

# CRISPR On-Target Efficiency

A CNN + attention model for sgRNA on-target editing efficiency, benchmarked against a reproduced Rule Set 2 gradient-boosted-tree baseline, with an attention interpretability layer and split conformal prediction intervals.

0.515

Rule Set 2 baseline, LOGO Spearman

0.507

CNN + attention, LOGO Spearman

0.495

CNN, gene-held-out test

90.2%

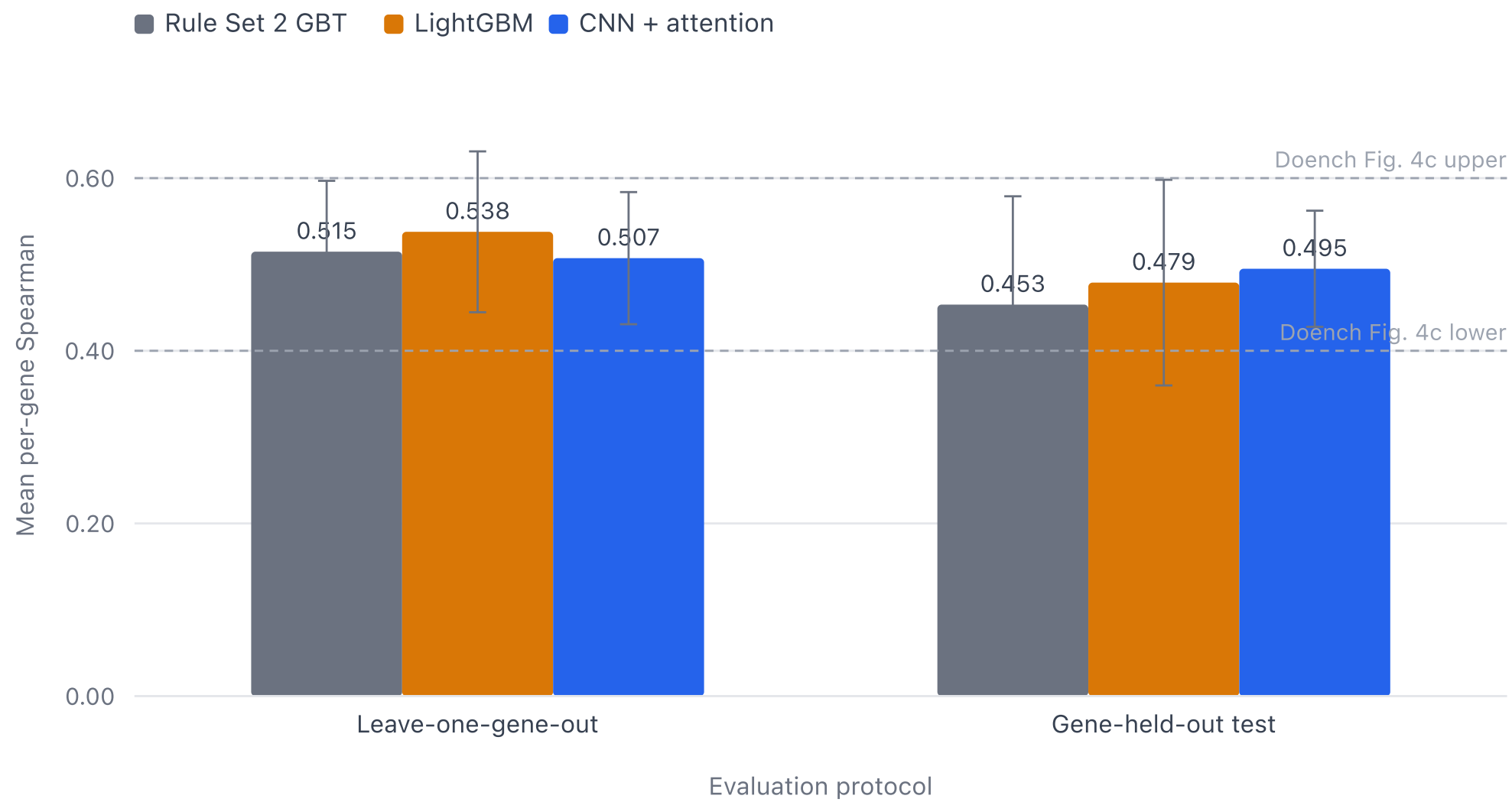
Coverage at 90% nominal

39/39

Validation checks passed

## Does the CNN beat the gradient-boosted-tree baseline?

**No. It ties under the paper's protocol and wins on the frozen test split, and neither result beats a tuned LightGBM on the same hand-crafted features.** On roughly 4,400 guides, a learned sequence representation buys nothing over well-chosen k-mer features.



