Chemical Engineering Journal 519 (2025) 165191

Contents lists available at ScienceDirect

Chemical Engineering Journal

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

Machine learning-guided engineering of T7 RNA polymerase and mRNA capping enzymes for enhanced gene expression in eukaryotic systems

Seung-Gyun Woo [a], Danny J. Diaz [b], Wantae Kim [c], Mason Galliver [a], Andrew D. Ellington [a,*]

[a] Department of Molecular Bioscience, University of Texas at Austin, Austin, TX 78712, USA
[b] Department of Chemistry and Institute for Foundations of Machine Learning, University of Texas at Austin, Austin, TX 78712, USA
[c] McKetta Department of Chemical Engineering, University of Texas at Austin, Austin, TX 78712, USA

ARTICLE INFO

Keywords:
T7 RNA polymerase (T7 RNAP)
mRNA capping enzyme (CE)
Machine learning (ML)
Protein engineering
Yeast expression system
Synthetic biology

ABSTRACT

The integration of synthetic biology tools into eukaryotic systems offers both significant opportunities and challenges, particularly in optimizing transcriptional and post-transcriptional processes. T7 RNA polymerase (T7 RNAP) and mRNA capping enzymes (CEs) have been fused to enable eukaryotic mRNA production within a single construct. However, the activity of the fusion construct between the African Swine Fever Virus capping enzyme (ASFVCE) and T7 RNAP was relatively low. To address this, we fused the Brazilian Marseillevirus capping enzyme (BMCE) to T7 RNAP and developed a machine learning (ML) pipeline to engineer greatly improved fusion variants. This approach enabled the additive integration of nine predicted single substitutions that improved gene expression in yeast, thereby generating fusion polymerases that exhibited over 10-fold improvements in gene expression efficiency relative to the original fusion enzyme. Not only were ML substitutions additive for gene expression, they could be further combined with variants identified via directed evolution for even higher activities. By allowing ML predictions to guide validations we could rapidly explore the sequence landscape for enzyme optimization, achieving superior results even when compared to directed evolution. The improved enzymes have potential impact for numerous synthetic biology applications, including metabolic engineering, mRNA therapeutics, and cell free systems.

1. Introduction

Over the past decade, the fields of synthetic biology and protein engineering have made remarkable strides, driving progress across therapeutic, molecular, and industrial biotechnology domains [1,2]. At the core of many of these advancements are RNA polymerases, which play a fundamental role in transcription—the first and indispensable step of gene expression [3,4]. Among these enzymes, bacteriophage T7 RNA polymerase (T7 RNAP) stands out as a versatile tool due to its exceptional efficiency and specificity in RNA synthesis. This has established T7 RNAP as a cornerstone in synthetic biology, particularly in applications involving in vitro transcription and mRNA production [5,6].

While T7 RNAP has proven highly effective in prokaryotic systems, its adaptation for eukaryotic expression presents substantial challenges. Unlike prokaryotic systems, eukaryotic cells require complex post-transcriptional modifications, such as the addition of 5′-m7G caps and poly(A) tails, to stabilize mRNA and promote efficient translation [7,8]. These additional requirements, coupled with processes like splicing and nuclear export, necessitate innovative approaches to expand T7 RNAP's functionality in eukaryotic contexts [9,10]. A promising solution has been the fusion of T7 RNAP with viral capping enzymes, such as the African Swine Fever Virus capping enzyme (ASFVCE), to facilitate co-transcriptional capping [11,12]. This approach has also been extended to the Faustovirus capping enzyme (FCE), where its fusion with T7 RNAP demonstrated up to 90% Cap-1 incorporation, significantly simplifying mRNA synthesis and enabling efficient production for therapeutic applications, including mRNA vaccines and protein therapeutics [13].

Although these fusion constructs generate abundant mRNA yields, they often fail to achieve the anticipated levels of protein expression, which emphasizes the need for further optimization. One way to overcome these limitations is engineering poly(A) polymerases to extend mRNA tails, resulting in moderate increases in protein yield [14]. Additionally, directed evolution of single-chain T7 RNAP variants fused with ASFV capping enzymes has enhanced activity by up to four-fold in

* Corresponding author.
E-mail address: ellingtonlab@gmail.com (A.D. Ellington).

https://doi.org/10.1016/j.cej.2025.165191
Received 19 March 2025; Received in revised form 18 June 2025; Accepted 20 June 2025
Available online 23 June 2025

mammalian cells [15].

The advent of machine learning (ML)-based protein engineering [16,17] provides opportunities for the rapid exploration of very large sequence spaces [18,19], even relative to methods like directed evolution, which is sometimes constrained by the limitations of local optima in complex fitness landscapes [20]. For example, EVOLVEpro [21] utilized AI-driven in silico evolution to engineer T7 RNAP variants optimized for in vitro applications, achieving enhanced RNA yield, improved mRNA translation efficiency, and reduced immunogenicity. In parallel, Rosetta-based structural modeling assessed the impact of amino acid substitutions on T7 RNAP stability and function [5]. This approach identified the G47A + 884G variant, which reduced immunostimulatory double-stranded RNA (dsRNA) formation by altering RNAP-RNA interactions. These changes lowered cytokine responses in mammalian systems and streamlined mRNA purification.

We have previously used directed evolution to identify improvements to a fusion between T7 RNAP and a cappase enzyme that improved gene expression in yeast [15], ultimately identifying an ASFVCE(443):T7 RNAP(443) variant that shows a 5-fold improvement in activity relative to the original wild-type fusion combination. Now, using a ML framework that integrated structure-based (MutCompute [22,23]), stability-based (Stability Oracle [24]), and evolution-based (MutRank [25]) predictions, we have generated a unique set of potential sequence substitutions relative to other algorithms [5,15,21]. Experimentally validating improvements to gene expression in yeast led to additive combinations that improved overall activity, allowing efficient and rapid enzyme optimization. In consequence, we identified a variant, EvoBMCE:EvoT7, that contains nine targeted substitutions and exhibits more than a 10-fold improvement in activity in yeast. These mutations, combined with those previously identified through directed evolution, yielded a polymerase fusion (EvoBMCE:EvoT7(443)) that demonstrated an almost 13-fold improvement in activity.

2. Materials and methods

2.1. Strains and culture media

For plasmid construction and maintenance, *Escherichia coli* NEB 5-alpha was cultured in lysogeny broth (LB) medium, which contained 10 g/L tryptone, 5 g/L yeast extract, and 10 g/L NaCl, at 37 °C with agitation at 225 rpm. When necessary, cultures were supplemented with 100 μg/mL ampicillin or 50 μg/mL kanamycin for plasmid selection. For yeast-based experiments, *Saccharomyces cerevisiae* BY4741 (MATα; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0) and its ΔGal2 derivative [15] were utilized. Yeast strains were cultivated in yeast extract peptone dextrose (YPD) medium, consisting of 10 g/L yeast extract, 20 g/L peptone, and 20 g/L glucose. For selective growth, cultures were maintained in synthetic defined (SD) medium (Takara Bio), supplemented with drop-out amino acids (BUFFERAD) at 30 °C with shaking at 225 rpm. Induction of gene expression was performed using yeast nitrogen base (YNB) medium lacking amino acids and carbon sources, supplemented with ammonium sulfate, raffinose, and galactose (Sigma-Aldrich). To ensure reproducibility, all chemicals and media were of the highest available purity.

2.2. Plasmid construction

Yeast-codon-optimized parts encoding all target proteins were obtained from Twist Bioscience (detailed sequences available in Table S1). Assembly of DNA fragments were performed using the NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs) in accordance with the manufacturer's protocol. PCR amplification was conducted with the Q5 High-Fidelity 2× Master Mix (New England Biolabs), and the resulting products were purified using the Monarch® PCR & DNA Cleanup Kit (New England Biolabs). When necessary, DNA fragments were excised and extracted using the Monarch® DNA Gel Extraction Kit (New England Biolabs). Primers were designed via SnapGene 8.0.1 and synthesized by IDT (USA). Plasmid preparations were carried out using the Monarch® Plasmid Miniprep Kit (New England Biolabs). Site-directed mutagenesis was introduced using the Q5 Site-Directed Mutagenesis Kit (New England Biolabs). To verify plasmid construction, sequencing was performed by Plasmidsaurus (USA).

2.3. Yeast transformation

Transformation of *S. cerevisiae* strains was performed using the Yeast Transformation Kit (YEAST1, Sigma-Aldrich) according to the manufacturer's protocol. For genomic integration, 2 μg of linearized plasmid (digested with NotI-HF, New England Biolabs) was introduced into 50 μL of competent yeast cells. In the case of plasmid-based transformation, 1 μg of circular plasmid DNA was used per 50 μL of yeast cells. Following transformation, the cells were spread on SD-URA-HIS selection plates and incubated at 30 °C for up to 72 h to enable colony formation. Successful genomic integration was verified through colony PCR, followed by sequencing analysis to confirm correct insertion.

2.4. Flow cytometry characterization

To evaluate fluorescence levels, engineered yeast strains were cultured in SD-URA-HIS medium at 30 °C for 48 h, allowing them to reach stationary phase. Cultures were subsequently diluted 1:10 in yeast nitrogen base (YNB) medium supplemented with 5% D-galactose for GAL1 promoter induction and 2% D-raffinose as an additional carbon source. Although 2% D-Galactose is commonly used for induction, 5% is used with the Gal2 transporter knockout to better control fusion protein expression [15]. The cultures were incubated overnight at 30 °C with agitation in a 96-well plate shaker. Following incubation, cells were harvested by centrifugation, washed twice with phosphate-buffered saline (PBS), and resuspended in fresh PBS. For flow cytometric analysis, samples were further diluted in PBS and transferred into a 96-well plate for automated analysis using an SA3800 Spectral Cell Analyser (Sony Biotechnology). Single-cell populations were gated based on forward scatter (FSC) and side scatter (SSC) parameters in a logarithmic scale, with 10,000 events recorded per sample. ZsGreen fluorescence was detected using a 488 nm excitation laser and a 495–510 nm emission filter. ZsGreen, a tetrameric green fluorescent protein derived from *Zoanthus* sp., was used as the reporter gene to monitor gene expression. Due to its high brightness and expression efficiency in *S. cerevisiae* [26], it was selected as a suitable reporter gene in our yeast-based screening system. The geometric mean fluorescence from three independent replicates was used for comparative analysis, with fold changes calculated as the ratio of mean fluorescence intensity of each T7 RNAP or BMCE variant relative to the wild-type control. Data analysis and processing were performed using FlowJo (version 10.10.0).

