Species-specific design of artificial promoters by transfer-learning based generative deep-learning model

Yan Xia[1], Xiaowen Du[1], Bin Liu[2], Shuyuan Guo[1*], Yi-Xin Huo[1,3*]

[1] *Key Laboratory of Molecular Medicine and Biotherapy, School of Life Science, Beijing Institute of Technology, 100081 Beijing, China;*

[2] *School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China*;

[3] *Tangshan Research Institute, Beijing Institute of Technology, 063611 Hebei, China;*

[*]Corresponding authors. E-mail addresses: guosy@bit.edu.cn (S. Guo) huoyixin@bit.edu.cn (Y.-X. Huo).

Abstract

Native prokaryotic promoters share common sequence patterns, but are species dependent. For understudied species with limited data, it is challenging to predict the strength of existing promoters and generate novel promoters. Here, we developed PromoGen, a collection of nucleotide language models to generate species-specific functional promoters, across dozens of species in a data and parameter efficient way. Twenty-seven species-specific models in this collection were finetuned from the pretrained model which was trained on multi-species promoters. When systematically compared with native promoters, the *Escherichia coli*- and *Bacillus subtilis*-specific artificial PromoGen-generated promoters (PGPs) were demonstrated to hold all essential features and distribution patterns of native promoters. A regression model was

developed to score generated either by PromoGen or by another competitive neural network, and the overall score of PGPs is higher. Encouraged by *in silico* analysis, we further experimentally characterized twenty-two *B. subtilis* PGPs *in vivo*, results showed that four of tested PGPs reached the strong promoter level while all were active. Furthermore, we developed a user-friendly website to generate species-specific promoters for 27 different species by PromoGen. This work presented an efficient deep-learning strategy for *de novo* species-specific promoter generation even with limited datasets, providing valuable promoter toolboxes especially for the metabolic engineering of understudied microorganisms.

Introduction

Regulating gene expression is essential to metabolic and biosynthetic engineering, as it enables the optimization of the production of desired metabolites, biofuels, pharmaceuticals, and other valuable compounds (1-4). Promoters are core regulatory elements that initiate the gene transcription, regulating gene expression, and affecting metabolic flux distribution in metabolic pathways (5,6). By gaining in-depth understanding and flexibly manipulating promoters, scientists could redesign cells creating gene circuits and synthetic biological systems tailored to specific requirements (7). Although native promoters have been utilized to regulate gene expressions, their regulatory scope is limited, lacking continuous regulatory strength (8). Therefore, many efforts have been made to develop artificial or hybrid promoters for regulating gene expressions (9-11).

43 The field of promoter engineering leverages techniques like random mutagenesis
44 and the re-assembly of promoter elements to develop libraries with new and varied
45 promoter sequences (12,13). Using random mutagenesis, bases could be mutated to
46 create a variety of expression levels in promoter region. For example, error-prone PCR
47 was applied to the PL-λ promoter in *Escherichia coli*, leading to a substantial 196-fold
48 variation in promoter activity identified through high-throughput screening (14). This
49 approach has also been successfully applied to other organisms, including *Lactococcus*
50 *lactis*, *Pichia pastoris*, and *Saccharomyces cerevisiae* Tef1 to generate active promoters
51 (15-17). Another strategy is the combination of core promoter elements, which has been
52 used to engineer synthetic promoter libraries with a wide range of transcriptional
53 activities up to 100-fold by altering native promoters like P_{veg}, P_{serA}, and P_{ymdA} in
54 *Bacillus subtilis*, specifically fixing the -10 and -35 regions (16,18). These methods are
55 pivotal for creating promoter libraries, but challenges persist in generating and
56 screening large-scale libraries for promoters with a broad and continuous range of
57 strengths (8).

58 Deep learning architectures have shown substantial promise for the manipulation
59 of both eukaryotic and prokaryotic promoter elements. These sophisticated models
60 excel in generating a lot of functional promoters by assimilating and analyzing
61 distribution patterns inherent in extensive native promoter sequences (19,20). A case in
62 point is the employment of Generative Adversarial Networks (GANs) for the *de novo*
63 generation of synthetic promoters, utilizing a 14,098 experimentally validated

64 promoters as a training dataset and 70.8% generated promoters showed activity in *E.*
65 *coli* (21). In a recent advancement, this model has been refined to not only produce
66 entire promoter sequences but also to engineer flanking regions while maintaining the
67 integrity of the pivotal -10 and -35 core regions in *E. coli* and mammalian cells (22). In
68 addition, a Variational Autoencoder (VAE) was trained using 3,712 promoter sequences,
69 as identified through differential RNA sequencing (dRNA-seq), to generated synthetic
70 promoters of varying efficacy in cyanobacteria (23). These models demonstrated robust
71 performance when supplied with extensive, high-quality experimental data. As the
72 scope of metabolic engineering broadens to encompass an array of non-model
73 prokaryotic organisms, the dearth of comprehensive and high-quality datasets emerges
74 as a formidable challenge.

75 Recently, the field of text generation has witnessed significant advancements
76 through the application of transfer learning, which encompasses the combination of
77 Generative Pre-trained Transformers (GPT) with instruction datasets for the purpose of
78 fine-tuning (24-26). This approach is exemplified by the success of ChatGPT (**Figure**
79 **1A**). In the initial phase, the model learns general patterns and distribution regularities
80 from a multitude of unlabeled datasets during pre-training. Thereafter, it undergoes a
81 fine-tuning process, steered by data related to specific tasks, thereby equipping the
82 model to adapt efficiently to a variety of application contexts (26). This methodology
83 is particularly beneficial in scenarios where data is limited, as transfer learning
84 effectively augments the precision of model and adaptability in such data-scarce

85 situations. Biological sequences including proteins and nucleotides are regarded similar
86 to natural languages as both of them use a set of symbols to represent information
87 (27,28). Some nucleotide language models were designed to understand the genetic
88 information, include DNABert, Nucleotide Transformers, GENA-LM, HyenaDNA
89 (27-29). By training on genome sequences, nucleotide language models learned
90 transferrable understandings of genomic sequences. However, most nucleotide
91 language models are BERT-like model trained to learn representations of nucleotide
92 sequences rather than generation (30). One of the current challenges in developing
93 generative model is to benchmark the quality of generated sequences. Typically, the
94 evaluation of generation quality relies on scores of third-party models *in silico* and
95 experimentally verification *in vivo* or *in vitro*. In silico evaluations commonly have a
96 much higher through-put than *in vivo* or *in vitro*, but the later are more reliable (31).

97 To address the challenge of insufficient promoter data of certain species, we
98 designed a pre-trained model based on Transformer, termed the PromoGen-pre. This
99 model was further refined with species-specific promoter data to yield 27 distinct
100 promoter-generating models for both model and non-model species. To validate the
101 effectiveness of our strategy, we evaluated the key characteristics of the generated
102 promoters including position weight matrix, 6-mer frequency correlation and -10 region
103 distribution in *E. coli* and *B. subtilis*. We developed a scoring model to evaluate the *E.*
104 *coli* promoters generated by our model and WGAN-GP, demonstrating the
105 competitiveness of our model. Most of the generated promoters exhibited confirmed

106 activities in *B. subtilis*. An online webserver was developed to enhance the accessibility
107 of the promoter-generative models where users could generate *de novo* promoters for
108 27 different species based on the PromoGen model. The related models and analysis
109 methods are listed in Supplementary Table S1. Via aforementioned transfer learning
110 approach for generating new promoters from scratch, we expanded the available
111 promoters for a specific species and offered potential candidates for use in metabolic
112 engineering even the available data is limited.

113 **Results**

114 **PromoGen-pre is a generative pre-trained model for promoter sequences**

115 Unsupervised language models have become the foundation of current advancement in
116 the natural language processing (NLP) field (32). Language model could be built and
117 trained based on the large number of unlabeled data, and be improved via transfer-
118 learning for performing downstream tasks with task-specific dataset. Transfer-learning
119 dependent strategy, which has been successfully exploited to generate functional
120 proteins across diverse families recently, has yet been reported to generate new
121 functional DNA elements. To investigate the potential of transfer learning on generating
122 DNA regulatory elements, we developed a decoder-only Transformer-based model,
123 termed PromoGen-pre, to specifically capture the characteristics of prokaryotic
124 promoter sequences (**Figure 1B**). Unlike the conventional byte-pair encoding (BPE) or
125 k-mer tokenization strategy (33,34), the DNA sequences were tokenized at the character
126 level in PromoGen-pre to ensure that the original location of each base in the DNA

127 sequence was kept within the position of each token, allowing the DNA generation
128 more controllable (**Supplementary Figure S1A**). The tokenized input sequence was
129 embedded as numerical representation $X_{emb} \in \mathbb{R}^{L \times D}$, where L is the sequence length
130 and D is the dimension of the hidden vector of each token. PromoGen model further
131 updated the numerical representation though GPT2Blocks with multi-head self-
132 attention (**Figure 1B, Supplementary Figure S1B**) (25).