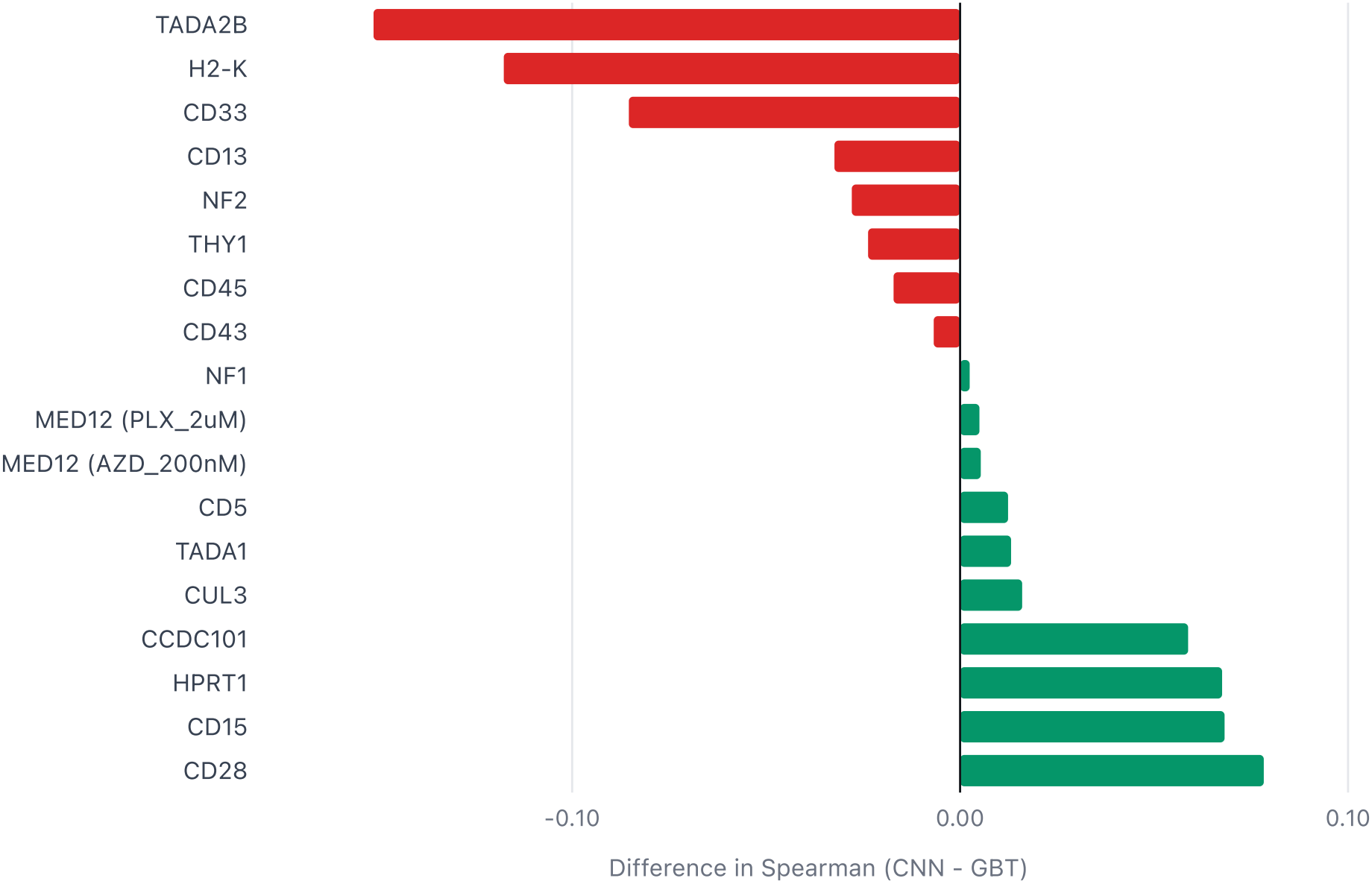
Mean Spearman across the 18 (gene, drug) evaluation groups; error bars are +/-1 SD across those groups. That spread (~0.08) is larger than every gap between models, which is the whole story. Dashed lines mark the 0.4-0.6 band Doench 2016 reports graphically for boosted regression trees in Fig. 4c.

### Are the differences real?

| COMPARISON                         | MEAN DIFF | 95% CI          | WILCOXON P | VERDICT            |
|------------------------------------|-----------|-----------------|------------|--------------------|
| ● CNN + attention - Rule Set 2 GBT | -0.0075   | [-0.036, 0.019] | 0.766      | Tie, CI spans zero |
| ● CNN + attention - LightGBM       | -0.0305   | [-0.063, 0.004] | 0.067      | Tie, CI spans zero |
| ● LightGBM - Rule Set 2 GBT        | +0.0230   | [-0.007, 0.047] | 0.006      | Tie, CI spans zero |

Paired across the 18 evaluation groups, with a 10,000-sample bootstrap on the mean difference. For LightGBM vs Rule Set 2 the two tests disagree: the sign is consistent across genes ( $p = 0.006$ ) while the CI on the mean spans zero, which means reliably a little better rather than decisively better.

### Per-gene difference, CNN minus Rule Set 2 GBT



Out-of-fold leave-one-gene-out predictions. The CNN wins on 10 of 18 groups and loses on 8. Mixed signs are what a statistical tie looks like.

Score a guide

Runs the **gene-held-out ensemble**, the same three-seed model behind the 0.495 test number rather than an in-sample refit, entirely in your browser, then wraps it in a split conformal interval calibrated on held-out genes.

Start from

Strong guide (measured top 5%)

30nt context sequence

CACTGCAGCCAATCCTGTGACAGACGGACC

4nt upstream + 20nt protospacer + 3nt NGG PAM + 3nt downstream.

Interval confidence

80%

90%

95%

☒ Use median gene position

The model consumes three gene-position features a bare 30mer cannot supply.

