Predicting function of evolutionarily implausible DNA sequences

Shiyu Jiang [1 2] **Xuyin Liu** [1] **Zitong Jerry Wang** [1]

Abstract

Genomic language models (gLMs) show potential for generating novel, functional DNA sequences for synthetic biology, but doing so requires them to learn not just evolutionary plausibility, but also sequence-to-function relationships. We introduce a set of prediction tasks called NULLSETTES, which assesses a model's ability to predict loss-of-function mutations created by translocating key control elements in synthetic expression cassettes. Across 12 state-of-the-art models, we find that mutation effect prediction performance strongly correlates with the predicted likelihood of the nonmutant. Furthermore, the range of likelihood values predictive of strong model performance is highly dependent on sequence length. Our work highlights the importance of considering both sequence likelihood and sequence length when using gLMs for mutation effect prediction.

1. Introduction

Genomic language models (gLMs) learn a probability distribution over DNA sequences, representing the evolutionary plausibility of genomic sequences (Benegas et al., 2025b; Consens et al., 2025). For example, gLMs would assign higher likelihood values for regulatory motifs that occur frequently in natural genomic sequences compared to random sequences.

Evolutionary plausibility can be a useful proxy for biological function. In both genomic and protein language models (gLMs and pLMs), higher model likelihoods often correlate with improved functional properties. For example, pLM likelihoods can guide the design of higher affinity antibodies (Hie et al., 2024), more efficient base editors (He et al., 2024), and brighter fluorescent proteins (Zhang et al., 2025). Similarly, gLM likelihoods enable genome-wide prediction of variant effects (Benegas et al., 2023) and 5' untranslated region (UTR) optimization (Chu et al., 2024).

However, current evaluations for gLMs on mutant effect prediction focus primarily on natural sequences. In contrast, synthetic biology often requires the design of functional sequences with little or no evolutionary precedent, such as sequences that confer novel functions, avoid crosstalk with native machinery, or push expression levels beyond natural limits. Examples include ultra-strong synthetic promoters (Schlabach et al., 2010), miRNA-based regulatory circuits that achieve dosage-compensated gene expression using elements orthogonal to native miRNAs (Du et al., 2024), and engineered metabolic pathways for production of small molecule drugs (Yan et al., 2023). For these applications, it is unclear whether gLMs trained on natural genomes can generalize to synthetic constructs, especially in the absence of deep mutational scanning data for DNA. If successful, gLMs could support more systematic and scalable approaches to genetic design.

We introduce a benchmark suite, NULLSETTES, to evaluate gLMs on their ability to predict loss-of-function (LOF) mutations in synthetic expression cassettes. NULLSETTES leverages well-established mechanisms of gene expression to systematically introduce virtual LOF mutations by rearranging key control elements, such as promoters and start codons, within functional cassettes. Using functional cassettes (referred to as nonmutant) curated from Massive Parallel Reporter Assay (MPRA) data (Lagator et al., 2022; de Boer et al., 2020; Kosuri et al., 2013; Zahm et al., 2024), including those with randomly generated, low-likelihood promoters, we show that Evo-2-7B outperforms 11 other state-of-the-art models in mutant effect prediction. However, all gLMs exhibit a sharp decline in predictive accuracy as the likelihood of the nonmutant sequence decreases. Moreover, the likelihood range at which each model performs optimally varies strongly with sequence length. These results suggest that both likelihood and length must be considered when applying gLMs to guide genetic design.

[1]Center for Interdisciplinary Studies, School of Science, Westlake University, Hangzhou, China [2]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA. Correspondence to: Shiyu Jiang <shiyujia@usc.edu>, Zitong Jerry Wang <jerry@westlake.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

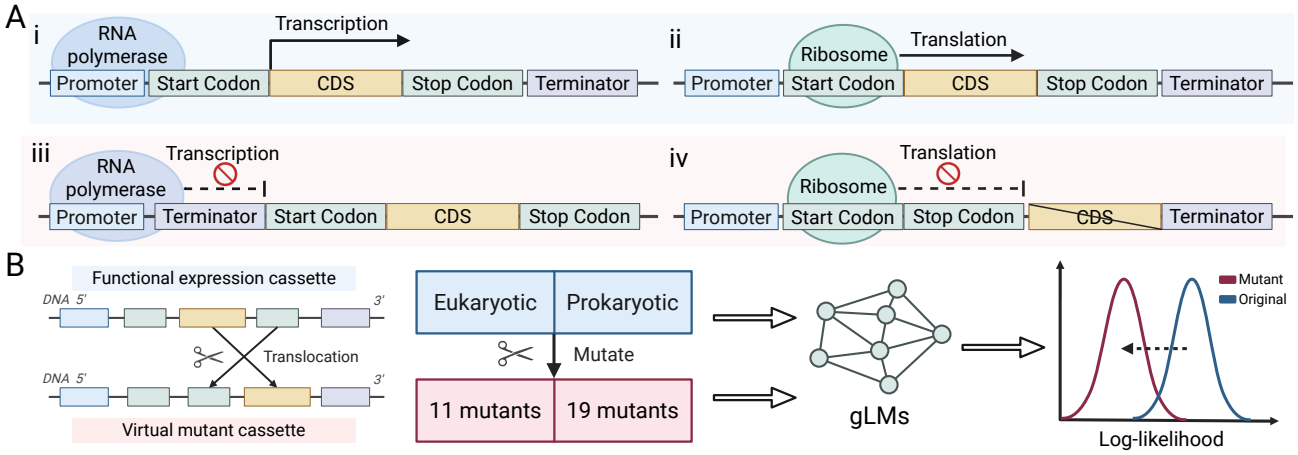


Figure 1. **Overview of NULLSETTES.** A) Schematic examples showing transcription (i) and translation (ii) becoming impaired when the order of control element such as terminator (iii) or stop codon (iv) are altered. B) NULLSETTES generates 11 and 19 mutant variants for each eukaryotic and prokaryotic expression cassette, respectively, by translocating control elements, and evaluates how well gLMs can predict these loss-of-function mutations based on changes in log-likelihood.

2. Methods

2.1. NULLSETTES Construction

To systematically evaluate gLMs' ability for zero-shot functional prediction on expression cassettes, we proposed NULLSETTES, a suite of virtual mutants that perturb transcription and translation of an expression cassette by shuffling the order of six key control elements. These elements are promoter, ribosome binding site (RBS), start codon, coding sequence (CDS), stop codon, and terminator.

