Sequence Display-Enabled Machine Learning for Protein Evolution

Linqi Cheng [a,§], Xinzhe Zheng [a,§] Shiyu (Jason) Jiang [a,§], Yu Hu [a], Yijie Liu [a], Mengxi Zhang [a], Kevin K. Yang [e], Han Xiao [a, b, c, d*]

[a]Department of Chemistry, [b]Department of Biosciences, [c]Department of Bioengineering, [d]SynthX Center, Rice University

[e]Microsoft Research

Lab website: xiao.rice.edu

Generation of Sequence–Activity Datasets for SlugCas9 Across Diverse PAMs Using Multiplexed-Barcoded Sequence Display

Evolution of SlugCas9 toward expanded PAM recognition using the Sequence Display platform

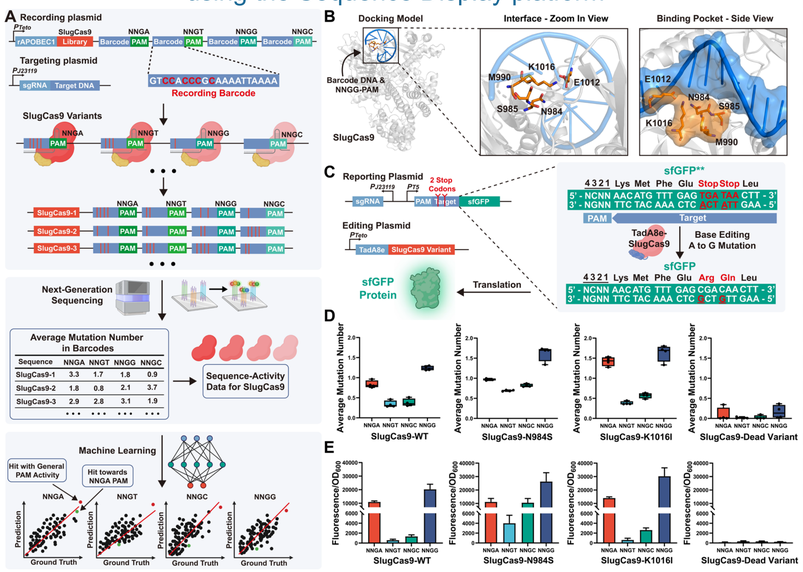

Fig. 2. Evolution of SlugCas9 toward expanded PAM recognition using the Sequence Display platform. **(A)** Pipeline for evolving SlugCas9 variants with broadened PAM recognition using the Sequence Display platform. **(B)** Docking structure of SlugCas9-WT in complex with DNA containing the recording barcode and an NNGG PAM. **(C)** TadA8e fluorescence assay for validating the activity of SlugCas9 variants. **(D)** Sequence Display results for SlugCas9-WT and variants.

Generating Large-scale sequence-activity datasets for SlugCas9 to enable machine learning-based SlugCas9 evolution

