GENERATIVE PROTEIN DESIGN FOR OVERLAPPING GENES

Chenling Antelope Xu, Jennifer Lynn Chlebeck, Dan Mcfarland Park, Jonathan E Allen
Lawrence Livermore National Lab
{xu26@llnl.gov}

Hunter Nisonoff
University of California, Berkeley

ABSTRACT

Successfully designed overlapping genes are protected from mutations and horizontal gene transfers, and an effective computational design method can have a large impact on stabilizing genetic constructs in synthetic biology. However, designing overlapping protein sequences in an alternative reading frame is a challenging task because it requires substantial sequence changes to both proteins, and the sequence space is constrained by their shared DNA sequences. We present an experimental dataset of highly divergent overlapping gene pair sequences with functional validation designed using a previously published method CAMEOS. We propose an iterative method for designing overlapping genes making use of a cutting-edge frontier generative model ESM3 and compare its output to the experimentally validated sequences. Our results highlight the surprising effectiveness of ESM3 at predicting *in vitro* protein fitness with only structure information. We used the new approach to generate over 2800 overlapping sequence designs with ESM3 computed scores higher than the minimum score of experimentally validated variants.

1 INTRODUCTION

1.1 OVERLAPPING GENE AS A UNIQUE CONSTRAINED PROTEIN DESIGN PROBLEM

Figure 1: **Introduction to Gene Overlap A**. SARS-CoV-2 ORF3 **B**. Protective effect of gene overlap

Genes are considered "overlapping" when multiple distinct protein sequences are encoded by the same DNA sequence. Although overlooked in most molecular biology courses, overlapping genes are common occurrences in the genome. One example is the SARS-CoV-2 open reading frame ORF3, which contains 4 overlapping proteins in two different reading frames (Figure1A) (Rubio et al., 2023). Overlapping genes are also pervasive in eukaryotic and bacterial genomes (Wright et al., 2022). One property of overlapping genes is that mutations in the DNA sequence will induce changes in both proteins, although one of them could be synonymous (Figure1B). This property can be used to decrease the mutation rate and prevent horizontal gene transfer (Chlebek et al., 2023; Blazejewski et al., 2019).
Currently, there is only one experimentally validated method for designing overlapping genes. CAMEOS (Blazejewski et al., 2019) uses a two-step process that first aligns two proteins to find

the best location for overlap, then optimizes amino acid mutations using a Potts model. This approach has successfully generated overlapping gene pairs that allowed experimental testing of gene overlap, but the success rate of generating overlapping gene pairs where both proteins remain functional is low, making it difficult to deploy gene overlap more widely. This has motivated an alternative approach RiBoSor that inserts a ribosomal binding site intoto an alternative reading frame of the upstream gene to creatoverlap of the coding sequence sequence (Decrulle et al., 2021). The downstream gene acquires extra amino acids from the out-of-frame sequence, but its original coding sequence remains the same. This paper also shows significant protection against mutation as a result of sequence overlap. It is easier to generate functional overlapping pairs using RiboSor due to the limited scope of protein sequence changes, but CAMEOS offers a higher level of mutational protection. Given the utility of gene overlap and recent progress in protein engineering, we aim to develop a new computational method combining the advantage of both methods using a generative approach.

