Machine learning-based promoter strength prediction derived from a fine-tuned synthetic promoter library in *Escherichia coli*

Mei Zhao[1,2,#], Shenghu Zhou[1,2,#], Longtao Wu[3], Yu Deng[1,2*]

[1] National Engineering Laboratory for Cereal Fermentation Technology (NELCF), Jiangnan University, 1800 Lihu Road, Wuxi, Jiangsu 214122, China

[2] Jiangsu Provincial Research Center for Bioactive Product Processing Technology, Jiangnan University, 1800 Lihu Road, Wuxi, Jiangsu 214122, China

[3] College of Physics and Optoelectronics, Taiyuan University of Technology, Taiyuan 030024, China

[*] Correspondence to:

Yu Deng: National Engineering Laboratory for Cereal Fermentation Technology (NELCF), Jiangnan University, 1800 Lihu Road, Wuxi, Jiangsu 214122, China.

Phone: +86-510-85329031, Fax: +86-510-85918309

E-mail: dengyu@jiangnan.edu.cn

[#]: Mei Zhao and Shenghu Zhou contributed equally.

Abstract:

Promoters are one of the most critical regulatory elements controlling metabolic pathways. However, in recent years, researchers have simply perfected promoter strength, but ignored the relationship between the internal sequences and promoter strength. In this context, we constructed and characterized a mutant promoter library of P_{trc} through dozens of mutation-construction-screening-characterization engineering cycles. After excluding invalid mutation sites, we established a synthetic promoter library, which consisted of 3665 different variants, displaying an intensity range of more than two orders of magnitude. The strongest variant was 1.52-fold stronger than a 1 mM isopropyl-β-D-thiogalactoside driven P_{T7} promoter. Our synthetic promoter library exhibited superior applicability when expressing different reporters, in both plasmids and the genome. Different machine learning models were built and optimized to explore relationships between the promoter sequences and transcriptional strength. Finally, our XgBoost model exhibited optimal performance, and we utilized this approach to precisely predict the strength of artificially designed promoter sequences. Our work provides a powerful platform that enables the predictable tuning of promoters to achieve the optimal transcriptional strength.

Keywords: Promoter, Fine-tune, Machine learning, XgBoost model

Introduction

The application of synthetic biology and metabolic engineering depends on essential biological regulatory elements, such as promoters[1, 2], ribosome binding sites (RBS)[3, 4] and terminators [5]. These important elements make genetic circuits more tunable, inducible, responsive, and/or coordinated[6, 7, 8]. Basic levels of transcriptional regulation occur at promoters to ensure natural and synthetic circuits or metabolic pathways[9, 10]. To increase regulatory gene expression efficiency, several promoters with gradient strengths were built and modified by optimizing important genetic elements, such as -35/-10 boxes, 5′-untranslated regions and transcription factor binding sites [11, 12, 13]. However, due to weak promoter strengths, low dynamic ranges (the highest strength/the lowest strength), limited library promoters, and inducers required, these promoters are often incapable of fine-tuning metabolic pathways. Thus, it is important to establish and characterize a comprehensive library consisting of hundreds of promoters, with continuous and broad dynamic ranges.

To overcome these limitations, Mey *et al.* constructed a synthetic promoter library with 75 variants using degenerated oligonucleotide primers, comprising a 57 bp length sequence of 20 random, 13 semi-conserved, and 24 conserved nucleotides [14]. Although the strength of this library was 0.14- to 275-fold that of the *Escherichia coli* constitutive promoter P_{LacI}, the library was small, and the strongest promoter was far lower than the commonly used P_{T7} promoter. To further extend library size, Zhou *et al.* [15] and Yang *et al.* [16] screened a hundred native promoters from *E. coli* and *Bacillus subtilis.* The transcriptional intensity ranged from 0.007%–4630% that of the P_{BAD} promoter, and 0.03–2.03-fold that of the P_{43} promoter at the transcriptional level. However, the strength of these promoters was still not comparative, or far lower than other well-studied promoters, such as P_{43} [17], P_{Veg} [17], P_{T7}, P_{trc} [18], and P_{Thl} [19].

The mutation, modification, or screening of existing promoters is difficult to obtain the desired ones, thus the *de novo* design of optimized promoters from sequences is a promising approach. To do this, the relationship between the promoter sequences and intensity should be established. However, few reports have contributed to this area. For example, Jensen *et al.* [13] revealed a simple statistical method to explore nucleotide positions, which exerted critical effect on promoter intensity in 69 $P_{L\text{-}\lambda}$ promoter variants in *E. coli*. Likewise, Mey *et al.* [14] and Liu *et al.* [17] observed similar results by analyzing a partial least squares (PLS) model in *E. coli* and *Bacillus subtilis* using 49 and 214 synthetic promoters, respectively. However, these reports suffered small data issues, single modeling, imperfect correlations and low dynamic ranges. Therefore, it is important to identify promoters with gradient strengths, broad dynamic ranges, and clear sequence profiles to explore and analyze relationships between promoter sequences and intensity, using huge data and model comparisons. Nowadays, significant advances have been made in machine and deep learning for big data analytics [2, 20], making promoter strength prediction achievable. In particular, Gradient Boosting Decision Tree (GDBT)[21], AdaBoost[22], Random Forest Regressor[23], Xgboost[24], and Recurrent Neural Network[25] by one-hot coding have provided efficient mechanisms to analyze big data in designing functional promoters.

It is well known that the core promoter (−35 box, spacer, and −10 box) and their flanking regions (down and up elements) of bacterial promoters are closely related with promoter intensity [26, 27]. To generate high strength constitutive promoters and broad dynamic range libraries, we randomized P_{trc}[18] regions by mutation-construction-screening-characterization (MCSC) engineering cycles (Fig. 1a). In doing so, P_{trc} variants we abolished its inherent limitations and derived a series of extremely high strength constitutive promoters. Based on this mutation library, we reconstructed and

characterized a *de novo* synthetic promoter library which can be used as a comprehensive synthetic biology toolbox for gene expression regulation over a broad range (Fig. 1b-d). Furthermore, we investigated a direct correlation between promoter base sequences and transcription strength of the synthetic promoter library, using machine learning models (Fig. 1e). Taken together, this work not only provided the constitutive promoters, whose strengths span from low to extremely high, but also established a promoter strength prediction model which could significantly reduce promoter characterization workload.