2.5. Fluorescence microscopy

Yeast cells were harvested by centrifugation, washed, and resuspended in distilled water (DW). The resulting cell suspensions were observed using a Zeiss Axiovert 200 M inverted microscope equipped with a 63× objective lens. Green fluorescence was detected using a FITC filter set, with an exposure time of 11 s. Imaging was performed at the Center for Biomedical Research Support Microscopy and Imaging Facility at The University of Texas at Austin (RRID:SCR_021756).

2.6. ML-predictions for T7 RNAP engineering

T7 RNA polymerase (T7 RNAP) transitions between two key conformational states during its catalytic cycle: the initiation complex (IC) and the elongation complex (EC). The IC is unstable and produces short RNA fragments, known as abortive transcripts, through a process called abortive cycling (PDB: 1QLN, 1CEZ, 2PI4, 2PI5), whereas the EC

is stable and processive (PDB: 1S0V, 1S76, 1S77, 1H38, 1MSW) [27,28]. We were agnostic regarding whether improvements to the IC or the EC might best enhance the overall activity of the polymerase in yeast, and thus used MutCompute models [22,23] to generate predictions by analyzing the nine crystal structures of T7 RNAP. Both MutCompute (a convolutional neural network [22], https://mutcompute.com) and MutComputeX (a residual neural network [23], https://github.com/danny305/MutComputeX) were used to analyze each of the nine PDB structures. Point mutants were selected based on high probability ratios, calculated as either the $\log_2$ ratio (for MutCompute) or the natural log ratio (for MutComputeX) of the probability score of the highest-ranked amino acid to that of the wild-type residue at the same position ($\log_2(\frac{mutAA\_probability}{wtAA\_probability})$ for MutCompute, $\log(\frac{mutAA\_probability}{wtAA\_probability})$ for MutComputeX). The choice of log base reflects differences in model design: MutCompute, optimized for discrete probability comparisons, applies $\log_2$ for intuitive interpretation in terms of fold changes, while MutComputeX, trained to predict continuous stability changes (e.g., ΔTm, ΔΔG), uses the natural log to better correlate with thermodynamic parameters. These values were averaged across multiple structures, irrespective of whether the structure was an IC or an EC (Table S2). We initially selected 20 mutants from the MutCompute predictions and 20 mutants from the MutComputeX predictions. In general, predictions by MutComputeX yielded more consistent results across multiple input structures. To explore whether simplifying the input representation might uncover additional functional variants, we tested a modified configuration of MutComputeX—referred to as MutComputeX.2—by excluding charge and solvent-accessible surface area (SASA) features. Although overall predictive performance was not enhanced, one beneficial variant (A584K) was identified and incorporated into high-performing combinations, suggesting the potential utility of this configuration in expanding sequence space coverage. All variants were experimentally tested, and mutations were stacked based on their relative activities in yeast. From the top five single substitutions, all possible double, triple, quadruple, and quintuple substitutions were assayed, and the best quadruple and quintuple substitutions were carried forward.

To further evaluate predictive models, we also employed the program MutRank [25] to identify mutations with potential functional benefits. MutRank leverages EvoRank [25], a self-supervised learning-based ranking framework, which incorporates evolutionary information from multiple sequence alignments (MSAs) to prioritize beneficial mutations. Unlike traditional wild-type accuracy-based models, which focus on recovering known amino acids, EvoRank ranks amino acid substitutions based on their evolutionary likelihood, improving its ability to predict functionally advantageous mutations. MutRank was run on ten PDB structures, including the nine used in MutCompute predictions (1QLN, 1CEZ, 2PI4, 2PI5, 1S0V, 1S76, 1S77, 1H38, and 1MSW), with the addition of 1ARO, which represents T7 RNAP complexed with T7 lysozyme. MutRank was implemented based on the methodology described in [25], without additional modification.

Variants were ranked based on their pred_prob values. MutRank employs a ranking-based learning framework, referred to as EvoRank loss, to identify amino acid substitutions that align more closely with evolutionary constraints and functional requirements. Unlike MutCompute and MutComputeX, which assess mutations through absolute probability ratios, MutRank instead estimates the relative ranking of amino acids at a given site. This is formulated as:

$$r_i(aa^+, aa^-) = \frac{P_j^{MSA}(aa^+)}{P_j^{MSA}(aa^+) + P_j^{MSA}(aa^-)} - \frac{1}{2}$$

where $P_j^{MSA}(aa^+)$ and $P_j^{MSA}(aa^-)$ denote the probability values derived from multiple sequence alignments (MSAs) for two competing amino acids at position j. Unlike previous models that focused on predicting individual amino acid probabilities, MutRank learns the evolutionary hierarchy of amino acids, enabling the model to capture functional fitness landscapes beyond conventional probability-based approaches. In this context, pred_prob values generated by MutRank do not represent absolute confidence scores but rather the relative probability that a given mutation would be observed in an evolutionary setting. A higher pred_prob suggests that the model assigns greater likelihood to the mutated residue over alternative substitutions, reflecting both evolutionary constraints inferred from sequence conservation patterns and functional fitness predicted from structural stability data. This transition from probability-based assessment to ranking-based evaluation facilitates a more biologically meaningful approach to prioritizing mutations, particularly in scenarios where evolutionary selection pressures extend beyond thermodynamic stability alone. Those variants with pred_prob scores higher than 2 were selected for manual examination via visualization software; twelve mutations were ultimately chosen based on their apparent ability to better fit into the chemical microenvironment (based on evaluation of hydrophobicity, solvent-accessibility, flexibility, and steric hindrance) compared to the wild-type amino acid. The MutRank predictions were further introduced into the best quadruple and quintuple mutant backgrounds, previously identified through MutCompute predictions, and subsequently assayed. Overall, a total of 72 single-substitution variants were predicted using four ML models—MutCompute (20 variants), MutComputeX (20), MutComputeX.2 (20), and MutRank (12)—and all were experimentally tested in yeast. These variants were selected based on model-specific scoring metrics and manually evaluated for structural plausibility. Their predicted scores and annotations are summarized in Table S2.

2.7. ML-predictions for BMCE engineering

To improve the activity of BMCE, we implemented a systematic engineering approach incorporating MutCompute [22], MutRank [25], and Stability Oracle [24] (https://github.com/danny305/StabilityOracle) predictions. This method facilitated the discovery of mutations with potential functional benefits by integrating machine learning-based structure-function analysis. For the initial stage, MutCompute was applied to analyze BMCE, using its AlphaFold 2 [29]-predicted structure as the input model. The predictions generated by MutCompute were ranked based on the log ratio comparing wild-type residues to potential substitutions. In addition, MutRank was employed to independently identify 20 single mutations from its prediction dataset, selecting variants with a high probability of contributing to BMCE activity. Finally, a distinct set of 20 mutations was selected using Stability Oracle [24], a graph-transformer-based model designed to assess the thermodynamic stability (ΔΔG) of protein variants. In contrast to previous methods such as Rosetta [30] and FoldX [31], which require explicit modeling of both wild-type and mutant structures, Stability Oracle predicts stability changes using a structure-informed approach that integrates local structural data with graph-based attention mechanisms. This method uses a graph-transformer architecture, where atoms are represented as nodes and atomic distances are treated as edges, guiding the attention mechanism to focus on relevant structural features. The core prediction is based on thermodynamic permutations (TP), a technique that enhances ΔΔG predictions by leveraging the Gibbs free energy state-function property. ΔΔG predictions indicate stabilizing or destabilizing mutations, where negative values represent stabilizing mutations and positive values indicate destabilizing mutations. The prediction model is based on the following equation:

$$\triangle\triangle G = W \cdot (e_{mut} - e_{wt})$$

In this equation, e_{mut} and e_{wt} represent the embedding vectors of the mutated and wild-type amino acids, respectively, capturing their structural and chemical properties. The difference between these embeddings is scaled by a learned weight matrix W, transforming the difference into a prediction of the thermodynamic stability change (ΔΔG). ΔΔG quantifies the relative impact of a mutation on stability,

considering the local structural context. A higher ΔΔG suggests a destabilizing effect, while a lower ΔΔG indicates stabilization. This transition from physics-based methods to graph-based learning provides a more biologically relevant approach for predicting mutations, particularly in cases where evolutionary pressures go beyond simple thermodynamic calculations. In total, the final BMCE dataset comprised 60 mutations, with 20 mutations selected from each of MutCompute, MutRank, and Stability Oracle. Among them, 58 unique mutations were chosen (with two overlapping across methods) to maximize BMCE activity and efficiency. A detailed list of selected mutations and their predicted scores can be found in Table S3. For BMCE, we employed complementary models—MutCompute, MutRank, and Stability Oracle—due to the absence of experimentally resolved structures, allowing us to integrate structure-, evolution-, and stability-based predictions.