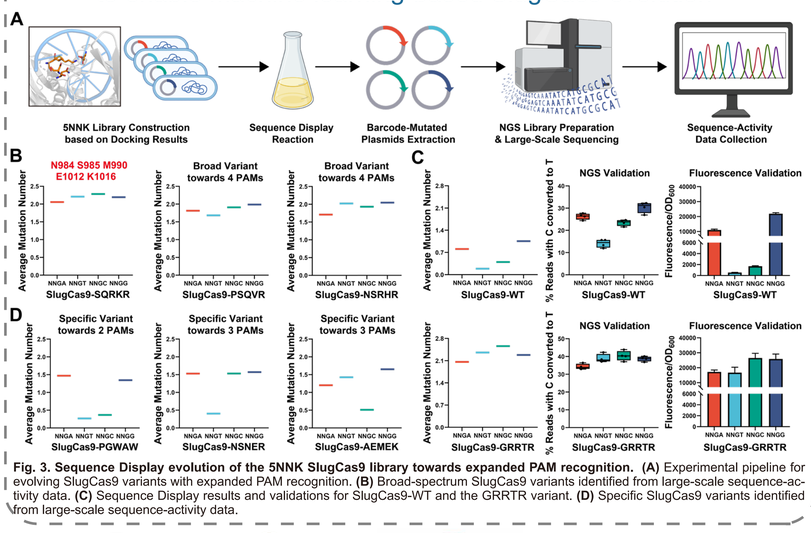

Fig. 3. Sequence Display evolution of the 5NNK SlugCas9 library towards expanded PAM recognition. **(A)** Experimental pipeline for evolving SlugCas9 variants with expanded PAM recognition. **(B)** Broad-spectrum SlugCas9 variants identified from large-scale sequence-activity data. **(C)** Sequence Display results and validations for SlugCas9-WT and the GRRTR variant. **(D)** Specific SlugCas9 variants identified from large-scale sequence-activity data.

Introduction

Sequence Display: Generation of large-scale sequence-activity datasets

Engineering proteins with desired functions remains challenging due to the labor-intensive nature of traditional methods and limited sequence–activity data for machine learning. Here, we present Sequence Display, a scalable platform that rapidly generates large-scale protein sequence–activity datasets, enabling machine learning-driven protein evolution. Sequence Display can be multiplexed to assess the specificity of individual mutants within a single experiment. By integrating these datasets with pre-trained protein language models, we construct fine-grained, variant-specific activity landscapes to identify high-performance variants. We demonstrate its broad applicability by generating datasets for cytosine deaminase, uracil glycosylase inhibitor, and a compact Cas9 nuclease. For Cas9, we produced over 32 million data points and evolved a variant with expanded PAM recognition, which outperformed a previously reported mutant from phage-assisted evolution. This study establishes Sequence Display as a powerful tool for mapping sequence–function relationships and accelerating the discovery of optimized proteins for biological and medical applications.

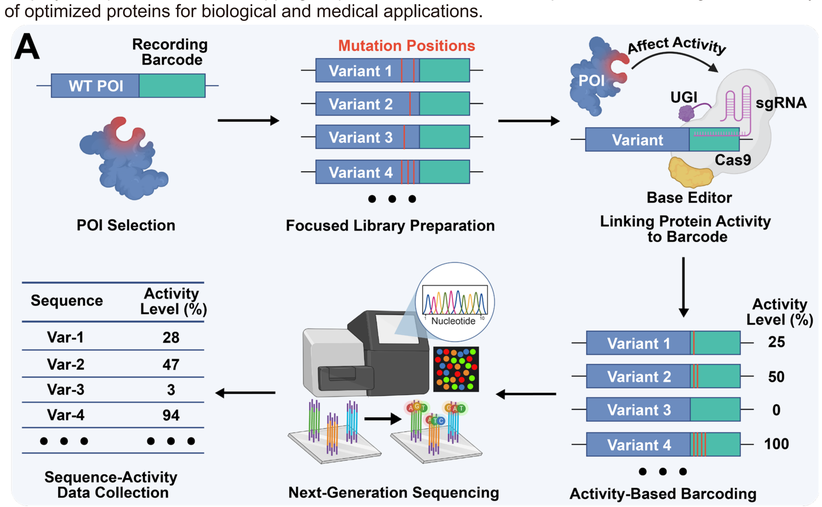

Fig. 1. Schematic overview of Sequence Display-enabled machine learning pipeline. **(A)** Sequence Display platform. **(B)** Machine learning framework for Sequence Display protein evolution. **(C)** Ensemble-based activity inference and landscape construction.