Results

Generation of the mutant library

We chose the P_{trc} promoter (74 bp in length) [18] as a template from the pTrc99a plasmid to obtain a mutant constitutive promoter library by *error-prone* PCR in this work. To extend span strength of the mutation library, iterative MCSC engineering cycles (Fig. 1a) was performed to obtain both strength enhanced and reduced promoters. After each round of mutagenesis, we counted the minimum and maximum fluorescence/OD_{600} and dynamic range. The minimum fluorescence/OD_{600} was relatively stable, remaining between 85–200, while the maximum fluorescence/OD_{600} showed an obviously increasing trend, and approached the highest after 40 rounds of MCSC engineering (Fig. S1 a-b). The dynamic range reached the maximum value after 82 rounds of MCSC engineering (Fig. S1 c). After each round of MCSC engineering, we selected the mutants whose fluorescence intensities changed more than 10 times (increased or decreased) as the templates for the next round of MCSC engineering. Finally, we obtained ~6×10^4 constructs with an intensity range ~509-fold difference between the strongest and weakest expression, and the strongest one had the expression

115 levels ~127- and 3.14-fold higher than the uninduced and induced P_{trc} (1 mM isopropyl-
116 β-D-thiogalactoside (IPTG) induction), respectively (Fig. 2a). Therefore, this mutant
117 library demonstrated the broad expression levels and excellent resolution.
118
119 **Sequence variances of the mutant P_{trc}-derived library**
120 After constructing the mutant P_{trc}-derived library, we used the high-throughput
121 sequencing (MiSeq) to understand the sequence variances of the mutant promoters (Fig.
122 1b) (SRA database # SRR11574455, https://www.ncbi.nlm.nih.gov/sra/SRR11574455).
123 The MiSeq results show that the mutations were mainly located in the core promoter
124 and the down element, indicating that the up element only contributes a little to the
125 promoter strength (Fig. 2b). Furthermore, a total of 66026 different mutants were found
126 in the library. In statistical analysis, 32.5% of the mutants contained mutations in both
127 the core promoter and down element region, 63.06% of the mutants contained
128 mutations only in the down element region, and 0.28% of the mutants contained
129 mutations only in the up element region. Therefore, we focused on both the core
130 promoter and down element to exclude the invalid mutations in the up element and
131 reconstruct a simpler synthetic promoter library (3665 out of 66026 mutant promoters).
132 The 3665 different P_{trc}-derived mutant promoters (74-bp covering core promoter and
133 down element) [28] were synthesized to form a synthetic promoter library (see Additional
134 file 2). The 3665 different promoter sequences had 1067 mutations in −35 and/or −10
135 boxes, 1313 mutations in spacers, and 3581 mutations in down elements, including 9
136 additions, 513 deletions, and 3143 substitutions. The diverse mutations caused the
137 diversity of promoter changes, which provides a guarantee for the evolution of
138 promoters.
139

Characterization of the synthetic promoter library

After high-throughput sequencing and reconstruction, the fluorescence intensity of all reconstructs in the synthetic promoter library were measured and analyzed. Three methods were adopted and evaluated for the synthetic promoter library. First, to standardize our synthetic promoter library, all synthetic promoters were compared with the induced P_{trc} and P_{T7} promoters in *E. coli*. To do this, P_{T7} promoter was used to replace P_{Trc} promoter on the backbone of vector pL0-sfGFP (Fig. 3a). To focus on comparing the strength of the promoter, all constructs were performed on the same backbone and introduced into *E. coli*. Expressions and comparison between fluorescent intensity of the synthetic promoters and the known promoters are shown in Fig. 3b. The maximum strength of our synthetic promoter library was ~114-fold that of uninduced P_{trc}, ~1.52-fold that of 1 mM IPTG induced P_{T7}, and ~2.83-fold that of 1 mM IPTG induced P_{trc}, with a ~454-fold difference between the strongest and weakest expressions.

To analyze the universality of the synthetic promoter library when expressing different genes, a set of synthetic promoters (selected based on multiples of control fluorescence intensity) [29] were chosen and used to control the expression of β-galactosidase (*lacZ*) and lactate dehydrogenase (*ldhA*). According to the fluorescence intensity data of the synthetic promoter library, 21 promoters (PL3153, PL757, PL2776, PL862, PL3034, PL2346, PL3293, PL2983, PL1078, PL3224, PL2169, PL1958, PL1260, PL3088, PL1456, PL2666, PL3227, PL948, PL3001, PL2986, and PL3147) with fluorescence intensity multiple that of the P_{trc} (0.5–100 times) in the synthetic promoter library were selected and used to express *lacZ* instead of sf*gfp* in the plasmid (Fig. S2a and Additional file 2). After the recombinant strains were constructed, the β-galactosidase activity was measured. The same trend was observed in the β-galactosidase activities as for the sfGFP expression driven by the same promoters (Fig.

S2b). The enzyme activity of *lacZ* correlated well with the fluorescence/OD_{600} (R^2=0.94) (Fig. 3c). The specific β-galactosidase activities spanned a ~188- and ~60-fold range relative to the Mlac0 (MG1655ΔlacZ) and Mlac1 (pTrc-lacZ), respectively. The corresponding fluorescence intensity spanned a ~184- and ~71-fold range relative to the Mlac0 (MG1655ΔlacZ) and ML1 (pTrc-sfGFP), respectively. Thus, to find the most suitable promoter for expression, there is only the need to choose different folds of promoters in the synthetic library according to fluorescence intensity. Furthermore, we also found that the transcriptional ($2^{-\Delta\Delta Ct}$) and translation level of β-galactosidase were highly correlated (Fig. 3d and Fig. S2c). The relative levels of the *lacZ* transcripts spanned a 13-fold range under the control of the selected promoters.

To test the effect of the synthetic promoters on the genome, we modified the lactate dehydrogenase gene (*ldhA*) in *E. coli* through replacing the native *ldhA* promoter by the synthetic promoters (Fig. S3a). We deleted the wild type *ldhA* promoter in *E. coli* MG1655 forming the strain Mldh0 as control. Five different promoters, PL1409, PL908, PL2436, PL3189, and PL1993, that were 49.52%, 110.61%, 556.41%, 4086.54%, and 9826.74% the strength of P_{ldhA}, respectively (Fig. S3b), were selected to substitute the P_{ldhA} in the chromosome of strain Mldh0 using CRISPR-Cas9, resulting in strains Mldh1409, Mldh2436, Mldh908, Mldh3189, and Mldh1993. The fluorescence strengths of PL2436 on the genome were almost the same as P_{ldhA}, indicating the expression level of lactate dehydrogenase was similar (Fig. S3b-c). The lactate dehydrogenase activity of MG1655, Mldh1409, Mldh2436, Mldh908, Mldh3189, and Mldh1993 was 1.29, 0.96, 1.33, 2.44, 3.16, and 3.63-fold of the control (Mldh0), respectively (Fig. S3c). As shown in Fig. S3d, the transcription levels of *ldhA* ($2^{-\Delta\Delta Ct}$) had a similar trend with fluorescence intensity and enzyme activities. Taken together, the fluorescence intensity of the same synthetic promoter candidates was tested, and the

activities of lactate dehydrogenase and transcription levels of *ldhA* showed similar results (Fig. S3b-d). These results provided robust, strain-independent, gene-independent regulation.

Machine learning-based rational design of promoters

Although a synthetic promoter library was obtained, it was extremely crucial to determine the functional relationships between promoter sequences and the strength to achieve the rational design of the desired promoter. PLS models [11, 13, 17] with multinomial statistics are often used for exploring nucleotide positions that have a significant effect on promoter intensity. After the promoter sequence is encoded, PLS compares the basic relationship between the two sequences. It tries to find the multi-dimensional direction of each promoter sequence to explain the multi-dimensional direction, with the largest intensity sequence variance. As such, we first tried the PLS model to predict promoter strength according to the given sequences. The data-set was randomly split into two parts: the training set, including 90% of the 3665 synthetic promoters, and the test set, including the remaining promoters [11]. The training set was utilized to construct the PLS model. In this procedure, 74 latent variables were determined as previously described [30] and retained in the PLS model (R^2=0.66) (Fig. 4a). The predicted and origin corresponded to the predicted promoter strength ($\log_{10}$(fluorescence/OD_{600})) by model and the observed promoter strength ($\log_{10}$(fluorescence/OD_{600})), respectively.

