Semantic mining of functional *de novo* genes from a genomic language model

Aditi T. Merchant[1,2], Samuel H. King[1,2], Eric Nguyen[1,2], and Brian L. Hie [*,1,3,4]

[1]Arc Institute, Palo Alto, CA; [2]Department of Bioengineering, Stanford University, Stanford, CA; [3]Department of Chemical Engineering, Stanford University, Stanford, CA; [4]Stanford Data Science, Stanford University, Stanford, CA

Abstract

Generative genomics models can design increasingly complex biological systems. However, effectively controlling these models to generate novel sequences with desired functions remains a major challenge. Here, we show that Evo, a 7-billion parameter genomic language model, can perform function-guided design that generalizes beyond natural sequences. By learning semantic relationships across multiple genes, Evo enables a genomic "autocomplete" in which a DNA prompt encoding a desired function instructs the model to generate novel DNA sequences that can be mined for similar functions. We term this process "semantic mining," which, unlike traditional genome mining, can access a sequence landscape unconstrained by discovered evolutionary innovation. We validate this approach by experimentally testing the activity of generated anti-CRISPR proteins and toxin-antitoxin systems, including *de novo* genes with no significant homology to any natural protein. Strikingly, in-context protein design with Evo achieves potent activity and high experimental success rates even in the absence of structural hypotheses, known evolutionary conservation, or task-specific fine-tuning. We then use Evo to autocomplete millions of prompts to produce SynGenome, a first-of-its-kind database containing over 120 billion base pairs of AI-generated genomic sequences that enables semantic mining across many possible functions. The semantic mining paradigm enables functional exploration that ventures beyond the observed evolutionary universe.

1. Introduction

While generative AI promises to accelerate the design of functional biological systems, precisely articulating "function" to a generative model is a challenging and often underspecified task. In natural language, the notion of distributional semantics hypothesizes that the meaning of words can be represented by word co-occurrence, i.e., that "you shall know a word by the company it keeps" (Firth, 1957; Harris, 1954) (**Figure 1A**). In biology, an emerging distributional hypothesis defines the function of a gene through its interactions with other genes, i.e., you shall know a gene by the company it keeps (Kwon et al., 2024).

In prokaryotic genomes, interacting genes are often positioned directly next to each other on the genome sequence in gene clusters or operons (Jacob and Monod, 1961; Overbeek et al., 1999). Researchers have exploited this property, in a process referred to as genome mining via guilt-by-association, to characterize unknown genes that neighbor functionally characterized genes (Aravind, 2000; Galperin and Koonin, 2000), resulting in the discovery of new molecular mechanisms (Lee et al., 2010; Pers et al., 2015) and important biotechnological tools for genome editing, natural product synthesis, and RNA modification (Medema et al., 2011; Shmakov et al., 2015; Doron et al., 2018; Gao et al., 2020; Blin et al., 2021). At its core, guilt-by-association leverages the distributional hypothesis of gene function to perform function-guided discovery.

A capable generative model of genomic sequences could learn a similar, distributional notion of function to perform function-guided design. Progress in long-context machine learning has enabled advanced generative models of genomic sequences at the multi-kilobase and multi-gene scale (Nguyen et al., 2024). These models

*Corresponding author: brianhie@stanford.edu.

are trained to predict the next base pair in a sequence, enabling them to "autocomplete" a partial genomic sequence by generating response sequences to a DNA sequence prompt (**Figure 1B**).

Given the success of guilt-by-association, we reasoned that prompt engineering of a generative genomic model with a sequence of known function could instruct the model to sample novel, functionally related genes in its response. We term this approach "semantic mining," a new paradigm that harnesses multi-gene relationships to generate novel DNA sequences enriched for targeted biological functions. Notably, unlike traditional genome mining, semantic mining is unconstrained by known evolutionary innovation, potentially allowing us to explore entirely new regions of functional sequence space.

Here, we show that Evo, a 7-billion parameter genomic language model, learns a distributional semantics over genes that enables function-guided design of proteins with high novelty. We first demonstrate that Evo enables in-context genomic design, enabling the successful completion of partial sequences of highly conserved genes and operons. We find that a new version of the previously described Evo model (Nguyen et al., 2024) that we pretrained on 50% more data, which we refer to as Evo 1.5, obtains the best performance on gene-level and operon-level completion tasks.

We then apply semantic mining with Evo to design proteins with both high novelty and specified functional activity. We first focus on the generation of toxin-antitoxin systems, which requires designing two genes that encode a protein-protein interaction. Using a sequential design strategy, we first generate novel toxins by prompting with genomic context around known toxin-antitoxin systems. We then use these Evo-generated toxins as prompts to design their conjugate antitoxin pairs, yielding functional toxin-antitoxin systems containing proteins with remote homology (< 30% sequence identity) to known proteins. We observed that 5 out of 10 designed proteins show antitoxin activity, representing a high experimental success rate of 50%.

To assess Evo's generalization capability beyond the design of known evolutionary homologs, we then performed semantic mining of systems with high evolutionary novelty. In particular, we prompt-engineered Evo to generate anti-CRISPR proteins that, in nature, have high diversity in both sequence and mechanism and may be produced by *de novo* gene birth (Eitzinger et al., 2020; Huang et al., 2021; Dong et al., 2022). Strikingly, Evo generates anti-CRISPR proteins that effectively inhibit SpCas9-mediated DNA cleavage despite possessing no clear sequence homology to known proteins and low-confidence structure predictions. Evo also designs variants of known anti-CRISPR proteins with stronger inhibitory activity compared to potent natural anti-CRISPRs, as well as an inhibitor with sequence similarity to a protein family not previously associated with CRISPR inhibition. For this task, we also observed a robust experimental success rate of 17%, with 14 out of 84 designs demonstrating anti-CRISPR activity. Together, these results indicate that Evo may be able to generalize beyond the distribution of known natural protein sequences while retaining higher-level, user-defined biological functions.

Given the experimental success of semantic mining in generating novel functional proteins, we then use Evo to generate SynGenome, the first AI-generated genomics database (`https://evodesign.org/syngenome/`), containing over 120 billion base pairs of synthetic DNA sequences derived from prompts spanning 1,705,667 UniProt entries, 37,108 species, 72,492 protein domains, and 8,914 functional terms. The diversity of coding sequences found in SynGenome recapitulates the natural distribution of protein families, while also containing many generations that go beyond the sequence landscape of the DNA prompts. We make SynGenome and Evo 1.5 openly available to the scientific community to facilitate semantic mining across diverse functions.

In total, Evo can successfully generate proteins with desired functions, often doing so by accessing a highly novel space of sequences that are unlike any known proteins. Semantic mining, with its generalizability and robust success rates, represents a promising new avenue for function-guided generative design. More broadly, we anticipate this work to mark a turning point in biological discovery in which mining of genetic variation becomes substantially augmented by AI-generated sequences that transcend known natural evolutionary constraints.

2. Results

2.1. Evo enables in-context genomic design

To effectively achieve semantic mining for function-guided design, a model must first understand not just individual gene sequences, but how genes relate to and interact with each other within their broader genomic context. Evo, a 7-billion-parameter genomic language model trained on prokaryotic sequences, is well-positioned to address this challenge through its ability to process long genomic sequences at single-nucleotide resolution (Nguyen et al., 2024).

Similar to how words in language derive meaning from their context (**Figure 1A**), DNA sequences take on functional significance when considered in the context of a gene, an operon, a pathway, or an entire organism. Evo's efficient long-context architecture can learn how sequence patterns at the level of individual nucleotides relate to a genomic context containing multiple kilobases. Because functionally related sequences cluster together on prokaryotic genomes, supplying an appropriate functional context could condition Evo to generate new genetic sequences with a desired function (**Figure 1B**).

As an initial experiment, we assessed Evo's ability to leverage genomic context by performing an "autocomplete" task in which we prompted the model with partial sequences of highly conserved prokaryotic genes. We tested three diverse and functionally important genes from bacteria and archaea: RNA polymerase sigma factor *rpoS* from *E. coli*, DNA gyrase subunit A *gyrA* from *S. typhimurium*, and cell division protein *ftsZ1* from *H. volcanii* (Dai and Lutkenhaus, 1991; Chiang and Schellhorn, 2010; Sada et al., 2022) (**Figure 1C**). We also tested three versions of the Evo model: Evo 1 8K, a model pretrained at 8,192 context length; Evo 1 131K, a model made by extending Evo 1 8K to 131,072 context length; and a model newly introduced in this study, which we call Evo 1.5, which extends the pretraining of Evo 1 8K (initially trained on 300 billion tokens) to 450 billion tokens, representing a 50% increase in training data (**Methods**). For each gene, we prompted each Evo model version with varying amounts (30%, 50%, and 80%) of the input sequence and evaluated each model's ability to complete the remainder of the gene.

Evo 1.5 consistently demonstrated the highest recovery of the native protein sequence, particularly at lower prompt lengths. For instance, with just 30% of the input sequence, Evo 1.5 achieved approximately 85% amino acid sequence recovery for *rpoS*, compared to 65% for Evo 1 131K. This performance advantage was maintained across all tested genes and prompt lengths, with Evo 1.5 achieving near-perfect sequence recovery at 80% input for all targets. These experiments are consistent with previous findings that longer pretraining can improve learning of long-range interactions and high-level concepts in sequence models of natural language and proteins (Lin et al., 2023; Kaplan et al., 2020; Brandes et al., 2022). Moving forward, we therefore selected Evo 1.5 for further investigation and all results attributed to Evo in this study were produced by the Evo 1.5 model.

We further tested Evo's understanding of genomic context at a multi-gene scale by evaluating its ability to predict gene sequences based on operonic neighbors (**Figures 1D** and **S1**). We prompted the model with sequences of genes either upstream or downstream of target genes in the well-characterized *trp* and *modABC* operons (Maupin-Furlow et al., 1995; Merino et al., 2008), leveraging DNA complementarity to generate sequences bidirectionally. Evo demonstrated robust predictive performance across all tested configurations, achieving over 80% protein sequence recovery for all target genes. Further, the model exhibited adaptability to genomic orientation, successfully generating downstream gene sequences when prompted with reverse complement sequences of upstream genes, and vice versa, while maintaining appropriate directionality across the operon. Notably, even with incomplete sequence recovery, the predicted protein structures of Evo's generations closely matched the native protein's conformations, suggesting preservation of functionally critical amino acid residues and structural motifs. These results indicate that Evo not only learns the primary sequence of genes but also captures the broader genomic organization of bacterial operons.

To assess whether Evo's generations go beyond trivial memorization of training sequences, we analyzed the per-position entropy of both amino acid and nucleotide sequences in the model's generations (**Figure 1E**). Using the *modABC* operon generations as a test case, we prompted the model with the sequence encoding *modA* from *E. coli* K-12 and analyzed the variability in the generated *modB* responses. The amino acid-level entropy analysis revealed selective conservation, with generally lower entropy at key structural positions and higher variability in less conserved regions. This pattern aligns with natural protein evolution, where functionally critical residues show high conservation while non-essential positions permit amino acid substitutions.

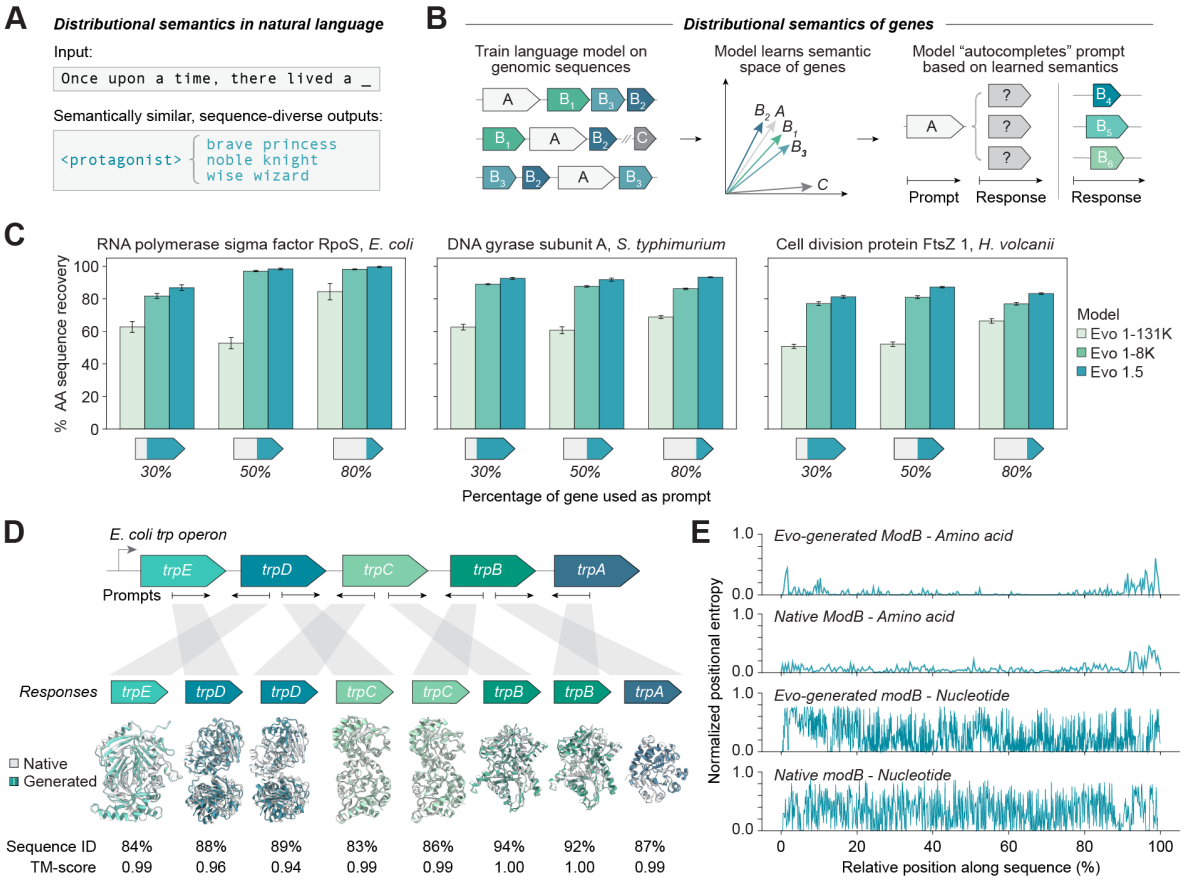


Figure 1 | **In-context genomic modeling and design with Evo. (A)** In natural language, distributional semantics leverages the property that functionally equivalent but lexically distinct words often occupy similar positions in a text with common sets of neighboring words. **(B)** In semantic mining, a genomic language model trained across multiple genes learns to map functionally related genes to similar semantic spaces, enabling generation of functionally related but sequence-diverse genes. **(C)** Sequence recovery assessment when using a genomic language model to autocomplete three conserved prokaryotic genes shows consistent improvement in performance from Evo 1 131K and Evo 1 8K to Evo 1.5, demonstrating enhanced ability to leverage genomic context. AA, amino acid; Bar height, mean; error bars, standard error; $n = 100$. **(D)** Bidirectional completion of conserved *E. coli trp* operon gene sequences demonstrates high sequence identity and structural conservation across the operon. **(E)** Positional entropy comparison between aligned natural and aligned Evo-generated ModB sequences at nucleotide and amino acid levels shows conservation of essential amino acid residues while maintaining high nucleotide diversity. $n = 500$.

At the nucleotide level, we observed substantially higher entropy across all positions, with variation even in regions coding for conserved amino acids. Given that a single prompt was used to generate all the response sequences, these results suggest that Evo is not simply reproducing memorized sequences but is rather synthesizing information from across its training set to reflect biological constraints while generating diversity in a manner reminiscent of natural evolution.

2.2. Semantic mining of functional protein-protein interactions with remote homology to nature

The ability of Evo to understand genomic context consisting of highly conserved genes and operons led us to explore if we could apply semantic mining to less well-conserved biology: phage and bacterial defense systems. These systems, which are shaped by the unrelenting evolutionary arms race between bacteria and phage (Murtazalieva et al., 2024), are some of the most rapidly evolving systems in nature. As a result, defense systems exhibit vast amounts of functional diversity while maintaining minimal sequence conservation across species (Beavogui et al., 2024; Tesson et al., 2024). Interestingly, defense systems of similar functionality also frequently cluster into defense islands, enabling the discovery of novel systems through guilt-by-association approaches (Makarova et al., 2011; Koonin, 2018). To date, some of the most fruitful discoveries of genome mining have been derived from defense systems, including CRISPR-Cas systems, restriction enzymes, and various antimicrobial compounds (Doron et al., 2018; Gao et al., 2020; Picton et al., 2021; Altae-Tran et al., 2023; Cheng et al., 2024). The remarkable success in repurposing these naturally occurring systems for biotechnology applications underscores the vast potential of bacterial defense mechanisms as a source of molecular tools.

Given the natural diversity and genomic co-localization of these systems, we sought to determine if semantic mining could be used to design new defense systems. As an initial test case, we focus specifically on type II toxin-antitoxin (T2TA) systems, some of which play a role in phage defense. Type II toxin-antitoxins consist of a stable toxin protein that can inhibit cell growth or induce cell death under conditions of stress, such as a phage infection, paired with an unstable antitoxin protein that binds and neutralizes the toxin's activity under homeostatic conditions (**Figures 2A,B**). These systems often maintain conserved high-level genomic architecture despite sequence divergence, with toxin and antitoxin genes typically arranged in adjacent positions (Chan et al., 2023; Guan et al., 2024).

To generate diversified T2TAs using Evo 1.5, we developed a prompting strategy that leveraged the characteristic co-localization of these systems (**Figure 2A**). In addition to prompting with sequences encoding known T2TAs, we compiled genomic sequences directly upstream and downstream of known T2TA systems. Following sampling using the compiled prompts, we filtered generations for sequences encoding toxin-antitoxin protein pairs that exhibited *in silico* predicted protein complex formation with AlphaFold 3 (Abramson et al., 2024) yet had limited sequence homology to known T2TA proteins.

We first tested Evo-generated toxins using a growth inhibition assay, in which growth arrest indicated successful toxin activity (**Methods**). From these experiments, we were able to identify a functional bacterial toxin, which we call EvoRelE1, that exhibited strong growth inhibition (approximately 70% reduction in relative survival) while possessing 71% sequence identity to a known RelE toxin (**Figures 2C,D**).

We subsequently prompted Evo 1.5 with the sequence of EvoRelE1, hypothesizing that the model could use the genomic context of the toxin sequence to generate a conjugate antitoxin (**Figure 2A**). Using this strategy, we identified a set of 10 antitoxin candidates exhibiting little to no sequence homology to natural protein sequences that were tested for their ability to block the toxin activity of EvoRelE1.