133 When predicting the next token, the long-distance relation across the sequence
134 would be taken account by PromoGen-pre via the self-attention mechanism.
135 PromoGen-pre was trained by autoregressive approach, where the probability of each
136 nucleotide in the sequence was conditioned solely on the upstream nucleotides. Cross-
137 entropy loss was utilized to evaluate and adjust the performance of PromoGen-pre
138 during the model training process. The 10 million tokens from more than 129k
139 sequences that containing experimentally verified promoters and Transcriptional Start
140 Sites (TSS) in the Prokaryotic Promoter Database (PPD), which is one of the most
141 comprehensive promoter sequence databases containing sequences from more than 27
142 species, were subjected to the pretraining of PromoGen **(Figure 1C, D)** to ensure the
143 high quality and reduced deviation (**Figure 2A, 2B**) (35).

144 After the above training, we observed that the promoters from the same species
145 were clustered close to each other when the sequence representations encoded by
146 PromoGen-pre model were projected to a two-dimensional plan, indicating that the
147 model has learned to distinguish the promoters by species in the absence of supervision

148 related to species information (**Figure 2C**). For example, the sequences from *Haloferax*
149 were notably well distinguished from other species by PromoGen, highlighting the
150 differences of regulatory elements between Archaea and Bacteria (**Figure 2C**) and
151 confirming that PromoGen-pre is able to encode meaningful promoter representations.
152 On the other hand, some projected points representing promoters from different species
153 were also gathered together in the Figure 2C, indicating that some promoters from
154 distinguished species are sharing common patterns which is in agreement with the well-
155 known observation that bacteria are able to recognize certain heterologous genetic
156 elements. This observation reinforced our rationale to tackle the low-data challenge by
157 transfer learning.

158 Since PromoGen-pre has been trained to obtain meaningful representations from
159 native promoters, we investigated whether the trained PromoGen-pre could generate
160 functional artificial promoters. The results showed that the projected points of artificial
161 promoters generated by PromoGen-pre (pink points in Figure 2C) was dispersedly
162 distributed within, and did not form obvious cluster outside of, the ones of native
163 promoters, indicating that the generated promoters were likely to be functional as the
164 native ones. So far, the generated promoters were not species specific although the pre-
165 training stage allowed PromoGen-pre to learn the comprehensive features of promoters
166 from different species, encompassing a broader spectrum and evolutionary information.

167 Although previous works developed a collection of benchmarks including
168 eighteen genomic related prediction tasks (27), we find it is still challenging to evaluate

169 whether nucleotide language models could understand prokaryotic promoters because
170 currently only mouse and human promoters were involved in the benchmarks. On the
171 other hand, only promoter sequences, but not genomic chunks, were provided to
172 PromoGen-pre for training, making PromoGen-pre being intrinsically capable of
173 distinguishing native promoters and random sampled genomic chunks, which was a
174 challenging task for other models. Therefore, a benchmark dataset based on
175 experimentally verified synthetic promoters was needed for PromoGen-pre.

176 We first evaluated the ability of PromoGen-pre to distinguish whether a generated
177 sequence is a promoter in a zero-shot manner (36). The log-likelihood of each sequence
178 was calculated by PromoGen-pre to a predicted score. As a comparison, we also
179 evaluated the performance of the nucleotide transformer family models (NTFM) which
180 have number of parameters range from 50 million (M) to 2.5 billion (B) NTFM enabled
181 precise prediction of molecular phenotypes based on learned patterns from DNA
182 sequences (27). We computed the area under the receiver operating characteristic
183 (auROC) curves to predict the binary label from scores (37). The results showed that
184 the PromoGen-pre, a small model with solely 7.4 M parameters, has log-likelihood
185 auROC of 0.60, equally well, (if not better than) as the same level of the 2.5 B model
186 (**Figure 2D**), better than the Nucleotide Transformer 50 M, 100 M, 250 M, and 500 M.
187 This result showed that a relatively small amount of project-related high-quality data
188 could fulfill the task, enabling PromoGen to reach the performance of large model
189 trained with large dataset.

190 **Finetune the pre-trained model based on datasets from 27 species**

191 PromoGen-pre has been trained on multiple prokaryotic promoter sequence datasets,
192 endowing it with the capability to generate promoters for various species. To adapt the
193 model for species-specific applications, we conducted fine-tuning based on the datasets
194 of 27 major species in the dataset, resulting in the development of 27 distinct promoter
195 models, encompassing species such as *E. coli*, *B. subtilis*, *Sinorhizobium*,
196 *Corynebacterium*, and so on. Such transfer learning methodology facilitated the
197 development of a robust promoter generative model even with relatively small-size
198 datasets.

199 Taking *E. coli* model as an example to evaluate the effectiveness of transfer
200 learning, we then tried to score and compare the generation quality of PromoGen-eco
201 and a GAN-based WGAN-GP model through independent predictive models which
202 were designed to predict activities of promoters. To select the most reliable predictive
203 model for predicting the activities of the generated promoters, we evaluated twelve
204 machine learning based predictive models by calculating metrics including MAE,
205 RMSE, Spearman correlation coefficient, and Pearson correlation coefficient
206 (**Supplementary Figure S2**). the Bayesian Bridge model was selected as our final
207 predictive model because it consistently demonstrated exceptional performance no
208 matter which calculation methods were used for evaluation. The Bayesian Bridge
209 model achieved a Spearman correlation coefficient of 0.69 or 0.58, and a Pearson
210 correlation coefficient of 0.70 or 0.60 on training and test data, respectively,

211 demonstrating the accuracy of this model (**Figure 3A, 3B**).

212 We generated 200 promoter sequences by PromoGen-eco and WGAN-GP,
213 individually, and used Bayesian Bridge prediction model to predict the promoter
214 activities of these 400 promoters (21). As depicted in figure 3C, the average score of
215 PromoGen-eco generated promoters was significantly higher compared to the ones
216 generated by WGAN-GP, indicating the strong capability of PromoGen-eco in *de novo*
217 promoter generation. In addition, another prediction model employed in previous study
218 (21), Deep Promoter CNN, was utilized to assess the performance of PromoGen-eco.
219 The results indicated that PromoGen-eco exhibited high competitiveness when
220 compared with WGAN-GP (**Supplementary Figure S3**). Furthermore, analysis
221 utilizing PWM revealed that the promoters generated by PromoGen-eco exhibited a
222 distribution identical to that observed in native promoters while the distribution of
223 promoters generated by WGAN-GP displayed slight variations from the native
224 distribution (**Supplementary Figure S4**). To evaluate the similarity between natively
225 occurring promoters and those synthesized by PromoGen-eco, we counted the hexamer
226 frequencies of the promoters. The resulting Pearson and Spearman correlation
227 coefficients were 0.954 and 0.965, respectively, indicating a high degree of similarity
228 in the 6-mer frequency between the generated and native promoters (**Figure 3D**). These
229 findings collectively suggested that PromoGen-eco has successfully captured the
230 characteristics and distribution patterns of *E. coli* promoters, showcasing its strong
231 capability in *de novo* promoter generation.

232 **Performance evaluation of PromoGen-bsu model**

233 To further validate the effectiveness of various species specific PromoGen models, we
234 performed promoter prediction and activity evaluation in *B. subtilis,* a species with
235 higher promoter selectivity and is an important Gram-positive bacterium used in protein
236 production and chemical manufacturing. We fine-tuned PromoGen-pre using 691
237 promoter sequences from *B. subtilis* to obtained PromoGen-bsu, which was then
238 utilized to generate *B. subtilis* specific promoters (**Supplementary Figure S5**). To
239 assess the quality of generated promoters, PromoGen-bsu generated promoters were
240 randomly sampled and compared with native ones and PromoGen-pre generated ones.
241 We utilized PWMs to depict the sequence motifs within the promoter regions derived
242 from native sources, as well as those synthesized by PromoGen-pre and PromoGen-bsu
243 (**Figure 4A**). Our investigation revealed a conserved motif within the -10 box,
244 extending from position -13 to -8 in the native promoters. The conservation of this motif
245 aligns with earlier studies conducted on a variety of bacterial species, including *B.*
246 *subtilis*, *E. coli*, and cyanobacteria. Notably, in the -10 region of the PromoGen-bsu-
247 generated promoters, we detected a sequence pattern mirroring that of the native
248 promoters, suggesting that PromoGen-bsu has successfully replicated the essential
249 features of this -10 motif. Additionally, the -35 region, located at the positions of -36 to
250 -31, exhibited partial conservation in both the natively occurring and PromoGen-bsu-
251 constructed promoter sequences. Contrasting with this, the sequence of PromoGen-pre
252 attributes exhibited substantial deviation from the native *B. subtilis* promoters,

253 exhibiting only a marginal preference for T, A bases within the -10 region and no
254 preference within -35 region.

