Bioresource Technology 438 (2025) 133206

Contents lists available at ScienceDirect

Bioresource Technology

journal homepage: www.elsevier.com/locate/biortech

Protein language model-assisted directed evolution of cyclodextrinase Enables Precision α-*O*-Oligosaccharide synthesis

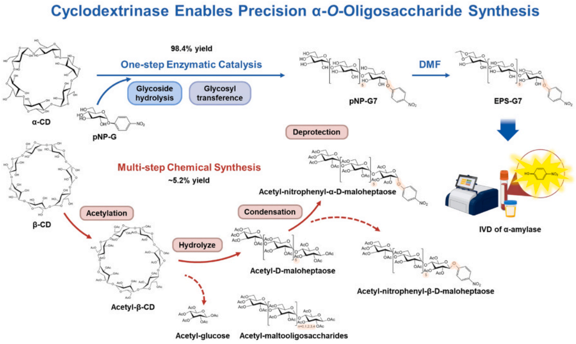
Ting Nie [a,b,1], Zhenxin Yan [c,1], Hao Liu [d], Xiudian Zhang [c], Weiwei Zhong [e], Qin Chen [e], Changbin Zhu [c], Liang Hong [f], Guang-yu Yang [a,b,*]

[a] *State Key Laboratory of Microbial Metabolism, Joint International Research Laboratory of Metabolic & Developmental Sciences, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China*
[b] *Institute of Key Biological Raw Material, Shanghai Academy of Experimental Medicine, Shanghai 201401, China*
[c] *Hzymes Biotechnology Co. Ltd., Hubei 430010, China*
[d] *Shanghai Matwings Technology Co., Ltd., Shanghai 200241, China*
[e] *Department of Food Science and Engineering, Zhejiang-Malaysia Joint Research Laboratory for Agricultural Product Processing and Nutrition, Zhejiang Provincial Key Laboratory of Animal Protein Food Intensive Processing Technology, Ningbo University, Ningbo 315800, China*
[f] *Shanghai National Center for Applied Mathematics (SJTU Center), & Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai 200240, China*

HIGHLIGHTS

- CDase: synthesis of α-*O*-oligosaccharides with controlled polymerization degrees.
- Novel CDase mined *via* ESBS probe, minimizes hydrolysis for transglycosylation.
- AI optimization: 8 × transglycosylation, 33.3 % less hydrolysis, 98.4 % selectivity.
- Industrial pNP-G7: 161 g/L (85 % yield), surpassing chemical synthesis (5.2 %).
- Engineered CDase offers a potential novel platform for transglycosylation.

GRAPHICAL ABSTRACT


Cyclodextrinase Enables Precision α-O-Oligosaccharide Synthesis
98.4% yield
One-step Enzymatic Catalysis
Glycoside hydrolysis
Glycosyl transference
α-CD
pNP-G
pNP-G7
DMF
EPS-G7
Deprotection
Multi-step Chemical Synthesis
~5.2% yield
Acetyl-nitrophenyl-α-D-maloheptaose
IVD of α-amylase
β-CD
Acetylation
Condensation
Hydrolyze
Acetyl-D-maloheptaose
Acetyl-nitrophenyl-β-D-maloheptaose
Acetyl-β-CD
Acetyl-glucose
Acetyl-maltooligosaccharides

ARTICLE INFO

Keywords:
α-*O*-oligosaccharide
Cyclodextrinase
Motif probe
General protein language model
Multi-objective optimization

ABSTRACT

Stereoselective synthesis of α-O-oligosaccharides remains a key challenge in glycobiology. While glycoside hydrolase-mediated transglycosylation is promising, current methods yield excessive byproducts and show low specificity. Here, we establish a glycoside hydrolase-based system for precise oligosaccharide synthesis using cyclodextrin as donor. Guided by an extra sugar binding space (ESBS) motif probe, a cyclodextrinase from *Paenibacillus* sp. MY03 was identified with a naturally high transglycosylation-to-hydrolysis (T/H) ratio. Using Pro-PRIME, a protein language model, we optimized three enzymatic properties—enhancing transglycosylation, reducing hydrolysis, and improving regioselectivity—based on minimal beneficial mutation data. Among 68 screened variants, the top mutant showed a 12-fold higher T/H ratio and improved 4-nitrophenyl-α-ᴅ-

[*] Corresponding author at: State Key Laboratory of Microbial Metabolism, Joint International Research Laboratory of Metabolic & Developmental Sciences, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China.

E-mail address: yanggy@sjtu.edu.cn (G.-y. Yang).

[1] Contributed equally to this work.

https://doi.org/10.1016/j.biortech.2025.133206
Received 31 May 2025; Received in revised form 22 August 2025; Accepted 23 August 2025
Available online 25 August 2025

maltoheptaoside (pNP-G7) yield from 63 % to 98 %. The engineered enzyme also showed broad substrate promiscuity, underscoring its utility for diverse biotechnological applications. This study advances α-O-oligosaccharide synthesis and demonstrates the power of language model-guided enzyme engineering to balance competing catalytic activities.

1. Introduction

The α-*O*-glycosylation reaction enhances the solubility, stability, bioavailability, and functionality of aglycones. Specifically, glycosylated oligosaccharides, such as tocopherol, capsaicin, and paclitaxel, have enhanced properties (Boltje et al., 2009; Shimoda et al., 2009; Shimoda et al., 2008; Shimoda et al., 2012). However, achieving stereoselective synthesis of α-glycosides, particularly α-oligosaccharides, remains a challenge in chemical synthesis (Wang & Huang, 2009). Enzymatic transglycosylation, which transfers glycosyl groups from donors to acceptors (Hatanaka et al., 1994), is favored for its simplicity, specificity, low cost, and mild conditions (Huang & Flitsch, 2019; Napiórkowska et al., 2017). Cyclodextrin glucanotransferase (CGTase, EC 2.4.1.19) is commonly used in oligosaccharide transglycosylation, where it transfers cyclic oligosaccharides (cyclodextrins) to acceptors through a ternary complex mechanism. However, CGTase also catalyzes cyclization, disproportionation, and hydrolysis, leading to numerous byproducts and limiting the synthesis of specific oligosaccharides with defined polymerization degrees (Lim et al., 2021). In contrast, cyclodextrinases (CDases, EC 3.2.1.54) from the glycoside hydrolase family 13 (GH13) have a natural preference for hydrolysis over cyclization while retaining a certain degree of transglycosylation activity, which may offer a potential new platform for transglycosylation (Park et al., 2000).

Some glycoside hydrolases (GHs) (e.g., CDase) can form glycosidic bonds via a double displacement mechanism, producing a covalent glycosyl intermediate (Chen & Arnold, 2020; Tao et al., 1989). The specificity of enzyme-catalyzed reactions is determined by the acceptor substrate: a water molecule in hydrolysis and a sugar hydroxyl group in transglycosylation (Iwamoto et al., 2017; Mangas-Sánchez & Adlercreutz, 2015). GH13-family cyclodextrinases exhibit distinct structural features that underpin their catalytic specificity. Their catalytic core adopts a canonical $(\beta/\alpha)_8$ barrel architecture containing the conserved Asp-His-Glu triad, while a unique *N*-terminal domain forms a dimerization-mediated substrate-binding pocket (Park et al., 2000). This structural configuration confers high selectivity for cyclic oligosaccharides: the deep, hydrophobic substrate-binding cleft accommodates cyclodextrins, while the narrow entrance excludes linear polysaccharides (Park et al., 2000). Kinetic analyses indicate moderate substrate affinity for α-cyclodextrin ($K_M = 0.1$–1 mM) and high catalytic turnover ($k_{cat} =$ 100–500 s^{-1}), positioning these enzymes as highly efficient biocatalysts for oligosaccharide synthesis (Park et al., 2000).

Transglycosylation reactions of GHs have enabled the *in vitro* synthesis of specific oligosaccharides (Iwamoto et al., 2017). For example, maltogenic amylases from *Geobacillus stearothermophilus* catalyze transglycosylation with acarbose and other acceptors to yield specific carbohydrate inhibitors (Hwa Park et al., 1998). GH-mediated transglycosylation has emerged as an effective enzymatic strategy for branched oligosaccharide production from starch (Kang et al., 1997). In this context, glycoside hydrolase (GH)-mediated transglycosylation is a valuable enzymatic strategy for oligosaccharide synthesis (Iwamoto et al., 2017; Mangas-Sánchez & Adlercreutz, 2015). Within this framework, GH13 cyclodextrinases (CDases)—traditionally known for cyclodextrin hydrolysis—also exhibit transglycosylation activity (Aroob et al., 2021).˙ This dual functionality is attributed to an auxiliary region within the active-site cleft termed the extra sugar-binding space (ESBS), which accommodates acceptor sugars such as maltose and competes with water as a nucleophile (Park et al., 2000). The ESBS's hydrophobicity significantly influences this competition: increased hydrophobicity reduces water accessibility, favoring sugar-mediated glycosidic bond formation and enhancing the transglycosylation-to-hydrolysis (T/H) ratio (Aroob et al., 2021). However, using inexpensive cyclodextrins as donors for synthesizing oligosaccharides of defined polymerization remains underexplored (Aroob et al., 2021). Additionally, engineering CDase to simultaneously achieve high transglycosylation activity, low hydrolytic activity, and strict regioselectivity remains a significant challenge in enzyme-directed evolution. Therefore, CDase-mediated precise oligosaccharide synthesis represents a key area for research.