In the canonical configuration of an expression cassette (Figure 1A i-ii), the proper ordering of promoter–start codon–CDS–stop codon–terminator from 5' to 3' end enables proper transcription (i) and translation (ii). However, translocations that reposition critical elements, such as placing the terminator between the promoter and the CDS (iii) or the CDS downstream of the stop codon (iv), result in transcriptional or translational failure, respectively.

NULLSETTES consists of single-element translocations that effectively eliminates any production of functional mRNA and protein (Figure 1B). Among the set of all possible single-element translocations, we subset for mutants where 1) the CDS expression cannot be rescued by any flanking sequences outside the cassette, 2) expression machinery can still act to generate useless products. These virtual mutant cassettes collectively make up the NULLSETTES, which consists of 11 mutations for eukaryotic and 19 mutations for prokaryotic cassettes (see Appendix Table 3 and 4 for full list). The 8 additional mutants for prokaryote cassettes are due to translocating the RBS. Given a functional cassette, we then compare its gLM log-likelihood (LL) with its corresponding NULLSETTES to assess how well a gLM can perform zero-shot functional prediction by consistently predicting lower likelihoods for the NULLSETTES.

2.2. Functional cassette curation

To assess mutant effect prediction using NULLSETTES, we curate expression cassettes from public MPRA datasets (Lagator et al., 2022; de Boer et al., 2020; Kosuri et al., 2013; Zahm et al., 2024). We specifically use cassettes with strong gene expression but low gLM likelihood representing low evolutionary plausibility. These cassettes have low likelihood due to containing elements from evolutionarily distant organisms. For example, the CDS, green fluorscent protein (GFP), is from the jellyfish *Aequorea victoria*, whereas promoter and terminators are sourced from bacteria, human and mice. Furthermore, cassettes curated from Lagator dataset and deBoer pTpA dataset use random DNA sequences as promoters, which further lowers gLM likelihood compared to those with natural promoters/motifs (See Appendix A.3). In the paper, we name datasets with random promoters with the suffix ("Low") and those with more natural promoters with the suffix ("High") (Figure 2).

2.3. Baseline Models

We benchmarked a diverse set of 12 self-supervised genomic foundation models that represent current state-of-the-art approaches to DNA language modeling. In general, the models can be categorized based on their tokenization schemes (fixed-length k-mers, single-nucleotide tokens, byte-pair encoding, and hybrid schemes), pretraining strategies (masked language modeling (Devlin et al., 2019) vs. autoregressive modeling (Brown et al., 2020)), training corpus diversity (human references, plant genomes, multispecies genomes,

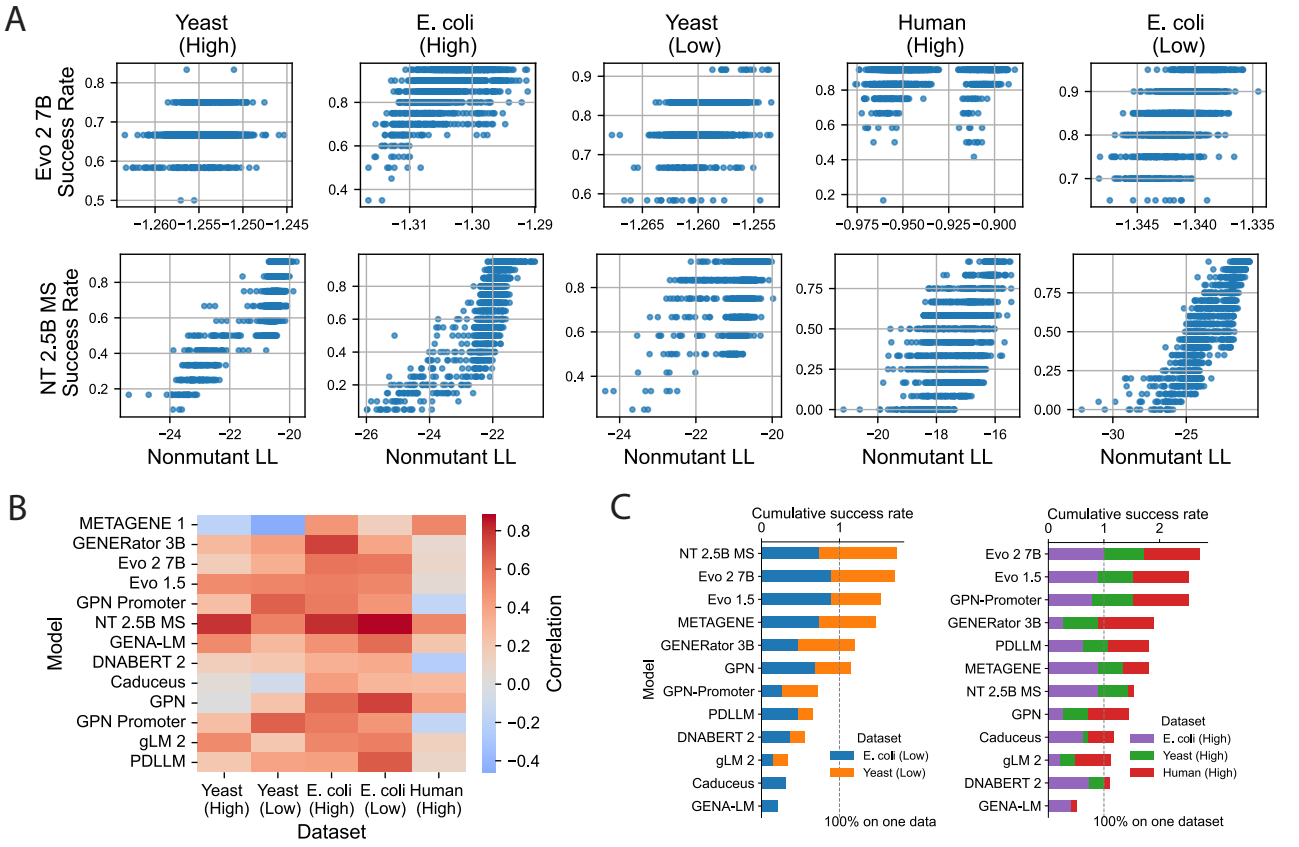


Figure 2. **Relationship between gLM likelihood and zero-shot performance on NULLSETTES prediction.** A) Each point represents an expression cassette. Scatterplots show the NULLSETTES prediction performance of Evo-2-7B and Nucleotide Tranformer-2.5B-MS for sequences from five datasets. B) Heatmap showing correlation between the gLM log likelihood (LL) of a sequence and gLM success rate on NULLSETTES prediction, across five datasets and 12 models. C) Stacked bar plot showing success rate of models computed using all sequences in each dataset, representing the proportion of NULLSETTES mutants that a model consistently outputs a lower LL compared to the nonmutant.

and large-scale metagenomic assemblies), and architectural paradigms (CNN-based models, transformer-based models (Consens et al., 2025), and emerging architectures like StripedHyena (Poli et al., 2023; Ku et al., 2025) and Mamba (Gu & Dao, 2023)). Detailed model specifications are concluded in Appendix Table 1 and 2.