255 We conducted a quantitative assessment of hexamer frequencies in both natively
256 occurring promoters and those generated by PromoGen-pre and PromoGen-Bsu. The
257 analysis revealed a moderate correlation between the hexamer frequencies of native
258 promoters and those generated by PromoGen-pre, as indicated by Pearson and
259 Spearman correlation coefficients of 0.508 and 0.199, respectively. Conversely, a
260 significantly higher correlation was observed in sequences produced by PromoGen-bsu,
261 with Pearson and Spearman coefficients of 0.900 and 0.733, respectively, illustrating a
262 closer alignment with the hexamer frequency patterns found in native promoters
263 (**Figure 4B**). We noted a substantial consistency in high-frequency 6-mers between the
264 native and PromoGen-bsu generated promoters, specifically including sequences like
265 TAAAAT and TATAAT that encompass the -10 box. We also investigated the distance
266 distribution of TAAAAT and TATAAT sequences relative to the transcription start site
267 (TSS) (**Figure 4C**). The analysis of the promoters synthesized by both PromoGen-pre
268 and PromoGen-bsu revealed a distinguishable accumulation of hexamers, particularly
269 in the -10 region. This region exhibited a significantly higher frequency of these
270 sequences compared to other areas, aligning with the distribution observed in native
271 promoters. The sequences generated by PromoGen-bsu having a higher portion of both
272 TATAAT and TAAAAT as the -10 region comparing to PromoGen-pre. These findings
273 underscored the capability of the PromoGen-bsu model to accurately replicate the

274 spatial distribution patterns of conserved motifs inherent to native promoters, thereby
275 affirming its efficacy in reflecting the native positional preferences of these sequences.

276 The t-SNE visualization in Figure 4D illustrated the arrangement of promoter
277 sequences, *i.e.* native, PromoGen-pre, and PromoGen-bsu, which were projected from
278 a high-dimensional landscape into a two-dimensional plane. In this representation, the
279 promoters generated by the pre-trained model, depicted as orange dots, were dispersed
280 throughout the space, predominantly occupying the right side, which suggestsed a
281 distinct pattern from the rest. In contrast, the native and PromoGen-pre -generated
282 promoters, represented by green and blue dots respectively, were intermingled and
283 uniformly situated on the left side. This indicated a shared distribution pattern among
284 them, showing the PromoGen-bsu has learned the related distribution in *B. subtilis*
285 promoters.

286 **The activity of generated promoters in *B. subtilis***

287 To assess the functionality of the synthesized promoters, we utilized PromoGen-bsu to
288 create 50 *de novo* designed promoters, each has 61 nucleotides in length (**Figure 5A**).
289 From this pool, 22 were arbitrarily chosen for experimental testing. As benchmarks, we
290 employed native promoters from *B. subtilis*, such as P_{43}, P_{veg}, and P_{lepA}, alongside three
291 random sequences to serve as controls, both positive and negative. We constructed 28
292 shuttle vectors expressing fluorescent proteins *sfGFP*, each harboring the replication
293 origins pBR322 and RepA for *E. coli* and *B. subtilis*, enabling *sfGFP* expression in both
294 hosts. The activities of the promoters were determined by the fluorescence intensity

295 emitted by *sfGFP*. Sampling was conducted during the logarithmic growth phase for
296 fluorescence measurement. A fluorescence reading higher or lower than that of the
297 randomly generated sequences indicated an active or inactive promoter, respectively.

298 As shown in Figure 5B, the results indicated that PromoGen-bsu could generate
299 active promoters although the strength of the promoters vary. Three random sequences
300 exhibited no activity. This result was expected because the range of promoter intensities
301 use in the training data vary also. Almost all of these promoters are active, with about
302 18% of them including A7, A9, A5, and A3 showed activities similar to the natively
303 strong promoters P_{43} and P_{veg}. Notably, the A7 promoter stands out, exhibiting an
304 activity rate 91% higher than that of P_{43}. The strength of the other promoters varies,
305 allowing for different levels of gene expression to be achieved. The bacterial strains
306 expressed the *sfGFP* gene were subjected to observation under a fluorescence
307 microscope. Notably, these organisms demonstrated notable fluorescence when
308 compared to counterparts utilizing random sequences as promoters (**Supplementary**
309 **Figure S6**). Furthermore, our investigation encompassed the analysis of t-SNE
310 distribution patterns, compared experimentally tested 22 promoters with 691 natively
311 promoters in *B. subtilis*. This analysis showed that the generated promoters were quite
312 dispersed and were significantly distant from strong promoters, indicating a
313 considerable difference (**Figure 5C**). Examination of the five most potent promoters
314 uncovered a shared sequence, TTG, within the -35 region, and adjacent sequences were
315 identified spanning positions 17-18 bp between the -35 and -10 regions (**Figure 5D**).

316 Additionally, the activities of these promoters were tested in *E. coli.* The experimental
317 results demonstrated that all promoters were active in *E. coli*, with 64% showing
318 enhanced activity compared to *B. subtilis* (**Supplemental Figure S7**). This finding
319 further validates that our model generates sequences that are not random but
320 functionally significant in biological contexts. Moreover, these results support the
321 concept of the universality of prokaryotic promoters, as discussed in studies from 2018.
322 Our research also indicates that promoters in *B. subtilis* exhibit greater selectivity
323 compared to those in other species.

324 To assess the similarity of the generated promoters to native promoters, a
325 comprehensive BLAST analysis was performed, comparing the experimentally
326 validated promoters against the entire genome of *B. subtilis*. The results showed that
327 the generated sequences do not have similarity. This finding implies that the sequences
328 were predominantly *de novo* designed by the models, rather than being a result of data
329 leaks. This suggested that PromoGen-bsu has effectively captured the distribution
330 patterns of active *B. subtilis* promoters, exploring a sequence space of 4^{61} and
331 showcasing the ability to generate new active species-specific promoters.

332 **Development of a user-friendly web sever**

333 To facilitate the rapid generation of promoter sequences, we have developed an online
334 web server (https://promogen1.cloudmol.org/) providing fine-tuned models for 27
335 prokaryotic organisms. As shown in **Fig. 5E**, the promoter sequences could be
336 generated after choosing the specific organism. The number and length of the generated

promoters could be flexibly customized according to the needs. The generated results were presented in a sequence that ranked them according to decreasing values of the log-likelihood. The repetition penalty, Top P, and temperature were commonly utilized hyperparameters for controlling the diversity and quality of text generation. By increasing the repetition penalty, we could prevent the generative model from producing multiple repeated tokens. This strategy not only mitigates the occurrence of hallucinations in the generative model but also reduces the complexity involved in constructing plasmids. By adjusting the temperature and Top P settings, the diversity of the generated sequences could be regulated, thereby mitigating the risk of converging on local minima. This online platform offers a user-friendly interface for creating customized promoters for diverse species, serving specific needs and thereby facilitating the efficiency of subsequent experimental and research activities.

Discussion

Regulatory elements play a crucial role in metabolic engineering (38). By selecting appropriate promoters, we could precisely control the gene expression of specific metabolic pathways, enhancing or inhibiting the fluxes of particular metabolic routes to increase the production of target chemicals. Identifying and verifying appropriate regulatory elements in understudied microbes is frequently a challenging and time-consuming task (18). Deep learning offers a promising approach for generating promoters from scratch or predicting the activity of existing promoters. However, the limited data available for understudied microbes poses a significant barrier to the

effective utilization of deep learning models. Here, we proposed a fast and convenient solution based on generative self-supervised nucleotide language model, capable of generating promoters with different strengths. We employed a transfer learning strategy, which pre-trained the model using promoter sequences mainly from 27 species to learn the general patterns and paradigms of the promoters. Subsequently, fine-tuning was performed using species-specific promoter data, resulting in promoter models applicable to 27 different species. Our PromoGen was pretrained on multi-species data, making its adaptation to other species easier than WGAN-GP, which was only trained to generate *E. coli* promoters. Even for *E. coli* promoters, our PromoGen worked as accurate as, if not better, than the existing WGAN-GP model. Furthermore, under core region analysis and experimental validation, PromoGen-bsu model exhibited exceptional performance, with all of the generated promoters showing activities. Consistent results from model evaluation, key region analysis, and experimental validation indicated that PromoGen successfully learned promoter distribution patterns and could be used for generating entirely new promoters. For user convenience, an online webserver was developed containing data for 27 species, allowing users to quickly generate promoter sequences for their desired species.