Ethylidene-4-nitrophenyl-α-ᴅ-maltoheptaoside (EPS-G7) is a critical α-*O*-oligosaccharide substrate for enzyme-linked assays detecting α-amylase in blood and urine, necessary for diagnosing pancreatic disorders (Lorentz, 2000). Its chemical synthesis is complicated by multiple reaction steps and complex separation (Zhong et al., 2023). Here, we developed a novel CDase that synthesizes oligosaccharides with a specific polymerization degree using cyclodextrin as a substrate. By ESBS sequence motif probe-guided mining, we identified a CDase with low hydrolytic activity. Leveraging prior knowledge and a protein language model, we optimized CDase for enhanced transglycosylation, minimized hydrolysis, and improved regioselectivity, resulting in the efficient production of 4-nitrophenyl-α-ᴅ-maltoheptaoside (PNP-G7), a precursor to EPS-G7. Finally, we elucidated the mechanism underlying the high transglycosylation/hydrolysis ratio of the optimized mutant and explored its activity with various substrates. This strategy represents an innovative approach for screening and evolving multifunctional enzymes for the synthesis of oligosaccharide glycosides.

2. Materials and methods

2.1. Reagents, strains, and plasmids

α-Cyclodextrin (α-CD) was purchased from Shanghai YuanYe. 4-nitrophenyl-α-ᴅ-glucopyranoside (pNP-G) was obtained from Jinan Shanmu Biotechnology Co., Ltd. The candidate CDases genes were synthesized by Genewiz (Suzhou, China). Peptone and yeast extract were acquired from Macklin. Restriction endonucleases and DNA ligase were procured from NEB. All inorganic salt reagents, of analytical grade with purity > 99.0 %, were supplied by Sinopharm Chemical Reagent Co., Ltd. The pET-28a(+) vector and *Escherichia coli* competent cells were purchased from Qingke Biotechnology. Sodium phosphate buffer (pH 7.5): 81 mL of 0.1 M disodium hydrogen phosphate was mixed uniformly with 19 mL of 0.1 M sodium dihydrogen phosphate.

2.2. Database mining of novel CDases

To discover such CDases, we performed a systematic search of the GenBank non-redundant database for sequences containing GH13 family-conserved regions (CSRs I–IV) and the CDase-specific CSRs VI–VII, using the "MPKLN" motif as a selection filter (Henrissat & Davies, 1997; Janecek, 1997). Candidate enzymes were further screened by calculating ESBS residue hydrophobicity scores. Iterative threshold adjustments were applied to optimize the detection of functionally relevant CDases. Final hydrophobicity cut-offs were set at 0.8 for ESBS-1 and 0.4 for ESBS-3, yielding 104 putative CDases. Multiple-sequence alignment was done using GREMLIN (https://gremlin.bakerlab.org/submit.php) and MAFFT v7.525.

2.3. Construction of mutant CDase production strains

The candidate CDases genes were inserted into the *Eco*R I and *Hind* III restriction sites of the pET-28a(+) plasmid, with the TAA sequence retained and no downstream His tag. Point mutants were constructed using plasmid pET-28a-cd as a template, while multiple mutants were constructed using plasmid pET-28a-cd with single or multiple mutants as templates. The design of mutant-related PCR primer sequences can be found in the supplementary materials. PCR was performed with primers on a Thermal Cycler instrument (Shanghai, China). The PCR products were treated with *Dpn* I and then inserted into *E. coli* DH5α for amplification. Plasmids containing the desired mutations were confirmed by DNA sequencing, and then inserted into *E. coli* BL21(DE3) for CDases production.

2.4. Production and Purification of CDase wild type and variants

The recombinant *E. coli* BL21 (DE3) strain was inoculated into 100 mL of fresh LB medium with 50 mg/L kanamycin and cultured until reaching an OD600 of 0.8. Protein expression was induced by adding a final concentration of 0.1 mM IPTG, followed by a 12-hour incubation at 18 °C. Cells were subsequently harvested through centrifugation at 10,000 × *g* for 20 min at 4 °C, washed with saline, and resuspended in 10 mL of sodium phosphate buffer (50 mM sodium phosphate buffer, pH 7.4, containing 20 mM imidazole and 500 mM NaCl). Cell disruption was performed using ultrasonic treatment (total duration of 5 min, with 2-second pulses and 3-second intervals at 25 % amplitude). After removing cellular debris by centrifugation (10,000 × *g*, 20 min, 4 °C), both His-tagged wild-type and mutant CDases were purified through immobilized metal affinity chromatography. The crude enzyme solutions were applied to Ni-NTA columns (GE, Shanghai) with sequential elution protocols. Initial elution removed contaminants using buffer A (25 mM Tris-HCl, 500 mM NaCl, pH 7.4) containing progressively increasing imidazole concentrations (0, 15, 30, and 45 mM). Target proteins were subsequently eluted with buffer B (buffer A containing 300 mM imidazole). Purified enzyme fractions were concentrated with 30-kDa cutoff ultrafiltration devices (Millipore, Shanghai) and dialyzed into 50 mM sodium phosphate buffer (pH 7.5). The protein concentration was determined using a Nanodrop 2000 spectrophotometer (Thermo Electron Co.) with an extinction coefficient of 1.701 (calculated via ExPASy ProtParam).

2.5. Cdase activity assays

Definition of Enzyme Activity Unit: One unit of enzyme activity is defined as the amount of enzyme required to hydrolyze 4-nitrophenyl-α-D-glucopyranoside (pNP-G) to produce 1 mmol of pNP per minute under the conditions of 40 °C, pH 7.5. The enzymatic activity of the recombinant cyclodextrinase was assayed in a 200 μL reaction mixture comprising 20 μL of 100 mM p-nitrophenyl-α-D-glucopyranoside (pNPG), 20 μL of 10 mg/mL enzyme solution (final enzyme concentration: 1 mg/mL) and 160 μL of 50 mM sodium phosphate buffer (pH 7.5). Reactions were incubated at 40 °C. After 15 min, the reactions were terminated using 200uL of 2 M Na_2CO_3. The samples were then collected, and their absorbance at 405 nm was measured using a UV spectrophotometer. The amount of pNP produced within the time interval (Δt) was calculated and used to determine enzyme activity by referencing an external standard curve, while setting a negative control at the same time. External Standard Curve Preparation: A 1 mM pNP stock solution was diluted 10-fold with 1 M Na_2CO_3, followed by two-fold serial dilutions with 1 M Na_2CO_3 to produce five different concentrations (2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, 2–5-fold). Starting from the lowest concentration, the absorbance of each pNP standard at 405 nm was measured using a microvolume spectrophotometer.

To determine the optimal time point for pNP-G7 production, a time-course analysis was conducted by comparing its peak area and its proportion among total reaction products. The HPLC detection conditions and gradient program are detailed (see supplementary materials). Transglycosylation activity was quantified as the cumulative peak area of all pNP-Gn products generated per minute. Hydrolytic activity was evaluated by the ratio of maltohexose (Glc6) to pNP-G7 peak areas. Regioselectivity was assessed by the proportion of the pNP-G7 peak area relative to the total pNP-Gn product peak area. The transglycosylation-to-hydrolysis ratio, calculated as the ratio of transglycosylation activity to hydrolytic activity, was used to evaluate catalytic preference.

2.6. Optimization of the production process for 4-nitrophenyl-α-D-maltoheptaoside by MY03 CDase

By adding 20, 30, 40, and 50 μL of 10 mg/mL CDase, yielding final concentrations of 0.54, 0.81, 1.08, and 1.35 mg/mL respectively, the relative maximum production range of pNP-G7 and the optimal enzyme concentration were investigated within a fixed reaction time. Using the optimal enzyme dosage, the effects of sodium phosphate buffers at different pH values (6.0, 6.5, 7.0, 7.5, 8.0, and 8.5) on the enzymatic reaction were validated to identify the optimal pH for maximizing pNP-G7 production and stabilizing the enzyme's activity. Subsequently, various buffer salts within the same pH value (pH = 7.5 ± 0.1) were prepared, including Tricine, TEA (triethanolamine), disodium hydrogen phosphate-citric acid, sodium dihydrogen phosphate-phosphoric acid, potassium dihydrogen phosphate-sodium hydroxide, Tris-HCl, to monitor the influence of different buffer salts on pNP-G7 accumulation. Additionally, metal ions (K^+, Li^+, Mg^{2+}, Ca^{2+}, Ba^{2+}, Cu^{2+}, Zn^{2+}, Mn^{2+}, and Fe^{3+}) and EDTA were selected as supplements and added to the reaction system at a final concentration of 10 mmol/L to investigate their effects on the reaction. Based on the summarized experimental data, we optimized the reaction substrates to determine the optimal addition amounts of α-CD and pNP-G, as well as production system conditions. The CDase activity assay used in this section refers to section 2.5.

2.7. Chemical synthesis of Ethylidene-4-nitrophenyl-α-D-maltoheptaoside using ethyl encapsulation

At room temperature, 150 g of pNP-G7, 41 g of *p*-toluenesulfonic acid, and 1.5 L of DMF were added to a 3 L three-neck flask, and the mixture was stirred evenly. A total of 21 g of glycolaldehyde was added, and the reaction mixture was protected with Ar and stirred. The reaction was carried out at 50 °C for 16 h. A total of 1.5 L of methyl *tert*-butyl ether was added to the reaction mixture and stirred with a paddle until the solution became clear, and a yellow oily solid precipitated out. The solution was decanted, and 1.5 L of MeOH was added and stirred for 1 h. Then, a mixture of 1.5 L of iPrOH and 3 L of PE was added, and the mixture was stirred for 3 h. The solid was collected by filtration, washed four times by resuspending and refiltering, and checked for the complete removal of *p*-toluenesulfonic acid by HPLC. The white solid EPS-G7 was collected after filtration and resuspension.

2.8. Pro-PRIME-assisted multi-point mutation prediction targeting multiple objectives

We trained models for three metrics based on a dataset of 68 single-point mutations: transglycosylation activity, hydrolysis activity, and the ratio of pNP-G7 to total products. These metrics were normalized using z-scores. During training, each labeled mutant sequence was transformed into a latent space representation (L × 1280) utilizing Pro-PRIME's transformer encoder, where L denotes the protein length (Jiang et al., 2024). The resulting 1280-dimensional amino acid vectors were averaged to generate a representative vector for each sequence. This vector served as input for a regression head comprising a multi-layer perceptron (MLP), a dropout layer, and a Tanh activation function, yielding a single prediction score for each mutant. The Adam optimizer was employed with mean squared error (MSE) as the loss function, a maximum of 200 training epochs, and early stopping applied if MSE did not improve over 10 epochs. All training was done on a single 80 GB

A100 GPU, using a learning rate of using a learning rate of 1×10^{-5} and a batch size of 4. Finally, we ranked the multi-point mutants, derived from combinations of the aforementioned 68 single-point mutations, based on a composite score calculated from transglycosylation activity plus the pNP-G7/products ratio, minus hydrolysis activity. Our screening process prioritized mutants with higher predicted transglycosylation activity, lower hydrolysis activity, and higher pNP-G7/products scores.

