Accelerating Protein Engineering with Machine Learning: A Few-Shot Transfer Learning Approach to Designing Novel Gene Activators

M. Zaki Jawaid*, Robin W. Yeo*, Aayushma Gautam, T. Blair Gainous, Daniel O. Hart, Timothy P. Daley

EpiCrispr Biotechnologies South San Francisco 94080 CA, USA

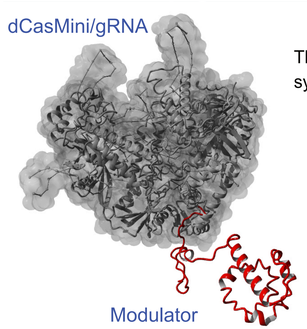
Abstract

Epigenomic CRISPR platforms in which modulator peptides are tethered to catalytically dead Cas molecules and targeted towards a genetic locus present an exciting new avenue for precise modulation of gene expression, thereby unlocking various therapeutic applications. However, the design of novel modulator peptides that are capable of gene activation and repression remains a slow and expensive process due to a variety of protein engineering challenges; in particular, the lack of reliable data that can motivate traditional protein engineering approaches like semi-rational design or binding affinity optimization.

We propose a few-shot machine learning approach to protein design that aims to accelerate the expensive wet lab testing cycle and is capable of leveraging a training dataset that is both small and skewed (~10^5 datapoints, < 1% positive hits).

Our approach is composed of two parts: a semi-supervised transfer learning approach to generate a discrete fitness landscape for desired protein function and a novel evolutionary Monte Carlo Markov Chain sampling algorithm to more efficient fitness landscape exploration. We demonstrate the performance of our approach by generating novel peptide sequences in silico and experimentally screening their ability to activate a fluorescent reporter locus using an engineered compact CRISPR-Cas system.

Gene Expression Modulation System for Epigenome Engineering

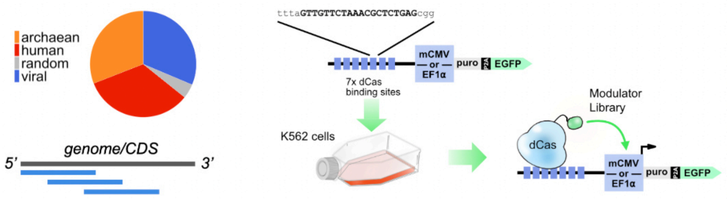


The gene expression and modulation system (GEMS) is composed of:

1) dCasMini (1): A compact, programmable DNA binding protein.
2) One or more guide RNAs.
3) Modulator peptide capable of activating or repressing gene transcription.

Identification of hypercompact transcriptional modulators by high-throughput screening

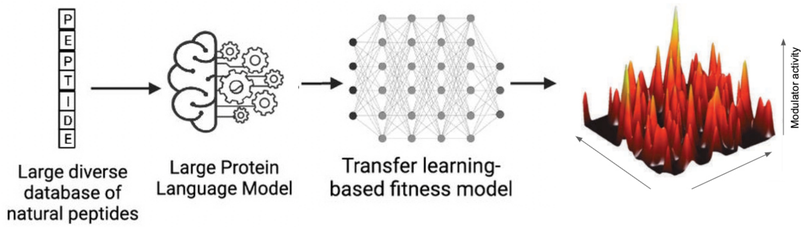
An initial library of 34217 85aa (85 amino acid long) putative modulator peptides from diverse biological origins were experimentally screened for their ability to activate a synthetic genetic locus using dCasMini-GEMS (2).



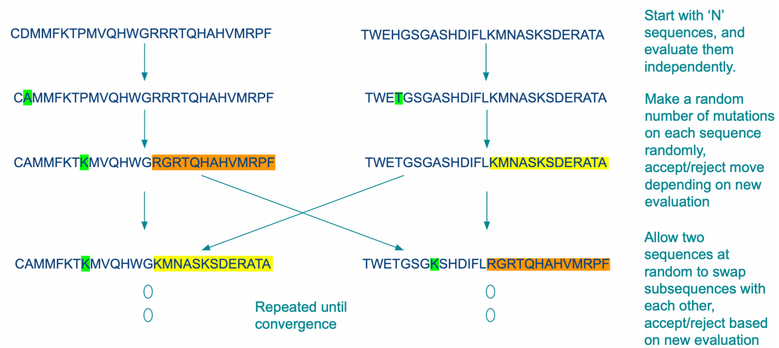
In an independent follow-up screen, a subset of these sequences were re-tested, resulting in 173 sequences that we classified as validated gene activators ("positive hits"), giving a hit rate of 0.51%. Thus, the full library of 34217 85aa peptides was used as our training data set.

