



TECHNICAL WHITE PAPER · VERSION 2.0 · OCTOBER 2026

Z-Screen: Combinatorial Perturbation Transcriptomics, Model Benchmarking and Multi-Atlas Phenomimicry

How a readable chemical grammar, same-well multimodal nanowell measurement and calibrated chemical-genetic phenomimicry turn whole-transcriptome screening into an engine for therapeutic discovery

ORGANIZATION

Zafrens, Inc.

DATASET ANALYZED

Z-Screen pilot package V2 · 162,914 compounds · 190,699 profiles · huggingface.co/datasets/Zafrens/zscreen-pilot

EDITION

Public edition · chemical structures withheld

COMPANION VOLUME

Recursive Self-Improvement in Therapeutics

In brief. Z-Screen pairs combinatorial small-molecule synthesis from 627 reusable building blocks with whole-transcriptome and high-content imaging readouts in up to 50,000 nanowells per run. This paper analyzes the 162,914-compound pilot dataset: cellular responses decompose by building block ($r = 0.89\text{--}0.94$), same-well imaging and RNA agree inside the nanowell (Mantel $r = 0.773$), and calibrated phenomimicry against three chemical atlases and 43 Perturb-seq sets yields 110 chemical-reference and 2,626 genetic-reference matches, which converge on 14 distinct mechanism hypotheses. The pilot is a sample of the platform's output, not its capacity: Z-Screen is designed to scale to millions of compounds and beyond.

Every number in this paper is traceable to a released evidence table (Section 7).

z-screen@zafrens.com · zafrens.com

Contents

— Executive summary

What this paper shows

1 Platform and combinatorial grammar

A high-throughput language connecting chemical parts to cellular states

2 ActiveSeq same-well multimodality and scaling

Coupling morphology and transcriptomics inside the same nanowell

3 Internal and external benchmarking

Where Z-Screen representations add predictive power, and where they do not

4 Cross-atlas small-molecule phenomimicry ("chemical twins")

De-orphaning combinatorial chemistry against LINCS, Tahoe-100M and CIGS

5 Cross-modality CRISPR Perturb-seq phenomimicry ("genetic twins")

Bridging chemical profiles to 43 Perturb-seq signature sets and ~18,800 genes

6 High-confidence hypotheses from the pilot dataset

Fourteen mechanism hypotheses from twenty convergent series

7 Definitions, scope boundaries and sources

Reading the evidence with the intended precision

Figures and tables. 11 figures and 7 tables; the generated figures are built directly from the evidence tables listed in Section 7 and are reproducible from `_build/figures.mjs`.

EXECUTIVE SUMMARY

What this paper shows

Z-Screen is Zafrens' nanowell screening platform. It pairs combinatorial small-molecule synthesis from reusable building blocks with whole-transcriptome and high-content imaging readouts, at up to 50,000 conditions per run.

This white paper analyzes the **Z-Screen pilot dataset** released as Package V2 (`ZScreen_Pilot_Package_v2.0.0`, available on Hugging Face): **162,914 unique compounds** assembled from **627 recurring building blocks**, profiled as **190,699 compound–context transcriptional profiles** across eight library–cell contexts, plus an **11,435-nanowell ActiveSeq** same-well imaging-and-RNA co-assay. The pilot is a sample of what the platform measures, not its ceiling: Z-Screen is designed to scale to millions, and potentially hundreds of millions, of compounds.

Five workstreams are reported, each with its saved evidence tables:

- 1. Platform and combinatorial grammar.** Cellular responses decompose by building block: computationally removing one block from a trained model reproduces its measured effect at median Spearman $r = 0.89$ – 0.94 , and fragment-level weights rebuild whole-block effects at $r = 0.97$ – 0.99 .
- 2. ActiveSeq same-well co-assay.** Morphology and RNA measured inside the same nanowell agree globally (Mantel $r = 0.773$) and the exact well link carries information that shuffled pairings lose (0.109 versus -0.001 held-out correlation).
- 3. Model benchmarking.** Simple recipe models and a reference transformer split the eight contexts four to four. On unseen building blocks the additive recipe model edges Chemprop, and retrospective active-learning replays recover the same responses with 33–52% fewer measurements.
- 4. Chemical twins.** Calibrated cross-atlas matching against LINCS L1000, Tahoe-100M and CIGS recovers reference controls (median panel rank 0.822 on 17 qualified controls) and yields 110 qualifying series–reference matches across 51 series.
- 5. Genetic twins.** Bridging to 43 Perturb-seq signature sets (~18,800 targets) yields 2,626 strict series–knockout matches in 178 disjoint program clusters, with zero synthetic false discoveries at $FDR < 0.05$.

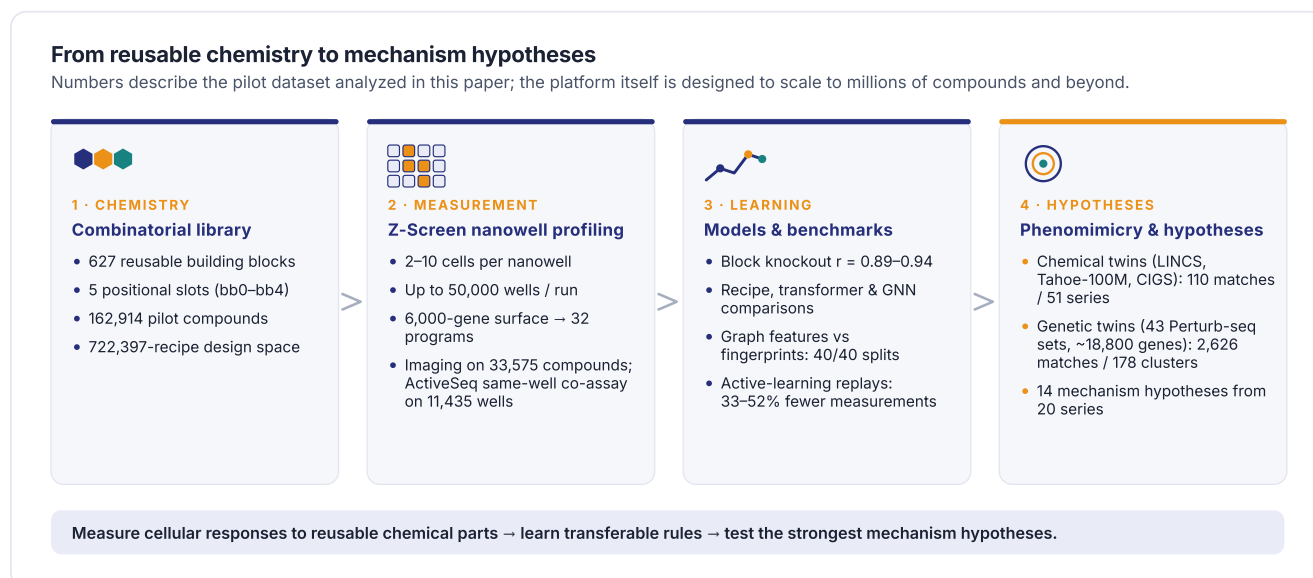


Figure 1. The Z-Screen discovery architecture. Reusable building blocks define a 722,397-recipe design space, of which the pilot dataset measures 162,914 compounds. Z-Screen profiles up to 50,000 nanowells per run on a 6,000-gene surface projected to 32 transcriptional programs, with library imaging on 33,575 compounds and the ActiveSeq same-well co-assay on 11,435 control wells. Calibrated models and phenomimicry engines turn those measurements into 110 chemical-reference and 2,626 genetic-reference matches, which converge on 14 distinct mechanism hypotheses.

162,914**pilot compounds**

190,699 compound–context profiles across 8 contexts; 627 building blocks

 $r = 0.89\text{--}0.94$ **block-level modularity**

in silico building-block knockout reproduces measured block effects

0.822**chemical-twin calibration**

median matched-panel rank on 17 CIGS reference controls; MZ1 chromatin-class AUROC 0.799

110 + 2,626**reference matches**

chemical-twin series matches and strict genetic-twin matches (178 clusters)

1 · PLATFORM AND COMBINATORIAL GRAMMAR

A high-throughput language connecting chemical parts to cellular states

Recurring chemical parts and rich cellular readouts let Z-Screen learn how specific substitutions change biological responses.

1.1 The Z-Screen platform

The Z-Screen microfluidic platform compartmentalizes 2 to 10 intact human cells per nanowell across arrays of up to 50,000 conditions per run. Whole-transcriptome mRNA is captured on barcoded beads, with high-content imaging acquired either across library runs or inside the exact same well through the ActiveSeq co-assay (Section 2).

Whether operating in sequencing mode, imaging mode, or conditionally paired library imaging and sequencing, the platform is called Z-Screen. The term ActiveSeq is reserved for experiments in which high-content imaging and RNA sequencing occur inside the same nanowell.