Due to the differences in mutant methods and huge databases, the final PLS training result was not perfect. Therefore, we started to try other machine and deep learning models, we chose the models Gradient Boosting Decision Tree (GDBT) [21], AdaBoost [22], Random Forest Regressor [23], XgBoost [24], and Recurrent Neural Network

[25] for research using the scikit-learn Python package. All models were modeled on 90% of the data, and then the predictions of this model were tested against mean were trained to eliminate the problems of multicollinearity, and efficiently verified using 10-fold cross-validation (i.e. by training surements for the remaining 10%) with the performance measured as the mean absolute error (MAE) of relative strength (cross_val_score and RepeatedKFold from the scikit-learn Python package, https://scikit-learn.org/stable/). Finally, the XgBoost was identified as the best model with the lowest average cross-validated MAE (test MAE) (0.204) and highest R^2 (0.77) for promoter prediction (Fig. 4a). Therefore, by applying XgBoost, an excellent relationship was found between promoter sequences and intensity. This XgBoost model can be a useful tool to rationally design a functional promoter to fine-tune gene expression. In this model, once the promoter sequence was input, the promoter strength was obtained.

XgBoost model verification

To further verify the performances of the established XgBoost model, we rationally designed a new promoter library. To do this, the mutation possibility at each position of strong and weak promoters (the P_{trc} promoter as an interval between strong and weak ones) in the synthetic promoter library was analyzed by a statistical analysis approach [25,27]. The positions 25, 28, 32, 41, 43, 46, 51, and 54 were identified as the most critical bases at which mutations could significantly influence promoter strength (Fig. 4b). In this regard, we randomized these eight sites (such as A) to different bases (*i.e.* C, T, G, or B), leaving the other 66 sites unchanged to form a new promoter library, with a size of 390625 (5^8 possible sequences) (Additional file 2). Predictive models of relative strength for the newly designed promoter assembly library were constructed, following a similar strategy based on one-hot encoding of features and the best model

regression and cross-validation way. To accumulate enough data to train an independent model, 100 promoters (Additional file 2) were randomly selected from the new promoter assembly library, and the promoter sequence was used as an additional input feature. In addition, 27 well-studied promoters (Additional file 2) [11, 12, 31, 32,29, 33] were entered into the model to test the universality. The intensity of the 127 promoters formed the test set and were predicted by the fully trained XgBoost model. The predicted intensity versus the observed intensity (fluorescence intensity) were shown in Fig. 4c. In the test set, the model generated good results. We evaluated model performance through the R criterion and the MAE (R^2 = 0.88, MAE=0.148). These results thus indicated a satisfactory correlation of promoter strength to the sequence. Based on these validation experiments, the established XgBoost model not only held a huge numbers of experimentally verified data, but also provided a robust predictive function for the expression levels of 127 unique expression vectors in numerous conditions. Therefore, we established a relationship between promoter sequence and strength in *E. coli*, an unprecedented discovery.

Discussion

Generally, gradient strength promoters are essential elements for pathway fine-tuning [9, 10]. However, existing promoters suffer with low strength, narrow strength span and limited numbers. Simply engineering or screening natural promoters is not only laborious, but also difficult to identify high strength promoters. To overcome these challenges, we iteratively evolved P_{trc} promoters based on MCSC engineering cycles, and reconstructed a synthetic promoter library, after MiSeq sequencing and mutation analysis. This synthetic promoter library consisted of 3665 gradient strength promoters; the strongest promoter was 1.52-fold the strength of a 1 mM IPTG induced P_{T7}

promoter. Using the synthetic promoter library as an input dataset, we built and optimized a series of promoter strength prediction models. In comparing models, the XgBoost model performed the optimally (R^2 = 0.77, MAE = 0.205). To further verify XgBoost model reliability, we compared the predicted and actual strength of a hundred rationally designed artificial promoters (R^2= 0.88). Taken together, we rationally designed and provided a powerful platform to enable predictable promoter tuning to transcriptional strength.

Although many advancements have been made in past decades, the strength and dynamic range of promoters were relatively low and narrow, respectively[17, 18]. In addition, promoter characterization in the literature is often performed using different genetic backgrounds and testing conditions, resulting in unquantifiable performances when applying them to the same host. Hence, several studies have screened hundreds of gradient strength constitutive promoters from *E. coli* [15], *B. subtilis* [16], *C. glutamicum* [34], and *S. cerevisiae* [35]. However, the strength of these constitutive promoters is still far lower than what is required for high expression levels, especially for protein over-overexpression. Single rounds of screening from mutation libraries are difficult to obtain for extremely high strength promoters[36]. Hence, the iterative generation of high strength promoters is a promising strategy to extend promoter strength. In this regard, the P_{trc} promoter was evolved by several MCSC engineering cycles, and a series of extremely high strength promoters were screened.

We further analyzed the mutation library and found that mutation sites were mainly distributed in the core promoter (−35 box, spacer, −10 box) and down element. The core promoter is known to have a great influence on gene expression [37, 38, 39, 40, 41, 42, 43, 44]. In-depth research found that changes in up elements [45, 46, 47, 48] and down elements [49] also had a significant contribution to gene expression, but the contribution was not as

great as that of the core promoter. However, there were only a few up element changes. To facilitate the study, we only explored the core promoter and down elements of mutant promoters. We then constructed a synthetic promoter library with 3665 mutant promoter constructs, which exhibited a 454.26-fold difference between the strongest and weakest expression. The strength of the synthetic promoters was much higher than other reported promoters and it was not necessary to tandem multiple promoters to achieve a higher intensity [17, 50].

Although the P_{trc} promoter was generally considered as a strong inducible promoter, it worked well in the absence of an inducer [51]. Thus, it could somehow work as a typical constitutive promoter. Originally, we searched for native constitutive promoters as the candidates for directed evolution. However, none of them was comparable with the P_{Trc} promoter in strength. In other words, the mutant native constitutive promoter might not meet our requirements of directed evolution. In this regard, we selected P_{trc} as the original promoter to establish comprehensive and constitutive promoter libraries. These libraries exhibited great stability in the expression of different reporter genes in both plasmids and the genome.

Previously, it was almost impossible to rationally predict promoter strength directly based on sequences. An ideal model was based on the predicted thermodynamics to predict the strength of the promoters. However, the thermodynamic model was too ideal to understand the promoters precisely and most of the time, the predicted promoters generated from the above models were far from the experimental results. Machine or deep learning models were independent of the “mechanisms” and thus provided a promising approach to predict promoter strength, without fully understanding mechanisms. Recently, Wang *et al*. successfully established a complicated AI model that could be used to rationally design and predict promoters[2].