Following co-expression with EvoRelE1, half of the Evo-generated antitoxin candidates were able to rescue cell growth (**Figures 2D** and **S2**). The most effective candidates, designated as EvoAT1 and EvoAT2, restored growth to 95-100% of normal cell survival, with candidates EvoAT3 and EvoAT4 demonstrating moderate rescue activity (70% and 90% relative survival, respectively). Sequence analysis of the successful antitoxins revealed discrete regions of conservation (**Figure 2E**), potentially highlighting key motifs required for toxin neutralization despite their overall sequence diversity. Furthermore, when tested against native RelE, MazF, and YoeB toxin systems, several of the Evo-generated antitoxins were able to rescue growth, with EvoAT2 showing inhibitory activity against all three toxins and EvoAT4 rescuing growth against native RelE and YoeB toxins (**Figure 2F**). This observation suggests that the EvoATs may be able to inhibit multiple toxins despite sharing limited sequence homology with their natural antitoxin counterparts (**Figures 2F** and **S3**). Interestingly, when co-folded using *in silico* structure prediction methods, several of the EvoATs had low-confidence pre-

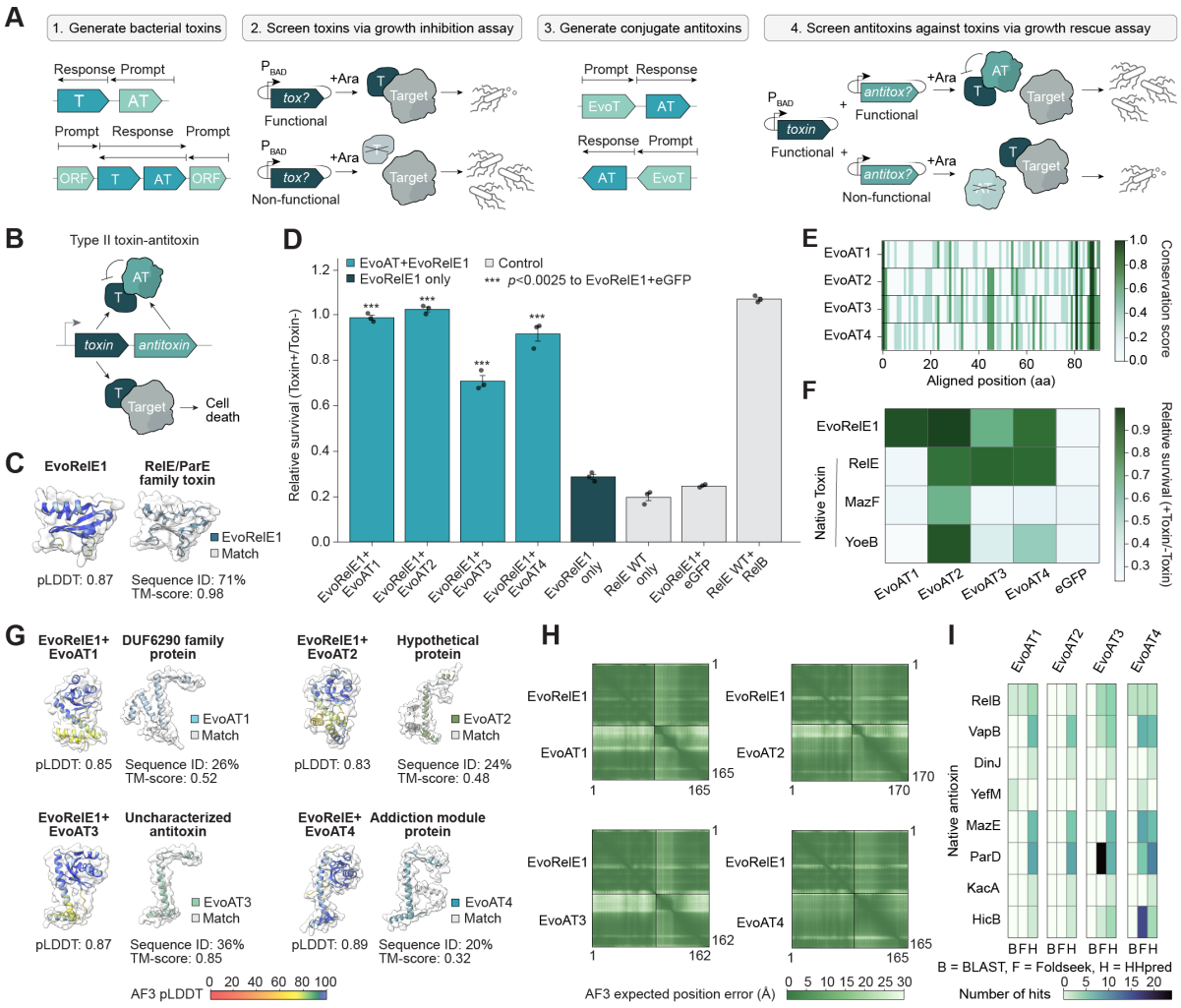


Figure 2 | **Evo generates functional toxin-antitoxin protein-protein interactions with remote homology to nature.** **(A)** To perform semantic mining of toxin-antitoxin systems, toxins (T) are first generated using toxin genomic context as prompts and tested via growth inhibition assay. Successful toxins (EvoT) then serve as prompts to generate cognate antitoxins (AT), which are finally validated through a growth recovery assay. **(B)** Type II toxin-antitoxin systems function via antitoxin binding and neutralization of toxin activity to prevent cell death. **(C)** AlphaFold 3 structure prediction comparison between functional toxin EvoRelE1 and its closest sequence match. pLDDT, predicted local distance difference test score; TM-score, template modeling score. **(D)** Bar plot showing relative survival rates for different Evo-generated toxin-antitoxin combinations. EvoAT1-4 achieve high growth rescue against EvoRelE1. Bar height, mean; error bars, standard error; circles, biological replicate values; $n = 3$. Statistical significance determined by one-sided Student's t-test comparing to EvoRelE1+eGFP ($^{***}p < 0.0025$) **(E)** Homology analysis of aligned amino acid sequences for EvoAT1-4, with darker green indicating higher conservation. **(F)** Heat map showing relative survival rates of EvoAT1-4 tested against native toxins. EvoAT2 and EvoAT4 appear to have inhibitory activity against multiple natural toxins. **(G)** AlphaFold 3 structure predictions of EvoAT1-4 in complex with EvoRelE1 and structure comparisons between Evo-generated antitoxins and their closest natural matches. EvoAT1-4 exhibit high confidence predicted structures despite limited sequence homology to their closest protein matches. **(H)** AlphaFold 3 predicted aligned error plots for EvoAT1-4 in complex with EvoRelE1 showing high confidence in predicted structures and interactions. **(I)** Heatmap showing number of structural and sequence matches between Evo-generated antitoxins and known antitoxin families. B, BLAST; F, FoldSeek; H, HHpred.

dicted complex formation with the natural toxins that they inhibited (EvoAT4 + YoeB ipTM = 0.14, EvoAT2 + YoeB ipTM = 0.47, EvoAT2 + MazF ipTM = 0.34) (**Figure S4**), highlighting the potential for semantic mining to identify novel molecular interactions that extend beyond the domain of state-of-the-art structure prediction models.

EvoAT1 through EvoAT4 all have sequence identities with natural proteins that fall near the "twilight zone" of sequence similarity (20% to 36%), a range where sequence similarity alone is insufficient to reliably predict shared structure or function (Rost, 1999). Both EvoAT1 and EvoAT2 only exhibit homology to proteins not actively annotated as antitoxins, showing just 26% sequence identity to a DUF6290 family protein from *Actinomyces* and 24% sequence identity to an uncharacterized *Magnetococcus sp*. YQC-5 hypothetical protein respectively (**Figure 2G**). Given that EvoAT2's closest match is an uncharacterized hypothetical protein that appears in a similar genomic context (**Figure S5**), our results suggest this protein may function as part of a toxin-antitoxin system. Similarly, the homology between EvoAT1 and the DUF6290 family reinforces a hypothesized potential role for this domain of unknown function in antitoxin activity (Blum et al., 2024). These findings suggest that Evo's understanding of genomic context enables functional annotation of previously uncharacterized proteins, as well as function-conditioned design that is unconstrained by existing annotations. The structural predictions for EvoAT1 and EvoAT2 revealed high-confidence folds (pLDDT scores of 0.85 and 0.83, respectively) (**Figure 2G**), and when co-folded with EvoRelE1, both antitoxins exhibited minimal predicted position error (**Figure 2H**). These structural characteristics and high functional activity are particularly noteworthy given the antitoxins' limited sequence homology to known antitoxin proteins (**Figure 2I**). Overall, these results further underscore the ability of Evo to generate functional proteins with remote homology to natural proteins.

Sequence and structural homology searches using BLAST, HHpred, and Foldseek (van Kempen et al., 2024; Söding et al., 2005; Gabler et al., 2020; Zimmermann et al., 2018; Sayers et al., 2022) revealed that the EvoATs show similarity to multiple independent antitoxin superfamilies, particularly ParD, MazE, and VapB (**Figure 2I**). This finding, coupled with the activity of EvoAT2 and EvoAT4 against multiple toxins, is especially notable because their cognate toxins employ fundamentally different mechanisms of action. In natural systems, these mechanistic differences have typically led to the evolution of distinct antitoxin structures specialized for neutralizing each type of toxin (Fraikin et al., 2020; Qiu et al., 2022). This suggests that Evo may have identified a broader functional compatibility between antitoxin and toxin families than previously recognized, highlighting the potential of semantic mining to reveal new insights into protein-protein interaction patterns in prokaryotic systems.

2.3. Semantic mining of functional *de novo* genes

Given the experimental success of Evo-generated functional proteins with remote homology to nature, we next explored whether semantic mining could learn to generate even greater evolutionary novelty. Anti-CRISPRs (Acrs) are proteins used by phages to neutralize bacterial CRISPR-Cas immune systems. Given their critical role in helping phages evade CRISPR-Cas mediated cleavage (**Figure 3A**), Acrs represent one of the most striking examples of rapid protein evolution in bacterial-phage arms races, with many Acrs appearing to be novel innovations lacking detectable homology to other protein families, including other Acrs (Davidson et al., 2020; Eitzinger et al., 2020; Huang et al., 2021). This remarkable diversity is reflected in their varied mechanisms of action, from direct CRISPR-Cas protein binding to DNA mimicry to transcriptional silencing (Shin et al., 2017; Camara-Wilpert et al., 2023; Choudhary et al., 2023; Trost et al., 2024), making them valuable tools for understanding protein evolution and developing synthetic control systems for CRISPR technologies (Pawluk et al., 2018).

Despite their sequence diversity, many Acr operons maintain a somewhat conserved operonic architecture, consisting of multiple Acr genes appearing together alongside neighboring anti-CRISPR associated (*aca*) genes (Stanley et al., 2019) (**Figure 3B**). This architectural conservation, combined with their frequent emergence as *de novo* genes, makes Acrs an ideal test case for assessing semantic mining's ability to generalize with respect to sequence and structure while retaining a desired higher-level function.

Using genomic sequences from known Cas9-targeting Acr genes and their associated operons as prompts, we used Evo 1.5 to generate 3,160 sequences totaling 3.16 million base pairs. After filtering for size, complexity, and secondary structure, we identified 468 potential Acr open reading frames (ORFs). To handle the high sequence diversity of Acr proteins complicating traditional MSA and structural filtering methods, we next

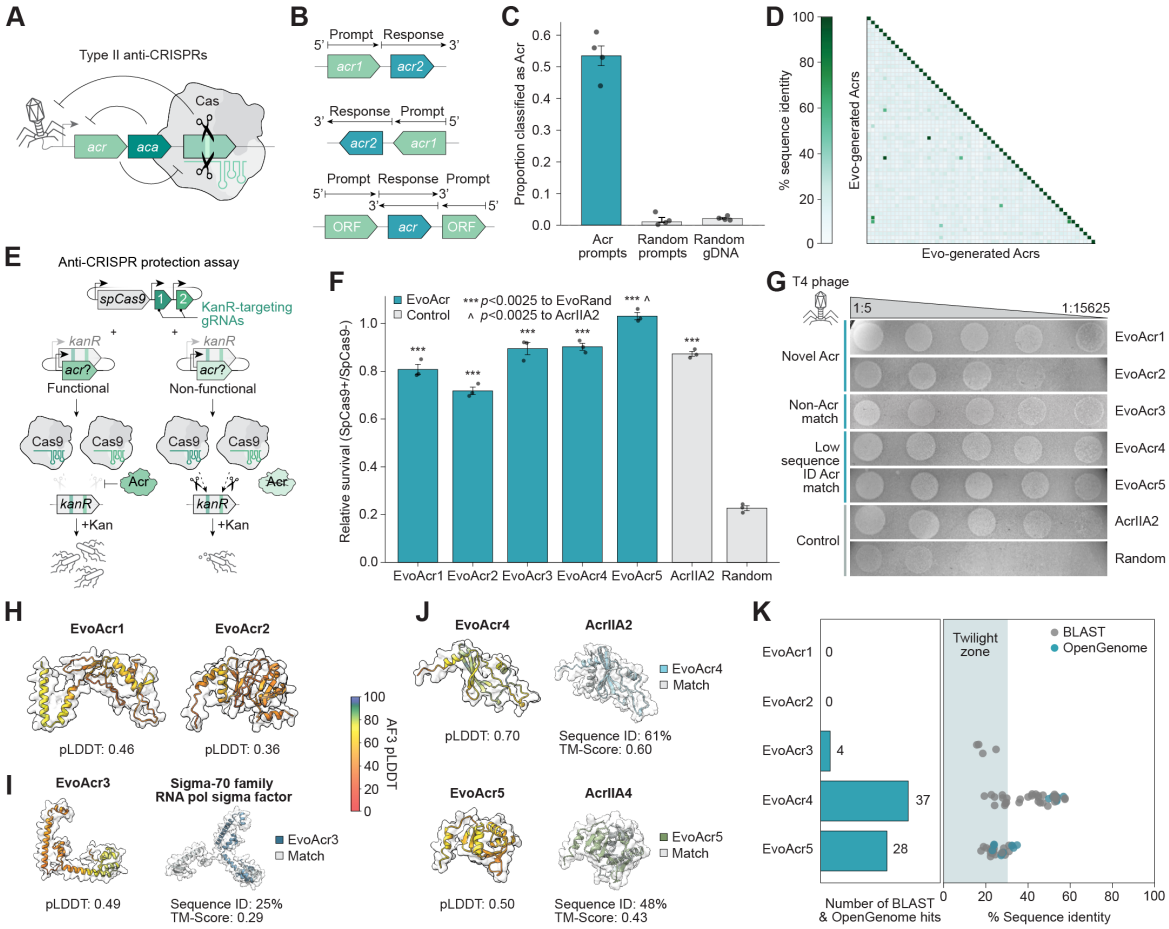


Figure 3 | **Evo generates functional anti-CRISPR proteins with no homology to known proteins.** **(A)** Type II anti-CRISPR systems involve an anti-CRISPR (*acr*) gene that encodes an Acr protein that inhibits type II-A Cas nuclease activity, often co-occurring with anti-CRISPR associated genes (*aca*). **(B)** To perform semantic mining of Acrs, known genomic contexts of type II anti-CRISPRs, including potential *aca* genes, are used as prompts. **(C)** PaCRISPR classification demonstrates significant enrichment of Acr-like sequences in generations from Acr prompts compared to controls. Bar height, mean; error bars, standard error; circles, different batches of generations; $n = 4$. **(D)** Sequence identity matrix demonstrating high diversity among random sample of Evo-generated Acrs. **(E)** In the anti-CRISPR protection assay, functional Acr proteins prevent SpCas9-mediated cleavage of a kanamycin resistance gene, enabling cell survival in kanamycin-supplemented media, while non-functional candidates allow SpCas9 targeting and cell death. **(F)** Bar plot showing relative survival rates for Evo-generated Acrs. EvoAcr1-5 achieve robust protection against SpCas9, with EvoAcr5 exceeding the activity of the AcrIIA2 positive control. Bar height, mean; error bars, standard error; circles, biological replicate values; $n = 3$. Statistical significance is determined by a one-sided Student's t-test compared to random control *(***$p < 0.0025$)* and AcrIIA2 ($\hat{}p < 0.0025$). **(G)** T4 phage plaque assays validating anti-CRISPR activity. *E. coli* cells co-expressing SpCas9 targeted to T4 phage and candidate Acrs were embedded in soft agar, and serial dilutions of T4 phage were spotted onto the bacterial lawn. Plaque formation indicates successful anti-CRISPR protection. Experiment performed in triplicate with representative images shown. **(H)** AlphaFold 3 structure predictions for EvoAcr1-2 showing low confidence predictions. pLDDT, predicted local distance difference test score; TM-score, template modeling score. **(I)** AlphaFold 3 predicted structure of EvoAcr3 and structure alignment between EvoAcr3 and its closest sequence match, revealing limited homology to a sigma-70 family protein. **(J)** AlphaFold 3 structure predictions comparing EvoAcr4-5 with their closest Acr sequence matches, showing moderate structural and sequence similarity. **(K)** Sequence similarity analysis of EvoAcr1-5 calculated by BLAST search against BLAST-nr and MMseqs2 search against OpenGenome. EvoAcr1-2 demonstrate no significant homology to known proteins, EvoAcr3 shows matches only in the "twilight zone" of sequence similarity (Rost, 1999), and EvoAcr4-5 show limited sequence matches to known Acrs.

used PaCRISPR, a machine learning model trained to identify potential Acr proteins, to evaluate our generated candidates (Wang et al., 2020). Consistent with our prompts providing successful functional conditioning, generations derived from Acr-containing genomic contexts were substantially more likely to be classified as potential Acrs by PaCRISPR compared to negative control sequences (**Figure 3C**). Approximately 55% of minimally filtered proteins generated from Acr-associated prompts were classified as likely Acrs, compared to less than 5% of proteins generated from either random prompts or random genomic sequences. Further, the distribution of sequence identities among the predicted Acrs showed a wide range of novelty, with most candidates showing low similarity to each other (**Figure 3D**) (median pairwise sequence identity = 23%). This enrichment for diverse Acr-like sequences demonstrates that semantic mining can effectively bias generations toward a desired function by leveraging genomic context alone, even in the absence of clear sequence or structural signatures.

To test the Acr candidates' protection ability against SpCas9, we co-transformed *E. coli* with plasmids encoding our candidate Acrs and a CRISPR-targeted kanamycin resistance gene, where functional Acr protection would enable bacterial growth in kanamycin-supplemented media through inhibition of CRISPR-mediated cleavage (Forsberg et al., 2019) (**Figure 3E**). From this initial experiment, we found that 17% of tested sequences exhibited some degree of anti-CRISPR activity (**Figure S6**), a high success rate when considering the lack of structure or sequence homology-guided design. From this pool, we further identified five promising proteins (EvoAcr1-5) that demonstrated strong protection against SpCas9-mediated targeting in both liquid culture survival assays (**Figure 3F**) and phage infection experiments (**Figures 3G** and **S7**).

Detailed bioinformatic analysis of these five Acrs revealed a high level of diversity in their sequence origins. EvoAcr4 and EvoAcr5 share moderate sequence similarity to known Acrs (61% and 48% to AcrIIA2 and AcrIIA4, respectively) and demonstrated robust protection against SpCas9, with EvoAcr5 showing activity exceeding that of the positive control AcrIIA2 in liquid culture assays (relative survival rates of 1.01 to 0.87 respectively) (**Figures 3F,H**). EvoAcr3 presented an intriguing case: while sharing only "twilight zone" sequence identity (Rost, 1999) and limited structural alignment (sequence ID = 25%, TM-score = 0.29) with a sigma-70 family RNA polymerase sigma factor, it maintained strong anti-CRISPR activity (relative survival of 0.89) (**Figures 3F,I**). This suggests a potential mode of CRISPR inhibition that is not described by existing functional databases.