The quality of the dataset plays a crucial role in the performance of deep learning models. Transfer learning offers an excellent solution for small-scale datasets. An increasing number of microorganisms are being used for various applications, but research on regulatory elements lags behind in non-model microorganisms. By utilizing

379 the pre-trained model PromoGen-pre, we are able to reduce the data requirements for
380 deep learning models, making it possible to *de novo* generate promoters for species
381 with limited data. Although there are some differences in promoters among different
382 species, the characteristics of the core region of prokaryotic promoters have been well-
383 learned by the pretrained model PromoGen-pre. The activity of generated promoters
384 could be enhanced by fixing key positions. Although the training data is limited
385 comparing with other models, the information about promoter sequences is highly
386 enriched. We observed DNA sequences differently from natural language in these
387 aspects because there is no evidence that there is a pre-organized grammar across all
388 species beside the Watson-Crick pairs. We also demonstrated that the zero-shot
389 prediction accuracy of PromoGen-pre on promoter activities were comparable or higher
390 than the ones of other nucleotide transformer models which are 6 to 330 times bigger
391 than PromoGen-pre. We anticipated that this information enrich strategy would be
392 helpful to projects with limited resources.

393 Although previous reports and our work demonstrated that it is effective to design
394 promoters using generative models, it is challenging to assess the quality of generated
395 promoters. The current methods for assessing generated sequences primarily
396 encompass wet lab verification and computational experiment evaluation. Wet lab
397 verification is the most precise approach, though large-scale experimental validations
398 are time-consuming and labor-intensive. The evaluation methods for computational
399 experiments include using comparative sample distributions and benchmark testing.

Deep learning is capable of learning shallow features of data, which facilitates easy overfitting to match the sample distribution. Therefore, biologists should focus on establishing high-quality assessment datasets that are comprehensive, covering a wide array of scenarios and conditions, to guarantee the accuracy of evaluations.

Despite the fact that employing promoters of varying strengths is one of the most crucial methods for regulating gene expressions, the integration of diverse regulatory elements such as Ribosome Binding Sites (RBS), terminators, insulators, and others, offers expanded options for gene expression modulation (39-41). Furthermore, synthetic promoters, which establish a predetermined expression intensity, present challenges in dynamically regulating gene expression during growth. Promoters that are inducible or dependent on sigma factors could dictate the timing and intensity of gene activation. In the future, the design and development of pertinent models carry substantial promise for applications within the realm of synthetic biology.

In summary, we implemented a transfer learning strategy to address the challenges posed by a limited dataset. By the combination of pre-training and finetune, 27 species-generated models were developed for *de novo* design of promoter sequences, which were evaluated by prediction model and experiments validation as promoters with exceptional performance. Furthermore, we have developed a user-friendly web server that enables rapid generation of promoters. Our work provided a novel strategy and a universal pre-trained model PromoGen-pre, showcasing the potential of deep learning methods to explore the sequence space of regulation elements, especially in situations

with limited data availability.

METHODS

Details of nucleotide language models

The vocabulary of PromoGen contains 7 tokens, including four bases A, T, C, G, and special tokens including [PAD], [MASK], and [UNK]. By applying one-hot encoding, the input sequences are transformed into numerical vectors that can be processed by neural networks. The one-hot vectors are transformed into hidden state vectors of 320 dimensions:

$$x_{emb} = E(x_i) \tag{1}$$

To distinguish the positions of each token, a learnable positional embedding is added:

$$PE(pos, 2i) = sin\left(pos/10000^{2i/d_{model}}\right) \tag{2}$$

$$PE(pos, 2i+1) = cos\left(pos/10000^{2i/d_{model}}\right) \tag{3}$$

where the pos is the position of the context, i is the index of the dimension in the encoding vector, and the d_{model} represents the dimensionality of the outputs in all layers of the model. This position embedding is followed the original design of the GPT2 model (24).

The longest sequence allowed is 128. The hidden vectors are updated by 6 layers of GPT2 Block (25). In a GPT2 block, the hidden vectors are updated with multi-head self-attention, and a feed forward layer, with residual connections:

$$Multihead\ Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{(QK^T)}{\sqrt{d_k}}\right)\mathbf{V} \tag{4}$$

Q is query matrix, **K** is key matrix, **V** is value matrix, and d_K is dimension of the key vectors. $\mathbf{W_1}$ and $\mathbf{W_2}$ is Weight matrices of the linear transformations. b_1 and b_2 is bias

terms of the linear transformations. The number of head in multi-head self-attention is set to 4.

$$FFN(x) = f(xW_1 + b_1)W_2 + b_2 \quad (5)$$

The feed forward layer (FFN) is consisted of two linear layers, to expend the dimension by a factor of 2 and activated using the Gaussian Error Linear Unit (GELU) function.

$$Output = LayerNorm\big(x + Sublayer(x)\big) \quad (6)$$

Residual connections and layer normalizations are applied in GPT2 block. The last hidden vectors are transformed to logits of tokens by the language modeling head, which is a linear layer has output dimension equal to the size of the vocabulary. The final output of the model is the probability of each base at the next position and generates promoter sequences by autoregression way from 5' to 3', one by one, which the next base has relied on all previously generated bases.

$$p(x) = \prod_{i=1}^{n} p(S_n | S_1, \dots, S_{n-1}) \quad (7)$$

Training of nucleotide language models

To train the PromoGen, we used 120k promoter sequences containing more than 27 species from the Prokaryotic Promoter Database (PPD). The PromoGen model is a model trained using the GPT-2 method (25). To efficiently train this model, the batch size was configured at 64, allowing for the simultaneous processing of 64 data samples. We predict the probability of each token based on previous context, comparing these predictions with actual data by computing cross-entropy loss to assess the prediction accuracy of model. To optimize the performance of model, the Adam optimizer was employed (42), renowned for its capability to dynamically adjust the learning rate,

466 thereby bolstering the training efficiency and overall efficacy. We set several key
467 parameters, including two decay rate parameters, β1 and β2, set to 0.9 and 0.999,
468 respectively, and a constant called epsilon, set to $5e^{-4}$. During the fine-tuning phase, the
469 weights from the pre-trained model are utilized as initial values. We configure the batch
470 size to be 32, and the learning rate for the Adam optimizer is set at $1e^{-5}$. Other
471 parameters remain consistent with those established in the PromoGen configuration.

472 **Training and selection of predictive models**

473 We systematically evaluated 12 machine learning models to accurately predict the
474 activity of promoters, including Extra Tree Regressor, Linear Regression, Gradient
475 Boosting Regressor, Ada Boost Regressor, Ridge Lasso K-Neighbors Regressor,
476 Random Forest Regressor, Bayesian Ridge, Support Vector Regression (SVR),
477 eXtreme Gradient Boosting (XGB) Regressor, and Decision Tree Regressor (43-50).
478 We extracted features from promoters using a nucleotide pre-trained model, employing
479 the features from the [CLS] token as the representation for the entire sequence, which
480 served as input to the machine learning models. To ensure a fair comparison of these
481 methods, we conducted 5-fold cross-validation across the entire dataset ensuring
482 independent predictions for all samples. At the same time, we calculated Mean Absolute
483 Error (MAE), Root Mean Square Error (RMSE), Spearman correlation, and Pearson
484 correlation coefficients (51,52).

485 To further enhance the accuracy of the predictive model, we conducted a
486 hyperparameter search for the Bayesian Bridge model. These hyperparameters included

maximum optimization iterations, alpha1, alpha2, lambda1, and lambda2, resulting in a total of 3^5 combinations. Setting the maximum optimization iterations to 300 and alpha1, alpha2, lambda1 as $1e^{-7}$, and lambda2 as $1e^{-5}$ yielded the best performance in cross-validation. These optimal hyperparameters were used to Bayesian Bridge model for predicting promoter strength, ensuring our model achieves the highest level of accuracy.

The model evaluation by position weight matrix (PWM) and 6-mer frequency analyses

We performed multiple sequence alignment separately for both native and model-generated promoter sequences. We calculated the weight for each position using a formula:

$$W_{i,j} = log_2(\frac{f_{i,j}}{p_i}) \tag{8}$$

where $W_{i,j}$ represents the weight score of the ith base at position j. $f_{i,j}$ indicates the relative frequency of the base at position j, and p_i represents the background frequency of that base. We utilized the obtained weight scores to create visualization using the online website WebLogo.

In the analysis of 6-mer frequencies, we conducted a comprehensive scan of all 6-mer sequences within both natural and synthetic promoters generated by PromoGen and PromoGen-bsu, subsequently computing the frequency of each individual 6-mer. During this process, particular attention was devoted to two pivotal 6-mer sequences,

507 TATAAT and TAAAAT, which frequently play crucial roles in the functionality of
508 promoters. To gain deeper insights into the representation of these key sequences across
509 various promoters, we specifically calculated their abundance distribution compared
510 with natural promoters.

