Designing *de novo* bacterial toxin-antitoxins using a generative genomic foundation model

Aditi Merchant[1], Santiago Mille-Fragoso[1], Samuel King[1], Chang M. Yun[2]

[1]Department of Bioengineering, Stanford University,
[2]Department of Chemical Engineering, Stanford University

Abstract

Toxin-antitoxin (TA) systems are natural kill switches in bacteria that consist of a toxin that inhibits an essential cellular process, and a corresponding antitoxin that counteracts it. TAs are important for bacterial adaptation, persistence, and defense, and have great potential in bioengineering and therapeutic applications, such as for antimicrobials, novel selection markers, and biocontainment. The recent development of genomic large language models (LLMs) promises to enable the generative design of novel TA systems to expand the current toolbox available. Here, we use Evo, a genomic foundation model capable of multi-element generation tasks, to generate *de novo* TAs by zero-shot prompting and after fine-tuning. We validate the viability of the new TA designs using multiple *in silico* metrics, including sequence homology and protein structure and interaction prediction, and show that our designs are likely confident candidates ready for experimental validation. Overall, this study demonstrates the first genomic LLM-based design pipeline for TAs and establishes a set of confidence metrics for similar future work.

Introduction

Toxin-antitoxin (TA) systems play important roles in bacterial persistence, adaptation, and multidrug tolerance [Page & Peti, 2016]. These systems consist of a pair of genes: a toxin, which interferes with vital cellular processes, and an antitoxin, which neutralizes the toxin during normal growth conditions. Under stress, the antitoxins are selectively degraded, freeing the toxins to inhibit vital cellular functions, such as DNA replication and protein translation, causing rapid growth arrest.

All known TA pairs are genetically linked (*i.e.*, one gene is directly adjacent to the other), and the extensive evolution of toxin and antitoxin genes has resulted in considerable diversity in TA combinations [Hayes & Melderen, 2011; Jurenas et al., 2022]. As a result, the intracellular targets of toxins are varied. For instance, some toxins function as sequence-specific endoribonucleases, making them attractive tools for basic research and bioengineering applications. On the other hand, antitoxins, which can be either proteins or small RNAs, employ different mechanisms to counteract toxin activity or inhibit toxin synthesis. The variety of mechanisms in which TAs act are classified into eight types, the majority of which are protein based [Jurenas et al., 2022]. Type II TAs consist of protein-toxin and protein-antitoxin pairs, and are also the most well studied class [Guan et al., 2024] (**Fig. 1a**).

To date, TAs have largely been employed as positive selection plasmid vectors or to produce novel expression systems aimed at massive overproduction of a target protein. Given the potent nature of TAs, their utility as antimicrobials, antivirals, biocontainment factors, and even cancer therapies are also actively being explored [Jurenas et al., 2022]. Their use as antimicrobials is especially pertinent, with the projected rise in prevalence of antibiotic-resistant pathogens [Frieri et al., 2017]. As such, we only stand to benefit from discovering or designing more TAs for uses in bioengineering and therapeutics.

While thousands of putative novel TAs have been discovered through homology-based searching of public databases [Akarsu et al., 2019; Guan et al., 2024; Shao et al., 2011; Xie et al., 2018], the search space is confined to direct homology with known TA sequences. Recent developments in genomic large language models (LLMs) have opened new possibilities in designing functional genetic systems based on learning evolutionary 'semantics' across species [Hwang et al., 2024; Nguyen et al., 2024; Ruffolo et al. 2024; Shao, 2023]. These models are typically trained on long context sequences of DNA with the goal of capturing multi-gene and coding-noncoding relationships. An added benefit of training on genomic DNA sequences is that multiple modalities of information are encoded in genomes, including DNA, RNA, and protein sequence information. This advantage was demonstrated by Evo, a genomic LLM based on the StripedHyena deep learning architecture [Nguyen et al., 2024]. Although the designs were only validated *in silico*, Evo was shown to be capable of multi-element generation tasks, such as RNA-protein complexes and entire transposable systems, as well as whole genome-like sequences up to 600 kilobases (kb) in length. The results suggest that Evo would be wholly capable of generating two-component systems such as TAs.

