ActiveSeq pairs same-well imaging and transcriptomics in a Z-Screen microchip to recover compound-resolved phenotypes

Abstract

High-throughput phenotypic discovery needs readouts that connect chemical identity to cell state without forcing a choice between scalable imaging and deeper molecular measurement. We evaluated ActiveSeq, the same-well image-plus-RNA arm of Z-Screen, in which cells in a one-bead-one-compound micro-well are imaged and then sequenced. Across two HEK293 pilots (ACT010 and ACT011), we analyzed 20,222 matched wells: 11,511 wells from a 35-compound control panel and 8,711 ZEL026 dummy-bead wells used only as negative-control QC. The image arm used an early 32-dimensional latent derived from generic image features and calibrated on this control-panel domain, so it tests a first assay representation rather than the full potential of raw multi-channel or target-aware imaging. Even with this early representation, both modalities recovered compound identity within and across runs: split-half top-1 retrieval reached 98.8% / 80.1% for RNA and 94.2% / 86.8% for image, and cross-run centroid transfer reached 71.4% for RNA and 85.7% for image. Image and RNA compound maps were aligned but non-identical, and merged image-plus-RNA features improved held-out classification to 55.5% balanced accuracy versus 41.4% for RNA and 29.2% for image. Same-compound shuffled fusion matched true same-well fusion on this task, while exploratory analyses detected modest above-random within-compound pairing structure. Low-pass RNA remained interpretable through compound-by-run pseudobulk signatures. These results establish ActiveSeq as a scalable same-micro-well multimodal readout linking morphology, RNA state, and compound provenance.

Significance

Most multimodal perturbation resources pair image and RNA profiles by treatment label across separate cell populations. ActiveSeq instead measures both readouts from the same small cell population in the same Z-Screen micro-well. In this pilot, a generic image-derived latent and low-pass RNA profile each recover compound-resolved information, and their merged representation improves held-out compound classification across a 35-compound control panel. The present image latent is an early representation, not the endpoint of the imaging strategy. The larger capability is that morphology, assay state, target-aware image channels, RNA state, and compound provenance can be linked by well in a scalable one-bead-one-compound screen.

Introduction

Phenotypic drug discovery is most useful when a screen can see both breadth and depth. Breadth is needed to explore many chemical structures; depth is needed to understand whether a response is reproducible, interpretable, and worth following. High-content imaging and Cell Painting make large phenotypic screens practical by measuring morphology and other visual features at high throughput [1]. Single-cell transcriptomic perturbation screens measure cell-state programs more directly, but usually at greater cost per perturbation and lower compound concurrency [3]. Recent public multimodal resources such as scGeneScope connect these readouts by profiling matched treatments across replicate populations, pairing Cell Painting and single-cell RNA sequencing profiles for the same treatment labels in U2-OS cells [4].

Z-Screen is organized around a different experimental unit: a high-density one-bead-one-compound microchip. ActiveSeq is the Z-Screen arm that reads image and RNA from the same micro-well. A small local cell population is imaged, the same well is then sequenced, and the resulting readouts can be linked to the well and its compound assignment. This design matters because it lets image-derived morphology and RNA state be interpreted as paired measurements from the same physical assay unit rather than as treatment-matched averages from separate wells.

The image arm should be understood positively but precisely. The present study does not ask whether a generic image embedding is the final imaging solution for Z-Screen. It asks whether an early image-derived representation, paired with low-pass RNA from the same micro-well, already carries reproducible compound information. Generic morphology can report assay state, colony state, quality-control structure, and low-cost triage signal. Future target-aware fluorescent or photoreactive image channels can be selected to report biology closer to the question being asked, while the same well still contributes RNA for molecular interpretation.

We tested ActiveSeq in two HEK293 pilot experiments, ACT010 and ACT011. The dataset contains 20,222 matched wells with both image and RNA measurements. The compound-response analyses focus on 11,511 matched wells assigned to a 35-compound control panel. A separate set of 8,711 matched wells labeled ZEL026 is used only as a dummy-bead negative-control QC condition; those wells support read-depth and detected-gene checks and are excluded from compound-response claims.

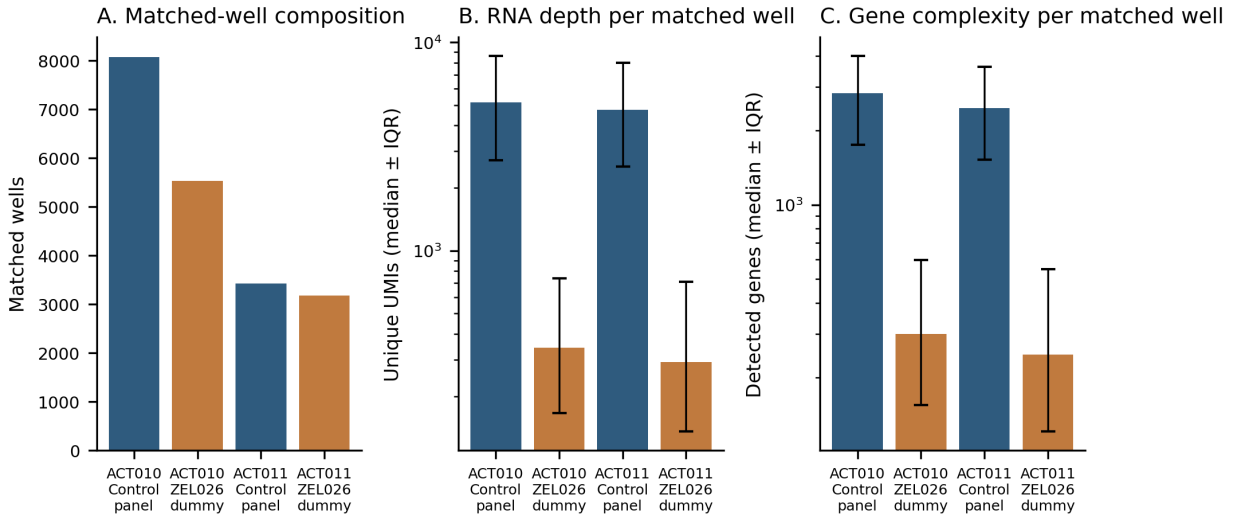
The analysis follows the decision path a scalable same-well workflow must pass. First, are the paired wells analyzable and is the image representation defined? Second, does each readout recover compound identity within and across experiments? Third, do image and RNA organize compounds in related but non-identical ways? Fourth, does merging the same-well readouts improve held-out compound classification? Fifth, how much exact well-to-well cross-modal structure is detectable after compound identity is held constant? Finally, does the low-pass RNA branch remain interpretable enough for triage?

Results

Are the same-well ActiveSeq readouts analyzable?

ActiveSeq produced 20,222 image-plus-RNA matched wells across ACT010 and ACT011. Of these, 11,511 wells were assigned to the 35-compound control panel used for compound-response analyses. The remaining 8,711 ZEL026 dummy-bead wells were used only as negative-control QC stress-test wells.

The QC contrast behaved as expected. Relative to control-panel wells, ZEL026 dummy-bead wells showed sharply reduced RNA recovery: median unique molecular identifiers (UMIs) decreased from 5,167 to 342 in ACT010 and from 4,750 to 293 in ACT011. Median detected genes decreased from 2,826 to 301 in ACT010 and from 2,463 to 248 in ACT011. This separation confirms that the same-well RNA branch detects the intended low-recovery QC condition while keeping ZEL026 outside compound-response interpretation.



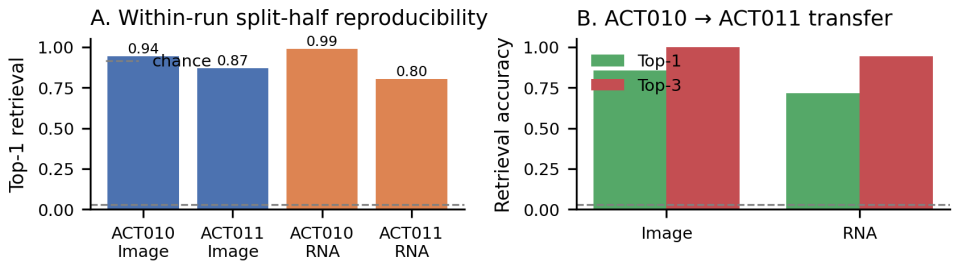
Supplementary Figure 1. QC overview for matched control-panel wells and ZEL026 dummy-bead negative-control QC stress-test wells in ACT010 and ACT011. ZEL026 wells are used only to show low RNA recovery in the QC stress-test condition and are excluded from compound-response analyses.

For the image branch, each well had an embedding derived from a frozen generic vision model. We analyzed a 32-dimensional calibrated image latent learned on the matched control-panel domain. The calibration objective retained compound-discriminating image information while reducing experiment-of-origin signal. This makes the image latent an early internal assay representation: it is useful for asking whether same-well imaging contributes compound-resolved information in this pilot, while leaving broader-library, cell-type, and imaging-channel generalization for larger studies.

Does each readout recover compound identity within and across experiments?

We first asked whether image and RNA independently encode reproducible compound information. Within each experiment, wells for each control compound were split into pseudo-replicates, and one half was used to retrieve the other among the 35 compounds. RNA latent space was highly stable, with split-half top-1 retrieval of 98.8% in ACT010 and 80.1% in ACT011. The calibrated image latent was also strong, with split-half top-1 retrieval of 94.2% in ACT010 and 86.8% in ACT011. Top-3 retrieval reached 100% for RNA in ACT010 and 99.6% for image.

Cross-run centroid transfer provided a stricter test. Compound centroids learned in ACT010 retrieved the same compounds in ACT011 at 71.4% top-1 and 94.3% top-3 for RNA, and at 85.7% top-1 and 100% top-3 for the calibrated image latent. Chance top-1 retrieval is 2.9%. Thus, both branches recover compound identity well above chance across independent ActiveSeq runs, and the image branch is not merely a QC accessory; even this early representation carries reproducible compound-level signal.


A. Within-run split-half reproducibility
Top-1 retrieval
1.00
0.75
0.50
0.25
0.00
0.94
chance 0.87
0.99
0.80
ACT010 Image
ACT011 Image
ACT010 RNA
ACT011 RNA
B. ACT010 → ACT011 transfer
Retrieval accuracy
Top-1
Top-3
Image
RNA

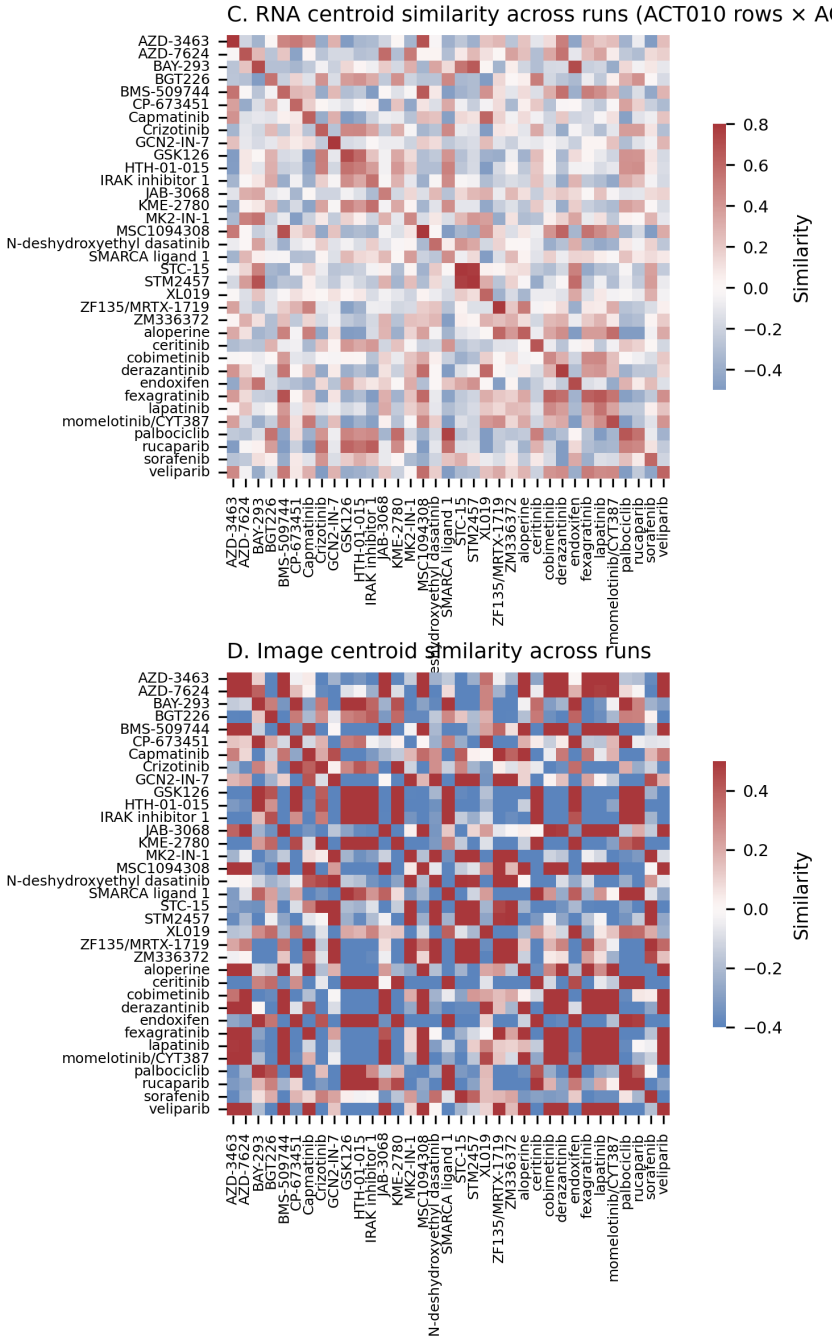

C. RNA centroid similarity across runs (ACT010 rows × ACT011 cols)
AZD-3463
AZD-7624
BAY-293
BGT226
BMS-509744
CP-673451
Capmatinib
Crizotinib
GCN2-IN-7
GSK126
HTH-01-015
IRAK inhibitor 1
JAB-3068
KME-2780
MK2-IN-1
MSC1094308
N-deshydroxyethyl dasatinib
SMARCA ligand 1
STC-15
STM2457
XL019
ZF135/MRTX-1719
ZM336372
aloperine
ceritinib
cobimetinib
derazantinib
endoxifen
fexagratinib
lapatinib
momelotinib/CYT387
palbociclib
rucaparib
sorafenib
veliparib
Similarity
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
D. Image centroid similarity across runs
Similarity
0.4
0.2
0.0
−0.2
−0.4