2.4. Evaluation Metrics

We computed the mean base-pair log-likelihood (LL) as the sequence-level LL score. To account for the distinct pretraining objectives of causal language models (CLMs) and masked language models (MLMs), we applied different LL computation strategies, as detailed in Appendix B.1. After obtaining the LL distributions for both the original cassette and its virtual mutants (NULLSETTES), we performed a one-sided paired permutation test to assess whether the mutant cassettes exhibit a significant decrease in LL compared to the original sequence (Figure 1B). Additional methodological details are provided in Appendix B.2.

3. Results

We benchmarked these gLMs and their variants across five datasets, as shown in Appendix C. Representative models were selected for further analysis in the following sections.

3.1. Zero-shot NULLSETTES prediction performance deteriorates for sequences with low gLM likelihood

The evolutionary plausibility of a sequence, as predicted by a gLM, strongly correlates with the gLM's performance in predicting mutation effects for the sequence. Figure 2A compares gLM prediction performance of individual functional cassettes and their predicted LL for two models, Evo-2-7B and NucleotideTransformer-2.5B-MS. Here success rate for each cassette is defined as the proportion of NULLSETTES (11 in eukaryote, 19 in prokaryote) whose LL score is lower

than the original, nonmutant cassette. For both models, we observe strong correlations between the nonmutant LL score and the gLM's ability to identify LOF mutants. For example, for the E.coli (Low) dataset (Lagator et al., 2022), zero-shot NULLSETTES prediction with NT-2.5B-MS is nearly impossible when the nonmutant LL is below -25, and easily accomplished for nonmutants with LL above -22. In fact, Figure 2B shows that all models, despite variations in architecture and training, exhibits a positive correlation between sequence likelihood and mutant prediction performance.

The Evo-2-7B model achieves the best overall performance in zero-shot NULLSETTES prediction. Evaluating zero-shot prediction using all sequences within a dataset (See Methods), Figure 2C shows that Evo-2-7B achieves the best overall performance across all five datasets. Specifically for the E.coli and yeast MPRA data using random promoters (left), which are less evolutionarily plausible than more natural sequences shown on the right (Appendix Figure 4).

3.2. Optimal likelihood range predictive of strong model performance varies with sequence length

Although mutant prediction performance correlates with sequence likelihood, the range of LL scores for which a model performs well depends highly on the length of the DNA sequence. The range of LL scores for which models achieve strong performance is higher for longer sequences, partly due to the fact that gLMs tend to assign higher LL scores to longer cassettes (Figure 3A). For example, Figure 3B shows that the NT-2.5B-MS model achieves strong performance for cassettes with log likelihood above -22 for both E. coli datasets, as indicated by the gray dotted line. However, for sequences from the Human GPRA dataset which are 2.5-folds longer, only LL score above -17 achieved strong performance. Similarly sequences from the yeast (low) dataset are 30% longer compared to sequences from the E. coli datasets. Thus, for GPN-promoter and Evo-1.5, the optimal likelihood range is also significant higher for the yeast (low) dataset.

Altogether, there does not appear to be a single optimal likelihood range within which a model would be able to make zero-shot mutation effect predictions, as sequence properties such as length can have a significant effect on optimal likelihood range.

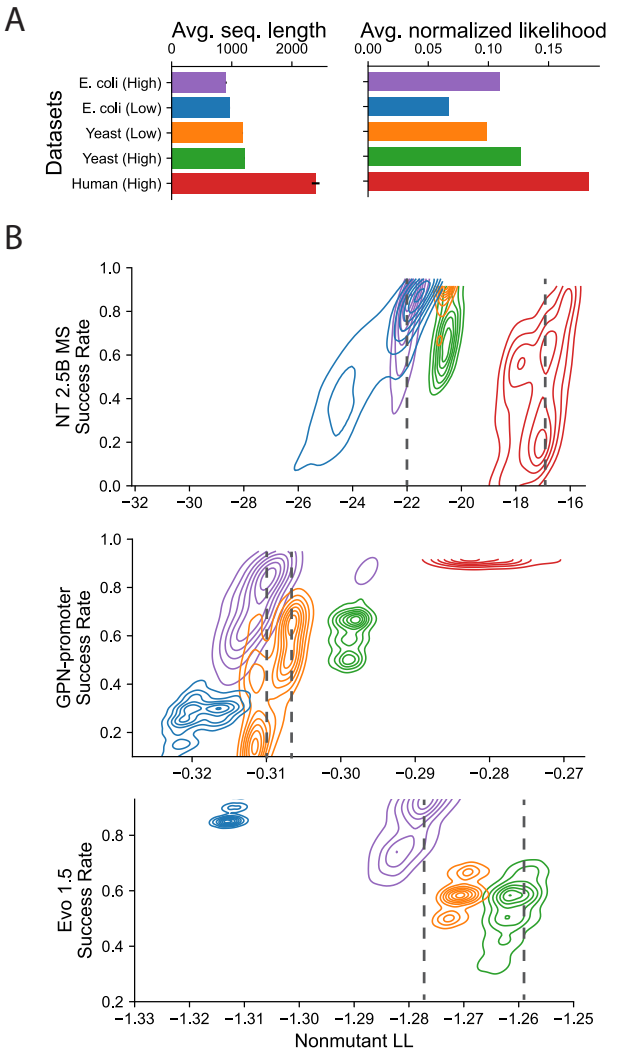


Figure 3. **Optimal likelihood range for NULLSETTES prediction varies with sequence length.** A) Length in base pairs and normalized gLM log likelihood of all nonmutant cassettes, averaged across all 12 models. B) Contour of scatter plots similar to those in Figure 2A, colored by dataset, dotted vertical lines are rough estimates of optimal LL value for sequences from different datasets. Human data for Evo 1.5 is cut off for better visual due to large LL score.