511 **Plasmid construction**

512 This investigation employed a variety of plasmids, detailed in Supplementary
513 Table S2. All primer sequences were synthesized from Genewiz, located in Tianjin,
514 China (Supplementary Table S3). The shuttle plasmid, pXY43-sfGFP, was utilized as a
515 template, incorporating both pBR322 ori and repA ori for compatibility with *E. coli* and
516 *B. subtilis* respectively. A series of 22 generated promoter sequences were integrated
517 upstream of the superfolder green fluorescent protein (*sfGFP*) gene, replacing the
518 existing P_{43} promoter via polymerase chain reaction (PCR) techniques. The amplified
519 PCR constructs were subsequently cultured on LB agar plates containing 50 μg ml^{-1} of
520 kanamycin to facilitate clone selection. Three clones were selected to Sanger
521 sequencing for sequence verification in each plate. Furthermore, corresponding
522 plasmids containing the P_{veg} and P_{lepA} promoters from *B. subtilis* were also engineered,
523 alongside a set of plasmids with random sequences as promoters, to serve as negative
524 controls in the study.

525 **Fluorescence measurements**

526 The activity of these generated promoters was assessed based on the fluorescence

527 intensity emitted by *sfGFP*. The relevant promoter-plasmid were transformed into *B.*
528 *subtilis* strains, with single colonies subsequently cultivated in LB medium
529 supplemented with 20 μg ml^{-1} kanamycin, maintained at 37°C and agitated at 220 r.p.m
530 (Supplementary Table S4). After an overnight growth, these cultures were then
531 transferred into fresh LB medium containing a 1% (v/v) inoculum, followed by an 8-
532 hour incubation at 37°C. The fluorescence intensity was quantitatively measured using
533 a microplate reader (BioTek) set at an excitation wavelength of 488 nm and an emission
534 wavelength of 507 nm. Concurrently, the cell optical density (OD) was detected at 600
535 nm. All these experimental procedures were replicated thrice to ensure reliability.

536 **Data availability:** All necessary data to assess the conclusions in the paper are
537 presented within the paper itself and/or the Supplementary Materials.

538 **Author contributions:** Y.-X.H., S.Y.G., and Y.X. conceptualized the project. Y.X.
539 designed, built, and implemented deep models in this study. Y.X., and X.W.D. carried
540 out the wet experiments. Y.-X.H., S.Y.G., and Y.X. analyzed the experiments data, and
541 wrote the manuscript.

542 **Acknowledgements:** This work was supported by the National Key R&D Program of
543 China (2021YFC2100500), and the Natural Science Foundation of China (grant
544 number 32370095). Part of the experiments was carried out in the Biological & Medical
545 Engineering Core Facilities of the Beijing Institute of Technology. We would like to
546 thank Jinyuan Sun for insightful discussions.

547 **Competing interests:** The authors declare no competing interests.

548 **References:**

549 1. Wang, H.H., Isaacs, F.J., Carr, P.A., Sun, Z.Z., Xu, G., Forest, C.R. and Church,
550 G.M. (2009) Programming cells by multiplex genome engineering and
551 accelerated evolution. *Nature*, **460**, 894-898.
552 2. Keasling, J.D. (2010) Manufacturing molecules through metabolic engineering.
553 *Science*, **330**, 1355-1358.
554 3. Xu, P., Li, L., Zhang, F., Stephanopoulos, G. and Koffas, M. (2014) Improving
555 fatty acids production by engineering dynamic pathway regulation and
556 metabolic control. *Proc Natl Acad Sci U S A*, **111**, 11299-11304.
557 4. Xu, X., Li, X., Liu, Y., Zhu, Y., Li, J., Du, G., Chen, J., Ledesma-Amaro, R. and
558 Liu, L. (2020) Pyruvate-responsive genetic circuits for dynamic control of
559 central metabolism. *Nat Chem Biol*, **16**, 1261-1268.
560 5. Cases, I. and de Lorenzo, V. (2005) Promoters in the environment:
561 transcriptional regulation in its natural context. *Nat Rev Microbiol*, **3**, 105-118.
562 6. Jones, J.A., Toparlak Ö, D. and Koffas, M.A. (2015) Metabolic pathway
563 balancing and its role in the production of biofuels and chemicals. *Curr Opin*
564 *Biotechnol*, **33**, 52-59.
565 7. Carr, P.A. and Church, G.M. (2009) Genome engineering. *Nature*
566 *Biotechnology*, **27**, 1151-1162.
567 8. Johns, N.I., Gomes, A.L.C., Yim, S.S., Yang, A., Blazejewski, T., Smillie, C.S.,
568 Smith, M.B., Alm, E.J., Kosuri, S. and Wang, H.H. (2018) Metagenomic mining
569 of regulatory elements enables programmable species-selective gene expression.
570 *Nat Methods*, **15**, 323-329.
571 9. Kang, M.K., Lee, J., Um, Y., Lee, T.S., Bott, M., Park, S.J. and Woo, H.M. (2014)
572 Synthetic biology platform of CoryneBrick vectors for gene expression in
573 Corynebacterium glutamicum and its application to xylose utilization. *Appl*
574 *Microbiol Biotechnol*, **98**, 5991-6002.
575 10. Markley, A.L., Begemann, M.B., Clarke, R.E., Gordon, G.C. and Pfleger, B.F.
576 (2015) Synthetic biology toolbox for controlling gene expression in the
577 cyanobacterium Synechococcus sp. strain PCC 7002. *ACS Synth Biol*, **4**, 595-

578 603.
579 11. Elmore, J.R., Furches, A., Wolff, G.N., Gorday, K. and Guss, A.M. (2017)
580 Development of a high efficiency integration system and promoter library for
581 rapid modification of Pseudomonas putida KT2440. *Metab Eng Commun*, **5**, 1-
582 8.
583 12. He, S., Zhang, Z. and Lu, W. (2023) Natural promoters and promoter
584 engineering strategies for metabolic regulation in Saccharomyces cerevisiae. *J*
585 *Ind Microbiol Biotechnol*, **50**.
586 13. Cazier, A.P. and Blazeck, J. (2021) Advances in promoter engineering: Novel
587 applications and predefined transcriptional control. *Biotechnol J*, **16**, e2100239.
588 14. Alper, H., Fischer, C., Nevoigt, E. and Stephanopoulos, G. (2005) Tuning
589 genetic control through promoter engineering. *Proc Natl Acad Sci U S A*, **102**,
590 12678-12683.
591 15. Hartner, F.S., Ruth, C., Langenegger, D., Johnson, S.N., Hyka, P., Lin-
592 Cereghino, G.P., Lin-Cereghino, J., Kovar, K., Cregg, J.M. and Glieder, A.
593 (2008) Promoter library designed for fine-tuned gene expression in Pichia
594 pastoris. *Nucleic Acids Res*, **36**, e76.
595 16. Redden, H. and Alper, H.S. (2015) The development and characterization of
596 synthetic minimal yeast promoters. *Nat Commun*, **6**, 7810.
597 17. Jensen, P.R. and Hammer, K. (1998) The sequence of spacers between the
598 consensus sequences modulates the strength of prokaryotic promoters. *Appl*
599 *Environ Microbiol*, **64**, 82-87.
600 18. Guiziou, S., Sauveplane, V., Chang, H.J., Clerté, C., Declerck, N., Jules, M. and
601 Bonnet, J. (2016) A part toolbox to tune genetic expression in Bacillus subtilis.
602 *Nucleic Acids Res*, **44**, 7495-7508.
603 19. Kotopka, B.J. and Smolke, C.D. (2020) Model-driven generation of artificial
604 yeast promoters. *Nat Commun*, **11**, 2113.
605 20. Zrimec, J., Fu, X., Muhammad, A.S., Skrekas, C., Jauniskis, V., Speicher, N.K.,
606 Börlin, C.S., Verendel, V., Chehreghani, M.H., Dubhashi, D. *et al.* (2022)
607 Controlling gene expression with deep generative design of regulatory DNA.
608 *Nat Commun*, **13**, 5099.
609 21. Wang, Y., Wang, H., Wei, L., Li, S., Liu, L. and Wang, X. (2020) Synthetic
610 promoter design in Escherichia coli based on a deep generative network.

Nucleic Acids Res, **48**, 6403-6412.

22. Zhang, P., Wang, H., Xu, H., Wei, L., Liu, L., Hu, Z. and Wang, X. (2023) Deep flanking sequence engineering for efficient promoter design using DeepSEED. *Nat Commun*, **14**, 6309.

23. Seo, E., Choi, Y.N., Shin, Y.R., Kim, D. and Lee, J.W. (2023) Design of synthetic promoters for cyanobacteria with generative deep-learning model. *Nucleic Acids Res*, **51**, 7071-7082.

24. Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018) Improving language understanding by generative pre-training.

25. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019) Language models are unsupervised multitask learners. *OpenAI blog*, **1**, 9.

26. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. and Askell, A. (2020) Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877-1901.