0.628

0.21 - 1.00

+/-0.0198

Predicted efficiency  
rank within a (gene, drug) group, 0-1

90% conformal interval  
+/-0.419

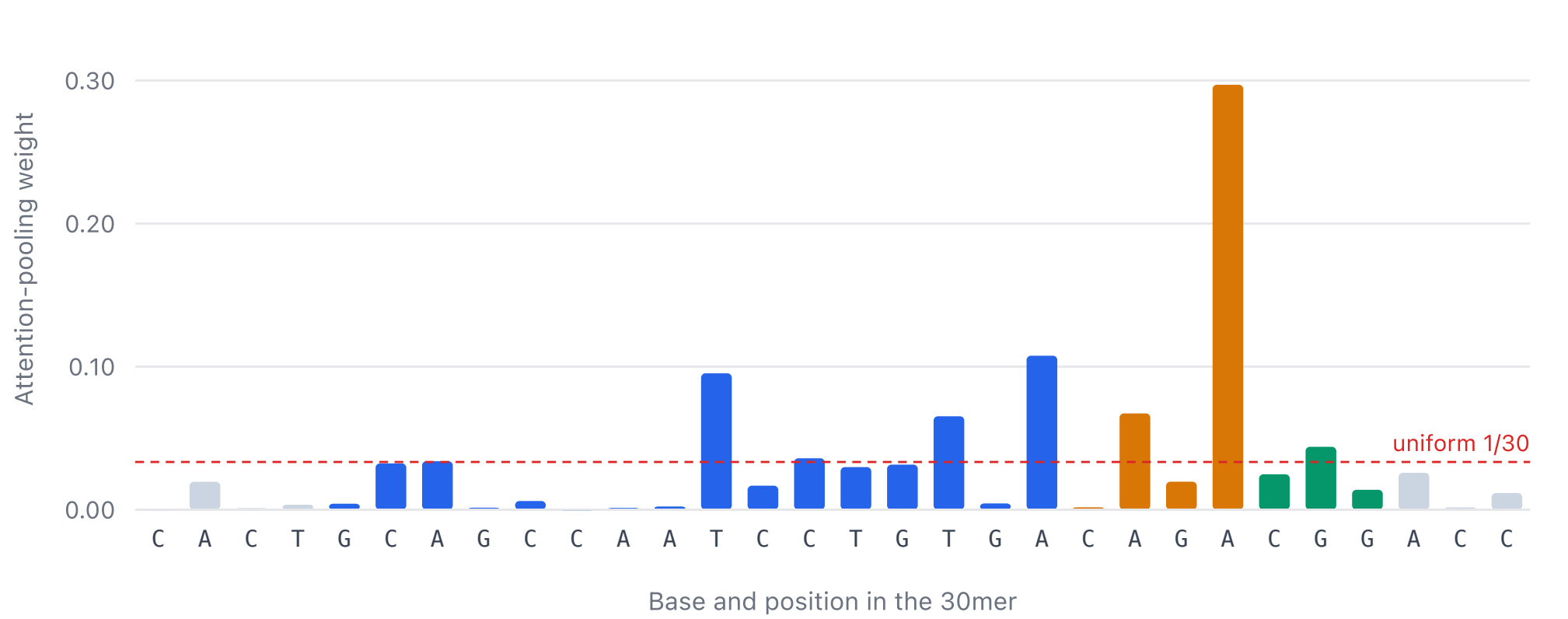
Ensemble disagreement  
spread across 3 seeds

● In the dataset: **CD43** (nodrug), measured activity **1.000**, ranked **1 of 145** in its group. It was in the **train** split.

**This interval is wide**  
It spans 0.79 of a (0, 1] scale. The intervals are calibrated, but wide enough to be useful mainly for flagging predictions not to trust rather than for ranking individual guides.

GC in protospacer 11/20 | Tm(30mer) 70.3 degC | per-seed 0.649, 0.601, 0.634

Where the model looked



Computed live in your browser for this exact sequence, averaged over the three ensemble seeds. Orange marks the seed region (protospacer 17-20), green the PAM, blue the rest of the protospacer, grey the flanking context.

Guide explorer

Every guide in the dataset with its measured activity and all three models' **out-of-fold** leave-one-gene-out predictions, each made by a model that never saw that guide's gene during training.

Gene

All genes

▼

Gene-held-out split

All splits

▼

Sequence contains

e.g. GGGTTT

5,310 of 5,310 guides

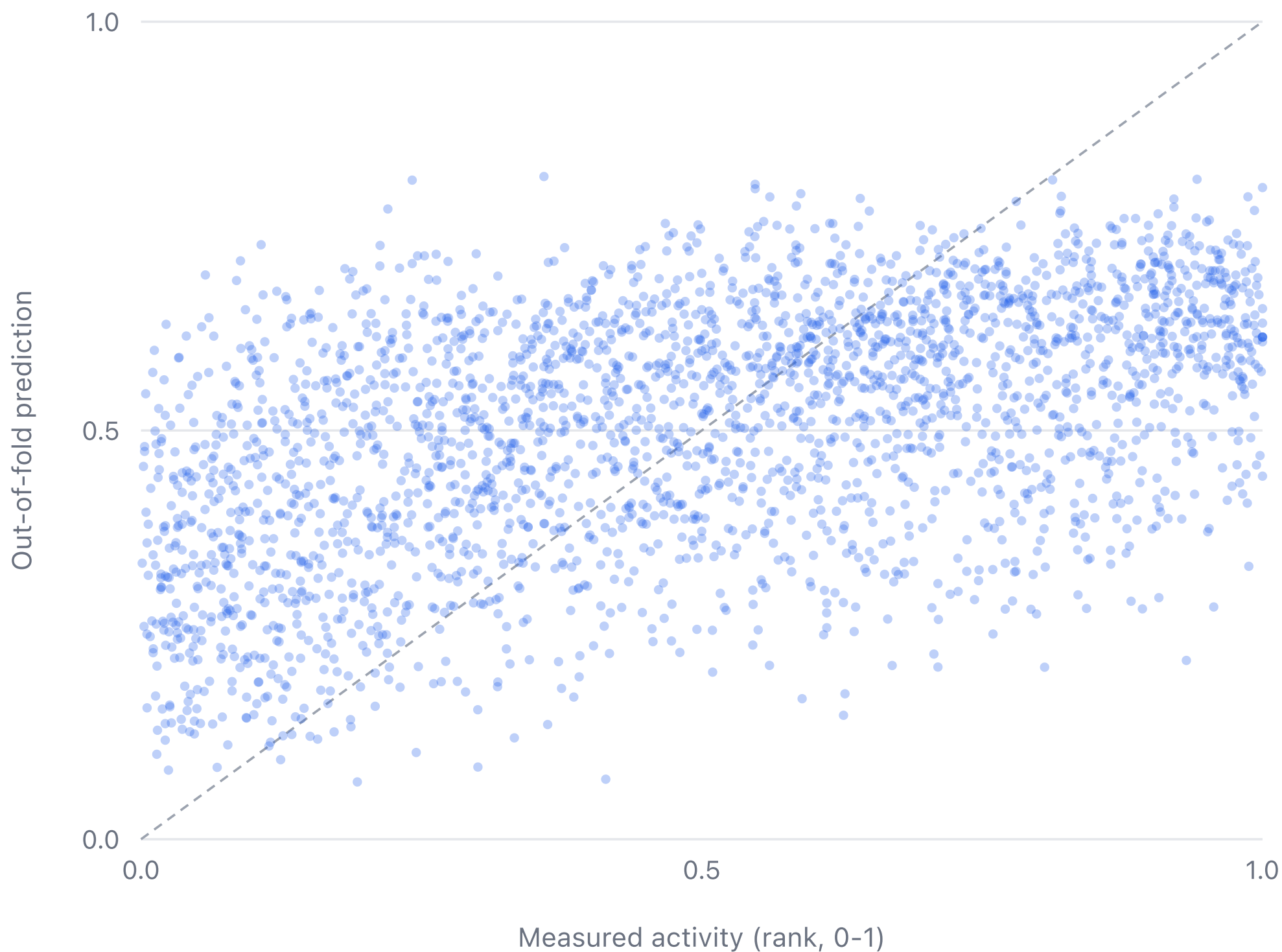
| 30MER                          | GENE  | DRUG      | MEASURED | RULE SET 2 GBT | LIGHTGBM | CNN   | SPLIT | RANK      |
|--------------------------------|-------|-----------|----------|----------------|----------|-------|-------|-----------|
| CGGCGATTATGCCCTGAACGTGACGGGCCC | CD13  | nodrug    | 0.712    | 0.583          | 0.595    | 0.712 | cal   | 118 / 455 |
| GCAGCTCACACCTTCACCTACACGGGGCTA | MED12 | AZD_200nM | 0.732    | 0.691          | 0.724    | 0.732 | train | 249 / 924 |
| TGTAGCCATTTAAAGAACAAATTTGGGGAG | MED12 | PLX_2uM   | 0.127    | 0.281          | 0.296    | 0.127 | train | 808 / 924 |
| TTCTTTGCTTCTGCACTTGGCTTGCGGATC | NF1   | PLX_2uM   | 0.114    | 0.145          | 0.060    | 0.114 | train | 653 / 736 |
| TACTGCTTGGTACGCAGAGCACCGAGGCCG | NF2   | PLX_2uM   | 0.780    | 0.786          | 0.899    | 0.780 | train | 50 / 223  |
| TCCAGGAAACAGGGGCTGGTGGCATGGGTG | CD15  | nodrug    | 0.381    | 0.413          | 0.276    | 0.381 | test  | 170 / 273 |
| GGGTCTGCTGAACCTCAATGTGACGGGCT  | CD13  | nodrug    | 0.675    | 0.573          | 0.625    | 0.675 | cal   | 147 / 455 |
| GTGGAAGTAGCGGCGATAGACCGCGGGGTT | CD15  | nodrug    | 0.564    | 0.408          | 0.492    | 0.564 | test  | 120 / 273 |
| CTGCTGCCACCTTGCTTTAGAAGAGGACC  | NF1   | PLX_2uM   | 0.480    | 0.567          | 0.525    | 0.480 | train | 384 / 736 |
| TCAACTGGACGCTCTCCTACCGGGCGGACT | CD15  | nodrug    | 0.656    | 0.746          | 0.895    | 0.655 | test  | 95 / 273  |
| GTCAGCAGAGACAGGGTCATCTTCTGGAAG | MED12 | PLX_2uM   | 0.282    | 0.416          | 0.396    | 0.282 | train | 664 / 924 |