1.2 The pilot dataset analyzed here

In the public Package V2 release, raw sequencing reads across 46,944 human genes are depth-normalized, background-adjusted with the documented context-specific baseline, harmonized onto a 6,000-gene high-variance expression surface, and projected onto a pinned $k = 32$ semi-nonnegative matrix factorization (semi-NMF) transcriptional program basis ($W \in \mathbb{R}^{32 \times 6000}$). Across four combinatorial libraries (ZEL024, ZEL028, ZEL031, ZEL039) and five human cell lines (HEK293, H1650, A549, THP1 and AEC7 endothelial cells), the pilot dataset comprises **190,699 compound–context profiles** spanning **162,914 unique compounds** assembled from **627 active positional building blocks** (bb0–bb4).

The release is public. The pilot dataset is published on Hugging Face as Zafrens/zscreen-pilot, with pinned per-file manifests and SHA-256 hashes; the evidence tables in Section 7 derive from that release. An interactive Functional Drug Atlas (<http://32.190.137.56/>) lets readers browse a portion of the measured responses.

Scope of the numbers in this paper

Every count, correlation and match reported here refers to the pilot dataset. It establishes that the platform's measurements are learnable and transferable; it does not bound the platform's capacity, which is designed to scale to millions of compounds and beyond.

1.3 Substructure modularity: cells read individual chemical parts

Unlike diversity libraries in which every molecule is an isolated scaffold, Z-Screen's combinatorial design means the median building block is observed across **461 distinct molecules**, and even a molecule's least-common synthon appears in roughly 1,100 other library members. Across the corpus, **5,077,542 measured compound pairs** differ by a single positional building-block substitution (207,956 pairs where both compounds sit in the held-out test fold). The building blocks define **722,397 unique chemical recipes** and **2,079,661 nominal compound–context states** across 104,385 Bemis–Murcko scaffolds, so the 190,699 measured profiles sample roughly 9.2% of the nominal combinatorial grid.

A compound is a recipe of positional building blocks

The same block recurs across hundreds of molecules and across libraries, so its cellular effect can be isolated. Structures are withheld in this public edition; block identities are schematic.

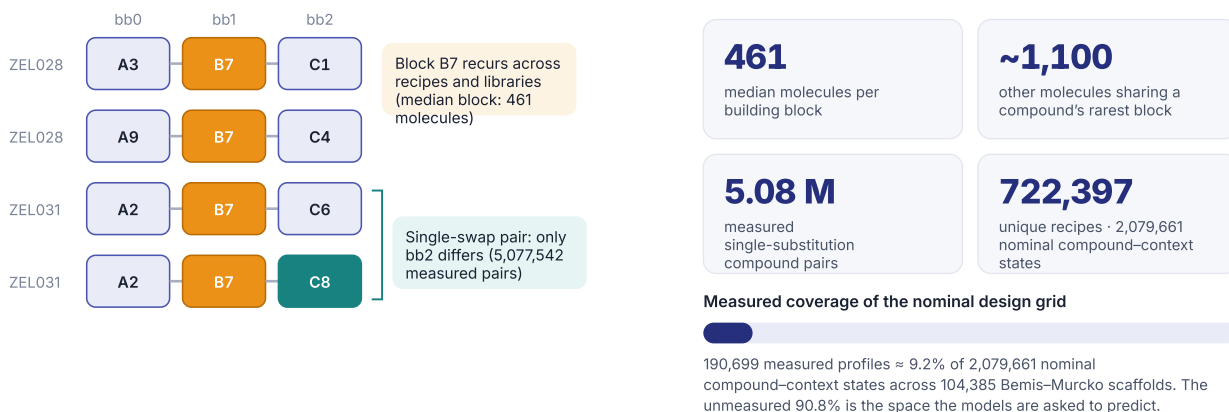


Figure 2. Recurring building blocks across recipes and libraries. A compound is a recipe of positional blocks (schematic; structures are withheld in this public edition). The same block recurs with different neighbors and across libraries, which isolates its contribution, and single-swap pairs expose activity cliffs. Right: reuse and coverage statistics for the pilot dataset.

Cellular transcriptional responses respect this chemical grammar. Computationally knocking out a single building block from a trained model's input reproduces that block's empirical marginal effect in cells at a **median Spearman $r = 0.89$ to 0.94** across the two most deeply measured contexts (ze1024_hek293 and ze1039_aec7, three splits each; $z = 13.8$ to 19.2 against a shuffled-label null near zero). Decomposing each building block into Morgan substructures and summing the learned fragment weights reconstructs the whole block's cellular effect at **$r = 0.97$ to 0.99** .

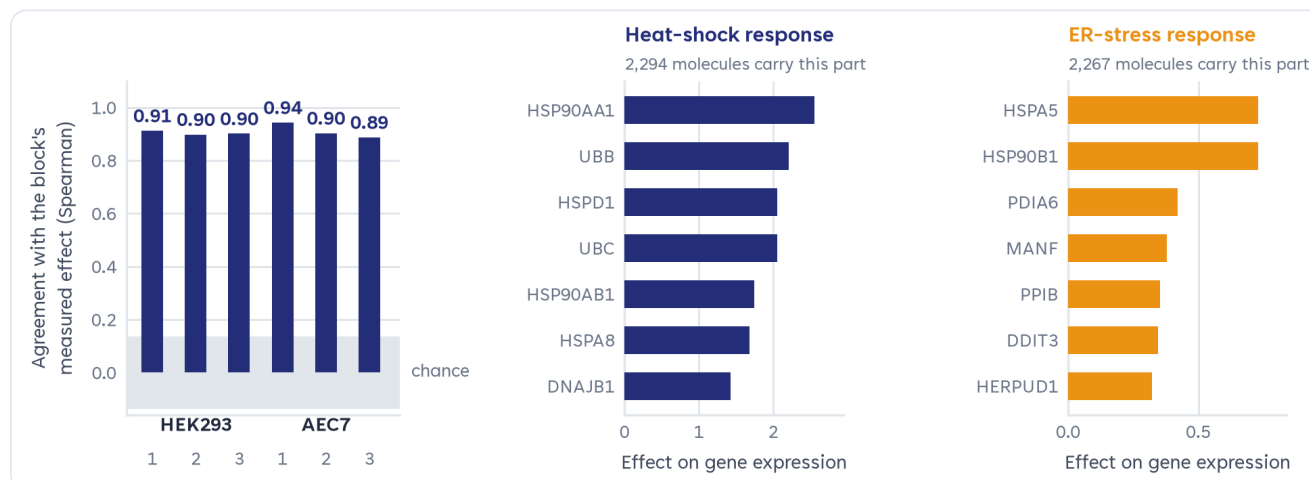


Figure 3. In silico building-block knockout and gene-program projection. Left: median Spearman correlation between computationally removing a building block from the model and its measured empirical effect, for about 100 blocks per split (HEK293 and AEC7; shaded band marks 2 SD of the shuffled null). Right: projecting individual block vectors onto the 6,000-gene surface separates biological states. One block carried by 2,294 molecules induces cytoplasmic heat-shock chaperones (*HSP90AA1*, *HSPA8*, *DNAJB1*); an alternative block at the same position (2,267 molecules) selectively activates endoplasmic-reticulum unfolded-protein stress (*HSPA5*/BiP, *HSP90B1*, *DDIT3*/CHOP).

2 · ACTIVESEQ SAME-WELL MULTIMODALITY AND SCALING

Coupling morphology and transcriptomics inside the same nanowell

Across the broader Z-Screen corpus, 33,575 library compounds have conditionally paired high-content imaging and transcriptomics (ZEL024: 10,129 compounds; ZEL031: 8,443; ZEL039: 15,003), with 448 CellProfiler morphological features per condition.

To test whether single-well morphological variation co-varies with transcriptional program usage inside the same cells, without batch or well-to-well confounding, a dedicated **11,435-nanowell ActiveSeq study** imaged each nanowell across brightfield and fluorescence channels immediately before lysis and barcoded mRNA capture, across 35 pharmacological control compounds.

- **Global cross-modality geometry.** Across the 35-control panel, pairwise distances in the 448-feature morphology space correlate with distances in the 32-program RNA space at Mantel $r = 0.773$ (exact permutation $p = 1/2,001 = 0.0005$).
- **The exact well link carries information.** On 2,287 held-out ActiveSeq wells evaluated across 72 models, preserving the nanowell link between image and RNA yields a QC-adjusted within-treatment image-to-RNA correlation of **0.109**; breaking the link within the same compound and batch collapses it to **-0.001**. Real pairs beat broken pairs in 35 of 35 compound averages and 64 of 70 compound-by-batch strata.
- **Population-level recognition.** After pooling 32 held-out wells per treatment, joint image-plus-RNA classification exceeds RNA alone by +2.86 percentage points (batch 1 $\rightarrow$ 2) and +10.48 points (batch 2 $\rightarrow$ 1). Single-well classification shows no consistent gain; the pooled result is a post-fit aggregation under a fixed classifier, not a learned fusion model.