3. Results

3.1. Engineering of T7 RNA polymerase:cappase based on structurally-aware machine learning predictions

Machine learning methodologies offer transformative opportunities to explore protein sequence space more deeply, enabling the identification of key residues critical for protein function, even when located outside the traditional active site. Many of these approaches model the relationship between protein structure and function, predicting the effects of mutations on protein activity or stability [16,32,33]. In this study, we employed MutCompute, a structure-informed machine learning algorithm for protein engineering [22,34,35], to optimize T7 RNAP for improved function in yeast. Both MutCompute [22], a convolutional neural network, and MutComputeX [23], a residual neural network, were applied to analyze multiple structural conformations from the Protein Data Bank (PDB), incorporating both initiation complex (IC) and elongation complex (EC) states (PDB IDs: 1QLN, 1CEZ, 2PI4, 2PI5 for IC; 1S0V, 1S76, 1S77, 1H38, and 1MSW for EC). To mitigate potential biases introduced by multiple structural representations within a single PDB file, we applied the model to each chain separately and averaged the resulting per-residue prediction scores across chains. We attempted to identify mutations that were predicted across multiple different structure files, including both the IC and the EC, and point mutants were selected based on high log ratio probability comparisons relative to the wild-type residue. To further improve the robustness of our predictions, we evaluated whether excluding charge and solvent-accessible surface area (SASA) as input features would impact the stability-focused predictions of T7 RNAP. This led to the development of MutComputeX.2, which generated an additional set of 20 predicted variants without these input features.

These predictions ultimately led to 60 single variants spanning 56 positions (20 from MutCompute, 20 from MutComputeX, and 20 from MutComputeX.2), with four positions (L534, V710, L853, and H854) having two, different predicted substitutions. These variants were then incorporated into a previously studied [11–13,15] fusion protein consisting of T7 RNAP and the African Swine Fever Virus (ASFV) NP868R mRNA capping enzyme (Table S2). Interestingly, we observed that MutCompute primarily predicted mutations in the IC (14 out of 20), whereas MutComputeX predictions were more balanced between the IC (9 out of 20) and EC (11 out of 20). It is possible that MutComputeX, which employs a residual neural network, may have captured higher-order structural dependencies that are relevant to both IC and EC states. IC and EC classifications were used solely to annotate the structural origin of each prediction and were not used to stratify or interpret the results of functional screening experiments.

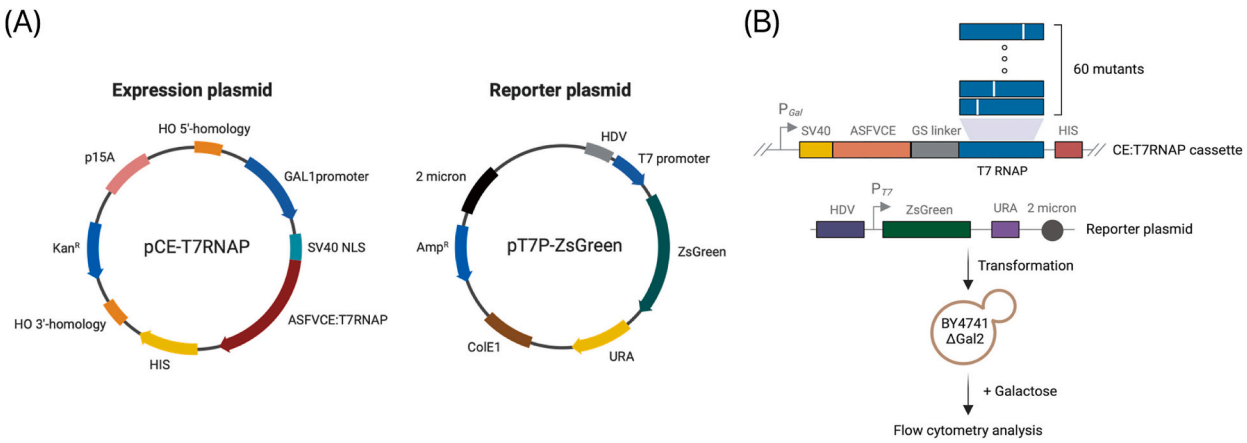
The ASFV NP868R cappase was initially chosen based on its demonstrated ability to enhance protein expression in both mammalian systems [11,12] and in yeast [15]. The fusion domains contained an SV40 nuclear localization signal (NLS), were linked by a GS linker, and were expressed under the control of the yeast GAL promoter [15] (Fig. 1A). To ensure proper transcriptional regulation, a hepatitis delta virus (HDV) ribozyme was positioned upstream of the P_{T7} promoter. This fusion architecture was adopted based on previous studies reporting that separate expression of mRNA capping enzymes and T7 RNAP resulted in reduced transcriptional and translational efficiency in eukaryotic systems. In contrast, physically linking the two enzymes improves co-transcriptional capping and significantly enhances protein expression [11,12]. Fusion proteins carrying the T7 RNAP variants were integrated into the HO locus of the yeast chromosome. To determine the activity of the fusion proteins in the ΔGal2 derivative yeast strain, induction was carried out with 5% D-Galactose. Following induction, the variants transcribed the *ZsGreen* reporter gene under the control of the T7 RNAP promoter on an episomal reporter plasmid (pT7P-ZsGreen; Fig. 1A and B). A null variant, the substitution Y639A [36,37], was used to validate the system as a negative control (Fig. 1C).

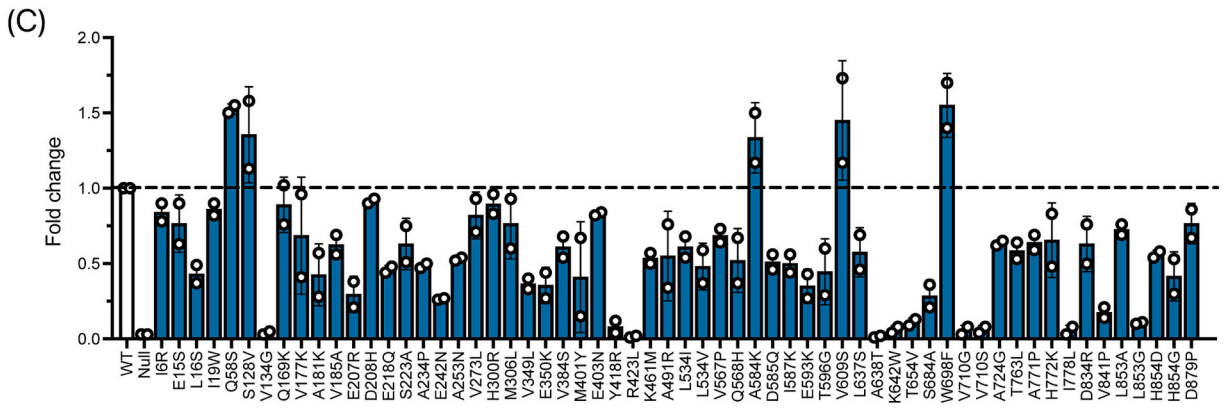
Out of the 60 single variants tested, five (Q58S, S128V, A584K, V609S, and W698F) led to up to 1.5-fold increases in expression of the *ZsGreen* reporter gene compared to wild-type (WT) T7 RNAP (Fig. 1C). These beneficial mutations originated from different MutCompute models: V609S from MutCompute; Q58S, S128V, and W698F from MutComputeX; and A584K from MutComputeX.2. We then systematically combined the five beneficial single variants (Q58S, S128V, A584K, V609S, and W698F) to generate all possible double, triple, quadruple, and quintuple combinations, resulting in 26 unique multi-mutant variants for further analysis (Fig. 1D). In general, the combinations showed increasing activity, with the quadruple variant T7 $RNAP^{Q58S/A584K/V609S/W698F}$ and the quintuple variant (T7 $RNAP^{Q58S/S128V/A584K/V609S/W698F}$) having 3.9- and 3.5-fold increases in activity, respectively, relative to WT ASFVCE:T7 RNAP (Fig. 1D). Fluorescence microscopy analysis was performed to qualitatively assess protein expression. The WT, null, quadruple (Q58S/A584K/V609S/W698F), and quintuple (Q58S/S128V/A584K/V609S/W698F) variants of ASFVCE:T7 RNAP were examined. Consistent with the quantitative fluorescence data, the quadruple and quintuple variants exhibited markedly stronger green fluorescence signals compared to the WT, whereas the null variant showed minimal fluorescence. These observations further confirm that the selected mutations enhance the expression of the *ZsGreen* reporter gene (Fig. S1).

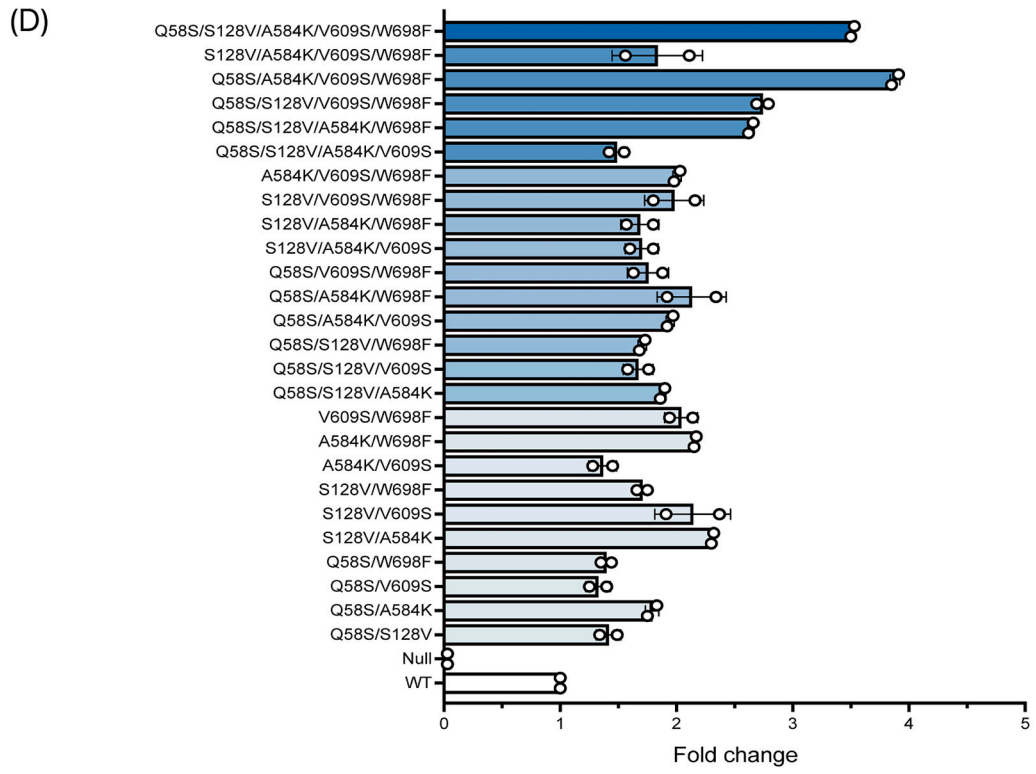
3.2. Further engineering of T7 RNAP based on evolutionarily aware machine learning predictions

As has been seen with numerous other proteins [22,23,34,35], MutCompute and other structure-aware machine learning algorithms were able to identify substitutions beneficial for activity, and these substitutions could generally be stacked in a roughly additive way. However, the exhaustive procedure we used for stacking (all combinations of the best variants) becomes less feasible as the number of mutations (and thus the number of paths for stacking) increases. We therefore employed MutRank, a self-supervised learning-based ranking framework that prioritizes beneficial mutations based on evolutionary likelihood [25] based on multiple sequence alignment (MSA) data. We hypothesized that MutRank might see additional mutations that MutCompute and other structurally aware algorithms could not.

