Westlake University
Center for Interdisciplinary Studies
Laboratory of Cell Ethology

Predicting function of evolutionarily implausible DNA sequences

Shiyu Jiang Xuyin Liu Zitong Jerry Wang
shiyujia@usc.edu, jerry@westlake.edu.cn

1 Introduction

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, our work highlights the importance of considering both sequence likelihood and sequence length when using gLMs for mutation effect prediction. Nullsettes dataset is publicly available on GitHub: `https://github.com/cellethology/GLM-Nullsette-Benchmark`.

2 Contributions

1. **Nullsettes benchmark**: Introduce a synthetic biology benchmark simulating loss-of-function mutations via control element translocations, enabling zero-shot evaluation of genomic language models.
2. **Likelihood–function link**: Find a strong correlation between model-assigned sequence likelihood and ability to detect LOF mutations, highlighting evolutionary plausibility as a proxy for function.
3. **Length-dependent range**: Show that optimal likelihood thresholds for mutation prediction vary with sequence length, stressing the need to account for both likelihood and length in synthetic design.

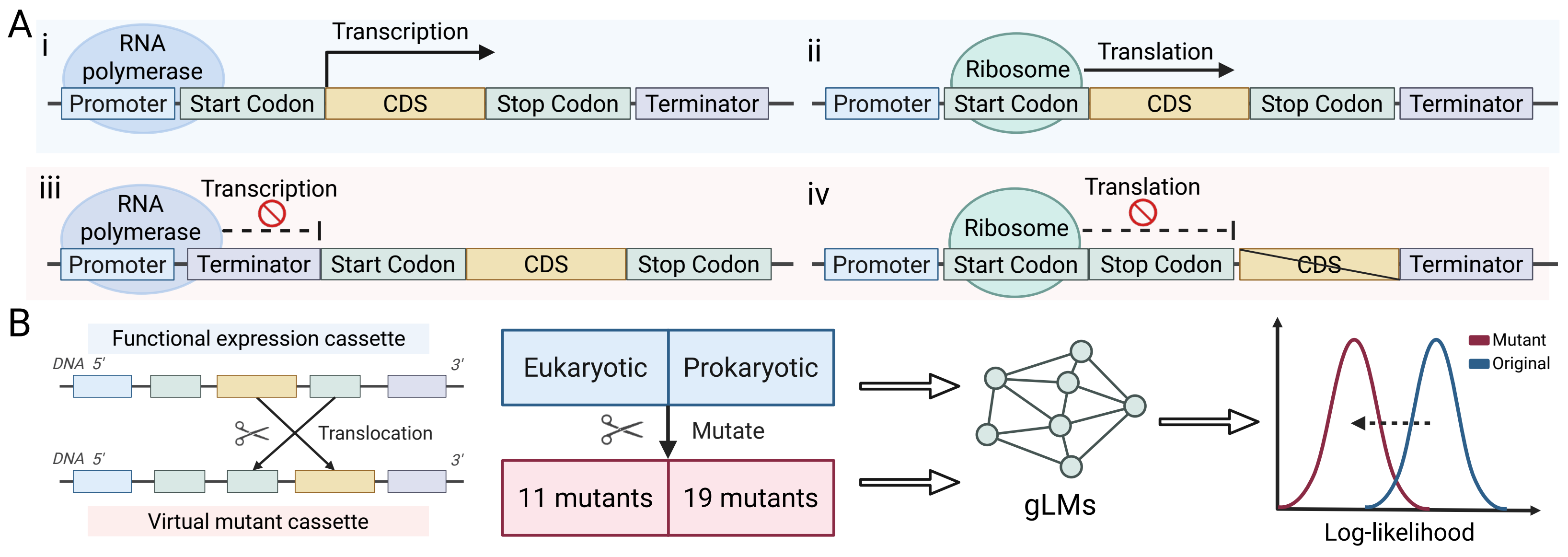

Figure 1: Nullsettes evaluation workflow

3 Methodology

3.1 Nullsettes construction

Nullsettes are synthetic expression cassettes created by permuting six regulatory elements—promoter, RBS, start codon, CDS, stop codon, and terminator—to disrupt transcription or translation. We systematically selected mutants where (1) CDS expression cannot be rescued by external sequences and (2) cellular machinery still generates nonfunctional transcripts or proteins. The set includes 11 eukaryotic and 19 prokaryotic variants, with 8 extra prokaryotic mutants arising from RBS-specific translocations.

3.2 Functional cassette curation

We curated expression cassettes from public MPRA datasets with high expression but low genomic language model (gLM) likelihood, reflecting low evolutionary plausibility. These include cassettes with cross-species elements (e.g., jellyfish GFP with bacterial or mammalian regulatory sequences) and synthetic constructs with random promoters (e.g., Lagator, deBoer pTpA). We label datasets with random promoters as "Low" and those with natural motifs as "High".

