STEM CELL INNOVATION CONTEST 2026
11TH EDITION | THEME: STEM CELLS & AI
ID: 411
CPD
ORGANIZER
SPONSORING ORGANIZATION

Predicting activator and repressor effector tiles for programmable control of stem-cell fate transcription factors

LE GIA CAT, NGUYEN ANH DUY, NGUYEN THE PHONG, VU NHU NAM PHUONG

COLOUR KEY — filled — what we build or train · teal outline — friend's CNN · activation · gene on · high · repression · gene off · low

ABSTRACT

Directed differentiation of human iPSCs is still done largely by flooding cells with transcription factors. However, a critical unmet need in the field is the development of titratable regulatory platforms capable of the precise, multiplexed, and temporally controlled bidirectional modulation of targeted transcription factors. Currently, the selection of effector domains mainly relies on structural annotations that lack the predictive power to determine actual transcriptional performance.

In this study, we present a pair of sequence models that read an 80-aa tile and predict how strongly it will switch a target gene on or off. The compact transcriptional effector domains, 80 amino acids long, are small enough to multiplex and fit the vectors that deliver them. Both models learn from high-throughput recruitment screens spanning over 100,000 tiles from 2,047 human transcription factors and chromatin regulators, and both are tested on held-out MMseqs2 homology groups, thus avoid close relative sequences in training.

Regarding activation tiles, 81% of the top five picks are genuinely active against 6% by chance, which are sorted into weak, medium and strong. Meanwhile, 45 of 50 top-ranked tiles are validated repressors of an endogenous gene, against 2 of 50 by chance. Domains that do both are flagged and filtered out. This in silico prediction will be further validated by investigating the function of candidate effector tiles using High-Throughput recruitment (HT-Recruit).

Our framework helps classify unrecognized tiles as potent activators or repressors but also predicts their context-specific performance at endogenous TF loci critical for stem cell maintenance and directed differentiation. By enabling the in silico screening of effector tiles, this computational pipeline hopes to reduce the exhaustive workload of empirical library testing, providing a scalable strategy to empower precise epigenetic reprogramming and advanced cellular therapies.

INTRODUCTION

- Transcription factors (TFs) set cell fate in human iPSCs, yet many TF pathways remain understudied.
- A compact **effector tile** carried to a TF locus by a DNA-binding domain switches that gene on or off.
- Potency is known only from exhaustive screening — sequence exists for every tile, measured activity does not.

Fig. 1. Effector-tile recruitment at a TF locus

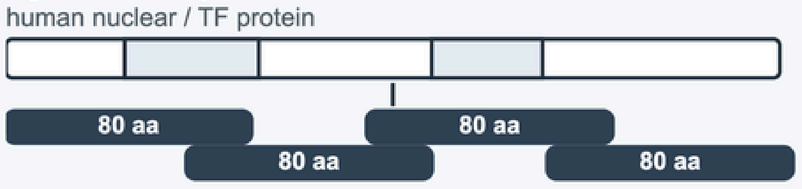


OBJECTIVES

1. Predict activation and repression potential of 80-aa tiles from sequence alone.
2. Beat composition-only baselines against a composition-matched negative control.
3. Test whether predictions hold when the DNA-binding domain changes.
4. Rank tiles at stem-cell fate TF loci for HT-Recruit validation.

METHOD

Fig. 2 – Tile library



Overlapping tiles from published HT-Recruit screens (Tycko et al., 2025; DelRosso et al., 2023) — 103,248 scored for repression, 77,915 of them also for activation.

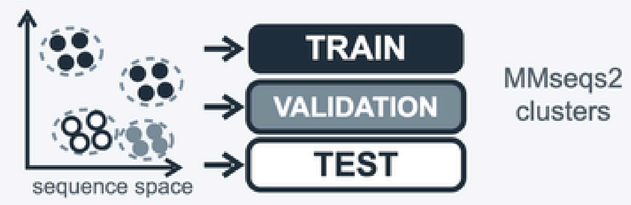
Fig. 3 – Target panel

↑ ACTIVATION SET: FGFR1 · ACVR1B · TGFBR1 · FZD7 · LRP5 · LRP6

↓ REPRESSION SET: BMPR1A · NOTCH1 · SMAD3 · RARA · RXRA

Stem-cell fate loci to be driven — FGF, Activin / TGF-β, Wnt, BMP, Notch and retinoic-acid pathways.