315 The model was based on the training of natural promoters which usually had moderate
316 strength. Although 70.8% of promoters exhibited activity, their strength was generally
317 low. In our study, the training of the XgBoost model was based on the high strength P_{trc}
318 promoter library. Hence, our model predicts the high strength promoters. Furthermore,
319 we found that the fluorescence intensity deviated from the trend line when
320 $\log_{10}^{(sfGFP/OD600)}$ lower than 2.5, which represents the low fluorescence intensity was not
321 well worked for the model. With the development of AI, we believe that models that
322 can precisely design desired promoters and predict the strength of a given promoter can
323 be established in the near future.
324
325 **Methods**
326 **Bacterial strains and cultivation**
327 *E. coli* JM109 was used for plasmid cloning and MG1655 K12, MG1655 (DE3)
328 were used for gene expression. Inoculates were cultured in 50 mL Luria Bertani (LB)
329 in 250 mL shaker flasks at 200 rpm. Assay strains were inoculated (10% working
330 volume) into M9 minimal medium [52], including 5 g/L D-glucose (M9G) and 0.1%
331 amino acids[12, 31] for the determination of fluorescence expression intensity. Gene
332 expression was induced initially by 1 mM IPTG or no inducer [33]. All other strains were
333 cultured at 37°C in LB medium, supplemented with 100 mg/mL ampicillin. All strains
334 and plasmids are listed in Table S1. All primers are listed in Table S2.
335 **Plasmid construction**
336 The sfGFP ORF was amplified from the pJKR-H plasmid [53] with pL1-sfGFP F
337 and pL1-sfGFP R primers and ligated into *Eco*R I/*Hin*d III sites of pTrc99a, resulting
338 in pL1-sfGFP. The negative control and backbone pL0-sfGFP (no promoter) was
339 constructed by whole plasmid PCR [54] from pL1-sfGFP, using the pL0-sfGFP F and R.

Products were digested by CIAP and ligated using DNA T4 ligase. P_{T7} promoter was amplified from pACYC-Duet-1 plasmid by primer pair of T7-F/T7-R. pT7-sfGFP plasmid was generated by whole plasmid PCR [54], as previously described.

Random mutagenesis, to generate novel promoters, was conducted by *error-prone* PCR [55] with Taq DNA polymerase in the existence of Mn^{2+}, $Mg^{2+,}$ and dNTP, using plasmid pTrc99a (Novagen, CA, USA) as the template along with the er-Trc F and er-Trc R. The primers mentioned above were used to perform 30 amplification cycles. The standard reaction conditions were as follows: 200 μl reaction volume; 10 pM each primer; 0.0625~3 mM $MnCl_2$ or 0.5~12 mM $MgCl_2$ or different ratios of 100 mM dNTP mixture; 2×Taq DNA polymerase. The cycle profile was: 1 min 94°C, 2 min 59°C, and 3 min 72°C. Then the SanPrep Column PCR Product Purification Kit (Sangon Biotech, Shanghai, China) was used to purify PCR products. The backbone was linearized by whole plasmid PCR with T0-sfGFP F and R, using plasmid pL1-sfGFP as the template. Following purification and digestion with *Dpn* I, the insert and backbone were assembled using Gibson method [28] and transformed into *E.coli* JM109. After colony PCR, these right recombinant plasmids were transferred into MG1655 and the fluorescence intensity was detected. The other constructions in this study also used the Gibson assembly method, as described above.

About ~6×10^4 colonies were visually screened from agar plates. A single colony from each plate was picked into M9G for fluorescence detection. After MiSeq, the reconstructed 3665 synthetic promoters were named PLN, forming the synthetic promoter library. They were transformed into MG1655 and called MLN. The promoters for predicting the XgBoost model were named pZN, transformed into MG1655, and called MZN. The synthetic promoters carrying lacZ were called pLacN, transformed into MG1655, and named MlacN. The different synthetic promoters replacing the ldhA

365 promoter in MG1655 were called MldhN.

366

367 **MiSeq sequencing**

368 Selected mutants were sequenced using primers Miseq F and Miseq R. A total of
369 ~6×10^4 single colonies obtained by MCSC were mixed and plasmids were extracted to
370 form a mixture sample. Samples were MiSeq sequenced by Sangon Biotech (Shanghai,
371 China). The original MiSeq data was submitted to the SRA database, under accession
372 number SRR11574455 (https://www.ncbi.nlm.nih.gov/sra/SRR11574455).

373

374 **Library screening using the sfGFP reporter assay**

375 Single colonies on agar plates were inoculated into 96-well plates including 200
376 μL LB. After 8–12 h, the inoculums were inoculated (2% working volume) into 180 μL
377 M9G. The cultures were grown at 37°C with 300 rpm. After 4–6 h, the fluorescence
378 and optical density were monitored on a plate reader (Tecan) at when OD_{600} reached
379 0.4–0.6. A 100 μL sample was transferred to a black 96-well plate, and the sfGFP
380 fluorescence was measured at 485 nm after excitation at 528 nm using a plate reader
381 (Tecan). Fluorescence was measured in arbitrary units (AFU) while optical density was
382 determined by absorbance (OD) at 600 nm. The intensity of sfGFP was characterized
383 and calculated by sfGFP fluorescence/OD_{600}. The negative controls are MG1655 K12
384 and ML0 (MG1655 carrying pL0-sfGFP).

385

386 **Genome manipulation**

387 The *lacZ* and promoter *ldhA* were knocked out separately in *E. coli* MG1655 by
388 CRISPR-Cas9 approach [56]. Gene insertion was a similar step to the knockout procedure
389 and the template introduced the fragment to be inserted. The template, which included

the upstream 500 bp, PL908, and downstream 500 bp, was obtained from pldh908. The rest of the procedure followed the same steps as with the knockout manipulation. The synthetic promoters PL1409, PL1993, PL2436, and PL3189 were inserted in *E. coli* MG1655 using the same method. The related sgRNA was designed and shown in Supplementary Table S3.

Analysis of transcriptional intensity

Total RNA was extracted using the Ultrapure RNA Kit (Novoprotein, Shanghai, China) and reverse transcribed using the SuperRT One-Step RT-PCR Kit (Novoprotein, Beijing, China). Real time quantitative PCR (qPCR) using a SuperRT One-Step RT-PCR Kit (Novoprotein, Beijing, China) was performed and analyzed according to the protocol [57, 58].

Enzyme activity assays

Cell crude extracts were obtained and analyzed to measure enzyme activity. β-galactosidase measurements were performed as described by Miller et al. [59, 60]. The lactate dehydrogenase assay was conducted according to previously published methods [61, 62]. Protein concentrations were performed using the Bradford method [63].

Model construction and prediction

The training set was generated by experimental parameters and calculated relative strengths. The training set contained 90% of the data, and testing predictions against measurements for the remaining 10%. Values in the training set were encoded into one-hot binary vectors. One-hot coding makes the discrete features continuous, allowing the model to optimally process data. On the other hand, through the representation of one-dimensional vectors, the purpose of expanding features is achieved, to a certain extent,

the features can be sparse to prevent overfitting. Each input sequence was encoded into an eigenvector of length 74 bp with this method. The models were trained (Gradient Boosting Decision Tree (GDBT)[21], AdaBoost[22], Random Forest Regressor[23], Xgboost[24], Recurrent Neural Network[25]) to eliminate multicollinearity interference, and were cross-validated by 10-fold cross-validation with the performance indicators as the MAE of relative strength (cross_val_score and RepeatedKFold from the scikit-learn Python package). Using this data processing method for coding promoters, we established a promoter sequence predictive model based on each algorithm, to predict promoter strength using the same strategy. This code can be found at https://github.com/YuDengLAB/Predictive-the-correlation-between-promoter-base-and-intensity-through-models-comparing.