To test this hypothesis, we selected twelve MutRank-predicted single variants (Table S2) and introduced them individually into the previous quadruple and quintuple backgrounds, which served as scaffolds for the next round of engineering. Of the twelve single substitutions added to the quadruple background (T7 $RNAP^{Q58S/A584K/V609S/W698F}$), five (C125I, C347P, Q404L, N419T, and Q786L) exhibited increased activity (Fig. 2A). An additional 12 combinations were evaluated, and three additional variants demonstrated significant improvements in activity, with the best variant containing both Q404L and Q786L, and having a 1.6-fold increase in activity compared to the original quadruple variant, corresponding to a 6.2-fold increase relative to ASFVCE(WT):T7 RNAP (WT) (Fig. 2A). We designated this variant EvoT7, and carried it forward







(caption on next page)

Fig. 1. Engineering T7 RNA polymerase using structure-aware machine learning predictions for enhanced target gene expression. (A) Two-plasmid system for screening T7 RNAP variants. The expression plasmid, containing the ASFVCE:T7 RNAP fusion protein, was selected using the HIS3 marker, while the reporter plasmid, encoding the *ZsGreen* reporter gene, was selected using the URA3 marker. Both plasmids were transformed into a ΔGal2 derivative of *S. cerevisiae* for screening gene expression. (B) Schematic representation of the ASFVCE:T7 RNAP expression cassette and reporter plasmid. T7 RNAP variants fused with ASFVCE via a GS linker were expressed under the control of the P_{Gal} promoter. The *ZsGreen* reporter gene, controlled by the P_{T7} promoter, was expressed from a plasmid maintained using the 2-micron system. (C) Screening of single-point mutations in T7 RNAP. A total of 60 single-point mutations were introduced into T7 RNAP fused with ASFVCE to evaluate their impact on reporter gene expression. (D) Screening of combinatorial mutations in T7 RNAP. A total of 26 combinatorial mutations, including double, triple, quadruple, and quintuple mutations, were introduced into T7 RNAP fused with ASFVCE to assess their effect on reporter gene expression. For each mutation, fold change was calculated as the ratio of ZsGreen expression, measured by flow cytometry, relative to WT T7 RNAP (set as 1) under identical galactose induction conditions. The "Null" construct, containing the Y629A mutation, was used as a negative control since this mutation impairs T7 RNAP activity. Each dot represents the average value obtained from a transformation experiment, with three biological replicates per transformation. Bar graphs indicate the mean and standard deviation (SD).

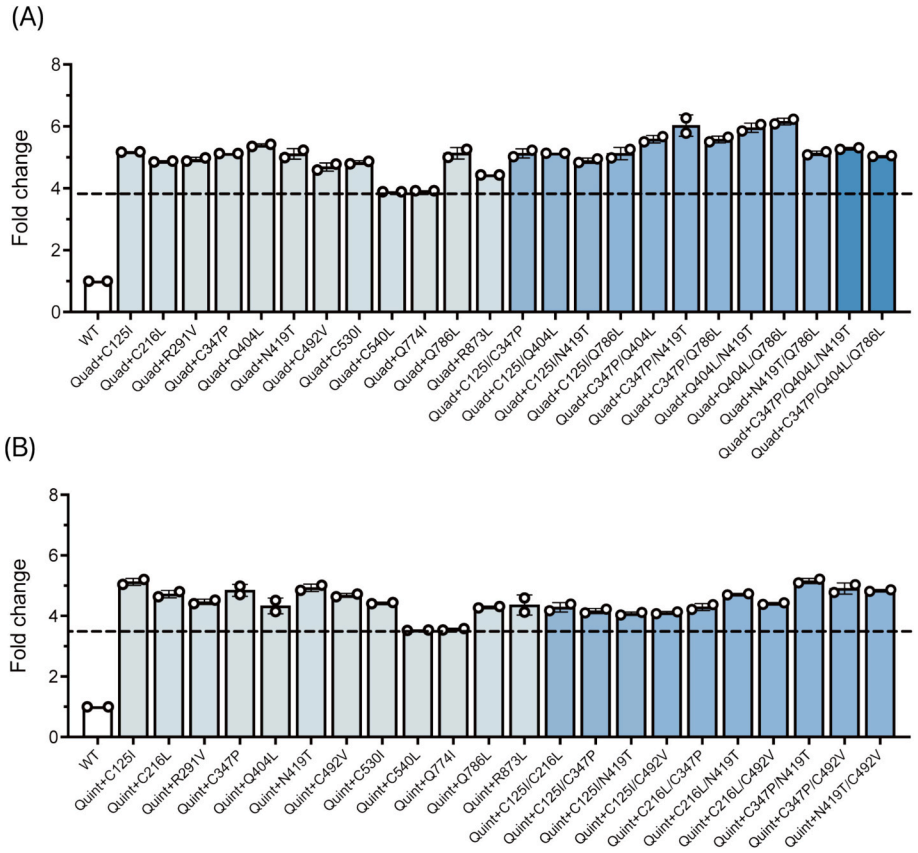


Fig. 2. Engineering quadruple and quintuple T7 RNAP variants using evolution-aware machine learning predictions. (A) Screening of single and combinatorial mutations in the quadruple T7 RNAP variant (T7 $RNAP^{Q58S/A584K/V609S/W698F}$). A total of 24 mutations, including single, double, and triple mutations, were introduced into the quadruple T7 RNAP variant fused with ASFVCE to assess their effect on reporter gene expression. (B) Screening of single and combinatorial mutations in the quintuple T7 RNAP variant (T7 $RNAP^{Q58S/S128V/A584K/V609S/W698F}$). A total of 22 mutations, including single and double mutations, were introduced into the quintuple T7 RNAP variant fused with ASFVCE to evaluate their impact on reporter gene expression. Fold change was determined by comparing ZsGreen fluorescence intensity, as measured via flow cytometry, relative to WT T7 RNAP (set as 1) under identical galactose induction conditions. The fold change of the quadruple and quintuple T7 RNAP variants is shown as a horizontal dashed line in (A) and (B), respectively. Each dot represents the average value from a transformation experiment, with three biological replicates per transformation. Bar graphs indicate the mean and standard deviation (SD).

into further experiments, as described below. A similar approach was applied using the quintuple variant as the parent enzyme, and quintuple$^{C347P/N419T}$ exhibited the highest activity, with a 5.2-fold increase relative to ASFVCE(WT):T7 RNAP(WT) (Fig. 2B).

3.3. Identification of a more highly active virus-derived mRNA capping enzyme

The addition of a 5′-m7G cap to eukaryotic mRNA is crucial, serving to prevent degradation and facilitate translation [38]. Previous research has identified several mRNA capping enzymes that can be used in mammalian cells and compared their activities through biochemical and functional assays [11]. Among these, ASFVCE exhibited the highest activity and was subsequently engineered to enhance its performance in yeast-based expression systems [15]. While these modifications improved its functionality, they were still based on a single viral source, prompting an investigation into whether alternative wild-type viral capping enzymes might exhibit even greater activity in yeast. To identify a single-subunit RNA capping enzyme with higher activity than the ASFVCE, we screened nine promising capping enzyme (CE) fusion candidates [39], again using our fluorescent reporter system (Fig. 3A and B). Among the nine tested enzymes, MACE and APMCE showed minimal or undetectable reporter gene expression, comparable to the null mutant of ASFVCE (ASFVCE^{K282N}) and the negative control (NC), which consisted of WT T7 RNAP without a capping enzyme (Fig. 3C and D). The enzymes FCE, PCE, BSVCE, GMCE, and NCE displayed activity levels

