of Labeled Proteins Prior to Flow Induced Dispersion Analysis. Anal. Chem. 91, 6702–6708 (2019).
60. Pedersen, M. E., Østergaard, J., Glintborg, B., Hetland, M. L. & Jensen, H. Assessment of immunogenicity and drug activity in patient sera by flow-induced dispersion analysis. Sci. Rep. 12, 4670 (2022).

3.3 Part II Small section - α-synuclein assembly by mRNA display

In this short section, we attempted the use of the mRNA display method to explore the sequence space of α-Syn related to Parkinson's Disease (PD). This study was planned to be conducted in two stages. The first stage involves establishing and optimizing the mRNA display method using wild-type α-synuclein. The second part would have involved employing large libraries of peptide species generated *in vitro* by mRNA display to explore the large sequence space of α-Syn. In the first part, we optimized key processes in mRNA display, including mRNA yield, puromycin primer linker, and the optimization of translation conditions. This could potentially enhance the versatility and applicability of mRNA display in studying a wide variety of amyloid proteins beyond α-Syn. Overall, we successfully established an mRNA display technical platform using wild-type α-synuclein. However, due to the pandemic and my inability to visit DTU until close to the end of the PhD period, I was unable to carry out the second part. This part will be continued in the laboratory of Prof. Alexander Büll in the framework of a recently awarded ERC consolidator grant.

α-synuclein assembly by mRNA display

Contributions: Lili Zhai performed the experiments for α-synuclein assembly by mRNA display on the bench, analyzed the data, and wrote this section.

Abstract

Relatively few studies have been conducted on the sequence space exploration of α-synuclein (α-Syn), in particular beyond the few handfuls of pathogenic variants. This lack of systematic large-scale studies limits a comprehensive understanding of α-Syn variants' influence on the pathogenic mechanisms of α-Syn in Parkinson's disease (PD). In this study, we attempted to use mRNA display technology to explore the sequence space of α-Syn for the first time at extremely high throughput of thousands or ultimately even millions of sequence variants. We optimized the mRNA display procedure so that it can be applied to studying amyloid proteins related to neurotoxic diseases, making it more widely accessible to the scientific community. In particular, we did not plan, as is customary, to immobilize the screening target to a surface, but to incubate the RNA-display library together with different forms of α-Syn aggregates. We established the mRNA displayed wild-type α-Syn successfully. Future work will focus on using extensive libraries of peptide species generated in vitro by mRNA display to explore the

sequence space of α-Syn related to neurodegenerative diseases.

Introduction

α-Synuclein (α-Syn) is a naturally disordered and unfolded presynaptic neuronal protein that can form filamentous aggregates, which are a prominent and presumably key etiological factor in Parkinson's disease (PD)[1]. Recently, mutations of α-Syn associated with early-onset familial PD are also known to modulate its aggregation, supporting a role in PD pathogenesis. These mutations include A53T, A30P, E46K, H50Q, G51D[2,3,4,5]. However, few studies have conducted mutagenesis studies ranging far beyond the small collection of disease-related variants on α-Syn and analyzed the characteristics of the mutant variants. This limitation significantly hinders our global and comprehensive understanding of the role of individual sequence regions and properties on α-Syn aggregation and of the pathogenic mechanisms of α-Syn in Parkinson's disease (PD).

The possibility, through the mRNA display approach, to search a larger fraction of sequence space for peptide species is one advantage compared to alternative methods (e.g., yeast display system) [6]. mRNA display generates large libraries of peptide species that are covalently attached to their own coding mRNA[7]. The mRNA display approach's advantages in terms of diversity, conceptual simplicity, directed evolution, functional characterization, and disease research make it a valuable tool in various fields of biological and biomedical research, such as protein-protein interactions[8], drugs discovery[9], and protein directed evolution[10]. Overall, mRNA display technology is a powerful tool for protein sequence exploration[11].

In this study, we employed an mRNA display technology approach with the ultimate aim to explore the sequence space of α-Syn, related to Parkinson's disease (PD). This platform allows us to generate a library that could in principle comprise over 10^{10} (ten billion) variants of α-Syn. The mRNA-displayed sequence library of α-synuclein can be characterized by physical properties such as solubility, aggregation levels, and hybrids that selectively bind aggregates/droplets from the soluble protein. In this study, we established the mRNA display approach using wild-type α-Syn successfully and will continue this study in the future. This application would be the first time to use the extremely high throughput nature of mRNA display for the study of α-synuclein aggregation, which is related to Parkinson's disease (PD). A study carried out at such high throughput will provide mechanistic insights into its role in neurodegenerative disorders at an unprecedented scale and contribute to unraveling the mechanistic intricacies of its involvement in neurodegenerative disorders.

Results and discussion

mRNA preparation for mRNA display

We optimized the DNA products by comparing three groups of primers (Primer S1, Primer S2, and Primer S3, shown in Table 1) at different annealing temperatures to amplify the α-Syn gene (shown in Figure 1A). Primer S2 F and R were the best primers to large-scale amplify α-Syn cDNA. Then, the α-Syn cDNA was transcribed into mRNA by T7 RNA polymerase. mRNA production was evaluated by comparing different incubation times (3 hours, 6 hours, and 12 hours). The results demonstrated that the highest accumulation was observed as the incubation time increased (Figure 1C), thus we chose 12 hours for incubation. Subsequently, four purification methods, including mRNA purification kit; lithium chloride precipitation; mRNA capture with oligo (dT) magnetic particles; and magnetic mRNA isolation, were performed to purify mRNA produced in the last step, with the most efficient one yielding over 80% recovery by using magnetic isolation from the reaction mixture (Figure 1B). Then, four primer-linkers carrying puromycin were designed to optimize the conjugation efficiency of the purified mRNA and puromycin, as shown in Table 1. Photo crosslinks were performed by using UV light (~365 nm) for 30 minutes. The puromycin molecule is linked to the 3' end of mRNA, which can enter the peptidyl transferase site after the ribosome reaches the end of the mRNA. During the process, the linker used is crucial as previous studies have shown that improper linkers (including length and components) can result in little or no hybrid formation[12]. The photo-crosslinked products were evaluated by running denatured Urea gel electrophoresis, resulting in > 50% crosslinked products with Primer-linker Puromycin D (Figure 1D and Table 1). As an alternative to photo-crosslinking, enzymatic ligation with T4 or T7 DNA ligase to attach the puromycin linker to the targeted mRNA was previously reported [11]. The main drawback of this enzymatic process is the requirement for a splint primer, which enhances the complexity of the process. Additionally, the splint primer must be fully removed to prevent inhibition of fusion formation[13]. Compared to enzymatic ligation, the photo-crosslinked mRNA puromycin conjugation method was easier to handle and more efficient in generating the mRNA crosslinked products. The cross-linked products were then purified from the crude photo-crosslinking reaction mixtures and added to the translation kit for fusion formation.