Measured versus predicted

CNN

Rule Set 2 GBT

LightGBM

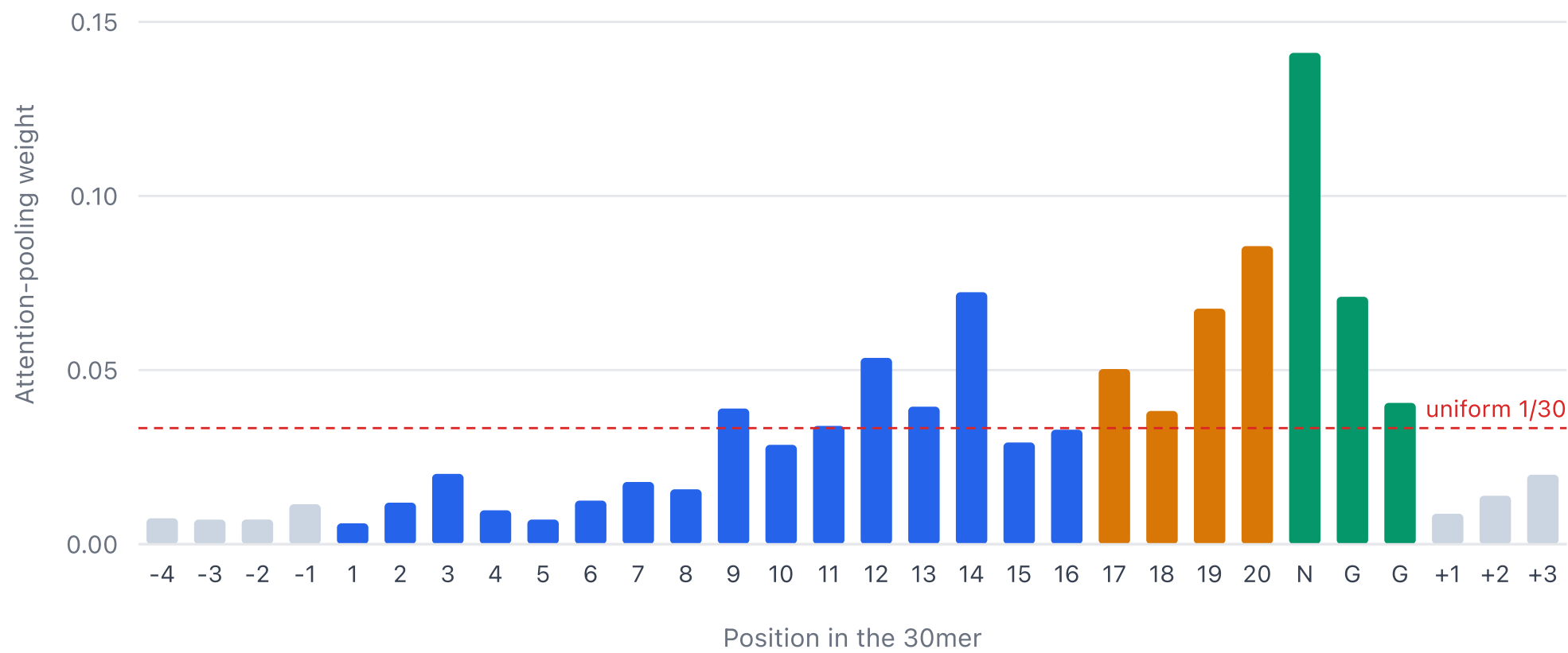


CNN on this selection: Spearman 0.487 over 5,310 guides (2,655 plotted). Dashed line is  $y = x$ . Predictions compress toward the middle because the models are fit with squared error on a rank target.