2.9. MD analysis

Using AlphaFold-Multimer (Jumper et al., 2021), the wild-type homodimer of MY03 CDase was predicted and scored for its structure. The three-dimensional structure of the wild-type MY03 CDase was used as a template, along with Rosetta software (Leaver-Fay et al., 2011), to generate a three-dimensional model of the CDase variant. GROMACS-2023(Páll et al., 2020) was utilized for classical molecular dynamics (MD) simulation analysis, simulating 20 ns at 298 K using CHARMM27 (protein residues) and TIP3P (water molecules). The trajectory of the

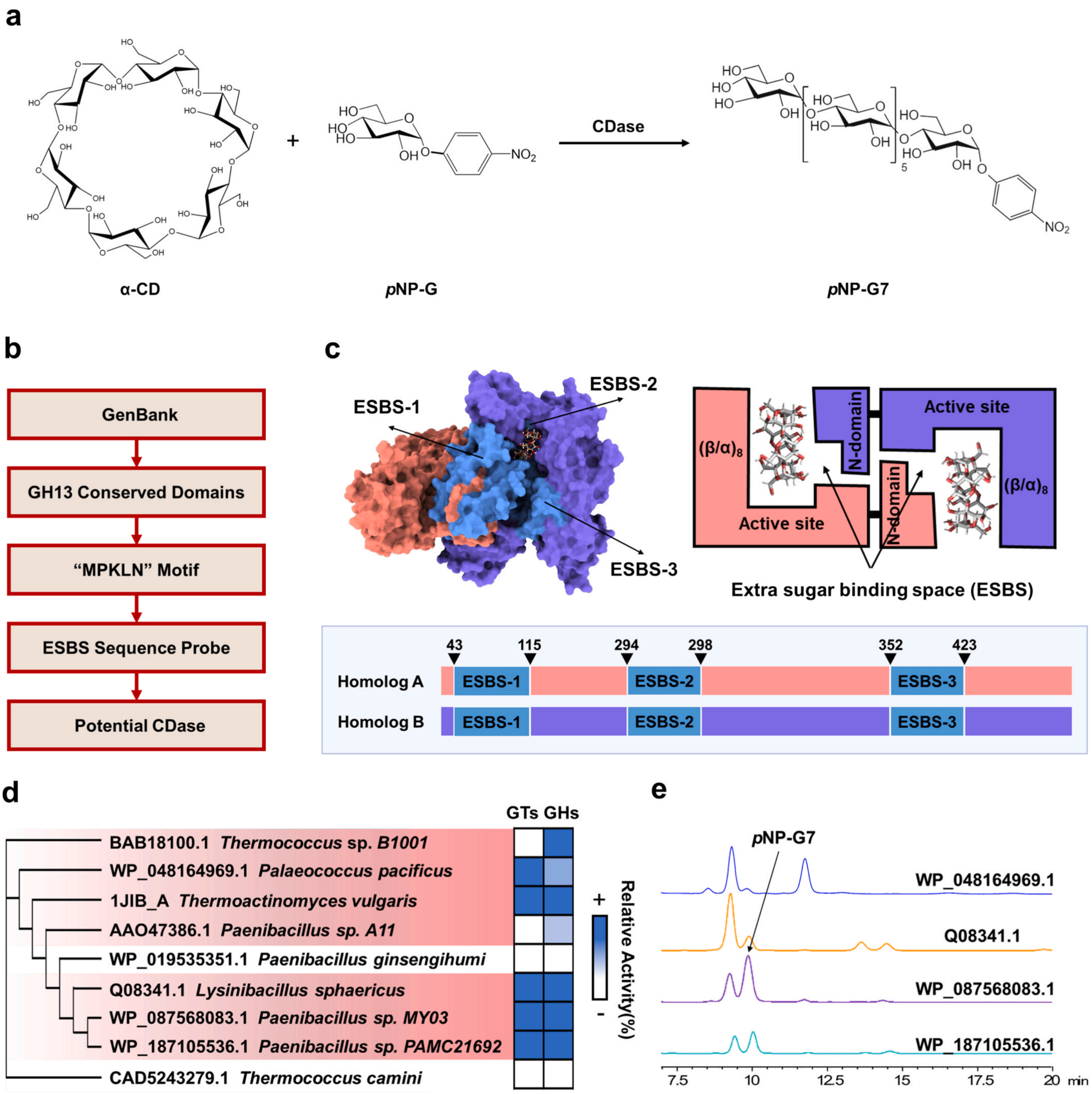

a
α-CD
+
pNP-G
CDase
pNP-G7
5
NO2
b
GenBank
GH13 Conserved Domains
“MPKLN” Motif
ESBS Sequence Probe
Potential CDase
c
ESBS-1
ESBS-2
ESBS-3
Active site
N-domain
(β/α)8
Active site
N-domain
(β/α)8
Extra sugar binding space (ESBS)
43
115
294
298
352
423
Homolog A
Homolog B
ESBS-1
ESBS-2
ESBS-3
d
GTs GHs
BAB18100.1 Thermococcus sp. B1001
WP_048164969.1 Palaeococcus pacificus
1JIB_A Thermoactinomyces vulgaris
AAO47386.1 Paenibacillus sp. A11
WP_019535351.1 Paenibacillus ginsengihumi
Q08341.1 Lysinibacillus sphaericus
WP_087568083.1 Paenibacillus sp. MY03
WP_187105536.1 Paenibacillus sp. PAMC21692
CAD5243279.1 Thermococcus camini
+
-
Relative Activity(%)
e
pNP-G7
WP_048164969.1
Q08341.1
WP_087568083.1
WP_187105536.1
7.5
10
12.5
15
17.5
20 min

Fig. 1. Discovery and functional analysis of novel CDase for 4-nitrophenyl-α-D-maltoheptaoside synthesis using extra sugar binding space motif probes a, Reaction scheme for 4-nitrophenyl-α-D-maltoheptaoside synthesis using CDase. **b,** Mining strategy for identifying novel CDase using ESBS sequence motif probes. **c,** Structural organization of CDase, highlighting its homodimeric form and ESBS sequence motif distribution, with ESBS-1: 43-115aa, ESBS-2: 294-298aa (containing the “MPKLN” motif), and ESBS-3: 352-423aa. **d,** Characterization of transglycosylation and hydrolysis of nine novel CDases. GTs: transglycosylation activity. GHs: hydrolysis activity. **e,** HPLC analysis of product spectra, showing the highest specificity for pNP-G7 in CDase from *Paenibacillus* sp. MY03.

CDase variant was analyzed using GROMACS and VMD's script, which tracks the coordinates and velocities of each atom at different times. The PyMOL software package (Delano, 2002) was used for visualizing the protein structure.

3. Results

3.1. Mining and characterization of novel CDase enzymes and products

EPS-G7 is a critical substrate for enzyme-linked assays used to detect α-amylase in blood and urine, which is vital for diagnosing pancreatic disease (Lorentz, 2000). Traditionally, EPS-G7 synthesis involves capping the reduced end of pNP-G7(Zhong et al., 2023). Given the limitations of the multi-step, low-yield chemical synthesis of pNP-G7, developing a one-step CDase-catalyzed synthesis method is a promising approach. The aim was to use α-cyclodextrin (α-CD) as a donor and CDase for transglycosylation with pNP-G to synthesize pNP-G7, serving as a model for multi-objective optimization of CDase (Fig. 1a). CDase from the GH13 family can transfer cyclodextrins to receptor sugars like pNP-G, forming α-*O*-oligosaccharides. However, CDase also exhibits high hydrolytic activity and complex regioselectivity, leading to glycosylation at different positions like α-1,3, α-1,4, and α-1,6, which can produce side reactions that decrease the yield of target α-*O*-oligomaltosides (Aroob et al., 2021).

To identify CDase candidates with a high transglycosylation/hydrolysis ratio, we mined the non-redundant GenBank database, focusing on the GH13 family, characterized by a $(\beta/\alpha)_8$/TIM barrel catalytic structure (Fig. 1b). We searched for sequence identifiers containing conserved sequence regions (CSRs I-IV) and unique sequences (CSRs VI-VII) of GH13. This yielded 54,624 potential CDase genes, narrowed to 6,842 by identifying the "MPKLN" motif associated with hydrolytic activity. CDases possess an extended ESBS within their active site cleft that accommodates receptor glycosides such as maltose. This ESBS facilitates transglycosylation by stabilizing acceptor molecules near the catalytic center, decreasing water accessibility, and suppressing hydrolytic activity. Screening for high hydrophobicity in ESBS residues identified 104 potential CDases with low hydrolytic activity (Fig. 1c).

Ten candidate genes were expressed in *E. coli* BL21(DE3), with nine producing soluble enzymes (Fig. 1d). Functional analysis with α-CD and pNP-G as substrates showed seven candidates had a high transglycosylation/hydrolysis ratio, two showed no detectable hydrolysis activity, and the remaining two candidates showed no detectable transglycosylation activity. Among these, *Paenibacillus* sp. MY03 CDase (NCBI reference sequence: WP_087568083.1) gave the highest pNP-G7 yield with minimal by-products, achieving a pNP-G7/products ratio of 63 % (Fig. 1e). The main product pNP-G7 and byproduct *P*1 were analyzed using HPLC and LC-MS/MS. Both the pNP-G7 standard and the main product can be rapidly and completely hydrolyzed by α-glucosidase, ultimately yielding pNP. In contrast, the byproduct *P*1 cannot be hydrolyzed by α-glucosidase under the same conditions, and the final accumulated product is pNP-G2. This demonstrates that *P*1 is an isomer of the main product pNP-G7 (see supplementary materials). MY03 CDase had an optimal temperature of 40 °C and pH of 7.5, maintaining stability after 8 h at 40 °C. Cu^{2+}, Zn^{2+}, and Fe^{3+} ions inactivated the enzyme, while other ions had minimal effects (see supplementary materials).