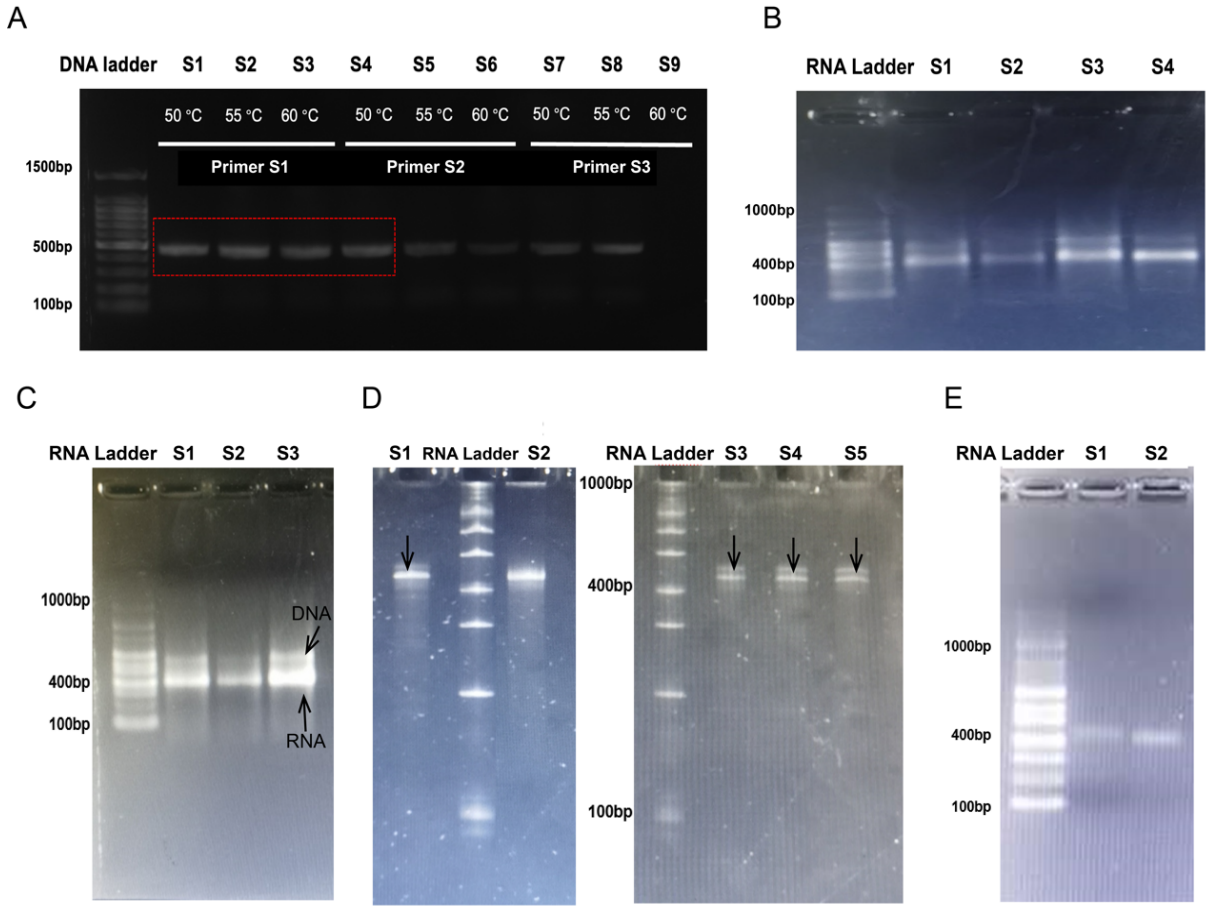


Figure 1. Gel analysis for mRNA display of WT α-Syn. A) The extensions of DNA products of WT α-Syn are normally performed using three groups of primers at different annealing temperatures: 55 °C, 60 °C, and 65 °C. Lanes S1 - S3 in this figure show extensions using primer S2; Lanes S3 - S6 show extensions using primer S3; Lanes S6 - SR9 show extensions using primer S1. The products in the red box in the figure indicated that the amplification is chosen with high productivity and high specificity. B) Four purification methods: mRNA purification kit (Invitrogen) (Lane S1); lithium chloride (LiCl) precipitation (Lane S2); mRNA capture with oligo (dT) magnetic particles (Thermo fisher) (Lane S3); and magnetic mRNA isolation (Lane S4). C) Optimized mRNA productions at 3 h (Lane S1), 6 h (Lane S2), and 12 h (Lane S3). D) Photo crosslink (UV light with $\lambda > 300$ nm for 30 minutes) between mRNA and puromycin linkers is measured by denaturing urea gel electrophoresis. Lane S2 shows the mRNA without a linker; Lane S1 (Primer-linker Puromycin A), Lane S3 (Primer-linker Puromycin B), Lane S4 (Primer-linker Puromycin C), and Lane S5 show the photo crosslink with different linkers. E) Separated photo-crosslinked mRNA from free mRNA uses oligo deoxythymidine (dT) magnetic particles. Lane S1 is the crosslinked mRNA and lane S2 is the free mRNA.

Table 1. Primer information.

Primer List	Sequence
Primer S1 F	TAATACGACTCACTATAGgcgACCATGGACGTGTTTATGAAAGGT
Primer S1 R	TTAAATAGCGGATGCatgatgatgatgatgatgTGCTTCCGGTTCATAATC
Primer S2 F	TAATACGACTCACTATAGgcgGCCATGGACGTGTTTATGAAAGGTCTGA
Primer S2 R	TTAAATAGCGGATGCatgatgatgatgatgatgTGCTTCCGGTTCATAATCCTGAT
Primer S3 F	TAATACGACTCACTATAGgcgACCATGGACGTGTTTATGAAAGGTCTGAGC AAAG
Primer S3 R	TTAAATAGCGGATGCatgatgatgatgatgatgTGCTTCCGGTTCATAATCCTGATA AC
Primer-linker Puromycin A	psoralen C6 TAGCGGATGCaaaaaaaaaaaaaaaaaTEGTEGACCC puromycin
Primer-linker Puromycin B	psoralen-TAGCGGATGC-(TEG)6 -dCdC-puromycin
Primer-linker Puromycin C	psoralen-TAGCGGATGCaaaaaaaaaaaaaa-(PEG)2-ACC-puromycin
Primer-linker Puromycin D	psoralen-TAGCGGATGCaaaaaaaaaaaaaa-(TEG)4-ACC-puromycin

cDNA-mRNA-protein hybrid generation by mRNA display

The crosslinked mRNA products from the last step were translated by using a translation kit to assess the production of translation and fusion products. Western blot analysis was performed using an anti-α-Syn monoclonal antibody to detect the presence of the protein. WT α-Syn protein as positive control demonstrated a 14 kDa band, corresponding to α-Syn-synuclein, in the Lanes S2 in Figure 2. In addition, free α-Syn mRNA was used as a positive control for protein translation, shown as Lanes S1 in Figure 2. The translated products under different salt concentrations were analyzed using WB (Figure 2), and the results indicated that high salt concentrations including Mg^{2+} and KCl improved the yield of mRNA-α-Syn. The most efficient mRNA-protein fusion formation was shown in Lane S3 in Figure 2, which was accomplished under the condition containing 50 mM Mg^{2+} and 200 mM KCl. In lanes S3-S6 in Figure 2, some bands indicated with black arrows might be non-full-length fusion products carrying fractional mRNA. The Western blot analysis confirmed the presence of full-length fusion products. These non-full-length products may form as a result of RNase degradation during gel processing or handling. To purify these products, puromycin-linker D containing oligo (dA) residues were used to generate a larger amount of translation that could bind to an oligo (dT) column. Subsequently, these fusion products were purified using a His-tag column to remove free mRNA and an oligo d(T) kit to remove free peptides. Then, the fusion molecules of mRNA-α-Syn were transformed into the DNA-mRNA-α-Syn hybrids by using reverse transcriptase. In general mRNA display technology, the target substance to be screened is immobilized on the

solid phase carrier, and the cDNA-mRNA-protein fusion containing the target protein can be separated by specific binding with the target substance[10]. In this study, the resulting DNA-mRNA-α-Syn hybrids are not immobilized but instead used for incubating with an excess of aggregating/phase-separating WT protein for subsequent characterization analysis.

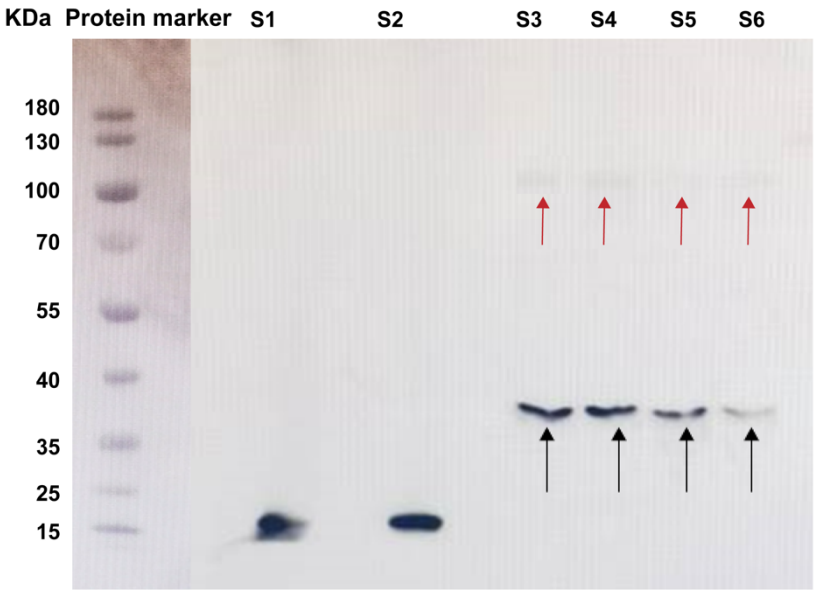


Figure 2. Hybrids of mRNA-α-Syn were formed in vitro using the Retic Lysate IVT™ Kit and identified by Western blotting using an anti-α-Syn antibody. The peptide synthesis from the free α-Syn mRNA template is shown in lane S1, and WT α-Syn protein as positive control is shown in lane S2 (2μM). Lanes S3 - S6 present the mRNA-protein fusion products. The full-length size of the target strips is about 140 KDa showing in top bands pointed by red arrows. The bands indicated by black arrows may not be full-length bands. Lane S3 and lane S4 show the fusion products employing high concentrations of salts such as Mg^{2+} and KCl (50 mM Mg^{2+} and 200 mM KCl; 50 mM Mg^{2+} and 400 mM KCl). Lane S5 shows the lower fusion with only 200 mM KCl; lane S6 shows the fusions without salt.

Future perspectives

Based on the study described above, we established the mRNA display method using wild-type α-Syn-synuclein successfully. Along with the study of the comprehensive understanding of the weak interactions between α-Syn and nucleic acids, we have confirmed the feasibility of the subsequent screening methods. Future studies will focus on using large libraries of peptide species generated in vitro by mRNA display to explore the sequence space of α-Syn in such a context for the first time. This approach will allow the simultaneous exploration of aggregation and phase separation of a large number of sequence variants by incubating an mRNA-displayed

sequence library of α-synuclein with an excess of aggregating/phase-separating WT protein. Then, physical separation of the aggregates/droplets from the soluble protein is performed, resulting in the formation of different variant pools, followed by sequencing of these different pools to determine the fraction of the library associated with each pool.

Therefore, the entire workflow can be divided into five stages, as shown in Figure 3: DNA library generation, generation of hybrids of DNA-mRNA-peptide using mRNA display, characterization analysis of hybrids of DNA-mRNA-peptide, physical separation, and sequencing analysis. First, we can construct a random library of the NAC region or N terminus of α-Syn and generate these hybrids of DNA-mRNA-peptide generated by mRNA display directly (Figures 3 A-D). The sequence space of more than 10^{10} peptide-aggregate interactions can be explored by the mutant design of specific regions in the original sequence of α-Syn. The mRNA-displayed sequence library of α-synuclein will be incubated with an excess of aggregating/phase-separating WT protein to identify different variant protein properties (Figure 3E). Then, these variants of hybrids can be physically separated by their properties such as solubility, aggregation levels, and hybrids that selectively bind aggregates/droplets from the soluble protein. The separation method includes centrifugation or microfluidic diffusion methods, which can be employed to separate these hybrids that selectively bind to aggregates or monomeric proteins (Figure 3F). The separation process generates different sequence variants pools, sequencing the different pools to determine the fraction of the library associated with each pool (Figure 3F). During the process, this approach may yield peptide species with diverse effects, such as solubilizing the original peptide and enhancing or inhibiting aggregation. Finally, these sequenced variants can be analyzed and summarized.