What the model learned to look at

CNN attention

GBT importance (independent control)



Mean attention-pooling weight over 1,066 held-out guides. Orange is the seed (protospacer 17-20), green the PAM, blue the rest of the protospacer, grey the flanking context. Weights sum to 1 across positions.

Weight by region, CNN attention

| REGION                                | MEAN WEIGHT | VS UNIFORM |
|---------------------------------------|-------------|------------|
| <div></div> PAM (NGG)                 | 0.0842      | 2.53x      |
| <div></div> Seed, protospacer 17-20   | 0.0604      | 1.81x      |
| <div></div> Rest of protospacer, 1-16 | 0.0269      | 0.81x      |
| <div></div> 3' context (+1 to +3)     | 0.0142      | 0.43x      |
| <div></div> 5' context (-4 to -1)     | 0.0083      | 0.25x      |

Seed versus the rest of the protospacer, paired across guides: Wilcoxon p ~ 3e-123.

The seed-region prior holds

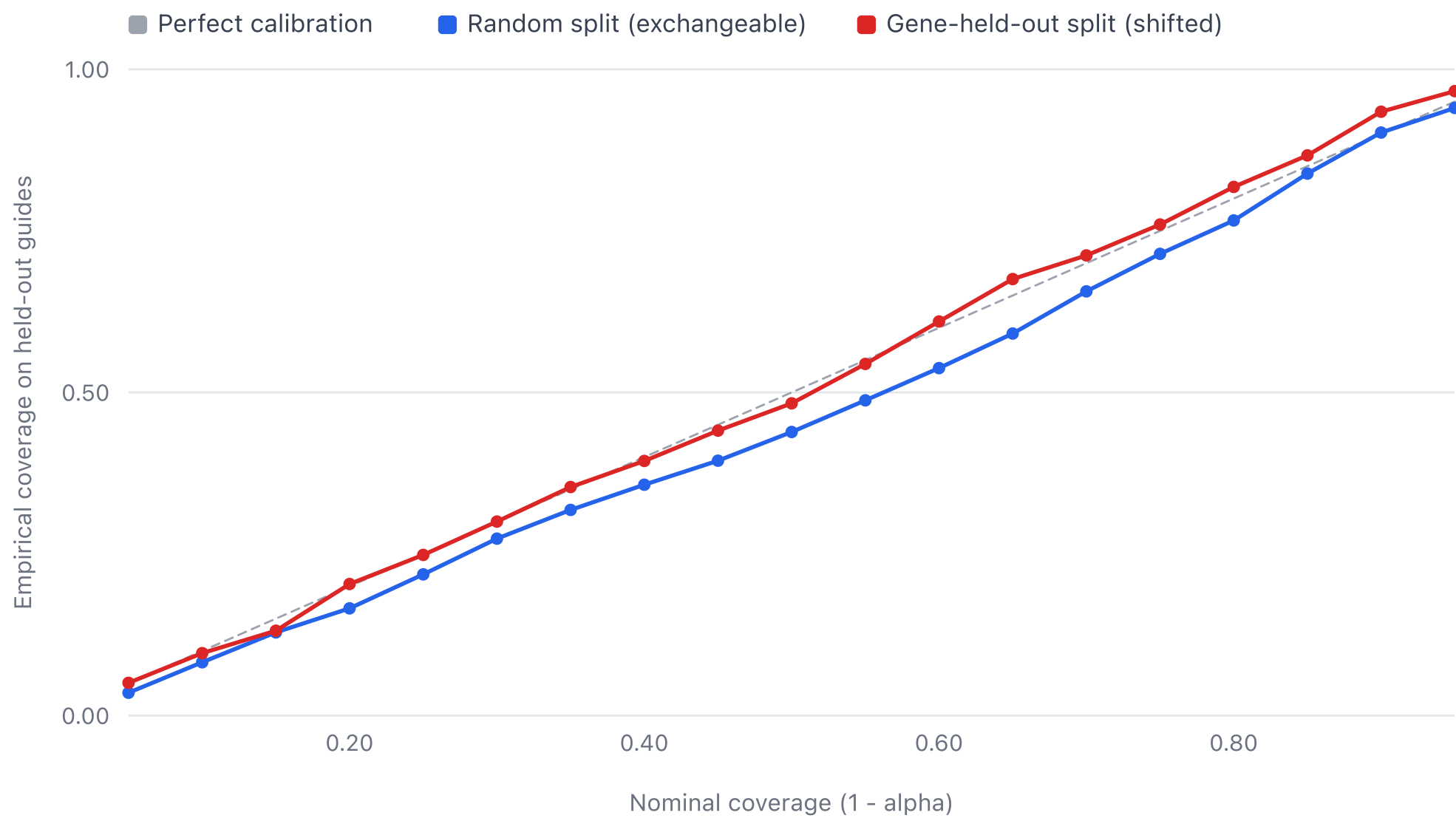
Attention rises along the protospacer toward the PAM and peaks at the PAM's variable N base, which is consistent with Cas9 requiring PAM recognition before it interrogates the protospacer, and with seed-proximal mismatches being the most disruptive.

Independent control agrees

Attention weights are easy to over-read, so the same question went to the GBT, which needs no attention mechanism to be interpretable. It places **54.8%** of its positional importance in the four seed positions, against 13.3% if it were flat. Two methods with nothing in common agree on where the signal is.

## Do the uncertainty intervals mean anything?

**Yes on exchangeable data, and conservatively on new genes.** Split conformal with absolute-residual nonconformity scores and the finite-sample  $\text{ceil}((n+1)(1-\alpha))$  quantile correction.