Figure 1. Compound-level reproducibility and cross-run centroid transfer in the 35-compound control panel. Both modalities are reproducible within run, and both transfer well above chance from ACT010 to ACT011. These metrics establish compound-level signal, not the incremental value of exact same-well pairing.

Do image and RNA encode related but non-identical compound maps?

A useful paired readout should not simply duplicate the same information twice. We therefore compared pairwise compound-distance matrices in image and RNA space within each run. The maps were aligned but distinct: Spearman correlations between image and RNA compound-distance matrices were 0.718 in ACT010 and 0.607 in ACT011. The strongest perturbations also recurred across experiments, with cross-run effect-magnitude rankings correlated at Spearman 0.799 for image and 0.662 for RNA.

Several compounds, including GCN2-IN-7, KME-2780, BGT226, ZF135 / MRTX-1719, and CP-673451, appeared among the stronger responses in both modalities in at least one run. The important result is the geometry, not any single example: same-well image and RNA measurements organize the control panel in related but non-identical ways. This is the pattern expected from complementary readouts of the same perturbation state.

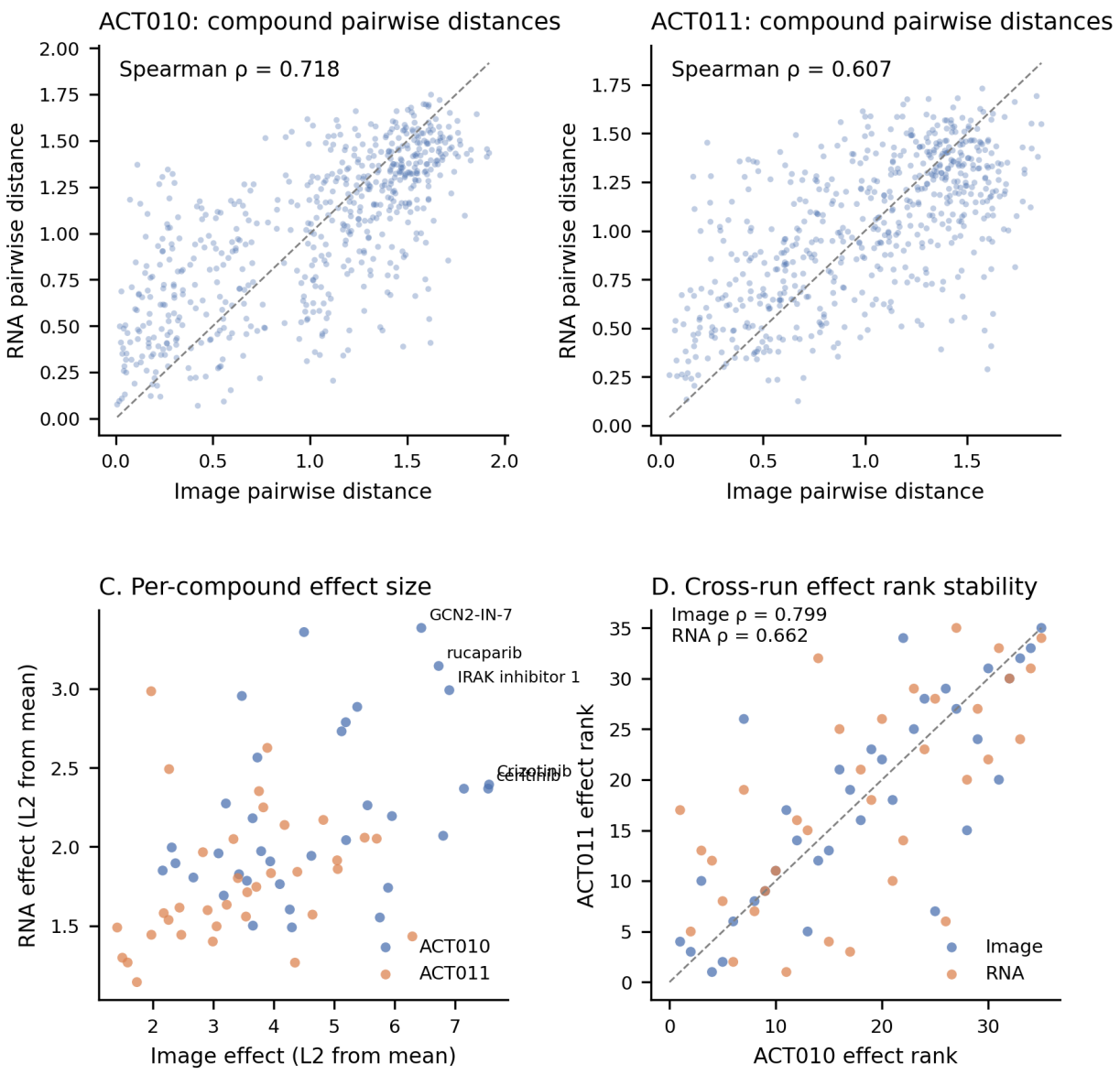


Figure 2. Image and RNA preserve a shared but non-identical compound geometry within run, and effect-size rankings are stable across runs. Pairwise compound-distance correlations are 0.718 in ACT010 and 0.607 in ACT011; cross-run per-compound effect magnitudes correlate at Spearman 0.799 for image and 0.662 for RNA.

Does combining same-well image and RNA improve held-out classification?

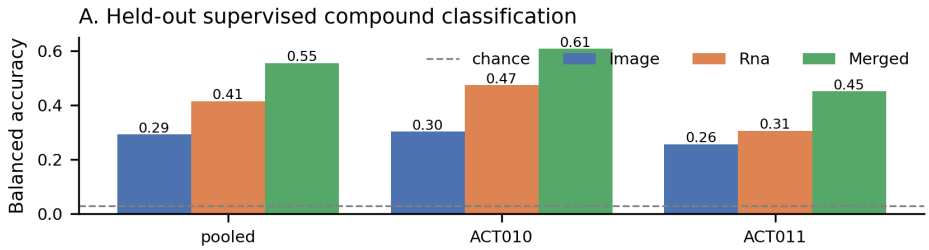
We next asked whether paired readouts help a decision model classify held-out wells. Each matched well was represented by one calibrated image vector, one RNA vector, or their concatenation. A linear discriminant model was trained on image alone, RNA alone, or merged image-plus-RNA features and evaluated on held-out wells.

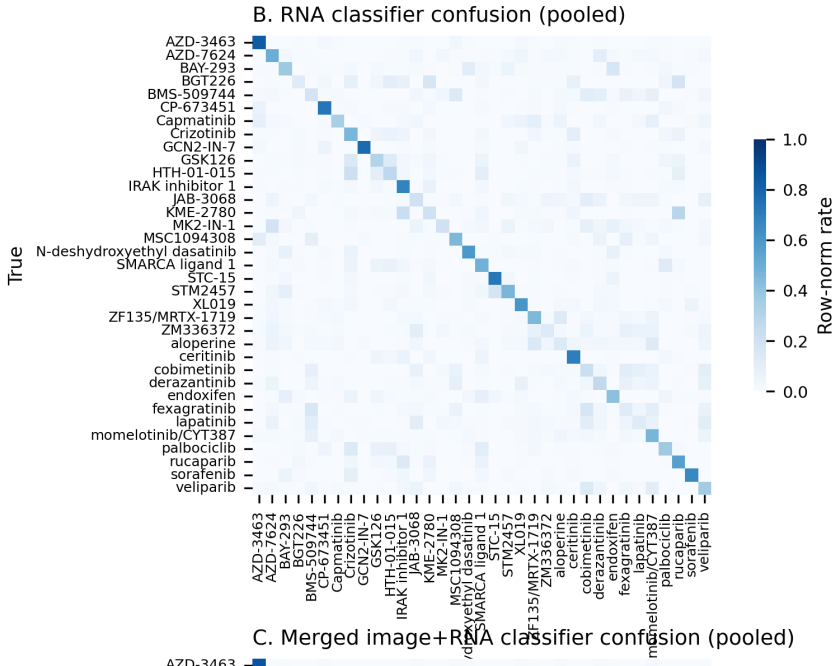
Merged features outperformed either single modality. In the pooled analysis, balanced accuracy

increased to 55.5% for merged features, compared with 41.4% for RNA alone and 29.2% for image alone. Within ACT010, merged features reached 60.8% balanced accuracy versus 47.3% for RNA and 30.3% for image. Within ACT011, merged features reached 45.1% versus 30.5% for RNA and 25.6% for image. Overall accuracy followed the same pattern, rising to 54.9% pooled, 60.2% in ACT010, and 45.5% in ACT011.

The gain was broad across the panel. Merged features improved per-compound recall for 35 of 35 compounds in the pooled analysis, 33 of 35 in ACT010, and 30 of 35 in ACT011. Confusion matrices show that the merged representation reduces many near-confusions rather than relying on a few easy compounds.

We then tested whether the classification gain required exact image/RNA pair identity. In a same-compound unpaired-fusion control, image and RNA vectors were shuffled among wells sharing the same compound and experiment before fitting the same classifier. This shuffled control matched the true same-well merged result: across 100 shuffles, pooled balanced accuracy averaged 55.4% (95% interval 54.7% to 56.1%), compared with 55.5% for true same-well fusion. ACT010 and ACT011 showed the same pattern. The classification result therefore supports a strong compound-level conclusion: same-well ActiveSeq produces complementary image and RNA features that improve held-out compound classification. On this label task, the added value comes from combining modality-level compound information, not from exact well-to-well pairing alone.