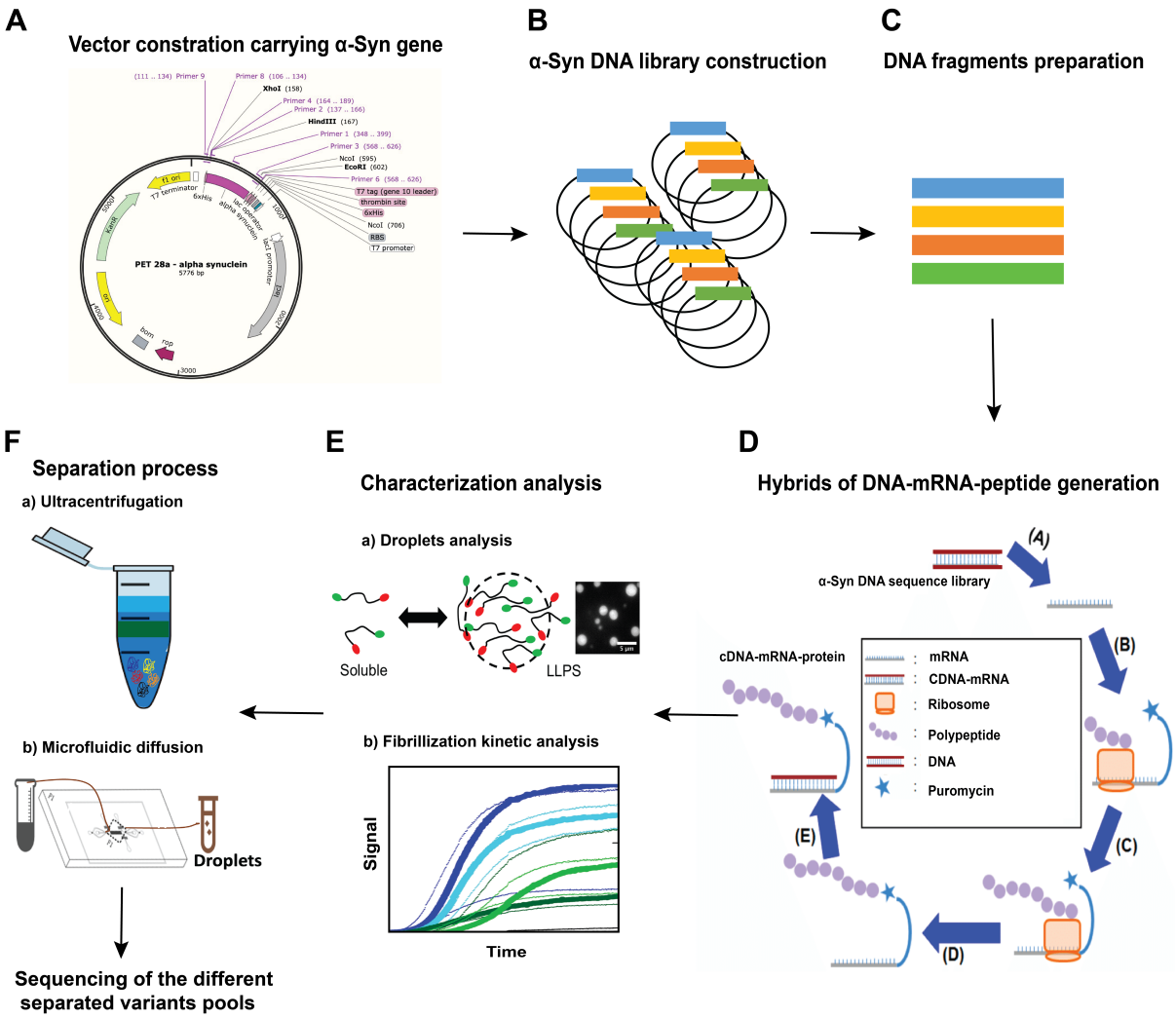


Figure 3. A schematic illustration of the workflow of the study of α-synuclein assembly using mRNA display is presented. A) The diagram depicts the constructed PET-28a Vector containing the α-Syn gene. B) DNA library of α-Syn is generated through random mutagenesis. C) DNA fragments are prepared for mRNA display. D) Hybrids of DNA-mRNA-peptide are generated via mRNA display. The overall process involves five steps: mRNA fragment preparation, puromycin conjugation with mRNA, mRNA translation with puromycin using the translation kit, purification, and generation of DNA-mRNA-peptide hybrids. E) The hybrids and α-Syn variants undergo characterization analysis by using multiple methods including a) Droplets of phase separation analysis, and b) Fibrillization kinetic measurements and analysis. F) The hybrids of DNA-mRNA-peptide are separated using a) ultracentrifugation and b) microfluidic diffusion. Then, the sequencing is performed for different separated variant pools.

Conclusion

This study outlines a comprehensive workflow for establishing an mRNA display technology

platform, utilizing wild-type α-Syn as the target protein. The optimized workflow for mRNA display involves the following seven steps: cDNA or library amplification preparation, in vitro transcription and DNase digestion, mRNA purification and conjugation to puromycin oligonucleotide linkers, in vitro translation/fusion formation, oligo(dT) mRNA Purification, reverse transcription, and protein affinity purification. In the limited time available for this work, we successfully established the generation of mRNA-displayed WT α-Syn. Future work in the Büll laboratory will continue the study of large libraries of peptide species generated in vitro by mRNA display to explore the sequence space of α-Syn. By exploring the sequence space of peptide-aggregate interactions, this study gained insights into the mechanisms underlying protein aggregation related to diseases such as Parkinson's. In addition, the workflow constructed in this study can also be employed for the study of other disordered proteins.

Methods

Materials

The human α-Syn gene was synthesized by GenScript and inserted into the plasmid PET 28a. High-fidelity Pfu DNA Polymerase was purchased from Promega Corporation and T7 RNA polymerase, RNase inhibitor, Retic Lysate IVT™ Kit, dNTP mix, magnetic grate, methionine, 0.2 nm PVDF membrane, SuperSignal™ West Pico PLUS chemiluminescent substrate, western blot buffer, BSA, DTT, $MgCl_2$, KCl, and Nuclease Free Water was purchased from Thermo Scientific. A handheld UV lamp (UVGL-55) was purchased from VWR International. Primers were also obtained from GenScript and listed in Table 1. The anti-α-synuclein antibody ([EPR20535] (ab212184)) was purchased from Abcam.

DNA products of WT α-Syn

DNA libraries can be constructed by random mutagenesis using the published protocol[14]. Here, we described the process for the PCR amplification of the whole length of the WT α-Syn gene. PCR reaction was started by adding 8 U of the Pfu DNA Polymerase, the reaction mixture including 10 µL of 10 X reaction buffer with MgSO4, 2 µL dNTP mix (10 mM each), 1 µg upstream primer, 1 µg downstream primer, 150 ng template, and nuclease-free water to a final volume of 100 µM. The reaction was assembled on ice and then amplified using a thermal cycler with the following protocol: 2 minutes at 95 °C for denaturation, followed by 30 cycles of 95 °C for 30 seconds, 50 - 60 °C for 30 seconds, and 72 °C for 1 minute for extension, followed by a final extension at 72 °C for 5 minutes and then stored at 4 °C. The amplified

DNA was then analyzed using a 20% agarose gel.

Transcription and purification

High quantities of mRNA were produced through in vitro transcription using T7 RNA polymerase. The reaction mixture included 1 X T7 RNA polymerase buffer, 25 mM $MgCl_2$, 5 mM NTP mixture, 500 nM initial DNA, 3 U T7 RNA polymerase, and nuclease-free DEPC water to reach the desired final volume. The mixture was incubated at 37 °C for 3 - 12 hours and then quenched by adding 10 mM Na-EDTA for 5 minutes at room temperature. DNA template was removed from the mRNA products through DNase digestion, which involved adding 2 µL RNase-free DNase and incubating at 37 °C for 1 hour, followed by adding 10 mM EDTA to stop the reaction. Different methods were used to purify the mRNA including lithium chloride (LiCl) precipitation[15]; mRNA purification kit (Thermo Fisher)[16]; mRNA capture with oligo(dT) magnetic particles[17], and magnetic mRNA isolation (RNAClean XP, Beckman) following the protocol from products instructions for the user.

Puromycin conjugation

There were several primer linkers reported for puromycin conjugation. We used four different primer linkers, as shown in Table 1, to perform the photocrosslinking reaction with puromycin, as psoralen-mediated UV crosslinking is simpler to conduct[13]. Psoralen was at the 5' end of the oligo linker and puromycin was connected to the 3' end. The oligo linkers (7.5 µM) were annealed to the target mRNAs (3 µM) in 25 mM Tris-HCl buffer, pH 7.0, 100 mM NaCl by heating at 85 °C for 10 minutes followed by cooling to 4 °C for 10 minutes. The reaction mixture was irradiated under ~365 nm UV light for 30 minutes. The photo-crosslinking efficiency was analyzed on a denaturing 6% TBE-urea gel and investigated using an imaging system (XR, Biorad). The photo-crosslinked mRNA was separated from free mRNA using oligo deoxythymidine $(dT)_{25}$ magnetic particles[17].

In vitro translation and Western Blot

In vitro, translation reactions were conducted in rabbit reticulocyte lysate for 1 hour at 30 °C. The final concentration of the translation reaction mixture contained 0.2 µM puromycin-linked mRNA, 200 mM to 300 mM KCl, 0 to 50 mM MgOAc, 25 µM amino acid mixture (Methionine minus), 1 µM Methionine, and 40% reticulocyte lysate, following the protocol[11]. The fusion

products were analyzed by electrophoresis on 4 -12% SDS-PAGE and then Western blot analysis was performed using an anti-α-Syn-synuclein antibody. The target hybrids were transferred to the PVDF membrane using eBlotTM L1 protein transfer (Biorad). After transfer, the membrane was rinsed twice in deionized water for 5 minutes with shaking to remove all transfer buffer and then blocked with blocking buffer for 1 hour at 37 °C, with the antibody solution completely covering the blot. The membrane was then rinsed 4 times by removing the buffer for 15 minutes with agitation in an incubator. The working solutions of chemiluminescent substrates were prepared in 2 ml final volume according to the manufacturer's instructions. The blot was incubated with the working solution for 1 minute in the dark. Finally, the blot was placed on a clear plastic wrap and any air bubbles were removed, and the blot was imaged under appropriate conditions using an XR imaging system. The fusion products were captured with oligo(dT) magnetic particles[17].