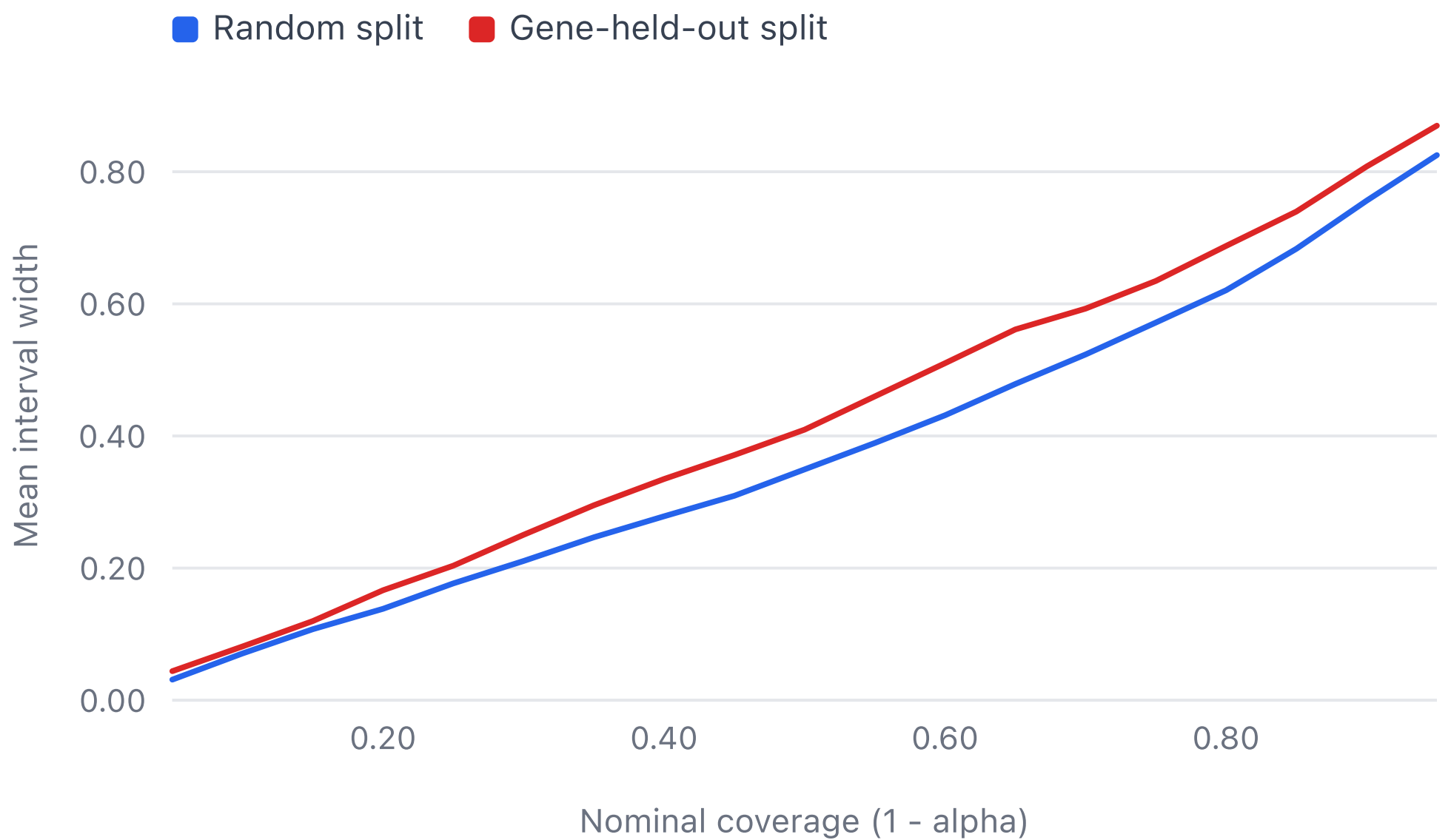


Above the dashed diagonal is conservative (intervals wider than needed); below is under-covering. Source: CNN + attention with 1,016 / 1,059 calibration guides.

### Coverage and width at key levels

| NOMINAL | RANDOM | GENE-HELD | HALF-WIDTH |
|---------|--------|-----------|------------|
| 95%     | 94.0%  | 96.6%     | +/-0.413   |
| 90%     | 90.2%  | 93.4%     | +/-0.378   |
| 80%     | 76.6%  | 81.8%     | +/-0.310   |
| 50%     | 43.9%  | 48.3%     | +/-0.175   |