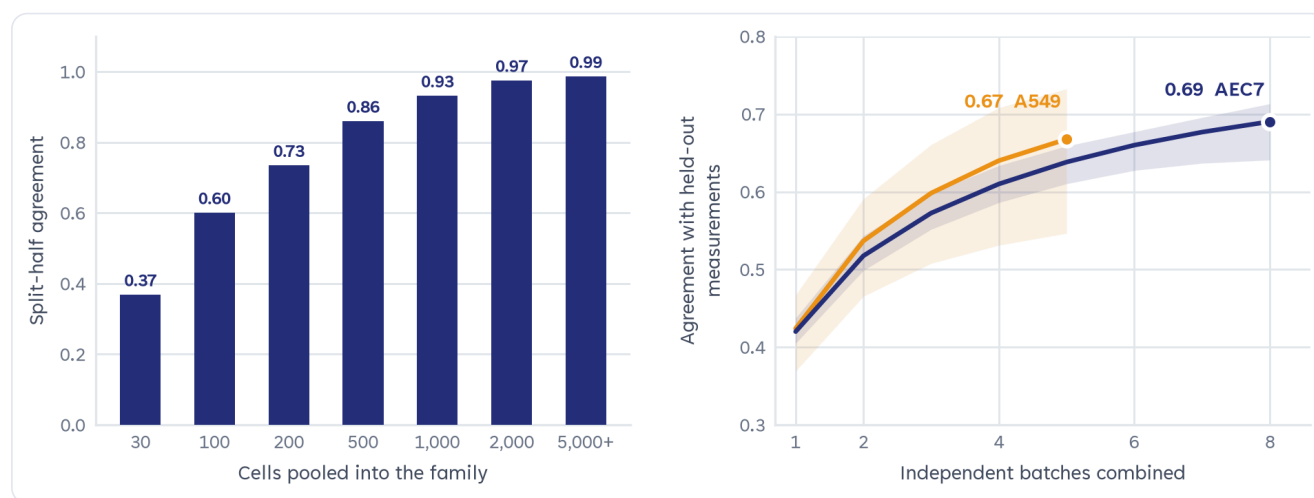


Figure 4. Measurement depth and batch accumulation. Left: across 1,562 endothelial chemical families (*ze1039_aec7*), split-half correlation between disjoint halves of a family's members rises from 0.37 (30–99 pooled cells) to 0.99 (more than 5,000 pooled cells; Spearman $r = 0.882$). Right: combining independent acquisition batches of control compounds raises gene-level agreement with held-out batches from 0.420 (1 batch) to 0.690 (8 batches) in AEC7 and from 0.424 to 0.668 in A549, in 20 of 20 partitions.

3 · INTERNAL AND EXTERNAL BENCHMARKING

Where Z-Screen representations add predictive power, and where they do not

Two internally generated benchmarking workstreams evaluated within-library generalization, cold-start chemistry, active learning, and transfer to public datasets. Each comparison below reports the saved test set it was scored on.

3.1 Within-library generalization on the released fold-0 split

In the released fold-0 benchmark, the reference transformer wins four of eight contexts and simple recipe models win the other four. Simple winners include additive building-block marginals and slot-embedding ridge regression.

HEK293/ZEL024 reaches transformer $r = 0.582$; AEC7/ZEL039 favors slot-embedding ridge at $r = 0.451$.

Within-library generalization on the released fold-0 split

Pearson r between predicted and measured 32-program response, per library-cell context. Simple recipe models and the reference transformer each win four of eight contexts.

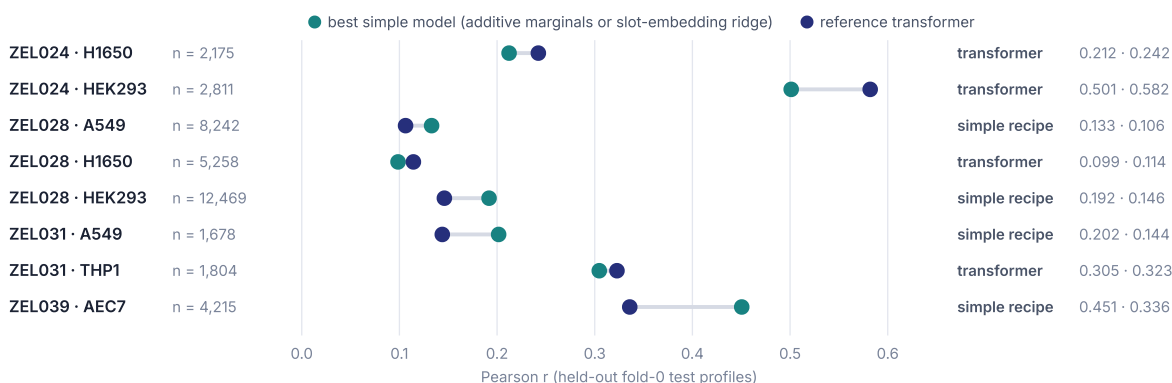


Figure 5. Fold-0 reference models by context. Pearson r on held-out fold-0 test profiles for the best simple recipe model (teal) and the reference transformer (navy). Fold 0 is the test fold; folds 1–4 train; fold 4 tunes the simple models. Source: [evidence/fold0_reference_model_summary.csv](#).

Table 1. Fold-0 comparison of the best simple model and the reference transformer, per context.

Context	Test profiles	Best simple model	Simple r	Transformer r	Winner
ZEL024 · H1650	2,175	additive BB marginals	0.212	0.242	transformer
ZEL024 · HEK293	2,811	additive BB marginals	0.501	0.582	transformer
ZEL028 · A549	8,242	slot-embedding ridge	0.133	0.106	simple
ZEL028 · H1650	5,258	slot-embedding ridge	0.099	0.114	transformer
ZEL028 · HEK293	12,469	slot-embedding ridge	0.192	0.146	simple
ZEL031 · A549	1,678	additive BB marginals	0.202	0.144	simple
ZEL031 · THP1	1,804	additive BB marginals	0.305	0.323	transformer
ZEL039 · AEC7	4,215	slot-embedding ridge	0.451	0.336	simple

3.2 Cold-start chemistry and active learning

A central question for medicinal chemistry is what happens when a candidate incorporates a building block the model has never seen (the `bb_disjoint` split). In a head-to-head benchmark on `zel024_hek293` fold 0 ($n = 7,544$ held-out molecules with unseen building blocks), an additive recipe model was compared with an end-to-end message-passing graph neural network (Chemprop v2.3.1 D-MPNN) and with Z-Screen's shared-context transformer.

- Unseen building blocks.** On the same 7,544 compounds, the additive recipe model reaches mean program $r = 0.363$, $R^2 = 0.109$ and 2.62× directional enrichment at a 5% selection budget, versus Chemprop at 0.337, 0.089 and 2.53×. Both use 5,721 training and 649 validation profiles. The three-seed shared-context transformer reaches $r = 0.398$ and 2.86× enrichment using eight contexts, but its mean program R^2 is -0.99 : ranking improves while effect-size calibration worsens.
- Pretrained graph features generalize.** Across all 8 contexts × 5 folds under strict `bb_disjoint` partitioning, frozen CheMeleon-2048 graph features (8-context mean $r = 0.264$, up to 0.408) outperform Morgan-2048 fingerprints (mean $r = 0.242$) in 40 of 40 splits. Adding Z-Screen recipe coordinates to frozen CheMeleon further improves top-5% selection enrichment in 8 of 8 contexts (up to 4.54×).
- Retrospective active learning.** Across five of six objectives, model-directed selection matched the response recovery of 1,400 random measurements using 667–933 selections: 33–52% fewer measurements. HSPA5 recovery reached 42.8%, versus 30.7% for a frozen model and 14.8% for random selection. These replays use previously measured compounds; prospective campaigns will test the laboratory savings.

Model comparison on unseen building blocks

HEK293 (ZEL024 fold 0) · 7,544 held-out compounds, each containing a building block absent from training · 5,721 training and 649 validation profiles for the matched models.

	Program $r \uparrow$	Program $R^2 \uparrow$	5% enrichment $\uparrow$
Chemprop v2.3.1D-MPNN end-to-end graph network, HEK293 only	0.3367	0.0888	2.534×
Additive recipe model building-block identities only, HEK293 only	0.3628	0.1094	2.624×
Shared-context transformer (3-seed ensemble) trained on all eight contexts	0.3978	-0.987 (worse calibration)	2.861×

Reading: the additive recipe model edges the end-to-end graph network on all three metrics with identical training data.