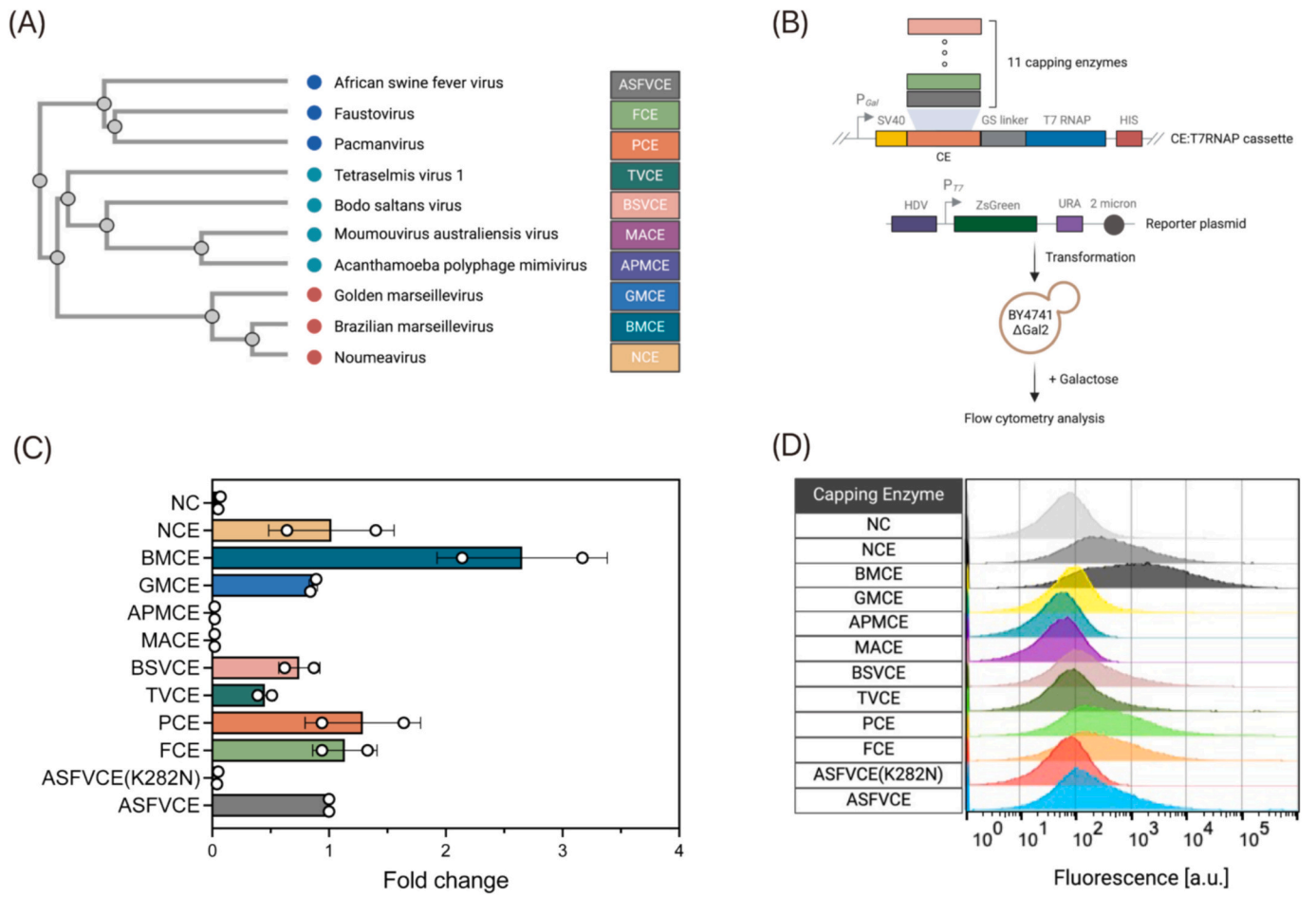


Fig. 3. Screening and characterization of single-subunit CEs. (A) Phylogenetic tree of single-subunit CE candidates. Phylogenetic relationships among 10 viral-derived single-subunit CEs, including ASFVCE and other enzymes, are shown. (B) Schematic representation of the CE:T7 RNAP expression cassette and reporter plasmid. Single-subunit CE variants were fused to WT T7 RNAP via a GS linker and expressed under the control of the P_{Gal} promoter. (C) Screening of single-subunit CEs. A total of 10 viral-derived single-subunit CEs, including WT ASFVCE and the ASFVCE (K282N) null mutant, were tested for their impact on reporter gene expression. Fold change in ZsGreen fluorescence was determined by normalizing each CE variant to the WT ASFVCE (set as 1), with measurements obtained under identical galactose induction conditions via flow cytometry. Each dot represents the average value obtained from three colonies per transformation, and the bar graphs indicate the mean and standard deviation (SD). (D) Evaluation of *ZsGreen* reporter gene fluorescence. Flow cytometry histograms illustrate the fluorescence intensity distribution, where a rightward shift represents increased CE activity relative to the negative control. ASFVCE(K282N) served as the null mutant control, displaying baseline fluorescence levels, while the negative control (NC) represented WT T7 RNAP expressed without a capping enzyme.

comparable to or slightly lower than ASFVCE, while BMCE (from Brazilian Marseillevirus) exhibited an activity level around 2-fold higher than the previously used ASFVCE.

As with T7 RNA polymerase, we pursued further enhancement of BMCE activity through ML-guided predictions. Starting from the AlphaFold 2-prediced structure of BMCE, we applied three ML models – MutCompute, MutRank, and Stability Oracle – to predict beneficial mutations. Each model identified 20 mutations, and after removing duplicates (S515G, H786F), a total of 58 unique single variants were generated (Table S3). To evaluate the activity of these variants, we employed the same two-plasmid system as in Fig. 1 (Fig. 4a). Among the 58 single variants, seven (D59E, M165L, S199A, N267W, Q426V, Q426Y, and D582F) exhibited up to 1.4-fold increased *ZsGreen* expression compared to WT BMCE (Fig. 4B). As before, the top-performing single mutants were combined to generate all possible double mutants, resulting in 20 variants (Fig. 4C). Two of the double mutants (M165L/Q426Y and S199A/Q426Y) yielded over a 2-fold increase in activity compared to WT BMCE. Additionally, four double mutants (M165L/Q426V, M165L/D582F, S199A/D582F, and N267W/Q426V) exhibited moderate improvements of over 1.5-fold (Fig. 4C). Based on the six selected double mutants that exhibited enhanced activity, we systematically generated 13 triple mutants. Three of these variants had additive improvements, up to a 3-fold final increase in activity. Further stacking of quadruple and quintuple mutants did not yield further improvements in activity.

3.4. Combining the best engineered cappase and T7 RNAP enzymes

The triple variant BMCE$^{S199A/N267W/Q426V}$ (hereafter referred to as EvoBMCE) was chosen for further testing with improved T7 RNAP variants. Similarly, in order to best combine previous directed evolution efforts with substitutions garnered from machine learning, we generated EvoT7(443), which contained 12 mutations (6 from T7 RNAP(443) and 6 from EvoT7). Fusion proteins were constructed using the most effective capping enzymes (ASFVCE(443) and EvoBMCE) and T7 RNAP variants (T7 RNAP(443), EvoT7, and EvoT7(443)), derived either from directed evolution, and their activities were assessed (Fig. 5A) relative to suitable controls. Gene expression was evaluated by measuring ZsGreen fluorescence normalized to OD_{600nm}, using ASFVCE(WT)-T7 RNAP(WT) as the baseline (set to 1).

Significant further enhancements in gene expression efficiency relative to previous cappase-polymerase combinations were found, and these were due to both improvements in the cappase and improvements in the polymerase (Fig. 5B). We had already observed a roughly 2-fold

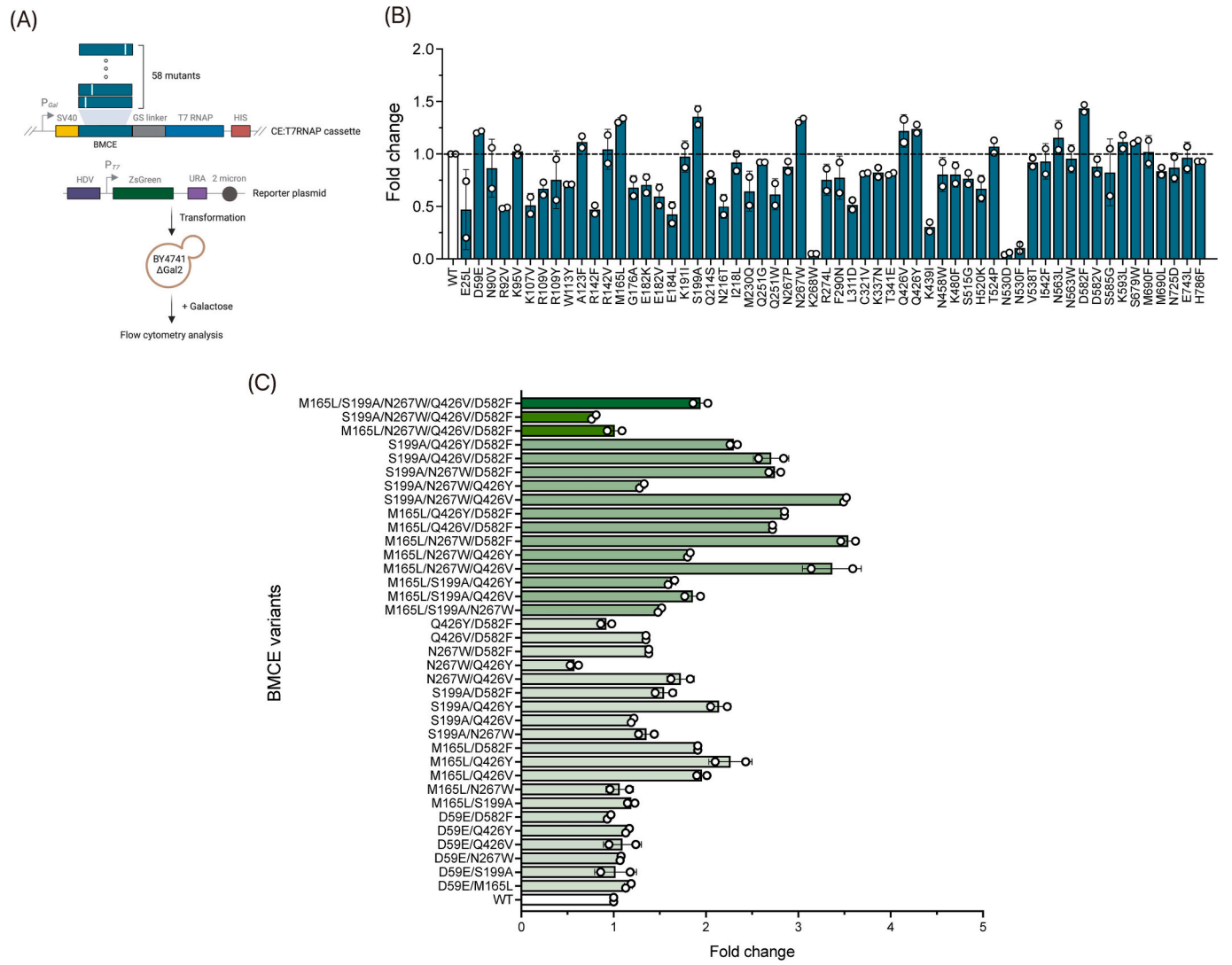


Fig. 4. Engineering BMCE for enhanced target gene expression. (A) Schematic representation of the BMCE:T7 RNAP expression cassette and reporter plasmid. BMCE variants fused to T7 RNAP via a GS linker were expressed under the control of the P_{Gal} promoter. (B) Screening of single-point mutations in BMCE. A total of 58 single-point mutations were introduced into BMCE fused with T7 RNAP to evaluate their impact on reporter gene expression. (C) Screening of combinatorial mutations in BMCE. A total of 36 combinatorial mutations, including double, triple, quadruple, and quintuple mutations, were introduced into BMCE fused with T7 RNAP to assess their effect on reporter gene expression. For each mutation, fold change was calculated as the ratio of ZsGreen expression, measured by flow cytometry, relative to WT BMCE (set as 1) under identical galactose induction conditions. Each dot represents the average value obtained from a transformation experiment, with three biological replicates per transformation. Bar graphs indicate the mean and standard deviation (SD).

