Supplementary Information

Design of synthetic promoters for cyanobacteria with a generative deep-learning model

Euijin Seo[1,†], Yun-Nam Choi[1,†], Ye Rim Shin[1], Donghyuk Kim[2,3] and Jeong Wook Lee[1,4,5,*]

[1] Department of Chemical Engineering, Pohang University of Science and Technology (POSTECH), 77 Cheongam-Ro, Nam-Gu, Pohang, Gyeongbuk 37673, Korea
[2] School of Energy and Chemical Engineering, Ulsan National Institute of Science and Technology (UNIST), 50 UNIST-Gil, Eonyang-Eup, Ulsan 44919, Korea
[3] Department of Energy Engineering, Ulsan National Institute of Science and Technology (UNIST), 50 UNIST-Gil, Eonyang-Eup, Ulsan 44919, Korea
[4] School of Interdisciplinary Bioscience and Bioengineering, Pohang University of Science and Technology (POSTECH), 77 Cheongam-Ro, Nam-Gu, Pohang, Gyeongbuk 37673, Korea
[5] Graduate School of Artificial Intelligence, Pohang University of Science and Technology (POSTECH), 77 Cheongam-Ro, Nam-Gu, Pohang, Gyeongbuk 37673, Korea

* To whom correspondence should be addressed. Email: jeongwook@postech.ac.kr (J.W. Lee)

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Supplementary Figure

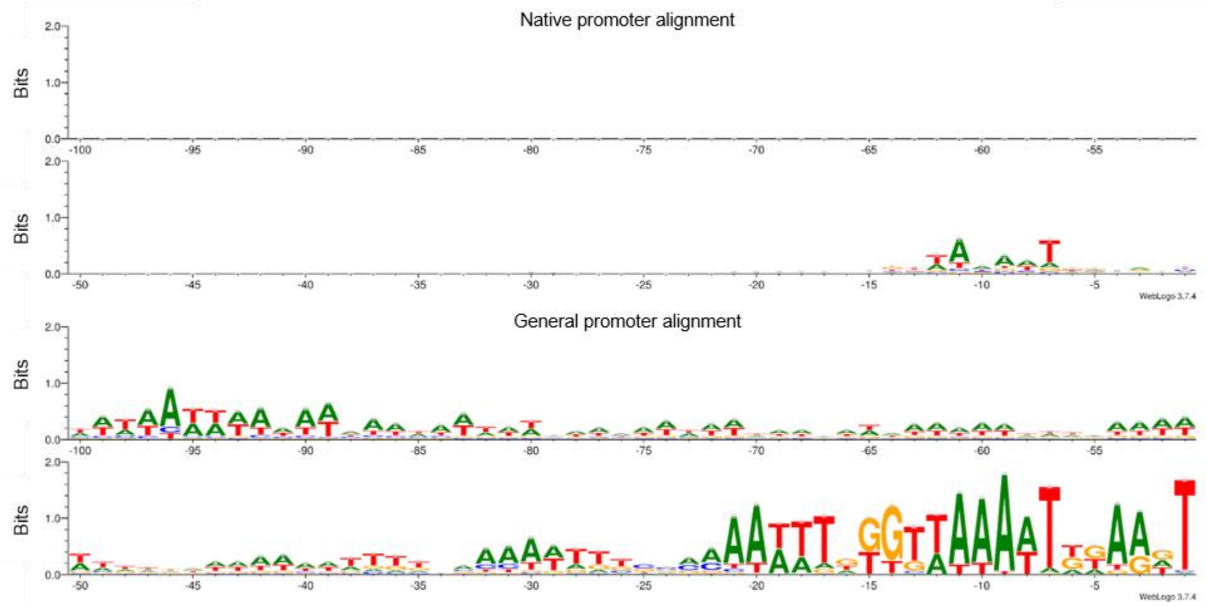

Figure S1. Sequence logos of native (top) and generated (bottom) promoters on promoter region between -100 and -1.