Most notably, EvoAcr1 and EvoAcr2 represented proteins that eluded both sequence and structural homology characterization, showing no significant identity to proteins in OpenGenome (the Evo training data) or to known proteins in the nr database by BLAST (**Figures 3J,K**). The AlphaFold 3-predicted structures of EvoAcr1 and EvoAcr2 also have low confidence pLDDT scores of 0.46 and 0.36, respectively (**Figures 3H** and **S8**). Despite this lack of sequence or structural similarity to known proteins, they demonstrated robust protection in both liquid culture and phage infection assays, with relative survival rates of 0.82 and 0.74, respectively. This experimental validation of novel, functional Acrs supports the ability of semantic mining to access entirely unexplored regions of sequence space while maintaining specific desired functions. Together with EvoAcr3-5, these results demonstrate that genomic context can guide the generation of diverse anti-CRISPR proteins, from sequence variants of known architectures to entirely new protein sequences.

2.4. SynGenome: A first-of-its-kind synthetic genomic database

Following our experimental validation that semantic mining can generate novel, functional proteins from genomic context alone, we reasoned that this approach could be systematically applied to generate functional genes across prokaryotic biology. To this end, we developed SynGenome, a comprehensive database of synthetic DNA sequences generated using Evo. In line with the function-guided design principles underlying semantic mining, we prompted the model with sequences derived from over 1.5 million prokaryotic and phage proteins to generate a large set of diverse sequences that could be enriched for novel genes functionally related to the prompt.

To construct SynGenome, we leveraged the UniProt database to identify protein-coding genes and their adjacent sequences from prokaryotic organisms and bacteriophages. For each coding sequence, we generated six distinct prompts: the upstream region, coding sequence (CDS), downstream region, and their respective reverse complements (**Figure 4A**). Using the Evo 1.5 model with standard autoregressive decoding, we generated multiple synthetic sequences for each prompt, resulting in a database containing over 120 billion DNA base pairs (**Methods**).

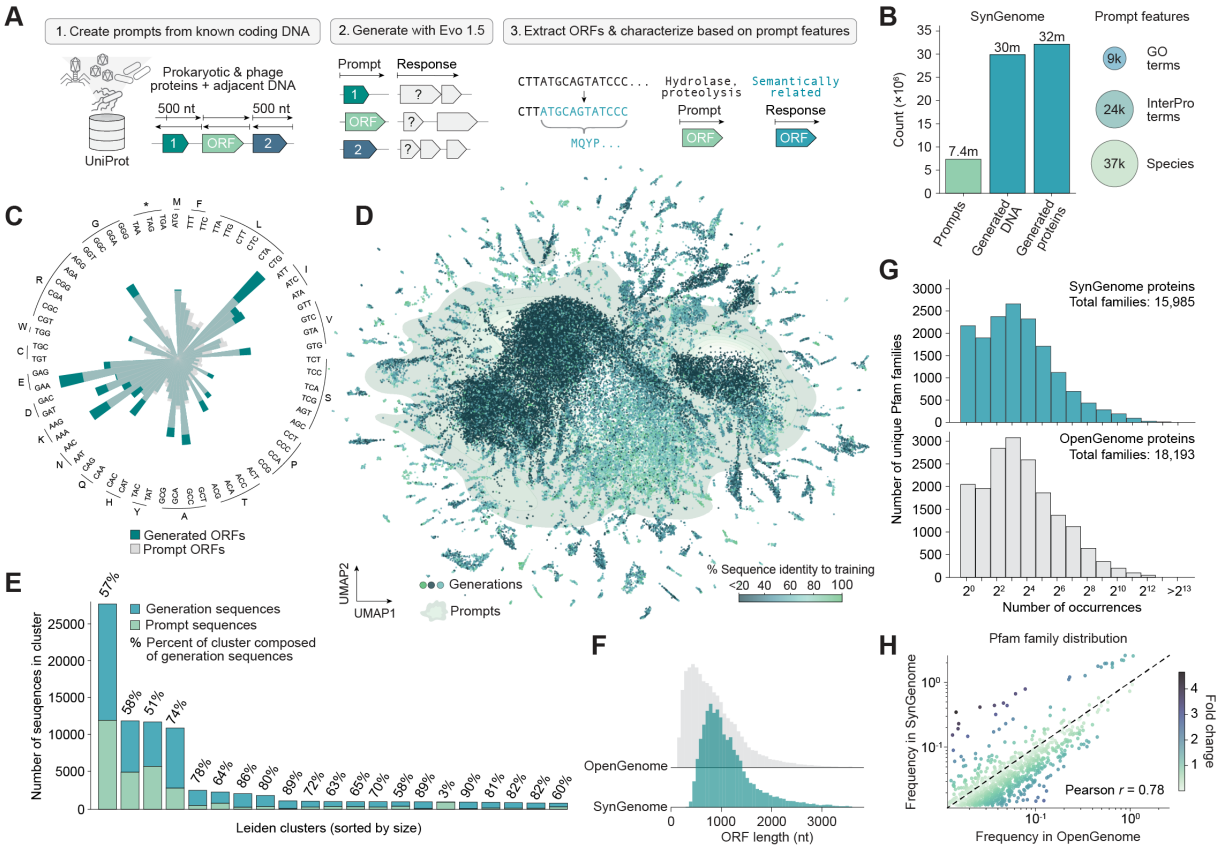


Figure 4 | **120 billion base pairs of AI-generated genomic sequences with SynGenome. (A)** To construct SynGenome, we created prompts from known-protein coding genes, generated synthetic DNA sequences using Evo, and bioinformatically characterized the generated sequences. **(B)** Number of prompts and generated sequences in SynGenome, and their associated features. GO, gene ontology. **(C)** Codon usage bias of Prodigal-predicted ORFs in prompt sequences and generated sequences, demonstrating preservation of codon preferences. **(D)** UMAP of jointly projected Evo embeddings of SynGenome and prompt sequences. Generation embeddings are colored by sequence identity to training data and prompt embeddings are shown in light green. **(E)** Bar plot showing relative proportions of generation and prompt sequences in most populous Leiden clusters, with percentages indicating fraction of generated sequences per cluster. **(F)** Prodigal-predicted ORF lengths in for 5000 nt representative samples of OpenGenome and SynGenome, with ORFs showing similar length distributions. nt, nucleotide. **(G)** Histogram showing number of Pfam protein family occurrences in SynGenome and OpenGenome, with both following a similar long-tailed distribution of family abundance. **(H)** Scatterplot showing correlation between protein family frequencies in SynGenome versus OpenGenome, demonstrating preservation of natural protein family abundance patterns.

To facilitate functional exploration of the database, we organized the synthetic sequences according to the Gene Ontology (GO) terms and InterPro domain annotations associated with their corresponding prompts (**Figure 4B**), as the generated sequences are likely to be enriched for functionally related elements. As generating sequences with language models can often be prone to repetitive generations, we implemented minimal filtering to remove highly repetitive low complexity sequences, for example, long stretches of sequence containing only a single base pair (**Methods**).

To characterize the generations in SynGenome, we first examined codon usage patterns between the generated sequences and the prompts. This analysis revealed that generated sequences closely mirrored their prompts, with the synthetic sequences maintaining similar codon preferences to natural coding sequences (**Figure 4C**). We further examined the relationship between prompts and their corresponding generated sequences in Evo embedding space (**Figure 4D**) by performing Leiden clustering and analyzing the relative proportions of each sequence type within clusters (**Figures 4E** and **S9**). Most Leiden clusters contain a mixture of generated sequences and natural prompt sequences, indicating that many generations occupy similar regions of the language model embedding space. However, we observed that 54 clusters representing 19% of the generated sequences are almost completely composed of generated sequences, potentially indicating regions of synthetic sequences that extend beyond the semantic space represented by natural sequence embeddings.

When compared to sequences from OpenGenome, we found that SynGenome-generated ORFs closely mirrored the natural prokaryotic ORF length distribution predicted by Prodigal (Hyatt et al., 2010) (**Figure 4F**). At the protein level, SynGenome matched both the global distribution of Pfam domain frequencies found in natural proteins (**Figure 4G**), with specific Pfam families also showing similar occurrence frequencies between synthetic and natural sequences (Mistry et al., 2021; Blum et al., 2024) (Pearson correlation coefficient, $r = 0.78$, (**Figure 4H**). These analyses demonstrate that SynGenome recapitulates the architecture and functional diversity of natural prokaryotic sequences.

Collectively, these results highlight the potential for SynGenome to become a valuable tool for exploring and expanding protein function through semantic relationships. By mining SynGenome, researchers could discover functionally related proteins that extend beyond natural sequence space, gain insights into potential functions of uncharacterized proteins based on genetic co-occurrence patterns, and create diverse screening libraries for engineering desired functions (**Figure S10**).

SynGenome is freely available as a public resource at `https://evodesign.org/syngenome/`, where researchers can search the database using protein names, UniProt IDs, InterPro domains, species names, or GO terms of interest. This allows for rapid identification of relevant synthetic sequences, enabling researchers to quickly find and explore generations that align with their research interests. We anticipate that SynGenome can serve as a practical tool that facilitates gene discovery and engineering with semantic mining for the broader scientific community.

3. Discussion

Advanced genomic sequence models, trained on hundreds of billions of DNA base pairs across the diversity of prokaryotic life, can enable unprecedented capabilities. We demonstrate that Evo enables controllable design of desired functions when prompted with sequences that are meant to elicit those functions, with high experimental success rates of 17-50% when testing only tens of variants in the laboratory. Many of the designed proteins also have no significant homology to proteins of similar function or, in some cases, to any known protein. These results blur the line between "*de novo*" protein design (Huang et al., 2016; Watson et al., 2023; Kortemme, 2024) and diversification based on evolutionary models (Madani et al., 2023; Hayes et al., 2024; Hie et al., 2024), providing an "existence proof" that sequence models without any structural conditioning can meaningfully generalize beyond natural sequence space.

Semantic mining represents a fundamentally new paradigm for protein design that is highly complementary to existing approaches. First, unlike design methods based on task-specific fine-tuning (Madani et al., 2023; Jiang et al., 2024; Nguyen et al., 2024), semantic mining requires no additional training or supervision that could bias generations toward the distribution of known examples. Second, in contrast with approaches that propose specifying function through natural language descriptions found in existing knowledge bases (Praljak et al., 2024), semantic mining accesses the vast functional diversity embedded within genomic se-

quences. This enables functional design that can leverage biological processes that are not yet documented in the scientific literature. For example, we generate antitoxins that suggest broader functional compatibility between diverse toxin and antitoxins systems than previously recognized (**Figure 2F**) and a protein with anti-CRISPR activity that maps to a protein family with a different putative function (**Figure 3I**). Third, by leveraging genomic context as functional conditioning, semantic mining does not require existing structural or mechanistic hypotheses; indeed, protein design pipelines that filter out low-confidence structure predictions (Verkuil et al., 2022; Pacesa et al., 2024) would have removed most of our high-activity designs. Genomic conditioning is also useful when specifying functions such as anti-CRISPR activity that could be accomplished by many structures and mechanisms (Choudhary et al., 2023). Semantic mining of genomic language models therefore represents a powerful and orthogonal approach to current biological design strategies.

Traditional genome mining via guilt-by-association, which motivates many of the ideas in this study, is constrained to discovered evolutionary diversity that was generated over billions of years. In contrast, semantic mining enables the generation of a near limitless supply of sequence diversity for a biological system of interest. To facilitate broader accessibility of this new source of evolutionary material, we report SynGenome, a database of over 120 billion base pairs of AI-generated genomic sequences, which we make publicly available. This resource enables researchers, especially those without the computing resources required to conduct large scale sampling from a generative model, to mine for sequences related to their function of interest. This data could potentially contain new molecular tools and provide insights into protein function and evolution.

While semantic mining represents a new level of sequence novelty and functional improvement for generative genomics, several fundamental limitations and challenges remain. Autoregressive generation is prone to sampling repetitive sequences or to hallucinating realistic but non-functional designs. Semantic mining therefore requires both *in silico* filtering and experimental testing to validate downstream functions. Semantic mining is also theoretically limited to functions that are found in nature, particularly those that are encoded in prokaryotic organisms. However, we note that only a small fraction of this functional diversity has been characterized by the biological literature and that mining of prokaryotic functional diversity has directly led to powerful biotechnologies such as polymerase chain reaction (PCR), optogenetics, genome editing, and genetic recombination (Chien et al., 1976; Nagel et al., 2003; Boyden et al., 2005; Govorunova et al., 2017; Altae-Tran et al., 2023; Durrant et al., 2024).

Looking forward, increased scaling of pretrained models and availability of training data from rapidly growing sequencing resources will likely reinforce the capabilities of semantic mining. We also anticipate combining the rich information learned by pretrained models with more advanced inference-time strategies will improve generation quality. Analogous to generative models of images that exhibit improved controllability when given prompts that are themselves generated by a natural language model (Esfandiarpoor and Bach, 2024; Mañas et al., 2024), synthetic prompts could be used to improve the generation quality of genomic language models as well; for example, EvoAT1-4 were all generated in the context of the previously designed EvoRelE1. Further, given the utility of chain-of-thought prompting in natural language (Wei et al., 2023), genomic language models that sequentially generate multicomponent systems could accelerate the development of synthetic biological circuits, molecular machines, metabolic pathways, or even complete genomes. Lastly, techniques for mechanistic interpretability of large language models (Bricken et al., 2023; Templeton et al., 2024) or post-training that leverages high-throughput experimental data collection (Ouyang et al., 2022; Rafailov et al., 2024; Rai et al., 2024) could improve the quality and diversity of generated functions. By leveraging semantic mining, exploration of synthetic genomic space may reveal biological discoveries that complement and extend beyond those discovered in natural organisms.

4. Code availability

Code for generation with Evo based on a user-defined prompt is available at https://github.com/evo-design/evo/. Scripts for semantic mining of toxin-antitoxin systems and anti-CRISPRs are available https://github.com/evo-design/evo/tree/main/semantic_mining. The Evo 1.5 model is available under an open source license at Hugging Face: https://huggingface.co/evo-design/evo-1.5-8k-base.

5. Data availability

SynGenome is explorable and searchable at https://evodesign.org/syngenome/. Raw data has been deposited to Hugging Face datasets at https://huggingface.co/datasets/evo-design/syngenome-uniprot.

6. Materials availability

DNA, RNA, and protein sequences used during our validation experiments are available in **Data S1**. All newly created materials are available upon reasonable request to the corresponding authors.

7. Acknowledgements

We thank Kevin Forsberg, Regan Russell, and Samantha Sakells for supplying plasmid backbones and for helpful discussions on anti-CRISPR experiments. We thank Matthew Durrant, Patrick Hsu, Julia Kazaks, David Li, Talal Widatalla, Brandon Ameglio, Sergey Ovchinnikov, April Pawluk, Brian Plosky, and Chiara Ricci-Tam for helpful discussions and assistance with manuscript preparation. A.T.M. acknowledges funding support from the Knight-Hennessy Graduate Scholarship Fund. A.T.M. and S.H.K. acknowledge funding support from the National Science Foundation Graduate Research Fellowship Program. B.L.H. acknowledges funding support from Arc Institute, Stanford Institute for Human-Centered Artificial Intelligence (HAI) Hoffman-Yee Research Grants, V. Gupta, and R. Tonsing.

8. Author Contributions

A.T.M., S.H.K., and B.L.H. conceived the project. B.L.H. supervised the project. E.N. and B.L.H. trained Evo 1.5. A.T.M. developed code for and performed the gene completion, operon completion, and entropy analyses. A.T.M. and S.H.K. compiled prompts for the Acr sampling. A.T.M developed code for and performed the analysis and filtering of the Acr candidates. A.T.M and S.H.K. experimentally tested the Acr generations. A.T.M compiled prompts for the toxin-antitoxin sampling. A.T.M developed code for and performed the analysis and filtering of the toxin-antitoxin candidates. A.T.M experimentally tested the toxin-antitoxin generations. A.T.M compiled the prompts for SynGenome. A.T.M and B.L.H. developed the code for and performed the sampling for SynGenome. B.L.H. developed the SynGenome website. A.T.M. S.H.K., and B.L.H. wrote the first draft of the manuscript. All authors wrote the final draft of the manuscript.

9. Competing Interests

B.L.H. acknowledges outside interest in Prox Biosciences as a scientific co-founder. A.T.M, S.H.K., and B.L.H. are named on a provisional patent application applied for by Stanford University and Arc Institute related to this manuscript. All other authors declare no competing interests. All other authors declare no competing interests.

10. Methods

Evo 1.5 Pretraining

We extended the pretraining of the Evo 1 model that was trained at a sequence context of 8,192 tokens with an initial learning rate of 0.003 after a warmup of 1,200 iterations; a cosine decay schedule with a maximum decay iterations of 120,000 and a minimum learning rate of 0.0003; a global batch size of 4,194,304 tokens; and 75,000 iterations (processing a total of 315 billion tokens). Other details on hyperparameters related to the model architecture and optimizer can be found in Nguyen et al. (Nguyen et al., 2024). We resumed the model's pretraining, including all model states, optimizer states, data loading schedule, and learning rate schedule, from 75,000 iterations to 112,000 iterations (processing a total of 470 billion tokens). We refer to the model trained up to 112,000 iterations as Evo 1.5.

Autoregressive sampling

To sample from Evo, we used a standard low-temperature autoregressive sampling algorithm (Chang and Bergen, 2023). We used the sampling code at `https://github.com/evo-design/evo/` based on the reference implementation at `https://github.com/togethercomputer/stripedhyena` that leverages kv-caching of Transformer layers and the recurrent formulation of hyena layers (Massaroli et al., 2023; Poli et al., 2023) to achieve efficient, low-memory autoregressive generation. Parameter optimization was performed across various temperatures (0.1-1.5, increments of 0.1), top-k values (1-4), and top-p values (0.1-1.0, increments of 0.1) using the Evo 1.5 model. For each parameter combination, one hundred 1,000-nt sequences were generated from a test set of 5 gene prompts encoding 50% of a highly conserved protein. Following identification of ORFs in generated sequences using Prodigal v2.6.3 with default parameters in metagenome mode (`-p meta`) (Hyatt et al., 2010), generated proteins were aligned against the full-length prompt protein sequence using MAFFT (v7.526) (Katoh et al., 2002) for sequence identity calculations. For evaluating sequence degeneracy, DustMasker (version 2.14.1+galaxy2) (Morgulis et al., 2006; Camacho et al., 2009; Cock et al., 2015) was run across the full-length generations using default parameters and the proportion of masked nucleotides was calculated. Final parameters (temperature = 0.7, top-k = 4, top-p= 1.0) were selected based on maximizing sequence completion accuracy while maintaining DustMasker proportions below 0.2, a value chosen to be just slightly higher than the typical frequency of non-coding DNA in prokaryotic genomes (Rogozin et al., 2002; Troyanskaya et al., 2002).