Here, we establish a training and *in silico* validation pipeline to produce *de novo* Type II TA sequences (**Fig. 1b**). We fine-tuned Evo on Type II TA sequences from the Toxin-Antitoxin Database (TADB) 3.0 using a Low-Rank Adaptation (LoRA) approach [Guan et al., 2024; Hu et al., 2021], and also performed zero-shot prompting to generate Type II toxins/antitoxins when prompted with an antitoxin/toxin (*i.e.,* the conjugate pair) sequence. We show through homology, structure prediction, and multimer prediction that Evo-generated TAs are viable candidates for experimental validation. The sequences produced from the fine-tuned models were largely similar in sequence to existing TAs, while the zero-shot prompting was able to produce more diversity in its generations. This work provides a new potential repertoire of TA systems and may expand existing antibiotic diversity, novel selection markers, and kill switches for synthetic biology applications.

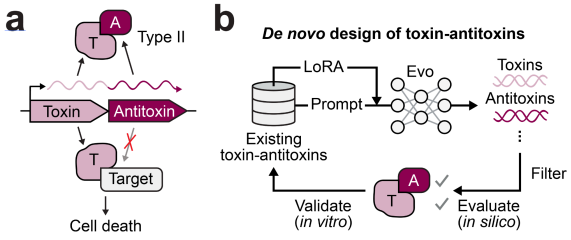

Figure 1. Strategy for *de novo* design of Type II toxin-antitoxins using a genomic foundation model. (**a**) Mechanism of Type II toxin-antitoxins. (**b**) High-level overview of the strategy for designing novel Type II toxin-antitoxins to expand the existing repertoire.

Results

Fine-tuning Evo on toxin-antitoxin systems from TADB 3.0

Evo is a language model-based genomic foundation model, with 7B parameters, based on a StripedHyena architecture (hybrid of rotary attention + long-range Hyena layers) for long-range context, up to 131k tokens, at single-nucleotide, byte-level resolution, and trained on 2.7M prokaryotic and phage genomes [Nguyen et al., 2024]. We aimed to fine-tune Evo to focus on prediction and design of a specific task: bacterial Type II TA sequence pairs. Deep learning is particularly appropriate for this problem, as the probabilistic model of a deep learning model enables generating entirely *de novo* sequences, previously unseen in nature. Additionally, a deep learning architecture enables fine-tuning from previously learned embeddings, transferring the learnings from a larger, more general dataset (here, 2.7M prokaryotic and phage genomes) to a specific, data-limited task (*e.g.*, thousands of known TA pairs).

The Toxin-Antitoxin Database (TADB) 3.0 is a publicly available database of TAs, with over 500 experimentally validated TAs from related publications, and over 200,000 putative TAs predicted by NCBI Basic Local Alignment Search Tool (BLAST) and Hidden Markov Model (HMMER3) via TAfinder Tool 2.0 [Guan et al., 2024]. Focusing on the most abundant TA type, 169,197 Type II TA sequences were located and preprocessed with toxin-antitoxin classification (CLS) tokens to prepare for training (**Fig. 2a–b**).

Low-Rank Adaptation (LoRA) is an efficient fine-tuning approach for large models, that models new weight updates as the matrix multiplication of two lower dimensional matrices [Hu et al., 2021]. We implemented LoRA on a frozen pretrained Evo 8K context model (**Fig. 2c**).