A. Held-out supervised compound classification
Balanced accuracy
0.0
0.2
0.4
0.6
chance
Image
Rna
Merged
0.29
0.41
0.55
0.30
0.47
0.61
0.26
0.31
0.45
pooled
ACT010
ACT011


B. RNA classifier confusion (pooled)
True
AZD-3463
AZD-7624
BAY-293
BGT226
BMS-509744
CP-673451
Capmatinib
Crizotinib
GCN2-IN-7
GSK126
HTH-01-015
IRAK inhibitor 1
JAB-3068
KME-2780
MK2-IN-1
MSC1094308
N-deshydroxyethyl dasatinib
SMARCA ligand 1
STC-15
STM2457
XL019
ZF135/MRTX-1719
ZM336372
aloperine
ceritinib
cobimetinib
derazantinib
endoxifen
fexagratinib
lapatinib
momelotinib/CYT387
palbociclib
rucaparib
sorafenib
veliparib
Row-norm rate
1.0
0.8
0.6
0.4
0.2
0.0

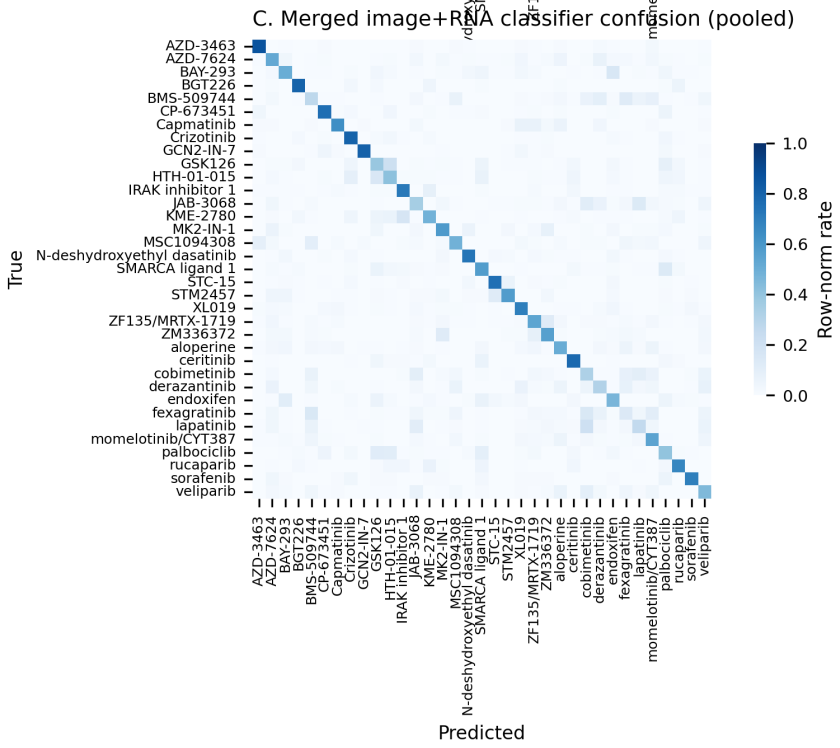

C. Merged image+RNA classifier confusion (pooled)
True
Predicted
Row-norm rate
1.0
0.8
0.6
0.4
0.2
0.0

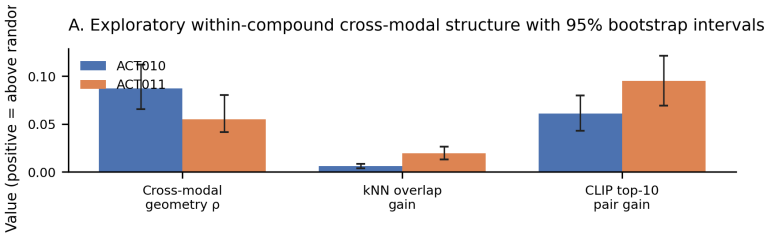
Figure 3. Same-well multimodal classification on held-out wells. Merged image-plus-RNA features outperform RNA alone in pooled and per-run analyses, and per-compound confusion matrices show the gain is from reduced class confusion across most compounds rather than from a few easy classes.

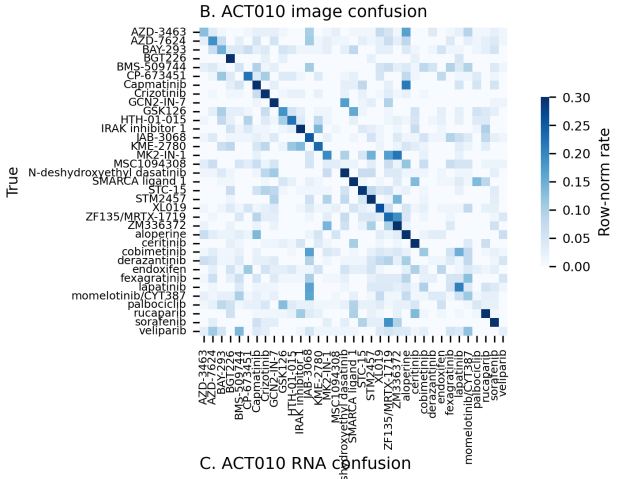
Is exact same-well structure detectable within compounds?

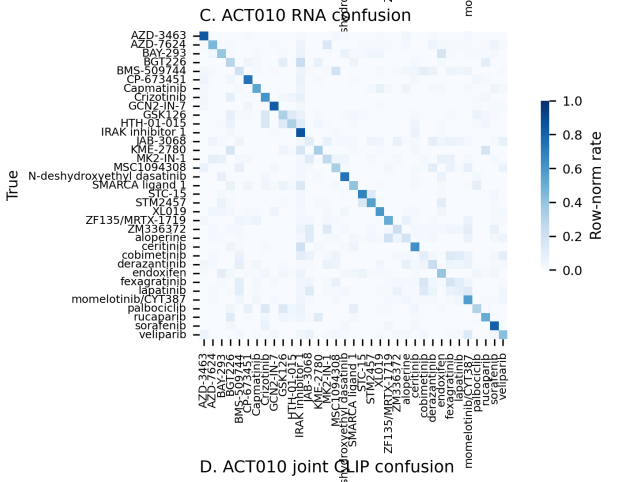
The previous analyses establish compound-level multimodal value. We next asked the stricter same-well question: after compound identity is held fixed, do image and RNA preserve matching well-to-well structure? This is a high bar because each well contains only a small local cell population, and many sources of biological and technical variation are not expected to align perfectly across modalities.

In a held-out centroid-retrieval task, image alone reached 25.1% / 21.7% top-1 across ACT010 / ACT011, RNA alone reached 44.8% / 24.3%, and naive image-plus-RNA concatenation reached 49.1% / 33.7%, exceeding RNA alone in both runs. A lightweight contrastive image-RNA pairing objective on top of the calibrated latent did not improve retrieval further, suggesting that the current representation already captures most of the readily available compound-identity information.

Exact pair retrieval and within-compound geometry showed small but consistent signal. Top-10 exact image-to-RNA retrieval after contrastive training was 2.1% in ACT010 and 2.5% in ACT011, compared with random expectations of 0.4% and 1.0%. Within each compound, image and RNA distance structure was positive in 34 of 35 compounds in both runs, with median cross-modal distance correlations of 0.087 in ACT010 and 0.055 in ACT011. Mean k-nearest-neighbor overlap gains were 0.006 and 0.020. These effects are modest in absolute size, but they indicate that same-well structure is measurable in the pilot. Larger panels, richer image channels, and perturbations designed to produce within-compound response heterogeneity should provide a clearer test of when exact pairing becomes operationally important.


A. Exploratory within-compound cross-modal structure with 95% bootstrap intervals
Value (positive = above randor
ACT010
ACT011
0.10
0.05
0.00
Cross-modal geometry ρ
kNN overlap gain
CLIP top-10 pair gain


B. ACT010 image confusion
True
Row-norm rate
0.30
0.25
0.20
0.15
0.10
0.05
0.00


C. ACT010 RNA confusion
True
Row-norm rate
1.0
0.8
0.6
0.4
0.2
0.0

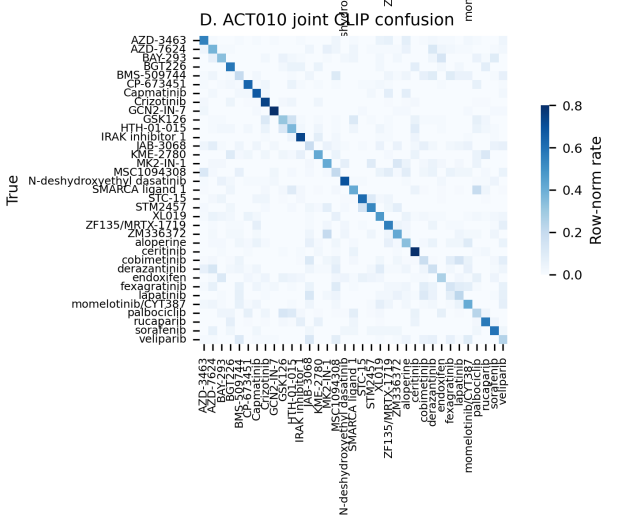

D. ACT010 joint CLIP confusion
True
Predicted
Row-norm rate
0.8
0.6
0.4
0.2
0.0

Figure 4. Exploratory same-well structure beyond supervised classification. Within-compound cross-modal agreement is usually above random but modest in magnitude. These data support the possibility of same-well response-state information, while leaving the incremental value of exact pairing to a dedicated control analysis.

Is the low-pass RNA branch interpretable for triage?

Classification alone is not enough for discovery; a useful RNA branch should also yield interpretable signatures. We aggregated raw counts from matched control wells into per-compound, per-run pseudobulk signatures, converted them to log-counts per million (log-CPM), and centered each run by subtracting its experiment-wide mean. This uses the compound-by-run pseudobulk as the robust unit at the present depth.

Centered signatures retrieved the correct compound across runs at 48.6% top-1 and 68.6% top-3. Representative compounds carried reproducible ranked gene signatures, including palbociclib (ARC, RASD1, CDKN1C, EGR1), BAY-293 (MT1X, MT1F, MT2A, SLC30A1), rucaparib (ITPRIP, SLC37A4, NACC2, COL4A1), crizotinib (SORCS2, ZNF34, AC253572.2), and GSK126 (DUSP10, SPATA6L, CCN2). These signatures are best interpreted as triage features rather than validated mechanisms in HEK293. They show that low-pass ActiveSeq RNA can cluster hits, flag recurring expression programs, and nominate follow-up assays while remaining paired to same-well imaging.

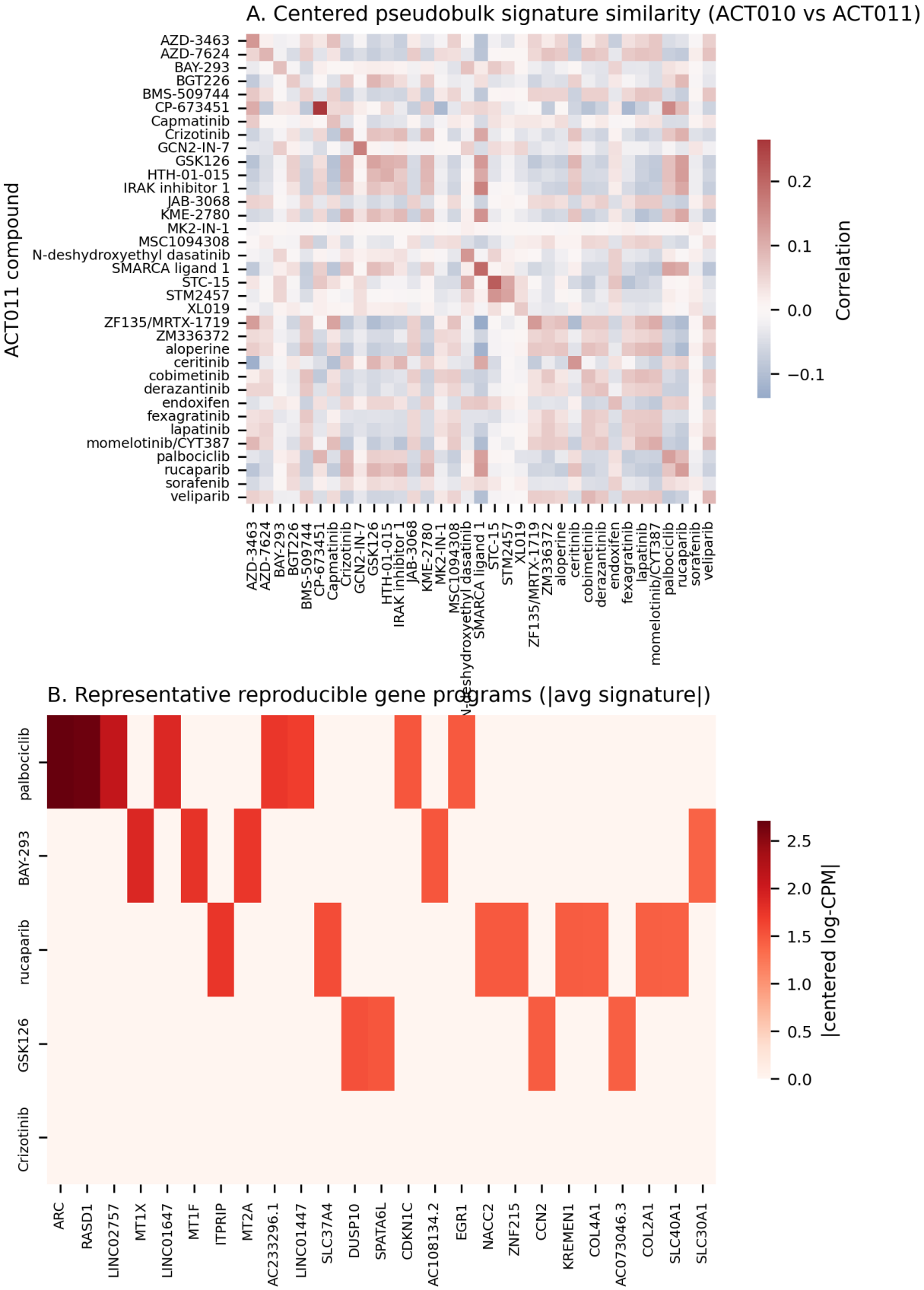

A. Centered pseudobulk signature similarity (ACT010 vs ACT011)
ACT011 compound
AZD-3463
AZD-7624
BAY-293
BGT226
BMS-509744
CP-673451
Capmatinib
Crizotinib
GCN2-IN-7
GSK126
HTH-01-015
IRAK inhibitor 1
JAB-3068
KME-2780
MK2-IN-1
MSC1094308
N-deshydroxyethyl dasatinib
SMARCA ligand 1
STC-15
STM2457
XL019
ZF135/MRTX-1719
ZM336372
aloperine
ceritinib
cobimetinib
derazantinib
endoxifen
fexagratinib
lapatinib
momelotinib/CYT387
palbociclib
rucaparib
sorafenib
veliparib
Correlation
0.2
0.1
0.0
−0.1
B. Representative reproducible gene programs (|avg signature|)
palbociclib
BAY-293
rucaparib
GSK126
Crizotinib
ARC
RASD1
LINC02757
MT1X
LINC01647
MT1F
ITPRIP
MT2A
AC233296.1
LINC01447
SLC37A4
DUSP10
SPATA6L
CDKN1C
AC108134.2
EGR1
NACC2
ZNF215
CCN2
KREMEN1
COL4A1
AC073046.3
COL2A1
SLC40A1
SLC30A1
|centered log-CPM|
2.5
2.0
1.5
1.0
0.5
0.0

Figure 5. Centered pseudobulk gene signatures preserve compound-specific ranked expression patterns across runs and recover the correct compound at 48.6% top-1 and 68.6% top-3 retrieval. The gene lists support RNA interpretability and triage rather than standalone mechanism calls.