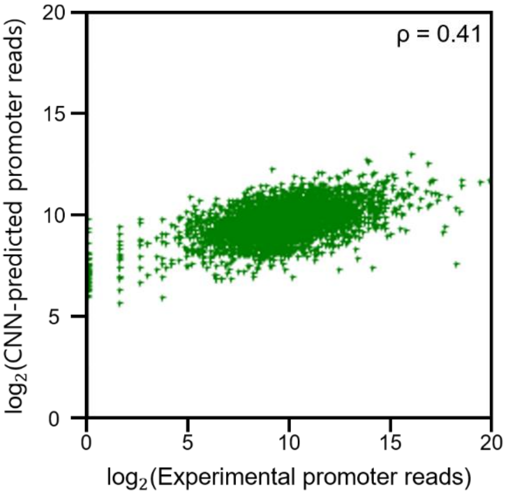

Figure S2. A scattered plot of experimental promoter reads versus CNN-predicted promoter reads. The x-axis represents experimental promoter reads, and the y-axis represents CNN-predicted promoter reads. ρ: Pearson correlation coefficient between the experimental- and the CNN-predicted promoter reads.

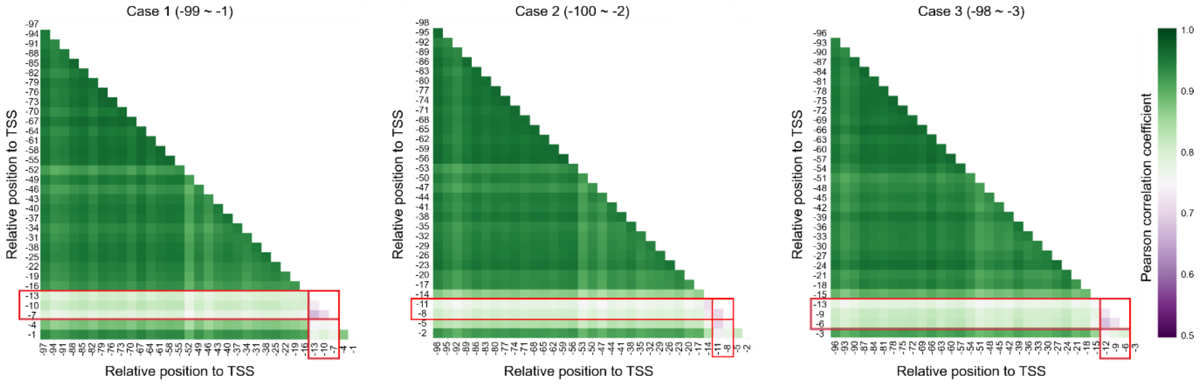

Figure S3. Heatmap showing the Pearson correlation coefficient between the predicted promoter strengths before and after mutations (Case 1: -99 to -1, Case 2: -100 to -2 Case 3: -98 to -3, respectively). The red box represents the subregion with the dramatic changes in the Pearson correlation coefficient.