Sequence completion prompt compilation and evaluation

Sequences of highly-conserved genes from across prokaryotic biology were downloaded in FASTA format from NCBI GenBank (Clark et al., 2015). Selected genes included *rpoS* from *E. coli* K-12 (GenBank: NC_000913.3, coordinates 2867551-2868753), *gyrA* from *S. typhimurium* LT2 (GenBank: NC_003197.2, coordinates 2337442-2339871), and *ftsZ1* from *H. volcanii* DS2 (GenBank: NC_013967.1, coordinates 1675932-1677149). Prompts were prepared by extracting 30%, 50%, and 80% sequence lengths from the 5' end.

Sequence completion performance was evaluated across varying prompt lengths (30%, 50%, and 80% of input sequence) using optimal sampling parameters for the Evo 1.5, Evo 1 8k (previous version of Evo trained with context length of 8,192 tokens), and Evo 1 131k (previous version of Evo trained with context length of 131,072 tokens) models (temperature = 0.7, top-k = 4, top-p = 1) (Nguyen et al., 2024). For each prompt, one hundred sequences of length 2,500 nt were generated. Prompts were subsequently appended to the start of each generated sequence and ORFs identified using Prodigal v2.6.3 with default parameters in metagenome mode (`-p meta`). Generated proteins were then aligned against their corresponding wild-type sequences using MAFFT (v7.526) using default parameters for sequence identity calculations.

Operon completion prompt compilation and evaluation

Sequences encoding the *trp* operon and *modABC* operon from *E. coli* K-12 (GenBank: NC_000913.3) were downloaded in FASTA format from NCBI GenBank. For *modABC*, prompts were prepared from the full coding sequences for *modA* (coordinates 795089-795862), *modB* (coordinates 795862-796551), *modC* (coordinates 796554-797612) and *acrZ* (coordinates 794773-794922) (Clark et al., 2015). For *trp*, prompts were prepared from the full coding sequences for *trpE* (coordinates c1322946-1321384), *trpD* (coordinates c1321384-

1319789), *trpC* (coordinates c1319788-1318427), *trpB* (coordinates c1318415-1317222), and *trpA* (coordinates c1317222-1316416). For bidirectional generation testing, the reverse complement of each prompt sequence was generated using Biopython (Cock et al., 2009).

Sequence completion performance was evaluated across the compiled operon completion prompts using previously identified optimal sampling parameters for Evo 1.5. For each prompt, 1000 sequences of length 2500 nt were generated. Following identification of ORFs using Prodigal, directional completion was assessed by searching *trpE*-prompted generations for *trpD*-like ORFs, *trpD* reverse complement-prompted generations for *trpE*-like ORFs, and similar pairing combinations across both *modABC* and *trp* operons. Protein sequences were then aligned against their corresponding wild-type proteins for sequence identity calculations using MAFFT (v7.526). Structural similarity was evaluated by generating protein structure predictions using AlphaFold 3 (Abramson et al., 2024) for both generated and wild-type sequences, with structural alignments and TM-score calculations performed using TM-align (Zhang and Skolnick, 2005). Native and predicted protein structures were subsequently visualized using ChimeraX (Pettersen et al., 2021).

Positional entropy evaluation

Per-position amino acid and nucleotide entropies were calculated from multiple sequence alignments of 500 generated and native *modB* sequences. Native *modB* sequences were fetched by querying 'modB' in NCBI protein, filtering by bacteria, and downloading the corresponding amino acid and nucleotide sequences in format 'FASTA' and 'FASTA CDS' respectively. Generated *modB* sequences were taken by selecting a random sample of 500 modB ORFs from the *modB* sequences generated by prompting with *modA* during the operon completion evaluation. First, nucleotide and amino acid sequences were aligned with MAFFT (v7.526) and trimmed to remove gaps appearing in >80% of sequences. For each position i, the Shannon entropy was then calculated as $H(x_i) = -\sum_{x_i \in \mathcal{X}} p(x_i) \log_2(p(x_i))$ and normalized by dividing the calculated entropy by the maximum Shannon entropy (2 for nucleotides meaning all 4 bases are equally present, 4.32 for amino acids meaning all 20 standard amino acids are equally present), where $p(x_i)$ represents the frequency of each amino acid or nucleotide x_i at that position and $\mathcal{X}$ indicates the vocabulary.

Toxin-antitoxin prompt compilation

Genomic loci and sequences encoding Type II toxin-antitoxin system sequences were obtained by downloading the nucleotide sequence information for all experimentally validated type II-TAs from the TADB 3.0 database (Guan et al., 2024). Using NCBI's Entrez Programming Utilities API (EFetch from the 'nuccore' database using the genomic loci from TADB 3.0) (Sayers et al., 2022), the 500 bp of upstream and downstream flanking sequence were extracted for each T2TA locus. In total, for each T2TA system, six types of prompts were prepared: (1) individual toxin sequences, (2) individual antitoxin sequences, (3) the upstream context of the toxin loci, (4) the downstream context of the toxin loci, (5) the upstream context of the antitoxin loci, and (6) the downstream context of the antitoxin loci. Following successful identification of an Evo generated toxin (see section "**Evaluation of toxin activity**" below), conjugate antitoxins were subsequently generated via prompting with the generated DNA sequence encoding the toxin.

Toxin-antitoxin sampling and filtering

To generate T2TA candidates, we first generated 57,936 sequences of 2,000 nucleotides each using Evo 1.5 (temperature = 0.7, top-k = 4, top-p = 1.0) with our compiled T2TA prompts. A multi-stage filtering pipeline was then applied to identify promising candidates. First, Prodigal was used to identify ORFs, excluding sequences containing proteins over 300 amino acids or those with only a single ORF. Next, SegMasker (version 2.14.1+galaxy2) with default parameters (Camacho et al., 2009; Cock et al., 2015) was employed to remove sequences containing low complexity regions with limited amino acid diversity.

We then assessed the protein-protein interaction potential of co-generated ORFs by co-folding all ORF pairs within each remaining generation using ESMFold. Generations were retained if they contained paired proteins with pDockQ (Basu and Wallner, 2016; Bryant et al., 2022) scores above 0.23 and high sequence dissimilarity between components. To identify novel candidates, the remaining sequences were searched against the non-redundant protein sequence database using BLAST (E-value cutoff 5×10^{-2}), selecting for generations containing at least one component with no significant BLAST hits to known toxins or antitoxins. Final candi-

dates were then selected based on high confidence interaction prediction using AlphaFold 2 (Jumper et al., 2021).

Following the identification of functional toxin candidates via experimental testing, the Evo-generated sequence encoding the strongest Evo-generated toxin, EvoRelE1, was used as a prompt to generate further diversified antitoxin candidates. After generating a total of 3,000 sequences from the EvoRelE1 prompt, generated ORFs were cleaned and filtered as above before being co-folded with EvoRelE1. As with the first round of generations, candidates were filtered for high pDockQ scores and low-sequence homology to known antitoxins before being evaluated for strong predicted co-folds using AlphaFold2. Final antitoxin candidates were further characterized using Foldseek web server (van Kempen et al., 2024) searches of the AlphaFold3 predicted structures (probability threshold of 0.6), blastp searches against the non-redundant protein database (E value threshold of 1), and HHpred searches (probability threshold of >90%) (Söding et al., 2005).

Cloning of toxin and antitoxin sequences

Sequences encoding Evo-generated toxins and antitoxins were codon optimized for *E. coli* expression using IDT's codon optimization tool and synthesized as eBlocks. The toxin plasmid backbone was prepared by PCR amplification (New England Biosciences) and gel purification (Qiagen) of pAraSpCas9+spMu (Forsberg et al., 2019) to exclude the Cas9, crRNA, tracrRNA, and gRNA sequences (**Data S1**), creating an empty arabinose-inducible vector with spectinomycin resistance. Codon-optimized toxin sequences were subsequently inserted into the modified backbone using Gibson Assembly (New England Biosciences) according to the manufacturer's recommendations.

For antitoxin cloning, the pZE21_tetR-AcrIIA4_kanR vector (Forsberg et al., 2019) (**Data S1**) was digested with EcoRI and HindIII (New England Biosciences) and gel purified (Qiagen) to remove the AcrIIA4 sequence. Codon-optimized antitoxin sequences were then inserted into the digested vector using Gibson Assembly (New England Biosciences).

Assembled plasmids were transformed into chemically competent Stellar *E. coli* HST08 cells (Takara Biosciences), and positive clones were selected on LB agar plates containing either 50 g/mL spectinomycin (for the toxin constructs) or 50 g/mL kanamycin (for the antitoxin constructs). Plasmid sequences were then confirmed by Sanger sequencing (Elim Biosciences).

Evaluation of toxin activity

Toxin activity was assessed through growth inhibition assays in *E. coli* NEB Turbo cells. Cells transformed with toxin constructs were first grown overnight in LB medium containing 50 g/mL spectinomycin. Cultures were then diluted 1:10 into 1 mL fresh LB supplemented with 50 g/mL spectinomycin and 2 mg/mL arabinose to induce toxin expression. After 2 hours of induction, cultures were further diluted 1:40 into 200 L of the same medium in triplicate wells of a 96-well plate. Growth was monitored using a Tecan Spark plate reader at 37°C with orbital shaking, measuring optical density (OD_{600}) at 30-minute intervals over an 8-hour period. Growth curves were analyzed to evaluate the extent of toxin-mediated growth inhibition.

Evaluation of antitoxin activity

For antitoxin evaluation, *E. coli* NEB Turbo cells were co-transformed with both toxin and antitoxin constructs and selected on LB plates containing 50 g/mL spectinomycin and 50 g/mL kanamycin. Co-transformed cells were grown overnight in LB medium containing both antibiotics. Cultures were then diluted 1:10 into 1 mL fresh LB supplemented with both antibiotics and 2 mg/mL arabinose to induce toxin expression. After 2 hours of induction, cultures were further diluted 1:40 into 200 L of the same medium in triplicate wells of a 96-well plate. Growth was monitored using a Tecan Spark plate reader at 37°C with orbital shaking, measuring optical density (OD_{600}) at 30-minute intervals over an 8-hour period. Growth curves were analyzed to evaluate antitoxin-mediated rescue of growth.

Statistical comparisons were done using a one-sided Student's *t*-test against the EvoRelE1+ eGFP negative control, with *p*-values for EvoAT1-4 being 1.61×10^{-7}, 1.42×10^{-7}, 2.30×10^{-5}, and 1.69×10^{-5} respectively.

Anti-CRISPR prompt compilation

Genomic sequences containing known Type II Cas9-targeting anti-CRISPR (*acr*) genes and their associated operons were obtained from previously characterized anti-CRISPR systems annotated in AcrDB. Using the Entrez API, we extracted Acr coding sequences along with 500 bp of flanking sequence both upstream and downstream of each Acr locus. For each Acr system, we prepared six types of prompts: (1) individual *acr* sequences, (2) anti-CRISPR associated (*aca*) gene sequences, (3) the upstream context of the Acr loci, (4) the downstream context of the Acr loci, (5) the upstream context of the *aca* loci, and (6) the downstream context of the *aca* loci.

Anti-CRISPR sampling and filtering

To generate Acr candidates, we sampled a total of 3,160 sequences of 1,000 nucleotides each using Evo 1.5 with our compiled Acr prompts. A multi-stage filtering pipeline was then applied to identify promising candidates. First, Prodigal was used to identify ORFs, excluding sequences containing proteins over 200 amino acids. Next, SegMasker was employed to remove sequences containing over 50% low complexity regions or limited amino acid diversity. Sequences were subsequently folded using ESMFold to remove any candidates with very low confidence folds (pLDDT < 0.25) or no secondary structure. Candidate sequences were then evaluated using PaCRISPR, a machine learning model trained to identify potential anti-CRISPR proteins based on sequence features (Wang et al., 2020). Following guidelines established by PaCRISPR, candidates scoring above 0.5 were considered potential anti-CRISPRs, with sequences scoring above 0.75 advancing to the next step. Generated sequences classified as potential Acrs by PaCRISPR were then searched against the non-redundant protein sequence database using blastp (E-value cutoff of 1) to identify candidates with varying degrees of sequence novelty. For comparison, we also applied this filtration pipeline to evaluate sequences generated by prompting with randomly generated DNA sequences and non-Acr related genomic sequences. We then assessed the sequence diversity among the predicted Acrs by performing pairwise alignments using MAFFT (v7.526) on a set of 56 randomly selected sequences that scored above 0.75 in PaCRISPR. The resulting pairwise sequence identities were visualized using a matplotlib heatmap.

Cloning of anti-CRISPR sequences

Sequences encoding Evo-generated anti-CRISPR proteins were codon optimized for *E. coli* expression using IDT's codon optimization tool and synthesized as eBlocks. To generate the cloning backbone, the pZE21_tetR-AcrIIA4-Coli_kanR vector (Forsberg et al., 2019) (**Data S1**) was digested with EcoRI and HindIII (New England Biosciences) to remove the AcrIIA4 sequence and gel purified (Qiagen). Codon-optimized Acr sequences were then inserted into the digested vector using Gibson Assembly (New England Biosciences) according to the manufacturer's recommendations. Assembled plasmids were then transformed into chemically competent Stellar *E. coli* HST08 cells (Takara Biosciences), and positive clones were selected on LB agar plates containing 50 g/mL kanamycin. Plasmid sequences were confirmed by Sanger sequencing (Elim Biosciences).

Liquid culture assay for measuring anti-CRISPR activity

Anti-CRISPR activity was assessed through protection assays against SpCas9-mediated DNA cleavage in *E. coli*. NEB Turbo cells were first co-transformed with both the Acr expression plasmid and the CRISPR targeting plasmid containing SpCas9 and a KanR targeting guide RNA (Forsberg et al., 2019) (pAraCas9 + Sp2 + Sp6 + I-SceI, **Data S1**) and grown overnight in LB medium containing both 50 g/mL kanamycin and spectinomycin.

For liquid culture survival assays, 30 L of overnight culture was diluted into 1000 L fresh LB medium containing spectinomycin and 0.2 mg/mL arabinose to induce SpCas9 expression and deplete the kanamycin resistance plasmid. After 7 hours of growth, cultures were normalized to equal optical density and diluted 1:40 into 200 L fresh LB medium containing 50 g/mL kanamycin and spectinomycin in a 96-well plate to select for cells with active anti-CRISPR proteins. Growth was monitored using a Tecan Spark plate reader at 37°C with orbital shaking, measuring optical density (OD_{600}) at 30-minute intervals over a 7-hour period. For each Acr, an uninduced control without arabinose was used to normalize growth values.

Statistical comparisons were done using a one-sided Student's t-test against the random sequence negative control and AcrIIA2, with p-values for EvoAcr1-5 being 8.31×10^{-6}, 6.19×10^{-6}, 8.19×10^{-6}, 1.43×10^{-6},

and 8.18×10^{-7} against the negative control respectively and the p-value for EvoAcr5 against AcrIIA2 being 5.07×10^{-4}.

Phage plaque assay for measuring anti-CRISPR activity

For phage protection assays, *E. coli* NEB Turbo cells were co-transformed with both the Acr expression plasmid and a modified SpCas9 plasmid containing a guide RNA targeting T4(GT7) phage (SpCas9-mrh2, **Data S1**) at a 5:1 Acr:Cas9 ratio. Co-transformed cells were grown overnight in LB medium containing 50 g/mL kanamycin and spectinomycin. Cultures were then diluted 1:10 into fresh LB medium supplemented with spectinomycin, kanamycin, and 0.2 mg/mL arabinose to induce Cas9 expression. When cultures reached an OD_{600} of 0.4, 300 L was mixed into 10 mL of 0.7% soft agar containing 50 mg/mL kanamycin, 50 mg/mL spectinomycin, and 0.02 mg/mL arabinose. Plates were allowed to harden and T4(GT7) phage at a titer of 1.7×10^8 PFU/mL was serially diluted 1:5 and spotted onto the bacterial lawn. Plates were then incubated overnight at 37°C before being imaged to visualize plaque formation.

Measurement of Evo-generated protein sequence diversity

Sequence and structural diversity of generated toxin-antitoxin pairs and anti-CRISPRs were assessed through a combination of sequence and structure-based searches. Protein sequences were searched against the NCBI non-redundant protein database using blastp (E-value threshold of 1.0) to identify potential homologs. An additional search against OpenGenome was performed using MMseqs2 (version 15.6f452) with maximum sensitivity (-s 7) and other parameters set to the default values (Steinegger and Söding, 2017) to evaluate similarity to sequences in the training data. Structural similarity was evaluated by searching AlphaFold 3-predicted structures against the BFVD, AFDB-Proteome, AFDB-Swissprot, AFDB-50, BFMD, CATH50, GMGCL_ID, MGNIFY_ESM30, and PDB100 protein databases using the Foldseek web server (`https://search.foldseek.com/search`) (van Kempen et al., 2024). For each generated protein, closest sequence and structural matches were identified from all three searches (BLAST, MMseqs2, and Foldseek) and evaluated to determine the degree of novelty compared to known proteins. Pairwise sequence identities were calculated from BLAST alignments, while structural similarities were quantified using Foldseek TM-scores.

SynGenome prompt compilation

Protein sequences from prokaryotic and phage organisms were retrieved from UniProt, with all Swissprot proteins with associated coding regions and a random sample of TrEMBL proteins with associated coding regions being used as starting points for prompts (Consortium, 2024). Genomic contexts were retrieved via NCBI's Entrez Programming Utilities (E-utilities) API, specifically using EFetch with the 'nuccore' database using the CDS annotations associated with the Protein Sequence accession in UniProt. For each protein's associated genomic identifier, we constructed three API calls: one to extract the coding sequence using sequence feature coordinates, and two to extract the flanking regions by calculating positions 500 nucleotides upstream and downstream of the CDS boundaries. For CDS regions that were longer than 500 nucleotides, the prompt was derived from the final 500 nucleotides in the coding sequence. For CDS regions that were shorter than 500 nucleotides, the CDS sequence was trimmed to the nearest 100 bp. Additionally, we generated reverse complement sequences for each region using the Biopython Seq module, resulting in six distinct prompts per protein: upstream, coding sequence (CDS), downstream, and their respective reverse complements. Associated functional annotations including Gene Ontology (GO) terms (Ashburner et al., 2000; The Gene Ontology Consortium et al., 2023), species names, and InterPro domains (Blum et al., 2024) were retrieved from UniProt's REST API for each protein using their UniProt accession numbers and linked to their corresponding prompts.

SynGenome sampling

Sequences were generated using Evo 1.5 with optimized sampling parameters (temperature = 0.7, top-k = 4, top-p = 1.0). For each prompt type, we generated 2 sequences of 5,000 nucleotides in length, yielding a total database size of over 120 billion base pairs. Generation was performed in parallel across multiple compute nodes to facilitate large-scale sequence production.

Data filtering of SynGenome sequences

Generated sequences underwent a multi-step filtering pipeline to remove low-complexity regions while preserving biologically plausible features. Initial validation removed any invalid characters from the nucleotide sequences. These sequences were then processed using DustMasker (NCBI BLAST+ v2.16.0) with a masking level of 30 (-level 30) and FASTA output format (-outfmt fasta) to identify low-complexity regions. Following DustMasker processing, we implemented two additional filtering steps: removal of successive 100 nucleotide chunks from the sequence end if they contained more than 40% masked bases, and elimination of any continuous masked regions longer than 800 nucleotides. Sequences were excluded from the final database based on several criteria: length below 100 nucleotides, masked base content exceeding 80% in sequences shorter than 2,000 nucleotides, complete masking of all bases, or empty/NaN values. The sequences passing these filtering criteria were converted to uppercase and retained in the final database.

