Accelerating the Discovery of Novel Hypercompact Transcriptional Activators with Machine Learning

M. Zaki Jawaid, Tyler Borrman, T. Blair Gainous, Chris Still, Aayushma Gautam, Dan O. Hart, Timothy P. Daley, Robin W. Yeo

Epicrispr Biotechnologies, South San Francisco 94080 CA, USA

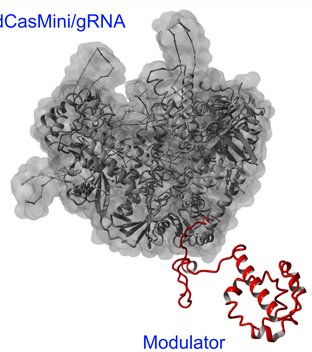
Abstract

Due to the programmability and versatility of CRISPR-Cas systems, gene therapy research is emerging as an increasingly promising strategy to treat genetic disease. However, many genetic diseases, such as those caused by haploinsufficiency, cannot be treated with traditional knock-out based CRISPR gene therapies, since they are caused by allelic loss-of-function. Due to its ability to both activate as well as suppress gene expression, epigenetic editing using CRISPR-dCas systems has the potential to address many genetic diseases that are unsuitable for traditional gene therapies. However, the versatility of epigenetic editing is constrained, especially in the context of gene activation, by the limited number and large size of most modulator peptides, severely limiting their therapeutic utility.

To address these limitations, we have built a generative AI platform capable of designing de novo hypercompact modulator peptides with the ability to transcriptionally upregulate a genetic locus. We first collected a large corpus of training data by performing high-throughput screens to discover novel transcriptional activators among peptides derived from human, viral, and archaeal proteomes. We then trained a machine learning ensemble model, composed of a decision tree model and a convolutional neural network, which leverages transfer learning (via large protein language model embeddings) to predict transcriptional activators based on peptide sequence alone. By exploiting a novel sampling algorithm, which we call evolutionary Monte Carlo search, to more efficiently traverse the predicted activator fitness landscape, we used this machine learning platform to generate a library of several thousand hypercompact peptides predicted to be transcriptional activators.

We experimentally screened these peptides and validated that our generative AI approach dramatically increased discovery rate (up to 45-fold) resulting in the discovery of hundreds of novel transcriptional activators sharing little sequence similarity with known naturally occurring peptides. We next investigated the evolutionary, biochemical, and biophysical properties of the synthetic activator library, revealing that validated activators consistently lack conserved functional domains but do share certain biochemical features, such as strong negative electrostatic potential. We subsequently selected 10 of our top synthetic activators for further characterization, and assessed their activation strength by screening them at an artificial GFP locus as well as at an endogenous human locus (CD45), comparing their potency to gold standard activators (e.g. vCD, VP64). These results demonstrate the capability of machine learning to accelerate the discovery of novel functional peptides to expand our toolbox of epigenetic modulators for future therapeutic applications.

Gene Expression Modulation System (GEMS)

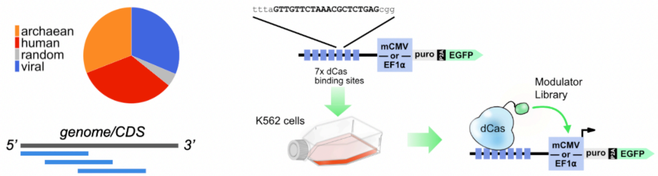


The gene expression and modulation system (GEMS) is composed of:

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

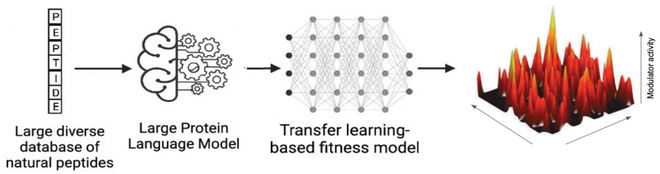
Identification of hypercompact transcriptional modulators by high-throughput screening

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]. After validation and retesting, we classified 173 sequences as gene activators ('positive hits'), thus giving a hit rate of 0.51%. Thus, the full library of 34217 85aa peptides was used as our training data set[3].