### Interval width grows with confidence



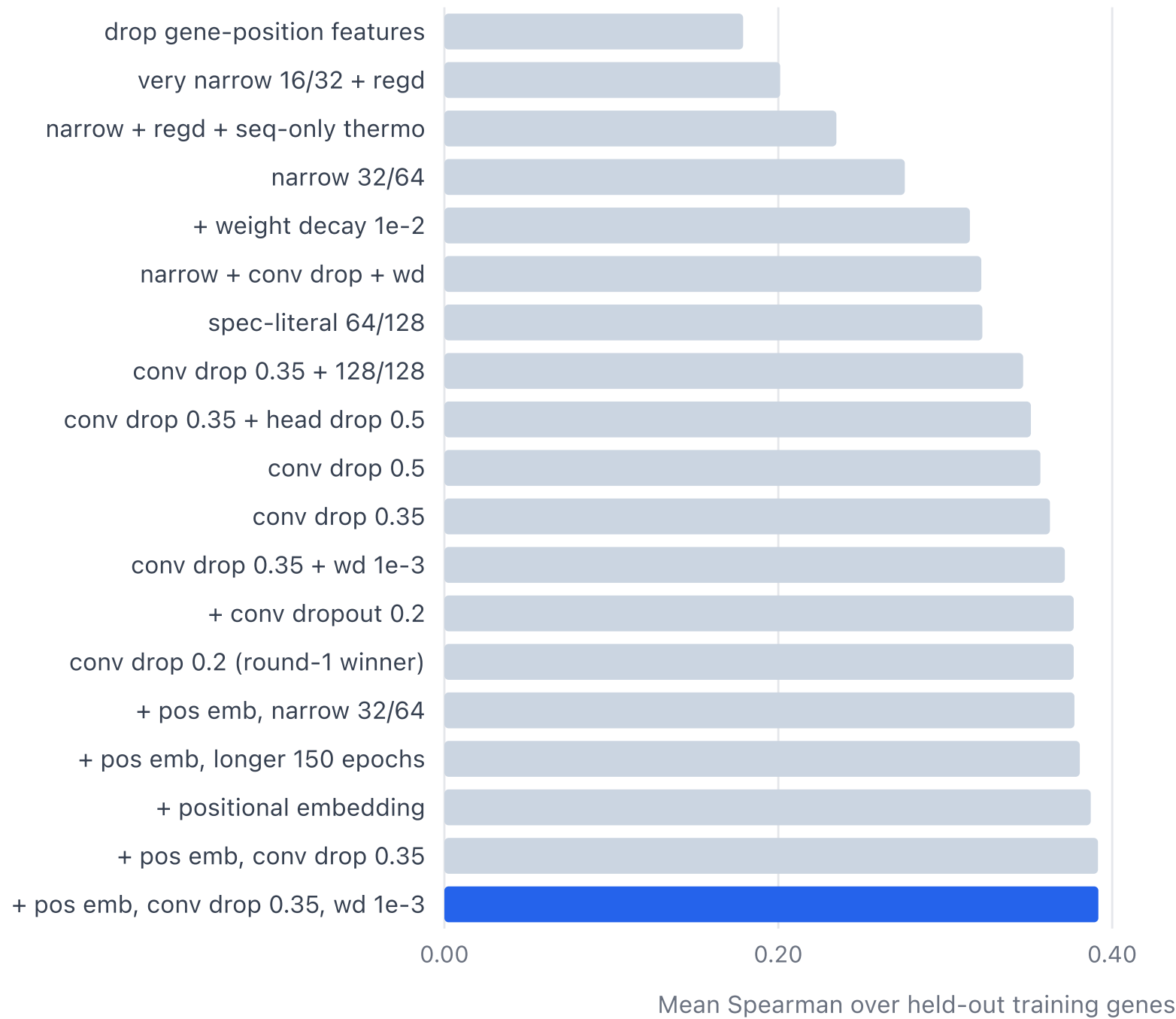
Mean width on a target bounded in  $(0, 1]$ , so a width of 1.0 is entirely uninformative.

#### The guarantee does not formally hold under gene shift

Conformal requires calibration and test data to be exchangeable. Under the random split they are, and coverage tracks nominal almost exactly, which validates the implementation. Under gene-held-out the calibration and test genes are disjoint, so the guarantee lapses; empirically it lands conservative (93.4% at nominal 90%), which is the safe direction but is luck rather than a theorem. Separately, a  $\pm 0.40$  interval on a  $(0, 1]$  target is wide enough that these are useful for flagging low-confidence predictions, not for ranking individual guides.

How the architecture was chosen

19 configurations, scored by leave-one-gene-out restricted to the **training genes only**. The test and calibration genes were never loaded during the sweep, so the reported test numbers stay honest.



Selected configuration in blue: + pos emb, conv drop 0.35, wd 1e-3 at 0.3917. The dashed line is the Rule Set 2 GBT measured on this same inner protocol, which is the bar that matters here.

What worked

**Feature-map dropout (+0.055).** The literal spec architecture reached 0.99 training Spearman against 0.21 held-out. Dropout1d was the single most effective fix.

**Learned positional embedding (+0.010).** Convolutions are translation-equivariant, so without one the network cannot tell position 4 from position 20, yet Rule Set 2 draws 58% of its Gini importance from position-specific nucleotide identity.

What did not work

**Dropping gene-position features (0.377 down to 0.179).** The hypothesis was that percent-peptide and cut position are gene-level shortcuts that cannot transfer across genes. Measurement said the opposite; they were kept.

**Narrowing the network.** Every narrower variant scored below the 64/128 stack. Capacity was not the problem; regularisation was.

Why 0.392 did not mean abandoning the model

The inner protocol trains on 5 genes instead of 16, so everything scores lower here. Re-measuring the GBT on the same protocol gave 0.505 rather than its 0.515 leave-one-gene-out figure, which established the correct bar. Given the full protocol, the CNN reached 0.507.

## Adversarial validation: 39 of 39 checks passing

Each check is written so it would actually fail if the corresponding bug were present, rather than restating what the code does. Run with `python scripts/08_validate.py`; it exits non-zero on any failure.

| CHECK                                                                | OBSERVED                                                                                                                 |
|----------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| ● FC rows within 5 of the paper's 1,841                              | got 1837; Azimuth drops a handful in the gene-position join (load_data notes '11 missing guides we were told to ignore') |
| ● RES unique guides == the paper's 2,549                             | got 2549 unique 30mers across 3473 rows; MED12 contributes 1848 rows under 2 drugs                                       |
| ● gene count == 17                                                   | -                                                                                                                        |
| ● split_gene (the reported protocol): no shared 30mer between splits | {'train/test': 0, 'train/cal': 0, 'cal/test': 0}                                                                         |
| ● split_random leaks duplicate sequences, as designed to demonstrate | {'train/test': 230, 'train/cal': 212, 'cal/test': 77} -- this split is reported only as a leakage contrast               |
| ● split_gene: no shared gene between splits                          | -                                                                                                                        |
| ● probe sequence is a valid 30mer with GG at [25:27]                 | seq=AAAAGGGGGCCCCCTTTTAAAAAAGGCTT                                                                                        |
| ● GC count uses the 20mer protospacer only                           | got 10.0, expected 10 (5 G + 5 C in seq[4:24])                                                                           |
| ● GC>10 and GC<10 indicators are both 0 at [4:24]                    | got 0.0, 0.0                                                                                                             |

### Two genuine findings, neither changing a conclusion

**The random split leaks duplicate sequences, not just gene context.** 230 test rows share a 30mer with a training row, because MED12 was screened under two drugs. Re-running conformal on only the test rows whose sequence the model never saw gives 89.2% coverage at 90% nominal, against 90.2% overall.

**Melting temperature on short segments is not physically meaningful.** Rule Set 2 computes Tm on 5nt and 8nt sub-segments, below the range where nearest-neighbour thermodynamics is valid, so the 5-mer values come out negative (-69 to +1 degC). This is inherited from the original design, not introduced here.