Prediction of SynGenome ORFs

Open reading frames were identified in the filtered sequences using Prodigal v2.6.3 with default parameters and metagenomic prediction mode (`-p meta`). Predictions were initially refined by excluding sequences shorter than 40 amino acids or longer than 1200 amino acids and sequences with incomplete protein sequences. Following basic filtration, low-complexity regions were identified using NCBI SegMasker (window size 15, locut 1.8, hicut 3.4), with sequences containing >20% masked regions being excluded. Additional complexity filters removed sequences with fewer than 12 unique amino acids or highly repetitive k-mer patterns (k=3-10, threshold >40% coverage). This multi-step filtering process ensured the retention of high-quality protein predictions while removing potentially spurious or low-complexity sequences.

Creation of SynGenome website

The SynGenome website was implemented with HTML, CSS, and JavaScript to provide a searchable web interface at https://evodesign.org/syngenome/. The interface allows users to query sequences using protein names, UniProt IDs, InterPro domains, species names, or GO terms. The database was structured to maintain associations between generated sequences and their corresponding prompt annotations. The raw SynGenome data is hosted on Hugging Face for public access.

Creation of SynGenome UMAP and Leiden Clusters

To generate the UMAP, first, a random sample of 50,000 sequences encoding at least one ORF with prompts derived from the CDS sequence was extracted from SynGenome. This random sample was subsequently filtered to remove generations with < 40% or > 60% GC content, resulting in a final set of 36,762 generations. Embeddings were generated for both the prompt and corresponding generated sequences by extracting activations from the MLP layer 3 in the 6th hyena block of Evo 1.5's architecture before being mean pooled along the sequence dimension to create fixed-length representations. These high-dimensional embeddings were subsequently reduced to two dimensions using UMAP (n_neighbors=15, min_dist=0.1). Sequences were colored according to their percent identity to sequences in the training data, calculated using MMseqs2 against a database of prokaryotic genomes used in model training. These high-dimensional embeddings were reduced to two dimensions using UMAP with ScanPy (version 1.10.3) (Wolf et al., 2018) default parameters. Sequences were colored according to their percent identity to sequences in the training data, calculated using MMseqs2 against OpenGenome. Graph-based clustering was performed using the default Leiden algorithm implemented in Scanpy (version 1.10.3). The resolution parameter was optimized by evaluating cluster stability across resolutions from 0.1 to 0.5, measuring both the number of clusters and coefficient of variation of cluster sizes. A final resolution of 0.2 was selected for clustering based on these metrics.

Comparison of SynGenome to native prokaryotic sequences

To evaluate how representative SynGenome was of natural prokaryotic sequences, first, a random sample of 36,762 5,000-nt generations was taken from OpenGenome to match the number of sequences used in the representative sample of SynGenome. Following the procedure used for prediction of ORFs in SynGenome (see section "Prediction of SynGenome ORFs"), Prodigal was used to identify potential ORFs in the sampled se-

quences and minimal filtering was applied to remove clearly incorrectly called sequences. Length distributions of predicted SynGenome ORFs were compared to those from the random OpenGenome sample. Codon usage bias of the prompts and SynGenome generations were analyzed by calculating the total frequency of each codon across all prompt ORFs and all generated ORFs and normalizing the frequencies to the total number of codons per dataset. Protein family domains were identified in both datasets using HMMER v3.3.2 hmmscan (hmmer.org) against the Pfam-A database (Paysan-Lafosse et al., 2024) with default parameters (E-value cutoff 0.01). Domain frequencies were compared between datasets using Fisher's exact test with Benjamini-Hochberg multiple testing correction (minimum occurrence threshold of 10 domains in both datasets). The relationship between domain frequencies was visualized using a log-scale scatter plot, with points colored by absolute $\log_2$ fold change between datasets. Correlation between domain frequencies was assessed using the Pearson correlation coefficient calculated using the pearsonr function in SciPy (Virtanen et al., 2020). The distribution of domain occurrence frequencies was analyzed by binning domains found in the scanned SynGenome and OpenGenome proteins into log-scale bins (2^n occurrences) and visualizing the count of unique domains per frequency bin for each.

References

J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. 630(8016):493–500, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://www.nature.com/articles/s41586-024-07487-w. Publisher: Nature Publishing Group.

H. Altae-Tran, S. Kannan, A. J. Suberski, K. S. Mears, F. E. Demircioglu, L. Moeller, S. Kocalar, R. Oshiro, K. S. Makarova, R. K. Macrae, E. V. Koonin, and F. Zhang. Uncovering the functional diversity of rare CRISPR-cas systems with deep terascale clustering. 382(6673):eadi1910, 2023. doi: 10.1126/science.adi1910. URL https://www.science.org/doi/10.1126/science.adi1910. Publisher: American Association for the Advancement of Science.

L. Aravind. Guilt by association: Contextual information in genome analysis. 10(8):1074–1077, 2000. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.10.8.1074. URL https://genome.cshlp.org/content/10/8/1074. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1546-1718. doi: 10.1038/75556. URL https://www.nature.com/articles/ng0500_25. Publisher: Nature Publishing Group.

S. Basu and B. Wallner. DockQ: A quality measure for protein-protein docking models. 11(8):e0161879, 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0161879. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161879. Publisher: Public Library of Science.

A. Beavogui, A. Lacroix, N. Wiart, J. Poulain, T. O. Delmont, L. Paoli, P. Wincker, and P. H. Oliveira. The defensome of complex bacterial communities. 15(1):2146, 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-46489-0. URL https://www.nature.com/articles/s41467-024-46489-0. Publisher: Nature Publishing Group.

K. Blin, S. Shaw, A. M. Kloosterman, Z. Charlop-Powers, G. P. van Wezel, M. Medema, and T. Weber. antiSMASH 6.0: improving cluster detection and comparison capabilities. 49:W29–W35, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab335. URL https://doi.org/10.1093/nar/gkab335.

M. Blum, A. Andreeva, L. Florentino, S. Chuguransky, T. Grego, E. Hobbs, B. Pinto, A. Orr, T. Paysan-Lafosse, I. Ponamareva, G. Salazar, N. Bordin, P. Bork, A. Bridge, L. Colwell, J. Gough, D. Haft, I. Letunic, F. Llinares-López, A. Marchler-Bauer, L. Meng-Papaxanthos, H. Mi, D. Natale, C. Orengo, A. Pandurangan, D. Piovesan, C. Rivoire, C. A. Sigrist, N. Thanki, F. Thibaud-Nissen, P. Thomas, S. E. Tosatto, C. Wu, and A. Bateman. InterPro: the protein sequence classification resource in 2025. page gkae1082, 2024. ISSN 0305-1048. doi: 10.1093/nar/gkae1082. URL https://doi.org/10.1093/nar/gkae1082.

E. S. Boyden, F. Zhang, E. Bamberg, G. Nagel, and K. Deisseroth. Millisecond-timescale, genetically targeted optical control of neural activity. 8(9):1263–1268, 2005. ISSN 1546-1726. doi: 10.1038/nn1525. URL https://www.nature.com/articles/nn1525. Publisher: Nature Publishing Group.

N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial. ProteinBERT: a universal deep-learning model of protein sequence and function. 38(8):2102–2110, 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac020. URL https://doi.org/10.1093/bioinformatics/btac020.

T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen,

B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah. Towards monosemanticity: Decomposing language models with dictionary learning. 2023.

P. Bryant, G. Pozzati, and A. Elofsson. Improved prediction of protein-protein interactions using AlphaFold2. 13(1):1265, 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28865-w. URL https://www.nature.com/articles/s41467-022-28865-w. Publisher: Nature Publishing Group.

C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. BLAST+: architecture and applications. 10(1):421, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-421. URL https://doi.org/10.1186/1471-2105-10-421.

S. Camara-Wilpert, D. Mayo-Muñoz, J. Russel, R. D. Fagerlund, J. S. Madsen, P. C. Fineran, S. J. Sørensen, and R. Pinilla-Redondo. Bacteriophages suppress CRISPR–cas immunity using RNA-based anti-CRISPRs. 623(7987):601–607, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06612-5. URL https://www.nature.com/articles/s41586-023-06612-5. Publisher: Nature Publishing Group.

W. T. Chan, M. P. Garcillán-Barcia, C. C. Yeo, and M. Espinosa. Type II bacterial toxin–antitoxins: hypotheses, facts, and the newfound plethora of the PezAT system. 47(5):fuad052, 2023. ISSN 0168-6445. doi: 10.1093/femsre/fuad052. URL https://doi.org/10.1093/femsre/fuad052.

Z. Cheng, B.-B. He, K. Lei, Y. Gao, Y. Shi, Z. Zhong, H. Liu, R. Liu, H. Zhang, S. Wu, W. Zhang, X. Tang, and Y.-X. Li. Rule-based omics mining reveals antimicrobial macrocyclic peptides against drug-resistant clinical isolates. 15(1):4901, 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-49215-y. URL https://www.nature.com/articles/s41467-024-49215-y. Publisher: Nature Publishing Group.

S. M. Chiang and H. E. Schellhorn. Evolution of the RpoS regulon: Origin of RpoS and the conservation of RpoS-dependent regulation in bacteria. 70(6):557–571, 2010. ISSN 1432-1432. doi: 10.1007/s00239-010-9352-0. URL https://doi.org/10.1007/s00239-010-9352-0.

A. Chien, D. B. Edgar, and J. M. Trela. Deoxyribonucleic acid polymerase from the extreme thermophile thermus aquaticus. 127(3):1550–1557, 1976. doi: 10.1128/jb.127.3.1550-1557.1976. URL https://journals.asm.org/doi/abs/10.1128/jb.127.3.1550-1557.1976. Publisher: American Society for Microbiology.

N. Choudhary, D. Tandi, R. K. Verma, V. K. Yadav, N. Dhingra, T. Ghosh, M. Choudhary, R. K. Gaur, M. H. Abdellatif, A. Gacem, L. B. Eltayeb, M. S. Alqahtani, K. K. Yadav, and B.-H. Jeon. A comprehensive appraisal of mechanism of anti-CRISPR proteins: an advanced genome editor to amend the CRISPR gene editing. 14:1164461, 2023. ISSN 1664-462X. doi: 10.3389/fpls.2023.1164461. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10328345/.

K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Research*, 44 (D1):D67–D72, 11 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv1276. URL https://doi.org/10.1093/nar/gkv1276.

P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. 25(11):1422–1423, 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp163. URL https://doi.org/10.1093/bioinformatics/btp163.

P. J. A. Cock, J. M. Chilton, B. Grüning, J. E. Johnson, and N. Soranzo. NCBI BLAST+ integrated into galaxy. 4(1):s13742–015–0080–7, 2015. ISSN 2047-217X. doi: 10.1186/s13742-015-0080-7. URL https://doi.org/10.1186/s13742-015-0080-7.

T. U. Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, page gkae1010, 11 2024. ISSN 0305-1048. doi: 10.1093/nar/gkae1010. URL https://doi.org/10.1093/nar/gkae1010.

K. Dai and J. Lutkenhaus. ftsZ is an essential cell division gene in escherichia coli. 173(11):3500–3506, 1991. doi: 10.1128/jb.173.11.3500-3506.1991. URL https://journals.asm.org/doi/10.1128/jb.173.11.3500-3506.1991. Publisher: American Society for Microbiology.

A. R. Davidson, W.-T. Lu, S. Y. Stanley, J. Wang, M. Mejdani, C. N. Trost, B. T. Hicks, J. Lee, and E. J. Sontheimer. Anti-CRISPRs: Protein inhibitors of CRISPR-cas systems. 89:309–332, 2020. ISSN 0066-4154, 1545-4509. doi: 10.1146/annurev-biochem-011420-111224. URL https://www.annualreviews.org/content/journals/10.1146/annurev-biochem-011420-111224. Publisher: Annual Reviews.

C. Dong, X. Wang, C. Ma, Z. Zeng, D.-K. Pu, S. Liu, C.-S. Wu, S. Chen, Z. Deng, and F.-B. Guo. Anti-CRISPRdb v2.2: an online repository of anti-CRISPR proteins including information on inhibitory mechanisms, activities and neighbors of curated anti-CRISPR proteins. 2022:baac010, 2022. ISSN 1758-0463. doi: 10.1093/database/baac010. URL https://doi.org/10.1093/database/baac010.

S. Doron, S. Melamed, G. Ofir, A. Leavitt, A. Lopatina, M. Keren, G. Amitai, and R. Sorek. Systematic discovery of antiphage defense systems in the microbial pangenome. 359(6379):eaar4120, 2018. doi: 10.1126/science.aar4120. URL https://www.science.org/doi/10.1126/science.aar4120. Publisher: American Association for the Advancement of Science.

M. G. Durrant, N. T. Perry, J. J. Pai, A. R. Jangid, J. S. Athukoralage, M. Hiraizumi, J. P. McSpedon, A. Pawluk, H. Nishimasu, S. Konermann, and P. D. Hsu. Bridge RNAs direct programmable recombination of target and donor DNA. 630(8018):984–993, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07552-4. URL https://www.nature.com/articles/s41586-024-07552-4. Publisher: Nature Publishing Group.

S. Eitzinger, A. Asif, K. E. Watters, A. T. Iavarone, G. J. Knott, J. A. Doudna, and F. Minhas. Machine learning predicts new anti-CRISPR proteins. 48(9):4698–4708, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa219. URL https://doi.org/10.1093/nar/gkaa219.

R. Esfandiarpoor and S. H. Bach. Follow-up differential descriptions: Language models resolve ambiguities for image classification, 2024. URL http://arxiv.org/abs/2311.07593.

J. R. Firth. Applications of general linguistics. 56(1):1–14, 1957. ISSN 1467-968X. doi: 10.1111/j.1467-968X.1957.tb00568.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-968X.1957.tb00568.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-968X.1957.tb00568.x.

K. J. Forsberg, I. V. Bhatt, D. T. Schmidtke, K. Javanmardi, K. E. Dillard, B. L. Stoddard, I. J. Finkelstein, B. K. Kaiser, and H. S. Malik. Functional metagenomics-guided discovery of potent cas9 inhibitors in the human microbiome. 8:e46540, 2019. ISSN 2050-084X. doi: 10.7554/eLife.46540. URL https://doi.org/10.7554/eLife.46540. Publisher: eLife Sciences Publications, Ltd.

N. Fraikin, F. Goormaghtigh, and L. Van Melderen. Type II toxin-antitoxin systems: Evolution and revolutions. 202(7):10.1128/jb.00763–19, 2020. doi: 10.1128/jb.00763-19. URL https://journals.asm.org/doi/10.1128/jb.00763-19. Publisher: American Society for Microbiology.

F. Gabler, S.-Z. Nam, S. Till, M. Mirdita, M. Steinegger, J. Söding, A. N. Lupas, and V. Alva. Protein sequence analysis using the MPI bioinformatics toolkit. 72(1):e108, 2020. ISSN 1934-340X. doi: 10.1002/cpbi.108. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.108. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpbi.108.

M. Y. Galperin and E. V. Koonin. Who's your neighbor? new computational approaches for functional genomics. 18(6):609–613, 2000. ISSN 1546-1696. doi: 10.1038/76443. URL https://www.nature.com/articles/nbt0600_609. Publisher: Nature Publishing Group.

L. Gao, H. Altae-Tran, F. Böhning, K. S. Makarova, M. Segel, J. L. Schmid-Burgk, J. Koob, Y. I. Wolf, E. V. Koonin, and F. Zhang. Diverse enzymatic activities mediate antiviral immunity in prokaryotes. 369(6507):1077–1084, 2020. doi: 10.1126/science.aba0372. URL https://www.science.org/doi/10.1126/science.aba0372. Publisher: American Association for the Advancement of Science.

E. G. Govorunova, O. A. Sineshchekov, E. M. Rodarte, R. Janz, O. Morelle, M. Melkonian, G. K.-S. Wong, and J. L. Spudich. The expanding family of natural anion channelrhodopsins reveals large variations in kinetics, conductance, and spectral sensitivity. 7(1):43358, 2017. ISSN 2045-2322. doi: 10.1038/srep43358. URL https://www.nature.com/articles/srep43358. Publisher: Nature Publishing Group.

J. Guan, Y. Chen, Y.-X. Goh, M. Wang, C. Tai, Z. Deng, J. Song, and H.-Y. Ou. TADB 3.0: an updated database of bacterial toxin–antitoxin loci and associated mobile genetic elements. 52:D784–D790, 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad962. URL https://doi.org/10.1093/nar/gkad962.

Z. S. Harris. Distributional structure. 10(2):146–162, 1954. ISSN 0043-7956. doi: 10.1080/00437956.1954.11659520. URL https://www.tandfonline.com/doi/abs/10.1080/00437956.1954.11659520. Publisher: Routledge.

T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, R. Badkundri, I. Shafkat, J. Gong, A. Derry, R. S. Molina, N. Thomas, Y. Khan, C. Mishra, C. Kim, L. J. Bartie, M. Nemeth, P. D. Hsu, T. Sercu, S. Candido, and A. Rives. Simulating 500 million years of evolution with a language model, 2024. URL https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1. Pages: 2024.07.01.600583 Section: New Results.

B. L. Hie, V. R. Shanker, D. Xu, T. U. J. Bruun, P. A. Weidenbacher, S. Tang, W. Wu, J. E. Pak, and P. S. Kim. Efficient evolution of human antibodies from general protein language models. 42(2):275–283, 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01763-2. URL https://www.nature.com/articles/s41587-023-01763-2. Publisher: Nature Publishing Group.

L. Huang, B. Yang, H. Yi, A. Asif, J. Wang, T. Lithgow, H. Zhang, F. Minhas, and Y. Yin. AcrDB: a database of anti-CRISPR operons in prokaryotes and viruses. 49:D622–D629, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa857. URL https://doi.org/10.1093/nar/gkaa857.

P.-S. Huang, S. E. Boyken, and D. Baker. The coming of age of de novo protein design. 537(7620):320–327, 2016. ISSN 1476-4687. doi: 10.1038/nature19946. URL https://www.nature.com/articles/nature19946. Publisher: Nature Publishing Group.

D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. 11(1):119, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-119. URL https://doi.org/10.1186/1471-2105-11-119.

F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. 3(3):318–356, 1961. ISSN 0022-2836. doi: 10.1016/S0022-2836(61)80072-7. URL https://www.sciencedirect.com/science/article/pii/S0022283661800727.

K. Jiang, Z. Yan, M. Di Bernardo, S. R. Sgrizzi, L. Villiger, A. Kayabolen, B. Kim, J. K. Carscadden, M. Hiraizumi, H. Nishimasu, J. S. Gootenberg, and O. O. Abudayyeh. Rapid in silico directed evolution by a protein language model with EVOLVEpro. 0(0):eadr6006, 2024. doi: 10.1126/science.adr6006. URL https://www.science.org/doi/10.1126/science.adr6006. Publisher: American Association for the Advancement of Science.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, Aug. 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Publisher: Nature Publishing Group.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020. URL http://arxiv.org/abs/2001.08361.