The shared-context transformer ranks best but its program R^2 is negative: ranking improves while effect-size calibration worsens.

Figure 6. Model comparison on unseen building blocks. HEK293 (ZEL024 fold 0, 7,544 held-out molecules); the transformer uses additional context supervision from all eight contexts. Source: [evidence/cold_start_metrics.csv](#).

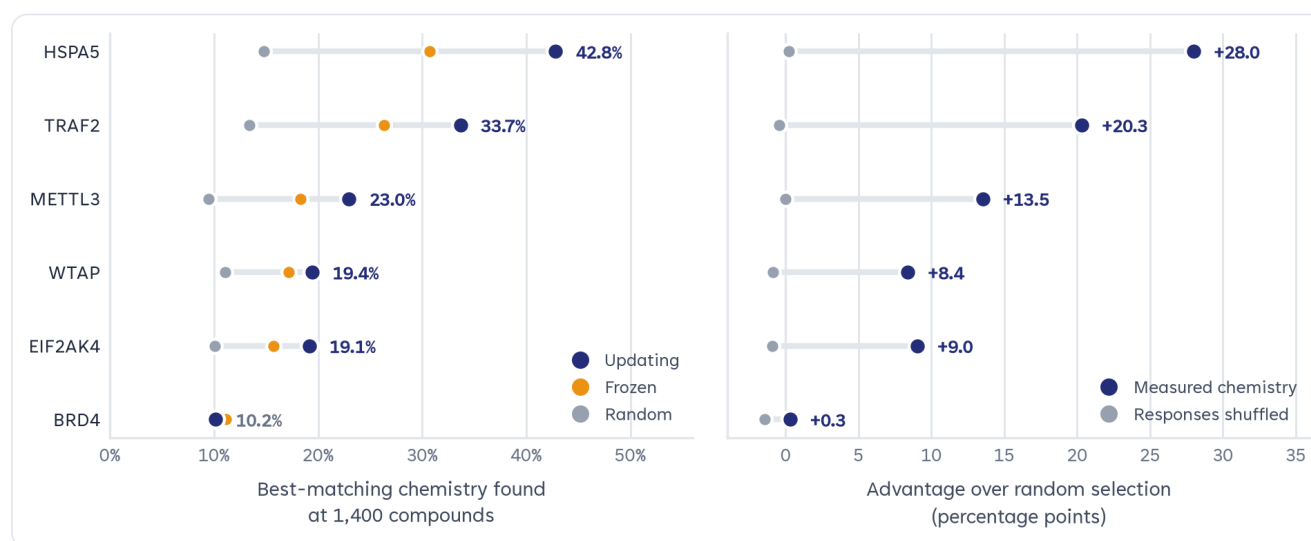


Figure 7. Closed-loop active learning across six objectives, with shuffled-response controls. Left: recovery of the full-universe top-5% response set at a 1,400-compound budget (200 initial plus six rounds of 200). Updating the recipe model after each round (navy) beats the frozen initial model (orange) and random selection (gray) on five of six objectives, reaching 42.8% on HSPA5. Right: the advantage over random selection vanishes when chemistry–response labels are shuffled (72 matched permutation runs collapse to about 0 points).

3.3 External check: the program basis on public primary cells

NeurIPS 2023 primary human PBMCs

On 614 PBMC profiles (146 drugs × 6 cell types), the pinned Z-Screen $k = 32$ program basis captures 17.59% of total differential-expression variance: 32.29× a random orthogonal basis and 6.40× a permuted basis ($p < 10^{-10}$). The basis was fitted on Z-Screen cell lines only; this is a coverage check on primary cells, not a prediction benchmark.

4 · CROSS-ATLAS SMALL-MOLECULE PHENOMIMICRY (“CHEMICAL TWINS”)

De-orphaning combinatorial chemistry against LINCS, Tahoe-100M and CIGS

An uncharacterized compound that induces the same signed transcriptional state as a validated drug, and whose structural analogs preserve that match across a chemical series, provides an immediate, testable mechanism-of-action hypothesis. We call such a match a chemical twin.

4.1 Baseline calibration and scoring

Cross-atlas matching compares signed transcriptional responses between platforms that were never designed to be compared, so three calibration choices determine whether pharmacological concordance is visible above noise:

- 1. Signed contrastive projections.** Up- and down-regulated genes are scored with their signs preserved. Averaging unsigned ranks across the top- $|z|$ genes would cancel a perfect phenomimic (up-regulated genes at the 95th percentile, down-regulated genes at the 5th) to an uninformative 50%.
- 2. Vehicle-based centering.** The internal control panel (ZIC008) is 90.7% active oncology, kinase and epigenetic drugs (35 drugs versus 4 vehicle sets). $\log_2$ fold changes and robust z-scores are therefore computed from raw UMI counts against 72 pure DMSO vehicle wells, or against the more-than-10,000-compound library median/MAD, rather than against the active-dominated panel mean, which would subtract the shared antiproliferative and cell-cycle pathways that carry much of the signal.
- 3. Plate-level standardization of CIGS.** Plate-based robust median/MAD standardization doubles CIGS replicate cosine (0.088 $\rightarrow$ 0.168). The analysis uses 3,261 exactly shared genes; the active reference panel contains 1,940 compounds.

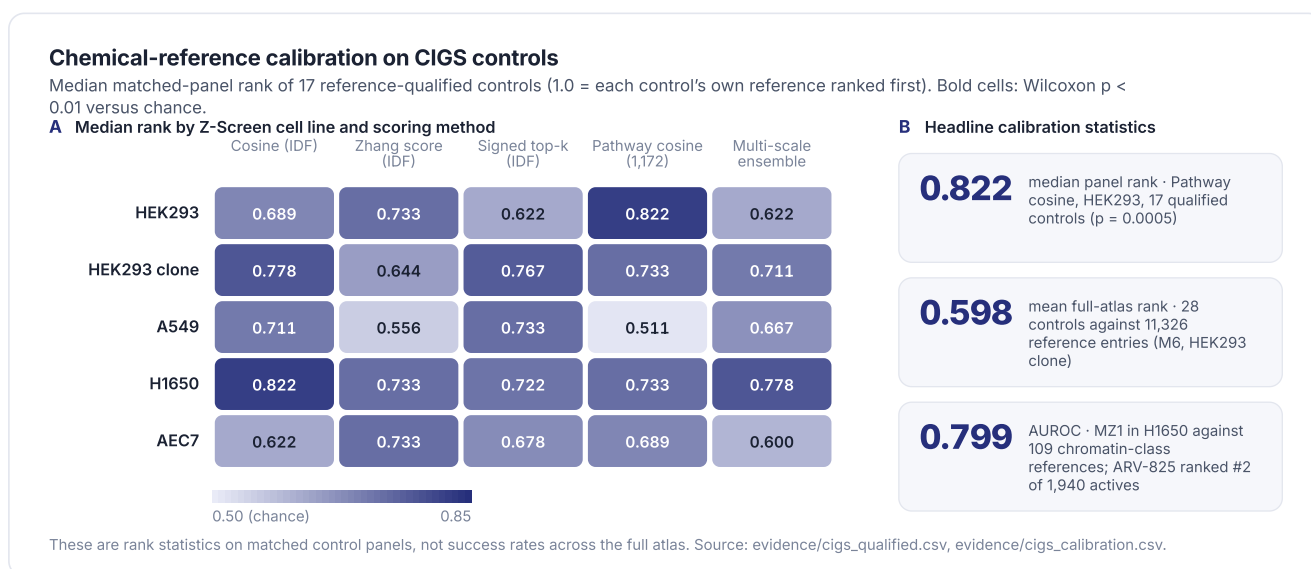


Figure 8. Chemical-reference calibration on CIGS controls. A: median matched-panel rank of 17 reference-qualified controls for each Z-Screen cell line and scoring method; bold cells are significant at Wilcoxon $p < 0.01$. B: headline statistics. The 0.822 median comes from pathway cosine scoring in HEK293 and is a best-matching-entry rank statistic, not an 82% success rate across the full atlas; full-atlas scoring (28 controls, 11,326 entries) gives a mean rank of 0.598. The MZ1 AUROC measures recovery of a broad 109-drug chromatin class in H1650, not BRD4 specificity; ARV-825 ranks second among 1,940 active references.