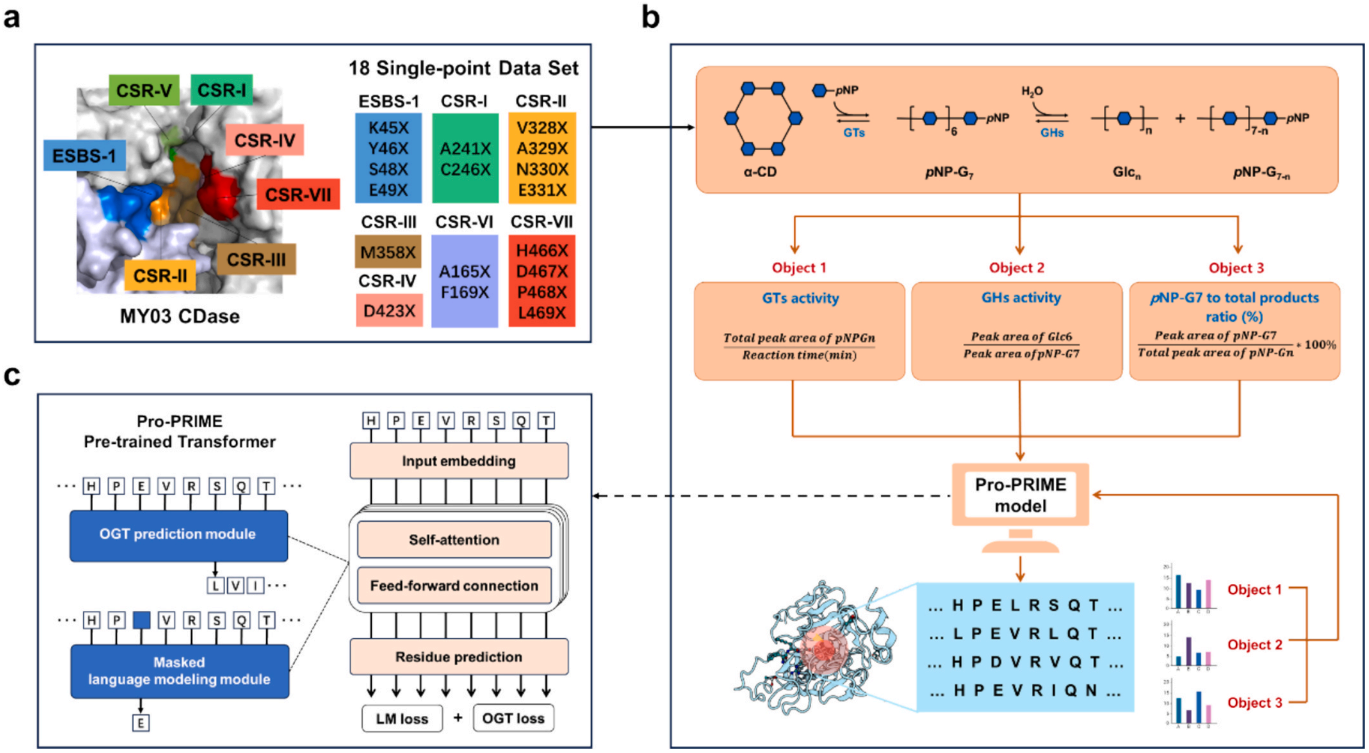


Fig. 2. Multi-objective optimization of MY03 CDase using the protein language pretraining model. a, Identification of hot spot residues in the ESBS and CSR regions of MY03 CDase. **b,** Optimization of transglycosylation activity (object 1), hydrolysis activity (object 2), and the 4-nitrophenyl-α-D-maltoheptaoside product ratio (object 3), with single-point mutations as inputs and two- and three-point mutations as outputs through the Pro-PRIME model. GTs denotes glycosyl transglycosylation activity, measured as the total peak area of pNP-Gn products per minute. GHs denotes hydrolysis activity, calculated as the ratio of maltohexose (Glc6) to pNP-G7 peak areas. The 4-nitrophenyl-α-D-maltoheptaoside ratio reflects CDase regioselectivity, calculated as the specific peak area of 4-nitrophenyl-α-D-maltoheptaoside over total pNP-Gn product areas. **c,** Architecture of Pro-PRIME. The core architecture of Pro-PRIME is a self-attention and feed-forward connection, with embedding utilized for initialization. Following the production of potential representations, Pro-PRIME is comprised of two primary modules: the language modeling (LM) module and the optimal growth temperature (OGT) prediction module.

3.2. Multi-objective optimization of MY03 CDase assisted by a protein language model

Optimizing MY03 CDase for a higher transglycosylation/hydrolysis ratio and increased regioselectivity (pNP-G7/products ratio) presents challenges for traditional directed evolution. To enhance MY03 CDase, we applied the Pro-PRIME protein language model, which integrates prior knowledge to guide efficient enzyme optimization (Jiang et al., 2024) (Fig. 2). A foundational dataset was constructed to support AI-assisted evolution. AlphaFold2 was used to predict the three-dimensional structure of MY03 CDase, with a focus on interactions with α-CD and pNP-G at the catalytic center, CSR, and ESBS (see supplementary materials). Eighteen amino acid residues within 5 Å of both α-CD and pNP-G were identified as hotspots and subjected to single-point mutagenesis (Fig. 2a). This yielded 68 mutants with varying transglycosylation and hydrolysis activities, as well as regioselectivity (see supplementary materials). Functional characterization of key single-point variants is shown in Fig. 3. Next, regression models predicting glycosylation/hydrolysis ratios and discriminative models for pNP-G7/product ratios were constructed by integrating prior knowledge with Pro-PRIME. These models enabled Pro-PRIME to evaluate the entire combinatorial mutant space, predicting transglycosylation and hydrolytic activity along with the pNP-G7/product ratio. Top 10 candidates identified by Pro-PRIME were experimentally validated for multi-site recombination, with the best-performing combinations shown in Fig. 2b, c, and Fig. 3.

Among these, the N330S (M10-1) and L469G (M15-1) mutants exhibited increased transglycosylation/hydrolysis ratios of 16 and 20, respectively, but showed decreases in pNP-G7/product percentages to 23 % and 54 %, respectively. Meanwhile, mutants K45R (M1-1), E49K (M4-1), A241G (M7-1), H466R (M12-1), D467E (M13-1), and P468A (M14-1) had enhanced pNP-G7/products percentages of 77 %, 73 %, 84 %, 83 %, and 92 %, respectively, though their transglycosylation/hydrolysis ratios dropped to 3, 4, 9, 4, and 14. Only the A165K (M5-1) and M358F (M11-1) mutants showed improvements in both the transglycosylation/hydrolysis ratio and pNP-G7/products percentage, achieving values of 28 and 68 % (M5-1), and 27 and 69 % (M11-1), respectively. Additionally, both mutations enhanced transglycosylation activity, though hydrolysis activity also saw a partial increase. This dataset of 68 single-point mutants, encompassing transglycosylation activity, hydrolysis activity, and pNP-G7/products percentage, was used to integrate prior knowledge into the Pro-PRIME model.

It has been demonstrated that language models trained with a masked language modeling objective can estimate sequence variants (Jiang et al., 2024) (Fig. 2c). Leveraging this ability, the language modeling module was used to assess protein sequence mutations. For each mutation, the amino acid in the wild-type protein was the reference state. The Pro-PRIME model, incorporating prior knowledge, was then used to predict the transglycosylation activity, hydrolysis activity, and pNP-G7/product distribution for all double and triple-point mutant variants. Testing double and triple mutants allowed us to validate the model's ability to predict the functionality of higher-order combinatorial mutants based on single-point mutation data. Our screening process prioritized mutants with higher predicted transglycosylation activity, lower hydrolysis activity, and higher pNP-G7/products percentages.

From the Pro-PRIME output, we selected the top 10 multi-point combinatorial mutants (M16-M25) (see supplementary materials), representing the highest-ranking double and triple-point combinations. All ten multi-point mutants showed significantly enhanced regioselectivity, with pNP-G7/products percentages exceeding 88 %. Furthermore, the transglycosylation/hydrolysis ratios of mutants M18-M19, M21-M22, and M24-M25 were markedly improved. Among these, the M25 mutant, a triple-point combination of S48W/M358F/D467E, achieved the highest performance, increasing the pNP-G7/product ratio from 63 % to 98 % and boosting the transglycosylation/hydrolysis ratio by 12-fold, from 13 to 160.