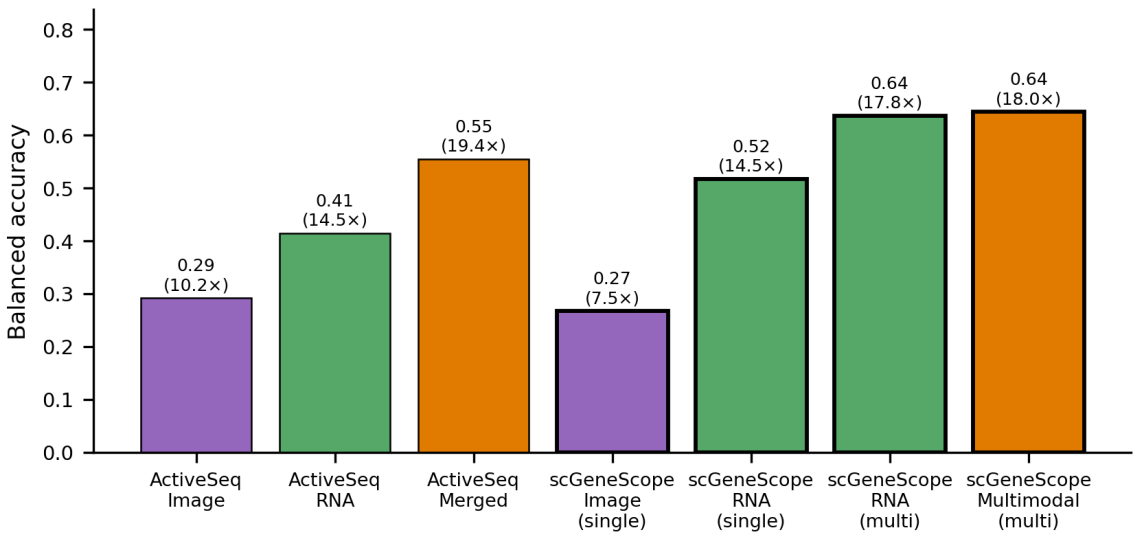
What context does scGeneScope provide?

scGeneScope is the closest public reference point for treatment-matched image and RNA perturbation modeling, but it is not a direct head-to-head benchmark. scGeneScope profiles approximately 716,767 single-cell images and 627,704 single-cell RNA profiles across 28 treatments in U2-OS cells, matching treatments across replicate populations [4]. ActiveSeq profiles 11,511 control-panel matched wells across two HEK293 experiments, each well measured as a 5- to 10-cell pseudobulk, with image and RNA read from the same micro-well and with a simpler early image representation.

The comparison is still useful for magnitude context. ActiveSeq trails the strongest scGeneScope multimodal reference in raw within-experiment balanced accuracy, consistent with the differences in scale, cell line, and imaging richness. After normalizing by random-chance label performance, pooled ActiveSeq RNA reaches 14.5 times chance, similar to the scGeneScope single-profile RNA value of 14.5 times chance. Pooled ActiveSeq image reaches 10.2 times chance, compared with 7.5 times for the scGeneScope single-profile image value. Pooled ActiveSeq merged features reach 19.4 times chance, compared with 18.0 times for the scGeneScope multimodal multiprofile reference.

The correct conclusion is that the same-well ActiveSeq pilot sits in the same broad performance range after chance normalization while measuring a different experimental unit. This supports scale-up and richer imaging rather than a claim of benchmark superiority.

A. Within-experiment compound classification (ActiveSeq filled, scGeneScope outlined)



B. Context for interpreting the gap

Aspect	**ActiveSeq (Z-Screen)**	**scGeneScope**
Single-cell scale	~60-110k cells (5-10 × 11.5k wells)	~717k images, ~628k RNA
Label space	35 compounds	28 treatments
Pairing	Same matched micro-well	Treatment-matched replicate
Imaging	Brightfield + calibrated latent	5-ch Cell Painting + ViT
Pooled image	29.2% (10.2× chance)	26.7% single (7.5×)
Pooled RNA	41.4% (14.5× chance)	51.7% single (14.5×)
Pooled multimodal	55.5% (19.4×)	64.4% multi (18.0×)
Best ActiveSeq run	ACT010 merged 60.8% (21.3×)	n/a
Takeaway	Smaller pilot; same range after chance no	Larger public benchmark on raw accuracy

Figure 6. Within-experiment compound classification for ActiveSeq versus the published scGeneScope reference, annotated with fold over random-chance performance, plus a context table summarizing the design differences that make the two benchmarks complementary rather than directly interchangeable.

Discussion

ActiveSeq establishes that the Z-Screen micro-well can support paired image and RNA readouts from the same small cell population. Across two HEK293 pilot experiments, both branches recovered compound identity within and across runs, image and RNA compound maps were aligned but non-identical, and merged features improved held-out compound classification across nearly the entire 35-compound panel. The same-compound shuffled-fusion control clarifies the mechanism of that gain: the current supervised classification improvement is driven by complementary compound-level image and RNA information rather than by exact well-to-well pair identity.

This is a positive result for scalable multimodal screening. A same-well workflow does not have to prove large exact-pair effects in a small control-panel classification task to be valuable. It first needs to show that both measurements can be collected from the same micro-well, that each readout is reproducible, that their compound maps are related but not redundant, and that their combination improves decisions. ActiveSeq passes those tests in this pilot.

The imaging result is especially important for the next version of the platform. The 32-dimensional latent used here is an early representation derived from generic image features and calibrated on the same control-panel domain. It already contributes reproducible compound information and improves classification when combined with RNA. At the same time, it should not define the ceiling for imaging in Z-Screen. Raw multi-channel images, assay-state features, colony morphology, and target-aware fluorescent or photoreactive channels can all be paired with the same RNA branch in future ActiveSeq screens. Target-aware channels are likely to be more informative for specific biological questions than a generic latent alone because they can report selected pathway, localization, abundance, binding, or activity states directly.

The RNA branch gives the paired screen interpretability. Low-pass per-well RNA is noisy, but compound-by-run pseudobulk signatures retrieved compounds across experiments and produced ranked gene signatures suitable for triage. These signatures do not establish mechanisms by themselves. They provide a way to cluster hits, recognize recurring cell-state programs, and choose follow-up assays while retaining the same-well connection to imaging.

The pilot also defines the next experiments. Broader control panels can test whether multimodal gains hold across more chemical and phenotypic diversity. New libraries can test how image calibration generalizes beyond the panel used here. Additional cell types can determine which compound-response programs are shared or context-specific. Image channels selected for target-relevant biology can ask whether exact same-well pairing becomes stronger when the image readout is designed to match the biological question. These are scale-up questions, not reasons to discount the present result.

In summary, ActiveSeq is the same-well image-plus-RNA arm of Z-Screen. In this first HEK293 pilot, it links morphology-derived signal, low-pass RNA state, and compound assignment in the same micro-well; both modalities recover compound-resolved phenotypes; and their combination improves held-out classification. The strategic value is the experimental unit: as Z-Screen scales, the same well can carry chemistry provenance, image-derived phenotypes, target-aware optical

measurements, and transcriptomic state in one linked assay record.

Methods

Dataset

The analysis uses the canonical Z-Screen ActiveSeq workspace under `data/ZScreen_Canonical_Dataset/`. Each ActiveSeq well is annotated with an experiment id (ACT010 or ACT011), a global well id, a compound assignment, an image embedding, and an RNA latent. Gene-level pseudobulk counts come from the matched H5AD object that backs the same-well RNA pipeline, repaired as documented in `DATA_REPAIR_REPORT.md`. The scGeneScope comparison uses published benchmark values transcribed from OpenReview Table 1 into `paper1/tables/scgenescope_reference.csv`.

Per-well image embedding calibration

The original per-well image embedding was produced by a frozen pretrained vision model on raw imaging tiles. That generic representation captures visual structure but does not guarantee that compound response is aligned with the principal directions of variation. Experiment-of-origin, position, microscopy, and colony-state effects can all contribute to the raw geometry.

A 32-dimensional control-panel-calibrated latent was learned from the original embedding by a lightweight domain-adversarial network trained on the matched control panel. The objective rewards compound discrimination and penalizes experiment-of-origin discrimination, encouraging compound-discriminating variation without retraining a vision model from scratch. Calibration was learned on the same control-panel domain studied here, so the calibrated-latent results should be read as an internal representation for this pilot. Future ActiveSeq studies can pair RNA with richer raw image representations and target-aware image channels.

Matched-well preprocessing

Imaging detections were aggregated by experiment_id and gwid so each matched well contributed one image vector. Wells with an assigned control name were treated as the 35-compound panel. ZEL026 wells were treated as dummy-bead negative-control QC stress-test wells and were excluded from compound-response analyses.

Reproducibility and transfer analyses

Within-run reproducibility was measured by repeated split-half pseudo-replication within each compound. Cosine similarity between pseudo-replicate centroids was computed after standardization within each matrix. Cross-run transfer used compound centroids learned in ACT010 and queried in ACT011.

Cross-modal geometry

For each run, pairwise distances among compound centroids were computed separately in image and RNA space. Agreement between distance matrices was summarized with Spearman correlation over

upper-triangular entries, which does not assume the two scales are matched. Cross-run effect-size stability used Spearman correlation over per-compound effect magnitudes.

Supervised multimodal classification

Held-out classification used a 50:50 compound-stratified split, either pooled across runs or fit within each run. Linear discriminant analysis with the eigen solver and automatic shrinkage was trained on calibrated image features alone, RNA features alone, or their raw concatenation. Performance was summarized with overall accuracy, balanced accuracy, macro F1, and per-compound recall.

To test whether the supervised gain depended on exact image/RNA pair identity, a same-compound unpaired-fusion control shuffled RNA vectors among wells sharing the same compound and experiment before fitting the same merged classifier. The control used 100 shuffle iterations with the same train/test split.

Same-well representation learning

A lightweight dual encoder with a symmetric contrastive image-RNA pairing objective was trained on same-well pairs from the training split, evaluated by nearest-centroid compound retrieval and by exact held-out pair retrieval. Within-compound analyses compared image and RNA distance structure among wells assigned to the same compound and summarized distance-correlation, k-nearest-neighbor overlap, and same-compound pair-retrieval gains over random baselines. These analyses were treated as exploratory because the absolute effects were small in the pilot.

Gene-level pseudobulk analysis

Raw counts were summed per experiment and control compound, converted to log-counts per million (log-CPM), and centered by subtracting the experiment-wide mean signature. Cross-run retrieval used correlation similarity between centered signatures. Per-cell differential expression was not performed because each well averages a 5- to 10-cell population; pseudobulk plus centering is the unit at which the signal is robust at the present depth.

External benchmark context

Within-run supervised metrics from `supervised_multimodal_metrics.csv` were compared to scGeneScope within-experiment values transcribed from OpenReview Table 1 into `scgenescope_reference.csv`. Because the label spaces differ (35 versus 28), the comparison reports both raw balanced accuracy and fold over random-chance performance.