3.3 Baseline models and evaluation metrics

We benchmarked 12 state-of-the-art genomic foundation models spanning diverse tokenization schemes (k-mers, single nucleotides, BPE, hybrids), pretraining objectives (masked vs. autoregressive), training corpora (e.g., human, plant, metagenomic), and architectures (CNNs, Transformers, StripedHyena, Mamba). Model performance was evaluated using mean base-pair log-likelihood (LL), computed separately for masked and causal models. We compared LL distributions between original cassettes and Nullsettes using one-sided paired permutation tests to quantify sequence disruption.

4 Results

We benchmarked 12 self-supervised gLMs that represent current SOTA approaches to DNA language modeling. Figure 2 shows that Evo series consistently achieves the competitive performance across four synthetic datasets. Figure 3 reveals that gLM performance is positively correlated with the log-likelihood of the original, non-mutant sequence, and that this correlation is consistent across models and datasets. Furthermore, Evo-2-7B achieving the best overall performance, especially on low-likelihood, synthetic sequences like E. coli and yeast with random promoters. Figure 4 demonstrates that the optimal likelihood threshold for accurate mutation prediction increases with sequence length, indicating no fixed likelihood cutoff across datasets. These results underscore the importance of considering both likelihood and length when applying gLMs to synthetic sequence design.