27. Dalla-Torre, H., Gonzalez, L., Revilla, J.M., Carranza, N.L., Grzywaczewski, A.H., Oteri, F., Dallago, C., Trop, E., Sirelkhatim, H., Richard, G. *et al.* (2023) The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv*, 2023.2001.2011.523679.

28. Ji, Y., Zhou, Z., Liu, H. and Davuluri, R.V. (2021) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, **37**, 2112-2120.

29. Fishman, V., Kuratov, Y., Petrov, M., Shmelev, A., Shepelin, D., Chekanov, N., Kardymon, O. and Burtsev, M. (2023) GENA-LM: A Family of Open-Source Foundational Models for Long DNA Sequences. *bioRxiv*, 2023.2006.2012.544594.

30. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

31. Khakzad, H., Igashov, I., Schneuing, A., Goverde, C., Bronstein, M. and Correia, B. (2023) A new age in protein design empowered by deep learning. *Cell Systems*, **14**, 925-939.

32. Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, O., Agirre,

E., Heintz, I. and Roth, D. (2023) Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Comput. Surv.*, **56**, Article 30.

33. Bostrom, K. and Durrett, G. (2020) Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*.

34. Sanabria, M., Hirsch, J. and Poetsch, A.R. (2023) Distinguishing word identity and sequence context in DNA language models. *bioRxiv*, 2023.2007.2011.548593.

35. Su, W., Liu, M.-L., Yang, Y.-H., Wang, J.-S., Li, S.-H., Lv, H., Dao, F.-Y., Yang, H. and Lin, H. (2021) PPD: a manually curated database for experimentally verified prokaryotic promoters. *Journal of Molecular Biology*, **433**, 166860.

36. Li, B., Gou, Y., Liu, J.Z., Zhu, H., Zhou, J.T. and Peng, X. (2020) Zero-shot image dehazing. *IEEE Transactions on Image Processing*, **29**, 8457-8466.

37. Hanley, J.A. and McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29-36.

38. Ko, Y.-S., Kim, J.W., Lee, J.A., Han, T., Kim, G.B., Park, J.E. and Lee, S.Y. (2020) Tools and strategies of systems metabolic engineering for the development of microbial cell factories for chemical production. *Chemical Society Reviews*, **49**, 4615-4636.

39. de Almeida, B.P., Reiter, F., Pagani, M. and Stark, A. (2022) DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nature Genetics*, **54**, 613-624.

40. Wang, Z., Liu, Z., Zhang, W., Li, Y., Feng, Y., Lv, S., Diao, H., Luo, Z., Yan, P., He, M. *et al.* (2023) De novo design and optimization of aptamers with AptaDiff. *bioRxiv*, 2023.2011.2025.568693.

41. Ding, N., Zhang, G., Zhang, L., Shen, Z., Yin, L., Zhou, S. and Deng, Y. (2023) Engineering an AI-based forward-reverse platform for the design of cross-ribosome binding sites of a transcription factor biosensor. *Computational and Structural Biotechnology Journal*, **21**, 2929-2939.

42. Kingma, D.P. and Ba, J. (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

43. Schneider, A., Hommel, G. and Blettner, M. (2010) Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Dtsch Arztebl Int*,

107, 776-782.

44. Natekin, A. and Knoll, A. (2013) Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, **7**, 21.

45. Ridgeway, G., Madigan, D. and Richardson, T.S. (1999), *Seventh International Workshop on Artificial Intelligence and Statistics*. PMLR.

46. Breiman, L. (2001) Random forests. *Machine learning*, **45**, 5-32.

47. McDonald, G.C. (2009) Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, **1**, 93-100.

48. Drucker, H., Burges, C.J., Kaufman, L., Smola, A. and Vapnik, V. (1996) Support vector regression machines. *Advances in neural information processing systems*, **9**.

49. Chen, T. and Guestrin, C. (2016), *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794.

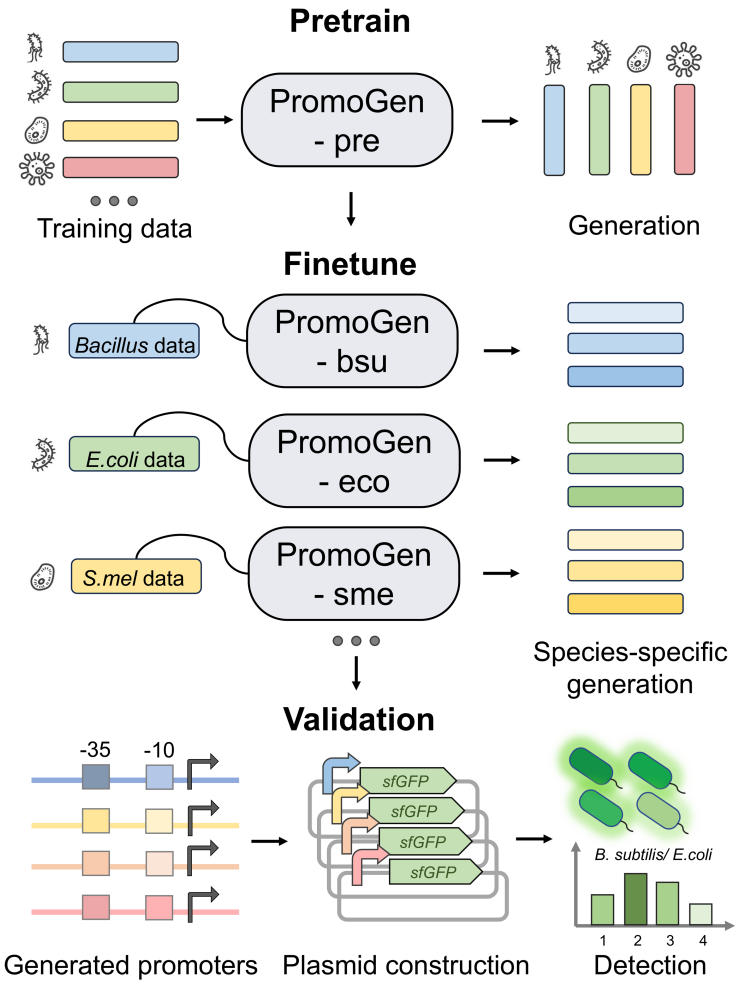
50. Loh, W.Y. (2011) Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, **1**, 14-23.

51. Willmott, C.J. and Matsuura, K. (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, **30**, 79-82.

52. Akoglu, H. (2018) User's guide to correlation coefficients. *Turkish journal of emergency medicine*, **18**, 91-93.

Figures

TOC


Pretrain
PromoGen - pre
Training data
Generation
Finetune
Bacillus data
PromoGen - bsu
E.coli data
PromoGen - eco
S.mel data
PromoGen - sme
Species-specific generation
Validation
-35
-10
sfGFP
sfGFP
sfGFP
sfGFP
B. subtilis/ E.coli
1
2
3
4
Generated promoters
Plasmid construction
Detection

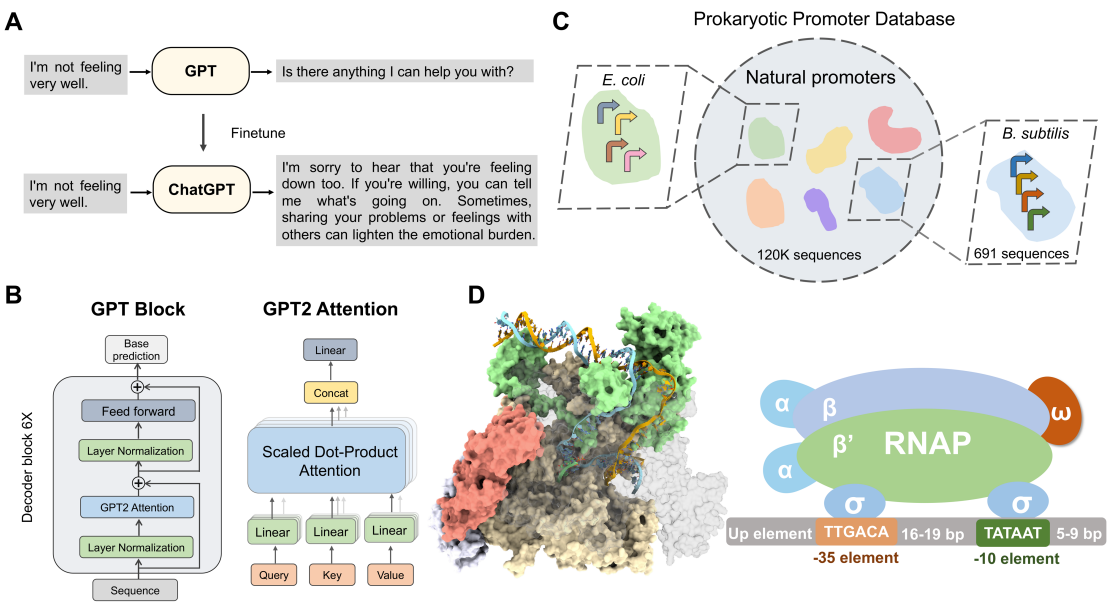


Figure 1. The framework of synthetic promoter design based on nucleotide language model. (A) A pretrained GPT model is capable of generating accurate sentences but may not always align with specific intentions or contexts. By finetuning a pretrained GPT, it can be adapted to not only produce correct sentences but also ensure alignment with the desired intent or context-specific requirements. (B) PromoGen-pre is a 7.4 million parameter neural network based on Transformer architecture, which relies on the self-attention mechanism to capture the distribution of tokens conditioned on context. (C) The PromoGen-pre is trained to generate artificial nucleotide sequence by minimizing the cross-entropy loss on the PPD database. (D) A diagrammatic representation depicts the formation of a transcription initiation complex, comprising promoters and RNA polymerase (RNAP). Promoters are primarily composed of up element, -35 element, flanking sequences, and -10 element. During the assembly of the transcription initiation complex, the α subunit is responsible for anchoring RNAP to the promoter region. The β and β' subunits are involved in recognizing the sigma factor, while the ω subunit assists in the assembly of RNA polymerase.