Limitations

This pilot uses one cell line (HEK293), 35 control compounds, and two matched experiments. The calibrated image latent was learned on the same control panel studied here, so broader libraries, additional cell lines, and richer image channels are needed to evaluate generalization. The supervised fusion result shows additive value for combined image and RNA readouts, while the same-compound

shuffled-fusion control indicates that exact well-to-well pair identity does not add measurable classification value on this compound-label task. Exact same-well cross-modal retrieval and within-compound structure are above random but modest in absolute terms. Gene-level signatures are summarized through pseudobulk aggregation rather than full differential-expression modeling with biological replicates. These limits define the next scale-up studies while preserving the central result: ActiveSeq supports reproducible same-micro-well multimodal profiling with interpretable RNA and useful image-plus-RNA compound classification.

Data availability

All data tables, derived latent representations, manuscript figures, and analysis inputs needed to reproduce this manuscript are organized in this repository under `paper1/` and `data/ZScreen_Canonical_Dataset/`. A persistent-archive deposition with an assigned DOI will accompany the corresponding preprint posting.

Code availability

Analysis and figure-generation scripts are in `paper1/scripts/`. Package-level dependencies are in the repository root `requirements.txt`. Tested with Python 3.14.3 on Windows.

References

1. Bray MA, Singh S, Han H, et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc. 2016;11:1757-1774. doi:10.1038/nprot.2016.105
2. Lopez R, Regier J, Cole MB, et al. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15:1053-1058. doi:10.1038/s41592-018-0229-2
3. Dixit A, Parnas O, Li B, et al. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell. 2016;167(7):1853-1866.e17. doi:10.1016/j.cell.2016.11.038
4. Dapello J, Nassar M, Eksi R, et al. scGeneScope: A treatment-matched single-cell imaging and transcriptomics dataset and benchmark for treatment response modeling. NeurIPS 2025 Datasets and Benchmarks. OpenReview. https://openreview.net/pdf/f7d541dae38bcf88a79789a4c6440aadfec123c7

Z-Screen maps combinatorial chemistry provenance to functional transcriptomic state

Abstract

Z-Screen is a high-throughput platform for linking one-bead-one-compound combinatorial chemistry to cellular RNA state. The public Z-Screen bundle is one of the largest public combinatorial chemistry transcriptomic datasets, containing 615,793 repaired RNA profiles, 615,721 valid scVI latent profiles, 12 combinatorial libraries, 4 cell lines, 142,187 unique hashed compounds with chemistry embeddings, explicit building-block and tuple provenance, and supporting imaging-derived features. We ask whether this provenance defines a functional chemistry-to-RNA map. Named controls were reproducible across A549, H1650, HEK293, and THP1 (median split-half cosine 0.968 to 0.993), and Z-Screen signatures were concordant with LINCS L1000 where compounds and cell lines overlapped (11 of 11 positive; median Spearman rho 0.490). Building-block and pair models predicted held-out tuple states, led by ZEL024 / HEK293 (median cosine 0.674 versus 0.397 baseline; 33.2% error reduction) and ZEL031 / THP1 (0.833 versus 0.783; 21.3% error reduction). Observed tuples recovered calibrated reference-state neighborhoods, including HEK293 ZF-104 and THP1 sorafenib, MZ1, and STC-15 phenomimics. Imaging-derived features added modest complementary signal in additive benchmarks. Z-Screen is not larger than Tahoe-100M by cell count; its contribution is chemistry-resolved functional readout tied to combinatorial grammar.

Significance

Z-Screen is designed so that each measured phenotype remains connected to the chemistry that made it. In this manuscript, a building block is a recorded chemical component at a defined library position, a tuple is the ordered set of building blocks specifying a library compound, a phenomimic is an observed tuple whose measured cell state resembles a named reference state after same-cell-line calibration, a rank signature is an ordered gene-level perturbation profile used for cross-platform comparison, and a multimodal readout links RNA with imaging-derived features from the same chemistry or well subset. With that language made explicit, the main result is practical: the public Z-Screen resource already contains enough reproducible RNA signal to predict held-out tuple responses, identify chemistry-resolved neighborhoods around known controls, and show that building-block effects can partly recur across cell lines.

Introduction

Early discovery screens need scale, biological resolution, and a usable description of the chemistry being tested. Scale is needed because most compounds are inactive or uninformative in a given cell context. Biological resolution is needed because a hit list is difficult to act on without information about the state each hit produced. Chemistry provenance is needed because discovery becomes more powerful when a measured phenotype can be traced back to the building-block choices that generated it.

Public perturbation resources have established several reference axes for the field. LINCS L1000 mapped small-molecule transcriptional responses at large scale using a reduced 978-gene representation and reported 1.3 million profiles [2]. Tahoe-100M is a larger modern single-cell drug-perturbation atlas, with over 100 million profiles across 50 cancer cell lines and roughly 1,100 to 1,200 drug perturbations [3]. Imaging assays such as Cell Painting provide high-throughput morphology [5], pooled single-cell perturbation studies show the value of scalable RNA readouts [6], and multimodal resources such as scGeneScope show that aligned imaging and RNA can improve treatment-response modeling when both readouts are available for matched perturbations [4].

Z-Screen occupies a different part of this landscape. It is not larger than Tahoe-100M by cell count, and it is not presented as a replacement for LINCS L1000, Cell Painting, or single-cell perturbation atlases. Its differentiating feature is combinatorial chemistry resolution: one-bead-one-compound libraries are profiled in 50,000-well microchips, and each low-pass RNA profile remains linked to compound identity, ordered building-block tuple, library context, and hashed chemistry embedding. The public bundle contains 615,793 repaired RNA profiles, 615,721 valid scVI latent profiles, 12 combinatorial libraries, 4 cell lines, 142,187 unique hashed compounds with chemistry embeddings, and supporting imaging-derived features. This makes Z-Screen one of the largest public combinatorial chemistry transcriptomic datasets with explicit chemistry provenance.

The gap addressed here is therefore not simply another transcriptomic atlas. It is the connection between transcriptomic state and combinatorial chemistry grammar. Each microwell contains approximately 5 to 10 cells and is linked to both a compound identity and a low-pass RNA readout. Because the libraries are combinatorial, each compound can be represented as an ordered tuple of building blocks. A building block is a recorded chemical component at a defined library position, such as bb0 or bb3. A tuple is the ordered list of building blocks that specifies a compound, with `NA` used where a library does not use all possible positions. Named controls are reference compounds with replicate wells; their measured RNA centroids provide reproducibility checks, comparison anchors, and candidate phenomimic neighborhoods.

The chemistry-to-RNA-map question is concrete: can a low-pass RNA screen recover a reproducible, chemistry-resolved relationship between building-block choices and transcriptomic state? Prediction, phenomimicry, and mechanism are kept separate throughout the paper. Prediction asks whether a model can estimate a held-out tuple's RNA state from its chemistry. Phenomimicry asks whether an observed tuple lands near a known compound's measured state after same-cell-line background calibration. Mechanism is a further biological hypothesis, requiring orthogonal evidence, and is not

assigned by cosine similarity alone.

The current manuscript evaluates eight tasks in order:

1. What is the public scale and field position of the Z-Screen chemistry-to-RNA resource?
2. Are named controls reproducible enough to support downstream modeling?
3. Do Z-Screen rank signatures show concordance with LINCS L1000 where the platforms overlap?
4. Do building-block identities and pairwise building-block terms predict held-out tuple RNA states?
5. Where does a structure-derived chemistry embedding add signal beyond building-block identity?
6. Do observed library tuples recover calibrated phenomimic neighborhoods around named reference states?
7. Do building-block effects partly recur across cell lines?
8. Does imaging add a complementary branch in same-resource additive benchmarks?

These tasks support a platform claim with defined boundaries. The strongest evidence comes from well-sampled systems, especially ZEL024 in HEK293, and the manuscript does not claim that structure-derived descriptors generalize across all libraries, that out-of-family chemical generalization is complete, or that phenomimic neighborhoods prove mechanism of action.

Results

Z-Screen links combinatorial chemistry provenance to RNA state at public-resource scale

The public Z-Screen bundle was assembled as a chemistry-to-RNA resource rather than as a compound list. It contains 615,793 repaired RNA profiles, 615,721 profiles with valid 32-dimensional scVI latent coordinates, 12 combinatorial libraries, 4 cell lines, and 142,187 unique hashed compounds with a 256-dimensional chemistry embedding. Each library compound is represented by explicit building-block and tuple annotations, allowing a measured RNA state to be traced back to the chemical grammar that produced it.

This scope places Z-Screen in a specific field position. LINCS L1000 remains a mature public small-molecule transcriptional catalog built around a reduced gene representation, and Tahoe-100M is much larger by single-cell profile count. Z-Screen's distinguishing axis is different: functional RNA readout at combinatorial-chemistry resolution, with supporting imaging-derived features where same-resource image linkage is available. The analyses below therefore ask whether the released data support a chemistry-to-RNA map: controls establish measurement credibility, building-block and structure-derived chemistry features predict held-out RNA states, observed tuples recover calibrated phenomimic neighborhoods around named controls, cross-cell comparisons test partial recurrence of building-block effects, and imaging provides a complementary phenotypic branch.

Control replicates establish reproducible RNA profiles

The first empirical question is whether the low-pass RNA readout is stable enough to support a chemistry-to-state analysis. To test this, wells containing the same named control compound were repeatedly split into pseudo-replicates, and cosine similarity was computed between the two scVI latent centroids for each split.

Control reproducibility was high in all four main cell lines. Median split-half cosine was 0.993 in A549, 0.991 in H1650, 0.993 in HEK293, and 0.968 in THP1. THP1 had the lowest median among these systems and also had sparser compound-control library coverage, so downstream THP1 results are interpreted with that sampling context in mind.

This result establishes a reproducible RNA measurement layer for downstream modeling. A model reaching median cosine 0.674 is operating well below the control-replicate ceiling, leaving enough headroom to distinguish model limitations from assay instability. The control panel also provides the reference states used later for phenomimic neighborhoods and for cross-platform comparison with LINCS L1000.

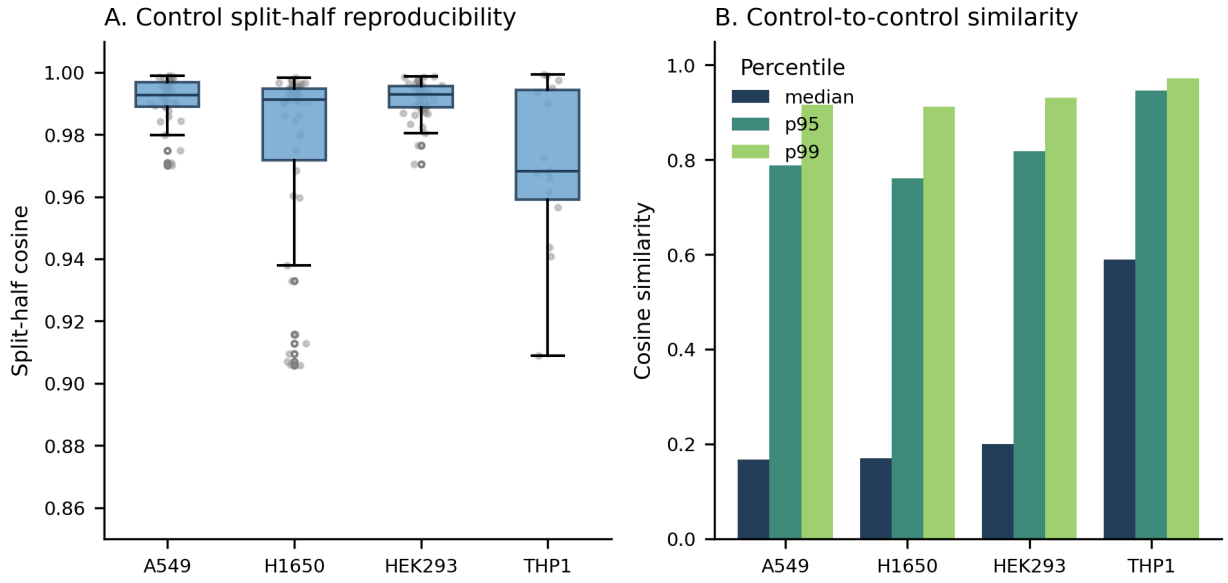


Figure 1. Named controls show high split-half reproducibility across all four cell lines, supporting downstream chemistry-to-RNA modeling.

LINCS L1000 concordance provides an external RNA check

We next asked whether Z-Screen named-control RNA signatures resemble an independent perturbational transcriptomics platform where compounds and cell lines overlap. This test used the Benchmarking module, which identified named-control overlap with the CMap LINCS 2020 L1000 build and compared rank signatures in the shared LINCS-visible gene universe.

Of 47 named Z-Screen control compounds, 14 (29.8%) matched LINCS L1000 compounds by name or alias. Three of four Z-Screen cell lines had LINCS sample coverage: HEK293, A549, and THP1;

H1650 was absent. After coverage filtering, 11 compound and cell-line pairs were compared, including 10 in A549 and 1 HEK293 pair through HEK293T mapping. All 11 matched pairs had positive Spearman correlation. The median rho was 0.490, the mean rho was 0.393, and 8 of 11 pairs exceeded rho 0.2.