Data availability

All experimental data were determined in triplicate, and error bars represent the standard deviation. The original MiSeq data has been submitted to the SRA database under accession number SRR11574455 (https://www.ncbi.nlm.nih.gov/sra/SRR11574455). The code to predict the correlation between promoter base and intensity through comparing models can be found at https://github.com/YuDengLAB/Predictive-the-correlation-between-promoter-base-and-intensity-through-models-comparing.

Reference

1. Meng H, Ma Y, Mai G, Wang Y, Liu C. Construction of precise support vector machine based models for predicting promoter strength. *Quantitative Biology* **5**, 90-98 (2017).

2. Wang Y, Wang H, Wei L, Li S, Liu L, Wang X. Synthetic promoter design in Escherichia coli based on a deep generative network. *Nucleic Acids Res*, (2020).

3. Bo Z*, et al.* Ribosome binding site libraries and pathway modules for shikimic acid synthesis with Corynebacterium glutamicum. *Microbial Cell Factories,14,1(2015-05-17)* **14**, 1-14 (2015).

4. Markley AL, Begemann MB, Clarke RE, Gordon GC, Pfleger BF. Synthetic biology toolbox for controlling gene expression in the cyanobacterium Synechococcus sp. strain PCC 7002. *Acs Synthetic Biology* **4**, 595 (2015).

5. Curran KA, Karim AS, Gupta A, Alper HS. Use of expression-enhancing terminators in Saccharomyces cerevisiae to increase mRNA half-life and improve gene expression control for metabolic engineering applications. *Metab Eng* **19**, 88-97 (2013).

6. Suong TW, Matthew Wook C. Development and characterization of AND-gate dynamic controllers with a modular synthetic GAL1 core promoter in Saccharomyces cerevisiae. *Biotechnology & Bioengineering* **111**, 144-151 (2013).

7. Ning J. (423b) Coordinated Induction of Multi-Gene Pathways in Saccharomyces Cerevisiae.

8. Afonso B, Silver PA, Ajofranklin CM. A synthetic circuit for selectively arresting daughter cells to create aging populations. *Nucleic Acids Research,38,8(2010-01-05)* **38**, 2727-2735 (2010).

9. Keaveney M, Struhl K. Activator-Mediated Recruitment of the RNA Polymerase II Machinery Is the Predominant Mechanism for Transcriptional Activation in Yeast. *Molecular Cell* **1**, 917-924 (1998).

10. Tirosh I, Barkai N, Verstrepen KJ. Promoter architecture and the evolvability of gene expression. *J Biol* **8**, 95-95 (2009).

11. De Mey M, Maertens J, Lequeux GJ, Soetaert WK, Vandamme EJ. Construction and model-based analysis of a promoter library for E. coli: an indispensable tool for metabolic engineering. *BMC Biotechnology* **7**, 34 (2007).

12. Hal A, Curt F, Elke N, Gregory S. Tuning genetic control through promoter engineering. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 12678-12683 (2005).

13. Jensen K, Alper H, Fischer C, Stephanopoulos G. Identifying Functionally Important Mutations from Phenotypically Diverse Sequence Data. *Applied and Environmental Microbiology* **72**, 3696-3701 (2006).

14. Mey MD, Maertens J, Lequeux GJ, Soetaert WK, Vandamme EJ. Construction and model-based analysis of a promoter library for E. coli : an indispensable tool for metabolic engineering. *Bmc Biotechnology* **7**, 34 (2007).

15. Zhou S, Ding R, Chen J, Du G, Li HZ, Zhou J. Obtaining a panel of cascade promoter-5'-UTR complexes in *Escherichia coli*. *ACS Synth Biol* **6**, 1065-1075 (2017).

16. Yang S, Du G, Chen J, Kang Z. Characterization and application of endogenous phase-dependent promoters in *Bacillus subtilis*. *Appl Microbiol Biot*, 1-11 (2017).

17. Liu D*, et al.* Construction, Model-Based Analysis, and Characterization of a Promoter Library for Fine-Tuned Gene Expression in Bacillus subtilis. *ACS Synthetic Biology* **7**, 1785-1797 (2018).

18. Yim SS, An SJ, Kang M, Lee J, Jeong KJ. Isolation of fully synthetic promoters for high‐level gene expression in Corynebacterium glutamicum. *Biotechnology and Bioengineering* **110**, 2959-2969 (2013).

19. Yang G*, et al.* Rapid Generation of Universal Synthetic Promoters for Controlled Gene Expression in Both Gas-Fermenting and Saccharolytic Clostridium Species. *ACS Synthetic Biology* **6**, 1672-1678 (2017).

20. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. *Journal of Big Data* **2**, 1 (2015).

21. Ding C, Wang D, Ma X, Li H. Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees. *Sustainability* **8**, (2016).

22. Zhao X, Ning B, Liu L, Song G. A prediction model of short-term ionospheric foF2 based on AdaBoost. *Advances in Space Research* **53**, 387-394 (2014).

23. Cootes TF, Ionita MC, Lindner C, Sauer P. Robust and Accurate Shape Model Fitting Using Random Forest Regression Voting. In: *Computer Vision – ECCV 2012* (ed^(eds Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C). Springer Berlin Heidelberg (2012).

24. Chen T, Guestrin C. *XGBoost: A Scalable Tree Boosting System* (2016).

25. Keren G, Schuller B. Convolutional RNN: An enhanced model for extracting features from sequential data. In: *2016 International Joint Conference on Neural Networks (IJCNN)* (ed^(eds) (2016).

26. Hammer K, Mijakovic I, Jensen PR. Synthetic promoter libraries--tuning of gene expression. *Trends Biotechnol* **24**, 53-55 (2006).

27. Rytter JV, Helmark S, Chen J, Lezyk MJ, Solem C, Jensen PR. Synthetic promoter libraries for *Corynebacterium glutamicum*. *Appl Microbiol Biotechnol* **98**, 2617-2623 (2014).

28. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343-345 (2009).

29. Jervis AJ*, et al.* SelProm: A Queryable and Predictive Expression Vector Selection Tool for Escherichia coli. *ACS Synthetic Biology* **8**, 1478-1483 (2019).

30. Li B, Morris J, B. Martin E. *Model selection for partial least squares regression* (2002).

31. Kyle J, Hal A, Curt F, Gregory S. Identifying functionally important mutations from phenotypically diverse sequence data. *Applied & Environmental Microbiology* **72**, 3696 (2006).

32. Zhou S, Ding R, Jian C, Du G, Li H, Zhou J. Obtaining a Panel of Cascade Promoter-5′-UTR Complexes in Escherichia coli. *Acs Synthetic Biology* **6**, (2017).

33. Lee TS*, et al.* BglBrick vectors and datasheets: A synthetic biology platform for gene expression. *Journal of Biological Engineering* **5**, 12 (2011).

34. Li N, Zeng W, Xu S, Zhou J. Obtaining a series of native gradient promoter-5′-UTR sequences in Corynebacterium glutamicum ATCC 13032. *Microb Cell Fact* **19**, 120 (2020).

35. Gao S, Zhou H, Zhou J, Chen J. Promoter library based pathway optimization for efficient (2S)-naringenin production from p-coumaric acid in Saccharomyces cerevisiae. *J Agr Food Chem*, (2020).

36. Blazeck J, Alper HS. Promoter engineering: Recent advances in controlling transcription at the most fundamental level. *Biotech J* **8**, 46-58 (2013).