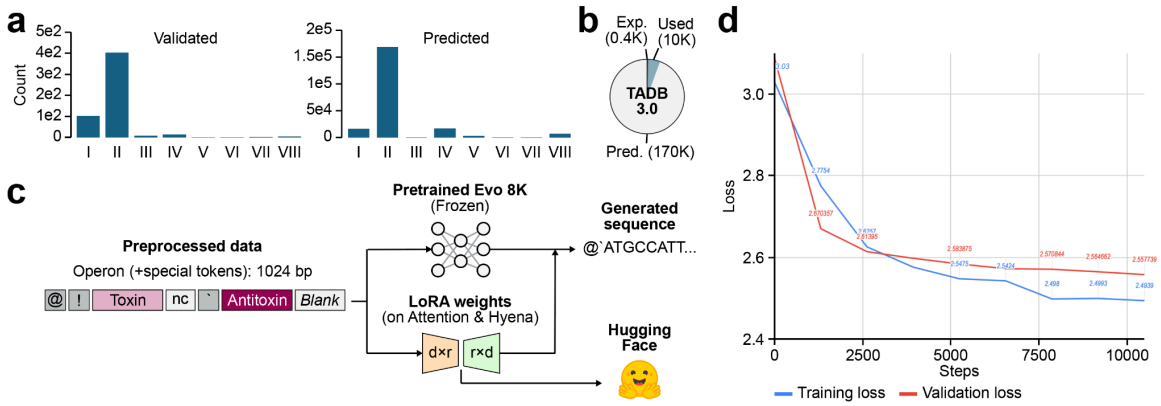

Figure 2. Fine-tuning Evo on Type II toxin-antitoxins. (**a**) Total counts of experimentally validated (left) and computationally predicted operons (right) of toxin-antitoxin types from the Toxin-Antitoxin Database (TADB) 3.0. (**b**) Proportion of Type II toxin-antitoxin operons used from TADB 3.0 for fine-tuning. Exp., experimentally validated; Pred., predicted.(**c**) Workflow for data preprocessing and fine-tuning using Low Rank Adaptation (LoRA). @ indicates Type II, ! indicates Toxin, and ` indicates Antitoxin. (**d**) Training and validation loss curves from LoRA fine-tuning.

Given that this was the first known time applying LoRA to a StripedHyena architecture, we performed a hyperparameter sweep to find an optimal setup. To identify optimal hyperparameters, we focused on testing key LoRA hyperparameters (number of epochs, learning rate, rank, alpha) and training data size (see **Methods** for final hyperparameters). Interestingly, a less explored "parameter" was the number of layers to fine-tune. To emulate a complete re-training of the model, we decided to fine-tune all linear layers of the model. Due to compute limitations, we decided to fine-tune the model on 10,000 Type II TA loci sequences (henceforth

referred to as the 10K model). For comparison, we performed downstream evaluations using models fine-tuned with 4,000 and 8,000 sequences (henceforth referred to as 4K and 8K models, respectively). The 10K fine-tuned model exhibited the lowest loss during training, where loss curves sharply reduced and flattened over ~10,000 steps (**Fig. 2d**).

Establishing an in silico validation pipeline for toxin-antitoxin designs

A variety of bioinformatics approaches exist that have been used to predict whether genes are toxins or antitoxins [Guan et al., 2024]. For example, multiple sequence alignment (MSA) can be used to identify sequence homology with known TA genes, while protein structure prediction (*e.g.*, AlphaFold, ESMFold) and interaction prediction (*e.g.*, AlphaFold Multimer) tools can be used to identify TA binding interactions [Evans et al., 2021; Jumper et al., 2021; Lin et al., 2023]. However, to date, these *in silico* validation metrics have not been applied to *de novo* TA designs.