Reference

1. Kuo, Y.-M. *et al.* Extensive enteric nervous system abnormalities in mice transgenic for artificial chromosomes containing Parkinson disease-associated α-synuclein gene mutations precede central nervous system changes. *Hum. Mol. Genet.* **19**, 1633–1650 (2010).
2. Zarranz, J. J. *et al.* The new mutation, E46K, of α-synuclein causes parkinson and Lewy body dementia: New α-Synuclein Gene Mutation. *Ann. Neurol.* **55**, 164–173 (2004).
3. Lesage, S. *et al.* G51D α-synuclein mutation causes a novel Parkinsonian-pyramidal syndrome: SNCA G51D in Parkinsonism. *Ann. Neurol.* **73**, 459–471 (2013).
4. Conway, K. A., Harper, J. D. & Lansbury, P. T. Accelerated in vitro fibril formation by a mutant α-synuclein linked to early-onset Parkinson disease. *Nat. Med.* **4**, 1318–1320 (1998).
5. Appel-Cresswell, S. *et al.* A-Syn-synuclein p.H50Q, a novel pathogenic mutation for Parkinson's disease: α- Synuclein p.H50q, A Novel Mutation For Pd. *Mov. Disord.* **28**, 811–813 (2013).
6. Newton, M. S., Cabezas-Perusse, Y., Tong, C. L. & Seelig, B. *In Vitro* Selection of Peptides and Proteins—Advantages of mRNA Display. *ACS Synth. Biol.* **9**, 181–190 (2020).
7. Seelig, B. mRNA display for the selection and evolution of enzymes from in vitro-translated protein libraries. *Nat. Protoc.* **6**, 540–552 (2011).
8. Blanco, C., Verbanic, S., Seelig, B. & Chen, I. A. High throughput sequencing of *in vitro* selections of mRNA-displayed peptides: data analysis and applications. *Phys. Chem. Chem. Phys.* **22**, 6492–6506 (2020).
9. Lipovsek, D. & Plückthun, A. In-vitro protein evolution by ribosome display and mRNA display. *J. Immunol. Methods* **290**, 51–67 (2004).
10. Takahashi, T. T., Austin, R. J. & Roberts, R. W. mRNA display: ligand discovery, interaction analysis and beyond. *Trends Biochem. Sci.* **28**, 159–165 (2003).
11. Barendt, P. A., Ng, D. T. W., McQuade, C. N. & Sarkar, C. A. Streamlined Protocol for mRNA Display. *ACS Comb. Sci.* **15**, 77–81 (2013).

12. Lipovsek, D. & Plückthun, A. In-vitro protein evolution by ribosome display and mRNA display. *J. Immunol. Methods* **290**, 51–67 (2004).
13. Kurz, M. Psoralen photo-crosslinked mRNA-puromycin conjugates: a novel template for the rapid and facile preparation of mRNA-protein fusions. *Nucleic Acids Res.* **28**, 83e–883 (2000).
14. McCullum, E. O., Williams, B. A. R., Zhang, J. & Chaput, J. C. Random Mutagenesis by Error-Prone PCR. in *In Vitro Mutagenesis Protocols* (ed. Braman, J.) vol. 634 103–109 (Humana Press, 2010).
15. Walker, S. E. & Lorsch, J. RNA Purification – Precipitation Methods. in *Methods in Enzymology* vol. 530 337–343 (Elsevier, 2013).
16. Mettel, C., Kim, Y., Shrestha, P. M. & Liesack, W. Extraction of mRNA from Soil. *Appl. Environ. Microbiol.* **76**, 5995–6000 (2010).
17. M, L. m R.N.A. Purifications Process: Affinity Chromato-Graphy, Magnetic Beads and Graphene Coated Large Scale Production and Toxicological Aspects. *Curr. Investig. Clin. Med. Res.* **2**, (2022).

3.4 Summary of α-synuclein aggregation studies

In this study, we intended to employ mRNA display for the first time to investigate the aggregation of α-synuclein (α-Syn), associated with Parkinson's Disease (PD). The extremely high-throughput nature of mRNA display would allow to screen and identify a very large number of α-synuclein variants in relation to PD. The screening process in this approach would involve incubating an mRNA-displayed sequence library of α-synuclein with an excess of aggregating/phase-separating WT protein and separations would then be conducted based on the distinct physical properties of α-synuclein variants, such as solubility or aggregation. The aggregates or droplets would then be separated from the soluble protein and both pools sequenced to identify those variants from the original sequence library associated with each pool. This would then in turn allow to determine which sequence properties are associated with a preference for the aggregated/condensed phase. The feasibility of this approach hinges on the nature and strength of the interactions between α-Syn and nucleic acids; if WT α-synuclein interacts very strongly with nucleic acids, the behavior of the mRNA-displayed protein constructs could be completely dominated by the RNA tail, leading to the lack of discrimination of the different sequence variants in the selection processes. We therefore set out to first study the interactions between nucleic acids (DNA in this case) and α-Syn, followed by first steps towards our intended mRNA display-based screening method.

Initially, we chose DNA as a substitute for RNA due to its ease of handling and to compare our findings with existing literature on α-Syn-DNA interactions. The investigation of the binding interaction between α-Syn and DNA was carried out in three parts, including quantification of the binding affinities between α-Syn and various types of DNA, evaluating the influence of DNA on the α-Syn fibrillization process, and investigating the effects of DNA on the phase separation of α-Syn. In the first part, we measured the binding affinities between α-Syn and four models of DNA, including nine different DNA sequences. The results demonstrated a weak binding interaction between the α-Syn monomer and all DNAs, with all binding affinities falling within the micromolar range which is several orders of magnitude higher (i.e. lower affinity) than the reported nanomolar range observed by others. In the second part, we selected five DNA sequences that displayed the highest affinities with α-Syn to evaluate the influence of DNA on the α-Syn fibrillization process monitored by two fluorescent dyes, namely 8-anilinonapthalene-1-sulfonic acid (ANS) and Thioflavin T (ThT). The results suggested that all DNA shows a weak inhibitory effect on the fibrillization process of α-synuclein, with observing varying degrees of delay and reduction in the lag and plateau phases of α-Syn aggregation

curves, respectively, in both the ThT and ANS experiments. In the third part, we investigated the influence of ssDNA and dsDNA on the phase separation of α-Syn. The results showed the formation of two distinct types of complex coacervates: Type-I coacervates with ssDNA and Type-II coacervates with dsDNA. Furthermore, these coacervates exhibited significantly different viscoelastic property changes over time. While α-Syn droplets exhibited aggregation and progressive solidification, the DNA-enriched droplets remained in a liquid state, suggesting that the presence of DNA has a minor impact on α-Syn aggregation under phase separation conditions. Taken together, these findings demonstrated that the binding interaction between α-Syn monomer and DNA molecules and the influence of DNA molecules on the α-Syn fibrillization process or α-Syn aggregates in phase separation was low. Therefore, we concluded that the long mRNA or mRNA-DNA tail of the mRNA-displayed sequence library of α-synuclein might be unlikely to interfere with the subsequent experiments.

Subsequently, we established the first steps of a workflow for implementing mRNA display technology in our laboratory, using α-Syn as the target protein. We simplified and improved the workflow to generate mRNA-displayed protein to make it applicable for studying amyloid proteins associated with neurodegenerative diseases, thereby widening its applicability in the scientific community. The optimized mRNA display workflow consists of seven steps: cDNA or library amplification preparation, in vitro transcription and DNase digestion, mRNA purification and conjugation to puromycin oligonucleotide linkers, in vitro translation/fusion formation, oligo(dT) mRNA Purification, reverse transcription, and protein affinity purification. This optimized approach provides the foundational and technical support necessary for constructing the mRNA display sequence library of α-Syn effectively. Future studies will focus on using large libraries of peptide species generated in vitro by mRNA display to explore the sequence space of α-Syn. In addition, these findings in this study can provide valuable insights into the binding interaction between nucleic acids and α-Syn, related to PD. Furthermore, the investigation of the protein sequence landscape of α-synuclein holds the potential to offer mechanistic elucidation of its involvement in neurodegenerative disorders on an unprecedented scale.

4 Conclusion and future perspectives

In this study, the relationship between protein sequence space and its functional and physical properties was explored, which focused on two proteins with different purposes: KOD DNA polymerase, a member of the archaeal B-family DNA polymerase, and α-synuclein, which is associated with neurodegenerative diseases such as Parkinson's Disease (PD).

Through systematically exploring the sequence space of the KOD protein, we successfully identified KOD variants with function improvement, which exhibited high catalytic activity towards modified nucleotides. We characterized and analyzed thousands of variants generated through semi-rational and rational designs to explore the relationship between the protein's sequence, structure, and functions. Firstly, we employed a semi-rational design strategy by integrating protein structure analysis and computer-assisted site prediction to selectively screen the sequence space of the KOD protein. By selectively introducing mutations and screening the mutants, we successfully obtained KOD mutants with functional improvement. Subsequently, based on the optimal mutants obtained from the semi-rational screening, we performed a rational screening approach to continually enhance their function towards catalytic efficiency. Finally, we engineered a highly improved functional KOD variant through rational screening and performed a comprehensive analysis to elucidate the correlation between its sequence, structure, and function.

In the semi-rational screening process, we employed the molecular dynamics (MD) simulation method in conjunction with the analysis of the KOD structure to explore a library of over 1800 sequences covering 93 sites. The binding energy of these 93 residues was calculated and compared using the MM-GBSA method. Subsequently, over 100 KOD sequences were experimentally validated to assess their functional improvements. In our experimental screening, we constructed multiple rounds of mutation libraries, with each subsequent round using the best-performing variant from the previous round as the parent sequence. This repetitive process continued until we identified the variant with favorable functional improvement, analogous to climbing a "fitness landscape". Finally, we successfully tested a KOD variant carrying eleven mutation sites from four rounds of screening, which exhibited excellent performance in the NGS application. This result highlights the strong correlation between sequence modifications and functional enhancements. Our semi-rational screening strategy not only obtained promising results but also provided further evidence for the effectiveness of employing a semi-rational design approach in exploring protein sequences for functional improvement.