4. Discussion

In this study, we introduce NULLSETTES, the first systematic benchmark for assessing the ability of genomic language models (gLMs) to predict loss-of-function mutations in synthetic expression cassettes. By enabling simple in silico translocation of sequences, NULLSETTES facilitates generalizable, cassette-independent evaluations of gLM functional performance. Beyond general performance benchmarking, our results reveal a strong correlation between gLM performance on mutation prediction tasks and the model's predicted likelihood of the original, non-mutant sequence. Moreover, we demonstrate that the optimal likelihood range for accurate mutant discrimination is not fixed but varies significantly with sequence length. Together, we hope this benchmark and its findings will inspire the development of gLMs that move beyond evolutionary priors to support the rational design of functional, out-of-distribution genetic sequences.

Software and Data

The source code for benchmarking gLM on NULLSETTES is publicly available at: https://github.com/cellethology/GLM-Nullsette-Benchmark.

Acknowledgements

The authors thank members of the Cell Ethology Lab for valuable comments and suggestions.

Impact Statement

This paper presents work whose goal is to advance the field of genomic language model and its application in synthetic biology. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Benegas, G., Batra, S. S., and Song, Y. S. Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44):e2311219120, 2023.

Benegas, G., Eraslan, G., and Song, Y. S. Benchmarking dna sequence models for causal regulatory variant prediction in human genetics. *bioRxiv*, pp. 2025–02, 2025a.

Benegas, G., Ye, C., Albors, C., Li, J. C., and Song, Y. S. Genomic language models: opportunities and challenges. *Trends in Genetics*, 2025b.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., et al. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, pp. 2025–02, 2025.

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

Chu, Y., Yu, D., Li, Y., Huang, K., Shen, Y., Cong, L., Zhang, J., and Wang, M. A 5′ utr language model for decoding untranslated regions of mrna and function predictions. *Nature Machine Intelligence*, 6(4):449–460, 2024.

Consens, M. E., Dufault, C., Wainberg, M., Forster, D., Karimzadeh, M., Goodarzi, H., Theis, F. J., Moses, A., and Wang, B. Transformers and genome language models. *Nature Machine Intelligence*, pp. 1–17, 2025.

Cornman, A., West-Roberts, J., Camargo, A. P., Roux, S., Beracochea, M., Mirdita, M., Ovchinnikov, S., and Hwang, Y. The omg dataset: An open metagenomic corpus for mixed-modality genomic language modeling. *bioRxiv*, pp. 2024–08, 2024.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, pp. 1–11, 2024.

de Boer, C. G., Vaishnav, E. D., Sadeh, R., Abeyta, E. L., Friedman, N., and Regev, A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature biotechnology*, 38(1):56–65, 2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Du, R., Flynn, M. J., Honsa, M., Jungmann, R., and Elowitz, M. B. miRNA circuit modules for precise, tunable control of gene expression. *BioRxiv*, 2024.

Fishman, V., Kuratov, Y., Shmelev, A., Petrov, M., Penzar, D., Shepelin, D., Chekanov, N., Kardymon, O., and Burtsev, M. Gena-lm: a family of open-source foundational dna language models for long sequences. *Nucleic Acids Research*, 53(2):gkae1310, 2025.

Gordon, C., Lu, A. X., and Abbeel, P. Protein language model fitness is a matter of preference. *bioRxiv*, pp. 2024–10, 2024.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

He, Y., Zhou, X., Chang, C., Chen, G., Liu, W., Li, G., Fan, X., Sun, M., Miao, C., Huang, Q., et al. Protein language models-assisted optimization of a uracil-n-glycosylase variant enables programmable t-to-g and t-to-c base editing. *Molecular Cell*, 84(7):1257–1270, 2024.

Hie, B. L., Shanker, V. R., Xu, D., Bruun, T. U., Weidenbacher, P. A., Tang, S., Wu, W., Pak, J. E., and Kim, P. S. Efficient evolution of human antibodies from general protein language models. *Nature biotechnology*, 42 (2):275–283, 2024.

Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., Endy, D., and Church, G. M. Composability of regulatory sequences controlling transcription and translation in escherichia coli. *Proceedings of*

the National Academy of Sciences, 110(34):14024–14029, 2013.

Ku, J., Nguyen, E., Romero, D. W., Brixi, G., Yang, B., Vorontsov, A., Taghibakhshi, A., Lu, A. X., Burke, D. P., Brockman, G., et al. Systems and algorithms for convolutional multi-hybrid language models at scale. *arXiv preprint arXiv:2503.01868*, 2025.

Lagator, M., Sarikas, S., Steinrueck, M., Toledo-Aparicio, D., Bollback, J. P., Guet, C. C., and Tkačik, G. Predicting bacterial promoter function and evolution from random sequences. *Elife*, 11:e64543, 2022.

Liu, G., Chen, L., Wu, Y., Han, Y., Bao, Y., and Zhang, T. Pdllms: A group of tailored dna large language models for analyzing plant genomes. *Molecular Plant*, 18(2): 175–178, 2025a.

Liu, O., Jaghouar, S., Hagemann, J., Wang, S., Wiemels, J., Kaufman, J., and Neiswanger, W. Metagene-1: Metagenomic foundation model for pandemic monitoring. *arXiv preprint arXiv:2501.02045*, 2025b.

Merchant, A. T., King, S. H., Nguyen, E., and Hie, B. L. Semantic mining of functional de novo genes from a genomic language model. *bioRxiv*, pp. 2024–12, 2024.

Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D. B., Bartie, L. J., Thomas, A. W., King, S. H., Brixi, G., et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723): eado9336, 2024.

Poli, M., Wang, J., Massaroli, S., Quesnelle, J., Carlow, R., Nguyen, E., and Thomas, A. StripedHyena: Moving Beyond Transformers with Hybrid Signal Processing Models, 12 2023. URL https://github.com/togethercomputer/stripedhyena.

Schiff, Y., Kao, C.-H., Gokaslan, A., Dao, T., Gu, A., and Kuleshov, V. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*, 2024.