Fig. 4 – Split by cluster



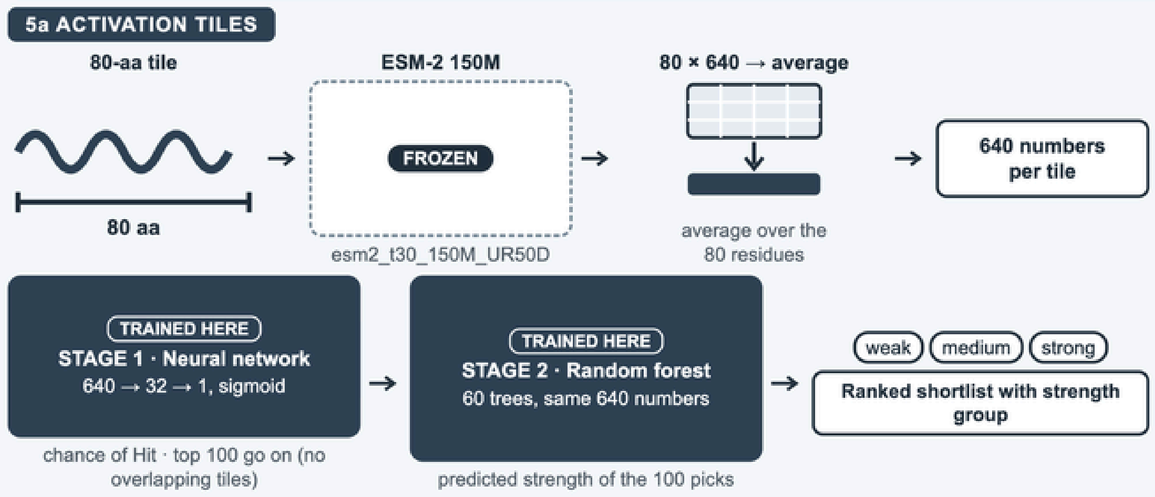
A whole cluster goes to train, validation or test.

CONTROLS — **Baseline:** rTetR carrying no tile · **Negative:** random 80-aa sequences · **Positive:** Oct4 for activation, ZNF10 KRAB for repression

Fig. 5 – Model architecture: two models, same tile library

Activation screen (K562, minCMV): 77,915 tiles · 819 Hits · ESM-2 150M · 10 splits

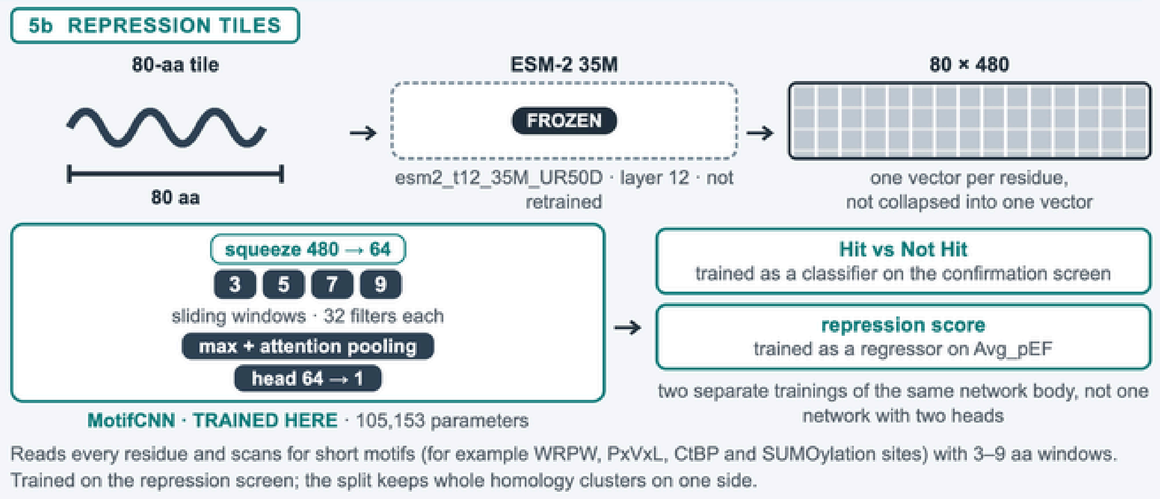
5a ACTIVATION TILES



One averaged ESM vector per tile. Stage 1 (20,609 parameters) picks the top 100 non-overlapping tiles; Stage 2 sorts them by strength. Composition baseline (20 amino-acid fractions) on the same splits.