K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. 30(14):3059–3066, 2002. ISSN 0305-1048. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC135756/.

E. V. Koonin. Hunting for treasure chests in microbial defense islands. 70(5):761–762, 2018. ISSN 1097-2765. doi: 10.1016/j.molcel.2018.05.025. URL https://www.sciencedirect.com/science/article/pii/S109727651830399X.

T. Kortemme. De novo protein design—from new structures to programmable functions. 187(3):526–544, 2024. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2023.12.028. URL https://www.cell.com/cell/abstract/S0092-8674(23)01402-2. Publisher: Elsevier.

J. J. Kwon, J. Pan, G. Gonzalez, W. C. Hahn, and M. Zitnik. On knowing a gene: A distributional hypothesis of gene function. 15(6):488–496, 2024. ISSN 2405-4712, 2405-4720. doi: 10.1016/j.cels.2024.04.008. URL https://www.cell.com/cell-systems/abstract/S2405-4712(24)00123-6. Publisher: Elsevier.

I. Lee, B. Ambaru, P. Thakkar, E. M. Marcotte, and S. Y. Rhee. Rational association of genes with traits using a genome-scale gene network for arabidopsis thaliana. 28(2):149–156, 2010. ISSN 1546-1696. doi: 10.1038/nbt.1603. URL https://www.nature.com/articles/nbt.1603. Publisher: Nature Publishing Group.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574. Publisher: American Association for the Advancement of Science.

A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, and N. Naik. Large language models generate functional protein sequences across diverse families. 41(8):1099–1106, 2023. ISSN 1546-1696. doi: 10.1038/s41587-022-01618-2. URL https://www.nature.com/articles/s41587-022-01618-2. Publisher: Nature Publishing Group.

K. S. Makarova, Y. I. Wolf, S. Snir, and E. V. Koonin. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. 193(21):6039–6056, 2011. doi: 10.1128/jb.05535-11. URL https://journals.asm.org/doi/10.1128/jb.05535-11. Publisher: American Society for Microbiology.

S. Massaroli, M. Poli, D. Fu, H. Kumbong, R. Parnichkun, D. Romero, A. Timalsina, Q. McIntyre, B. Chen, A. Rudra, C. Zhang, C. Ré, S. Ermon, and Y. Bengio. Laughing hyena distillery: Extracting compact recurrences from convolutions. 36:17072–17116, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/371355cd42caaf83412c3fbef4688979-Abstract-Conference.html.

J. A. Maupin-Furlow, J. K. Rosentel, J. H. Lee, U. Deppenmeier, R. P. Gunsalus, and K. T. Shanmugam. Genetic analysis of the modABCD (molybdate transport) operon of escherichia coli. 177(17):4851–4856, 1995. doi: 10.1128/jb.177.17.4851-4856.1995. URL https://journals.asm.org/doi/10.1128/jb.177.17.4851-4856.1995. Publisher: American Society for Microbiology.

O. Mañas, P. Astolfi, M. Hall, C. Ross, J. Urbanek, A. Williams, A. Agrawal, A. Romero-Soriano, and M. Drozdzal. Improving text-to-image consistency via automatic prompt optimization, 2024. URL http://arxiv.org/abs/2403.17804.

M. H. Medema, K. Blin, P. Cimermancic, V. de Jager, P. Zakrzewski, M. A. Fischbach, T. Weber, E. Takano, and R. Breitling. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. 39:W339–W346, 2011. ISSN 0305-1048. doi: 10.1093/nar/gkr466. URL https://doi.org/10.1093/nar/gkr466.

E. Merino, R. A. Jensen, and C. Yanofsky. Evolution of bacterial trp operons and their regulation. 11(2):78–86, 2008. ISSN 1369-5274. doi: 10.1016/j.mib.2008.02.005. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2387123/.

J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, and A. Bateman. Pfam: The protein families database in 2021. 49:D412–D419, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa913. URL https://doi.org/10.1093/nar/gkaa913.

A. Morgulis, E. M. Gertz, A. A. Schäffer, and R. Agarwala. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. 13(5):1028–1040, 2006. ISSN 1066-5277. doi: 10.1089/cmb.2006.13.1028.

K. Murtazalieva, A. Mu, A. Petrovskaya, and R. D. Finn. The growing repertoire of phage anti-defence systems. 0(0), 2024. ISSN 0966-842X, 1878-4380. doi: 10.1016/j.tim.2024.05.005. URL https://www.cell.com/trends/microbiology/abstract/S0966-842X(24)00136-7. Publisher: Elsevier.

G. Nagel, T. Szellas, W. Huhn, S. Kateriya, N. Adeishvili, P. Berthold, D. Ollig, P. Hegemann, and E. Bamberg. Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. 100(24):13940–13945, 2003. ISSN 0027-8424. doi: 10.1073/pnas.1936192100. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC283525/.

E. Nguyen, M. Poli, M. G. Durrant, B. Kang, D. Katrekar, D. B. Li, L. J. Bartie, A. W. Thomas, S. H. King, G. Brixi, J. Sullivan, M. Y. Ng, A. Lewis, A. Lou, S. Ermon, S. A. Baccus, T. Hernandez-Boussard, C. Ré, P. D. Hsu, and B. L. Hie. Sequence modeling and design from molecular to genome scale with Evo. 386 (6723):eado9336, 2024. doi: 10.1126/science.ado9336. URL https://www.science.org/doi/10.1126/science.ado9336. Publisher: American Association for the Advancement of Science.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, 2022. URL http://arxiv.org/abs/2203.02155.

R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. 96(6):2896–2901, 1999. doi: 10.1073/pnas.96.6.2896. URL https://www.pnas.org/doi/10.1073/pnas.96.6.2896. Publisher: Proceedings of the National Academy of Sciences.

M. Pacesa, L. Nickel, J. Schmidt, E. Pyatova, C. Schellhaas, L. Kissling, A. Alcaraz-Serna, Y. Cho, K. H. Ghamary, L. Vinué, B. J. Yachnin, A. M. Wollacott, S. Buckley, S. Georgeon, C. A. Goverde, G. N. Hatzopoulos, P. Gönczy, Y. D. Muller, G. Schwank, S. Ovchinnikov, and B. E. Correia. BindCraft: one-shot design of functional protein binders, 2024. URL https://www.biorxiv.org/content/10.1101/2024.09.30.615802v1. Pages: 2024.09.30.615802 Section: New Results.

A. Pawluk, A. R. Davidson, and K. L. Maxwell. Anti-CRISPR: discovery, mechanism and function. 16(1):12–17, 2018. ISSN 1740-1534. doi: 10.1038/nrmicro.2017.120. URL https://www.nature.com/articles/nrmicro.2017.120. Publisher: Nature Publishing Group.

T. Paysan-Lafosse, A. Andreeva, M. Blum, S. R. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, M. L. Bileschi, F. Llinares-López, L. Meng-Papaxanthos, L. J. Colwell, N. V. Grishin, R. D. Schaeffer, D. Clementel, S. C. E. Tosatto, E. Sonhammer, V. Wood, and A. Bateman. The Pfam protein families database: embracing AI/ML. *Nucleic Acids Research*, page gkae997, Nov. 2024. ISSN 0305-1048. doi: 10.1093/nar/gkae997. URL https://doi.org/10.1093/nar/gkae997.

T. H. Pers, J. M. Karjalainen, Y. Chan, H.-J. Westra, A. R. Wood, J. Yang, J. C. Lui, S. Vedantam, S. Gustafsson, T. Esko, T. Frayling, E. K. Speliotes, M. Boehnke, S. Raychaudhuri, R. S. N. Fehrmann, J. N. Hirschhorn, and L. Franke. Biological interpretation of genome-wide association studies using predicted gene functions. 6(1): 5890, 2015. ISSN 2041-1723. doi: 10.1038/ncomms6890. URL https://www.nature.com/articles/ncomms6890. Publisher: Nature Publishing Group.

E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, and T. E. Ferrin. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. 30(1):70–82, 2021. ISSN 0961-8368. doi: 10.1002/pro.3943. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7737788/.

D. M. Picton, Y. A. Luyten, R. D. Morgan, A. Nelson, D. Smith, D. T. F. Dryden, J. C. D. Hinton, and T. Blower. The phage defence island of a multidrug resistant plasmid uses both BREX and type IV restriction for complementary protection from viruses. 49(19):11257–11273, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab906. URL https://doi.org/10.1093/nar/gkab906.

M. Poli, J. Wang, S. Massaroli, J. Quesnelle, R. Carlow, E. Nguyen, and A. W. Thomas. StripedHyena: Moving beyond transformers with hybrid signal processing models, 2023. URL https://github.com/togethercomputer/stripedhyena. original-date: 2023-11-21T15:56:04Z.

N. Praljak, H. Yeh, M. Moore, M. Socolich, R. Ranganathan, and A. L. Ferguson. Natural language prompts guide the design of novel functional protein sequences, 2024. URL https://www.biorxiv.org/content/10.1101/2024.11.11.622734v1. Pages: 2024.11.11.622734 Section: New Results.

J. Qiu, Y. Zhai, M. Wei, C. Zheng, and X. Jiao. Toxin–antitoxin systems: Classification, biological roles, and applications. 264:127159, 2022. ISSN 0944-5013. doi: 10.1016/j.micres.2022.127159. URL https://www.sciencedirect.com/science/article/pii/S0944501322001999.

R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL http://arxiv.org/abs/2305.18290.

K. Rai, Y. Wang, R. W. O'Connell, A. B. Patel, and C. J. Bashor. Using machine learning to enhance and accelerate synthetic biology. 31:100553, 2024. ISSN 2468-4511. doi: 10.1016/j.cobme.2024.100553. URL https://www.sciencedirect.com/science/article/pii/S2468451124000333.

I. B. Rogozin, K. S. Makarova, D. A. Natale, A. N. Spiridonov, R. L. Tatusov, Y. I. Wolf, J. Yin, and E. V. Koonin. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. 30(19):4264–4271, 2002. ISSN 0305-1048. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC140549/.

B. Rost. Twilight zone of protein sequence alignments. 12(2):85–94, 1999. ISSN 1741-0126. doi: 10.1093/protein/12.2.85. URL https://doi.org/10.1093/protein/12.2.85.

M. Sada, H. Kimura, N. Nagasawa, M. Akagawa, K. Okayama, T. Shirai, S. Sunagawa, R. Kimura, T. Saraya, H. Ishii, D. Kurai, T. Tsugawa, A. Nishina, H. Tomita, M. Okodo, S. Hirai, A. Ryo, T. Ishioka, and K. Murakami. Molecular evolution of the pseudomonas aeruginosa DNA gyrase gyrA gene. 10(8):1660, 2022. ISSN 2076-2607. doi: 10.3390/microorganisms10081660. URL https://www.mdpi.com/2076-2607/10/8/1660. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

E. W. Sayers, E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. C. Comeau, R. Connor, K. Funk, C. Kelly, S. Kim, T. Madej, A. Marchler-Bauer, C. Lanczycki, S. Lathrop, Z. Lu, F. Thibaud-Nissen, T. Murphy, L. Phan, Y. Skripchenko, T. Tse, J. Wang, R. Williams, B. W. Trawick, K. D. Pruitt, and S. T. Sherry. Database resources of the national center for biotechnology information. 50:D20–D26, 2022. ISSN 1362-4962. doi: 10.1093/nar/gkab1112.

J. Shin, F. Jiang, J.-J. Liu, N. L. Bray, B. J. Rauch, S. H. Baik, E. Nogales, J. Bondy-Denomy, J. E. Corn, and J. A. Doudna. Disabling cas9 by an anti-CRISPR DNA mimic. 3(7):e1701620, 2017. doi: 10.1126/sciadv.1701620. URL https://www.science.org/doi/10.1126/sciadv.1701620. Publisher: American Association for the Advancement of Science.

S. Shmakov, O. O. Abudayyeh, K. S. Makarova, Y. I. Wolf, J. S. Gootenberg, E. Semenova, L. Minakhin, J. Joung, S. Konermann, K. Severinov, F. Zhang, and E. V. Koonin. Discovery and functional characterization of diverse class 2 CRISPR-cas systems. 60(3):385–397, 2015. ISSN 1097-2765. doi: 10.1016/j.molcel.2015.10.008. URL https://www.sciencedirect.com/science/article/pii/S1097276515007753.

S. Y. Stanley, A. L. Borges, K.-H. Chen, D. L. Swaney, N. J. Krogan, J. Bondy-Denomy, and A. R. Davidson. Anti-CRISPR-associated proteins are crucial repressors of anti-CRISPR transcription. *Cell*, 178(6):1452–1464.e13, Sept. 2019.

M. Steinegger and J. Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. 35(11):1026–1028, 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL https://www.nature.com/articles/nbt.3988. Publisher: Nature Publishing Group.

J. Söding, A. Biegert, and A. N. Lupas. The HHpred interactive server for protein homology detection and structure prediction. 33:W244–W248, 2005. ISSN 0305-1048. doi: 10.1093/nar/gki408. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1160169/.

A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

F. Tesson, E. Huiting, L. Wei, J. Ren, M. Johnson, R. Planel, J. Cury, Y. Feng, J. Bondy-Denomy, and A. Bernheim. Exploring the diversity of anti-defense systems across prokaryotes, phages, and mobile genetic elements, 2024. URL https://www.biorxiv.org/content/10.1101/2024.08.21.608784v1. Pages: 2024.08.21.608784 Section: New Results.

The Gene Ontology Consortium, S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris, D. P. Hill, R. Lee, H. Mi, S. Moxon, C. J. Mungall, A. Muruganugan, T. Mushayahama, P. W. Sternberg, P. D. Thomas, K. Van Auken, J. Ramsey, D. A. Siegele, R. L. Chisholm, P. Fey, M. C. Aspromonte, M. V. Nugnes, F. Quaglia, S. Tosatto, M. Giglio, S. Nadendla, G. Antonazzo, H. Attrill, G. dos Santos, S. Marygold, V. Strelets, C. J. Tabone, J. Thurmond, P. Zhou, S. H. Ahmed, P. Asanitthong, D. Luna Buitrago, M. N. Erdol, M. C. Gage, M. Ali Kadhum, K. Y. C. Li, M. Long, A. Michalak, A. Pesala, A. Pritazahra, S. C. C. Saverimuttu, R. Su, K. E. Thurlow, R. C. Lovering, C. Logie, S. Oliferenko, J. Blake, K. Christie, L. Corbani, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, C. Smith, A. Cuzick, J. Seager, L. Cooper, J. Elser, P. Jaiswal, P. Gupta, P. Jaiswal, S. Naithani, M. Lera-Ramirez, K. Rutherford, V. Wood, J. L. De Pons, M. R. Dwinell, G. T. Hayman, M. L. Kaldunski, A. E. Kwitek, S. J. F. Laulederkind, M. A. Tutaj, M. Vedi, S.-J. Wang, P. D'Eustachio, L. Aimo, K. Axelsen, A. Bridge, N. Hyka-Nouspikel, A. Morgat, S. A. Aleksander, J. M. Cherry, S. R. Engel, K. Karra, S. R. Miyasato, R. S. Nash, M. S. Skrzypek, S. Weng, E. D. Wong, E. Bakker, T. Z. Berardini, L. Reiser, A. Auchincloss, K. Axelsen, G. Argoud-Puy, M.-C. Blatter, E. Boutet, L. Breuza, A. Bridge, C. Casals-Casas, E. Coudert, A. Estreicher, M. Livia Famiglietti, M. Feuermann, A. Gos, N. Gruaz-Gumowski, C. Hulo, N. Hyka-Nouspikel, F. Jungo, P. Le Mercier, D. Lieberherr, P. Masson, A. Morgat, I. Pedruzzi, L. Pourcel, S. Poux, C. Rivoire, S. Sundaram, A. Bateman, E. Bowler-Barnett, H. Bye-A-Jee, P. Denny, A. Ignatchenko, R. Ishtiaq, A. Lock, Y. Lussi, M. Magrane, M. J. Martin, S. Orchard, P. Raposo, E. Speretta, N. Tyagi, K. Warner, R. Zaru, A. D. Diehl, R. Lee, J. Chan, S. Diamantakis, D. Raciti, M. Zarowiecki, M. Fisher, C. James-Zorn, V. Ponferrada, A. Zorn, S. Ramachandran, L. Ruzicka, and M. Westerfield. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, May 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad031. URL https://doi.org/10.1093/genetics/iyad031.

C. N. Trost, J. Yang, B. Garcia, Y. Hidalgo-Reyes, B. C. M. Fung, J. Wang, W.-T. Lu, K. L. Maxwell, Y. Wang, and A. R. Davidson. An anti-CRISPR that pulls apart a CRISPR–cas complex. 632(8024):375–382, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07642-3. URL https://www.nature.com/articles/s41586-024-07642-3. Publisher: Nature Publishing Group.

O. G. Troyanskaya, O. Arbell, Y. Koren, G. M. Landau, and A. Bolshoy. Sequence complexity profiles of prokaryotic genomicsequences: A fast algorithm for calculating linguisticcomplexity. 18(5):679–688, 2002. ISSN 1367-4811, 1367-4803. doi: 10.1093/bioinformatics/18.5.679. URL https://academic.oup.com/bioinformatics/article/18/5/679/199697.

M. van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C. L. M. Gilchrist, J. Söding, and M. Steinegger. Fast and accurate protein structure search with foldseek. 42(2):243–246, 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0. URL https://www.nature.com/articles/s41587-023-01773-0. Publisher: Nature Publishing Group.

R. Verkuil, O. Kabeli, Y. Du, B. I. M. Wicky, L. F. Milles, J. Dauparas, D. Baker, S. Ovchinnikov, T. Sercu, and A. Rives. Language models generalize beyond natural proteins, 2022. URL https://www.biorxiv.org/content/10.1101/2022.12.21.521521v1. Pages: 2022.12.21.521521 Section: New Results.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

J. Wang, W. Dai, J. Li, R. Xie, R. A. Dunstan, C. Stubenrauch, Y. Zhang, and T. Lithgow. PaCRISPR: a server for predicting and visualizing anti-CRISPR proteins. 48:W348–W357, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa432. URL https://doi.org/10.1093/nar/gkaa432.

J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh,

I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker. De novo design of protein structure and function with RFdiffusion. 620 (7976):1089–1100, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL https://www.nature.com/articles/s41586-023-06415-8. Publisher: Nature Publishing Group.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL http://arxiv.org/abs/2201.11903.

F. A. Wolf, P. Angerer, and F. J. Theis. SCANPY: large-scale single-cell gene expression data analysis. 19 (1):15, 2018. ISSN 1474-760X. doi: 10.1186/s13059-017-1382-0. URL https://doi.org/10.1186/s13059-017-1382-0.

Y. Zhang and J. Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. 33 (7):2302–2309, 2005. ISSN 0305-1048. doi: 10.1093/nar/gki524. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1084323/.

L. Zimmermann, A. Stephens, S.-Z. Nam, D. Rau, J. Kübler, M. Lozajic, F. Gabler, J. Söding, A. N. Lupas, and V. Alva. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. 430(15):2237–2243, 2018. ISSN 0022-2836. doi: 10.1016/j.jmb.2017.12.007. URL https://www.sciencedirect.com/science/article/pii/S0022283617305879.