Schlabach, M. R., Hu, J. K., Li, M., and Elledge, S. J. Synthetic design of strong promoters. *Proceedings of the national academy of sciences*, 107(6):2538–2543, 2010.

Wang, A. and Cho, K. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019.

Wu, W., Li, Q., Li, M., Fu, K., Feng, F., Ye, J., Xiong, H., and Wang, Z. Generator: A long-context generative genomic foundation model. *arXiv preprint arXiv:2502.07272*, 2025.

Yan, X., Liu, X., Zhao, C., and Chen, G.-Q. Applications of synthetic biology in medical and pharmaceutical fields. *Signal transduction and targeted therapy*, 8(1):199, 2023.

Zahm, A. M., Owens, W. S., Himes, S. R., Fallon, B. S., Rondem, K. E., Gormick, A. N., Bloom, J. S., Kosuri, S., Chan, H., and English, J. G. A massively parallel reporter assay library to screen short synthetic promoters in mammalian cells. *Nature Communications*, 15(1): 10353, 2024.

Zhang, Q., Chen, W., Qin, M., Wang, Y., Pu, Z., Ding, K., Liu, Y., Zhang, Q., Li, D., Li, X., et al. Integrating protein language models and automatic biofoundry for enhanced protein evolution. *Nature Communications*, 16(1):1553, 2025.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., and Liu, H. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

A. Datasets

A.1. Prokaryote

To evaluate the regulatory grammar of prokaryotes on genomic language models, we utilized two benchmark datasets derived from synthetic expression constructs in *Escherichia coli*. The first, the Kosuri dataset (Kosuri et al., 2013), comprises combinatorial assemblies of promoters and ribosome binding sites (RBS) upstream of a superfolder GFP (sfGFP) reporter, designed to dissect transcriptional and translational contributions to gene expression. The second, the Lagator dataset (Lagator et al., 2022), features a large-scale library of random promoter sequences with experimentally measured activity, enabling evaluation of models' ability to generalize to highly diverse, out-of-distribution regulatory inputs.

A.2. Eukaryote

To evaluate the capacity of genomic language models to generalize to eukaryotic regulatory logic, we utilized two large-scale massively parallel reporter assay (MPRA) datasets. The Zahm dataset (Zahm et al., 2024) comprises a library of 6,144 synthetic promoters constructed by combining transcriptional response elements (TREs) from 229 human and mouse transcription factors with minimal promoters. These constructs were assayed across multiple human cell lines and stimulatory conditions to quantify dynamic, stimulus-specific gene expression. The deBoer dataset (de Boer et al., 2020) includes a comprehensive collection of 100 million randomized promoter sequences in yeast, enabling high-resolution dissection of cis-regulatory grammar under out-of-distribution scenario.

A.3. Promoter selection

To construct decoupled expression cassettes, we selected 1,500 promoters or promoter–RBS pairs from four distinct datasets.

A.3.1. Kosuri and Lagator datasets

For the Kosuri dataset (Kosuri et al., 2013), we selected candidate promoter–RBS pairs with protein expression levels exceeding $\mu + 1.5\sigma$ (10,165) across the full distribution. In contrast, given the limited number of highly active constructs in the Lagator dataset, we directly selected the top 1,500 promoter sequences ranked by protein output. We then compared the log-likelihood (LL) distributions of promoter–RBS groups using genomic language models including Evo1 8k, GENERator 3B, GPN, and Nucleotide Transformer 2.5B multispecies. As the Lagator dataset (Lagator et al., 2022) consists of randomly generated promoters, whereas the Kosuri library is more rationally designed, we prioritized Kosuri promoter–RBS pairs whose LL distributions surpassed those from Lagator (Appendix Figure 4a), yielding 1,500 expression cassettes from each dataset.

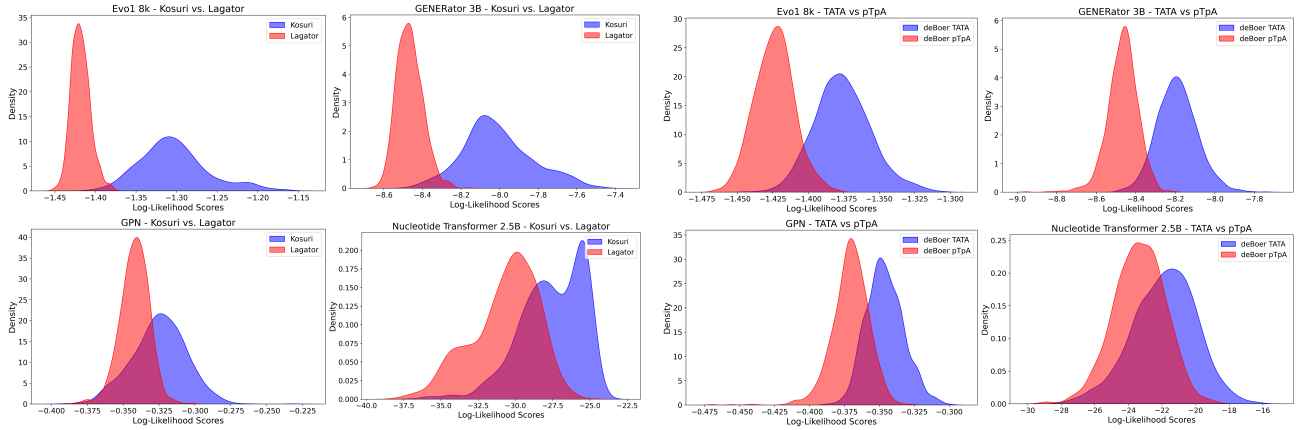
A.3.2. deBoer dataset

deBoer dataset (de Boer et al., 2020) contains two promoter libraries: Abf1TATA and pTpA. Abf1TATA is designed by embedding conserved transcription factor binding sites such as Abf1 and a canonical TATA box, mimicking features of natural yeast promoters; while pTpA consists of synthetic promoters constructed with a simple poly-T–poly-A architecture, lacking specific transcription factor motifs and serving as a minimal, randomized control library. Similar to Kosuri vs Lagator, we selectively constructed 1,500 active promoters for each dataset based on LL distribution where pTpA is left-shifted comparing to Abf1TATA as Abf1TATA is more conservative (Appendix Figure 4b). The source can be found at NCBI's GEO: GSE104878[1].