In the rational screening process, we employed machine learning（ML and MD simulation to explore a library of over 200 sequences covering 9 sites. To construct our machine learning model, we initially explored a library of 349 KOD sequences using traditional mutagenesis methods, such as site-directed mutagenesis and random mutagenesis, for the purpose of function improvement. With the screening results of 349 KOD variants, we then generated a medium-sized dataset for constructing our machine-learning model to generate the virtual library. Then, we screened 45 variants from the virtual library predicted by the ML model for experimental validation in order to assess their functional improvements. Seven KOD variants showed a significant increase in functional improvement, exhibiting kinetic performance that exceeded ten-fold improvement compared to the parent variant. In addition, we performed an analysis of the mechanism underlying the enhancement in function observed in the best KOD variant. We concluded that the function improvement observed in the best KOD variant is a comprehensive result of the synergy between dual-site mutations, which leads to conformational dynamics shifting within the local enzyme subdomain and concerted dynamical motion of the bound DNA duplex. Our rational screening strategy successfully engineered an improved KOD protein, as well as contributing to a better understanding of the relationship between sequence, structure, and protein function, thus facilitating the development of more efficient protein sequence space exploration strategies. In addition, these results provided new evidence highlighting the efficacy of employing machine learning techniques in enzyme engineering for comprehensive analysis of enzyme sequence-function relationships.

Additionally, there exist some disadvantages in our semi-rational and rational screening approaches. We employed the specific modified nucleotide (modified dATP) during the selection process, the enhancement of protein function might be confined to the improvement of the incorporation of this substrate. In NGS application, besides the need to incorporate four types of modified nucleotide, the DNA polymerase should also meet multiple requirements including high polymerization speed, low error rate, and good compatibility with sequencing reagents. Therefore, future studies could consider conducting catalytic activity tests on various modified substrates to ensure the applicability of proteins in various application scenarios. Furthermore, in our semi-rational screening, some combinatorial mutant libraries lack a comprehensive evaluation, which may lead to local optima in subsequent evolution, i.e., climbing up a small hill in the evolutionary landscape without reaching the highest peak. Future studies could increase the number and diversity of variants in the combinatorial mutant library, and optimize the screening process, as well as the machine learning models, to enhance the efficiency of exploring protein sequence space.

In a separate study, we made first steps towards the development of an mRNA display method to explore the sequence space of α-synuclein for mechanistic studies. The mRNA display method will allow the simultaneous study of large libraries of α-Syn related to Parkinson's Disease (PD) in such a context for the first time. This methodology allows the exploration of aggregation and phase separation of a very large number of sequence variants at the same time, which is achieved by co-incubating an mRNA-displayed sequence library of α-synuclein with an excess of wild-type protein that undergoes aggregation or phase separation. Then, the separation of the aggregates/droplets from the soluble protein can be performed, followed by sequencing of the separated pools to determine the fraction of the library associated with each pool. Therefore, an important aspect of the feasibility of this approach is to investigate the interactions between α-synuclein and nucleic acids. We performed the evaluation for the interactions between α-synuclein and DNA. We selected DNA as a surrogate for RNA in our study due to its ease of handling and the availability of comparison of prior literature on α-synuclein-DNA interactions.

Therefore, this study can be mainly divided into three stages, including the investigation of the interaction between α-synuclein and DNA, the optimization of the mRNA display method using wild-type α-synuclein, and the exploration of large libraries of peptide species generated in vitro by mRNA display. In the first stage, we initially investigated the interactions between α-synuclein and various types of DNA through three parts of experiments. The three parts included the binding affinity measurements of α-synuclein and DNA, evaluating the influence of DNA on the α-synuclein fibrillization process, and investigating the influence of DNA in α-synuclein phase separation. The results demonstrated that there are binding interactions between α-synuclein and various types of nucleic acids, and the binding interactions are weak. Thus, we concluded that the mRNA tail of the mRNA-displayed sequence library of α-synuclein might not influence the subsequent experiments. In the second stage, we successfully optimized and established an mRNA display technical platform using wild-type α-synuclein. The successful establishment and optimization of the mRNA display platform using wild-type α-synuclein represents a significant step forward in the field of neurodegenerative disorder studies. The third stage of exploring large libraries of peptide species generated in vitro by mRNA display was not completed in this study, due to lack of time. In the future, we plan to continue this study to explore the sequence space of α-Syn related to PD, by mRNA display approach.

Consequently, by developing a semi-rational and rational screening strategy, we successfully enhanced the efficiency of exploring the sequence space of the KOD protein and identified

purposeful variants with function improvement. In addition, using the mRNA display method, future studies could explore the sequence space of α-Syn to identify these variants with high specificity and affinity for well-defined soluble or aggregated states, related to PD, on an unprecedented scale. Overall, these findings and protein sequence exploration strategies have significant implications for the design and improvement of functional properties or mechanism studies in other proteins.

5 References

1. Vilchez, D., Saez, I. & Dillin, A. The role of protein clearance mechanisms in organismal ageing and age-related diseases. Nat. Commun. 5, 5659 (2014).
2. Sinz, A. Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. J. Mass Spectrom. 38, 1225–1237 (2003).
3. Li-Chan, E. C. Y. Properties of proteins in food systems: an introduction. in Proteins in Food Processing 2–26 (Elsevier, 2004). doi:10.1533/9781855738379.2.
4. Gustafsson, C., Govindarajan, S. & Emig, R. Exploration of sequence space for protein engineering. J. Mol. Recognit. 14, 308–314 (2001).
5. Helbert, W. et al. Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space. Proc. Natl. Acad. Sci. 116, 6063–6068 (2019).
6. Dryden, D. T. F., Thomson, A. R. & White, J. H. How much of protein sequence space has been explored by life on Earth? J. R. Soc. Interface 5, 953–956 (2008).
7. Sakai, A. STUDY OF ENZYME EVOLUTION WITHIN THE MLE SUBGROUP FOCUSING ON CHARACTERIZING MEMBER ENZYMES FOR THE PURPOSE OF DISCOVERING RELATIONSHIPS AMONG SEQUENCE, STRUCTURE, AND FUNCTION. 481.
8. Aschenbrenner, J. & Marx, A. DNA polymerases and biotechnological applications. Curr. Opin. Biotechnol. 48, 187–195 (2017).
9. Coco, W. M. et al. DNA shuffling method for generating highly recombined genes and evolved enzymes. Nat. Biotechnol. 19, 354–359 (2001).
10. Porter, J. L. et al. Directed Evolution of New and Improved Enzyme Functions Using an Evolutionary Intermediate and Multidirectional Search. ACS Chem. Biol. 10, 611–621 (2015).
11. Tobin, M. B., Gustafsson, C. & Huisman, G. W. Directed evolution: the 'rational' basis for 'irrational' design. Curr. Opin. Struct. Biol. 10, 421–427 (2000).
12. Jin, F. J., Maruyama, J., Juvvadi, P. R., Arioka, M. & Kitamoto, K. Development of a novel quadruple auxotrophic host transformation system by argB gene disruption using adeA gene and exploiting adenine auxotrophy in Aspergillus oryzae. FEMS Microbiol. Lett. 239, 79–85 (2004).
13. Dvorak, P. et al. Exacerbation of substrate toxicity by IPTG in Escherichia coli BL21(DE3) carrying a synthetic metabolic pathway. Microb. Cell Factories 14, 201 (2015).
14. Wu, L. et al. Nonpeptide-Based Small-Molecule Probe for Fluorogenic and Chromogenic Detection of

Chymotrypsin. Anal. Chem. 89, 3687–3693 (2017).
15. Nishikawa, T., Sunami, T., Matsuura, T., Ichihashi, N. & Yomo, T. Construction of a Gene Screening System Using Giant Unilamellar Liposomes and a Fluorescence-Activated Cell Sorter. Anal. Chem. 84, 5017–5024 (2012).
16. Leconte, A. M. et al. Directed Evolution of DNA Polymerases for Next-Generation Sequencing. Angew. Chem. 122, 6057–6060 (2010).
17. Planas-Iglesias, J. et al. Computational design of enzymes for biotechnological applications. Biotechnol. Adv. 47, 107696 (2021).
18. Siegel, J. B. et al. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. Science 329, 309–313 (2010).
19. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. Proc. Natl. Acad. Sci. 116, 8852–8858 (2019).
20. Li, G. et al. Significantly improving the thermostability of a hyperthermophilic GH10 family xylanase XynAF1 by semi-rational design. Appl. Microbiol. Biotechnol. 105, 4561–4576 (2021).
21. Kintses, B. et al. Picoliter Cell Lysate Assays in Microfluidic Droplet Compartments for Directed Enzyme Evolution. Chem. Biol. 19, 1001–1009 (2012).
22. Tubeleviciute, A. & Skirgaila, R. Compartmentalized self-replication (CSR) selection of Thermococcus litoralis Sh1B DNA polymerase for diminished uracil binding. Protein Eng. Des. Sel. 23, 589–597 (2010).
23. Ellefson, J. W. et al. Synthetic evolutionary origin of a proofreading reverse transcriptase. Science 352, 1590–1593 (2016).
24. Bruno, J. G., Carrillo, M. P., Phillips, T. & Andrews, C. J. A Novel Screening Method for Competitive FRET-Aptamers Applied to E. coli Assay Development. J. Fluoresc. 20, 1211–1223 (2010).
25. Ong, J. L., Loakes, D., Jaroslawski, S., Too, K. & Holliger, P. Directed Evolution of DNA Polymerase, RNA Polymerase and Reverse Transcriptase Activity in a Single Polypeptide. J. Mol. Biol. 361, 537–550 (2006).
26. Takahashi, F. et al. Activity-based in vitro selection of T4 DNA ligase. Biochem. Biophys. Res. Commun. 336, 987–993 (2005).
27. Zhang, L. DNA-Directed DNA Polymerases Evolve From Reverse Transcriptase. 103.
28. Eriksen, D. T., Lian, J. & Zhao, H. Protein design for pathway engineering. J. Struct. Biol. 185, 234–242 (2014).
29. Cherf, G. M. & Cochran, J. R. Applications of Yeast Surface Display for Protein Engineering. in Yeast Surface Display (ed. Liu, B.) vol. 1319 155–175 (Springer New York, 2015).
30. Skirgaila, R., Pudzaitis, V., Paliksa, S., Vaitkevicius, M. & Janulaitis, A. Compartmentalization of destabilized enzyme-mRNA-ribosome complexes generated by ribosome display: a novel tool for the directed evolution of enzymes. Protein Eng. Des. Sel. 26, 453–461 (2013).
31. Lipovsek, D. & Plückthun, A. In-vitro protein evolution by ribosome display and mRNA display. J. Immunol. Methods 290, 51–67 (2004).
32. Li, R., Kang, G., Hu, M. & Huang, H. Ribosome display: a potent display technology used for selecting and evolving specific binders with desired properties. https://peerj.com/preprints/26702v1 (2018) doi:10.7287/peerj.preprints.26702v1.