37. Hawley DK, McClure WR. Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic acids research* **11**, 2237-2255 (1983).

38. Youderian P, Bouvier S, Susskind MM. Sequence determinants of promoter activity. *Cell* **30**, 843-853 (1982).

39. Deuschle U, Kammerer W, Gentz R, Bujard H. Promoters of Escherichia coli: a hierarchy of in vivo strength indicates alternate structures. *EMBO J* **5**, 2987-2994 (1986).

40. Kammerer W, Deuschle U, Gentz R, Bujard H. Functional dissection of Escherichia coli promoters: information in the transcribed region is involved in late steps of the overall process. *EMBO J* **5**, 2995-3000 (1986).

41. Mandecki W, Reznikoff WS. A lac promoter with a changed distance between -10 and -35 regions. *Nucleic acids research* **10**, 903-912 (1982).

42. Stefano JE, Gralla JD. Spacer mutations in the lac ps promoter. *Proceedings of the National Academy of Sciences of the United States of America* **79**, 1069-1072 (1982).

43. Russell C. Chronic Pancreatitis. *Surgery (Oxford)* **20**, 231-236 (2002).

44. Aoyama T*, et al.* Essential structure of E. coli promoter effect of spacer length between the two consensus sequences on promoter function. *Nucleic Acids Research* **11**, 5855-5864 (1983).

45. Rhodius VA, Mutalik VK, Gross CA. Predicting the strength of UP-elements and full-length E. coli σE promoters. *Nucleic Acids Research* **40**, 2907 (2012).

46. Ross W, Aiyar SE, Salomon J, Gourse RL. Escherichia coli promoters with UP elements of different strengths: modular structure of bacterial promoters. *Journal of bacteriology* **180**, 5375-5383 (1998).

47. Estrem ST, Gaal T, Ross W, Gourse RL. Identification of an UP element consensus sequence for bacterial promoters. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 9761-9766 (1998).

48. Yan Q, Fong SS. Study of in vitro transcriptional binding effects and noise using constitutive promoters combined with UP element sequences in Escherichia coli. *Journal of biological engineering* **11**, 33-33 (2017).

49. Burke TW, Kadonaga JT. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes & development* **10**, 711-724 (1996).

50. Li M*, et al.* A strategy of gene overexpression based on tandem repetitive promoters in Escherichia coli. *Microbial Cell Factories* **11**, 19 (2012).

51. Balzer S, Kucharova V, Megerle J, Lale R, Brautaset T, Valla S. A comparative analysis of the properties of regulated promoter systems commonly used for recombinant gene expression in Escherichia coli. *Microbial cell factories* **12**, 26 (2013).

52. Denman AM. Molecular Cloning: a Laboratory Manual. *Immunology* **49**, 411 (1983).

53. Rogers JK, Guzman CD, Taylor ND, Srivatsan R, Kelley A, Church GM. Synthetic biosensors for precise gene control and real-time monitoring of metabolites. *Nucleic Acids Research* **43**, 7648-7660 (2015).

54. Miyazaki K. Creating Random Mutagenesis Libraries by Megaprimer PCR of Whole Plasmid (MEGAWHOP). *Methods Mol Biol* **231**, 23-28 (2003).

55. Zhou YH, Zhang XP, Ebright RH. Random mutagenesis of gene-sized DNA molecules by use of PCR with Taq DNA polymerase. *Nucleic Acids Research* **19**, 6052 (1991).

56. Jiang Y, Chen B, Duan C, Sun B, Yang J, Yang S. Multigene editing in the *Escherichia coli* genome via the CRISPR-Cas9 system. *Appl Environ Microbiol* **81**, 2506-2514 (2015).

57. Geyer CN*, et al.* Evaluation of CTX-M steady-state mRNA, mRNA half-life and protein production in various STs of Escherichia coli. *Journal of Antimicrobial Chemotherapy* **71**, 607-616 (2016).

58. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.

59. Israelsen H, Madsen S, Vrang A, Hansen E, Johansen E. *Cloning and partial characterization of regulated promoters from Lactococcus lactis TN917-LacZ integrants with the new promoter probe vector, pAK80* (1995).

60. H. Miller J. *Experiments In Molecular Genetics* (1972).

61. Mat-Jan F, Alam KY, Clark DP. Mutants of Escherichia coli deficient in the fermentative lactate dehydrogenase. *Journal of Bacteriology* **171**, 342-348 (1989).

62. Tarmy EM, Kaplan NO. Chemical Characterization of d-Lactate Dehydrogenase from Escherichia coli B. *Journal of Biological Chemistry* **243**, 2579-2586 (1968).

63. Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry* **72**, 248-254 (1976).

Acknowledgements

This work was supported by the National Key R&D Program of China (2019YFA0905502), the National Natural Science Foundation of China (21877053, 31900066), the Top-Notch Academic Programs Project of Jiangsu Higher Education Institutions (TAPP), the National First-class Discipline Program of Light Industry Technology and Engineering (LITE2018-24), the Fundamental Research Funds for the Central Universities (JUSRP51705A)**.**

Author contributions

Y.D. and M.Z. conceived the study; M.Z. designed and performed the experiments; S.Z. assisted the fluorescence measurement; L.W. assisted with data analysis and programs; M.Z., Y.D, and S.Z.. wrote the draft; S.Z., and M.Z. drew the figures; Y.D., S.Z., and M.Z. discussed and provided suggestions for this study. All authors reviewed, approved, and contributed to the final version of the manuscript.

Competing financial interests

The authors declare no competing financial interests.

