# A deep-learning predictor of RNA-binding-protein footprints for antisense-oligonucleotide screen design

**BajaCLIP Predict™** reads RNA sequence and predicts, base-region by base-region, where each of **170 RNA-binding proteins (RBPs)** attaches to a transcript. Its purpose is to localize the regulatory footprints.

| **170** | **48** | **0.96** | **112×** |
|---|---|---|---|
| RBPs predicted per scan (168 ENCODE eCLIP + 2 POSTAR3) | reliable RBPs (held-out AUROC ≥ 0.90) | TDP-43 accuracy (AUROC, chromosome-held-out) | smaller ASO library when the map focuses the screen |

## HOW IT WORKS

A convolutional neural network tiles a **64-nt window** along any pre-mRNA and returns a per-RBP binding probability for each position. It was retrained on the complete ENCODE crosslinking compendium (6.9 million binding sites) with calibrated backgrounds, and evaluated on entire held-out chromosomes to exclude leakage.

## PRIMARY APPLICATION — COST-EFFECTIVE ASO SCREENS

### ~5,500 → ~40 ASOs

Effective steric-blocking ASOs are found by empirically tiling overlapping 20-mers across a target. Tiling all of `STMN2` (55 kb) needs ~5,500 oligos; focusing the walk on the model's ~500-bp predicted hot-spot needs ~40 — a plate-sized, affordable screen.

The model **sizes and places the screen**; the winning ASO is still found in the assay. Its region-level accuracy is matched to this region-level decision.

## VALIDATION

- Recovers TDP-43's canonical **UG-rich motif** from sequence alone.
- **STMN2:** 71% of predicted TDP-43 sites fall in the experimentally defined intron-1 element (6.8× enrichment) — blind to those coordinates.
- Across 12 published TDP-43 cryptic-exon genes, focal enrichment is elevated vs housekeeping controls (median 12× vs 4×).

## INPUTS & OUTPUTS

- **In:** a gene symbol or RNA/pre-mRNA sequence (GRCh38).
- **Out:** per-RBP hit maps ( `TSV` / `BED` ), footprint-cluster coordinates, per-isoform hit-map figures, and single-base sensitivity (saturation) tracks.

## RELIABLE RBP COVERAGE (AUROC ≥ 0.90)

TARDBP, FUS, EWSR1, MATR3-class hnRNPs (HNRNPM), PUM1, FMR1, FXR2, KHDRBS1, SRSF1, TRA2A, U2AF1, PCBP1, ELAC2, DDX3X, DDX1/24/43/51/52, DROSHA, RBM15, NONO, PPIG, SF3A3, EIF4E, EIF4G2, EIF3H, XRN2, FTO, and others — **48 of 137 evaluated channels**. Full list: `reliable_rbps_enhanced_model.tsv` .

## INTENDED USE & LIMITS

- A **discovery-stage prioritization tool**, not a diagnostic or clinical device.
- Trained on **non-neuronal** (K562/HepG2) data; predicts sequence-intrinsic binding — confirm in a neuronal context (iPSC-neurons, minigene, targeted RNA-seq).
- Reliable at the **region** level; it sizes screens, it does not call individual variants or pick the single active ASO.
- Best on its 48 reliable RBPs; weak/ubiquitous channels are excluded.

**BajaCLIP Predict™** — research-use draft model. Not validated for clinical or diagnostic use; predictions require experimental confirmation. "™" denotes an unregistered mark.