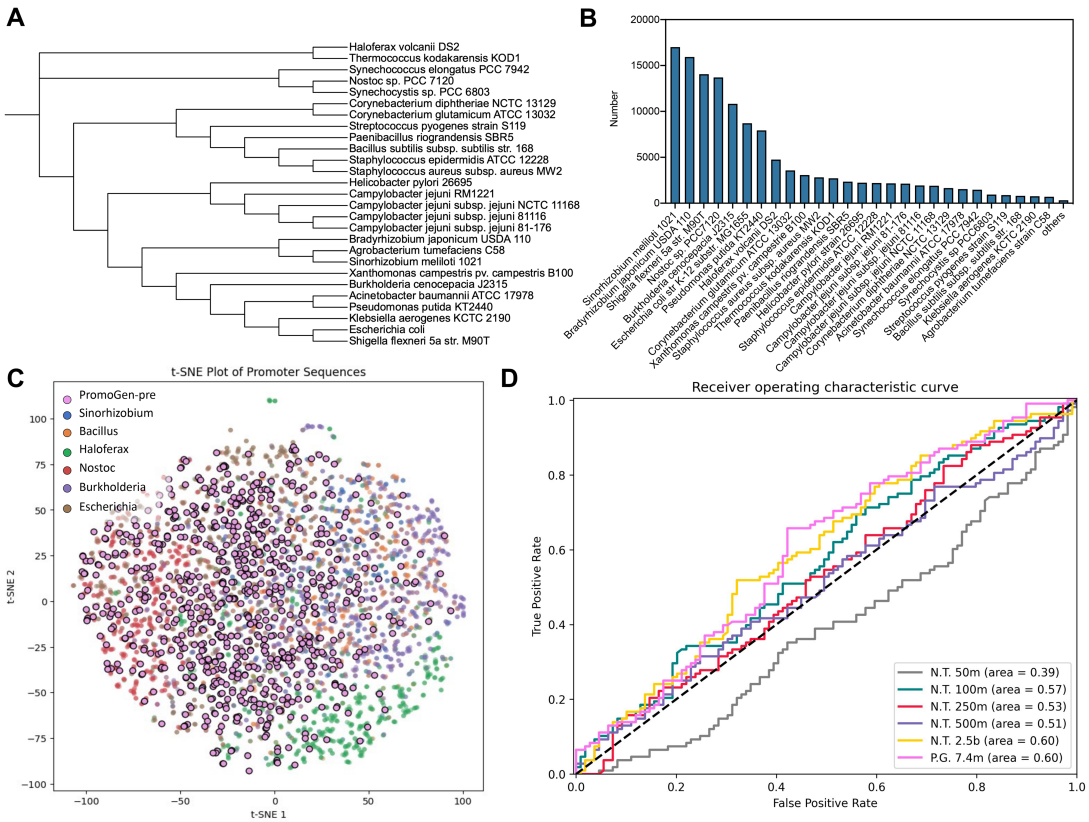


Figure 2. The training of PromoGen-pre. (A) Evolutionary tree of 27 major species included in PPD. (B)The number of promoter sequences contained in each species within PPD. The others indicate the sum of promoters from species except the 27 major ones. (C) The t-SNE, employed as a dimensionality reduction technique, was utilized to illustrate the spatial distribution of both native and generated promoters in a two-dimensional space. The distances between points represent the similarity. Blue, orange, green, red, purple, and grey color points represent the promoter sequences of *Sinorhizobium*, *Bacillus*, *Haloferax*, *Nostoc*, *Burkholderia*, and *Escherichia*, respectively. The pink color points represent the generated sequences of PromoGen. (D) The model likelihoods of PromoGen-pre can accurately predict the activity of generated promoters compared with different parameter sizes of pre-trained nucleotide transformers.

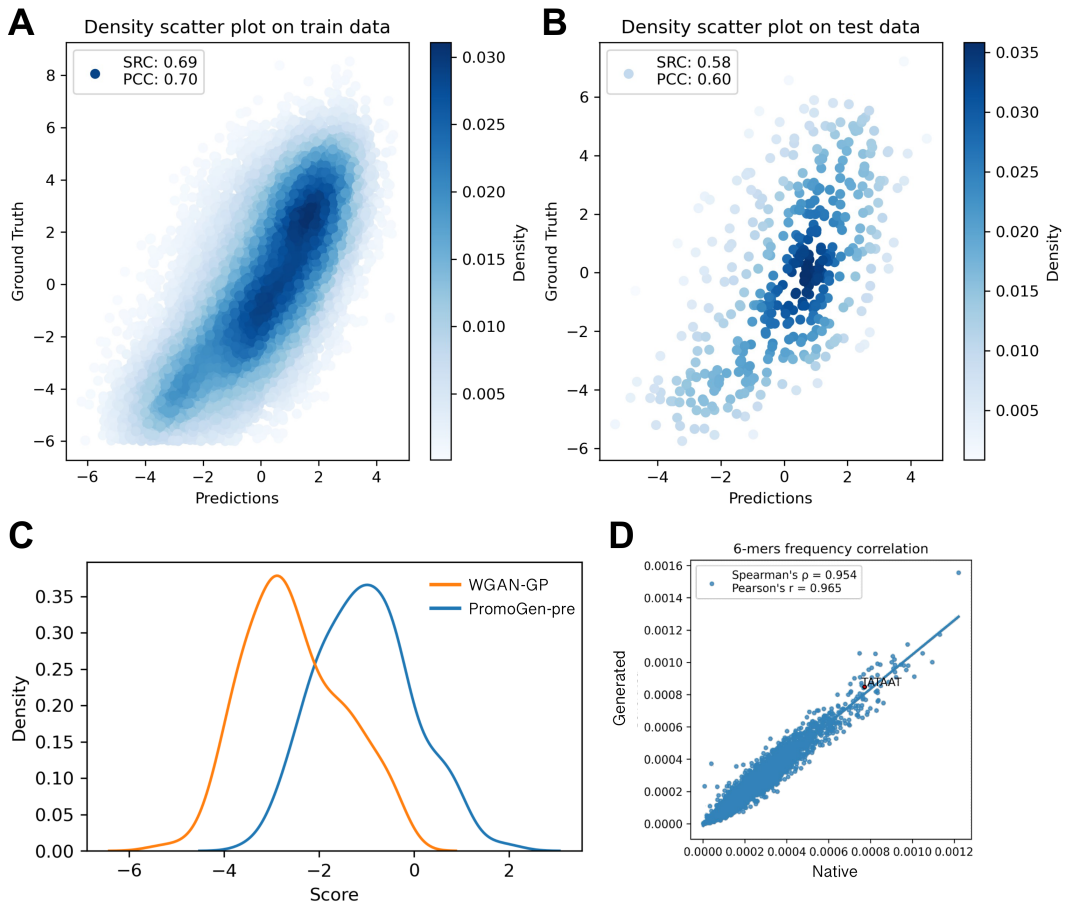


Figure 3. Comparison of PromoGen-eco with WGAN-GP model. (A) Density scatter plot on train data by Bayesian Bridge prediction model. The depth of color points represents density, with the x-axis and y-axis indicating predicted and actual promoter strengths, respectively. (B) Density scatter plot on test data by Bayesian Bridge prediction model. (C) The distribution of scores in PromoGen-eco and WGAN-GP. The orange and blue color linear represent WGAN-GP and PromoGen-eco, respectively. (D) The 6-mers frequency correlation between native and PromoGen-eco generated promoter sequences. Each data point represents a specific 6-mers, with the x-axis and y-axis indicating the frequency of the 6-mer in native and generated promoters, respectively.