663 **Additional information**

664 **Additional file 1:**

665 Fig. S1 The promoter library profiles of each round of MCSC engineering cycles.

666 Fig. S2 Comparing the fluorescence, LacZ activity, and transcriptional level of the

667 Ptrc-derived synthetic promoters.

668 Fig. S3 Comparing the fluorescence, LdhA activity, and transcriptional level of the

669 Ptrc-derived synthetic promoters.

670 Table S1 Strains and plasmids used in this study

671 Table S2 Primers used in this study

672 **Additional file 2:**

673 3665 synthetic promoters

674 3665 recombinant strains

675 27 literature promoters

676 100 random assembly promoters

677 390625 assembly promoters

678

679

Figure legends

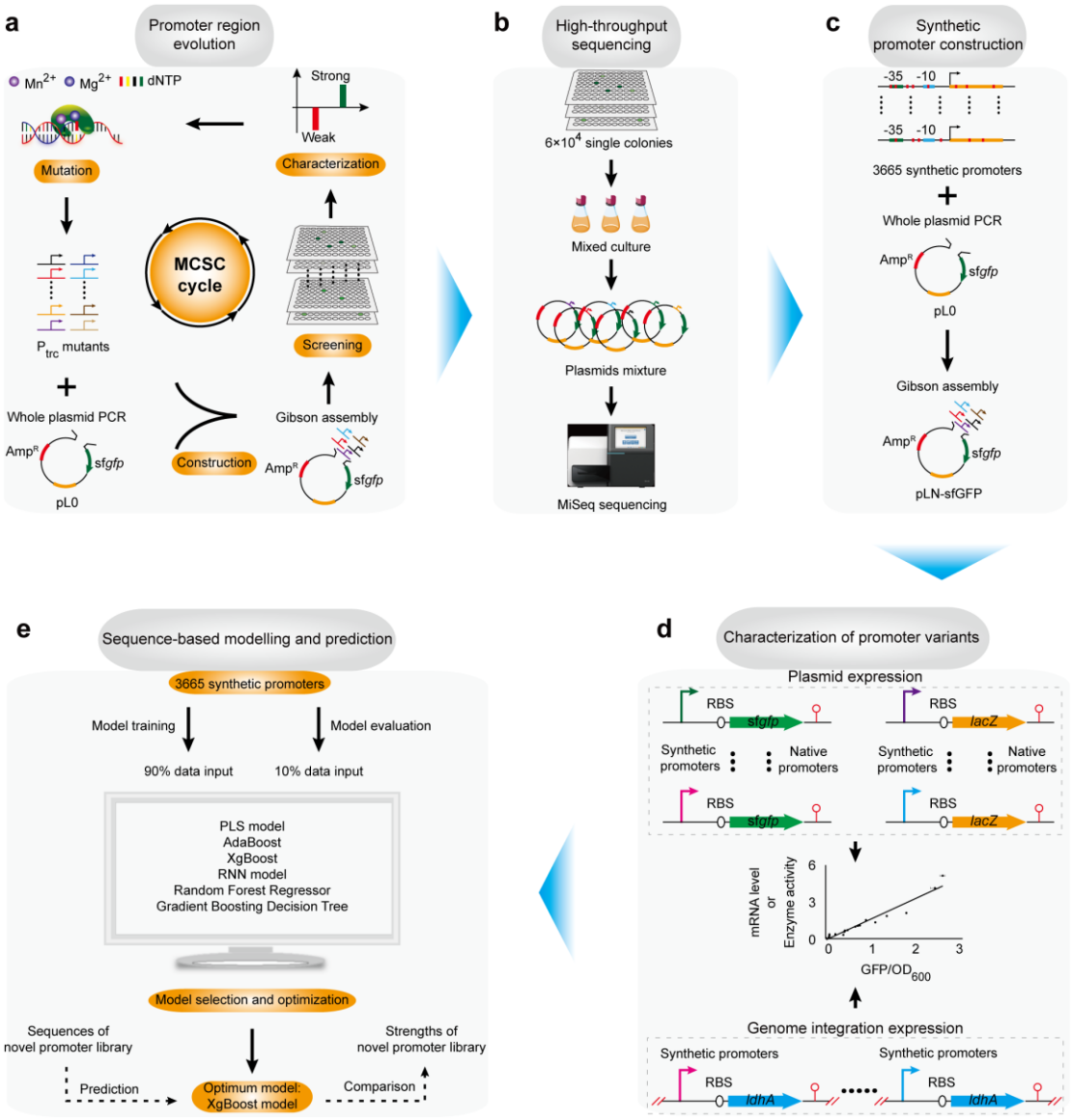

Fig. 1 Schematic diagram of the strategies based on the promoter library to predict promoter strength.

(a) Schematic illustration of the screening procedure for the promoter region evolution by MCSC engineering approach. (b) High-throughput sequencing for all mutant promoters. (c) Reconstruction of the synthetic promoter library based on MiSeq. (d) Characterization of promoter variants using different reporters. (e) Modeling and prediction based on base sequence and intensity.

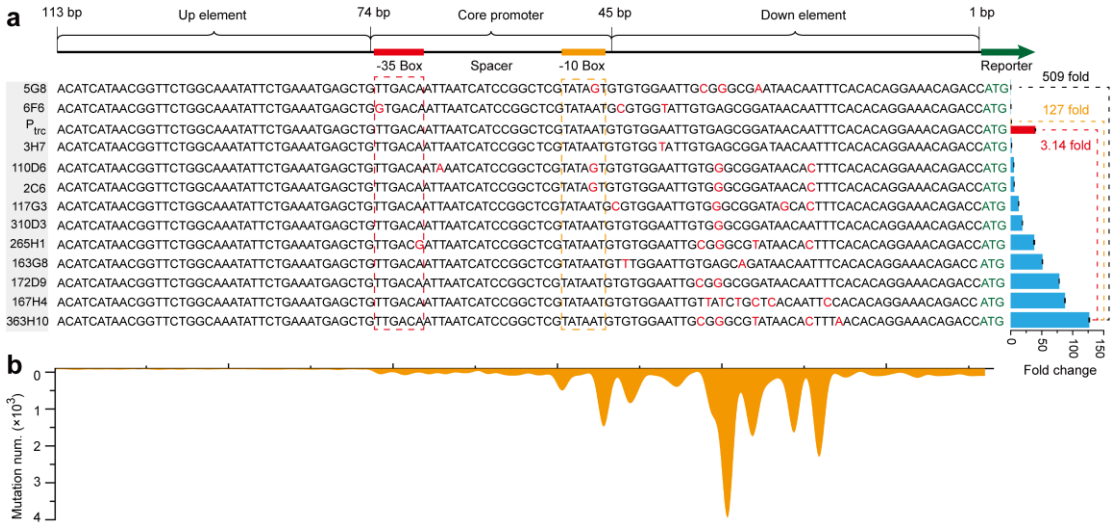

Fig. 2 Promoter abundance distribution map.

(a) Detection of fluorescence intensity and sanger sequencing for mutants during MCSC engineering. The structure of the entire promoter region included up element, core promoter, down element. 5G8 represents the strain's position was in G8 wells of the fifth 96 plates. Other similar strains were named as described above. Red column represents 1 mM IPTG induced P_{trc} promoter. Orange column represents uninduced P_{trc} promoter. (b) Distribution of mutation positions in high-throughput sequencing.