Five A549 compounds reached empirical p <= 0.048 against a same-cell-line unmatched-compound permutation null: crizotinib (rho 0.679, p 0.048), ZM-336372 (rho 0.625, p 0.022), rucaparib (rho 0.530, p 0.023), sorafenib (rho 0.505, p 0.023), and veliparib (rho 0.490, p 0.046). Sorafenib in HEK293T mapped to HEK293 reached rho 0.597, and the HEK293 matched-versus-unmatched cell-line summary reached one-sided Mann-Whitney p = 0.009. Several compounds fell in the bulk of the unmatched distribution, including palbociclib, BGT226, and GSK126; the GSK126 result is consistent with slow transcriptional onset of EZH2 inhibition under the 6 h LINCS treatment window.

This benchmark does not claim that Z-Screen outperforms LINCS or matches its scale. It is an external RNA check: where the same named compounds and related cell lines overlap, low-pass Z-Screen rank signatures recover compound-specific transcriptional structure visible in an independent platform.

Building-block grammar predicts held-out tuple RNA states

The next question is whether the chemistry grammar of a library carries predictive information about RNA state. The test held out tuple identities, fit building-block models on the remaining tuple centroids, and compared predicted RNA state with the measured held-out state. The baseline was the library and cell-line centroid; the main models used additive building-block terms or additive terms plus pairwise building-block interactions.

The strongest result was ZEL024 in HEK293. A building-block plus pair model reached median cosine 0.674 against a centroid-baseline median cosine of 0.397 and reduced mean squared error by 33.2%. ZEL031 in THP1 also improved over baseline, reaching median cosine 0.833 versus 0.783 and reducing error by 21.3%. ZEL031 in A549 and ZEL024 in H1650 showed smaller positive gains with the best available building-block model.

The interpretation is prediction within a measured library grammar. The model learns that particular building blocks and building-block pairs are associated with recurrent RNA responses among tuples drawn from the same library vocabulary. This is the core chemistry-to-RNA map in the manuscript: compound provenance is not only recorded metadata, but a predictive coordinate system for functional cell state. Stricter prospective transfer to unseen building blocks, new scaffold families, or unrelated chemistry spaces requires separate held-chemistry and cross-library evaluation.

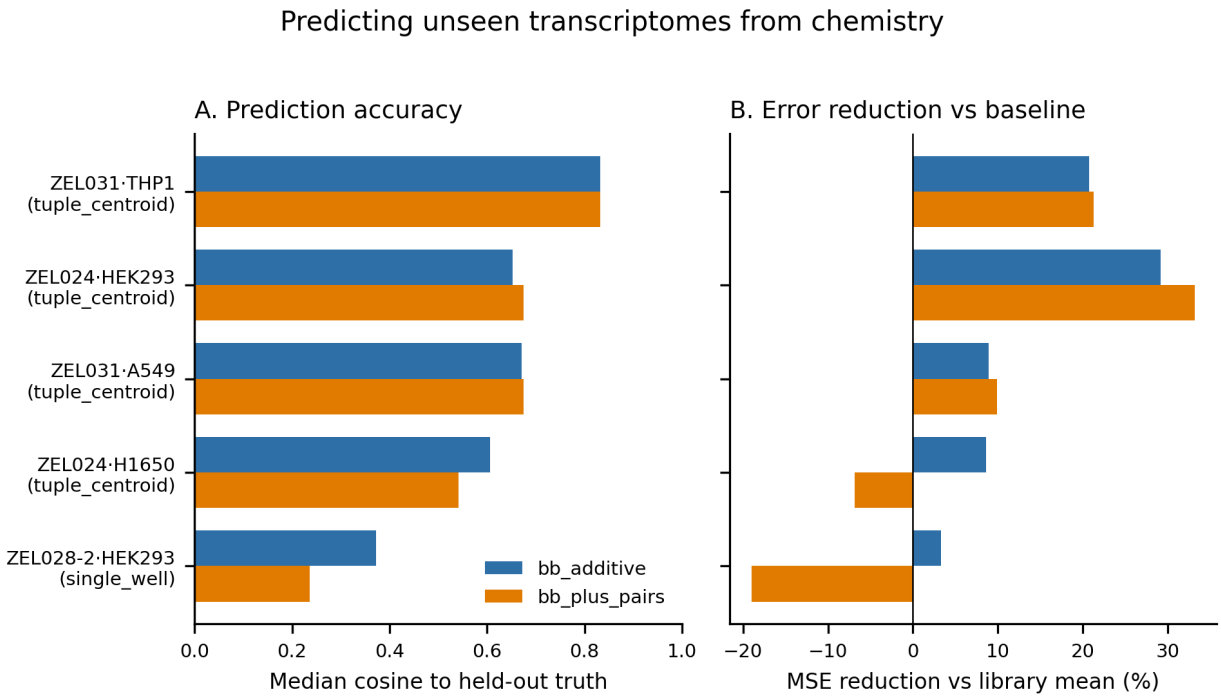


Figure 2. Building-block models improve prediction of held-out transcriptomic states across multiple library and cell-line systems, with the strongest gains in ZEL024 / HEK293 and ZEL031 / THP1.

Structure-derived chemistry features refine the map in the strongest sampled setting

We then asked whether information derived from full compound structure improves prediction beyond building-block identity. Because the public bundle does not expose raw Z-Screen SMILES, this benchmark used a fixed 256-dimensional chemistry embedding derived privately from ECFP4 fingerprints and joined through `smiles_hash` in the public package.

ZEL024 / HEK293 was the clearest positive case. In this system, the structure-derived embedding alone reached mean cosine 0.611 with 27.8% mean squared error reduction, building-block identity alone reached mean cosine 0.611 with 27.5% error reduction, and the combined building-block plus embedding model reached mean cosine 0.621 with 29.3% error reduction. This is the strongest sampled evidence that substructure-derived information adds to the explicit library grammar.

The effect was not general across the present pilot. In ZEL024 / H1650, ZEL031 / A549, and ZEL031 / THP1, adding the chemistry embedding to building-block identity did not improve mean error reduction over building blocks alone. In sparse ZEL028-2 / HEK293, all chemistry-only variants performed below the mean-RNA baseline. The conclusion is therefore specific and useful: building-block provenance is the most reliable design coordinate today, while structure-derived descriptors can add substructure signal when library coverage and cell context are favorable, with ZEL024 / HEK293 as the strongest sampled positive case.

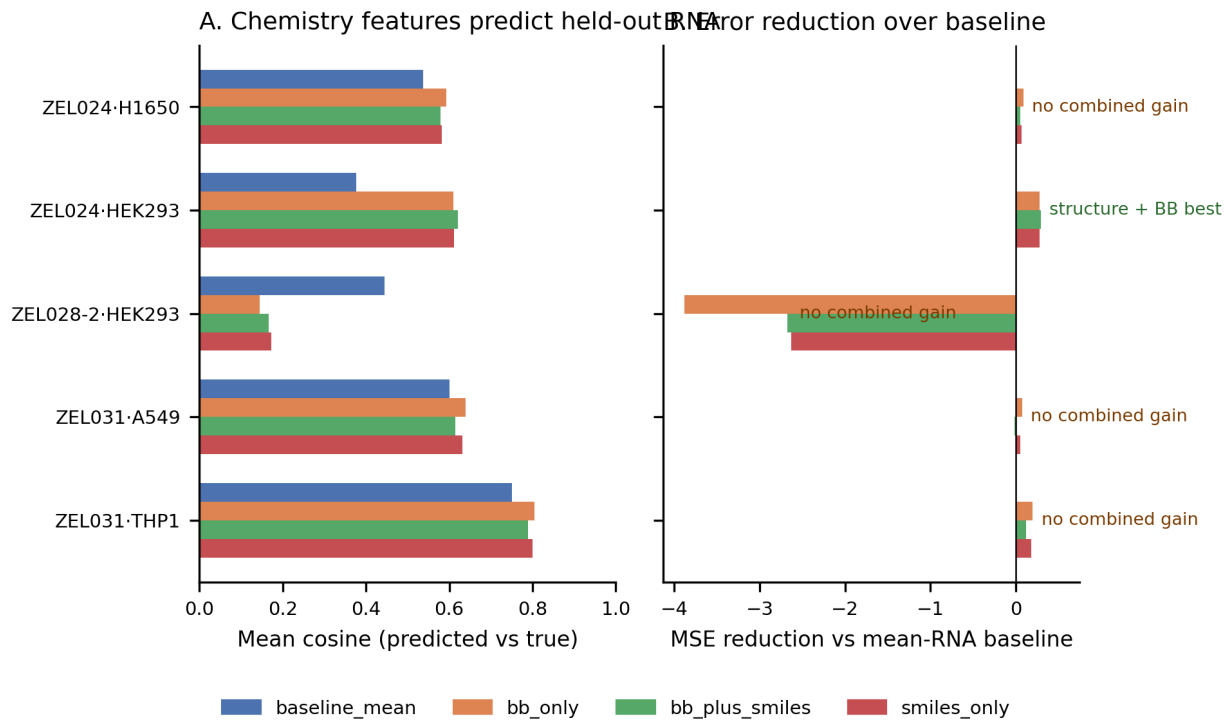


Figure 3. Chemistry-only models capture predictive signal unevenly across systems. Building-block provenance is the most reliable coordinate, and adding structure-derived descriptors improves performance most clearly in ZEL024 / HEK293 while failing to improve or hurting in weaker and sparser settings.

Library tuples recover calibrated phenomimic neighborhoods

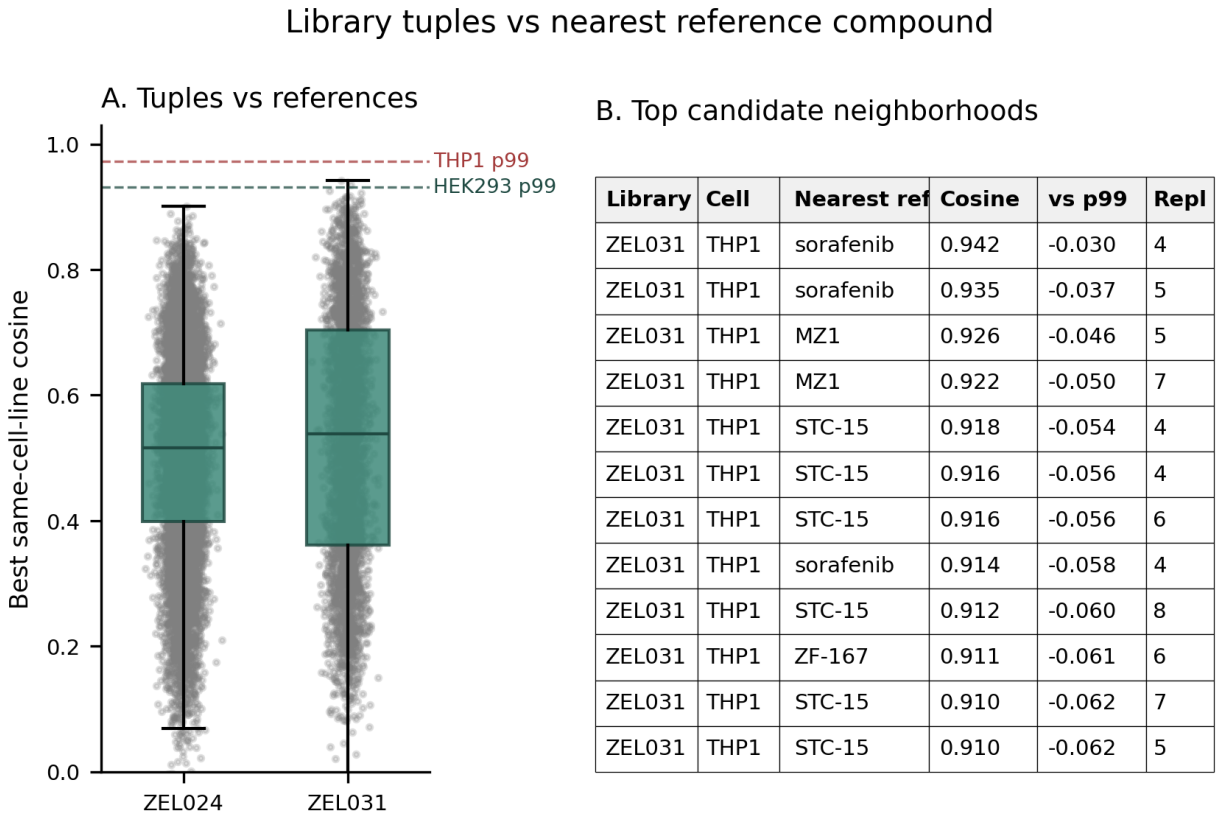
We next asked whether observed library tuples land near known compound states in a way that is useful for triage. The test compared tuple RNA centroids with named-control centroids within the same cell line, then calibrated each apparent match against the same-cell-line distribution of unrelated control-control similarities. This calibration is essential because a high cosine in one cell line may be ordinary in another.