4.2 Library-wide chemical-twin discovery: 110 matches across 51 series

Using a five-channel similarity ensemble coupled with bipartite mutual specificity (Z_{mutual}), well-count-stratified empirical permutation nulls ($B = 2,000$), Benjamini–Hochberg FDR $q < 0.01$, and at least one qualifying orthogonal support channel, 110 qualifying (series, reference) matches were identified across 51 chemical series (47 distinct variable cores). Of these, 33 have both internal-control and CRISPR support, including all ten highest-ranked chemical-twin series (ZS-SAR-01 to ZS-SAR-10).

Table 2. The ten highest-ranked chemical-twin series. Z_{emp} is the series Z against the empirical permutation null; “mw” counts multi-well active analogs.

Series	Context · variable core	Size / actives	External twin (CIGS)	Z_{emp} (FDR q)	Internal control	Same-MOA CRISPR KO
ZS-SAR-01	A549 · ZEL028 BB_2172915349 (+BB_3460866978)	132 / 15 (0 mw)	Roniciclib CDK4/6/7/9 kinase	7.15 (1.6e–8)	Palbociclib Z = +5.14	<i>CDK9</i> KO Z = +5.27
ZS-SAR-02	A549 · ZEL031 BB_3103020399	83 / 14 (13 mw)	Endoxifen ER/AR nuclear receptor	5.75 (1.1e–8)	Endoxifen Z = +6.40	<i>AR</i> KO Z = +6.43
ZS-SAR-03	H1650 · ZEL028 BB_3467217614	213 / 11 (0 mw)	Ceritinib ALK/MET/IGF1R RTK	5.97 (8.4e–6)	AZD-3463 Z = +6.26	<i>ALK</i> KO Z = +3.56
ZS-SAR-04	THP1 · ZEL031 BB_3717371053	87 / 5 (3 mw)	AZD5153 BRD4/BET reader	4.90 (8.4e–5)	MZ1 Z = +4.10	<i>BRD2</i> KO Z = +3.49
ZS-SAR-05	H1650 · ZEL024 BB_5272889191	131 / 29 (11 mw)	HTH-01-015 NUAK1/ROCK kinase	7.79 (7.9e–9)	HTH-01-015 Z = +2.08	<i>NUAK1</i> KO Z = +3.11
ZS-SAR-06	H1650 · ZEL028 BB_8596154155	223 / 15 (0 mw)	GSK3368715 type I PRMT inhibitor	4.97 (7.3e–4)	STM2457 Z = +3.58	<i>EZH2</i> KO Z = +4.28
ZS-SAR-07	AEC7 · ZEL039 BB_0302650651 & BB_7334588199	37 / 5 (1 mw)	LY2857785 CDK9/7 elongation	3.94 (6.1e–9)	Palbociclib Z = +4.16	<i>CDK9</i> KO Z = +6.70
ZS-SAR-08	A549 · ZEL031 BB_5574910360	88 / 17 (16 mw)	Ailanthone AR/ER chaperone axis	5.63 (2.5e–8)	Endoxifen Z = +6.70	<i>AR</i> KO Z = +6.83
ZS-SAR-09	H1650 · ZEL024 BB_3472530680 & BB_2122574568	76 / 17 (17 mw)	ZM 336372 RAF/MEK/MAPK	4.34 (6.9e–5)	ZM 336372 Z = +2.96	<i>MAP2K1</i> KO Z = +3.99
ZS-SAR-10	H1650 · ZEL024 BB_7372899989 & BB_8873714086	136 / 23 (20 mw)	ZM 336372 RAF/MEK/MAPK	4.45 (5.5e–5)	Sorafenib Z = +2.51	<i>RAF1</i> KO Z = +4.49

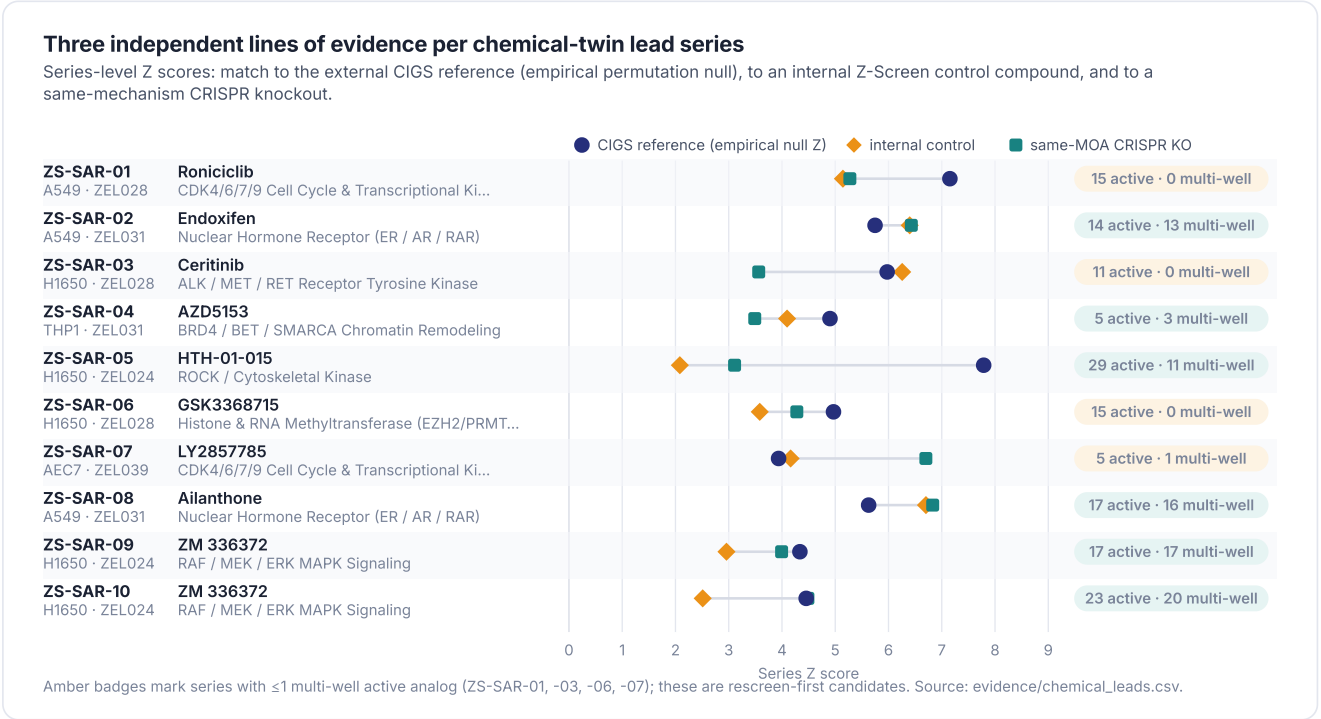


Figure 9. Three lines of evidence per chemical-twin series. For each prioritized series: the series Z against the empirical permutation null for the external CIGS reference (circle), the Z of the matching internal Z-Screen control compound (diamond), and the Z of a same-mechanism CRISPR knockout (square). Series with one or no multi-well active analogs (ZS-SAR-01, -03, -06, -07) are flagged as rescreen-first candidates. Chemical structures are withheld in this public edition; compound identifiers and measured results are retained in [evidence/chemical_leads.csv](#).

5 · CROSS-MODALITY CRISPR PERTURB-SEQ PHENOMIMICRY ("GENETIC TWINS")

Bridging chemical profiles to 43 Perturb-seq signature sets and ~18,800 genes

Small-molecule phenomimicry is bounded by targets that already have tool compounds. Bridging Z-Screen directly to genome-wide CRISPR Perturb-seq extends chemical search to genetic reference responses beyond existing tool compounds.

5.1 Reference harmonization and scoring calibration

We harmonized **43 Perturb-seq signature sets** (about 18,800 target genes and 8 million single cells, drawn from X-Atlas Orion 2025 HEK293T and HCT116; Replogle et al., *Cell* 2022, K562 and RPE1; Nadig et al., *Nature Genetics* 2025, HepG2 and Jurkat; Jiang et al., *Nature Biotechnology* 2025, 30 cytokine-by-cell matrices; Zhu et al., primary human CD4⁺ T cells; and the 2025 Virtual Cell Challenge) and constructed a **1,225-target multi-atlas reference panel** (1,068 verified multi-atlas knockdowns with target-gene $z \leq -1.0$, 132 priority therapeutic targets, and 25 obligate multiprotein complexes).

Perturb-seq pseudobulks across 25,000 genes are sparse: roughly 45% of genes have a median absolute deviation of zero, so an unfloored z-score lets a handful of sparse genes dominate cosine similarity. Scoring methods were therefore calibrated on 25 matched Tier-1 intracellular compounds across genome-wide and essential-gene reference datasets, using the compound-median rank of each compound's direct target (0.5 = chance). A naive unfloored cosine scores 0.485. Enforcing a positive MAD floor with a pure-vehicle baseline raises the median to 0.541 (paired median change +0.087, $p = 0.045$), and the selected method, an unblended control-contrast bimodal Z_{mutual} ensemble, reaches 0.580 ($p = 0.043$). The 1,225-entry consensus panel is then used for library discovery.