Generating a protein fitness landscape for desired protein function using semi-supervised transfer learning

An ensemble model (XGBoost/CNN) was trained on the sequence embeddings of the 650M ESM2 large protein language model to generate a fitness function.

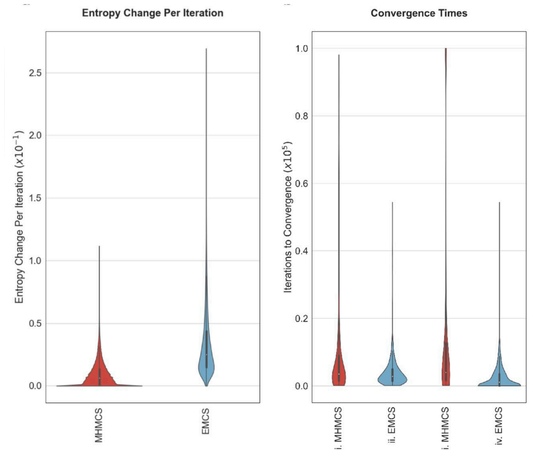


Fitness landscape exploration using novel Evolutionary Monte Carlo Search (EMCS)



EMCS outperforms standard MHMCS sampling

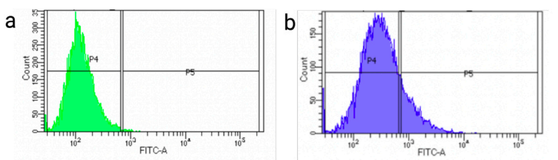
EMCS involves multiple monte carlo chains run simultaneously, In addition to random single mutations in each iteration, we allow genetic crossover events to access more of the fitness landscape.



When compared to Metropolis-Hastings Monte Carlo, EMCS allows greater sequence diversity per iteration as measured by entropy change per iteration, which also results in faster convergence times.

4600 Novel Sequences with high predicted fitness were experimentally validated

Representative FACS histograms illustrating cell counts within GFP_OFF (P4) and GFP_ON (P5) gates in un-infected cells (a) and cells infected with the validation library (b).

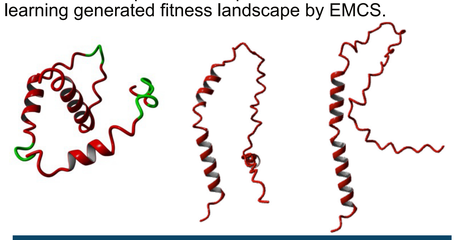


Experimental Results: Sampling Algorithm				
Algorithm	Initialization	Total Sequences	Number of Hits	Hit Percentage
EMCS	All	2600	338	13%
MHMCS	All	2000	18	0.9%
EMCS	Known	1310	270	20.6%
EMCS	Random	1290	68	5.3%
MHMCS	Random	2000	18	0.9%
Negative Controls	n/a	300	1	0.33%

Final hit percentage of novel sequences sorted by choice of sampling algorithm. Initialization: Notes the sampling algorithm starting sequence as either randomly initialized, or known positive. Total Sequences: Number of sequences that were identified as high fitness by the machine learning model. Number of Hits: Number of sequences that validated experimentally.

ESMFold structures of validated hits share structural similarities with training dataset

Despite significant sequential diversity between the validated hits and the training dataset positives (hamming distance of 40-65), the validated hits had a similar secondary structure profile to the training dataset positives. Left: ESMFold structure of a transcriptional modulator in training dataset. Center/Right: ESMFold structures of sequences sampled from the machine learning generated fitness landscape by EMCS.



Conclusions

1. Using EMCS to sample a protein fitness landscape approximated by a semi-supervised transfer learning (ESM Embeddings on XGBoost/CNN) model, we were able to improve our base hit rate of 0.51% to 20.6%, a 40x improvement.
2. EMCS outperforms standard MHMCMC by 4x in final hit rate of experimentally validated molecules.
3. This approach is easily generalizable to various protein engineering tasks.

References

1) Xiaoshu Xu et al. Engineered miniature CRISPR-Cas system for mammalian genome regulation and editing, Molecular Cell, Volume 81 , Issue 20, DOI: 10.1016/j.molcel.2021.08.008.2021
2) Zeming Lin et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science***379**,1123-1130(2023).DOI:10.1126/science.ade2574
3) Carosso et al. Discovery and engineering of hypercompact epigenetic modulators for durable gene activation, BioRxiv 2023.06.02. DOI: 10.1101/2023.06.02.543492
4) Jawaid et al. Improving few-shot learning-based protein engineering with evolutionary sampling. BioRxiv 2023.05.23. DOI: 10.1101.2023.05.23.541997

10.1101/2023.06.02.543492 10.1101/2023.05.23.541997