Construction of fine-grained, variant-specific activity landscapes to identify high-performance variants

Modeling the Relationship between SlugCas9 Variants and their Activities using pLMs

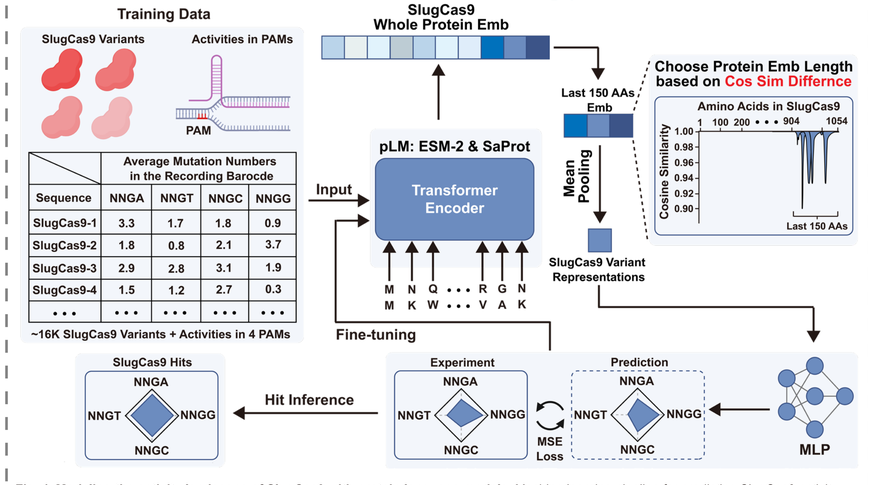

Fig. 4. Modeling the activity landscape of SlugCas9 with protein language models. Machine learning pipeline for predicting SlugCas9 activity. pLMs, including ESM-2 and SaProt, are used to extract local embeddings based on cosine similarity calculations for each amino acid in the SlugCas9 sequence.

Ensemble strategy for Constructing SlugCas9 Activity Landscapes and PAM-Specific Activity Inference

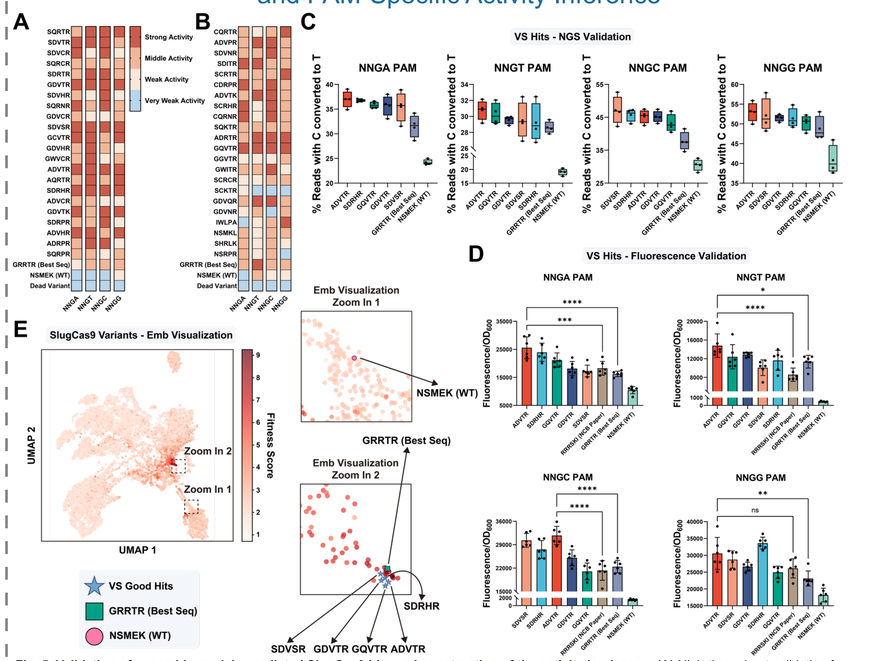

Fig. 5. Validation of ensemble models-predicted SlugCas9 hits and construction of the activity landscape. **(A)** High-throughput validation for top-predicted SlugCas9 variants. **(B)** High-throughput validation of an additional batch of top-predicted variants. **(C)** NGS-based validation of top-performing variants. **(D)** Fluorescence-based validation of top-performing variants. **(E)** UMAP visualization of the SlugCas9 variant activity landscape.

Cancer Prevention & Research Institute of Texas

The Welch Foundation

John S. Dunn Foundation