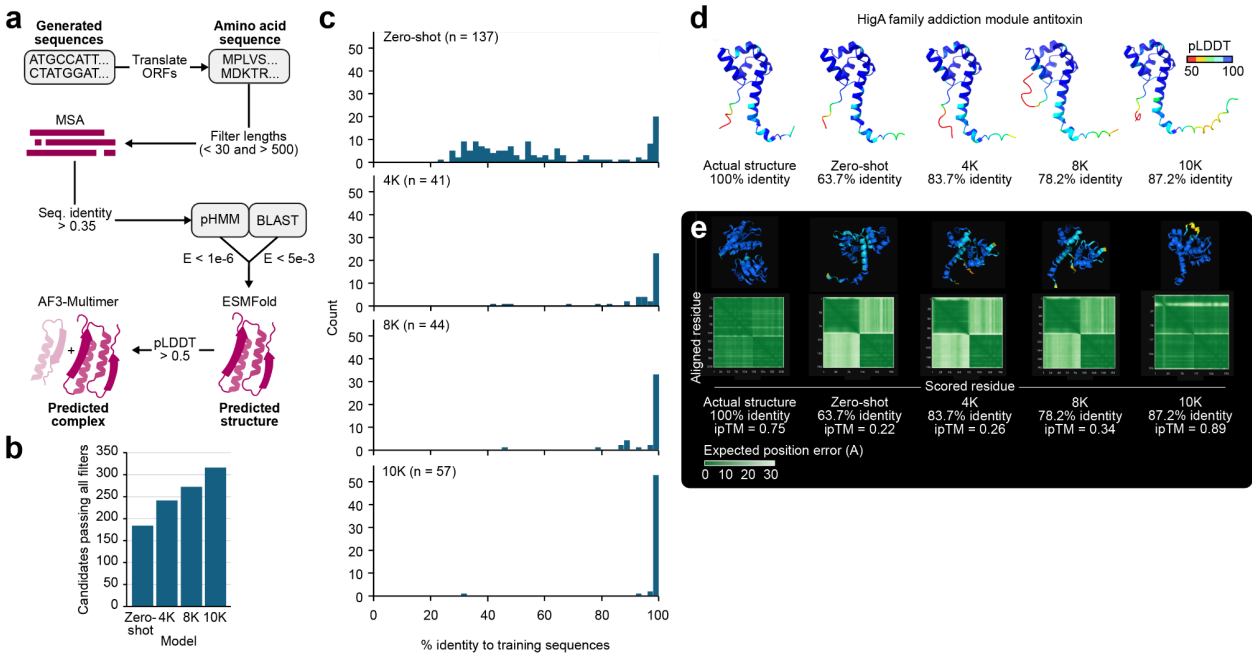

Figure 3. *In silico* evaluation of Evo-generated toxin-antitoxins. (**a**) Filtering and evaluation pipeline for generated sequences. ORF, open reading frame; MSA, multiple sequence alignment; Seq., sequence; pHMM, profile hidden Markov model; BLAST, Basic Local Alignment Search Tool; pLDDT, predicted local distance difference test; AF3, AlphaFold3. (**b**) Number of candidates passing all filters for zero-shot prompting, and the 4K, 8K, and 10K fine-tuned models. (**c**) Percent identities of generated sequences with high confidence BLAST hits compared to the training sequences for zero-shot prompting, and the 4K, 8K, and 10K fine-tuned models. (**d**) ESMFold-predicted structures of the most confident sequences that BLASTed to the HigA family antitoxin from zero-shot prompting, 4K, 8K, and 10K fine-tuned models. The percent identities to the actual sequence are shown. (**e**) Top: AF3-Multimer-predicted structures of the sequences from (**d**) in complex with the wildtype HigB toxin. Middle: Predicted Aligned Error (pAE) plots for each multimer prediction. Bottom: The percent identities to the actual sequence and the interface predicted template modeling (ipTM) scores are shown.

To determine the feasibility of Evo's generations, we used a similar structure and sequence-informed approach, employing a pipeline of filters that removed sequences with incorrect sizes, low pairwise sequence alignment between any existing TA, no hits on BLAST or on a TA profile hidden Markov model (pHMM), and low structural identity (**Fig. 3a**). This resulted in a total of 1,013 final generations across three fine-tuned models (4K, 8K, 10K) and

zero-shot prompting, from 800 initial sequences generated per model. Each Evo-generated sequence was 1,000 nucleotides (nt) long (in accordance with TA sequences typically being 240-900 nt long) and contained an average of 3.3 ± 0.6 ORFs per generation. Of all the filtration steps, the BLAST and pHMM filter resulted in the greatest reduction of sequences, culling roughly 72% of the total number of starting ORFs.