A.3.3. Zahm dataset

The Zahm dataset (Zahm et al., 2024) comprises synthetic promoters built by coupling one of three minimal promoters—minCMV, minProm, or minTK—with diverse transcriptional response element (TRE) units. To construct a representative set of mammalian expression cassettes, we selected the top 1,500 promoter-TRE combinations exhibiting the highest transcriptional activity. The source can be found at NCBI's GEO: GSE271608[2].

[1]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104878
[2]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE271608



(a) Kosuri vs Lagator

(b) deBoer TATA vs deBoer pTpA

Figure 4. Comparison of promoter log-likelihood distribution

A.4. Virtual mutant

Translocation mutants were systematically engineered by permuting the canonical order of key regulatory and coding elements—promoter, ribosome binding site (RBS), start codon, coding sequence (CDS), stop codon, and terminator in prokaryotic systems; and promoter, start codon, CDS, stop codon, and terminator in eukaryotic systems. Each permutation aims to disrupt the natural transcription-translation flow, creating syntactically or functionally invalid constructs. A total of 19 translocation variants were designed for prokaryotic systems and 11 for eukaryotic systems. All constructs were rendered completely non-functional by design. All translocation combinations can be found in Appendix Table 3 for prokaryote and Appendix Table 4 for eukaryote.

B. Evaluation Metrics

B.1. Mean Log-likelihood

For casual language models (CLM), the log-likelihood of a nucleic acid sequence $X = (x_1, x_2, ..., x_n)$ is computed using logits output from the model. Since CLM operates in an autogressive manner, each token x_t is predicted based on all previous token $x < t$. Given the model's $logits_t$ at each position t, the probability of the ground-truth token is obtained via softmax:

$$P(x_t \mid x_{<t}) = \frac{\exp(\text{logits}_{t,x_t})}{\sum_{v \in V} \exp(\text{logits}_{t,v})} \tag{1}$$

where V represents the vocabulary. The mean log-likelihood is then computed as:

$$\text{LL}_{\text{CLM}}(X) = \frac{1}{n-1} \sum_{t=2}^{n} \log P(x_t \mid x_{<t}) \tag{2}$$

Masked language models (MLMs) differ from autoregressive models in their inference behavior, as they are not inherently designed for sequential generation. To address this, Wang & Cho (2019) introduced the pseudo log-likelihood (PLL) approach, wherein each token in a sequence is masked individually to compute token-wise conditional probabilities. More recently, Gordon et al. (2024) proposed Single-Inference Pseudo Log-Likelihood, an efficient approximation that enables linear-time inference for BERT-style MLMs by exploiting training-time masking dynamics.

Rather than computing pseudo log-likelihood (PLL), we assessed relative changes in mean token-wise log-probabilities derived from unmasked logits for all MLM-based gLMs. This method, though lacking strict probabilistic interpretation, provides a scalable and effective proxy for quantifying model sensitivity to syntactic or functional disruptions. The formula is as follows:

Given the $logits_t$ at each position t, the probability of the correct token is computed as:

$$P(x_t \mid X) = \frac{\exp(\text{logits}_{t,x_t})}{\sum_{v \in V} \exp(\text{logits}_{t,v})} \tag{3}$$

Then the log-likelihood for the entire sequence is computed over all positions:

$$\text{LL}_{\text{MLM}}(X) = \frac{1}{n} \sum_{t=1}^{n} \log P(x_t \mid X) \tag{4}$$

where all positions contribute to the likelihood computation since no masking is performed during inference.

B.2. One-sided Paired Permutation Test

To assess significance between paired conditions (e.g., original vs. mutant), we performed as one-sided paired permutation test. Given differences $d_i = {x_i}^{mutant} - {x_i}^{original}$ for each pair i, we computed the observed mean $\bar{d}_{\text{obs}}$. Under the null hypothesis of no effect, signs of d_i are exchangeable. We generated $N = 10,000$ random permutations by sampling $s_i \in \{-1, 1\}$ and computing $\bar{d}^{(j)} = \frac{1}{n} \sum_{i=1}^{n} s_i d_i$. The one-sided p-value was calculated as:

$$p = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\left(\bar{d}^{(j)} \leq \bar{d}_{\text{obs}}\right) \tag{5}$$

for the alternative hypothesis that the mutant is smaller than the original.

C. Model Performance Evaluation

We systematically evaluated the impact of tokenization strategies and model sizes on the performance of gLMs. As shown in Appendix Figure 5, we benchmarked the PDLLM model series (Liu et al., 2025a), which differ only in their tokenization choices. Our analysis revealed that the choice of tokenization profoundly influences model success rates: byte pair encoding (BPE) shows a stable performance across models. In contrast, single-token schemes performed poorly, while the 6-mer representation exhibited intermediate performance, with model-specific variability. To assess the effect of model size, we compared models that vary only in parameter size. Notably, we found no consistent relationship between model size and task performance: in several cases, smaller models performed on par with, or even outperformed, their larger counterparts.

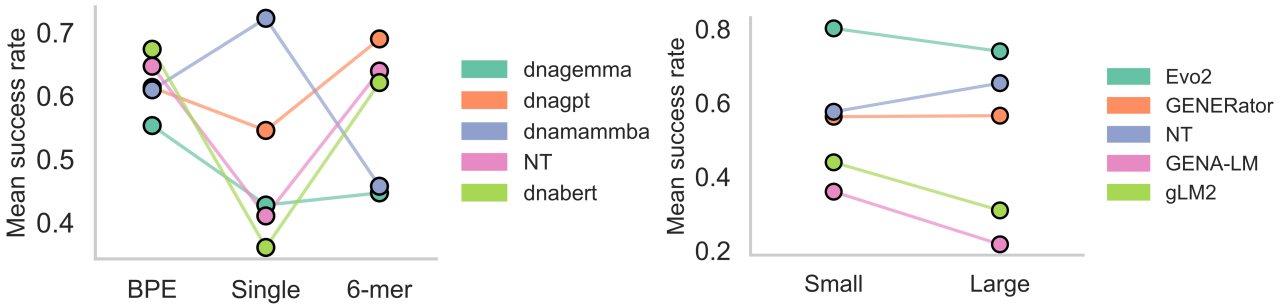


Figure 5. Comparative evaluation of gLMs across tokenization methods (left) and model sizes (right).