improvement of activity relative to wild-type when BMCE cappase was used in place of the ASFVCE enzyme (Fig. 3C). The introduction of the improved EvoBMCE further improved transcriptional efficiency, up to 5.6-fold above the previously developed wild-type enzyme combination of ASFVCE:T7 RNAP. These improvements were similar to those seen in fusions between the evolved ASFVCE(443) cappase and the wild-type enzyme. The use of improved T7 RNAP variants also proved to be roughly additive, with the directed evolution (443) and machine learning (EvoT7) variants both leading to 2- to 3-fold improvements over parental enzyme combinations. Remarkably, the directed evolution and machine learning variants could be stacked together (EvoBMCE-EvoT7(443)) to attain even higher activity, upwards of an overall 7-fold improvement over the wild-type enzyme combination (BMCE-T7 RNAP (WT)) and greater than 12-fold improvement over the previously studied ASFVCE(WT)-T7 RNAP(WT) fusion. While direct comparisons were not made in the same experiment, stepwise gains observed between the non-fusion control (NC), the original fusion enzyme, and the final construct suggest a cumulative enhancement exceeding 100-fold.

4. Discussion

In this study, we employed a machine learning (ML)-driven approach to optimize fusion constructs between cappase enzymes (either ASFVCE or BMCE) and T7 RNA polymerase (T7 RNAP) for improved transcription and gene expression in yeast. MutCompute, a structure-based ML model, was initially used to predict beneficial mutations by evaluating multiple T7 RNAP conformations derived from Protein Data Bank (PDB) structures. Specifically, MutCompute, a convolutional neural network-based model, and MutComputeX, a residual neural network-based approach, were applied to predict mutations that could enhance polymerase function by analyzing structural features of both the initiation complex and elongation complex states. The model's predictions were subsequently filtered to select the most promising variants based on a log-ratio scoring system, prioritizing mutations predicted to be more favorable than the wild-type residue. Additionally, MutComputeX.2 was developed by modifying input features—excluding charge and solvent-accessible surface area (SASA)—to assess whether this adjustment

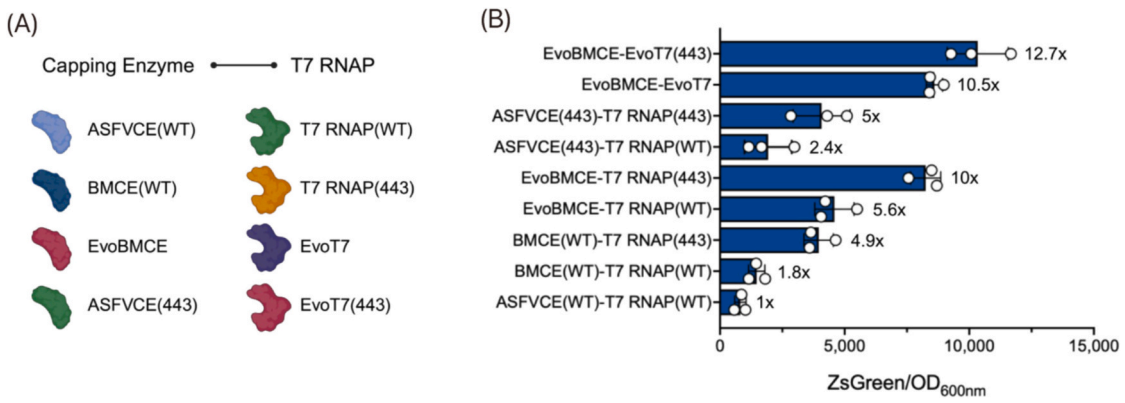


Fig. 5. Comparative activity of combinatorial capping enzyme (CE) and T7 RNAP variants on reporter gene expression. (A) Schematic representation of combinatorial mutations. Four capping enzyme (CE) variants (ASFVCE(WT), BMCE(WT), EvoBMCE, and ASFVCE(443)) and four T7 RNAP variants (T7 RNAP(WT), T7 RNAP(443), EvoT7, and EvoT7(443)) were combined to generate multiple CE-T7 RNAP fusion constructs for activity assessment. EvoBMCE corresponds to the triple variant BMCE$^{S199A/N267W/Q426V}$, and EvoT7(443) comprises 12 mutations (six from T7 RNAP(443) and six from EvoT7), integrating both directed evolution and machine learning-derived substitutions. (B) Characterization of combinatorial variants for reporter gene expression. Reporter gene expression was quantified by measuring ZsGreen fluorescence normalized to OD_{600nm}. ZsGreen/OD_{600nm} values were assessed for each combinatorial CE-T7 RNAP construct and compared to the baseline ASFVCE(WT)-T7 RNAP(WT), which was set to 1. Each dot represents the average value obtained from three colonies per transformation, with bar graphs indicating the mean and standard deviation (SD).

could improve stability-focused predictions.

All machine learning models—including MutCompute, MutComputeX, MutRank, and Stability Oracle—were used solely for inference, without any retraining or fine-tuning. Each model was applied using its default settings as described in the original publications or associated repositories, and all predictions were performed using standard laboratory computing resources. These ML algorithms contributed distinct yet complementary insights, and overall described a broad fitness landscape for experimental validation. MutCompute and its variants focused on structural constrains, MutRank prioritized evolutionarily favorable changes, and Stability Oracle identified stability-enhancing substitutions. Each program eventually identified useful, non-overlapping substitutions. For T7 RNAP, one improved substitution originated from MutCompute, three were from MutComputeX, and one was from MutComputeX.2. The relative underperformance of MutComputeX.2 compared to MutComputeX suggests that removing charge and SASA features, both of which play critical roles in protein stability and function [40,41], may have impeded the selection of beneficial substitutions. Similarly, for the BMCE cappase variants, seven beneficial mutations were found: one from MutCompute, one from MutRank, and five from Stability Oracle. Notably, Stability Oracle was not applied to T7 RNAP in this study, as multiple high-resolution crystal structures enabled structure-rich modeling using MutCompute, MutComputeX, and MutRank. In contrast, it was applied to BMCE, which lacked an experimentally determined structure. The minimal overlap among top-ranked variants suggests that each model encodes distinct selection preferences.

Beyond model-specific variation, our ML-guided designs also diverged substantially from prior studies. Interestingly, our machine-learning-based optimizations of T7 RNAP led to significantly different results compared to previous studies (Table S4). In particular, EVOLVEpro-assisted directed evolution led to epT7 (T3M/G47A/E643G) [21]; structure-based rational engineering with Rosetta resulted in G47A/884G [5]; and high-throughput mutagenesis coupled with FACS-based selection led to our previous best cappase-polymerase combination, v443 (N131K/L261M/H300R/R307H/Q648R//H772R) [15]. Along the paths to each of these polymerases, numerous single mutations were tested and either incorporated or discarded (Fig. S2): epT7 tested 42 single substitutions, while G47A/884G tested 21 single substitutions; v443 accumulated mutations iteratively over the course of directed evolution. Only one mutation (H300R) overlapped between our EvoT7 and v443, while no exact residue matches were observed between EvoT7 and epT7 or G47A/884G (Fig. S2A). When considering only potential sites for substitution, as opposed to exact substitutions (Fig. S2B), EvoT7 contained 68 unique mutation sites, epT7 had 33, G47A/884G had 8, and v443 accumulated 20 by the conclusion of the directed evolution process. EvoT7 shared three mutation sites (V134, V177, and V273) with epT7, one with v443 (H300), and none with G47A/884G.

To better understand the structural basis for the improved performance of EvoT7 and EvoBMCE, we generated an in-silico enzyme structure with AlphaFold 3 [42] and examined the atomic interactions gained or lost compared to either the crystal structure of WT T7 RNAP (PDB: 1H38) or the AF 3 wild-type BMCE structure (Fig. S3). AlphaFold 2-predicted structures were used as inputs for ML-guided variant prediction, as these predictions were performed prior to the release of AlphaFold 3 and were optimized for compatibility with model input formats. In contrast, AlphaFold 3 was used post hoc for structural visualization and comparison to help interpret atomic-level changes in evolved variants. Thus, AlphaFold 2 and AlphaFold 3 served distinct purposes in the study. To visualize the distribution of predicted mutations, structural models of EvoT7 and EvoBMCE were annotated with ML model-specific color codes (Fig. S3).