When comparing the generations from the fine-tuned vs. actual models, the 10K model resulted in the highest number of final TA candidates, with a total of 316 final generated sequences (**Fig. 3b**). The 8K and 4K model resulted in a similar number of sequences, with 272 and 241 generations respectively. Understandably, the zero-shot prompting approach resulted in 184 final generations, which, while lower than the fine-tuned model, suggested that even just a zero-shot based approach was enough to generate a diverse list of TAs. Of the final generations that had BLAST hits, the generations from the fine-tuned models were highly similar to sequences within the training set, with the vast majority having > 95% sequence identity to a TA in the training data (**Fig. 3c**). Meanwhile, the zero-shot prompting approach produced a much more diverse array of toxin/antitoxin sequences, with a bimodal distribution of percent identities centered around 50% and >95%. Interestingly, over half of the sequences with % identities less than 70% in the zero-shot batch had query covers higher than 80%, suggesting that the model was adding well-distributed diversification to the amino acid sequences. For the sequences that did not have any significant BLAST hits but still passed the pHMM filter, there were much higher numbers of TA generations in the fine-tuned batch than the zero-shot batch, suggesting that the fine-tuned models were better at generating sequences that still resembled TAs despite not directly BLASTing to sequences in the training set.

At a structural level, the majority of sequences that passed the pHMM and BLAST filters had high structural confidence, likely due to their alignment with existing TA sequences. However, when evaluating their ability to bind their conjugate wild-type toxin or antitoxin pair, small sequence differences resulted in vastly different docking outcomes. This was clearly demonstrated in a group of generations that mapped to a HigA family addiction module antitoxin, which across all models and zero-shot, closely resembled the wild-type structure despite drastically different sequence identities (**Fig. 3d**), but had extremely varied ipTM scores in AlphaFold3 when co-folded with the wild-type conjugate toxin, HigB (**Fig. 3e**). While the generations from the zero-shot, 8K, and 4K model resulted in very low ipTM scores, the generation from the 10K model actually had a higher ipTM than the wild-type antitoxin, highlighting the ability of each of the models to produce highly diversified generations with varied outcomes on TA interactions.

Discussion

We established the first generative design workflow for TAs, showing the promise of genomic LLMs for expanding the functional space of multi-component genetic systems. After performing fine-tuning Evo with up to 10,000 TA operons and zero-shot prompting, a total of 1,013 sequences passed all of our filtering steps across all models and appear likely to be TAs.

While our results were promising, a major obstacle when working with a large 7B parameter model such as Evo was the intensive computational requirement for fine-tuning to a satisfactory level. It is possible that we could have produced better or more viable TA designs with an optimized fine-tuning approach. To avoid requiring too much compute, we decided to use a LoRA approach whereby the model weights are frozen and the parameter space is reduced

to a significantly lower dimension, fine-tuned, and projected back to the original dimension and factored into the model's outputs [Hu et al., 2021]. The compute cost could be further reduced by fine-tuning on only select layers. Given that this study, to the best of our knowledge, represents the first time a LoRA fine-tuning approach has been applied to the StripedHyena architecture, we believe that it is worth exploring fine-tuning various combinations and amounts of layers in the model. Unsatisfactory fine-tuning would result in either overfitting the model or underutilization of the TA data, which could potentially be aided by gradual unfreezing and fine-tuning on previous layers and training for more or less epochs until an appropriate model performance is reached.