Semantic mining of functional *de novo* genes from a genomic language model

Supplementary Material

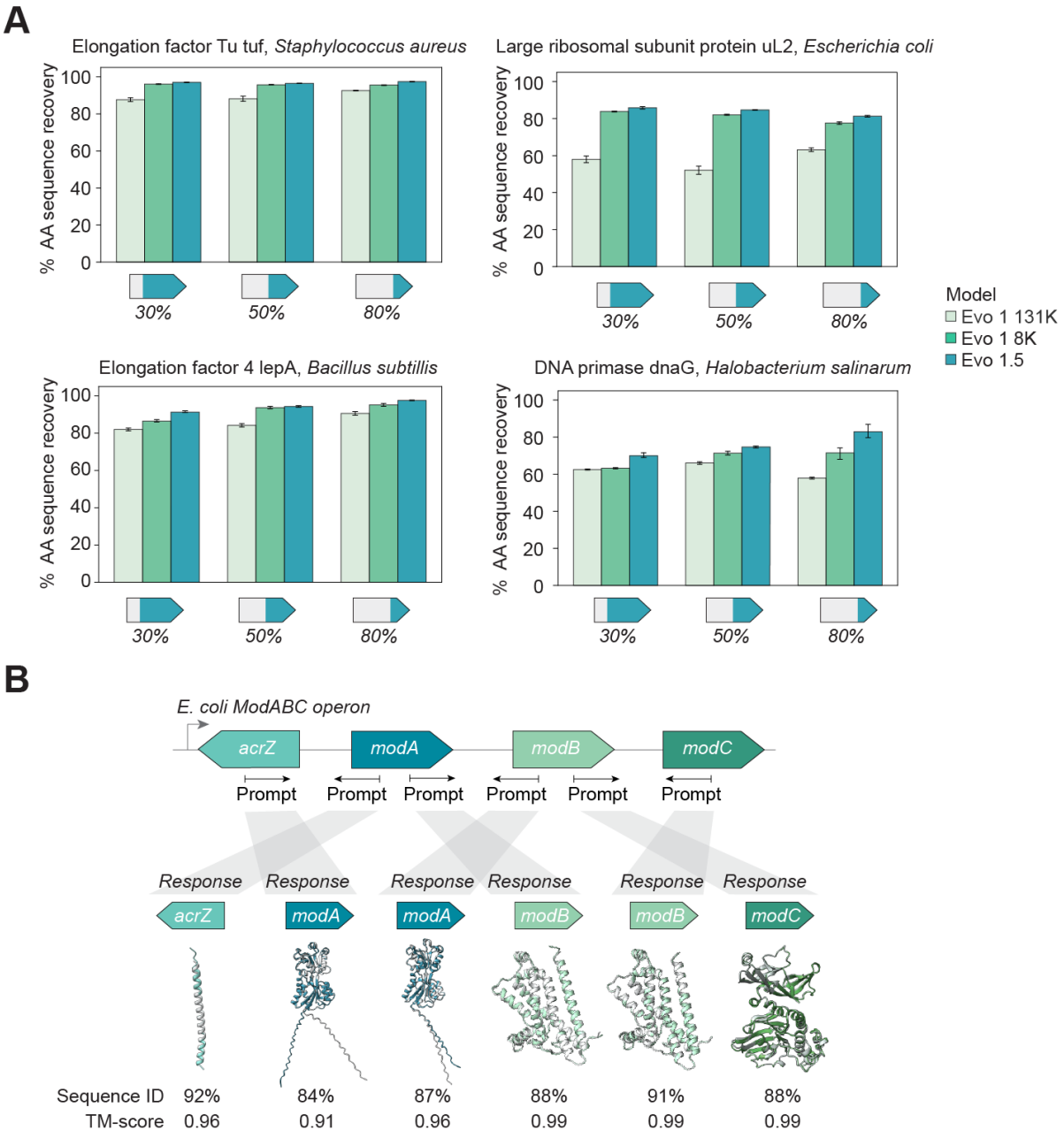


Figure S1 | **Additional gene and operon completion evaluations. (A)** Bar graph showing gene completion of conserved genes *EF-Tu*, *uL2*, *lepA*, and *dnaG* highlights the ability of Evo to understand genomic localization from prompts alone. AA, amino acid; Bar height, mean; error bars, standard error; $n = 100$. **(B)** Bidirectional completion of conserved E. coli modABC operon gene sequences demonstrates high sequence identity and structural conservation across the operon.

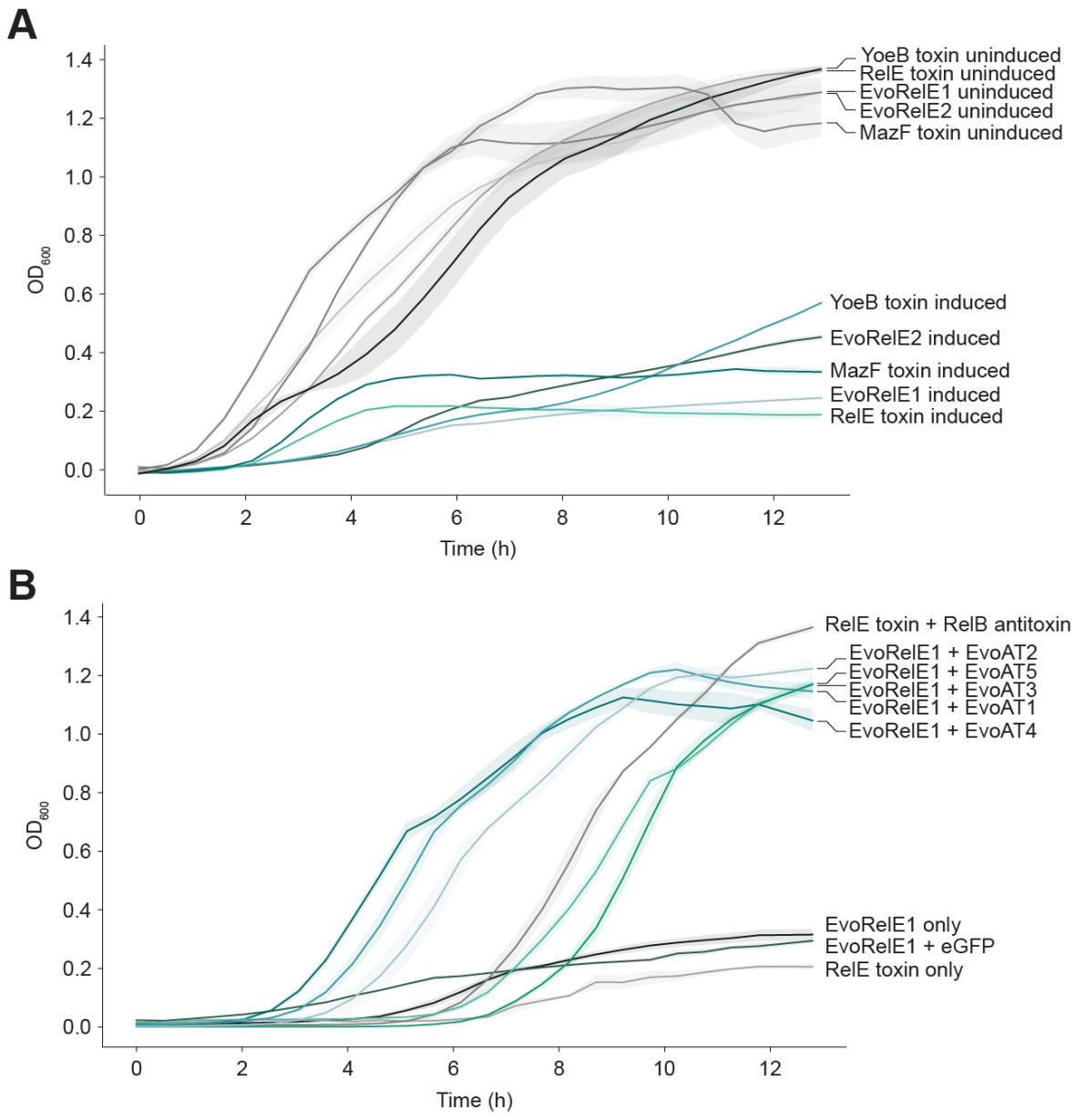


Figure S2 | **Growth curves for toxin and antitoxin growth inhibition and rescue assays. (A)** Growth curves show growth arrest in E. coli when toxins EvoRelE1, EvoRelE2, and all native toxins are induced. **(B)** Growth curves show growth rescue when EvoAT1-5 are present in combination with EvoRelE1 and growth arrest when only EvoRelE1 is present.

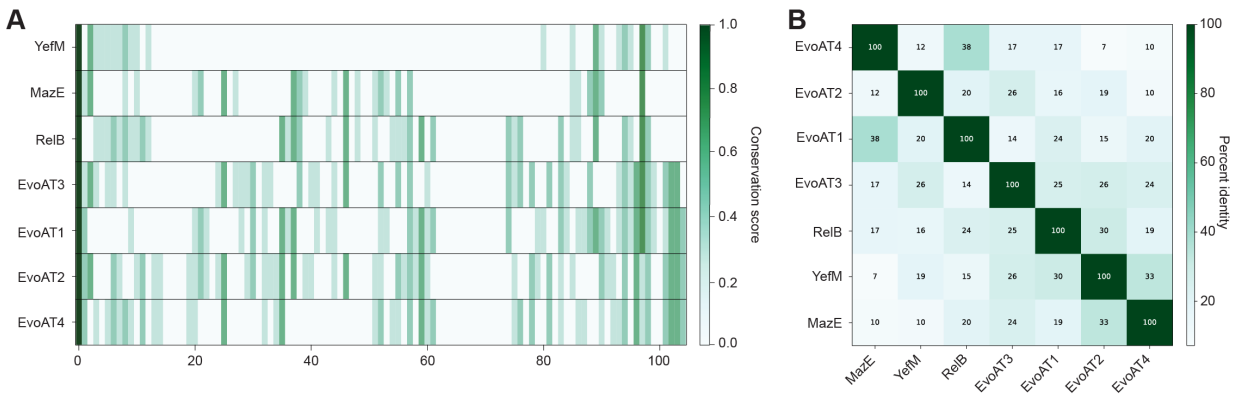


Figure S3 | **Alignment of Evo antitoxins with native antitoxins. (A)** Homology analysis of aligned amino acid sequences for EvoAT1-4 against *E. coli* MazE, YefM, and RelB, with darker green indicating higher conservation. **(B)** Percent identity matrix for EvoAT1-4 against against E. coli MazE, YefM, and RelB, with darker green representing higher percent identity. EvoAT1-4 share limited homology with wild-type antitoxins.

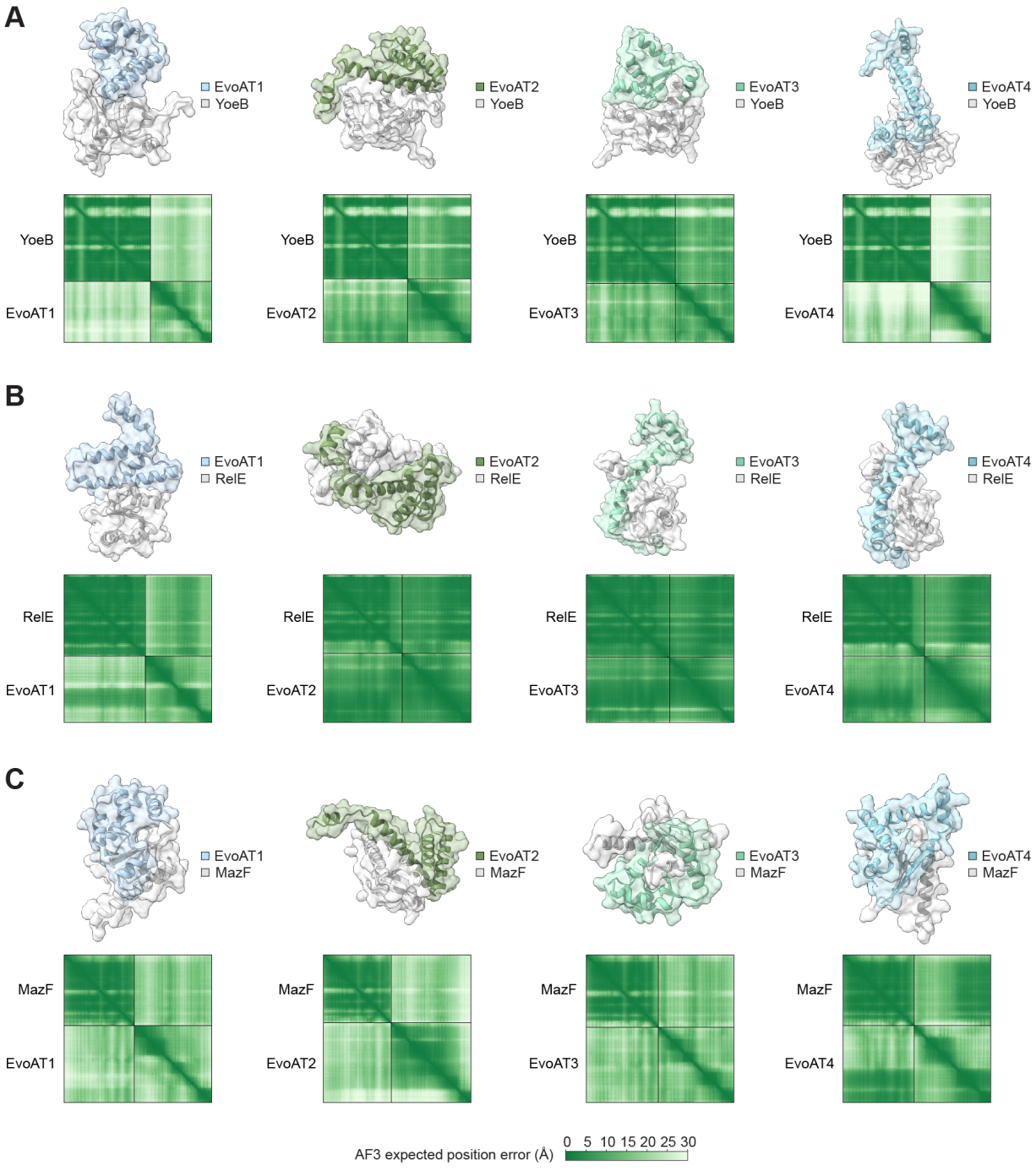


Figure S4 | **AlphaFold 3 structure predictions for Evo antitoxins in complex with native toxins. (A)** Evo antitoxins in complex with YoeB **(A)**, RelE **(B)**, and MazF **(C)**.

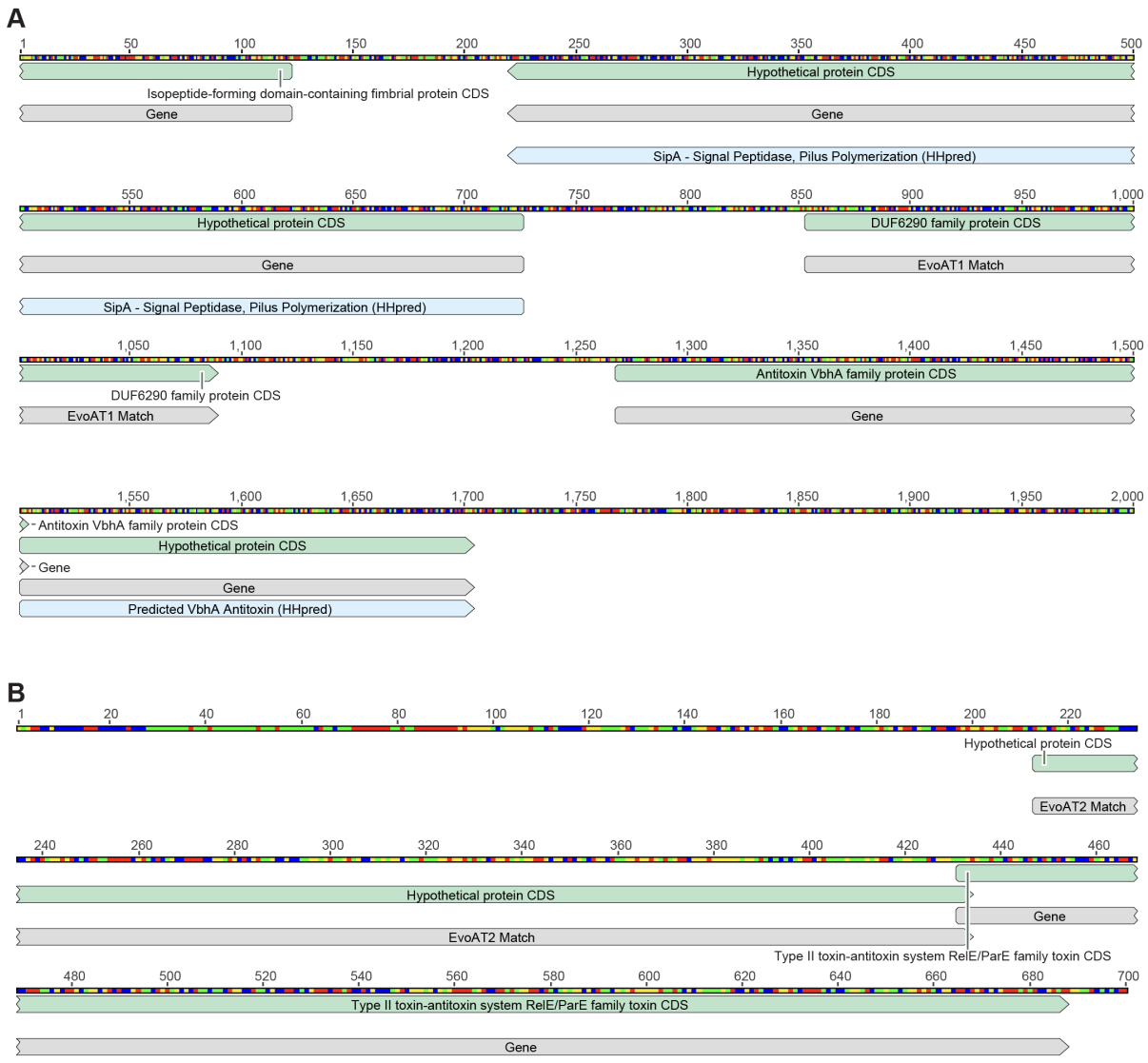


Figure continued on next page

Figure S5 | **Operonic context of Evo antitoxin sequence matches. (A-D)** Genes known and predicted to be related to toxin-antitoxin systems can be found in the immediate vicinity of the closest sequence identity matches for EvoAT1-4.