The six mutation sites (Q58S/Q404L/A584K/V609S/W698F/Q786L) in EvoT7 were distributed throughout T7 RNAP, and possible structural improvements span a gamut of chemistries (Fig. S3A). It seems that the Q58S substitution may create a stable hydrogen bond with Ser58 (Fig. S3B), and V609S may form hydrogen bonds with Asn592 and Gln669 (Fig. S3B). The Q404L substitution enhances the hydrophobic core formed by Phe400, Phe408, and Phe432 (Fig. S3B), while Q786L places a hydrophobic residue within the greasy pocket formed by Met549, Thr729, Pro730, Phe782, and Val841 (Fig. S3B). In contrast, A584K mutation installs a positively charged lysine within the solvent-exposed domain and allows polar interactions with nearby polar residues (Glu580, Asp585, and Asn588) or water molecules (Fig. S3B).

It is more difficult to interpret improvements to EvoBMCE, as there is no known structural determination of this enzyme. In the two predicted BMCE structures (EvoBMCE and WT BMCE), S199A, N267W, and Q426V appear to enhance the hydrophobic core of the enzyme (Fig. S3C and D). It is interesting to note that the mutations S199A, N267W, and Q426V were exclusively predicted by the Stability Oracle model, which leverages a structure-based deep learning approach to identify thermodynamically favorable substitutions. Notably, Stability Oracle predicted the majority (5 out of 7) of the beneficial mutations for BMCE cappase variants. The improvements in enzyme activity for EvoBMCE are

particularly impressive when we recall that predictions were done entirely computationally, without an experimentally determined structure.

It is instructive that the different algorithms and models predicted diverse, mostly non-overlapping variants. The fact that different methods all work roughly equally well in improving enzymes is probably a function of the fact that these programs were all trained on roughly the same data: the preponderance of information that we have about wild-type proteins from nature. In essence, the different machine learning approaches to gleaning a 'Platonic ideal' of a wild-type protein, the best protein that accords with what nature has already done, will provide variable but similarly useful insights into how to engineer a protein. There is no one best way to do machine learning for protein improvement. That said, our methods (especially the ability to test in a yeast background, where basal transcription was very low, and improved variants could be more easily identified than in bacteria) allowed us to explore much larger landscapes than previous studies that focused on a limited number of beneficial mutations, and the catalytic superiority of EvoT7 is likely due in part to this broadened exploration. Fortunately, and as we have previously observed [22,23,34,35], the improved single substitutions identified by our structure-based machine learning algorithms could generally be combined to generate variants with increasingly higher activities. We now also observe that machine learning can be utilized to build on the results of previous directed evolution experiments, as evidenced by the combination of the original T7 RNAP(443) with predicted substitutions.

Overall, we have used a diverse, ML-driven strategy to engineer highly active variants of the fusion protein between T7 RNA polymerase (T7 RNAP) and the Brazilian Marseillevirus capping enzyme (BMCE), achieving significant improvements in gene expression in yeast over previous directed evolution efforts. To our knowledge, this study represents the first application of machine learning to co-optimize both components of a transcriptional fusion enzyme, rather than treating each component independently. By leveraging diverse and orthogonal ML models—structure-based (MutCompute), evolution-based (MutRank), and stability-based (Stability Oracle)—we systematically identified beneficial substitutions under different structural constraints, enabling the design of functionally synergistic multi-domain constructs. While the current study focused on engineering a single fusion enzyme, future work may explore co-expression of independently optimized domains to further dissect the mechanistic contributions of domain linkage to transcriptional and translational efficiency.

Our ML-guided fusion engineering strategy expands the scope of ML-guided protein design to more complex, modular systems. Through iterative integration of ML predictions with high-throughput experimental validation, we developed the 9-tuple substitution EvoBMCE ($BMCE^{S199A/N267W/Q426V}$):EvoT7 ($T7\ RNAP^{Q58S/Q404L/A584K/V609S/W698F/Q786L}$) which has greater than 10-fold enhancement in gene expression efficiency compared to wild-type cappase-polymerase combinations, and which could be further combined with substitutions identified by directed evolution to yield even higher activity. In this study, we employed a domain-wise optimization strategy. In future work, applying machine learning directly to full-length fusion constructs may provide additional benefits, potentially enabling the discovery of cross-domain synergistic substitutions if suitable structural data become available. The improved fusion protein has the potential to be assayed and improved in important non-model yeast strains with unique metabolic traits, such as *Yarrowia lipolytica*, *Pichia pastoris*, and *Kluyveromyces marxianus* [43,44], and in mammalian cells [45], potentially leading to high-efficiency gene expression platforms for synthetic biology, mRNA therapeutics, and even cell free systems [46–48].

CRediT authorship contribution statement

Seung-Gyun Woo: Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization, Writing – review & editing, Writing – original draft. **Danny J. Diaz:** Investigation, Formal analysis, Writing – review & editing. **Wantae Kim:** Investigation, Formal analysis, Writing – review & editing. **Mason Galliver:** Investigation, Formal analysis, Writing – review & editing. **Andrew D. Ellington:** Supervision, Project administration, Funding acquisition, Conceptualization, Writing – review & editing, Writing – original draft.

Funding

This work was supported by the Levy-Longenbaugh Fund and Texas Biologics; the Robert J. Kleberg, Jr. and Helen C. Kleberg Foundation [FA00003446]; the National Institutes of Health [1R01EB0277202-01A1]; the Welch Foundation [F-1654]; and the Basic Science Research Program of the National Research Foundation of Korea (NRF), funded by the Ministry of Education [RS-2024-00407966].

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The authors declare that they have filed an invention disclosure related to this work (Invention Disclosure No. 8628 ELL).

Acknowledgements

We thank Shaunak Kar and Jimmy Gollihar for their valuable advice and expert assistance in experimental work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cej.2025.165191.

Data availability

The data supporting the findings of this work are available within the paper and the Supplementary information file.

References

[1] O.G. Ndochinwa, Q.Y. Wang, O.C. Amadi, T.N. Nwagu, C.I. Nnamchi, E.S. Okeke, A.N. Moneke, Current status and emerging frontiers in enzyme engineering: an industrial perspective, Heliyon 10 (11) (2024) e32673, https://doi.org/10.1016/j.heliyon.2024.e32673.
[2] R. Buller, S. Lutz, R.J. Kazlauskas, R. Snajdrova, J.C. Moore, U.T. Bornscheuer, From nature to industry: harnessing enzymes for biocatalysis, Science 382 (6673) (2023) eadh8615, https://doi.org/10.1126/science.adh8615.
[3] Y. Wang, Q. Li, P. Tian, T. Tan, Charting the landscape of RNA polymerases to unleash their potential in strain improvement, Biotechnol. Adv. 54 (2022) 107792, https://doi.org/10.1016/j.biotechadv.2021.107792.
[4] S. Borukhov, E. Nudler, RNA polymerase: the vehicle of transcription, Trends Microbiol. 16 (3) (2008) 126–134, https://doi.org/10.1016/j.tim.2007.12.006.
[5] A. Dousis, K. Ravichandran, E.M. Hobert, M.J. Moore, A.E. Rabideau, An engineered T7 RNA polymerase that produces mRNA free of immunostimulatory byproducts, Nat. Biotechnol. 41 (4) (2023) 560–568, https://doi.org/10.1038/s41587-022-01525-6.
[6] W. Wang, Y. Li, Y. Wang, C. Shi, C. Li, Q. Li, R.J. Linhardt, Bacteriophage T7 transcription system: an enabling tool in synthetic biology, Biotechnol. Adv. 36 (8) (2018) 2129–2137, https://doi.org/10.1016/j.biotechadv.2018.10.001.
[7] W. Wang, X. An, K. Yan, Q. Li, Construction and application of orthogonal T7 expression system in eukaryote: an overview, Adv. Biol. (Weinh.) 7 (2) (2023) e2200218, https://doi.org/10.1002/adbi.202200218.
[8] D.R. Gallie, The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency, Genes Dev. 5 (11) (1991) 2108–2116, https://doi.org/10.1101/gad.5.11.2108.
[9] J. Katahira, Nuclear export of messenger RNA, Genes (Basel) 6 (2) (2015) 163–184, https://doi.org/10.3390/genes6020163.
[10] S. Millevoi, S. Vagner, Molecular mechanisms of eukaryotic pre-mRNA 3′ end processing regulation, Nucleic Acids Res. 38 (9) (2010) 2757–2774, https://doi.org/10.1093/nar/gkp1176.
[11] P.H. Jais, E. Decroly, E. Jacquet, M. Le Boulch, A. Jais, O. Jean-Jean, H. Eaton, P. Ponien, F. Verdier, B. Canard, S. Goncalves, S. Chiron, M. Le Gall, P. Mayeux,