Table 1 provides the functional characterization of the three single-point mutations (S48W, M358F, and D467E) in the optimal mutation combination M25. The M358F mutation increased both transglycosylation and hydrolysis activities, resulting in a modest rise in the

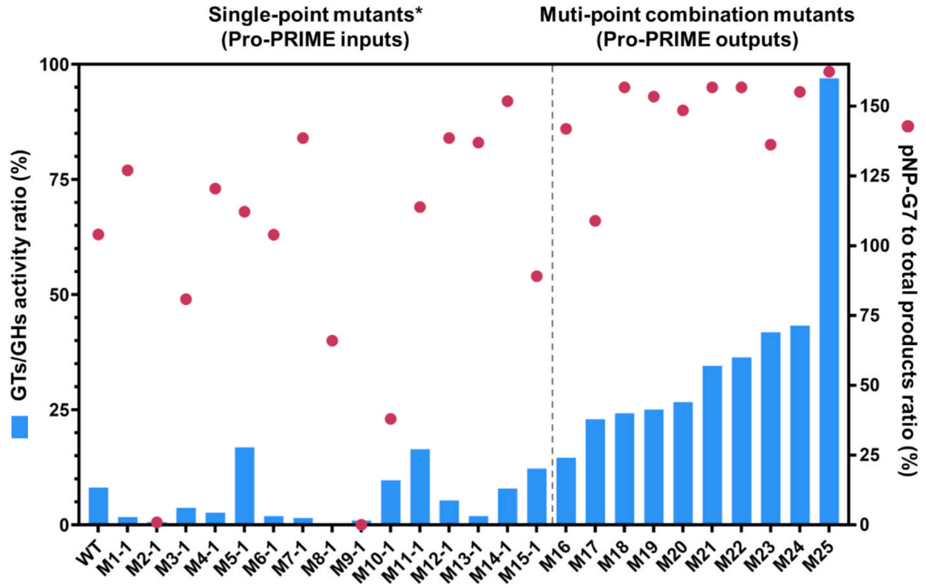


Fig. 3. PLM-guided multi-objective optimization of MY03 CDase variants following a single round of directed evolution. Transglycosylation activity, hydrolytic activity, and the ratio of 4-nitrophenyl-α-D-maltoheptaoside (pNP-G7) to total products were quantified for 68 single-point mutants (only a subset is shown in the figure; full dataset available in the supplementary materials) and used as prior knowledge to train Pro-PRIME. Based on its outputs, the top 10 combinatorial variants (M16–M25) were selected for multi-objective experimental validation. Glycosyltransferase to hydrolase (GTs/GHs) activity ratios are represented by light blue bars, while the pNP-G7-to-total product ratios are shown as deep red dots. *Note: Only a portion of the single-point mutant data is displayed. With just these 68 data points as input, Pro-PRIME successfully identified the high-performance M25 variant, a triple mutant (S48W/M358F/D467E). M25 increased the pNP-G7 product proportion from 63 % to 98 % and enhanced the transglycosylation/hydrolysis ratio twelvefold, from 13 to 160. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Multi-objective optimization and 4-nitrophenyl-α-D-maltoheptaoside yield of WT, single-point, and M25 variants.

Mutants	Transglycoside activity[a]	Hydrolysis activity[b]	Transglycosylation /hydrolysis activity ratio	4-nitrophenyl-α-D-maltoheptaoside /products ×100 %	4-nitrophenyl-α-D-maltoheptaoside yield(g/L)
WT	200	15	13	63	2
S48W (M3-1)	158	26	6	49	3
M358F (M11-1)	732	27	27	69	6
D467E (M13-1)	41	13	3	83	4
S48W/M358F/ D467E (M25)	1600	10	160	98	6

[a] Transglycoside activity, defined as the total peak area of pNP-Gn products divided by the reaction time (min).

[b] Hydrolysis activity, defined as the peak area of maltohexose (Glc6) divided by the peak area of 4-nitrophenyl-α-D-maltoheptaoside.

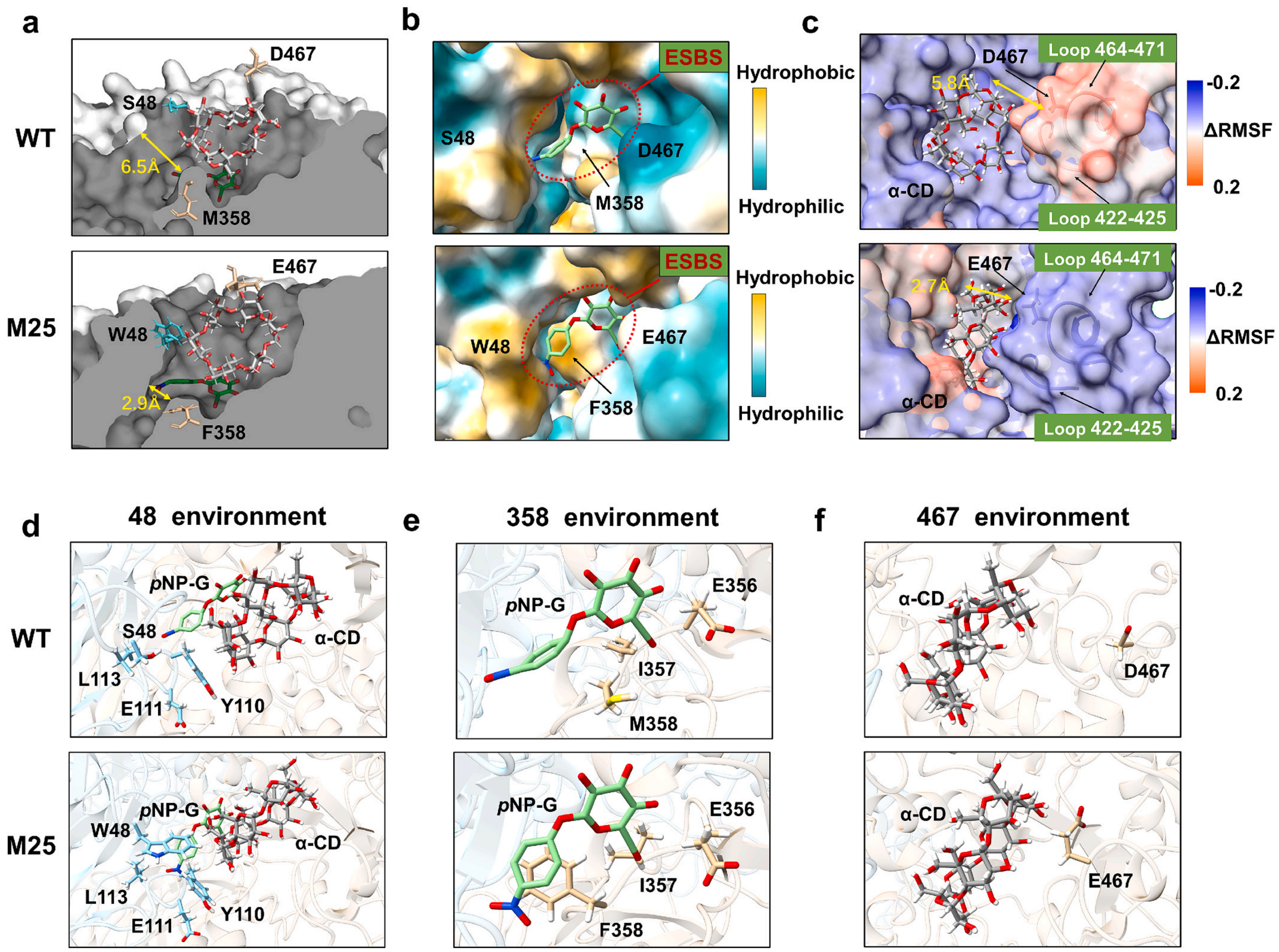


Fig. 4. Mechanistic insights into MY03 CDase transglycosylation and hydrolysis. a, Comparison between cross-sections of the active channels of WT and M25. The S48W mutation reduces the receptor substrate channel diameter from 7 Å to 3 Å, enabling the binding of sugar molecules and reducing water binding. **b,** The M358F mutation enhances the local hydrophobicity of the ESBS region, providing a hydrophobic environment for the receptor site, helping in sugar molecule binding, and reducing water binding. The red dashed circle represents ESBS. The protein surface is represented by a yellow color to indicate hydrophobic regions and a light blue color to indicate hydrophilic regions. **c,** The D467E mutation results in ΔRMSF changes in the surrounding Loop422-425 and Loop464-471. The D467E mutation improved the flexibility of Loop422-425 and Loop464-471, resulting in a tighter binding with α-CD, decreasing pocket size. The protein surface is presented in silver, with the protein cartoon colored blue for low ΔRMSF values and salmon for high ΔRMSF values. **d,** Local environments before and after mutations at site 48. The large hydrophobic indole side chain of the S48W mutation restricts access to the substrate channel, thereby reducing the entry of water molecules into the channel. **e,** Local environments before and after mutations at site 358. The aromatic side chain of the M358F mutation produces a π-π stacking with the pNP-G substrate, increasing the affinity of CDase for the substrate. **f,** Local environments before and after mutations at site 467. D467E exhibits stronger polarity, enabling it to form stronger polar interactions with α-CD. α-CD is depicted in grey, pNP-G is colored in light green, the side chains of homodimer A residues are colored in tan, and the side chains of homodimer B residues are colored in light blue. All structures are represented as stick models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

transglycosylation/hydrolysis activity ratio from 13 to 27 and a slight increase in regioselectivity from 63 % to 69 %. In contrast, the D467E mutation sharply reduced the transglycosylation activity and led to a slight decrease in hydrolysis activity, resulting in a dramatic drop in the transglycosylation/hydrolysis ratio from 13 to 3, alongside an improvement in regioselectivity from 63 % to 83 %. The S48W mutation did not significantly affect transglycosylation activity, hydrolysis activity, or regioselectivity. Compared to these individual mutations, the three-point combination in M25 achieved remarkable improvements in transglycosylation activity, hydrolysis activity, and regioselectivity, reaching a maximum transglycosylation activity of 1,600, the lowest hydrolysis activity of 10, and the highest regioselectivity of 98 %. Additionally, the pNP-G7 yield reached a peak of 6 g/L in a reaction system containing 5 g/L CDase, 400 g/L α-CD, and 150 g/L pNP-G. Due to the complex interactions between the three sites (S48W, M358F, and D467E), navigating the evolutionary path for multi-point optimization is challenging with traditional directed evolution, which focuses on single-point mutations. Leveraging prior knowledge, Pro-PRIME effectively identifies high-dimensional interactions in multi-objective optimization, overcoming the inefficiency of traditional directed evolution for multi-point recombination.