Furthermore, while representative models were selected for the main analysis, we conducted an extensive evaluation of various model versions across all datasets. As shown in Appendix Figure 6, the Evo series models (Nguyen et al., 2024; Merchant et al., 2024; Brixi et al., 2025) consistently outperformed all others in the NULLSETTES tasks, with nearly every Evo variant ranking at the top. This advantage likely reflects the superior architectural design and scaling strategy employed by the Evo models. Among non-Evo models, METAGENE-1 (Liu et al., 2025b) achieved the highest performance, likely due to its training on over 1.5 trillion base pairs of DNA and RNA sequences derived from wastewater metagenomes. The unique breadth and diversity of metagenomic data may provide this model with an enhanced ability to generalize across sequence contexts. The Nucleotide Transformer (NT) (Dalla-Torre et al., 2024) series followed, ranking third overall. These findings highlight the impact of both model architecture and training data diversity on generalization performance across regulatory sequence prediction tasks.

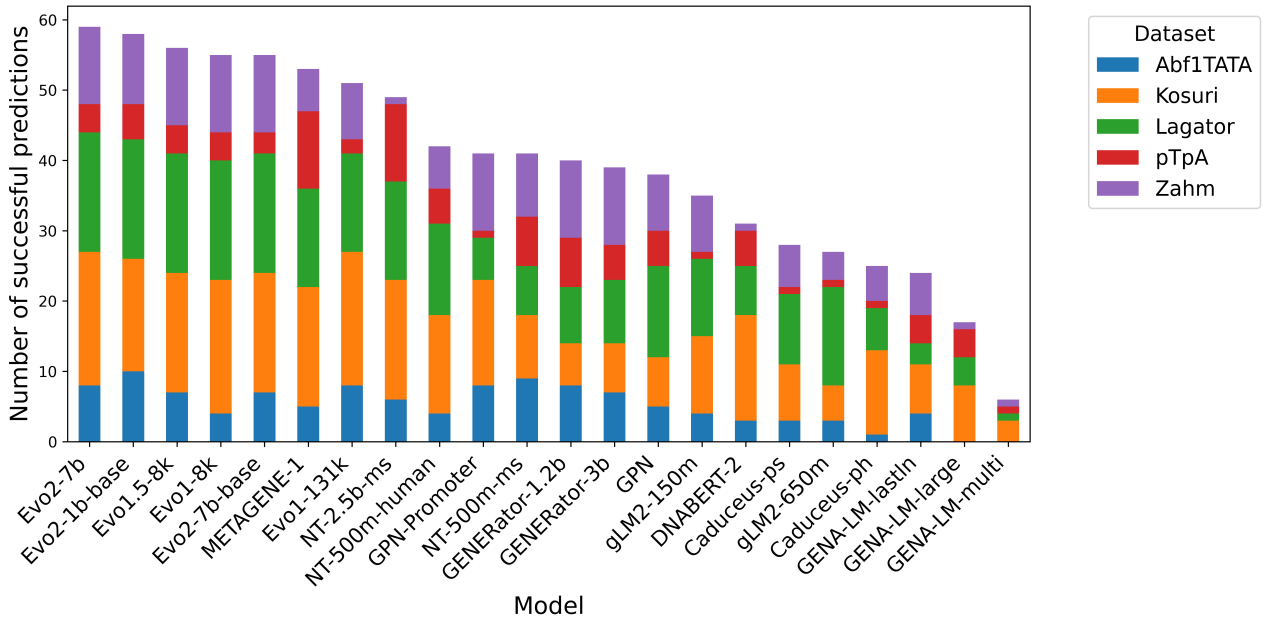


Figure 6. The number of cumulative successful predictions made by each model series across four datasets: Abf1TATA and pTpA (from deBoer), Zahm, Kosuri, and Lagator.

Model name	Pretraining dataset	Pretraining method	Tokenization	Architecture	Input length
Nucleotide Transformer 2.5B multispecies (Dalla-Torre et al., 2024)	850 whole genomes from NCBI, plants and viruses are not included, resulting in a total of 174B nucleotides, i.e. roughly 29B tokens.	MLM	6-mers	Transformer, 2.5B params	1000 bp
Nucleotide Transformer 500M multispecies (Dalla-Torre et al., 2024)	850 whole genomes from NCBI, plants and viruses are not included, resulting in a total of 174B nucleotides, i.e. roughly 29B tokens.	MLM	6-mers	Transformer, 500M params	1000 bp
Nucleotide Transformer 500M Human ref (Dalla-Torre et al., 2024)	GRCh38 human reference genome, resulting in 3B nucleotides, i.e. roughly 500M 6-mer tokens.	MLM	6-mers	Transformer, 500M params	1000 bp
Evo1 8k (Nguyen et al., 2024)	OpenGenome: Prokaryotic whole-genomes dataset (300B tokens).	CLM	Single-nucleotide tokens	StripedHyena, 7B params	Pretrained with 8,192 context
Evo1 131k (Nguyen et al., 2024)	OpenGenome: Prokaryotic whole-genomes dataset (300B tokens).	CLM	Single-nucleotide tokens	StripedHyena, 7B params	Pretrained with 131,072 context using Evo1 8k as the base model
Evo1.5 (Merchant et al., 2024)	50% increase in training data comparing to Evo1	CLM	Single-nucleotide tokens	StripedHyena, 7B params	Pretrained with 8,192 context using Evo1 8k as the base model
Evo2 1b base (Brixi et al., 2025)	OpenGenome2: a dataset containing 8.8 trillion tokens from all domains of life.	CLM	Single-nucleotide tokens	StripedHyena 2, 1B params	Pretrained with 8192 context length
Evo2 7b base (Brixi et al., 2025)	OpenGenome2: a dataset containing 8.8 trillion tokens from all domains of life.	CLM	Single-nucleotide tokens	StripedHyena 2, 7B params	Pretrained with 8192 context length
Evo2 7b (Brixi et al., 2025)	OpenGenome2: a dataset containing 8.8 trillion tokens from all domains of life.	CLM	Single-nucleotide tokens	StripedHyena 2, 7B params	Pretrained with 1M context using Evo2 7b base as the base model
GENERator 3b (Wu et al., 2025)	Eukaryotic genomes (386B bp), plants, fungi, protozoa, mammlian, vertebrate, invertebrate.	CLM	6-mer	Llama-based decoder, 3B params	Pretrained with a context length of 98k bp
GENERator 1.2b (Wu et al., 2025)	Eukaryotic genomes (386B bp), plants, fungi, protozoa, mammlian, vertebrate, invertebrate.	CLM	6-mer	Llama-based decoder, 1.2B params	Pretrained with a context length of 98k bp
METAGENE-1 (Liu et al., 2025b)	1.5T base pairs of DNA and RNA sequences from human wastewater samples.	CLM	Byte-pair encoding	Llama-2, 7B params	Pretrained with 512 sequence length