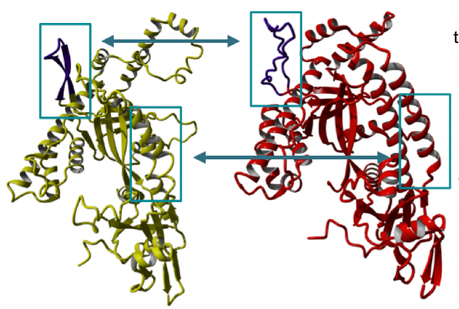


Generating a protein fitness landscape for desired protein function using semi-supervised transfer learning

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)

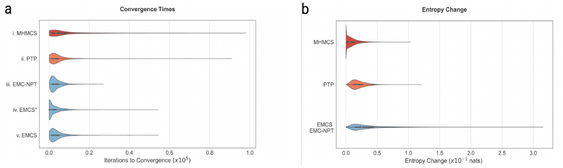


The EMCS algorithm has the following components:

1. Parallel Metropolis-Hastings Monte Carlo (MHMCS) runs.
2. Temperature Ladder implementation (Parallel Tempering).
3. Domain swapping between peptide chains run in parallel (EMCS).

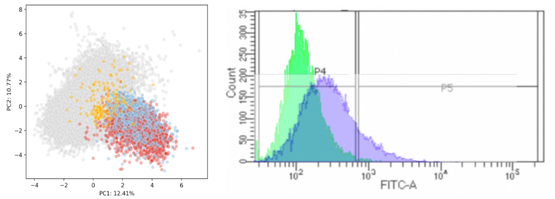
EMCS draws inspiration from genetic swapping events to allow domain swaps to occur between parallel Metropolis-Hasting Monte Carlo runs, thereby enabling more efficient exploration of the fitness landscape.

EMCS outperforms standard Metropolis-Hastings Monte Carlo Sampling (MHMCS)



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. PTP (Parallel Tempering) and EMC-NPT (EMCS without parallel tempering) were run for ablation studies.

4600 sequences with high predicted fitness were experimentally validated

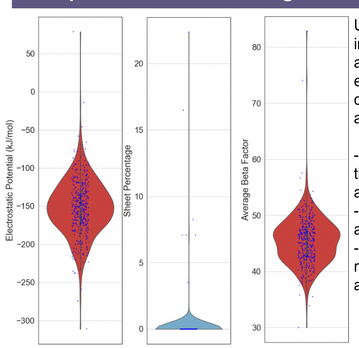


Experimental Results: Sampling Algorithm

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%

For experimental validation, we used EMCS and MHMCS to design 4600 novel sequences that were largely distinct from the sequence space occupied by the training data. We then experimentally assayed the peptides for their ability to activate a synthetic genetic locus. Using a standard differential expression pipeline (DESeq2), we found that 357 of the 4600 sequences (7.51% hit rate) significantly activated the genetic reporter over background florescence. For contrast, our training dataset had a hit rate of 0.51%. Using our best sampling method (EMCS from known sequences), we were able to achieve a hit rate of 20.6%. From top left to bottom, we show the training dataset with the generated sequences in principal component space, representative FACs histograms, and statistics from differential expression analysis.

Biochemical analysis of experimentally validated sequences revealed a range of characteristic features

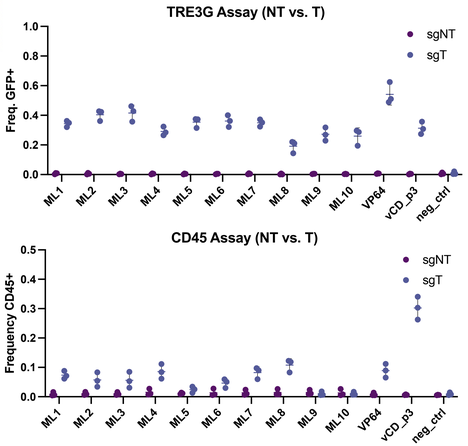


Using ESMFold, we investigated the biochemical and structural properties of experimentally validated structures. In general, we found that activators were:

- Enriched with acidic residues that contributed to the solvent accessible surface area.
- Moderately flexible, having an average beta factor close to 45.
- Lacking beta sheets, and were mostly made up of random coils and alpha helices.

Further validation in HEK293 cells: Activation at synthetic and endogenous loci

We individually screened 10 top activators (chosen via a UMAP cluster based analysis on ESM embeddings) at synthetic and endogenous loci (CD45) and compared their potency to gold standard activators such as VP64 and vCD. To further test the robustness of these activators, we used a different cell type (HEK293).



Summary and Conclusions

1. Using EMCS to sample a protein fitness landscape approximated by a semi-supervised transfer learning (ESM embeddings on XGBoost/CNN) ensemble model, we were able to improve our base hit rate of 0.51% to 20.6%, a 40x improvement.
2. EMCS outperforms MHMCMC by 4xm in final hit rate of experimentally validated molecules.
3. EMCS outperforms standard MHMCMC convergence times by 1.5-5x, depending on choice of parameters.
4. Activation of synthetic and endogenous loci was also observed for a subset of sequences in a different cell line.
5. Despite high sequence dissimilarity, biochemical analysis of the 4600 sequences showed similar trends such as strong negative electrostatic potential, as well as the presence of disordered domains.
6. 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