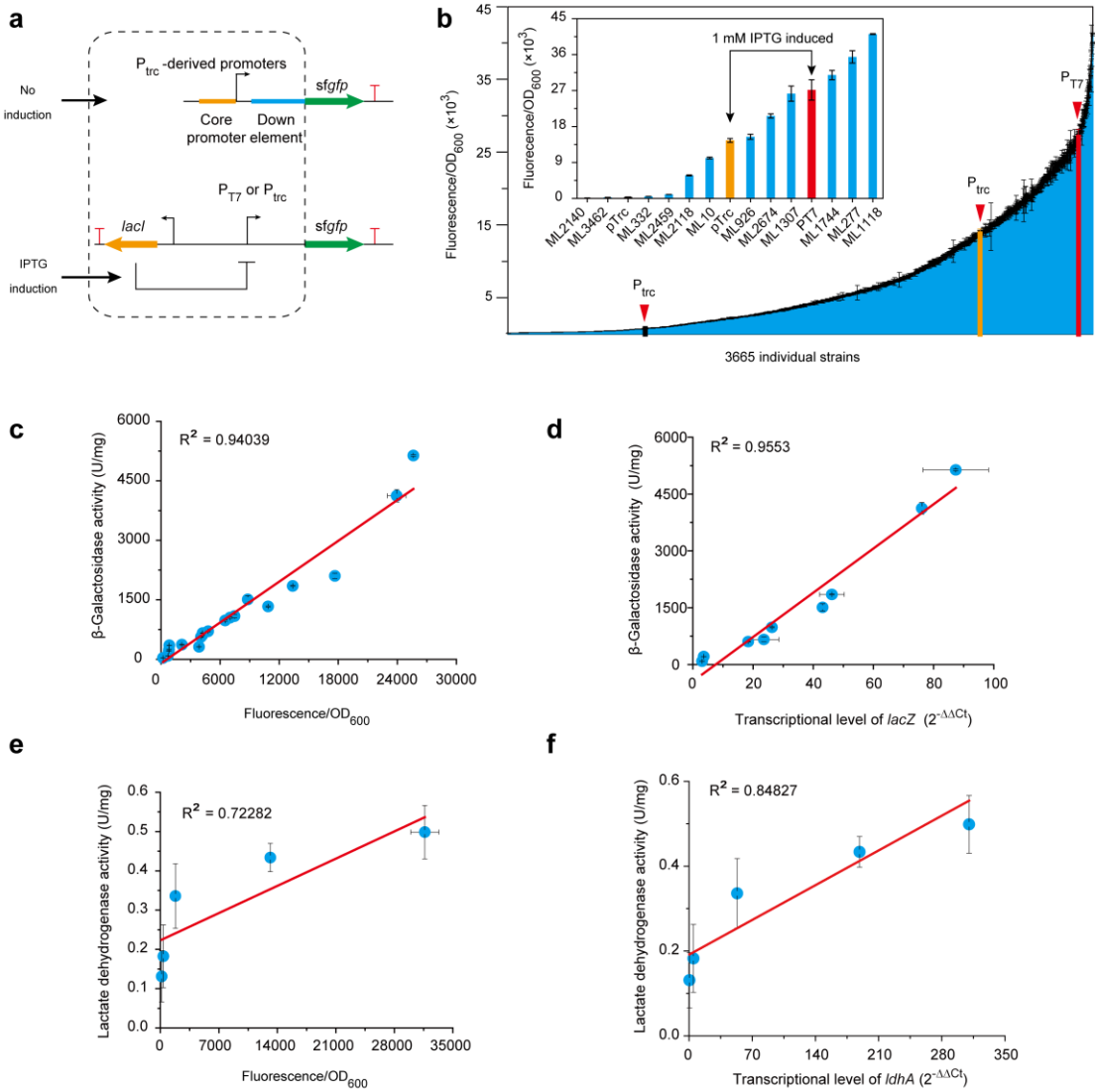

Fig. 3 Construction and characterization of the promoter clusters using different reporters.

(a) Schematic diagram of different promoters expressing the sfGFP protein. (b) Expressions and comparison of fluorescent intensity of different promoters. Data are means ± standard deviation for three independent experiments. P_{T7} and P_{trc} promoters that were induced by 1 mM IPTG were colored by red and yellow. Uninduced P_{trc} promoter was colored by black. The embedded figure represents part of the gradient strength promoters in the synthetic promoter library. (c) Relationship between β-galactosidase activity and sfGFP expression levels. (d) Cognation of the activity of promoter candidates at the transcriptional and expression level of *lacZ*. Level of

709 changes of mRNA level ($2^{-\Delta\Delta Ct}$) of *lacZ* measured by real-time fluorescence
710 quantitative PCR. (e) Relationship between lactate dehydrogenase activity and sfGFP
711 expression levels in genome. (f) Cognation of the activity of promoter candidates at the
712 transcriptional and expression level of *ldhA*. mRNA level($2^{-\Delta\Delta Ct}$) of ldhA was measured
713 by real-time fluorescence quantitative PCR. All experiments were performed three
714 times and the error bars represent standard deviation.
715
716

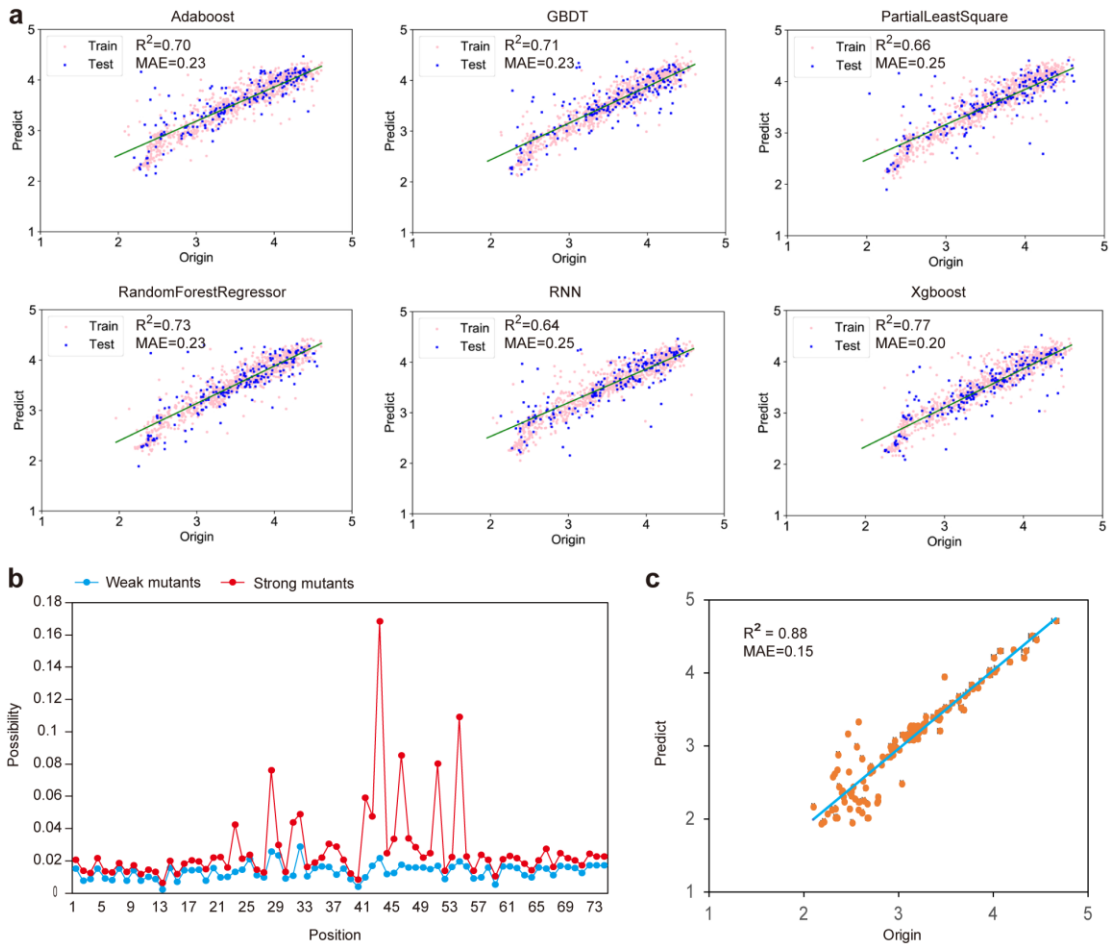

Fig. 4 Accurate prediction of the correlation between the promoter sequence and intensity by the machine learning model.

(a) Comparison and establishment of different models based on the P_{trc}-derived synthetic promoter library. (b) Statistical distribution of mutations and their effects on mutant fluorescence. The red and blue curves represent strong and weak mutants, respectively. (c) The predicted promoter strength (predict, $\log_{10}^{(\text{fluorescence/OD600})}$) versus the observed promoter strength (origin, $\log_{10}^{(\text{fluorescence/OD600})}$) of the training set and the test set. The R criterion was given as the relative error sum of squares. MAE represent mean absolute error.