2 METHOD

2.1 FUNCTIONAL ASSAY FOR CAMEOS DESIGNED ENTANGLEMENT PAIR INFA/AROB

To generate an initial data set for model training, we initiated a DBTL campaign of an entanglement pair composed of infA and aroB. The infA gene encodes a 72-amino-acid translation initiation factor, essential for growth, while aroB encodes a 362-amino-acid enzyme required for aromatic amino acid biosynthesis. We used an updated version of CAMEOS (Martí et al., 2024) to generate 130,000 entanglement designs that included infA entangled at various locations within the longer gene, aroB. Among these, 2000 designs were selected for experimental testing and split evenly between the two most common entanglement positions (ERP017 and ERP766). The designed variants of overlapping infA/aroB pair are distinct because of their high divergence from known wild-type sequences. Compared to the initial wild-type sequences, the average number of changes (including deletions) is 47.5 (65.9%) for infA and 52.8 (14.4%) for aroB. The sequences in the library are highly divergent from other sequences in the library as well, averaging 5 amino acid changes to the nearest neighboring sequence. The functionality of infA and aroB was assessed separately through growth-based assays. Enrichment factors, calculated as the log ratio of normalized read counts before and after selection, were used to define functional sequences. Thresholds of -1 for infA and -2 for aroB were set based on enrichment relative to control sequences. The final enrichment factor was the median of three replicates, with the wild-type factor subtracted. The median log enrichment factor was 2.67 for aroB and 2.26 for infA. Thus, a -2 threshold for aroB indicates a 2-fold increase in sequencing reads, while -1 for infA indicates a 3-fold increase.

2.2 EVALUATION OF COMPUTATIONAL METRIC FOR PROTEIN FITNESS IN HIGHLY DIVERGENT VARIANTS

ESM3(Hayes et al., 2025) version esm3-medium-2024-08 was used as an inverse folding model by providing ESM3 with atom 3D coordinate and fully or partially masked amino acid sequence. We did not use a crystal structure because they were not available for infA and aroB for the *Pseudomonas protegens Pf-5* wild-type sequence. Instead we generated structure from wild-type amino acid sequences using the Alphafold server (Abramson et al., 2024) with default parameters. To compute an ESM3-based sequence score for each experimentally-validated sequence, we summed over the logits at each position given the structure with fully masked amino acid sequence. This can be considered as the cross entropy loss between the sequence and the model prediction because the normalizing factor is the same for every sequence and can be ignored. The second metric we computed was the Potts model energy function of all aligned positions of tested variants. The Potts model parameters were inferred using CCMpred from a MSA alignment built by phmmer. Homologous sequence were identified by searching the Uniref100 sequence database and keeping sequences with sequence bit score threshold above $0.5 \times sequence\ length$. The Potts model energy score was computed using custom python code. The Potts model is recomputed for evaluation, and is not the same one used by CAMEOS for design.

2.3 GENERATING OVERLAPPING SEQUENCES USING ESM3

For the design task, we start with two amino acid sequences and identify amino acid positions that cannot be coded by the same DNA sequence. In these positions we keep the starting sequence in one of the genes, and mask the identity of those positions in the other protein. During this process we also identify a DNA sequence that encodes all unmasked amino acids in both sequences in alternate frames. We then generate new sequences conditioned on existing overlapping amino acids and the protein structure using ESM3 for the masked gene. Since all compatible positions are unmasked and remain unchanged in the generation process, and newly generated sequence can become compatible with the fixed sequence, the number of compatible positions increase iteratively. The end result are two protein sequences encoded by the same DNA molecule in alternative frames. One of these proteins are generated using ESM3, and the other one is partially generated by ESM3 in the last iteration while the masked position takes the amino acids the shared DNA molecule translates to. We refer to the two sequences as the "generated" and "translated" sequence. We improved the efficiency of this sampling process by only accepting new sequences with overall improved Potts energy, and running multiple threads from best variants from previous iterations.

3 RESULTS

3.1 EXPERIMENTAL VALIDATION OF THE INFA/AROB PAIR

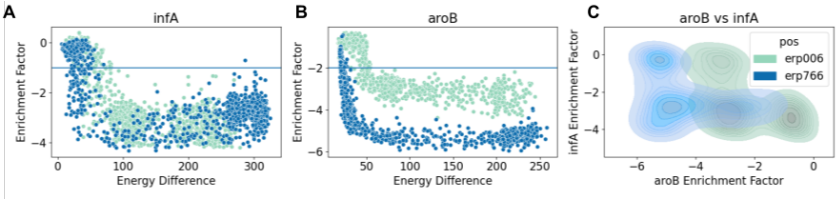

Figure 2: **Enrichment Factor for infA aroB A**. infA enrichment factor as a function of the Potts model energy difference between the variant and wild-type control. The blue line is the enrichment factor threshold for a functional variant. **B**. Same as A for aroB. **C**. Kernel density estimate plot for infA and aroB enrichment factor joint distribution for two different insertions sites.