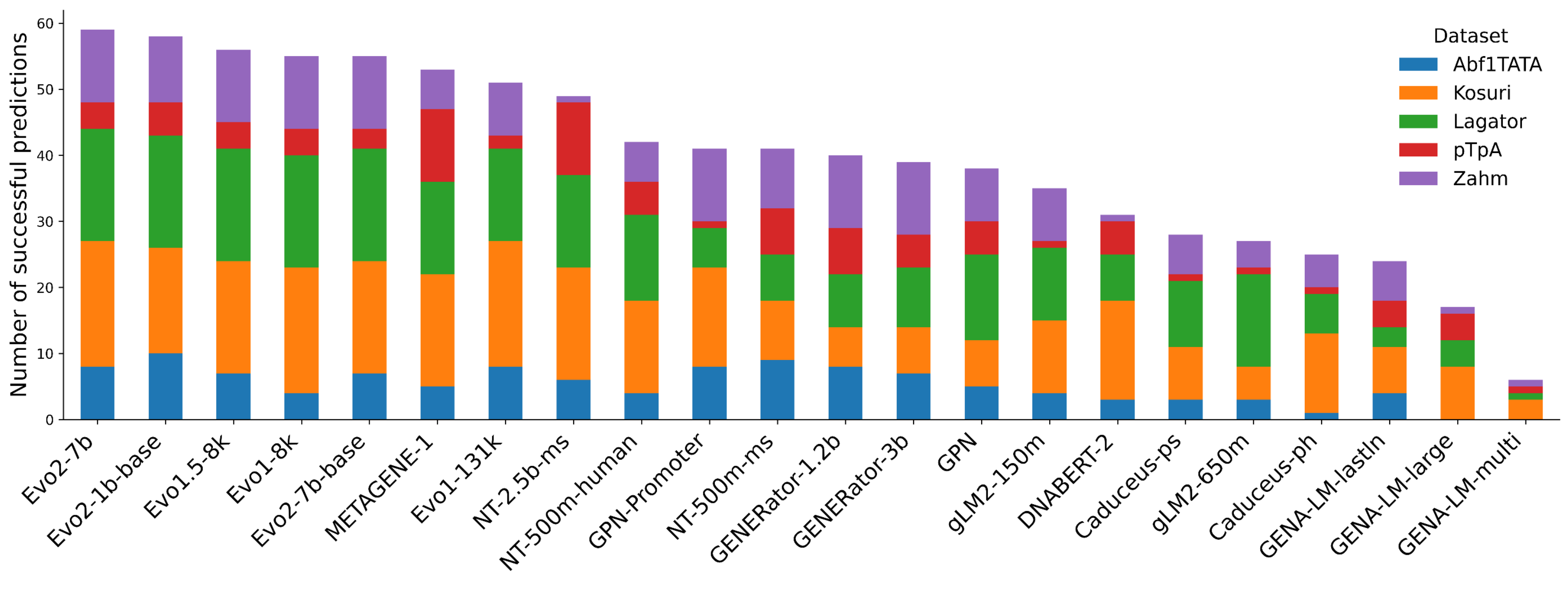

Figure 2: Successful predictions made by each model series across four datasets

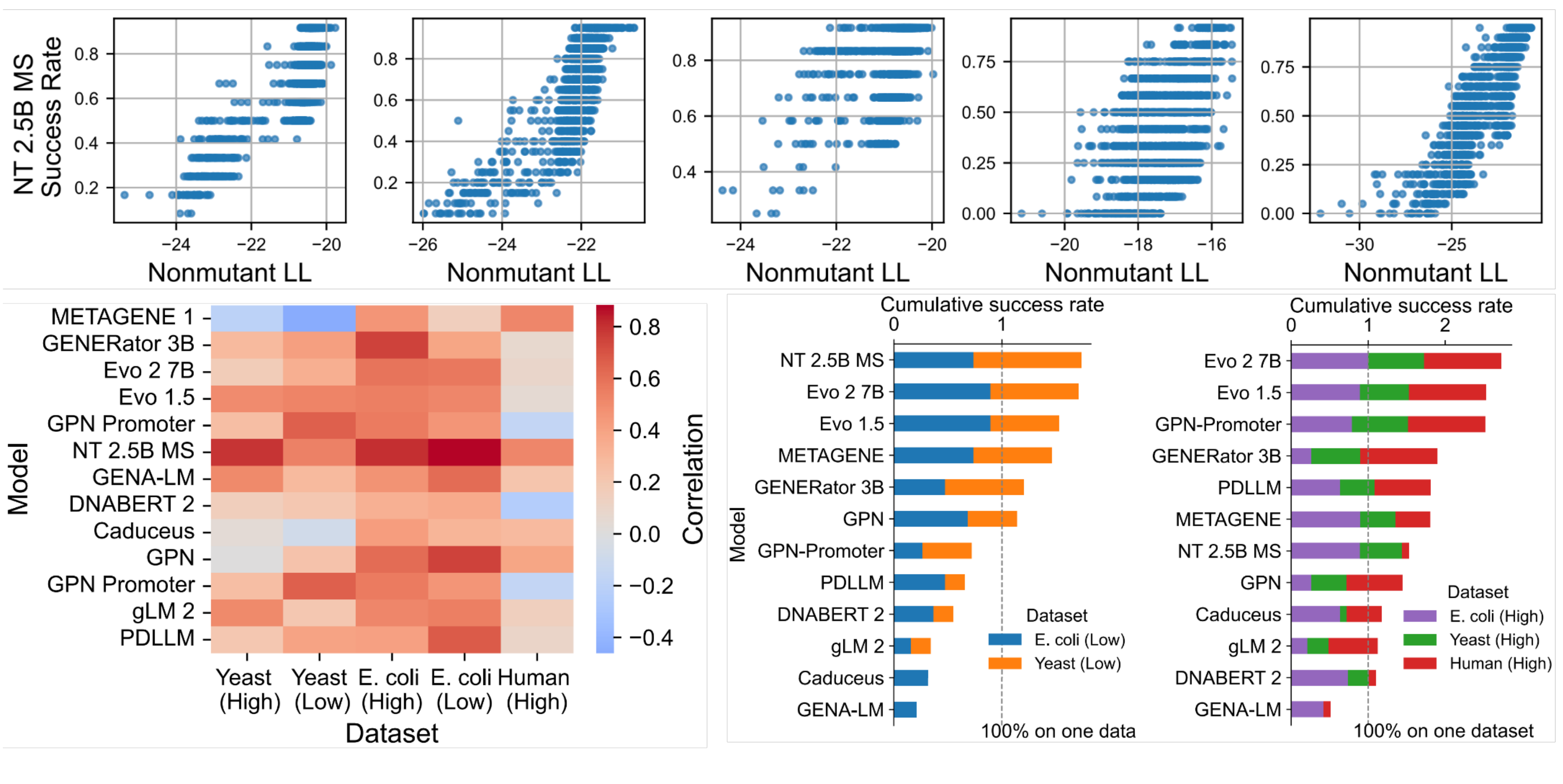

Figure 3: Relationship between gLM likelihood and performance on Nullsettes prediction

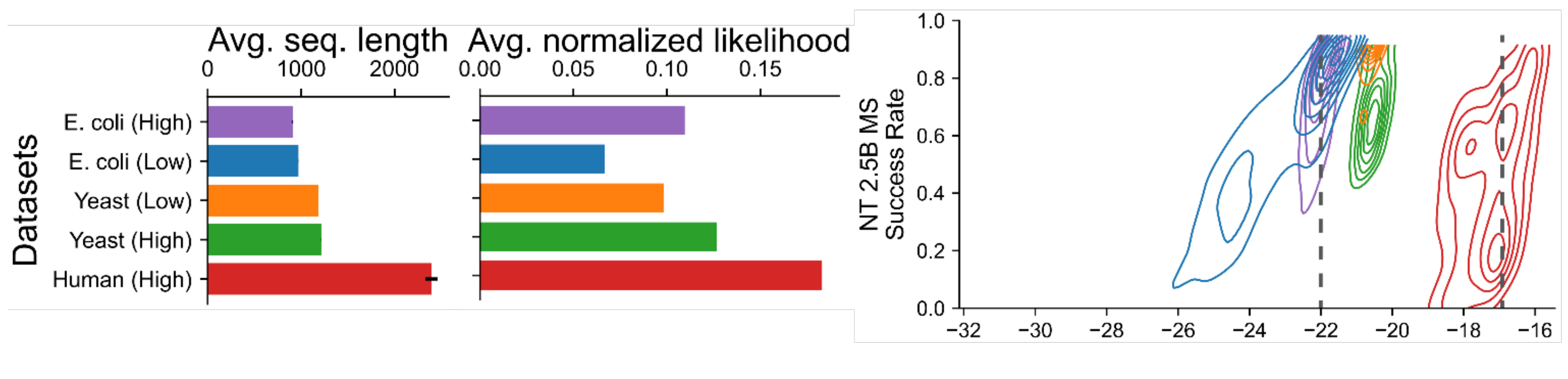

Figure 4: Optimal likelihood range for Nullsettes prediction varies with sequence length