M. Shmulevitz, C3P3-G1: first generation of a eukaryotic artificial cytoplasmic expression system, Nucleic Acids Res. 47 (5) (2019) 2681–2698, https://doi.org/10.1093/nar/gkz069.
[12] C. Qin, Y. Xiang, J. Liu, R. Zhang, Z. Liu, T. Li, Z. Sun, X. Ouyang, Y. Zong, H. M. Zhang, Q. Ouyang, L. Qian, C. Lou, Precise programming of multigene expression stoichiometry in mammalian cells by a modular and programmable transcriptional system, Nat. Commun. 14 (1) (2023) 1500, https://doi.org/10.1038/s41467-023-37244-y.
[13] S.H. Chan, L. Tirole, D. Kneller, T.M. Kelley, N. Dai, G.B. Robb, Co-transcriptional capping using an RNA capping enzyme-T7 RNA polymerase fusion protein, bioRxiv (2023) 2023.10.28.564488, https://doi.org/10.1101/2023.10.28.564488.
[14] M. Le Boulch, E. Jacquet, N. Nhiri, M. Shmulevitz, P.H. Jais, Rational design of an artificial tethered enzyme for non-templated post-transcriptional mRNA polyadenylation by the second generation of the C3P3 system, Sci. Rep. 14 (1) (2024) 5156, https://doi.org/10.1038/s41598-024-55947-0.
[15] S. Kar, E.C. Gardner, K. Javanmardi, D.R. Boutz, R. Shroff, A.P. Horton, T.H. Segall-Shapiro, A.D. Ellington, J. Gollihar, Directed evolution of an orthogonal transcription engine for programmable gene expression in eukaryotes, iScience 28 (1) (2025), https://doi.org/10.1016/j.isci.2024.111541.
[16] C.R. Freschlin, S.A. Fahlberg, P.A. Romero, Machine learning to navigate fitness landscapes for protein engineering, Curr. Opin. Biotechnol. 75 (2022) 102713, https://doi.org/10.1016/j.copbio.2022.102713.
[17] T. Yu, A.G. Boob, N. Singh, Y. Su, H. Zhao, In vitro continuous protein evolution empowered by machine learning and automation, Cell Syst. 14 (8) (2023) 633–644, https://doi.org/10.1016/j.cels.2023.04.006.
[18] B.J. Wittmann, K.E. Johnston, Z. Wu, F.H. Arnold, Advances in machine learning for directed evolution, Curr. Opin. Struct. Biol. 69 (2021) 11–18, https://doi.org/10.1016/j.sbi.2021.01.008.
[19] N.E. Siedhoff, U. Schwaneberg, M.D. Davari, Machine learning-assisted enzyme engineering, Methods Enzymol. 643 (2020) 281–315, https://doi.org/10.1016/bs.mie.2020.05.005.
[20] S. Towers, J. James, H. Steel, I. Kempf, Learning-based estimation of fitness landscape ruggedness for directed evolution, bioRxiv (2024) 2024.02.28.582468, https://doi.org/10.1101/2024.02.28.582468.
[21] K. Jiang, Z. Yan, M. Di Bernardo, S.R. Sgrizzi, L. Villiger, A. Kayabolen, B.J. Kim, J. K. Carscadden, M. Hiraizumi, H. Nishimasu, J.S. Gootenberg, O.O. Abudayyeh, Rapid in silico directed evolution by a protein language model with EVOLVEpro, Science 387 (6732) (2025) eadr6006, https://doi.org/10.1126/science.adr6006.
[22] R. Shroff, A.W. Cole, D.J. Diaz, B.R. Morrow, I. Donnell, A. Annapareddy, J. Gollihar, A.D. Ellington, R. Thyer, Discovery of novel gain-of-function mutations guided by structure-based deep learning, ACS Synth. Biol. 9 (11) (2020) 2927–2935, https://doi.org/10.1021/acssynbio.0c00345.
[23] S. d'Oelsnitz, D.J. Diaz, W. Kim, D.J. Acosta, T.L. Dangerfield, M.W. Schechter, M. B. Minus, J.R. Howard, H. Do, J.M. Loy, H.S. Alper, Y.J. Zhang, A.D. Ellington, Biosensor and machine learning-aided engineering of an amaryllidaceae enzyme, Nat. Commun. 15 (1) (2024) 2084, https://doi.org/10.1038/s41467-024-46356-y.
[24] D.J. Diaz, C. Gong, J. Ouyang-Zhang, J.M. Loy, J. Wells, D. Yang, A.D. Ellington, A. G. Dimakis, A.R. Klivans, Stability Oracle: a structure-based graph-transformer framework for identifying stabilizing mutations, Nat. Commun. 15 (1) (2024) 6170, https://doi.org/10.1038/s41467-024-49780-2.
[25] C. Gong, A. Klivans, J.M. Loy, T. Chen, D.J. Diaz, Evolution-inspired loss functions for protein representation learning, in: Forty-First International Conference on Machine Learning, 2024.
[26] M. Kaishima, J. Ishii, T. Matsuno, N. Fukuda, A. Kondo, Expression of varied GFPs in Saccharomyces cerevisiae: codon optimization yields stronger than expected expression and fluorescence intensity, Sci. Rep. 6 (2016) 35932, https://doi.org/10.1038/srep35932.
[27] G.A. Diaz, M. Rong, W.T. McAllister, R.K. Durbin, The stability of abortively cycling T7 RNA polymerase complexes depends upon template conformation, Biochemistry 35 (33) (1996) 10837–10843, https://doi.org/10.1021/bi960488+.
[28] C.T. Martin, D.K. Muller, J.E. Coleman, Processivity in early stages of transcription by T7 RNA polymerase, Biochemistry 27 (11) (1988) 3966–3974, https://doi.org/10.1021/bi00411a012.
[29] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S.A. A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature 596 (7873) (2021) 583–589, https://doi.org/10.1038/s41586-021-03819-2.
[30] E.H. Kellogg, A. Leaver-Fay, D. Baker, Role of conformational sampling in computing mutation-induced changes in protein structure and stability, Proteins 79 (3) (2011) 830–838, https://doi.org/10.1002/prot.22921.
[31] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, L. Serrano, The FoldX web server: an online force field, Nucleic Acids Res. 33 (Web Server issue) (2005) W382–W388, https://doi.org/10.1093/nar/gki387.
[32] K.K. Yang, Z. Wu, F.H. Arnold, Machine-learning-guided directed evolution for protein engineering, Nat. Methods 16 (8) (2019) 687–694, https://doi.org/10.1038/s41592-019-0496-6.
[33] R. Corral-Corral, J.A. Beltran, C.A. Brizuela, G. Del Rio, Systematic identification of machine-learning models aimed to classify critical residues for protein function from protein structure, Molecules 22 (10) (2017), https://doi.org/10.3390/molecules22101673.
[34] H. Lu, D.J. Diaz, N.J. Czarnecki, C. Zhu, W. Kim, R. Shroff, D.J. Acosta, B. R. Alexander, H.O. Cole, Y. Zhang, N.A. Lynd, A.D. Ellington, H.S. Alper, Machine learning-aided engineering of hydrolases for PET depolymerization, Nature 604 (7907) (2022) 662–667, https://doi.org/10.1038/s41586-022-04599-z.
[35] I. Paik, P.H.T. Ngo, R. Shroff, D.J. Diaz, A.C. Maranhao, D.J.F. Walker, S. Bhadra, A.D. Ellington, Improved Bst DNA polymerase variants derived via a machine learning approach, Biochemistry 62 (2) (2023) 410–418, https://doi.org/10.1021/acs.biochem.1c00451.
[36] P.A. Osumi-Davis, N. Sreerama, D.B. Volkin, C.R. Middaugh, R.W. Woody, A. Y. Woody, Bacteriophage T7 RNA polymerase and its active-site mutants. Kinetic, spectroscopic and calorimetric characterization, J. Mol. Biol. 237 (1) (1994) 5–19, https://doi.org/10.1006/jmbi.1994.1205.
[37] R. Sousa, R. Padilla, A mutant T7 RNA polymerase as a DNA polymerase, EMBO J. 14 (18) (1995) 4609–4621, https://doi.org/10.1002/j.1460-2075.1995.tb00140.x.
[38] A. Ramanathan, G.B. Robb, S.H. Chan, mRNA capping: biological functions and applications, Nucleic Acids Res. 44 (16) (2016) 7511–7526, https://doi.org/10.1093/nar/gkw551.
[39] S.H. Chan, C.N. Mole, D. Nye, L. Mitchell, N. Dai, J. Buss, D.W. Kneller, J. M. Whipple, G.B. Robb, Biochemical characterization of mRNA capping enzyme from Faustovirus, RNA 29 (11) (2023) 1803–1817, https://doi.org/10.1261/rna.079738.123.
[40] N. Pokala, T.M. Handel, Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation, Protein Sci. 13 (4) (2004) 925–936, https://doi.org/10.1110/ps.03486104.
[41] C. Savojardo, M. Manfredi, P.L. Martelli, R. Casadio, Solvent accessibility of residues undergoing pathogenic variations in humans: from protein structures to protein sequences, Front. Mol. Biosci. 7 (2020) 626363, https://doi.org/10.3389/fmolb.2020.626363.
[42] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A.J. Ballard, J. Bambrick, S.W. Bodenstein, D.A. Evans, C.C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Zemgulyte, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A.I. Cowen-Rivers, A. Cowie, M. Figurnov, F.B. Fuchs, H. Gladman, R. Jain, Y.A. Khan, C.M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E.D. Zhong, M. Zielinski, A. Zidek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J.M. Jumper, Accurate structure prediction of biomolecular interactions with AlphaFold 3, Nature 630 (8016) (2024) 493–500, https://doi.org/10.1038/s41586-024-07487-w.
[43] M.P. Lacerda, E.J. Oh, C. Eckert, The model system Saccharomyces cerevisiae versus emerging non-model yeasts for the production of biofuels, Life (Basel) 10 (11) (2020), https://doi.org/10.3390/life10110299.
[44] X. Yi, H.S. Alper, Considering strain variation and non-type strains for yeast metabolic engineering applications, Life (Basel) 12 (4) (2022), https://doi.org/10.3390/life12040510.
[45] F. Lienert, J.J. Lohmueller, A. Garg, P.A. Silver, Synthetic biology in mammalian cells: next generation research tools and therapeutics, Nat. Rev. Mol. Cell Biol. 15 (2) (2014) 95–107, https://doi.org/10.1038/nrm3738.
[46] R. Mehrotra, K. Renganaath, H. Kanodia, G.J. Loake, S. Mehrotra, Towards combinatorial transcriptional engineering, Biotechnol. Adv. 35 (3) (2017) 390–405, https://doi.org/10.1016/j.biotechadv.2017.03.006.
[47] M. Recktenwald, E. Hutt, L. Davis, J. MacAulay, N.M. Daringer, P.A. Galie, M. M. Staehle, S.L. Vega, Engineering transcriptional regulation for cell-based therapies, SLAS Technol. 29 (2) (2024) 100121, https://doi.org/10.1016/j.slast.2024.100121.
[48] J.G. Perez, J.C. Stark, M.C. Jewett, Cell-free synthetic biology: engineering beyond the cell, Cold Spring Harb. Perspect. Biol. 8 (12) (2016), https://doi.org/10.1101/cshperspect.a023853.