Another strategy to mitigate compute costs would be to only perform zero-shot prompting with the Evo-8k or Evo-131k model. Prompting with a complete or partially masked toxin sequence to generate a corresponding antitoxin, or vice versa, is a relatively easy way to generate novel designs, and the complementarity of DNA makes it possible to generate operons in either direction despite Evo's autoregressive nature. Indeed, when we performed zero-shot prompting, we were able to produce 184 TA candidates that were more diverse than the candidates produced from the fine-tuned models. Prompting with the sequence directly upstream of a known TA pair or with a consensus sequence made by aligning all the sequences of interest are two other viable strategies for generating *de novo* TA designs.

The current state-of-the-art for candidate TA validation is by sequence homology [Akarsu et al., 2019; Guan et al., 2024; Shao et al., 2011; Xie et al., 2018]. We added to this metric by modeling both the confidence of protein structure predictions and the protein-protein interactions using ESMFold and AlphaFold3 [Abramson et al., 2024; Lin et al., 2023; Jumper et al., 2021]. Although these models are not perfect, they have shown to be effective oracles for predicting protein fitness and viable protein interaction partners [Hie et al., 2024; Lim et al., 2023].

We found that many of the Evo-generated sequences were noisy, which made our generations more "diverse" but resulted in low pAE and ipTM scores in downstream analyses and limited actual functionality as a toxin or antitoxin. To address this from a hyperparameter perspective, we could adjust the top k value to prioritize greedy sampling and use a lower temperature. Alternatively, we could set the pHMM threshold to be more stringent or perform a direct MSA against the database of TA sequences and set a high percent sequence alignment filter, therefore ensuring that any remaining candidates are reasonable simply by means of their similarity to existing TA sequences. Finally, sequence generations could be improved by using more training data (*e.g.*, the entire TADB 3.0 dataset of ~170K Type II TA pairs as opposed to only the 10K that we used).

We validated the Evo-generated TA sequences by computational metrics and did not perform any *in vitro* experiments to test the functionality of our designs. We were able to produce 1,013 high-confidence candidates that could be synthesized and tested in E. coli by a simple killing assay. Briefly, each toxin and antitoxin could be synthesized in inducible expression vectors and transformed in pairs. The transformed cells would be grown in media containing the inducers and the antitoxin inducer could be subsequently removed. If the cells die (*i.e.*, the number of colonies or concentration of bacteria decreases), it would be indicative that the TA operon works.

While TA sequences are useful for a variety of bioengineering and therapeutic applications, such as for antibiotics, biosecurity measures, and molecular biology protocols, it is conceivable that the same design methods we have described here could be easily applied to other genetic systems. Two-component systems consisting of coding sequences such as Type II

TAs are especially intuitive to design and test with the suite of protein modeling tools available today, but we imagine other useful operons such as anti-CRISPRs and diversity generating retroelements could similarly fit into our design pipeline. We hope that other creative applications with models like Evo will continue to be established to expand the capabilities of generative genomics.

Methods

Data curation. We collected TA sequences from the Toxin-Antitoxin Database (TADB) 3.0, a public access database of TAs, last updated in January 2024 [Guan et al., 2024]. TADB 3.0 contains over 500 experimentally validated TAs and over 200,000 putative TAs predicted by NCBI BLASTp and HMMER3 via TAfinder Tool 2.0. Data are labeled by TA type (I-VIII), coding sequence, organism, and genome accession number. We selected the TA type with the largest available dataset: TA Type II, with a total of 169,192 TA pairs.

Pre-processing. Given computational constraints, we selected 10,000 out of the 169,192 Type II TA pairs in TADB 3.0 [Guan et al., 2024] for fine-tuning. The training data was prepared as follows: Each TA locus position was identified from its respective genome (including toxin gene, antitoxin gene, and TA intergenic region). The locus sequence (on the genome forward strand) was retrieved from NCBI Entrez Molecular Sequence Database [Sayers et al., 2022]. Three classification (CLS) tokens were prepended to each locus component: '@' at the start of the Type II locus, '!' at the start of the toxin gene, and '`' at the start of the antitoxin gene. The processed sequences were filtered below 1024 bases, appended with spaces up to 1024 bases, and 10,000 samples were selected at random.