The strongest sampled reference-neighborhood result was ZEL024 / HEK293. One tuple, `BB-0000031|BB-0000007|BB-0001916|BB-0000082-002|NA`, matched the internal reference ZF-104 at cosine 0.901. The HEK293 cross-control background had median cosine 0.200, 95th percentile 0.818, and 99th percentile 0.931. The ZF-104 neighborhood therefore sits far above the bulk of unrelated controls but remains below the most extreme control-control tail. Additional ZEL024 / HEK293 tuples reached cosine 0.85 to 0.89 against references including ZEL029-25, ZEL029-28, ZEL029-24, and ZF-104.

ZEL031 / THP1 also produced high-cosine neighborhoods, including tuples near sorafenib at cosine 0.942, MZ1 at 0.926, and STC-15 at 0.918. THP1, however, had a much higher cross-control background: median 0.589, 95th percentile 0.946, and 99th percentile 0.972. These THP1

phenomimics are plausible candidates for follow-up, but the higher background makes them less specific than the ZEL024 / HEK293 case.

The important result is not a single nearest-neighbor match. It is the repeated appearance of chemistry-resolved tuple neighborhoods around known reference states, showing that Z-Screen can turn a large combinatorial screen into a prioritized follow-up map. A phenomimic match says that an observed tuple produced an RNA state close to a named reference under the same cell-line measurement conditions. It nominates chemistry for follow-up target-engagement, dose-response, and orthogonal phenotyping assays; it does not prove that the tuple and reference share a target or mechanism.



Library	Cell	Nearest ref	Cosine	vs p99	Repl
ZEL031	THP1	sorafenib	0.942	-0.030	4
ZEL031	THP1	sorafenib	0.935	-0.037	5
ZEL031	THP1	MZ1	0.926	-0.046	5
ZEL031	THP1	MZ1	0.922	-0.050	7
ZEL031	THP1	STC-15	0.918	-0.054	4
ZEL031	THP1	STC-15	0.916	-0.056	4
ZEL031	THP1	STC-15	0.916	-0.056	6
ZEL031	THP1	sorafenib	0.914	-0.058	4
ZEL031	THP1	STC-15	0.912	-0.060	8
ZEL031	THP1	ZF-167	0.911	-0.061	6
ZEL031	THP1	STC-15	0.910	-0.062	7
ZEL031	THP1	STC-15	0.910	-0.062	5

Figure 4. Multiple library tuples land near transcriptomic neighborhoods centered on known reference compounds. These are calibrated candidate neighborhoods, not mechanism assignments. The strongest current case is HEK293 (ZF-104 at cosine 0.901, against a HEK293 99th-percentile cross-control background of 0.931); THP1 hits to sorafenib, MZ1, and STC-15 occur in a system with a higher cross-control baseline.

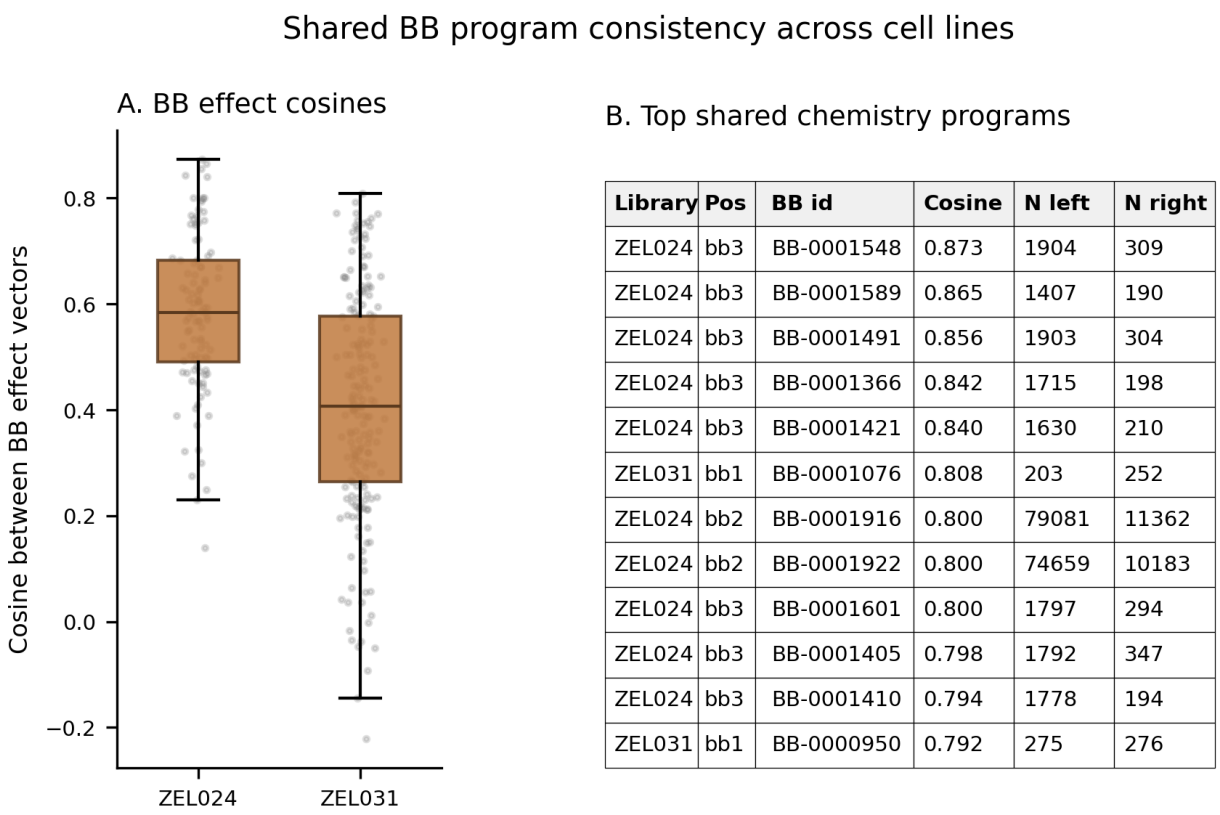
Building-block effects partly recur across cell lines

We then asked whether chemistry-associated RNA effects remain directionally similar when the cell line changes. The test estimated per-building-block effect vectors within paired library and cell-line systems, then compared matched building blocks across cell lines by cosine similarity.

Only chemistry-resolved positions are emphasized. In ZEL024, bb3 contains 83 building blocks and each bb3 effect averages over a median of 168 distinct tuples, making it the most interpretable position for cross-cell chemistry effects. In ZEL031, bb0 and bb1 form a two-position library; each single-position effect still averages over roughly 70 to 100 partner chemistries and is interpretable as a position-level building-block effect. By contrast, ZEL024 bb0, bb1, and bb2 have very few unique values in this architecture, so their averages collapse hundreds to thousands of tuples and are treated as diagnostics rather than headline chemistry-resolved results.

Across the chemistry-resolved positions, building-block effects partly recurred across cell lines. In ZEL024 bb3, the median HEK293-to-H1650 cross-cell cosine was 0.603 across 83 building blocks, with an interquartile range of 0.505 to 0.687 and a maximum of 0.873. In ZEL031, the A549-to-THP1 median was 0.355 at bb0 across 105 building blocks, with interquartile range 0.230 to 0.475 and maximum 0.747. At bb1, the median was 0.526 across 84 building blocks, with interquartile range 0.318 to 0.670 and maximum 0.808.

These values are below the within-cell reproducibility ceiling, so they are not evidence of perfect cell-line transfer. They do show that some building blocks carry partially conserved RNA directionality across contexts, which is the prerequisite for reusing chemistry lessons beyond a single assay system and for deciding where deeper profiling is likely to be informative. The lower-resolution ZEL024 bb0, bb1, and bb2 comparisons remain available as diagnostics in `paper2/tables/bb_effect_consistency.csv`.



Library	Pos	BB id	Cosine	N left	N right
ZEL024	bb3	BB-0001548	0.873	1904	309
ZEL024	bb3	BB-0001589	0.865	1407	190
ZEL024	bb3	BB-0001491	0.856	1903	304
ZEL024	bb3	BB-0001366	0.842	1715	198
ZEL024	bb3	BB-0001421	0.840	1630	210
ZEL031	bb1	BB-0001076	0.808	203	252
ZEL024	bb2	BB-0001916	0.800	79081	11362
ZEL024	bb2	BB-0001922	0.800	74659	10183
ZEL024	bb3	BB-0001601	0.800	1797	294
ZEL024	bb3	BB-0001405	0.798	1792	347
ZEL024	bb3	BB-0001410	0.794	1778	194
ZEL031	bb1	BB-0000950	0.792	275	276

Figure 5. Building-block effects partially recur across cell lines at chemistry-resolved positions, with positive median transfer and a subset of building blocks showing strong cross-context recurrence.

Imaging provides a complementary branch of the platform

Finally, we asked whether derived image features add signal in this RNA-first manuscript. ActiveSeq is the same-well image plus transcriptome workflow within Z-Screen, and the current paper2 public image features should be read as an early derived representation of that branch rather than as the full imaging opportunity. The test used simple additive classifiers on image features, RNA features, chemistry features, and concatenated feature sets in image-linked subsets. This framing was chosen because it makes each feature axis interpretable and avoids attributing gains from more complex multimodal models to a single modality without evidence.

In the ZEL024 control benchmark, image-only classification reached balanced accuracy 0.213, RNA-only reached 0.524, image plus RNA reached 0.540, and image plus RNA plus chemistry reached 0.611 across nine classes. This ordering is consistent with the broader paper: RNA and chemistry provenance carry the dominant signal, while imaging adds a smaller but positive increment in the current derived-feature pipeline.

The smaller ZEL031 recurrent-tuple benchmark followed the same image-versus-RNA ordering at lower absolute accuracy: 0.127 for image only, 0.146 for RNA only, and 0.180 for image plus RNA across seven classes. Chemistry-only and chemistry-containing models reached perfect accuracy in that task, but the task has only seven top-tuple classes with about 30 held-out examples; it is therefore reported as a diagnostic, not as evidence of broad multimodal generalization.

The interpretation is deliberately scoped. Imaging is a complementary branch in this manuscript, while the primary evidence for a Z-Screen design map comes from RNA state linked to chemistry provenance. Even in this derived-feature form, imaging contributes directionally useful signal in additive tasks. The larger platform opportunity is richer than these benchmarks: raw multi-channel images, target-aware fluorescent or photoreactive channels, and transcriptomes can be linked by well and chemistry to support assay state, morphology, QC, colony-state, and lower-cost triage.

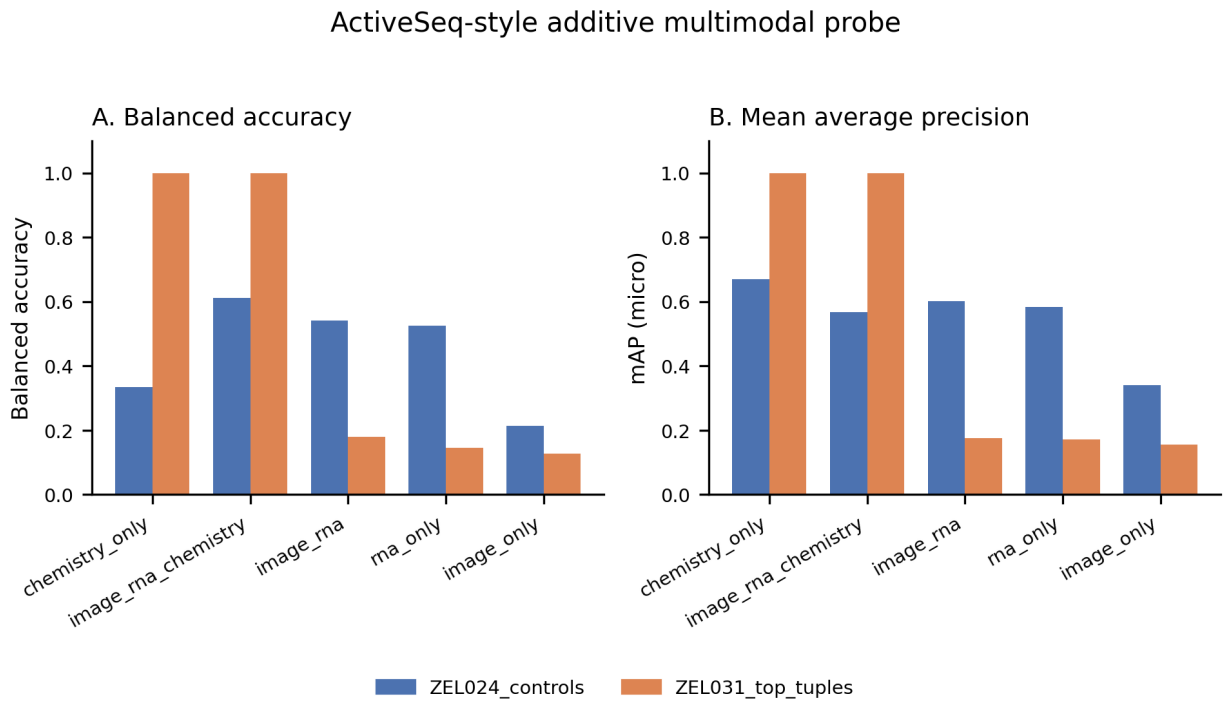


Figure 6. In a simple additive benchmark, image plus RNA modestly outperforms either single modality alone in ZEL024 controls, while RNA and chemistry provenance remain the dominant signals in the current dataset.