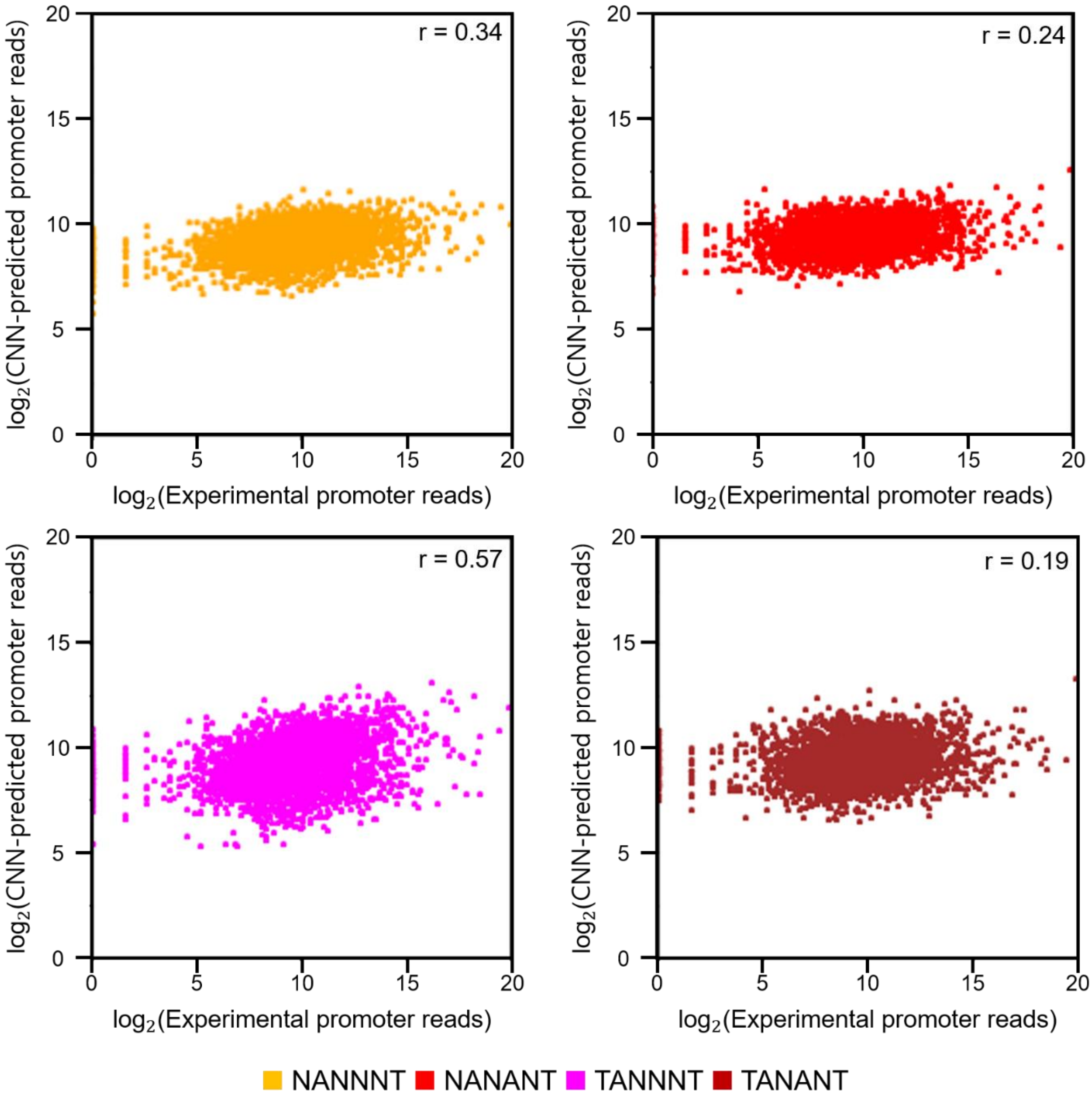

Figure S4. A scattered plot of experimental promoter reads versus predicted promoter reads by CNN trained with differently refined promoter datasets. Each dataset is composed of native promoter sequences that contain NANNNT, NANANT, TANNNT, or TANANT 6-mers at -13 and -6 regions, respectively. r: Pearson correlation coefficient between the experimental- and the CNN-predicted promoter reads.

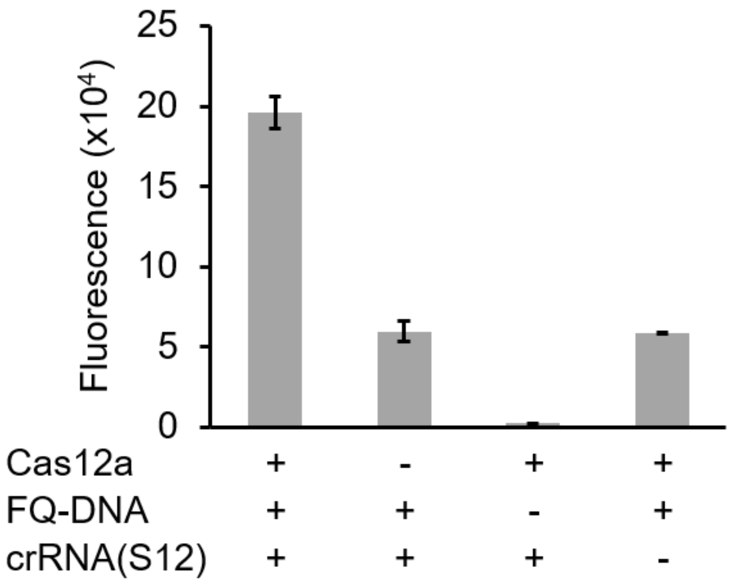

Figure S5. Control experiments for CRISPR/Cas12a-based nucleic acid detection.