33. Amstutz, P., Forrer, P., Zahnd, C. & Plückthun, A. In vitro display technologies: novel developments and applications. Curr. Opin. Biotechnol. 12, 400–405 (2001).
34. Smith, G. P. Filamentous Fusion Phage: Novel Expression Vectors That Display Cloned Antigens on the Virion Surface. Science 228, 1315–1317 (1985).
35. Clackson, T., Hoogenboom, H. R., Griffiths, A. D. & Winter, G. Making antibody fragments using phage display libraries. Nature 352, 624–628 (1991).
36. Hanes, J. & Plückthun, A. In vitro Selection and Evolution of Functional Proteins by Using Ribosome Display. Proc. Natl. Acad. Sci. U. S. A. 94, 4937–4942 (1997).
37. Gai, S. A. & Wittrup, K. D. Yeast surface display for protein engineering and characterization. Curr. Opin. Struct. Biol. 17, 467–473 (2007).
38. Dong, M. et al. Optimization of production conditions of rice α-galactosidase II displayed on yeast cell surface. Protein Expr. Purif. 171, 105611 (2020).
39. Seelig, B. mRNA display for the selection and evolution of enzymes from in vitro-translated protein libraries. Nat. Protoc. 6, 540–552 (2011).
40. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. Nature 410, 715–718 (2001).
41. Zhang, L., Kang, M., Xu, J. & Huang, Y. Archaeal DNA polymerases in biotechnology. Appl. Microbiol. Biotechnol. 99, 6585–6597 (2015).
42. Rothwell, P. J. & Waksman, G. Structure and mechanism of DNA polymerases. in Advances in Protein Chemistry vol. 71 401–440 (Elsevier, 2005).
43. Patel, P. H. & Loeb, L. A. Getting a grip on how DNA polymerases function. Nat. Struct. Biol. 8, 4 (2001).
44. Ishino, S. & Ishino, Y. DNA polymerases as useful reagents for biotechnology the history of developmental research in the field. Front. Microbiol. 5, (2014).
45. Terpe, K. Overview of thermostable DNA polymerases for classical PCR applications: from molecular and biochemical fundamentals to commercial systems. Appl. Microbiol. Biotechnol. 97, 10243–10254 (2013).
46. Hashimoto, H. et al. Crystal structure of DNA polymerase from hyperthermophilic archaeon Pyrococcus kodakaraensis KOD111Edited by R. Huber. J. Mol. Biol. 306, 469–477 (2001).
47. Hoshino, H. et al. Consecutive incorporation of functionalized nucleotides with amphiphilic side chains by novel KOD polymerase mutant. Bioorg. Med. Chem. Lett. 26, 530–533 (2016).
48. Kuroita, T. et al. Structural Mechanism for Coordination of Proofreading and Polymerase Activities in Archaeal DNA Polymerases. J. Mol. Biol. 351, 291–298 (2005).
49. Kropp, H. M., Betz, K., Wirth, J., Diederichs, K. & Marx, A. Crystal structures of ternary complexes of archaeal B-family DNA polymerases. PLOS ONE 12, e0188005 (2017).
50. LinWu, S.-W. et al. Thermococcus sp. 9°N DNA polymerase exhibits 3′-esterase activity that can be harnessed for DNA sequencing. Commun. Biol. 2, 224 (2019).
51. Wynne, S. A., Pinheiro, V. B., Holliger, P. & Leslie, A. G. W. Structures of an Apo and a Binary Complex of an Evolved Archeal B Family DNA Polymerase Capable of Synthesising Highly Cy-Dye Labelled DNA. PLoS ONE 8, e70892 (2013).

52. Wynne, S. A., Pinheiro, V. B., Holliger, P. & Leslie, A. G. W. Structures of an Apo and a Binary Complex of an Evolved Archeal B Family DNA Polymerase Capable of Synthesising Highly Cy-Dye Labelled DNA. PLoS ONE 8, e70892 (2013).
53. Huber, C. & Marx, A. Variants of sequence family B Thermococcus kodakaraensis DNA polymerase with increased mismatch extension selectivity. PLOS ONE 12, e0183623 (2017).
54. Gardner, A. F. et al. Therminator DNA Polymerase: Modified Nucleotides and Unnatural Substrates. Front. Mol. Biosci. 6, 28 (2019).
55. Nikoomanzar, A., Vallejo, D. & Chaput, J. C. Elucidating the Determinants of Polymerase Specificity by Microfluidic-Based Deep Mutational Scanning. ACS Synth. Biol. 8, 1421–1429 (2019).
56. Killelea, T., Saint-Pierre, C., Ralec, C., Gasparutto, D. & Henneke, G. Anomalous electrophoretic migration of short oligodeoxynucleotides labelled with 5′-terminal Cy5 dyes: Nucleic Acids. ELECTROPHORESIS 35, 1938–1946 (2014).
57. Kropp, H. M., Diederichs, K. & Marx, A. The Structure of an Archaeal B-Family DNA Polymerase in Complex with a Chemically Modified Nucleotide. Angew. Chem. Int. Ed. 58, 5457–5461 (2019).
58. Huber, C., von Watzdorf, J. & Marx, A. 5-methylcytosine-sensitive variants of Thermococcus kodakaraensis DNA polymerase. Nucleic Acids Res. gkw812 (2016) doi:10.1093/nar/gkw812.
59. Chim, N., Shi, C., Sau, S. P., Nikoomanzar, A. & Chaput, J. C. Structural basis for TNA synthesis by an engineered TNA polymerase. Nat. Commun. 8, 1810 (2017).
60. Elshawadfy, A. M. et al. DNA polymerase hybrids derived from the family-B enzymes of Pyrococcus furiosus and Thermococcus kodakarensis: improving performance in the polymerase chain reaction. Front. Microbiol. 5, (2014).
61. Pinheiro, V. B. Engineering-driven biological insights into DNA polymerase mechanism. Curr. Opin. Biotechnol. 60, 9–16 (2019).
62. Horhota, A. et al. Kinetic Analysis of an Efficient DNA-Dependent TNA Polymerase. J. Am. Chem. Soc. 127, 7427–7434 (2005).
63. Sabat, N. et al. Towards the controlled enzymatic synthesis of LNA containing oligonucleotides. Front. Chem. 11, 1161462 (2023).
64. Ellefson, J. W. et al. Synthetic evolutionary origin of a proofreading reverse transcriptase. Science 352, 1590–1593 (2016).
65. Huang, J. et al. A reference human genome dataset of the BGISEQ-500 sequencer. GigaScience 6, (2017).
66. Cyranoski, D. Chinese genomics giant BGI plots commercial path. Nat. Biotechnol. 30, 1159–1161 (2012).
67. Lang, J. et al. Evaluation of the MGISEQ-2000 Sequencing Platform for Illumina Target Capture Sequencing Libraries. Front. Genet. 12, 730519 (2021).
68. Drmanac, S. et al. CoolMPS TM : Advanced massively parallel sequencing using antibodies specific to each natural nucleobase. http://biorxiv.org/lookup/doi/10.1101/2020.02.19.953307 (2020) doi:10.1101/2020.02.19.953307.
69. Jeon, S. A. et al. Comparison between MGI and Illumina sequencing platforms for whole genome sequencing. Genes Genomics 43, 713–724 (2021).

70. Guo, J. et al. Four-color DNA sequencing with 3′- O -modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proc. Natl. Acad. Sci. 105, 9145–9150 (2008).
71. Kropp, H. M., Diederichs, K. & Marx, A. The Structure of an Archaeal B-Family DNA Polymerase in Complex with a Chemically Modified Nucleotide. Angew. Chem. Int. Ed. 58, 5457–5461 (2019).
72. Wu, J. et al. 3′- O -modified nucleotides as reversible terminators for pyrosequencing. Proc. Natl. Acad. Sci. 104, 16462–16467 (2007).
73. Rashid, N. & Aslam, M. An overview of 25 years of research on Thermococcus kodakarensis, a genetically versatile model organism for archaeal research. Folia Microbiol. (Praha) 65, 67–78 (2020).
74. Chen, C.-Y. DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. Front. Microbiol. 5, (2014).
75. Leconte, A. M. et al. Directed Evolution of DNA Polymerases for Next-Generation Sequencing. Angew. Chem. 122, 6057–6060 (2010).
76. Kennedy, E. M., Hergott, C., Dewhurst, S. & Kim, B. The Mechanistic Architecture of Thermostable Pyrococcus furiosus Family B DNA Polymerase Motif A and Its Interaction with the dNTP Substrate. Biochemistry 48, 11161–11168 (2009).
77. Bergen, K., Betz, K., Welte, W., Diederichs, K. & Marx, A. Structures of KOD and 9°N DNA Polymerases Complexed with Primer Template Duplex. ChemBioChem 14, 1058–1062 (2013).
78. Sawai, H. et al. Expansion of structural and functional diversities of DNA using new 5-substituted deoxyuridine derivatives by PCR with superthermophilic KOD Dash DNA polymeraseElectronic supplementary information (ESI) available: sequencing of the PCR products (108mer DNA) from substrates 1 and 2. See http://www.rsc.org/suppdata/cc/b1/b107838k/. Chem. Commun. 2604–2605 (2001) doi:10.1039/b107838k.
79. Hottin, A. & Marx, A. Structural Insights into the Processing of Nucleobase-Modified Nucleotides by DNA Polymerases. Acc. Chem. Res. 49, 418–427 (2016).
80. Chen, F. et al. The History and Advances of Reversible Terminators Used in New Generations of Sequencing Technology. Genomics Proteomics Bioinformatics 11, 34–40 (2013).
81. Bergen, K., Betz, K., Welte, W., Diederichs, K. & Marx, A. Structures of KOD and 9°N DNA Polymerases Complexed with Primer Template Duplex. ChemBioChem 14, 1058–1062 (2013).
82. Schaser, A. J. et al. Alpha-synuclein is a DNA binding protein that modulates DNA repair with implications for Lewy body disorders. Sci. Rep. 9, 10919 (2019).
83. Pacheco, C., Aguayo, L. G. & Opazo, C. An extracellular mechanism that can explain the neurotoxic effects of α-synuclein aggregates in the brain. Front. Physiol. 3, (2012).
84. Fan, T.-S., Liu, S. C.-H. & Wu, R.-M. Alpha-Synuclein and Cognitive Decline in Parkinson Disease. Life 11, 1239 (2021).
85. Mehra, S., Gadhe, L., Bera, R., Sawner, A. S. & Maji, S. K. Structural and Functional Insights into α-Synuclein Fibril Polymorphism. Biomolecules 11, 1419 (2021).
86. Kuo, Y.-M. et al. Extensive enteric nervous system abnormalities in mice transgenic for artificial chromosomes containing Parkinson disease-associated α-synuclein gene mutations precede central nervous system changes. Hum. Mol. Genet. 19, 1633–1650 (2010).
87. Xu, L. & Pu, J. Alpha-Synuclein in Parkinson's Disease: From Pathogenetic Dysfunction to Potential