3.3. Molecular mechanism of transglycosylation and hydrolysis in the optimal MY03 CDase mutant

A structural comparison between the wild-type (WT) and M25 mutant is shown (see supplementary materials). Molecular dynamics (MD) simulations were conducted for 20 ns for both WT and M25 at room temperature, as detailed in the Materials and Methods section. Additional MD data collection and analysis are provided (see supplementary materials). Structural analysis of the MY03 CDase-substrate complex revealed that the S48W mutation introduces an aromatic hydrophobic side chain into the substrate channel, reducing its diameter from 7 Å to 3 Å (Fig. 4a). This modification restricts solvent accessibility, limiting nucleophilic water attack on the enzyme–substrate intermediate. Experimentally, S48W reduced hydrolytic activity to 26 while retaining substantial transglycosylation activity, resulting in a pNP-G7 yield of 3 g/L (Table 1). As shown in Fig. 4d, the hydrophobic side chain of W48 acts as a large “lid” over the ESBS region, further restricting the entry of water molecules. In Fig. 4c, the S48W/M358F/D467E mutation combination enhances hydrophobicity near pNP-G, strengthening nonpolar interactions between pNP-G and the ESBS region. Equilibrium conformational statistics indicate that the S48W/M358F/D467E mutation reduces the number of solvent water molecules within 3 Å of pNP-G by 65 %. Specifically, as shown in Fig. 4e, the aromatic side chain of M358F forms a π-π stacking interaction with the benzene ring of pNP-G, enhancing the binding affinity between pNP-G and the ESBS region and promoting transglycosylation. Experimental findings confirm that M358F increased relative transglycosylation activity by 3.7-fold and improved pNP-G7 yield to 6 g/L (Table 1), demonstrating its role in substrate recognition.

The root mean square fluctuation (ΔRMSF) analysis highlights changes in structural flexibility between the WT and M25 mutant. The M25 mutant exhibited a lower RMSD than the WT, indicating reduced atomic or residue displacement and increased stability. The ΔRMSF between M25 and WT is notably reduced, especially in regions mediated by D467E (Loop 422-425 and 464-471), suggesting that these areas become more stable in M25, resulting in stronger binding with substrate α-CD. This reduces the channel diameter for water molecules to enter the ESBS region from 6 Å to 3 Å, thus decreasing hydrolytic activity. Experimental data further demonstrate a 133-fold reduction in hydrolytic activity for D467E relative to WT. The D467E mutant data corroborate this finding, showing a 13-fold decrease in hydrolytic activity while improving regioselectivity to 83 % (Table 1). Additionally, when substrate α-CD binds, Glu467, acting as an acid, provides a stronger proton donation to the glycosidic bond oxygen. The S48W/M358F/D467E mutations collectively enhance hydrophobicity in these residues, decreasing the likelihood of water attacking the donor substrate intermediate, which in turn lowers hydrolytic activity. The synergistic effect of these mutations is evident in the M25 triple mutant, which achieves a remarkable 160 transglycosylation/hydrolysis ratio (12-fold improvement over WT) and 98 % pNP-G7 product ratio, with a final yield of 6 g/L (Table 1). This validates our database mining approach to increase surface hydrophobicity near the ESBS region, thereby reducing hydrolytic activity.

3.4. Optimization of production conditions for the enzymatic synthesis of 4-nitrophenyl-α-D-maltoheptaoside

We assessed the pNP-G7 synthesis capability of WT and M25at a concentration of 5 g/L. As shown in Fig. 5a, M25 had a significantly higher pNP-G7 synthesis ability than WT, reaching 0.4 g/L in just 30 min—surpassing WT’s maximum synthesis capacity—and sustaining high production efficiency up to 360 min. The WT yield plateaued after 180 min, likely due to competitive hydrolytic activity hindering conversion rates. Fig. 5b tracks the reaction process, revealing that the main product, pNP-G7, maintains high regioselectivity with minimal by-products. The most abundant by-product, pNP-G6, constituted only about 10 % at 150 min. Subsequent optimization of the production conditions for M25′s pNP-G7 synthesis, detailed in Supplementary (see supplementary materials), increased M25′s transglycosylation activity eightfold, boosting the pNP-G7 yield from 73 g/L (WT) to 161 g/L (M25) (Fig. 5c). The optimized reaction system contained CDase (5 g/L), α-CD (400 g/L), and pNP-G (150 g/L). It was incubated at 40 °C under precisely controlled pH conditions. Product analysis using LC-MS/MS, ^{1}H NMR, ^{13}C NMR, and 2D NMR confirmed that the product pNP-G7 exhibited the required purity and correct structure (see supplementary materials).

As shown in Fig. 5d, M25′s one-step synthesis of pNP-G7 demonstrates clear advantages, raising the pNP-G conversion rate from 55 % to 85 % and the space–time yield from 31 g/L × h to 64 g/L × h. In comparison, the chemical synthesis of pNP-G7 achieves only a 5 % yield, with significant metal-catalyzed pollution (Zhong et al., 2023) from multi-step reactions, typically using the non-reducing end of pNP-G7 maltoside. In contrast, our study achieved a 91 % yield of EPS-G7. The CDase-mediated transglycosylation method for EPS-G7 synthesis thus offers significant cost advantages (see supplementary materials). Product analysis using LC-MS/MS, ^{1}H NMR, ^{13}C NMR, and 2D NMR confirmed that the final product EPS-G7 had the required purity and correct structure (see supplementary materials).

3.5. Glycosylated receptor substrate profile of the optimal MY03 CDase mutant

We further examined the transglycosylation activity of MY03 CDase on various glycoside compounds. M25 had a broad substrate range across valuable glycoside compounds, with conversion rates for different acceptor substrates as follows: α-arbutin at 22 %, β-arbutin at 25 %, indolyl β-D-glucoside at 32 %, rhodioloside at 19 %, and polydatin at 36 % (Fig. 6). Detailed product analysis data are available in the supplementary materials. These conversion rates were measured under suboptimal conditions; further optimization of reaction conditions through engineering could significantly improve MY03 CDase’s receptor substrate conversion rates in the future.

4. Discussion

This study addresses challenges in the α-*O*-oligosylation reaction (Boltje et al., 2009). We successfully developed a novel CDase-mediated transglycosylation approach to synthesize oligosaccharides with specific degrees of polymerization using cyclodextrin as a substrate. The targeted evolution method based on the ESBS sequence motif probe and

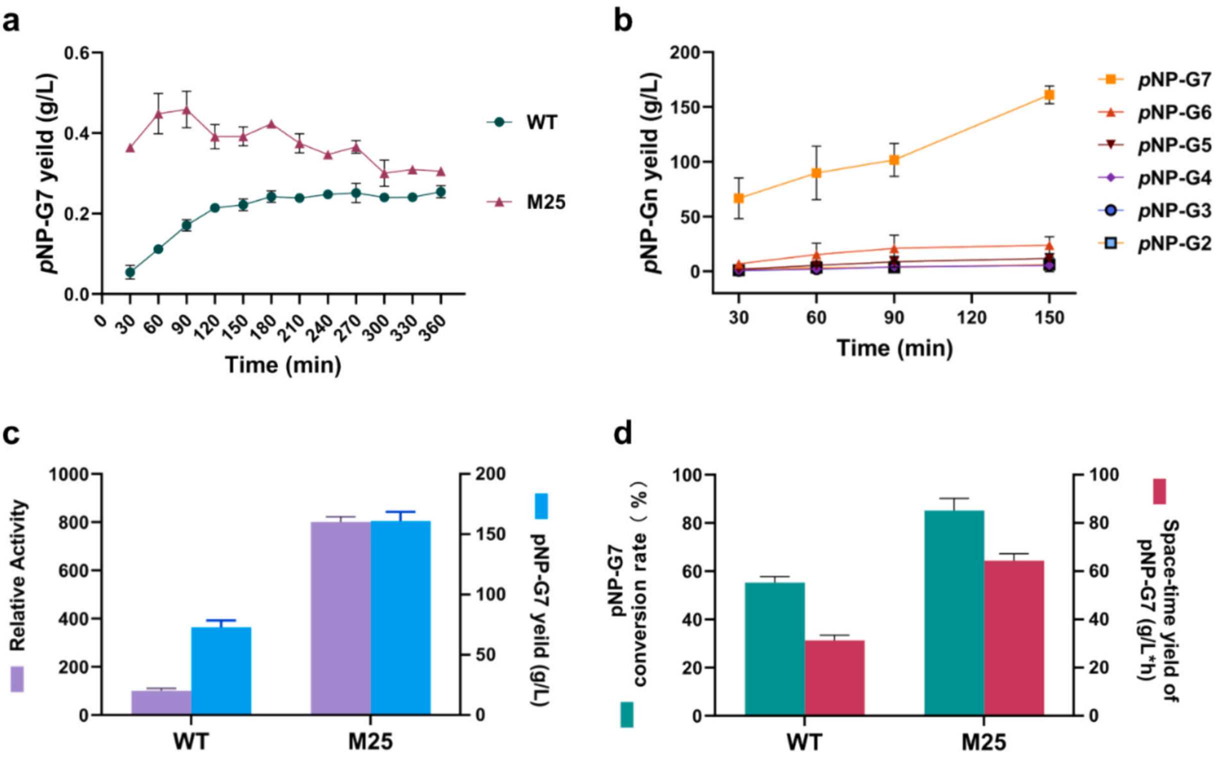


Fig. 5. Enzymatic production of 4-nitrophenyl-α-D-maltoheptaoside by WT and M25 CDase variants. a, Yield of 4-nitrophenyl-α-D-maltoheptaoside(g/L) at various times during the reaction process of WT and M25 variants. **b,** Product generation throughout the reaction process of the M25. 4-nitrophenyl-α-D-maltoheptaoside is the specific product, whereas the others are byproducts. **c,** Relative transglycosylation activity and total 4-nitrophenyl-α-D-maltoheptaoside yield (g/L) under production conditions for WT and M25. **d,** Conversion rates of pNP-G and 4-nitrophenyl-α-D-maltoheptaoside space–time yield (g/L × d) for WT and M25. CDase (5 g/L), α-CD (400 g/L), and pNP-G (150 g/L) were incubated at 40 °C.