Table 1. Genomic language model details

Model name	Pretraining dataset	Pretraining method	Tokenization	Architecture	Input length
Caduceus-ph (Schiff et al., 2024)	Human reference genome (35M tokens / nucleotide base pairs).	MLM	Single-nucleotide tokens	BiMamba, 7.7M params	Pretrained with 131,072 sequence length (reverse complement augmentation)
Caduceus-ps (Schiff et al., 2024)	Human reference genome (35M tokens / nucleotide base pairs).	MLM	Single-nucleotide tokens	BiMamba, 7.7M params	Pretrained with 131,072 sequence length (No reverse complement augmentation)
GENA-LM human (Fishman et al., 2025)	T2T vs. GRCh38.p13 human genome assembly.	MLM	Byte-pair encoding	BERT/BigBird, 110M params	About 4,500 nucleotides (512 BPE tokens)
GENA-LM multi (Fishman et al., 2025)	T2T vs. GRCh38.p13 human genome assembly augmented by sampling mutations from 1000-genome SNPs (gnomAD dataset) + Multispecies genomes from ENSEMBL release 108.	MLM	Byte-pair encoding	BERT/BigBird, 110M params	About 4,500 nucleotides (512 BPE tokens)
GENA-LM large (Fishman et al., 2025)	T2T vs. GRCh38.p13 human genome assembly.	MLM	Byte-pair encoding	BERT/BigBird, 336M params	About 4,500 nucleotides (512 BPE tokens)
GPN (Benegas et al., 2023)	Brassicales reference genome from NCBI Genome. Took the union of exons (with a small intronic flank), promoters (1,000 bp upstream of transcription start sites) as well as an equivalent amount of random windows from the whole genome.	MLM	Single-nucleotide tokens	CNN	512 bp
GPN-Promoter (Benegas et al., 2025a)	Animal promoter, genomes of 434 animal species.	MLM	Single-nucleotide tokens	ByteNet, 152M	Pretrained on 512 bp sequences centered at TSSs of protein-coding genes
DNABERT 2 (Zhou et al., 2023)	Human genome dataset (2.75B nucleotide bases) + Multispecies genome dataset (from 135 species, spread across 6 cetogories). The dataset includes 32.49B nucleotides bases, excluding all sequence with N and retain only ATCG.	MLM	Byte-pair encoding	BERT, 117M params	Pretrained on 700 bp length sequences
gLM2 (Cornman et al., 2024)	OMG: encodes genomic scaffolds with both amino-acid (CDS) and DNA tokens (315B tokens).	MLM	Char-level tokens (DNA: lowercase, AA: uppercase)	Transformer, 650M params	Pretrained with a 4096 token context window
PDLLM-DNAMamba-6mer (Liu et al., 2025a)	Plant reference genomes	CLM	6-mer	SSM, 130M params	Pretrained max token length is 512

Table 2. Genomic language model details (continued)

Mutant ID	Description
Translocation-1	CDS - Promoter - RBS - Start codon - Stop codon - Terminator
Translocation-2	Promoter - CDS - RBS - Start codon - Stop codon - Terminator
Translocation-3	Promoter - RBS - CDS - Start codon - Stop codon - Terminator
Translocation-4	Promoter - RBS - CDS - Stop codon - Start codon - Terminator
Translocation-5	Promoter - RBS - CDS - Stop codon - Terminator - Start codon
Translocation-6	Promoter - RBS - Start codon - Stop codon - CDS - Terminator
Translocation-7	Promoter - RBS - Start codon - Stop codon - Terminator - CDS
Translocation-8	Promoter - RBS - Start codon - Terminator - CDS - Stop codon
Translocation-9	Promoter - RBS - Terminator - Start codon - CDS - Stop codon
Translocation-10	Promoter - Start codon - CDS - RBS - Stop codon - Terminator
Translocation-11	Promoter - Start codon - CDS - Stop codon - RBS - Terminator
Translocation-12	Promoter - Start codon - CDS - Stop codon - Terminator - RBS
Translocation-13	Promoter - Start codon - RBS - CDS - Stop codon - Terminator
Translocation-14	Promoter - Terminator - RBS - Start codon - CDS - Stop codon
Translocation-15	RBS - Promoter - Start codon - CDS - Stop codon - Terminator
Translocation-16	RBS - Start codon - CDS - Promoter - Stop codon - Terminator
Translocation-17	RBS - Start codon - CDS - Stop codon - Promoter - Terminator
Translocation-18	RBS - Start codon - Promoter - CDS - Stop codon - Terminator
Translocation-19	Start codon - Promoter - RBS - CDS - Stop codon - Terminator

Table 3. **Prokaryotic virtual mutant cases.** Translocation-3,4,5 cannot be compensated by a later start codon as there are no other in-frame start codon.

Mutant ID	Description
Translocation-1	CDS - Promoter - Start codon - Stop codon - Terminator
Translocation-2	Promoter - CDS - Start codon - Stop codon - Terminator
Translocation-3	Promoter - CDS - Stop codon - Start codon - Terminator
Translocation-4	Promoter - CDS - Stop codon - Terminator - Start codon
Translocation-5	Promoter - Start codon - Stop codon - CDS - Terminator
Translocation-6	Promoter - Start codon - Stop codon - Terminator - CDS
Translocation-7	Promoter - Start codon - Terminator - CDS - Stop codon
Translocation-8	Promoter - Terminator - Start codon - CDS - Stop codon
Translocation-9	Start codon - CDS - Promoter - Stop codon - Terminator
Translocation-10	Start codon - CDS - Stop codon - Promoter - Terminator
Translocation-11	Start codon - Promoter - CDS - Stop codon - Terminator

Table 4. **Eukaryotic virtual mutant cases.** Translocation-2,3,4 cannot be compensated by a later start codon as there are no other in-frame start codon.