Clinical Application. Park. Dis. 2016, 1–10 (2016).
88. Meade, R. M., Fairlie, D. P. & Mason, J. M. Alpha-synuclein structure and Parkinson's disease – lessons and emerging principles. Mol. Neurodegener. 14, 29 (2019).
89. Dent, S. E. et al. Phosphorylation of the aggregate-forming protein alpha-synuclein on serine-129 inhibits its DNA-bending properties. J. Biol. Chem. 298, 101552 (2022).
90. Li, S. & Le, W. Milestones of Parkinson's Disease Research: 200 Years of History and Beyond. Neurosci. Bull. 33, 598–602 (2017).
91. Padrick, S. B. & Miranker, A. D. Islet Amyloid: Phase Partitioning and Secondary Nucleation Are Central to the Mechanism of Fibrillogenesis. Biochemistry 41, 4694–4703 (2002).
92. Uversky, V. N. Looking at the recent advances in understanding α-synuclein and its aggregation through the proteoform prism. F1000Research 6, 525 (2017).
93. Zarranz, J. J. et al. The new mutation, E46K, of α-synuclein causes parkinson and Lewy body dementia: New α-Synuclein Gene Mutation. Ann. Neurol. 55, 164–173 (2004).
94. Appel-Cresswell, S. et al. Alpha-synuclein p.H50Q, a novel pathogenic mutation for Parkinson's disease: α- Synuclein p.H50q, A Novel Mutation For Pd. Mov. Disord. 28, 811–813 (2013).
95. Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. Science 357, eaaf4382 (2017).
96. Hyman, A. A., Weber, C. A. & Jülicher, F. Liquid-Liquid Phase Separation in Biology. Annu. Rev. Cell Dev. Biol. 30, 39–58 (2014).
97. Ray, S. et al. α-Synuclein aggregation nucleates through liquid–liquid phase separation. Nat. Chem. 12, 705–716 (2020).
98. Mukherjee, S. et al. Liquid-liquid Phase Separation of α-Synuclein: A New Mechanistic Insight for α-Synuclein Aggregation Associated with Parkinson's Disease Pathogenesis. J. Mol. Biol. 435, 167713 (2023).
99. Mukherjee, S. et al. Liquid-liquid Phase Separation of α-Synuclein: A New Mechanistic Insight for α-Synuclein Aggregation Associated with Parkinson's Disease Pathogenesis. J. Mol. Biol. 167713 (2022) doi:10.1016/j.jmb.2022.167713.
100. Sawner, A. S. et al. Modulating α-Synuclein Liquid–Liquid Phase Separation: Published as part of the Biochemistry virtual special issue "Protein Condensates". Biochemistry 60, 3676–3696 (2021).
101. Srinivasan, E. et al. Alpha-Synuclein Aggregation in Parkinson's Disease. Front. Med. 8, 736978 (2021).

6 Appendix A

Data Set: Screening Results of Medium-Sized Preliminary Library (Part II) in Chapter I

(1): The kinetic performance of each KOD variant was calculated with equation 1 (eq. 1) of Part II. The kinetic was measured by enzyme kinetic assay (details in the **Materials and Methods** section of Part II). The data presented represent the mean values obtained from independent experiments.

Mutants	Mutation sites	E(Mut)/ P/T-2Cy5 (1)	Mutants	Mutation Sites	E(Mut)/ P/T-2Cy5 (1)
MPL_1	A485I	2.9	MPL_147	Y496M	1.3
MPL_2	S707G	1.1	MPL_148	Y496S	0.2
MPL_3	S707L	0.6	MPL_149	K531E	0.5
MPL_4	S707K	0.7	MPL_150	E485I+A490H	0.9
MPL_5	S707T	1.0	MPL_151	E485I+R379H	2.7
MPL_6	K705Q	0.8	MPL_152	E485I+I610G	0.9
MPL_7	L630F	0.6	MPL_153	E485I+T605V	2.8
MPL_8	I610L	0.8	MPL_154	E485I+T605W	3.1
MPL_9	I610V	0.5	MPL_155	E485I+S347V	2.2
MPL_10	E385R	0.5	MPL_156	E485I+R379Y	3.0
MPL_11	A574L	0.7	MPL_157	E485I+K477E	5.9
MPL_12	A574E	0.9	MPL_158	E485I+T470S	1.9
MPL_13	A574H	0.7	MPL_159	E485I+K465Q	2.6
MPL_14	K531R	0.5	MPL_160	E485I+Q461A	3.8
MPL_15	L367M	0.6	MPL_161	E485I+G454Q	1.3
MPL_16	T723I	0.8	MPL_162	E485I+L453V	2.4
MPL_17	T723S	0.8	MPL_163	D480N+D718Q	0.9
MPL_18	T723G	0.9	MPL_164	D480N+K584T	1.4
MPL_19	T723F	0.7	MPL_165	D480N+R379Y	0.0
MPL_20	D718T	0.9	MPL_166	D480N+T267D	1.3
MPL_21	K726S	0.5	MPL_167	D480N+L453V	2.1
MPL_22	K726D	0.6	MPL_168	K465Q+D718Q	2.6
MPL_23	T723C	0.8	MPL_169	K465Q+S707G	1.2
MPL_24	D718Q	1.0	MPL_170	K465Q+K584T	1.6
MPL_25	T723V	0.9	MPL_171	R379Y+R380D	0.7

MPL_26	I610K	0.5	MPL_172	S451Q+L453V	1.3
MPL_27	I610H	0.8	MPL_173	D480N+R482K	1.3
MPL_28	I610G	0.4	MPL_174	L457M+Q461A	1.5
MPL_29	E609Y	0.0	MPL_175	S451Q+E485I	3.2
MPL_30	E609F	0.0	MPL_176	D480N+S451Q	2.3
MPL_31	E609G	0.0	MPL_177	D480N+E485I	2.8
MPL_32	E609A	0.0	MPL_178	Y493H+Y497W	0.7
MPL_33	T605V	0.4	MPL_179	Y493L+Y497W	1.0
MPL_34	T605M	0.2	MPL_180	Y493H+Y497L	1.6
MPL_35	T605W	0.3	MPL_181	Y493H+Y497F	1.7
MPL_36	T605C	0.2	MPL_182	Y493H+Y497M	1.2
MPL_37	F545T	0.4	MPL_183	Y493H+Y497V	0.7
MPL_38	F545Q	0.9	MPL_184	Y493F+Y497L	7.3
MPL_39	F545K	0.8	MPL_185	Y493F+Y497I	1.4
MPL_40	F545I	0.4	MPL_186	Y493F+Y497F	1.0
MPL_41	D542S	0.0	MPL_187	Y493F+Y497M	1.7
MPL_42	D542R	0.0	MPL_188	K676V+M680T	0.4
MPL_43	D542K	0.0	MPL_189	K676A+M680T	0.8
MPL_44	S539H	0.0	MPL_190	K676I+M680V	0.6
MPL_45	N491T	0.0	MPL_191	K676M+M680V	0.6
MPL_46	N491S	0.2	MPL_192	K676R+M680V	0.5
MPL_47	N491I	0.0	MPL_193	K676H+M680V	0.4
MPL_48	N491E	0.0	MPL_194	K676L+M680V	0.7
MPL_49	A490L	0.0	MPL_195	K676V+M680V	1.0
MPL_50	A490F	0.3	MPL_196	K676V+V682M	1.2
MPL_51	A490D	0.2	MPL_197	K676I+M680L	2.0
MPL_52	R379H	0.2	MPL_198	K676I+M680I	1.2
MPL_53	R379I	0.3	MPL_199	K676S+M680I	9.4
MPL_54	R379D	0.9	MPL_200	K676S+M680L	0.9
MPL_55	S347V	0.8	MPL_201	K676M+M680L	1.6
MPL_56	S347Q	0.9	MPL_202	K676R+M680I	6.0
MPL_57	S347I	0.0	MPL_203	K676L+M680L	4.8
MPL_58	S539Q	0.9	MPL_204	K676H+M680I	3.9
MPL_59	A490H	0.2	MPL_205	K676A+M680L	1.3
MPL_60	E609Q	0.9	MPL_206	K676V+M680I	2.1