protein language models enabled the efficient synthesis of pNP-G7 by the MY03 CDase variant M25. Compared to the wild type, M25 had significantly enhanced transglycosylation activity, hydrolysis activity, and regioselectivity. Current chemical synthesis routes for pNP-G7 involve a four-step process with a total yield of only 5 %, including acetylation to protect hydroxyl groups (99 % yield), acid hydrolysis of O-glycosidic bonds (44 % yield), condensation isomerization (33 % yield), and deprotection (36 % yield). These methods are complex and can produce unwanted β-isomers (Zhong et al., 2023). Additionally, glycosyltransferase-mediated oligosaccharide synthesis, as seen with CGTases, suffers from low efficiency due to enzyme-specific side reactions. Strompen et al. synthesized O-glycosidic bonds using CGTase, yielding diverse products (pNP-G2, pNP-G3, pNP-G4, and pNP-G5) but failing to produce specific pNP-G7 oligosaccharides (Hancock et al., 2006; Lim et al., 2021; Zhong et al., 2023). Our innovative α-*O*-oligosaccharide synthesis strategy overcomes the limitations of both traditional chemical methods and glycosyltransferase-based approaches, achieving a 12-fold improvement in the transglycosylation/hydrolysis ratio compared to the wild type, while increasing regioselectivity for pNP-G7 from 63 % to 98 %. The breakthrough industrial application potential is demonstrated by the M25 mutant achieving 160 g/L pNP-G7 yield at 5 g/L enzyme loading, with a space–time yield of 64.38 g/L·h (8-fold higher than wild-type) at the 100L scale, with a reaction time of less than 2 h. While maintaining excellent regioselectivity, this single-step enzymatic process improves yield from 5.2 % to 85 % compared to conventional multi-step chemical synthesis. MY03 CDase also had broad substrate promiscuity, showing high modification efficiency for substrates like arbutin, rhodioloside, polydatin, and indole-β-D-glucoside, making it a potential platform enzyme for precise oligosaccharide modification.

Glycoside hydrolases (GHs) suitable for transglycosylation applications require high transglycosylation-to-hydrolysis ratios, allowing enzymes to compete effectively with water during deglycosylation of the sugar intermediate (Mangas-Sánchez & Adlercreutz, 2015; Napiórkowska et al., 2017). The thermodynamically unfavorable direction of the reaction and the enzymatic degradation of reaction products limit the yield of GH-mediated methods (Boltje et al., 2009). To further improve the transglycosylation/hydrolysis ratio, low-hydrolysis variants were prioritized. Consistent with Kim et al., our findings support the role of ESBS in facilitating glycosyl transfer by enabling sugar molecules to outcompete water as nucleophiles attacking the intermediate (Kim et al., 1999). In our study, using high ESBS surface hydrophobicity as a screening criterion, we experimentally validated nine randomly selected CDase candidates, discovering that seven had low hydrolytic activity. At the same time, the remaining two showed no detectable hydrolytic activity. This confirmed the feasibility of the ESBS sequence motif probe strategy for discovering novel CDases with high transglycosylation/hydrolysis ratios.

Recent advances in protein language models (PLMs), such as ESM-2 (Lin et al., 2023), have revolutionized enzyme engineering by enabling zero-shot prediction of functional mutations. However, these models primarily focus on single-objective optimization (e.g., thermostability or activity). In contrast, our Pro-PRIME framework extends PLMs to multi-objective optimization, simultaneously balancing transglycosylation, hydrolysis, and regioselectivity—a critical challenge for glycosidases. The optimized MY03 CDase variant, M25, functions via two distinct but synergistic mechanisms. First, the M358F mutation enhances surface hydrophobicity in the ESBS. Second, the S48W mutation reduces solvent access by narrowing the substrate channel, thereby suppressing hydrolysis. This dual-mode strategy overcomes the trade-offs typical in multifunctional enzyme engineering. This dual modification synergistically reduced water infiltration into the active site, effectively suppressing hydrolysis. These observations align with prior studies demonstrating that ESBS hydrophobicity governs CDase reaction

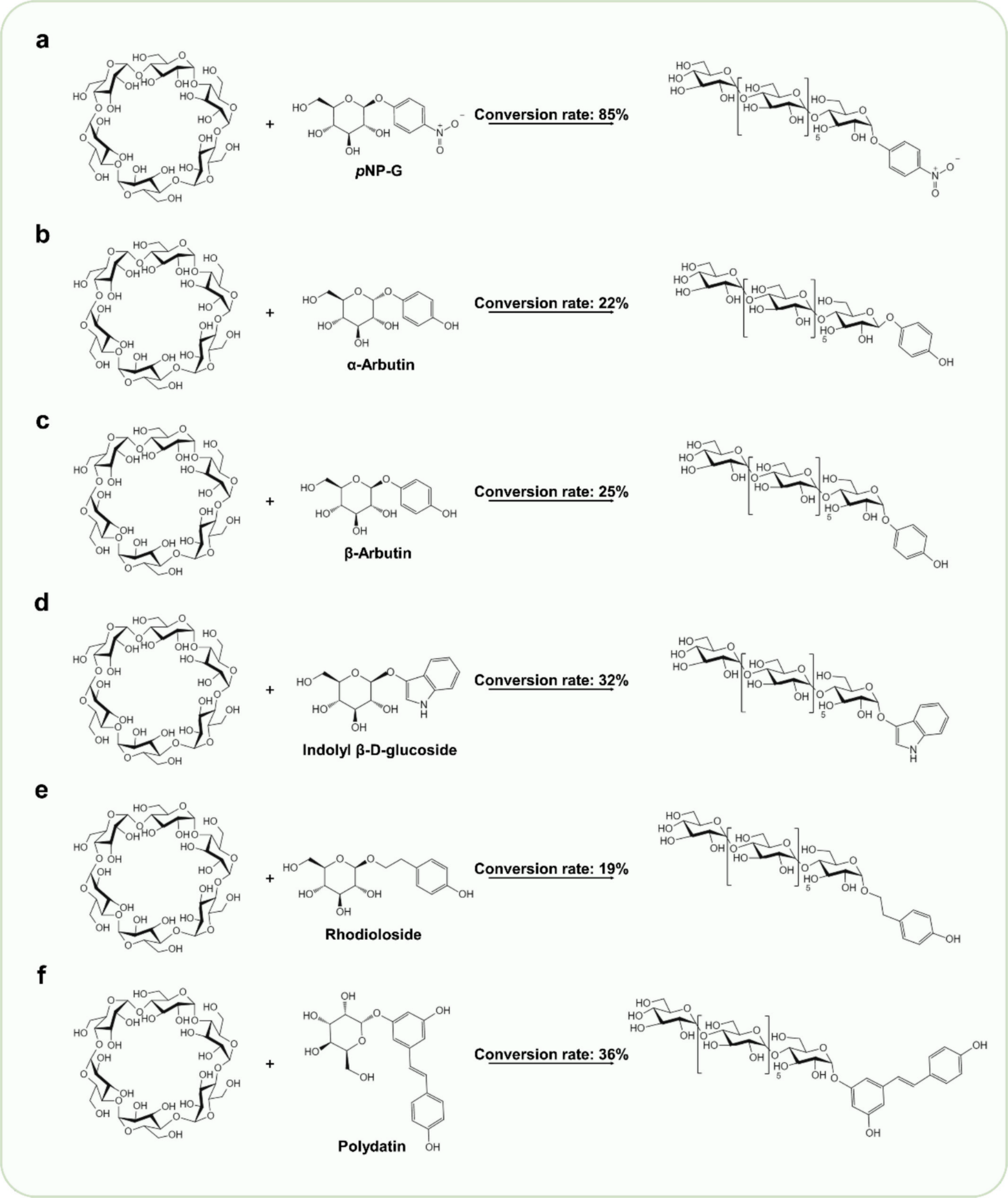


Fig. 6. Receptor substrates of M25 and conversion rates. Reactions were done with 45 μL of 10 mg/mL CDase, 50 μL of 15 % α-CD, and 5 μL of 2 M receptor substrate, incubated at 40 °C for 180 min. The reaction was quenched with 190 μL of 75 % methanol, and the products were analyzed via LC at 270 nm.

specificity (Kim et al., 1999). 2. Strengthened Substrate Binding: The M358F mutation reinforced nonpolar interactions between the enzyme and the acceptor substrate (pNP-G), enhancing substrate affinity and transglycosylation efficiency. Molecular dynamics analysis revealed markedly reduced root mean square fluctuation (ΔRMSF) in M25 compared to wild-type, particularly in regions flanking the D467E mutation (Loops 422–425 and 464–471). This increased structural stability facilitated tighter binding with the donor substrate α-cyclodextrin (α-CD), further optimizing catalytic performance. These insights provide valuable guidance for future engineering of CDase and other GH13 enzymes, highlighting the importance of factors such as ESBS hydrophobicity and substrate binding to achieve desired catalytic properties, and offering a blueprint for balancing competing enzymatic activities.

Our study introduces a novel AI-assisted approach for multi-target enzyme evolution that enhances transglycosylation activity while reducing hydrolysis activity and improving regioselectivity by identifying CDase candidates with high transglycosylation/hydrolysis ratios, thus filling gaps in previous studies (Vu et al., 2023; Yu et al., 2023). Multi-target enzyme optimization is a complex and underexplored area in protein engineering (Gumulya et al., 2012; Romero & Arnold, 2009; Wiltschi et al., 2020). We combined AI-assisted predictions with experimental validation to simultaneously enhance transglycosylation activity for desired oligosaccharides, reduce hydrolysis activity, and improve specific product ratios. Our approach effectively navigates multi-target optimization challenges by integrating protein language models with experimental data, leveraging limited beneficial single-point mutation data as prior knowledge. This approach allows for the simultaneous optimization of multiple enzyme parameters—transglycosylation activity, hydrolytic suppression, and regioselectivity—surpassing the constraints of traditional single-target methods. Compared to prior AI-assisted directed evolution studies, our work highlights the power of protein language models in complex, multi-objective enzyme engineering (Lin et al., 2023).