Fine-tuning Evo. To refine Evo to generate novel TA systems, we used Hugging Face's (HF) *transformers* and *PEFT* libraries to load, configure and re-train the model using LoRA. Using `AutoModelForCausalLM` and `AutoTokenizer` we loaded the fixed ("1.1_fix") version of the 8k context Evo model using its tokenizer as specified in the Evo's HF page. Interestingly, the model was loaded using PyTorch's dfloat16 datatype to avoid bit overflowing. Using the pre-processed data, the final dataset was generated by tokenizing each of the sequences with padding enabled using Evo's byte-level, single-nucleotide tokenizer. The dataset was subsequently split into training (70%, 7000) and validation (30%, 3000) sets. A rank of 128 and an alpha (α) of 128 were used to create an modified Evo model containing the LoRA weight adapters. Briefly, PEFT's *`LoraConfig`* was used with the following parameters: r=128, alpha=128, dropout=0.1 and lora_init_weights='gaussian'. Importantly, LoRA was applied to all the linear components of the Evo model (see the code supplied below), including the three attention layers in StrippedHyena and all MLPs in the Hyena operator. Next, HF's `TrainingArguments` and *`Trainer`*, were used to define the training hyperparameters and perform fine-tuning. The model was tuned over 3 epochs using a learning rate of 3e-4 and a batch size of 1. These hyperparameters, as well as the ones used to set up LoRA, were chosen based on compute limitations as well as a manual hyperparameter sweep performed using a smaller dataset of 400 sequences. The complete list of preliminary fine-tuned models with descriptions of the hyperparameters used can be found in: https://huggingface.co/lsmille. To assess the impact of different hyperparameters, validation and training loss were monitored

during training to find the ones that decreased training and validation loss faster without leading to overfitting (*i.e.*, decrease of training loss and increase of the validation loss).

Sampling and validation. Following satisfactory fine-tuning, the model was sampled to generate 1) toxins, 2) antitoxins, and 3) Type II TA pairs using their respective prompt tokens (!,`, and @ respectively). To evaluate the quality of the generations, first, ORFs within the generated sequences were identified using Prodigal and translated into amino acid sequences [Hyatt et al., 2010]. These sequences were then filtered based on size in alignment with filtration metrics from TADB 3.0, with sequences > 30 and < 500 amino acids being kept. From there, these sequences were searched against a toxin pHMM and antitoxin pHMM generated using HMMER [Eddy, 2011], with sequences with E < 1 × 10−6 being kept. Concurrently, the sequences were searched against a custom BLAST database containing only TAs from the 10K training dataset for hits with E < 0.005. Any sequences passing pHMM filtration criteria or BLAST filtration criteria were then folded using ESMFold [Lin et al., 2023]. Predicted structures with average pLDDT scores > 0.5, in alignment with the average pLDDT from known TA structures, were kept as final candidates. For a random sample of final candidates, AlphaFold3 was used to evaluate binding between TA pairs (using the closest existing toxin or antitoxin match for antitoxin- or toxin-only generations, respectively), allowing for an evaluation of the final candidates' binding abilities to their conjugate pair.

To further evaluate the final candidates, the generations with BLAST hits were clustered on their sequence similarity to existing toxins or antitoxins, allowing for the determination of how "novel" Evo's high-quality generations were compared to existing sequences. For the actual sampling process, a temperature and top_k sweep was performed to determine the optimal sampling conditions, with a temperature of 0.7 and top_k of 4 being used for all generations discussed in the results section. Additionally, to serve as a point of comparison, the base model was sampled on a set of prompts intended to generate TAs (providing the sequence of the toxin, antitoxin, reverse complement of the toxin, and reverse complement of the antitoxin). These zero-shot generations were then passed through the same filtration process as what was used for the fine-tuned generations, allowing for comparison with the fine-tuned models.