MPL_61	E609I	0.0	MPL_207	K676V+M680L	2.4
MPL_62	G386W	0.0	MPL_208	K676I+M680A	0.7
MPL_63	G386N	0.8	MPL_209	K676M+M680A	2.0
MPL_64	G386F	0.0	MPL_210	K676I+M680T	1.1
MPL_65	N351M	0.0	MPL_211	K676S+M680A	1.5
MPL_66	N351Y	0.9	MPL_212	K676L+M680T	1.3
MPL_67	G387L	0.6	MPL_213	K676H+M680T	2.2
MPL_68	G386V	0.5	MPL_214	K676L+M680A	1.9
MPL_69	E385F	1.0	MPL_215	E325N+F326T	0.8
MPL_70	Q382L	0.4	MPL_216	E325Q+F326T	0.8
MPL_71	R381N	0.9	MPL_217	E325H+F326T	0.9
MPL_72	R381G	0.9	MPL_218	E325H+F326Y	0.8
MPL_73	R380Y	1.1	MPL_219	E325D+F326Y	0.8
MPL_74	R380D	1.3	MPL_220	E325N+F326Y	0.8
MPL_75	R379Y	1.4	MPL_221	E325Q+F326Y	0.7
MPL_76	R379V	0.2	MPL_222	E325K+F326Y	0.7
MPL_77	V353D	1.1	MPL_223	E325D+F326V	0.7
MPL_78	N351S	0.6	MPL_224	E325Q+F326I	0.4
MPL_79	N351C	0.6	MPL_225	E325Q+F326V	0.6
MPL_80	G350R	0.2	MPL_226	E325N+F326V	0.7
MPL_81	G350I	0.7	MPL_227	E325D+F326M	0.9
MPL_82	T349G	1.0	MPL_228	V353D+Y496T	1.4
MPL_83	T349A	0.6	MPL_229	R379K+R380G	1.1
MPL_84	S348N	0.4	MPL_230	R379E+R380D	0.9
MPL_85	S347R	0.8	MPL_231	R379Q+R380N	1.1
MPL_86	S347D	1.4	MPL_232	R379E+R380K	0.9
MPL_87	P271N	0.7	MPL_233	R379Q+R380K	0.6
MPL_88	P271I	0.7	MPL_234	R379Q+R380S	1.0
MPL_89	L270N	0.8	MPL_235	R379E+R380G	0.8
MPL_90	T267D	1.7	MPL_236	R379K+R380K	0.8
MPL_91	T267E	0.8	MPL_237	R379E+R380S	1.2
MPL_92	L489M	0.7	MPL_238	R379K+R380S	1.0
MPL_93	I488V	0.9	MPL_239	R379Q+R380G	0.9
MPL_94	I488M	0.8	MPL_240	R379K+R380E	1.8
MPL_95	R484K	0.7	MPL_241	R379Q+R380D	1.4

MPL_96	R482K	2.2	MPL_242	K375E+A378S	0.4
MPL_97	R482E	0.4	MPL_243	K375A+A378S	0.9
MPL_98	Y481F	0.3	MPL_244	K375A+A378R	0.7
MPL_99	D480N	3.9	MPL_245	K375T+A378R	0.7
MPL_100	L479M	1.1	MPL_246	K375N+A378S	0.7
MPL_101	L479I	0.6	MPL_247	K375G+A378K	1.0
MPL_102	L478M	0.8	MPL_248	K375E+A378K	1.4
MPL_103	L478I	0.5	MPL_249	K375A+A378E	1.0
MPL_104	K477R	0.7	MPL_250	K375E+A378E	1.0
MPL_105	K477E	0.7	MPL_251	K375T+A378E	0.9
MPL_106	R476F	0.5	MPL_252	K375I+A378Q	1.2
MPL_107	R476H	0.4	MPL_253	K375A+A378Q	1.0
MPL_108	E475Q	0.5	MPL_254	K375T+A378K	0.9
MPL_109	E475D	0.8	MPL_255	R379H+R380Y	0.7
MPL_110	I474M	0.9	MPL_256	R379H+R380F	0.7
MPL_111	I474E	0.7	MPL_257	R379Y+R380Y	0.7
MPL_112	D472E	0.5	MPL_258	R379Q+R380F	0.6
MPL_113	D472N	0.7	MPL_259	R379K+R380F	0.7
MPL_114	I471A	0.7	MPL_260	R379Q+R380Y	0.8
MPL_115	I471L	0.4	MPL_261	R379E+R380F	0.8
MPL_116	T470S	0.7	MPL_262	R379K+R380Y	0.6
MPL_117	T470Q	0.4	MPL_263	R379H+R380N	0.8
MPL_118	A469S	0.5	MPL_264	R379Y+R380G	0.7
MPL_119	A469M	0.4	MPL_265	R379H+R380G	0.9
MPL_120	K466R	0.5	MPL_266	R379C+R380S	0.5
MPL_121	K466N	0.9	MPL_267	R379H+R380D	0.9
MPL_122	K465R	0.6	MPL_268	H589R+T590S	0.8
MPL_123	K465Q	1.3	MPL_269	H589G+T590S	0.5
MPL_124	I463V	0.7	MPL_270	H589S+T590S	0.5
MPL_125	K462R	0.8	MPL_271	H589L+T590N	0.4
MPL_126	K462D	0.5	MPL_272	H589I+T590N	0.4
MPL_127	Q461S	1.1	MPL_273	H589R+T590N	0.5
MPL_128	Q461A	1.3	MPL_274	K676L+M680I	6.4
MPL_129	E459S	0.7	MPL_275	K676A+M680I	1.2
MPL_130	E459D	0.7	MPL_276	I474T+L478R	0.9

MPL_131	E458Q	0.4	MPL_277	I474T+L478N	1.8
MPL_132	E458D	0.7	MPL_278	I474T+L478D	0.0
MPL_133	L457M	1.0	MPL_279	I474T+L478C	2.2
MPL_134	L457I	1.1	MPL_280	I474T+L478I	0.9
MPL_135	D455S	0.5	MPL_281	I474T+L478G	1.6
MPL_136	D455E	0.2	MPL_282	I474T+L478H	1.7
MPL_137	G454Q	1.0	MPL_283	I474K+L478Y	1.3
MPL_138	G454N	0.7	MPL_284	I474K+L478C	0.0
MPL_139	L453V	1.8	MPL_285	I474K+L478F	0.9
MPL_140	L453I	1.4	MPL_286	I474K+L478H	1.4
MPL_141	L452V	1.1	MPL_287	I474K+L478G	1.2
MPL_142	S451Q	2.3	MPL_288	I474K+L478V	1.1
MPL_143	S451N	0.7	MPL_289	I474K+L478I	1.0
MPL_144	A467S	2.7	MPL_290	H589Q+T590K	0.5
MPL_145	A467M	1.7	MPL_291	H589D+T590K	0.7
MPL_146	A467Q	0.7	MPL_292	K674Q+M680V	0.8
MPL_293		D480N+S451Q+D718Q			2.3
MPL_294		D480N+S451Q+Q461A			1.9
MPL_295		D480N+S451Q+K465Q			1.5
MPL_296		D480N+S451Q+A485I			2.0
MPL_297		D480N+S451Q+R379Y			1.7
MPL_298		R484G+EI485+I486I			1.2
MPL_299		R484G+EI485I+I486L			0.8
MPL_300		K465Q+D718Q+S707G			1.2
MPL_301		K465Q+D718Q+485I			2.9
MPL_302		K465Q+D718Q+R379Y			1.3
MPL_303		K465Q+D718Q+T267D			0.0
MPL_304		K465Q+D718Q+D480N			2.0
MPL_305		K465Q+D718Q+L453V			1.2
MPL_306		K465Q+K584T+S707G			0.0
MPL_307		K465Q+K584T+R379Y			1.2
MPL_308		K465Q+K584T+T267D			1.1
MPL_309		K465Q+K584T+D480N			0.9
MPL_310		E485I+K676L+M680L			1.1
MPL_311		K465Q+K584T+L453V			1.2

MPL_312	K465Q+K584T+485I	3.0
MPL_313	L457M+Q461A+T267D	3.0
MPL_314	L457M+Q461A+485I	2.6
MPL_315	L479M+D480N+R482K	0.7
MPL_316	L453V+L452V+G454Q	0.7
MPL_317	S451Q+L453V+L452V	1.3
MPL_318	L457M+Q461A+K465Q	1.6
MPL_319	E485I+K676H+M680I	1.5
MPL_320	E485I+K676L+M680I	1.0
MPL_321	K676L+M680T+V698I	3.3
MPL_322	E325Q+F326T+E426G	0.9
MPL_323	R379K+R380D+A551V	1.0
MPL_324	E485I+K676S+M680I	2.4
MPL_325	S451Q+R482K+K676T	1.1
MPL_326	Y493F+Y497L+E485I	2.4
MPL_327	S451Q+K674L+M680V	0.9
MPL_328	S451Q+R482K+M680L	0.8
MPL_329	D480N+K674Q+M680I	1.6
MPL_330	S451Q+K674M+K676T	1.3
MPL_331	S451Q+L453V+K676T	1.0
MPL_332	K465Q+D718Q+L457M+Q461A	0.0
MPL_333	K465Q+D718Q+D480N+E485I	1.8
MPL_334	K465Q+K584T+L457M+Q461A	0.4
MPL_335	D480N+E485I+K674Q+M680I	1.0
MPL_336	S451Q+R482K+K674Q+M680L	1.1
MPL_337	S451Q+L479M+R482K+K674M+M680L	1.1
MPL_338	E485I+K676L+M680T+V698I	1.7
MPL_339	S451Q+L452V+K674L+M680L	1.0
MPL_340	S451Q+L452V+L479M+D480N+K674H+M680I	1.0
MPL_341	S451Q+L453V+K674L+M680V	0.8
MPL_342	S451Q+L453I+R482K+K574L+M680L	0.8
MPL_343	S451Q+R482K+K674M+M680V	0.8
MPL_344	S451Q+L452V+L453I+L479M+K674L+M680V	0.9
MPL_345	L479M+D480N+R482K+K676T	0.6
MPL_346	S451Q+L452V+D480N+R482K+K674Q+M680L	1.5

MPL_347	K676S+M680I+Y493F+Y497L	1.5
MPL_348	K465Q+ D718Q+D480N+E485I	1.3
MPL_349	K465Q+ K584T+L457M+Q461A	0.7