Repression screen (K562, pEF): 103,248 tiles · ESM-2 35M · 70/15/15 split by cluster

5b REPRESSION TILES



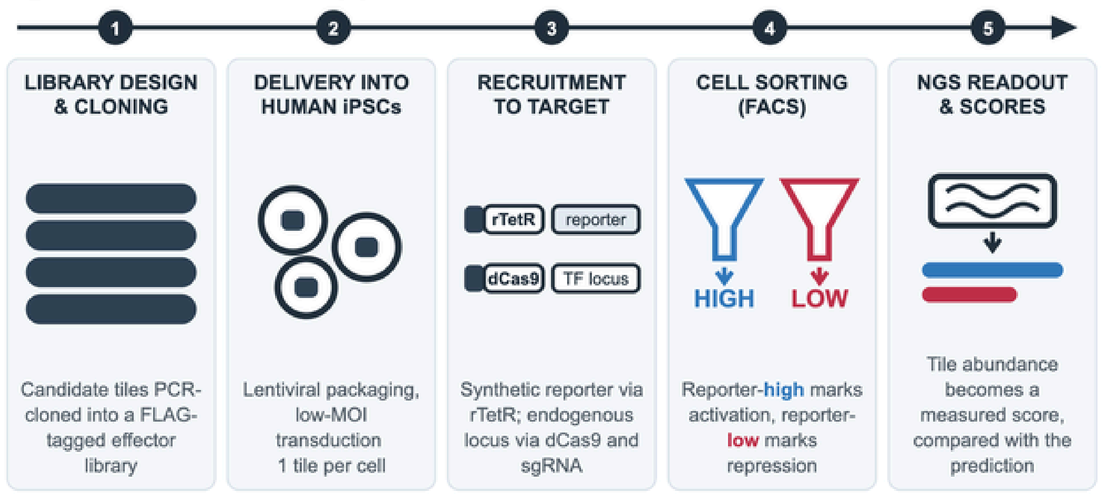
Reads every residue and scans for short motifs (for example WRPW, PxVxL, CtBP and SUMOylation sites) with 3–9 aa windows. Trained on the repression screen; the split keeps whole homology clusters on one side.

Benchmarked against: ridge on 20 amino-acid frequencies, ridge on 400 dipeptide frequencies, mean-pool MLP and attention-pool MLP — same split, same seeds.

5a: scored on 10 held-out splits, about 12,000 test tiles and about 120 Hits each. 5b: scored on held-out homology clusters of the repression screen.

VALIDATION USING HT-RECRUIT

Fig. 6 – HT-Recruit validation pipeline



EXPECTED RESULTS

Cards 1–2: mean ± sd over 10 held-out splits · Card 3: repression screen, own split

STAGE 1 · Logistic regression finds Hits	STAGE 2 · Random forest sorts Hits by strength	CNN · repression own screen, labels and split
ROC-AUC 0.951 ± 0.041	Rank correlation ρ 0.462 ± 0.069	ROC-AUC 0.813 ± 0.016
AUPRC 0.654 ± 0.095	3-group Accuracy 0.475 ± 0.060	AUPRC 0.290 ± 0.031

Our two-stage model ranks held-out activation tiles well (AUPRC 0.654 against 0.022 for random) and orders Hit strength moderately (rank correlation 0.46). The friend's CNN is applied to the repression screen, with its own labels and split, so its results answer a separate question and are not compared head to head with ours.

REFERENCES

1. DelRosso N, Tycko J, Suzuki P, et al. Large-scale mapping and mutagenesis of human transcriptional effector domains. Nature 616, 365–372 (2023).
2. Tycko J, DelRosso N, Hess GT, et al. Development of compact transcriptional effectors using high-throughput measurements in diverse contexts. Nature Biotechnology 43, 1525–1538 (2025).
3. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130 (2023).