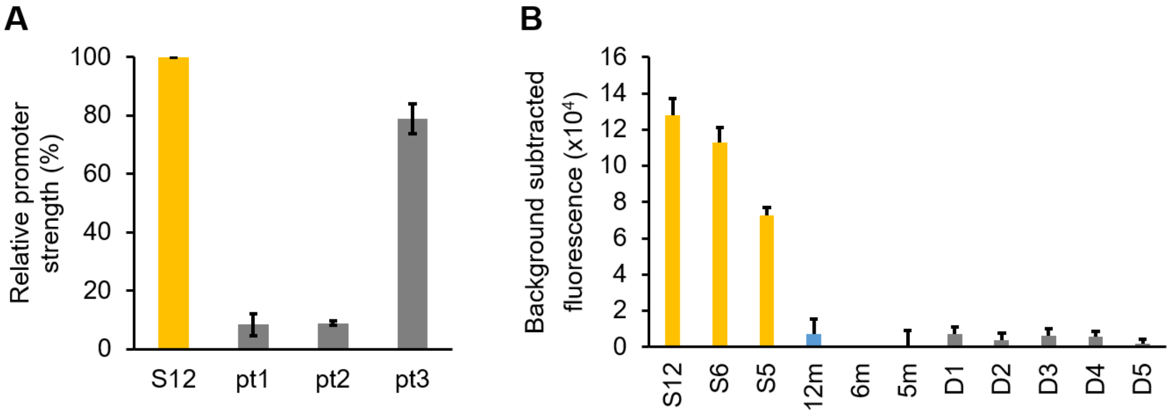

Figure S6. Validation of a single -10 box sequence motif on the synthetic promoters**.** (A) No transcriptional activity of a partial sequence derived from S12 promoter. pt1: 50-bp partial sequence from -100 to -51, pt2: 50-bp partial sequence from -75 to -24, pt3: 50-bp partial sequence from -50 to -1. (B) Replacement of the -10 element of S12, S6, and S5 promoter with the corresponding sequences of D1, D2, and D3 (12m, 6m, 5m).

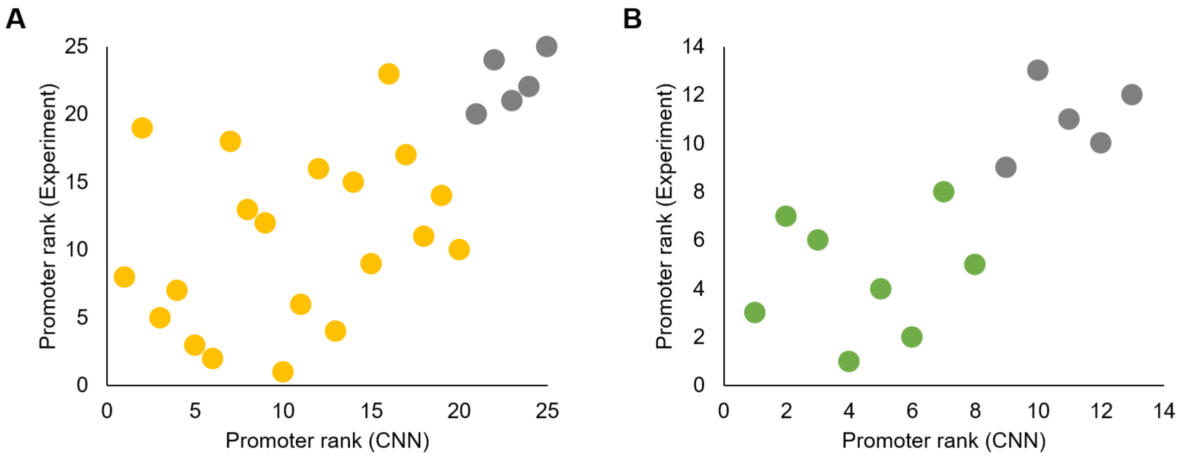

Figure S7. Comparison of the CNN-predicted promoter strength and experimental promoter strength. (A) Scatter plot of the rank orders of CNN-predicted promoter strength versus the rank orders of their *in vitro* experimental results. The x-axis represents the rank orders of CNN-predicted promoter strength, and the y-axis represents the rank orders of *in vitro* experimental results. The yellow dots represent the rank orders of the top 20 synthetic promoters; grey dots represent the rank orders of the five dummy sequences. (B) Scatter plot of the rank orders of CNN-predicted promoter strength versus the rank orders of their *in vivo* experimental results. The x-axis represents the rank orders of CNN-predicted values, and the y-axis represents the rank orders of *in vivo* experimental results. The yellow dots represent the rank orders of the eight synthetic promoters; grey dots represent the rank orders of the five dummy sequences.