The relationship between enrichment factor and Potts model energy in the experimental validation for CAMEOS is visualized in Figure2. For both infA and aroB the Potts model energy function is predictive of functionality, where only variants with energy value close to the wildtype have the potential to be functional *in vivo* (Figure2AB). Around 17% of infA variants and 19% of aroB variants were enriched in the surviving population, indicative of protein function. Due to the trade-off of the two genes, no overlapping pair was functional as can be seen in the lack of sequences in the upper right corner in Figure2C. This highlights the difficulty of the problem and is consistent with the cysJ-infA library tested in the CAMEOS paper where 6 unique functional overlapping pairs were identified from 7500 unique designs.

3.2 GENERATIVE APPROACH

An alternative approach is to design overlapping sequences with generative machine learning models. We show the feasibility of this approach using ESM3, but in theory this approach could be used with any other inverse-folding models. CAMEOS generated sequences can be thought of as samples with overlapping constraint amino acid residues from the site and pairwise site frequency (Potts model) of natural sequences. We can have a wider choice of sequence distributions using state-of-the-art generative models. Here we sampled sequences conditioning on the 3D structure using the ESM3 model. We validated this approach using the previously mentioned experimental dataset. We computed the AUC score using either the ESM3-based cross entropy score or the Potts energy to predict the functionality of protein variants. Using a double-sided t-test, ESM3-based cross entropy score has a higher AUC score than Potts energy with p-value 9.6e-32 for aroB, while there is no significant difference between the two scores for infA. This

is particularly impressive because ESM3 is a zero-shot model. This assured us that conditioning on structure provides sufficient information for protein function for the following design process.

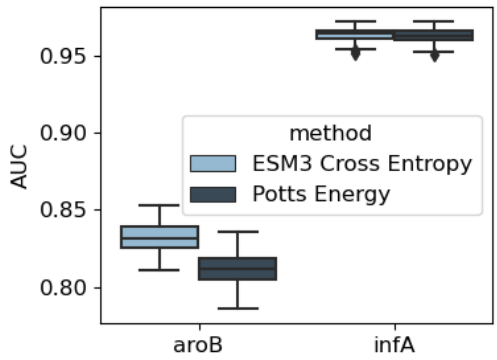

Figure 3: **Area Under Curve (AUC) Score for Predicting Sequence Function using ESM3 Cross Entropy and Potts Energy** The uncertainty is generated by bootstrap.

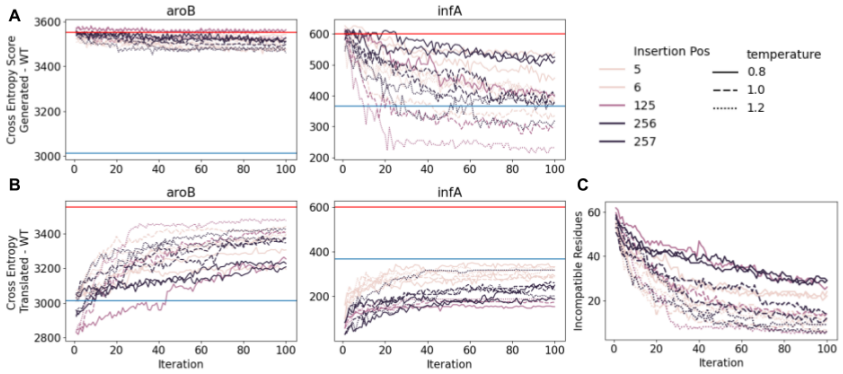

Figure 4: **Quality of generated overlapping sequence over iterations** Red line indicates the value of wild-type sequence, and blue line indicates the minimum value corresponding to all experimentally validated functional sequences. The x-axis indicates the timestep in the iterative process and the color and line type indicates the parameters used in the sequence generation. **A**. Cross entropy score of generated aroB and infA sequences. **B**. Cross entropy score of translated aroB and infA sequence. **C** The number of incompatible sites between the two generated protein sequences at each iteration.