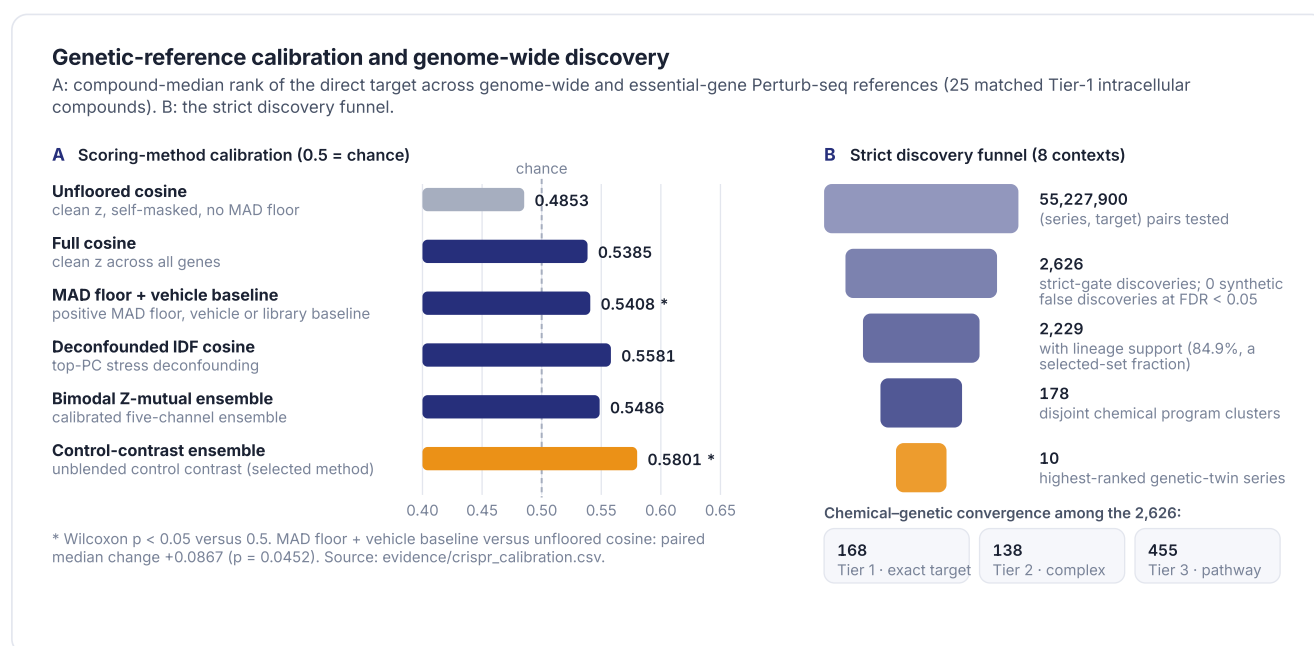


Figure 10. Genetic-reference calibration and the discovery funnel. A: compound-median rank of the direct target for six scoring variants on the 25-compound matched benchmark (0.5 = chance); the selected method is highlighted. B: from 55.2 million tested (series, target) pairs to 2,626 strict discoveries, 2,229 with lineage support, 178 disjoint program clusters and the ten highest-ranked genetic-twin series. Lineage support participates in selection, so 84.9% is a selected-set fraction, not a prospective replication rate. Source: evidence/crispr_calibration.csv.

5.2 Genome-wide CRISPR-SAR discoveries and chemical-genetic convergence

Evaluating **55,227,900 (series, target) pairs** against well-count- and magnitude-stratified empirical nulls, with a synthetic negative-control pass yielding **zero false discoveries at FDR < 0.05 across all eight contexts**, identified **2,626 strict-gate (series, CRISPR knockout) discoveries** across 293 family identifiers, 251 core building blocks and 178

disjoint chemical program clusters, with 2,229 (84.9%) passing the decoupled lineage-support gate. Cross-referencing against 4,102 pre-staged chemical-twin pairs produced 168 Tier 1 (exact direct target), 138 Tier 2 (obligate physical complex) and 455 Tier 3 (broad pathway) convergent rows.

Table 3. The ten highest-ranked genetic-twin series. Z_{mag} is the series Z against the magnitude-matched null; “Disc → Holdout” gives the discovery-partition Z and the decoupled holdout-lineage Z.

Rank · context	Core variable block	Primary CRISPR target (partner)	Z_{mag} (FDR q)	Disc → Holdout Z	Cliff Tanimoto · chemical twin (tier)
#1 · A549 · ZEL031	bb0 = BB_3717371053 N = 87, 33 mw	<i>BRD2</i> partner <i>ELL2</i> , $Z_{\text{corr}} = 2.67$	16.37 (4.2e-27)	18.46 → 9.20	0.873 · dBET6 / AZD5153 (Tier 1 exact)
#2 · AEC7 · ZEL039	bb1 = BB_0164651813 (+BB_2085420374) N = 51, 15 mw	<i>PSMA3</i> partner <i>PSMB5</i> , $Z_{\text{corr}} = 7.94$	10.70 (8.8e-14)	9.82 → 9.76	0.875 · Ixazomib citrate (Tier 2 complex)
#3 · HEK293 · ZEL024	bb3 = BB_6048600856 N = 167, 12 mw	<i>ATR</i> / <i>CHEK1</i> / <i>WEE1</i> <i>ATR</i> Z = 7.94	11.12 (2.6e-20)	7.03 → 13.36	0.912 · Prexasertib / AZD-7762 (Tier 2)
#4 · A549 · ZEL031	bb0 = BB_7209113661 N = 87, 13 mw	<i>ACTL6A</i> / BAF53A SWI/SNF	11.86 (6.7e-22)	8.09 → 8.94	0.868 · dBET6 / AZD5153 (Tier 3 pathway)
#5 · HEK293 · ZEL024	bb3 = BB_1193003673 (+BB_5290020531) N = 84, 12 mw	<i>TYK2</i> partner <i>STAT5B</i> , $Z_{\text{corr}} = 2.91$	6.61 (1.3e-10)	3.74 → 4.06	0.800 · Ilginatinib, JAK-STAT (Tier 3)
#6 · AEC7 · ZEL039	bb0 = BB_1722497045 & bb1 = BB_4318016053 N = 35, 5 mw	<i>CDK9</i> partner <i>SUPT5H</i> , $Z_{\text{corr}} = 7.85$	4.18 (6.8e-11)	3.96 → 3.51	0.808 · Dinaciclib, CDK9 (Tier 1 exact)
#7 · A549 · ZEL031	bb1 = BB_4729940582 N = 82, 4 mw	<i>RPL28</i> partner <i>ATF4</i> , $Z_{\text{corr}} = 1.81$	5.30 (4.1e-5)	3.10 → 5.12	0.847 · Anisomycin, ribosome/ISR (Tier 1)
#8 · THP1 · ZEL031	bb1 = BB_9265511036 N = 112, 3 mw	<i>PLK1</i> / <i>AURKA/B</i> / <i>CDK1</i> <i>CDK1</i> Z = 5.04	6.07 (3.2e-4)	7.46 → 2.66	0.976 · SCH-1473759, mitotic (Tier 3)
#9 · H1650 · ZEL024	bb3 = BB_9434530945 N = 136, 9 mw	<i>METTL3</i> / <i>METTL14</i> / <i>WTAP</i> m^6A writer	4.58 (2.7e-3)	4.20 → 4.47	0.821 · epigenetic / RNA methyltransferase (Tier 3)
#10 · THP1 · ZEL031	bb0 = BB_8117313079 N = 70, 4 mw	<i>TOP1</i> topoisomerase I	4.50 (1.3e-4)	5.49 → 2.32	0.587 · Daunorubicin, DNA-damage pathway (Tier 3)

Activity cliffs without structures: active analogs versus inactive controls

Each dot is a measured compound's resemblance (Z_{mutual}) to the named CRISPR knockout response. Active analogs share the core block; inactive controls differ by a single substitution.

