Scientific value and novelty of OncoSwitch

OncoSwitch is an engineering cycle of "design → MPRA → model → redesign" that purposefully finds and collects short DNA regulators that are activated in tumors and silent in normal cells. We use pairs of "tumor vs. normal" cells as a biological contrast enhancer: in tumors, key pathways (e.g., HIF, TEAD/YAP, TCF/LEF, ER/AR, AP-1, etc.) are active, while in normal cells they are silenced. We input pools of short inserts (starting at 50 bp), measure reporter activity in both lines, taking into account RNA/DNA normalization and barcodes (MPRA/STARR-seq paradigm) [1], train the model on the obtained data, and draw the next batch of sequences. We repeat this 10 cycles until the selectivity gain plateaus [2].

How the idea of random start and ~4000 sequences per cycle works.

In a 50-bit insertion, the number of windows of length k is equal to (50−k+1). For k=6, this is 45 windows; for k = 8, it is 43. For 4000 insertions, we get:

- k=6: 4,000×45 = 180,000 windows per chain; expectation of coincidences with a given 6-dimensionality: $180{,}000/4^6 \approx 43.9$ (on both chains ≈ 88). The probability of at least one hit in one insertion ≈ $45/4^6 \approx 1.1\%$; that is, ≈ 44 insertions out of 4,000 contain a specific 6-dimensional on one chain.
- k=8: 4,000×43 = 172,000 windows; expectation for an exact 8-mer: $172{,}000/4^8 \approx 2.6$ (both chains ≈ 5.2).

Real factors recognize not just one "hard" k-mer, but degenerate motifs and flanks, so the real saturation is higher; MPRA models do indeed extract such "grammar" [5], and reviews emphasize the importance of context and motif syntax [2].

How the cycle turns chance into engineering.

Each round consists of:

1. **Design** – a mixture of pure randomness, motif-biased randomness (we embed HIF/TEAD/TCF cores, etc.) and "micro-logic" in 50 bp (simple AND/NOT at intervals of 5–12 bp);
2. **MPRA** – plasmid mode, RNA/DNA normalization by barcodes, ≥3 bioreplicas per "tumor" and "norm" [1];

3. **Model** – convolutional-attentive architecture for 50–200 p.n.: filters converge to motifs, the next level captures distances/order ("regulatory grammar") – examples of DL approaches to activity prediction and motif interpretation see [5], for an overview, see [2];
4. **Redesign** – exploitation (point corrections of leading sequences), crossovers (gluing together successful fragments), research (areas of high model uncertainty), and movement towards 200 p.n. with explicit AND/OR/NOT logic (2–3 "tiles" + linkers, repressor at the promoter, miRNA-NOT in 3′UTR); examples of targeted DL design of synthetic promoters – [3], as well as a thematic review of scalable design strategies – [4]

Why AND/OR/NOT logic?

OR provides coverage of subclones: if any of the tumor pathways are activated, there is a signal. AND sets the threshold: two similar motifs are silent separately, but together they produce a surge. NOT insures the background: repressor sites at the promoter and/or miRNA targets in 3′UTR extinguish activity in normal conditions. Within ≤200 p.n., this provides the necessary "on/off" function: high in tumors, low in normal conditions [2.5].

Why our design is rational.

With a 6-dimensional cut alone, we collect ~180,000 local observations per cycle; over 10 cycles – ~1.8 million (excluding 8-dimensional, flanks, and multi-position edits). This volume is sufficient for the model to learn the grammar for a specific "tumor-normal" pair [2.5]. Active learning saves the library: instead of 30-50 thousand variants per round, we spend 4,000 on the maximum increase in information and Δ(tumor/normal). Expected: by the 2nd–3rd cycle – lead selectivity ≥10× (MVP), by the 5th–7th – 15–20×; then comes polishing and stability testing; examples of successful DL-assisted design – [3,4].

Validation and thresholds.

Start – plasmid MPRA with RNA/DNA normalization and tech/bioreplicas [1].

MVP goal: by month 8, obtain at least one switch with ≥10× tumor:normal selectivity and reproducible results. Then continue iterations: point corrections of leading sequences, assembly of AND/OR/NOT logic to ≤200 bp, addition of "double NOT" (repressor at the promoter + miRNA targets in 3′UTR), stress tests for leaks.

After pooling, we perform clonal verification of leads to prevent leaks in the norm. For guidance on modern approaches and design scaling, see [2,4].

Risks and control.

- Epistemic artifacts: limit the expression window (24–48 hours), confirm effects in an integrative and clonal context [2,4].
- Retraining: separate splits by constructs/barcodes, regularization, ensembles, active learning on uncertainty (overview guidelines – [2]).
- Leaks are normal: "double NOT", negative controls, and "red list" of sequences; fundamental principles and methods for interpreting effects – [2,5].Научная новизна.

We do not stop at the standard "measure → train model" approach, but close the loop until we obtain target regulators with explicit AND/OR/NOT logic and transparent interpretation ("DNA instead of weights": model filters = motifs, importance maps, and mutation scans provide readable rules). Compared to classic MPRA/STARR-seq and predictive DL works, the main result here is a reproducible process for producing selective regulators "for the task," not just another promoter found.

Final value.

OncoSwitch transforms regulator design from a manual search into a controlled, measurable procedure. Each measurement makes the model smarter and the next design cleaner.

References:

1. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013 Mar 1;339(6123):1074-7. doi: 10.1126/science.1232542. Epub 2013 Jan 17. PMID: 23328393.
2. La Fleur A, Shi Y, Seelig G. Decoding biology with massively parallel reporter assays and machine learning. Genes Dev. 2024 Oct 16;38(17-20):843-865. doi: 10.1101/gad.351800.124. PMID: 39362779; PMCID: PMC11535156.
3. Fu ZH, He SZ, Wu Y, Zhao GR. Design and deep learning of synthetic B-cell-specific promoters. Nucleic Acids Res. 2023 Nov 27;51(21):11967-11979. doi: 10.1093/nar/gkad930. PMID: 37889080; PMCID: PMC10681721.
4. Gosai SJ, Castro RI, Fuentes N, Butts JC, Kales S, Noche RR, Mouri K, Sabeti PC, Reilly SK, Tewhey R. Machine-guided design of synthetic cell type-specific cis-regulatory elements. bioRxiv [Preprint]. 2023 Aug 9:2023.08.08.552077. doi:

10.1101/2023.08.08.552077. Update in: Nature. 2024 Oct;634(8036):1211-1220. doi: 10.1038/s41586-024-08070-z. PMID: 37609287; PMCID: PMC10441439.

5. de Almeida BP, Reiter F, Pagani M, Stark A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. Nat Genet. 2022 May;54(5):613-624. doi: 10.1038/s41588-022-01048-5. Epub 2022 May 12. PMID: 35551305.