Code. All code and models are publicly available, and can be found at:
https://github.com/chang-m-yun/CS273B_TA_Evo

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., ... & Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 1-3.

Akarsu, H., Bordes, P., Mansour, M., Bigot, D. J., Genevaux, P., & Falquet, L. (2019). TASmania: a bacterial toxin-antitoxin systems database. *PLoS Computational Biology*, 15(4), e1006946.

Bryant, P., Pozzati, G., & Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, *13*(1), 1265.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7*(10), e1002195.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., ... & Hassabis, D. (2021). Protein complex prediction with AlphaFold-Multimer. *biorxiv*.

Frieri, M., Kumar, K., & Boutin, A. (2017). Antibiotic resistance. *Journal of Infection and Public Health*, *10*(4), 369-378.

Google Cloud Compute Engine Pricing. https://cloud.google.com/products/calculator

Guan, J., Chen, Y., Goh, Y. X., Wang, M., Tai, C., Deng, Z., ... & Ou, H. Y. (2024). TADB 3.0: an updated database of bacterial toxin–antitoxin loci and associated mobile genetic elements. *Nucleic Acids Research*, *52*(D1), D784-D790.

Hayes, F., & Van Melderen, L. (2011). Toxins-antitoxins: diversity, evolution and function. *Critical Reviews in Biochemistry and Molecular Biology*, *46*(5), 386-408.

Hie, B. L., Shanker, V. R., Xu, D., Bruun, T. U., Weidenbacher, P. A., Tang, S., ... & Kim, P. S. (2024). Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, *42*(2), 275-283.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hwang, Y., Cornman, A. L., Kellogg, E. H., Ovchinnikov, S., & Girguis, P. R. (2024). Genomic language model predicts protein co-regulation and function. *Nature Communications*, *15*(1), 2880.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 1-11.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.

Jurėnas, D., Fraikin, N., Goormaghtigh, F., & Van Melderen, L. (2022). Biology and evolution of bacterial toxin–antitoxin systems. *Nature Reviews Microbiology*, 20(6), 335-350.

Nguyen, E., Poli, M., Durrant, M. G., Thomas, A. W., Kang, B., Sullivan, J., ... & Hie, B. L. (2024). Sequence modeling and design from molecular to genome scale with Evo. *bioRxiv*, 2024.

Lim, Y., Tamayo-Orrego, L., Schmid, E., Tarnauskaite, Z., Kochenova, O. V., Gruar, R., ... & Walter, J. C. (2023). In silico protein interaction screening uncovers DONSON's role in replication initiation. *Science*, *381*(6664), eadi3448.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123-1130.

Page, R., & Peti, W. (2016). Toxin-antitoxin systems in bacterial growth arrest and persistence. *Nature Chemical Biology*, *12*(4), 208-214.

Ruffolo, J. A., Nayfach, S., Gallagher, J., Bhatnagar, A., Beazer, J., Hussain, R., ... & Madani, A. (2024). Design of highly functional genome editors by modeling the universe of CRISPR-Cas sequences. *bioRxiv*.

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 50(D1), D20–D26.

Shao, B. (2023). A long-context language model for the generation of bacteriophage genomes. *bioRxiv*.

Shao, Y., Harrison, E. M., Bi, D., Tai, C., He, X., Ou, H. Y., ... & Deng, Z. (2011). TADB: a web-based resource for Type 2 toxin–antitoxin loci in bacteria and archaea. *Nucleic Acids Research*, 39, D606-D611.

Xie, Y., Wei, Y., Shen, Y., Li, X., Zhou, H., Tai, C., ... & Ou, H. Y. (2018). TADB 2.0: an updated database of bacterial Type II toxin–antitoxin loci. *Nucleic Acids Research*, 46(D1), D749-D753.