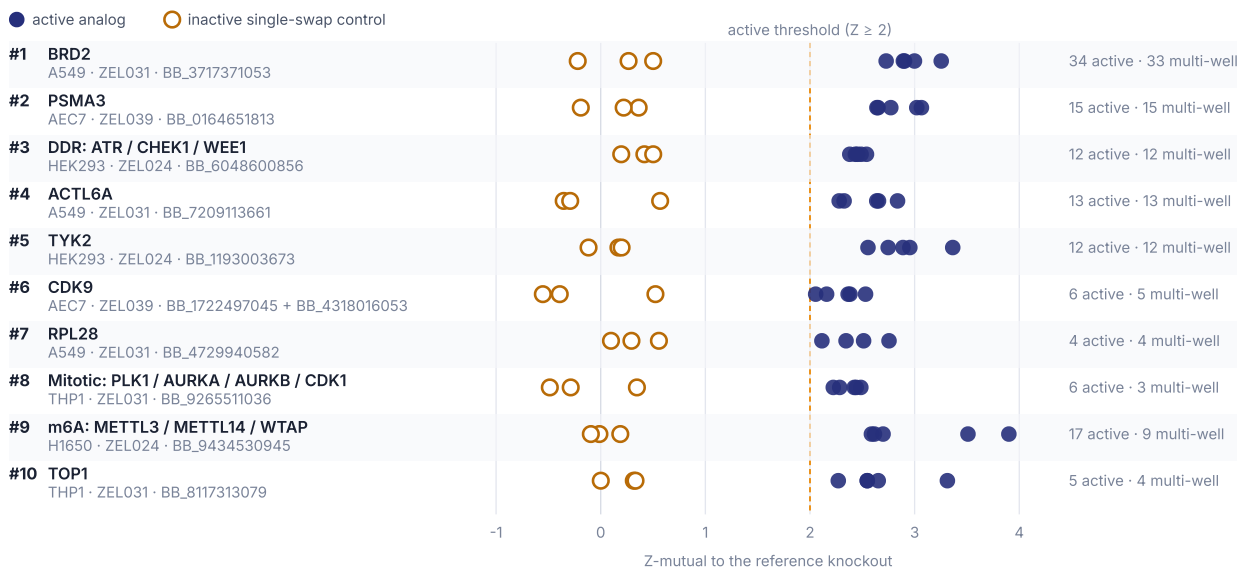


Figure 11. Activity cliffs without structures. For each prioritized genetic-twin program, the measured resemblance (Z_{mutual}) of the top active analogs (filled) and of single-substitution inactive controls (hollow) to the named knockout response. Substituting one partner block while holding the core block fixed toggles the phenomimicry on or off. Structures are withheld; compound identifiers, wells and devices are in evidence/genetic_members.csv.

6 · HIGH-CONFIDENCE HYPOTHESES FROM THE PILOT DATASET

Fourteen mechanism hypotheses from twenty convergent series

The two discovery tracks contribute 20 series. Grouped by mechanism they describe 14 distinct hypotheses: ten with a precedented drug class or with none at all (Table 4), and four cytotoxic or stress-response mechanisms that the engine recovers but that are not ranked. Three mechanisms are reached by more than one series, and the BET core BB_3717371053 is found by both tracks. These are high-confidence hypotheses from the pilot dataset, offered for collaborators to test. They are not Zafrens pipeline programs, and their order is not a statement of which targets Zafrens is advancing.

6.1 Ranked mechanism hypotheses

Hypotheses are ordered by convergent evidence (independent series, knockout rank in the 1,225-target panel, orthogonal channels passed, multi-well replication) and then by how crowded the precedented drug class is. The order is editorial; the statistics behind every row are in Tables 3 and 4 and in the released tables chemical_leads.csv and genetic_leads.csv. The last four rows are cytotoxic or stress-response mechanisms that the engine recovers without being told; they show calibration rather than opportunity and are not ranked.

Table 4. Ranked mechanism hypotheses from the pilot dataset. Ranks 1–10 group the twenty series by mechanism; “Experiment” is the assay that would test the hypothesis, not a planned Zafrens study.

Rank	Mechanism hypothesis · therapeutic context	Series · context · core block	Convergent evidence	Experiment that would test it
1	METTL3 / METTL14 / WTAP m⁶A writer AML and solid tumours; one clinical compound in the class	CRISPR-SAR #9 · H1650 · ZEL024 · BB_9434530945 (17 active, 9 multi-well, 54 cliff controls); supporting series ZS-SAR-06 · H1650 · ZEL028 · BB_8596154155 (15 active, 74 cliffs)	m ⁶ A-writer complex #3 of 1,225 targets; partner METTL14 Z 5.63; regioisomer cliff 0.82. ZS-SAR-06 mimics GSK3368715; STM2457 control Z 3.58; EZH2 KO Z 4.28	METTL3/14 MTase-Glo; m ⁶ A/LC-MS/MS
2	BRD2 / BRD4 BET bromodomain haematological malignancies; crowded precedented class	CRISPR-SAR #1 · A549 · ZEL031 · BB_3717371053 (34 active, 33 multi-well, 19 cliffs); ZS-SAR-04 · THP1 · ZEL031 · same core (5 active, 3 multi-well)	BRD2 KO #2 of 1,225, Z 16.4; partner ELL2; AZD5153 mimic with MZ1 control Z 4.10; one-carbon linker cliff 0.87; the only mechanism found by both tracks	BRD2/4 BD1/BD2 TR-FRET; MYC repression
3	CDK9 / P-TEFb transcriptional kinase MYC- and MCL1-dependent leukaemia and lymphoma	ZS-SAR-01 · A549 · ZEL028 (15 active, 47 cliffs); ZS-SAR-07 · AEC7 · ZEL039 (5 active); CRISPR-SAR #6 · AEC7 · ZEL039 (6 active, 5 multi-well)	ronidociclib and LY2857785 mimics; palbociclib control Z 5.14 and 4.16; CDK9 KO Z 5.27 and 6.70; SUPT5H partner Z 7.85; three series in two cell lines	CDK9/cyclin T1 NanoBRET; p-RNAPII Ser2
4	TYK2 / STAT5B oral immunology; the only non-oncology hypothesis in the set	CRISPR-SAR #5 · HEK293 · ZEL024 · BB_1193003673 with linker BB_5290020531 (12 active, 12 multi-well, 34 cliffs)	TYK2 KO #6; STAT5B partner Z 2.91; three orthogonal channels; cliff 0.80	TYK2 JH1/JH2 TR-FRET; p-STAT1/5 flow
5	AR / ER axis with the p23/HSP90 co-chaperone breast and prostate cancer; resistance mechanism rather than another antagonist	ZS-SAR-02 · A549 · ZEL031 · BB_3103020399 (14 active, 13 multi-well); ZS-SAR-08 · A549 · ZEL031 · BB_5574910360 (17 active, 16 multi-well)	endoxifen and aianthone mimics; endoxifen control Z 6.40 and 6.70; AR KO Z 6.43 and 6.83; best multi-well replication among the chemical twins	ERα/AR PolarScreen; p23/HSP90 client stability; FKBP5/TFF1 qPCR
6	RAF1 / MEK MAPK signalling KRAS- and BRAF-driven tumours; crowded precedented class	ZS-SAR-09 and ZS-SAR-10 · H1650 · ZEL024 (17 and 23 active; 17 and 20 multi-well)	ZM 336372 mimic; ZM336372 and sorafenib controls; MAP2K1 KO Z 3.99; RAF1 KO Z 4.49	RAF1/MEK1 TR-FRET; p-ERK1/2 AlphaLISA
7	ACTL6A (BAF53A) / SWI-SNF chromatin remodelling ACTL6A-amplified squamous cancers; no precedented drug class	CRISPR-SAR #4 · A549 · ZEL031 · BB_7209113661 (13 active, 13 multi-well, 18 cliffs)	ACTL6A KO #1 of 1,225, Z 11.9; cliff 0.87; partner ARID1A weak (Z 1.09), pathway-level support	SWI/SNF remodelling ATPase; ATAC-seq
8	NUAK1 / ROCK cytoskeletal kinase LKB1- and KRAS-mutant tumours, fibrosis; tool compounds only in the class	ZS-SAR-05 · H1650 · ZEL024 · BB_5272889191 (29 active, 11 multi-well, 33 cliffs)	HTH-01-015 mimic and control Z 2.08; NUAK1 KO Z 3.11; cliff 0.90	NUAK1/ROCK1 radiometric IC ₅₀ ; p-MYPT1
9	ALK / MET / IGF1R receptor tyrosine kinase fusion-driven NSCLC; precedented class	ZS-SAR-03 · H1650 · ZEL028 · BB_3467217614 (11 active, no multi-well active, 87 cliffs)	ceritinib mimic; AZD-3463 control Z 6.26; ALK KO Z 3.56; rescreen before any target assay	dose-response rescreen, then ALK/IGF1R ADP-Glo
10	ATR / CHK1 / WEE1 checkpoint DDR-deficient tumours; crowded precedented class	CRISPR-SAR #3 · HEK293 · ZEL024 · BB_6048600856 (12 active, 12 multi-well, 74 cliffs)	DDR complex #6; Z 11.1; ATR partner Z 7.94; cliff 0.91	ATR/CHK1 kinase IC ₅₀ ; p-CHK1 Ser345, γ-H2AX
—	26S proteasome PSMA3 / PSMB5 recovered, not ranked · established cytotoxic class (bortezomib, ixazomib)	CRISPR-SAR #2 · AEC7 · ZEL039 (15 active, 15 multi-well)	PSMA3 KO #2, Z 10.7; PSMB5 partner Z 7.94	20S chymotrypsin-like activity
—	Ribosomal stress → ATF4 (RPL28) recovered, not ranked · stress response at the ribosome	CRISPR-SAR #7 · A549 · ZEL031 (4 active, 4 multi-well)	RPL28 KO #23; ATF4 partner Z 1.81	OP-Puro translation; p-eIF2α / ATF4