We demonstrated the effect of the iterative overlapping sequence generation on protein fitness in (Figure4). At each iteration we obtain a DNA sequence that encodes the generated sequence faithfully and the translated sequence with minimal number of changes. Using 15 different parameter sets (five insertion location and three sequence generation temperatures), we generated and computed the cross entropy scores for the generated and translated sequence at each iteration. As expected, cross entropy scores decrease with the number of iterations(Figure4A). The translated sequences consistently score lower, but improve over time. (Figure4B). The optimization process of 100 iterations takes about 1 hour using the forge client provided by EvolutionaryScale. A minimum of 5 incompatible sites out of 74 total overlapping codons remain (Figure4C) after 100 iteration steps. However, this result is achieved with higher temperatures, which increase the chance of finding compatible amino acids, but decrease the generated cross entropy score. The position of the insertion has an effect on protein fitness but the effect depends on the gene. Compared to aroB, infA is more severely compromised in the translated variant compared to the generated variant. We highlighted the minimum cross entropy score in functional variants in blue, and wild-type cross

entropy score in red. Any sequences above the blue line are expected to be functional according to the experimental data. According to these results, DNA sequences that encode the generated variant of infA and the translated variant of aroB, with infA inserted in position 5 or 6 of aroB have the highest rate of success. Multiple runs led to translated aroB above the functional threshold (Figure 4B), with the corresponding generated infA variants also above the threshold (Figure 4A), suggesting functional overlapping gene pairs. One potential concern is that when using Potts energy as the evaluation metric, the translated aroB sequences do not meet the functional threshold (data not shown until next section), so more testing might be needed to validate this result.

3.3 Strategies to Increase Sampling Efficiency and Diversity

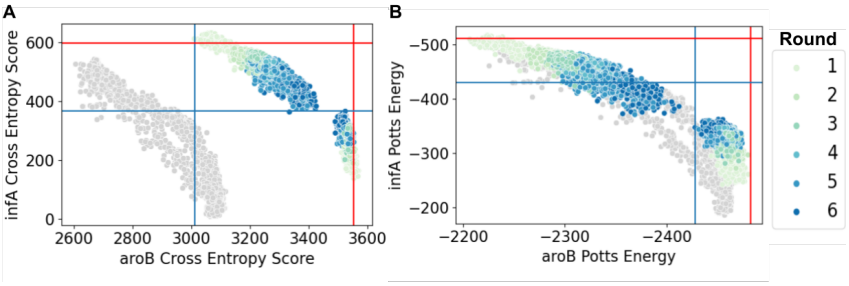

Figure 5: **Sequence quality metrics comparison between CAMEOS-designed and ESM3-generated sequence pairs** The blue line is the minimum score in experimentally validated functional sequences, and the redline is the wild-type sequence score. Any sequence pairs in the top right blue quardrants are expected to be functional. **A**. aroB and infA cross entropy scores, where the gray dots represent the CAMEOS-designed pairs and the ESM3-generated pairs are colored based on the number of rounds taken to reach the design. **B**. Same as Panel A but using Potts energy as the sequence quality metric.