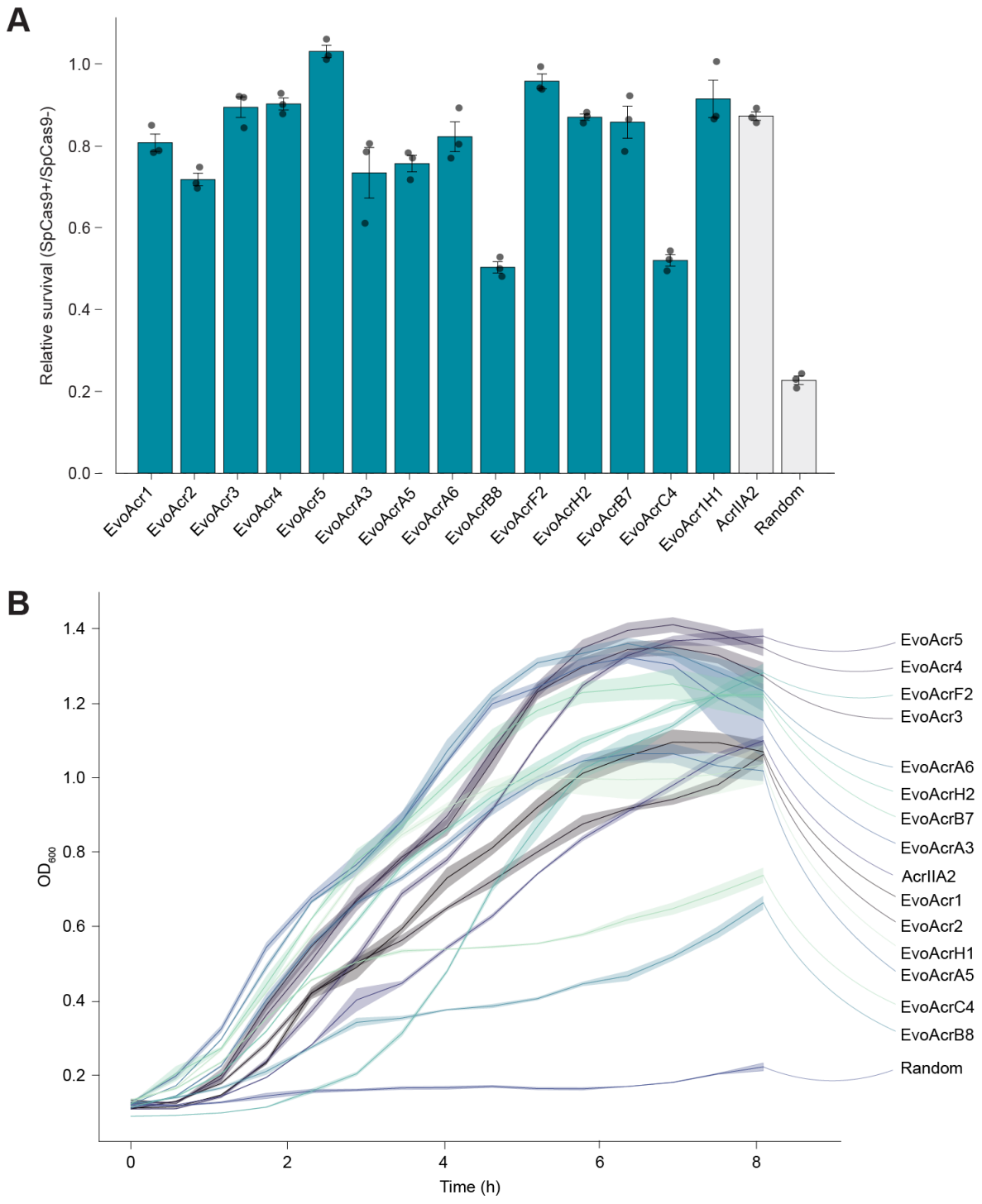


Figure S6 | **Growth curves for Acr protection assay.** **(A)** Bar plot showing relative survival rates for all successful Evo-generated Acrs. Bar height, mean; error bars, standard error; circles, biological replicate values; $n = 3$. **(B)** Growth curves show growth rescue when functional EvoAcrs are present with KanR-targeted SpCas9 and cell death when only non-functional sequences are present.

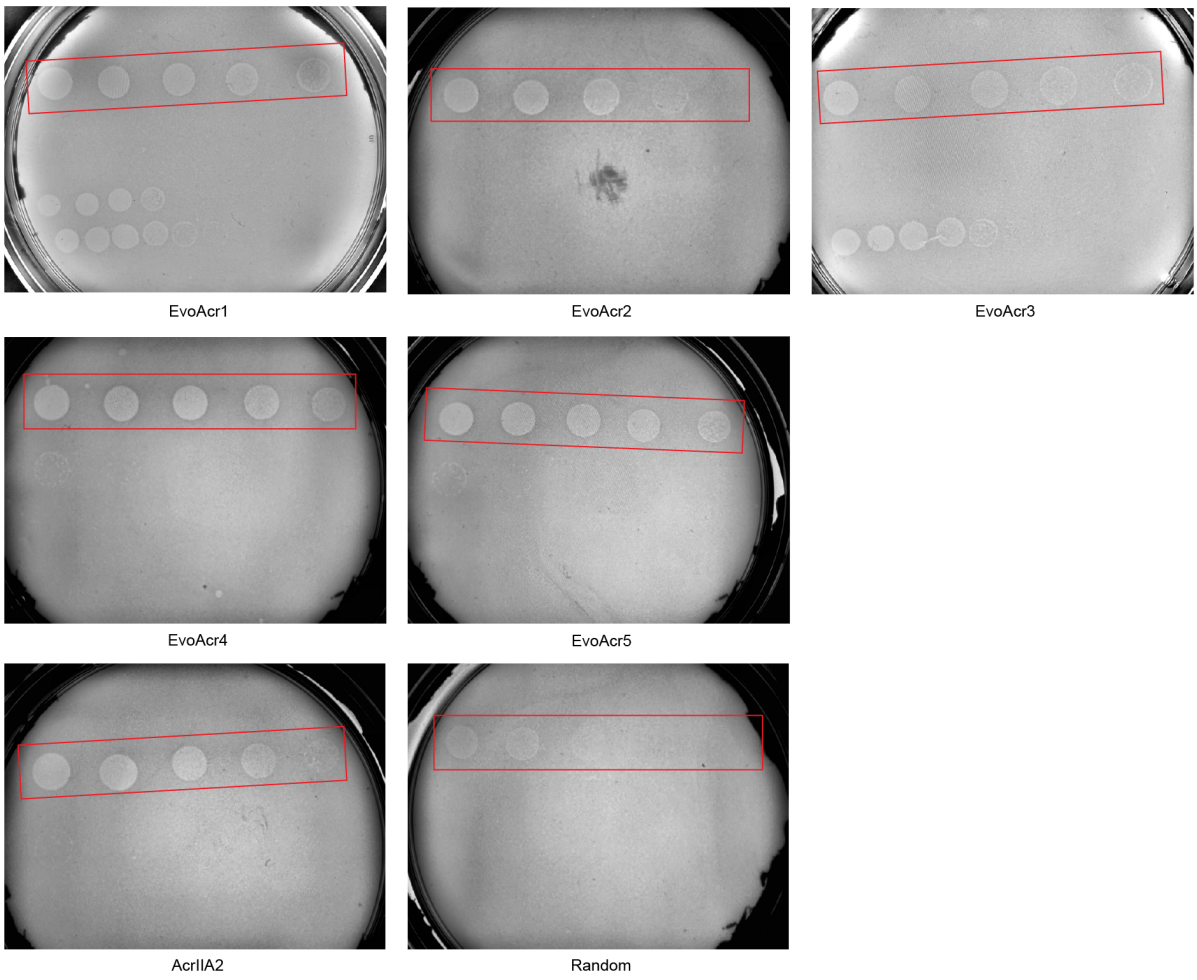


Figure S7 | **Phage plaque images for Acr protection assay.** Uncropped phage plaque images depicting formation of plaques in response to EvoAcr1-5, AcrIIA2, and a random sequence.

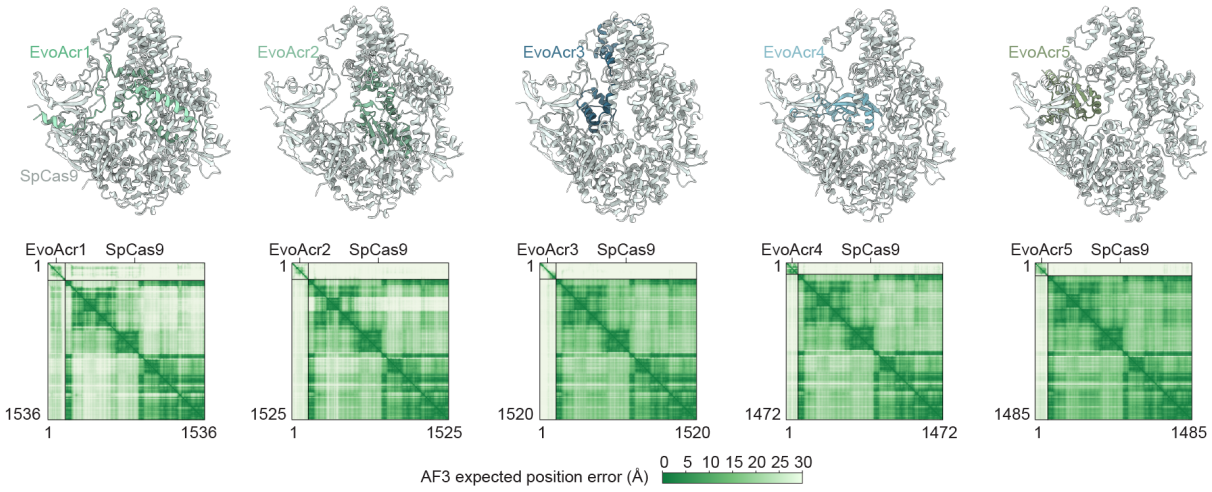


Figure S8 | **AlphaFold 3 structure predictions for Evo Acrs in complex with SpCas9.** AlphaFold 3 structure predictions of Evo Acrs in complex with SpCas9 (top) and their corresponding predicted aligned error (PAE) plots (bottom).

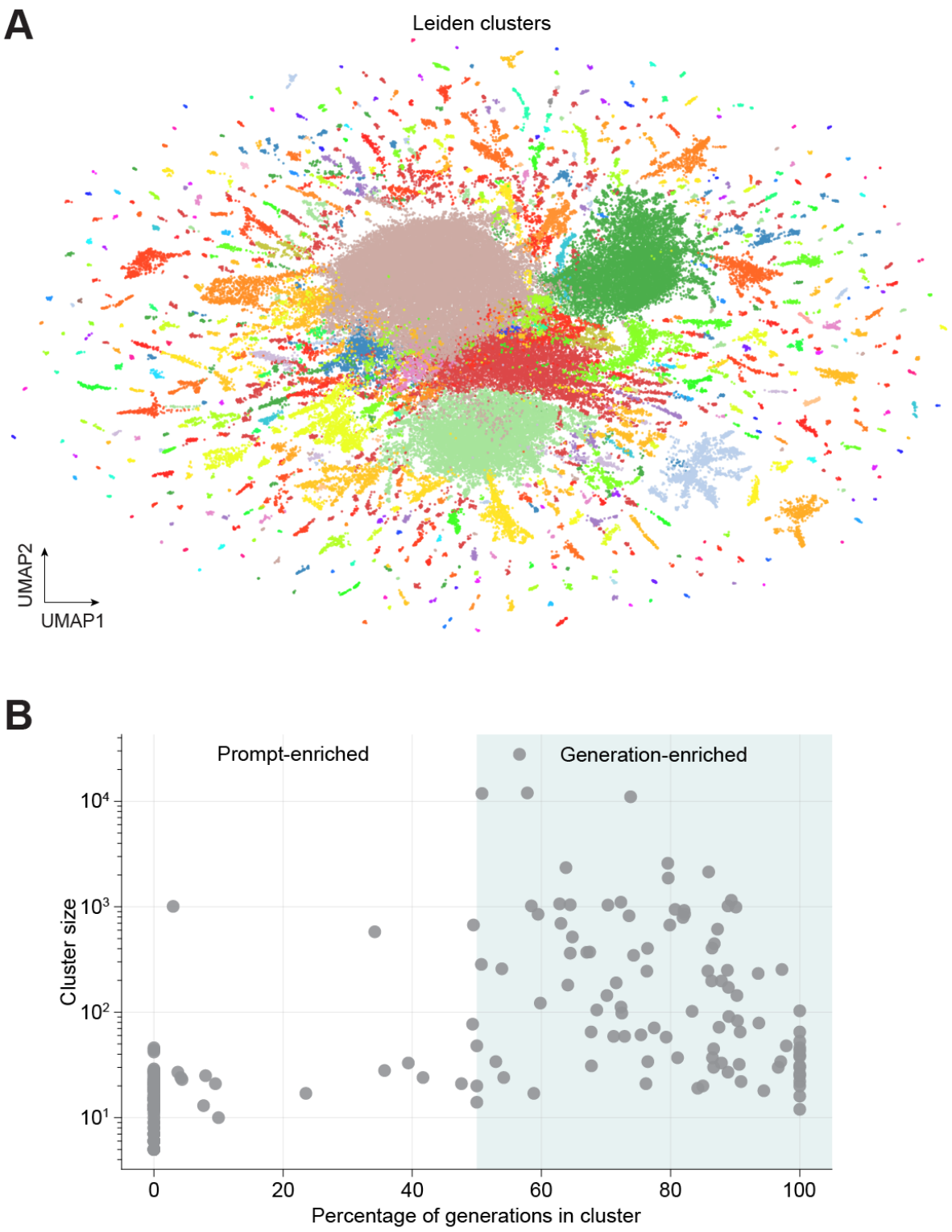
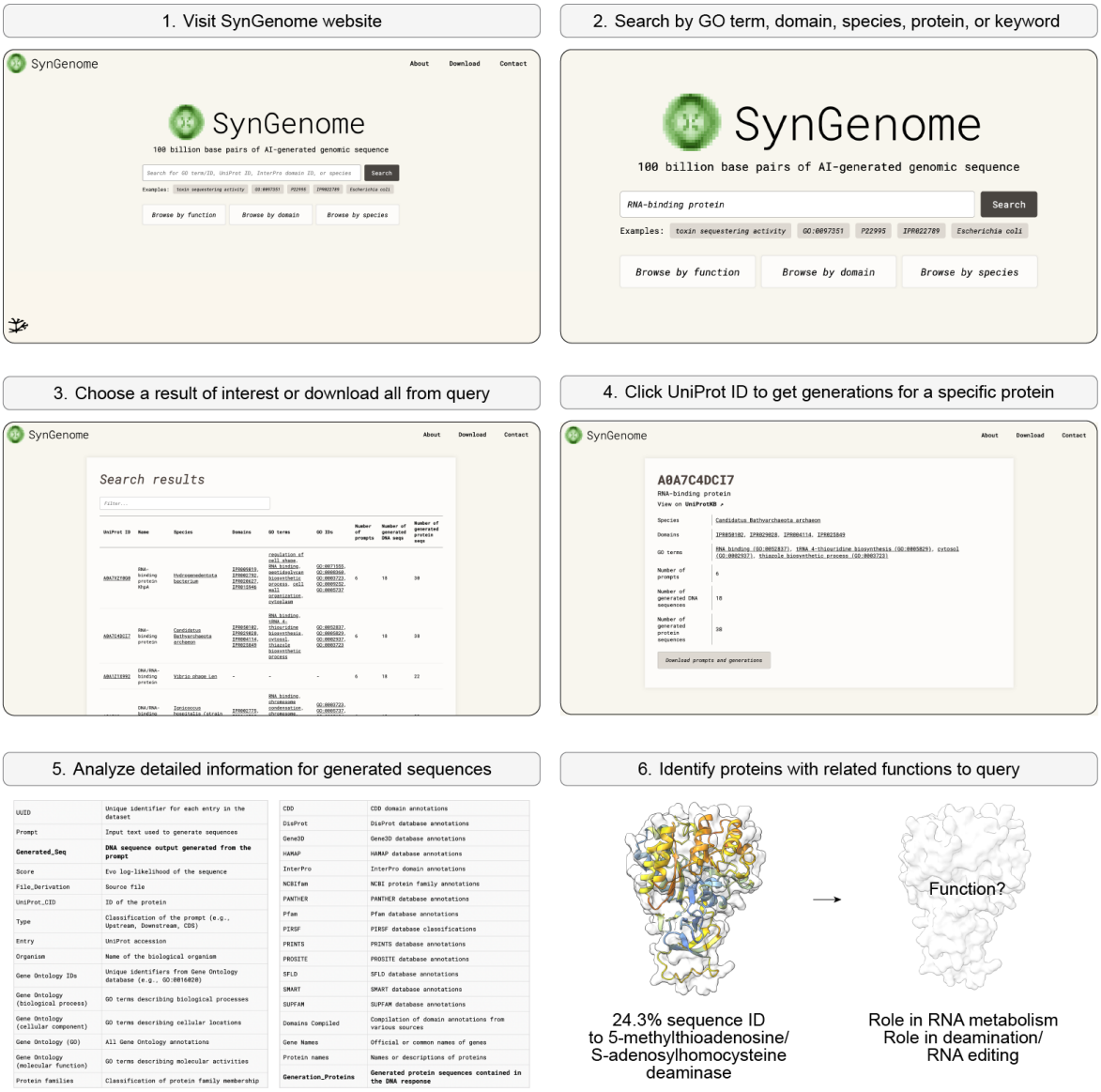


Figure S9 | **Extended Leiden cluster analysis of SynGenome. (A)** UMAP of Leiden clusters of Evo embeddings of prompt sequences and generated sequences. **(B)** Distribution of prompts and generations across Leiden clusters.



UUID	Unique identifier for each entry in the dataset
Prompt	Input text used to generate sequences
Generated_Seq	**DNA sequence output generated from the prompt**
Score	Evo log-likelihood of the sequence
File_Derivation	Source file
UniProt_CID	ID of the protein
Type	Classification of the prompt (e.g., Upstream, Downstream, CDS)
Entry	UniProt accession
Organism	Name of the biological organism
Gene Ontology IDs	Unique identifiers from Gene Ontology database (e.g., GO:0016020)
Gene Ontology (biological process)	GO terms describing biological processes
Gene Ontology (cellular component)	GO terms describing cellular locations
Gene Ontology (GO)	All Gene Ontology annotations
Gene Ontology (molecular function)	GO terms describing molecular activities
Protein families	Classification of protein family membership

CDD	CDD domain annotations
DisProt	DisProt database annotations
Gene3D	Gene3D database annotations
HAMAP	HAMAP database annotations
InterPro	InterPro domain annotations
NCBIfam	NCBI protein family annotations
PANTHER	PANTHER database annotations
Pfam	Pfam database annotations
PIRSF	PIRSF database classifications
PRINTS	PRINTS database annotations
PROSITE	PROSITE database annotations
SFLD	SFLD database annotations
SMART	SMART database annotations
SUPFAM	SUPFAM database annotations
Domains Compiled	Compilation of domain annotations from various sources
Gene Names	Official or common names of genes
Protein names	Names or descriptions of proteins
Generation_Proteins	**Generated protein sequences contained in the DNA response**

Figure S10 | **Example walkthrough of the SynGenome web-interface workflow.** SynGenome can easily be queried with a domain, function, species, or keyword of interest to find relevant generated sequences, which can then be downloaded to identify potential functionally-related protein candidates. We demonstrate the applicability of SynGenome by showing its generation of a potential RNA modification enzyme in response to a RNA binding protein-derived prompt.