5. Conclusions

This study presents the first CDase-mediated method for precise oligosaccharide conjugate synthesis using pNP-G as substrate, validating the use of ESBS motif-based probes for identifying CDases with favorable transglycosylation/hydrolysis activity ratios. The engineered variant M25 achieved optimal performance across multiple traits in a single Pro-PRIME iteration, enabling efficient pNP-G7 synthesis via transglycosylation. This establishes an AI-guided framework for CDase-based α-O-oligosaccharide synthesis with glucoside acceptors. However, further evolution will be required to extend this strategy to the synthesis of other oligoglycosides. E-supplementary data are available in the online version of this paper.

Ethics statement

No animals or humans were involved in this study.

CRediT authorship contribution statement

Ting Nie: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zhenxin Yan:** Writing – review & editing, Investigation, Formal analysis, Data curation, Conceptualization. **Hao Liu:** Validation, Software. **Xiudian Zhang:** Validation, Data curation. **Weiwei Zhong:** Formal analysis, Data curation. **Qin Chen:** Validation. **Changbin Zhu:** Investigation. **Liang Hong:** Supervision. **Guang-yu Yang:** Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by Key-Area Research and Development Program of Guangdong Province, China (2022B1111050001), National Key R&D Program of China (2020YFA0907902), National Natural Science Foundation of China (32030063 and 32371483), Shanghai Pilot Program for Basic Research and Shanghai Municipal Science and Technology Major Project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biortech.2025.133206.

Data availability

Benchmark datasets used for the Pro-PRIME model are available within the article (https://doi.org/10.1126/sciadv.adr2641) and the website (https://github.com/ai4protein/Pro-Prime); The data supporting the findings of MY03 CDase are available within the article and/or its supplementary materials.

References

Aroob, I., Ahmad, N., Rashid, N., 2021. Cyclodextrin-preferring glycoside hydrolases: properties and applications. Amylase 5 (1), 23–37.

Boltje, T.J., Buskas, T., Boons, G.-J., 2009. Opportunities and challenges in synthetic oligosaccharide and glycoconjugate research. Nat. Chem. 1 (8), 611–622.

Chen, K., Arnold, F.H., 2020. Engineering new catalytic activities in enzymes. Nat. Catal. 3 (3), 203–213.

Delano, W.L. 2002. PyMOL: An Open-Source Molecular Graphics Tool.

Gumulya, Y., Sanchis, J., Reetz, M.T., 2012. Many pathways in laboratory evolution can lead to improved enzymes: how to escape from local minima. Chembiochem 13.

Hancock, S.M., Vaughan, M.D., Withers, S.G., 2006. Engineering of glycosidases and glycosyltransferases. Curr. Opin. Chem. Biol. 10 (5), 509–519.

Hatanaka, K., Ito, Y., Ishio, K., Uryu, T., 1994. Glycosyl transfer reaction to the oligosaccharide chain on the synthetic polymer. Polym. J. 26 (11), 1295–1297.

Henrissat, B., Davies, G., 1997. Structural and sequence-based classification of glycoside hydrolases. Curr. Opin. Struct. Biol. 7 (5), 637–644.

Huang, K., Flitsch, S.L., 2019. Glyco-enzymatic cascades get protection. Nat. Catal. 2 (6), 479–480.

Hwa Park, K., Jeong Kim, M., Seob Lee, H., Soo Han, N., Kim, D., Robyt, J.F., 1998. Transglycosylation reactions of Bacillus stearothermophilus maltogenic amylase with acarbose and various acceptors. Carbohydr. Res. 313 (3), 235–246.

Iwamoto, S., Kasahara, Y., Yoshimura, Y., Seko, A., Takeda, Y., Ito, Y., Totani, K., Matsuo, I., 2017. Endo-α-Mannosidase-Catalyzed Transglycosylation. Chembiochem 18.

Janecek, S.t. 1997. α-amylase family: Molecular biology and evolution. *Progress in Biophysics and Molecular Biology*, **67**(1), 67-97.

Jiang, F., Li, M., Dong, J., Yu, Y., Sun, X., Wu, B., Huang, J., Kang, L., Pei, Y., Zhang, L., Wang, S., Xu, W., Xin, J., Ouyang, W., Fan, G., Zheng, L., Tan, Y., Hu, Z., Xiong, Y., Feng, Y., Yang, G., Liu, Q., Song, J., Liu, J., Hong, L., Tan, P., 2024. A general temperature-guided language model to design proteins of enhanced stability and activity. Sci. Adv. 10 (48), eadr2641.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873), 583–589.

Kang, G.-J., Kim, M.-J., Kim, J.-W., Park, K.H., 1997. Immobilization of thermostable maltogenic amylase from bacillus stearothermophilus for continuous production of branched oligosaccharides. J. Agric. Food Chem. 45 (10), 4168–4172.

Kim, J.-S., Cha, S.-S., Kim, H.-J., Kim, T.-J., Ha, N.-C., Oh, S.-T., Cho, H.-S., Cho, M.-J., Kim, M.-J., Lee, H.-S., Kim, J.-W., Choi, K.Y., Park, K.-H., Oh, B.-H., 1999. Crystal structure of a maltogenic amylase provides insights into a catalytic versatility *. J. Biol. Chem. 274 (37), 26279–26286.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W., Davis, I.W., Cooper, S., Treuille, A., Mandell, D.J., Richter, F., Ban, Y.E., Fleishman, S.J., Corn, J.E., Kim, D.E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J.J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J.J., Kuhlman, B., Baker, D., Bradley, P., 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 487, 545–574.

Lim, C.H., Rasti, B., Sulistyo, J., Hamid, M.A., 2021. Comprehensive study on transglycosylation of CGTase from various sources. Heliyon 7 (2).

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A., 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379 (6637), 1123–1130.

Lorentz, K., 2000. Routine α-Amylase Assay using Protected 4-Nitrophenyl-1,4-α-d-maltoheptaoside and a Novel α-Glucosidase. Clin. Chem. 46 (5), 644–649.

Mangas-Sánchez, J., Adlercreutz, P., 2015. Enzymatic preparation of oligosaccharides by transglycosylation: a comparative study of glucosidases. J. Mol. Catal. B Enzym. 122, 51–55.

Napiórkowska, M., Boilevin, J., Sovdat, T., Darbre, T., Reymond, J.-L., Aebi, M., Locher, K.P., 2017. Molecular basis of lipid-linked oligosaccharide recognition and processing by bacterial oligosaccharyltransferase. Nat. Struct. Mol. Biol. 24 (12), 1100–1106.

Páll, S., Zhmurov, A., Bauer, P., Abraham, M., Lundborg, M., Gray, A., Hess, B., Lindahl, E., 2020. Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. J. Chem. Phys. 153 (13), 134110.

Park, K.-H., Kim, T.-J., Cheong, T.-K., Kim, J.-W., Oh, B.-H., Svensson, B., 2000. Structure, specificity and function of cyclomaltodextrinase, a multispecific enzyme of the α-amylase family. Biochim. Biophys. Acta Protein Struct. Mol. Enzymol. 1478 (2), 165–185.

Romero, P.A., Arnold, F.H., 2009. Exploring protein fitness landscapes by directed evolution. Nat. Rev. Mol. Cell Biol. 10 (12), 866–876.

Shimoda, K., Akagi, M., Hamada, H., 2009. Production of beta-maltooligosaccharides of alpha- and delta-tocopherols by Klebsiella pneumoniae and cyclodextrin glucanotransferase as anti-allergic agents. Molecules 14 (8), 3106–3114.

Shimoda, K., Hamada, H., Hamada, H., 2008. Chemo-enzymatic synthesis of ester-linked taxol–oligosaccharide conjugates as potential prodrugs. Tetrahedron Lett. 49 (4), 601–604.

Shimoda, K., Kubota, N., Akagi, M., 2012. Synthesis of capsaicin oligosaccharides and their anti-allergic activity ——synthesis of capsaicin oligosaccharides as anti-allergic food-additives. Adv. Chem. Eng. Sci. 2012, 45–49.

Tao, B.Y., Reilly, P.J., Robyt, J.F., 1989. Detection of a covalent intermediate in the mechanism of action of porcine pancreatic α-amylase by using 13C nuclear magnetic resonance. Biochim. Biophys. Acta Protein Struct. Mol. Enzymol. 995 (3), 214–220.

Vu, M.H., Akbar, R., Robert, P.A., Swiatczak, B., Sandve, G.K., Greiff, V., Haug, D.T.T., 2023. Linguistically inspired roadmap for building biologically reliable protein language models. Nat. Mach. Intell. 5 (5), 485–496.

Wang, L.-X., Huang, W., 2009. Enzymatic transglycosylation for glycoconjugate synthesis. Curr. Opin. Chem. Biol. 13 (5), 592–600.

Wiltschi, B., Cernava, T., Dennig, A., Galindo Casas, M., Geier, M., Gruber, S., Haberbauer, M., Heidinger, P., Herrero Acero, E., Kratzer, R., Luley-Goedl, C., Müller, C.A., Pitzer, J., Ribitsch, D., Sauer, M., Schmölzer, K., Schnitzhofer, W., Sensen, C.W., Soh, J., Steiner, K., Winkler, C.K., Winkler, M., Wriessnegger, T., 2020. Enzymes revolutionize the bioproduction of value-added compounds: from enzyme discovery to special applications. Biotechnol. Adv. 40, 107520.

Yu, T., Boob, A.G., Volk, M.J., Liu, X., Cui, H., Zhao, H., 2023. Machine learning-enabled retrobiosynthesis of molecules. Nat. Catal. 6 (2), 137–151.

Zhong, X., Ying, D., Huang, Y., Zhang, L., Wang, H.-Y., Zhu, Q., Chen, E., Yu, Y., Chen, W., 2023. BF 3 · CH 3 CN promotes a formation of Aryl α-O -glycosides from pyranoses: a facile approach to oligosaccharide EPS-G7. ChemistrySelect.