Rank	Mechanism hypothesis · therapeutic context	Series · context · core block	Convergent evidence	Experiment that would test it
—	Mitotic spindle PLK1 / AURKA / CDK1 recovered, not ranked · generic antimitotic	CRISPR-SAR #8 · THP1 · ZEL031 (6 active, 3 multi-well)	complex #11; CDK1 partner Z 5.04; cliff 0.98	mitotic kinase panel; p- histone H3
—	TOP1 recovered, not ranked · payload-class cytotoxic mechanism	CRISPR-SAR #10 · THP1 · ZEL031 (5 active, 4 multi-well)	TOP1 KO #10; one orthogonal channel; cliff 0.59	TOP1 relaxation; ICE cleavage complex

6.2 Replication status

All ten genetic-twin series have at least three multi-well active analogs and pass the lineage-score gate. The chemical-twin series vary: ZS-SAR-01, -03 and -06 have no multi-well active analog and ZS-SAR-07 has one, so those four series should be rescreened before any target assay. Exported maximum active–cliff Tanimoto similarities span 0.449–0.917 for the chemical-twin series and 0.587–0.976 for the genetic-twin series; higher values mean the inactive control is a closer structural neighbor of the active compound and therefore a sharper cliff.

Table 5. Replication status of the twenty series behind the hypotheses. “Active” counts series members with $Z \geq 2$; “multi-well active” counts those measured in more than one nanowell. Source: [evidence/chemical_leads.csv](#), [evidence/genetic_leads.csv](#).

Series	Track	Context · reference	Family size	Active	Multi-well active	Max cliff Tanimoto	Status
ZS-SAR-01	chemical twin	A549 · ZEL028 Roniciclib	132	15	0	0.570	rescreen first
ZS-SAR-02	chemical twin	A549 · ZEL031 Endoxifen	83	14	13	0.481	ready for target assay
ZS-SAR-03	chemical twin	H1650 · ZEL028 Ceritinib dihydrochloride	213	11	0	0.584	rescreen first
ZS-SAR-04	chemical twin	THP1 · ZEL031 AZD5153	87	5	3	0.701	ready for target assay
ZS-SAR-05	chemical twin	H1650 · ZEL024 HTH-01-015	131	29	11	0.901	ready for target assay
ZS-SAR-06	chemical twin	H1650 · ZEL028 GSK3368715	223	15	0	0.576	rescreen first
ZS-SAR-07	chemical twin	AEC7 · ZEL039 LY2857785	37	5	1	0.907	rescreen first
ZS-SAR-08	chemical twin	A549 · ZEL031 Ailanthone	88	17	16	0.449	ready for target assay
ZS-SAR-09	chemical twin	H1650 · ZEL024 ZM 336372	76	17	17	0.917	ready for target assay
ZS-SAR-10	chemical twin	H1650 · ZEL024 ZM 336372	136	23	20	0.791	ready for target assay
CRISPR-SAR #1	genetic twin	A549 · ZEL031 BRD2	87	34	33	0.873	ready for target assay
CRISPR-SAR #2	genetic twin	AEC7 · ZEL039 PSMA3	51	15	15	0.875	ready for target assay
CRISPR-SAR #3	genetic twin	HEK293 · ZEL024 DDR / ReplicationStress / ATR / CHEK1 / WEE1	167	12	12	0.912	ready for target assay
CRISPR-SAR #4	genetic twin	A549 · ZEL031 ACTL6A	87	13	13	0.868	ready for target assay
CRISPR-SAR #5	genetic twin	HEK293 · ZEL024 TYK2	84	12	12	0.800	ready for target assay
CRISPR-SAR #6	genetic twin	AEC7 · ZEL039 CDK9	35	6	5	0.808	ready for target assay
CRISPR-SAR #7	genetic twin	A549 · ZEL031 RPL28	82	4	4	0.847	ready for target assay

Series	Track	Context · reference	Family size	Active	Multi-well active	Max cliff Tanimoto	Status
CRISPR-SAR #8	genetic twin	THP1 · ZEL031 MitoticSpindle / PLK1 / AURKA / AURKB / CDK1	112	6	3	0.976	ready for target assay
CRISPR-SAR #9	genetic twin	H1650 · ZEL024 m6A / Writer / METTL3 / METTL14 / WTAP	136	17	9	0.821	ready for target assay
CRISPR-SAR #10	genetic twin	THP1 · ZEL031 TOP1	70	5	4	0.587	ready for target assay

6.3 Experiments that would test each hypothesis

Transcriptional phenomimicry generates mechanism hypotheses; it does not prove binding. These experiments are described so that collaborators can test the hypotheses; they are not a Zafrens development plan. For each hypothesis the sequence is: fresh dose-response measurement of active and inactive-cliff compounds side by side; a direct target assay matched to the mechanism axis (for example TR-FRET or NanoBRET binding, ADP-Glo or radiometric kinase activity, proteasome chymotrypsin-like activity, or MTase-Glo methyltransferase activity); a proximal cellular readout (for example phospho-flow or western blot of the pathway node, MYC repression for the BET hypothesis, or m⁶A/A quantification for the METTL3 hypothesis); and a genetic perturbation experiment in which knockout or rescue of the nominated target shifts the compound's transcriptional response. A measured false-positive rate requires known experimental positives and negatives and is a deliverable of these experiments, not of this paper.

7 · DEFINITIONS, SCOPE BOUNDARIES AND SOURCES

Reading the evidence with the intended precision

- **Pilot dataset versus platform.** Counts in this paper (162,914 compounds, 190,699 profiles, 627 building blocks) describe the released pilot package. They are not platform limits.
- **Phenomimicry versus target engagement.** Transcriptional phenomimicry measures alignment of cellular state directions. It is a calibrated mechanism-of-action hypothesis generator, not proof of direct physical binding, enzymatic IC₅₀ or target selectivity.
- **Evaluation units and overlapping series.** A three-slot molecule belongs to up to three one-slot series and three two-slot series, so overlapping series are not independent samples. Headline counts report both total qualifying series and strictly disjoint program clusters (178 disjoint CRISPR-SAR clusters; 10 prioritized entries per modality, with overlap between modalities).
- **Calibration metrics and affine R^2 .** Internal R^2 on `ze1024_hek293` (0.493) incorporates a post-hoc linear affine calibration fitted on validation predictions.
- **Selected-set statistics.** The 84.9% lineage-support fraction is computed after selection with an OR gate that includes holdout support. The CIGS 0.822 median is a matched-panel rank on 17 qualified controls. Neither is a prospective replication rate.
- **Public edition.** Chemical structure images and SMILES/scaffold fields are withheld. Tables retain compound identifiers and measured results. Download hashes refer to the sanitized files; source hashes are recorded separately in `evidence/provenance.json`.
- **Data availability.** The pilot dataset is published on Hugging Face at huggingface.co/datasets/Zafrens/zscreen-pilot; the evidence tables below are derived from that release. An interactive Functional Drug Atlas covering a portion of the dataset is available at <http://32.190.137.56/>.

Evidence register

Table 6. Source tables released with this paper. Each file is linked from the portal's evidence register with its SHA-256 hash and source path.

File	What it supports
platform_counts.json	Pilot-dataset scale: compounds, profiles, contexts, programs, gene panel
fold0_reference_model_summary.csv	Exact per-context simple-model and transformer results (Table 1)
cold_start_metrics.csv	Unseen-building-block head-to-head on 7,544 HEK293 compounds
replay_savings.csv	Retrospective active-learning measurement savings
activeseq_treatment_results.txt	ActiveSeq treatment-recognition and population-pooling results
cigs_qualified.csv , cigs_calibration.csv	CIGS qualified-control panel metrics and full-atlas calibration
chemical_leads.csv , chemical_qualified.csv	The ten highest-ranked chemical-twin series and the 110 qualifying matches
crispr_calibration.csv	Compound-level and evaluation-row CRISPR calibration
genetic_leads.csv , genetic_members.csv	The ten highest-ranked genetic-twin series and their measured active and inactive analogs
provenance.json	Source paths and SHA-256 hashes for every released table