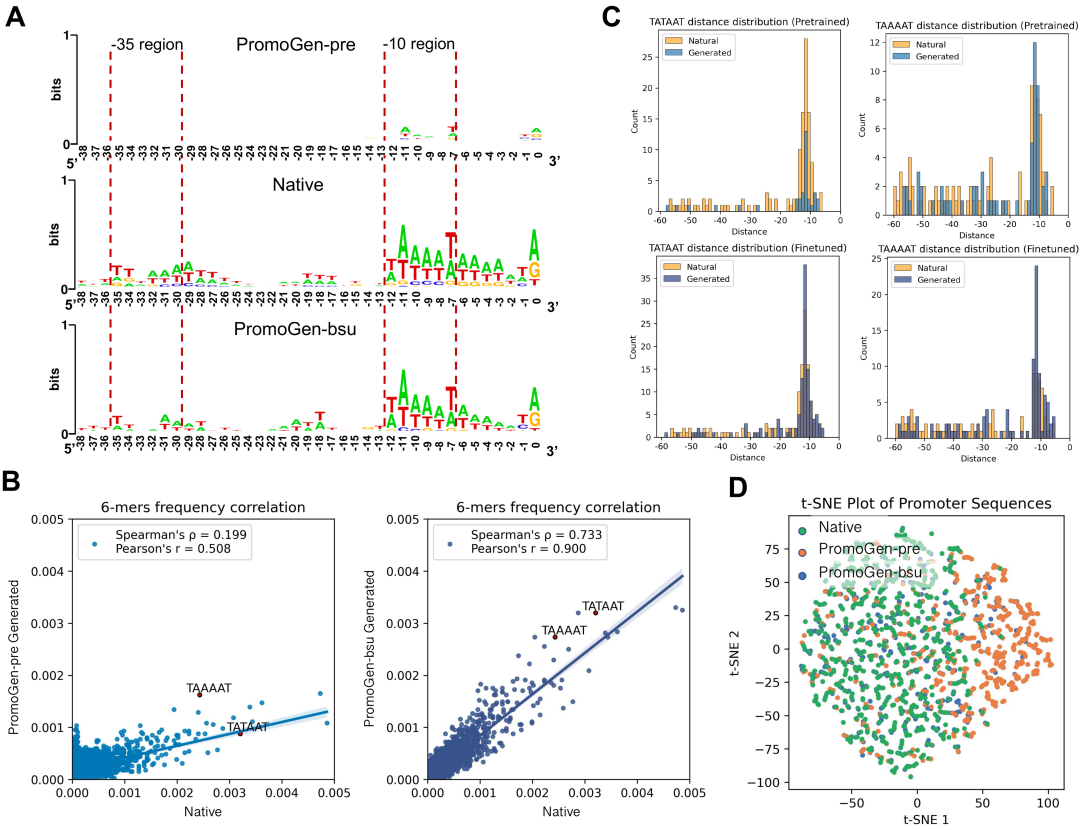


Figure 4. Evaluation of the PromoGen-pre and PromoGen-bsu generated promoters. (A) The sequence logos of the PromoGen-pre-generated, native, and PromoGen-bsu-generated promoters. The 0 represents the transcriptional start site (TSS). The key regions -10 and -35 have been marked by red dotted lines. (B) Spearman and Pearson values were used to assess the frequency of 6-mers between native and generated promoters. Each data point represents a specific 6-mers, with the x-axis and y-axis indicating the frequency of the 6-mer in native and generated promoters, respectively. (C) The distribution patterns of the top two 6-mers in PromoGen-pre and PromoGen-bsu within the -10 regions. The x-axis represents the distance from the TSS, and the y-axis represents the count of the 6-mer at that specific location. (D) The native and generated promoters are visualized in a two-dimensional space, with high-dimensional

data condensed through t-SNE. The distances between points reflect the degree of relevance. Native promoters are depicted as green dots, while PromoGen-pre and PromoGen-bsu generated promoters are represented by orange and blue dots, respectively.

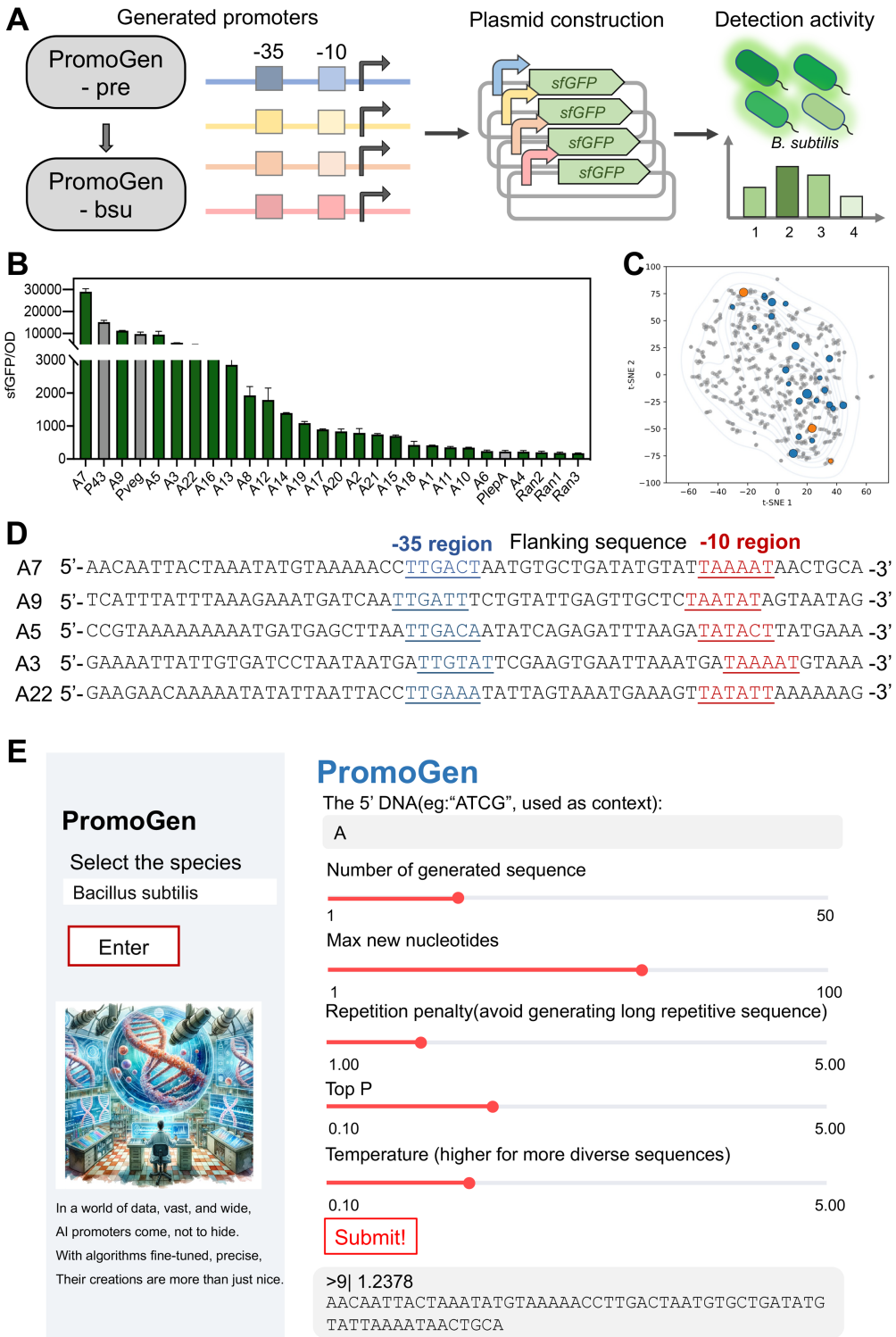


Figure 5. Validation of the generated promoter strength by *sfGFP* expression in *B. subtilis*. (A) Schematic representation of *sfGFP* expression and detection. (B) Strength of promoters were assessed by measuring the fluorescence intensity of *sfGFP* in *B.*

764 *subtilis*. Twenty-two PromoGen-bsu generated promoters were detected, alongside
765 three native promoters (P_{43}, P_{veg}, P_{lepA}) serving as positive controls, and three dummy
766 sequences as negative controls. (C) The native and validation promoters in the
767 embedding spaces. The grey dots represent the 691 native promoters in the dataset. The
768 orange and blue dots represent the native and generated validation promoters. The size
769 of the dots shows the activity of the promoter. (D) The five most active promoters in
770 the experiment were listed. Red and blue characters represent the -10 and -35 regions,
771 respectively. (E) The web server of generate promoter sequences in 27 prokaryotes.
772 When selecting the typical species, certain parameters could be adjusted, including the
773 quantity and length of generated promoter sequences, base repetition penalty, top P and
774 temperature.