Due to the randomness in generating entangled sequences, we ran multiple replicates for insertion at position 5 with temperature 0.8. The top 10 sequence pairs were selected every 20 iterations as starting points for the next round, boosting efficiency and diversity. We compared cross entropy and Potts energy scores of generated infA/aroB variants to experimentally validated CAMEOS sequences (Figure 5). Later generations showed higher translated aroB and lower generated infA scores, with 2805 out of 2807 gene pairs expected to be functional, compared to none in CAMEOS (Figure5A). We also obtained one pair consisting of a translated infA with close to functional cross entropy score and a generated aroB with close to wild-type cross entropy score in the last round. Unfortunately when evaluating the ESM3 generated sequences with Potts energy, we did not obtain any sequence pairs with Potts energy above the functional threshold, although the generative approach slightly outperformed CAMEOS in obtaining higher infA Potts energy with the same aroB Potts energy (Figure 5B, bottom right).

4 Conclusion and Discussion

Generative design has been shown to perform well in designing highly divergent protein sequences. We first showed that ESM3 model cross entropy score is highly predictive of overlapping gene functionality, then sampled from the model to obtain new overlapped sequences conditional on protein structures. Although it is time-consuming to generate a fully overlapped protein pair by sampling, simply translating an imperfect DNA sequence can produce overlapped protein sequences that perform similarly in terms of Potts energy compared to CAMEOS generated variants in only a few iterations. The advantage of using a generative approach is its ability to integrate additional information beyond evolutionary conservation, in our case protein structure, but in theory we can sample proteins conditioned on other properties such as ligand binding, enzymatic activity, or thermostability provided there is sufficient training data. The disadvantage is the amount of repeated computation in sampling from the pretrained models because the codon constraints are not considered in the sampling step, but rather added post-hoc. Future work could incorporate the number of

incompatible overlapped positions as a sampling constraint in the sampling step, potentially using diffusion models with guidance((Nisonoff et al., 2024; Li et al., 2024)).

ACKNOWLEDGMENTS

This research is funded by the BioSecure: Building Robust Biocontainment Strategies for Evolutionarily Stable Microorganisms Science Focus Area grant. The generative model and compute power is provided by EvolutionaryScale Forge.
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DEAC52-07NA27344 (LLNL-CONF-2002368)

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Tomasz Blazejewski, Hsing-I Ho, and Harris H. Wang. Synthetic sequence entanglement augments stability and containment of genetic information in cells. *Science*, 365(6453), August 2019.

Jennifer L Chlebek, Sean P Leonard, Christina Kang-Yun, Mimi C Yung, Dante P Ricci, Yongqin Jiao, and Dan M Park. Prolonging genetic circuit stability through adaptive evolution of overlapping genes. *Nucleic Acids Research*, 51:7094–7108, July 2023.

Antoine L. Decrulle, Antoine Frénoy, Thomas A. Meiller-Legrand, Aude Bernheim, Chantal Lotton, Arnaud Gutierrez, and Ariel B. Lindner. Engineering gene overlaps to sustain genetic constructs in vivo. *PLoS Computational Biology*, 17:e1009475, October 2021. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1009475.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-Free Guidance in Continuous and Discrete Diffusion Models with Soft Value-Based Decoding, October 2024. arXiv:2408.08252 [cs].

Jose Manuel Martí, Chloe Hsu, Charlotte Rochereau, Chenling Xu, Tomasz Blazejewski, Hunter Nisonoff, Sean P Leonard, Christina S Kang-Yun, Jennifer Chlebek, Dante P Ricci, et al. Gentangle: integrated computational design of gene entanglements. *Bioinformatics*, 40(7):btae380, 2024.

Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking Guidance for Discrete State-Space Diffusion and Flow Models, October 2024. URL `http://arxiv.org/abs/2406.01572`.

Alejandro Rubio, Maria de Toro, and Antonio J. Pérez-Pulido. The most exposed regions of SARS-CoV-2 structural proteins are subject to strong positive selection and gene overlap may locally modify this behavior. *mSystems*, 9(1):e00713–23, December 2023. Publisher: American Society for Microbiology.

Bradley W Wright, Mark P Molloy, and Paul R Jaschke. Overlapping genes in natural and engineered genomes. *Nature Reviews Genetics*, 23(3):154–168, 2022.