Dataset and splits

5,310

Guide x drug rows

4,379

Unique 30mers

17

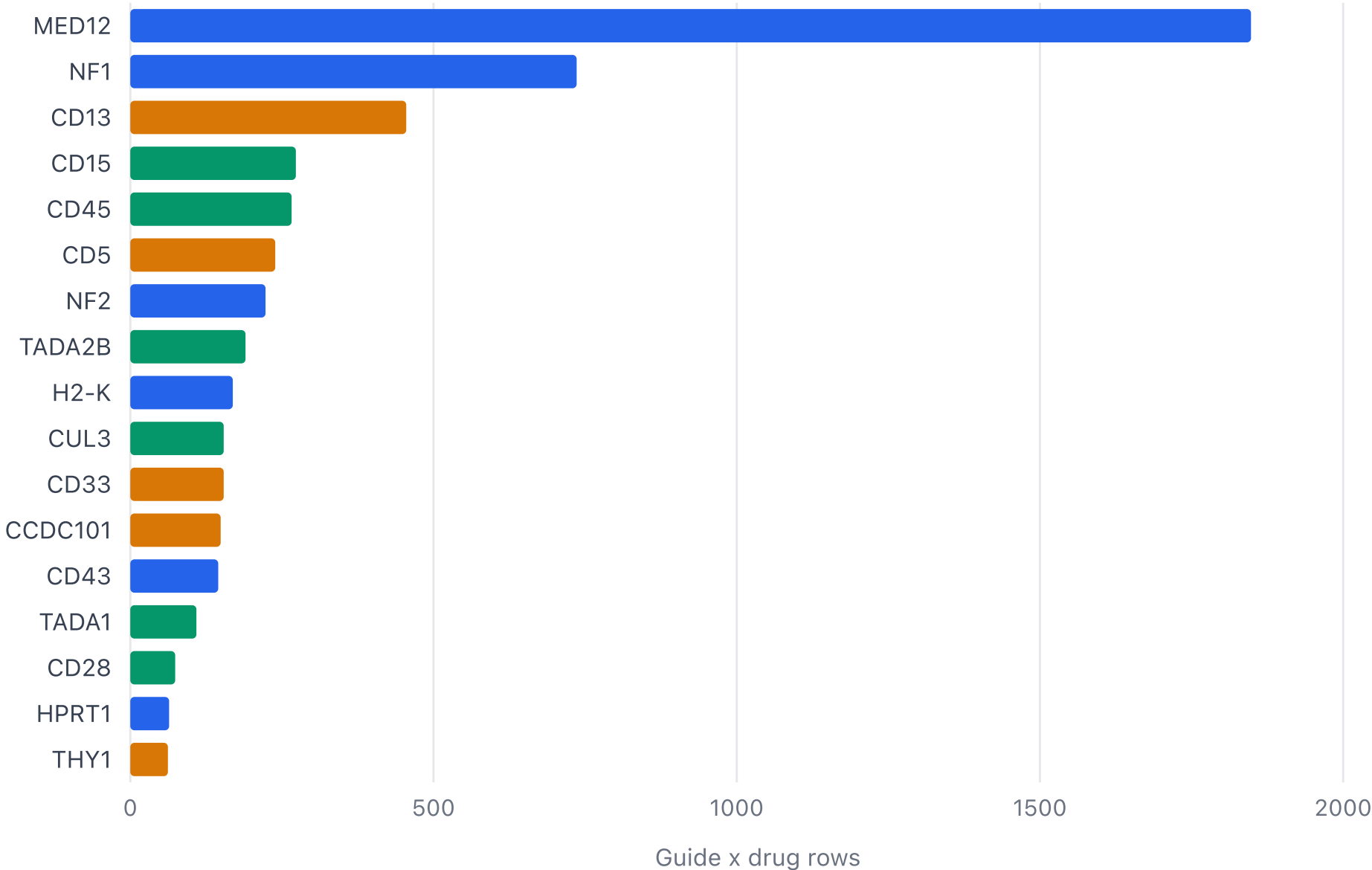
Genes

1,059

Calibration guides

Doench et al. 2016 (Nat Biotechnol 34:184) FC+RES, via MicrosoftResearch/Azimuth. Target is `score_drug_gene_rank`, activity rank-transformed within each (gene, drug) group, scaled to (0, 1].

Rows per gene, coloured by split



Blue train, orange calibration, green test. MED12 alone is 35% of all rows, so genes are assigned largest-first to whichever split is furthest below its quota rather than at random.

Why three protocols

**Leave-one-gene-out (primary).** Azimuth's own default and how Doench 2016 Fig. 4 reports Spearman. The paper-comparable number.

**Frozen gene-held-out 60/20/20.** Whole genes to train, calibration, and test. Needed for conformal, which requires a disjoint calibration set.

**Random row-level.** Reported only as a leakage contrast. It inflates the GBT from 0.454 to 0.544, roughly the size of the entire effect being measured.

Verified against the raw source

The target was independently re-derived from `V1_data.xlsx` and `V2_data.xlsx` by porting Azimuth's Python 2 rank-transform logic. Maximum absolute difference across all 5,310 rows: **5e-10**.

Chromatin accessibility channel: cut, with evidence

| CELL LINE                         | ROWS  | ENCODE ATAC-SEQ / DNASE-SEQ        |
|-----------------------------------|-------|------------------------------------|
| ● A375 (RES half)                 | 3,473 | 0 experiments, only RNA-based data |
| ● NB4 + TF1 + MOLM-13 (FC, human) | 882   | TF1 none; NB4 1; MOLM-13 3         |
| ● Mouse lines (FC)                | 955   | different genome entirely          |

Three independent blockers, any one sufficient

A375 generated 65% of the rows and has **zero** chromatin accessibility experiments on ENCODE; its only ENCODE data is RNA-based.

The remaining rows span three more human lines plus mouse, so a per-row match would need several tracks across two genomes, and TF1 has no ENCODE data at all.

The Azimuth release carries no genomic coordinates, only Ensembl transcript IDs and transcript-relative cut positions, so there is nothing to query a bigWig with.