Discussion

This manuscript positions Z-Screen as a public combinatorial chemistry-to-transcriptome platform. The central contribution is not only the number of RNA profiles, but the fact that each measured state remains connected to an ordered building-block tuple and a hashed chemistry representation. That linkage converts a high-throughput screen into a reusable map from chemical grammar to functional RNA response.

The evidence chain is sequential. The released resource is one of the largest public combinatorial chemistry transcriptomic datasets with supporting imaging-derived features and explicit chemistry provenance. Named controls show that the low-pass RNA readout is stable, and LINCS L1000 overlap provides an external check on named-control rank signatures. Building-block models predict held-out tuple states, structure-derived embeddings add signal in the strongest sampled setting, observed tuples recover calibrated phenomimic neighborhoods around reference compounds, some building-block effects recur across cell lines, and imaging-derived features add a modest complementary branch.

The most important hierarchy is between prediction, phenomimicry, and mechanism. Prediction is a supervised modeling result: chemistry features estimate held-out RNA states. Phenomimicry is a measured-neighbor result: an observed tuple lies near a named control state after same-cell-line background calibration. Mechanism is a biological claim about target engagement or pathway

causality. This manuscript supports the first two in sampled settings and treats the third as a follow-up hypothesis to be tested with dose response, target engagement, CRISPR comparison, and orthogonal phenotyping.

The strongest sampled positive case is ZEL024 / HEK293. It has high control reproducibility, the largest building-block prediction gain, the clearest added value from the structure-derived chemistry embedding, and the most specific phenomimic example against a relatively low HEK293 cross-control background. The manuscript therefore does not argue for a universal full-structure model across Z-Screen. Instead, it shows where current structure-derived descriptors help and where explicit building-block provenance remains the more reliable design coordinate.

Relative to public resources, Z-Screen should be read on the right axis. Tahoe-100M is much larger by cell count and is the stronger reference for single-cell atlas scale. LINCS L1000 remains a mature public small-molecule transcriptional catalog. scGeneScope and Cell Painting define useful imaging and multimodal benchmarks [2-5]. Z-Screen is differentiated by combinatorial chemistry grammar linked to functional RNA readout: each tuple can be decomposed into building-block choices, and each measured RNA state can be traced back to that tuple.

The next experiments follow directly from the platform map. Sparse libraries need deeper sampling so that building-block and pair effects can be estimated with less shrinkage. Prospective held-building-block and cross-library tests are needed to determine how far the learned chemistry-to-RNA relationship extends beyond an individual library vocabulary. Phenomimic neighborhoods need orthogonal validation, including dose response, target engagement, and follow-up phenotyping. Imaging should move beyond current derived features into raw multi-channel and target-aware same-well designs. For discovery use today, the main practical value is already clear: a Z-Screen campaign can leave behind a reusable, chemistry-resolved RNA model rather than only a list of active wells.

Methods

Dataset and preprocessing

The analysis uses the repaired canonical Z-Screen public workspace, which contains RNA profiles, scVI latent coordinates, chemistry annotations, and selected derived image features across 12 combinatorial libraries and 4 cell lines. The repaired RNA aggregate contains 615,793 rows. Latent-space analyses use the 615,721 rows with valid scVI coordinates; the 72 ZIC004 / A549 rows without valid coordinates remain in the RNA count table but are excluded from latent-space modeling. The public chemistry embedding table contains 142,187 unique hashed compounds. This matches the row-count and chemistry-feature distinctions documented in the package README.

Named controls were retained for reproducibility, reference-state neighborhood analyses, and external concordance checks. Transcriptomic models were evaluated on library and cell-line subsets with sufficient compound or tuple coverage. Tuple-level analyses used centroids where replication supported them; sparse single-well systems were interpreted separately.

scVI latent representation

Most modeling analyses used a 32-dimensional scVI latent representation rather than full gene-level expression [1]. Low-pass per-well RNA-seq is noisy at the level of individual genes because each well contains a small pseudobulk of approximately 5 to 10 cells. The scVI latent space reduces this sampling noise and provides a common coordinate system for within-library prediction and cross-cell comparisons. Gene-level or rank-based profiles were used where they were the appropriate unit, including external LINCS concordance.

Control reproducibility

Control reproducibility was computed from named reference compounds with replicate wells. For each compound and cell line, wells were repeatedly partitioned into two random halves, scVI latent centroids were computed for each half, and split-half cosine similarity was recorded. The reported cell-line summaries are medians across named controls.

Building-block predictive modeling

Held-out tuple prediction used transcriptomic centroids as targets. Models were trained after excluding tuple identities assigned to the test set. The baseline predicted the training centroid for the relevant library and cell line. Building-block models used additive one-hot terms for each building-block position, and the pair model added pairwise building-block interactions where sampling supported them. Performance was summarized by cosine similarity between predicted and measured RNA centroids and by mean squared error reduction relative to the centroid baseline.

Structure-derived chemistry prediction

Chemistry benchmarks compared building-block identity, a structure-derived chemistry embedding, and the concatenation of both. The public package does not include raw Z-Screen SMILES. Instead, it includes `chem_embed.parquet`, a 256-dimensional per-compound embedding keyed by `smiles_hash`. The embedding was generated privately by canonicalizing SMILES, computing 2048-bit ECFP4 fingerprints, applying a fixed-seed Gaussian random projection from 2048 to 256 dimensions, and L2-normalizing the result. The projection matrix and raw SMILES are held privately. The public embedding preserves useful neighborhood structure for modeling while preventing reconstruction of Z-Screen substructures from the released package. Claims from these models were interpreted per library and cell line, not pooled into a general structure-to-RNA claim.

Reference-state neighborhood recovery and cross-control calibration

Reference-state neighborhood recovery compared observed library tuple centroids with named-control centroids in the same cell line. For each tuple, the nearest named-control state was recorded by cosine similarity. To calibrate these similarities, unrelated named-control centroids were compared within each cell line, producing the same-cell-line cross-control background distribution. Candidate neighborhoods were interpreted relative to that background, using the median, 95th percentile, and

99th percentile as calibration points. This analysis nominates follow-up hypotheses and was not used as proof of shared mechanism.

Cross-cell building-block consistency

Cross-cell building-block consistency estimated the direction of each building-block effect in paired library and cell-line systems, then compared matched building blocks across cell lines by cosine similarity. Results were emphasized only for chemistry-resolved positions with enough unique building blocks and partner tuples to support interpretation. Positions that collapsed many distinct chemistries into a small number of averaged values were retained as diagnostics but not treated as headline evidence.

Additive multimodal benchmark

The multimodal probe evaluated derived image features, RNA features, chemistry features, and concatenated feature sets on held-out class prediction tasks in image-linked subsets. These image-linked subsets represent the ActiveSeq branch available in the paper2 public artifacts, not the full future imaging design space. Performance was summarized with balanced accuracy and mean average precision. The benchmark used a conservative additive framing so that improvements could be traced to feature axes rather than to a complex multimodal architecture.

External LINCS L1000 concordance

The LINCS concordance analysis used the repository's Benchmarking module. Z-Screen named controls were matched to LINCS L1000 CMap LINCS 2020 compounds by name or alias. Z-Screen rank signatures were restricted to the 12,328-gene LINCS-visible universe, and LINCS level-5 consensus signatures were drawn from NCBI GEO GSE70138 and GSE92742. Spearman rank correlation was computed between matched Z-Screen and LINCS signatures for adequately covered compound and cell-line pairs. Empirical p values were computed against same-cell-line unmatched-compound null distributions.

External benchmark context

Tahoe-100M, scGeneScope, L1000, Cell Painting, and Z-Screen statistics were assembled into a comparison table covering RNA scale, image scale, cell-line coverage, perturbation diversity, pairing structure, and representative results. The comparison was used for field positioning and was not used to claim parity with larger public atlases.

Limitations

The current evidence comes from a set of public pilot-scale experiments, not a saturated atlas. Z-Screen is one of the largest public combinatorial chemistry transcriptomic datasets, but it is not larger than Tahoe-100M by cell count. The strongest sampled positive case is ZEL024 / HEK293, and conclusions from weaker or sparser systems are correspondingly narrower. Building-block

prediction is strongest within sampled library vocabularies; prospective held-building-block and cross-library transfer remain harder tests. Structure-derived embeddings add clear value mainly in ZEL024 / HEK293 and should not be treated as a universal full-structure model. Phenomimic neighborhoods are calibrated against same-cell-line cross-control backgrounds, but even the strongest matches require orthogonal validation before any mechanism claim. Imaging uses derived feature artifacts rather than raw microscopy in this public bundle and is a complementary branch rather than the primary evidence for the chemistry-to-RNA map.

Data availability

Data tables, derived image features, manuscript figures, and analysis inputs needed to reproduce this manuscript are organized in this repository under `paper2/`, `Benchmarking/`, `data/ZScreen_Canonical_Dataset/`, and `data/paper2_artifacts/`. Raw microscopy images are not included in the shareable bundle; image-linked analyses use derived parquet feature files documented in the package README. A persistent-archive deposition with an assigned DOI will accompany the corresponding preprint posting.

Code availability

Analysis and figure-generation scripts are in `paper2/scripts/`. LINCS overlap and concordance scripts are in `Benchmarking/scripts/`. Package-level dependencies are in the repository root `requirements.txt`. The package was tested with Python 3.14.3 on Windows.

References

1. Lopez R, Regier J, Cole MB, et al. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15:1053-1058. doi:10.1038/s41592-018-0229-2
2. Subramanian A, Narayan R, Corsello SM, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171(6):1437-1452.e17. doi:10.1016/j.cell.2017.10.049
3. Tahoe Therapeutics. Tahoe-100M: A 100-million single-cell perturbational atlas across 50 cancer cell lines. Dataset card and February 2025 release. https://huggingface.co/datasets/tahoe-bio/Tahoe-100M
4. Dapello J, Nassar M, Eksi R, et al. scGeneScope: A treatment-matched single-cell imaging and transcriptomics dataset and benchmark for treatment response modeling. NeurIPS 2025 Datasets and Benchmarks. OpenReview. https://openreview.net/pdf/f7d541dae38bcf88a79789a4c6440aadfec123c7
5. Bray MA, Singh S, Han H, et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc. 2016;11:1757-1774. doi:10.1038/nprot.2016.105
6. Dixit A, Parnas O, Li B, et al. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell. 2016;167(7):1853-1866.e17. doi:10.1016/j.cell.2016.11.038

A functional generalization roadmap for designing cell states from combinatorial chemistry

Abstract

Drug-discovery machine learning has advanced most visibly in settings where the output is target binding, docking geometry, or DNA-encoded-library enrichment. Z-Screen asks a complementary question: can chemistry be modeled directly against functional cellular state? We define a five-rung train/test ladder for chemistry-to-transcriptome prediction, separating missing-tuple interpolation, held-building-block extrapolation, chemical-neighborhood holdout, and cross-library scaffold-family transfer. Applied to Z-Screen, a 50,000-well one-bead-one-compound microchip assay with compound provenance, imaging, and mRNA-sequencing readouts, the ladder resolves three useful design regimes. ZEL031 in THP1 cells provides genuine held-building-block extrapolation: under strict multi-axis holdout, ridge regression on chemistry features reached 0.765 cosine to the measured transcriptomic state, exceeded chemistry nearest-neighbor retrieval by 0.168, won on 90% of test compounds, and remained positive in all 10 random draws. ZEL024 in HEK293 cells provides high-value dense-grid completion inside a saturated 12 by 7 by 2 by 83 combinatorial design, where held-out compounds often have close measured neighbors. Retrospective cross-library transfer is not yet a present-tense capability; it defines the prospective L5 frontier, with the best same-cell scaffold-family hop reaching 0.194 cosine and a 0.096 gain over retrieval. The ladder turns "generalization" from a single ambiguous score into a roadmap for designing chemistry toward desired cell states.

Significance

Most computational chemistry benchmarks ask whether a model can find, rank, dock, or design binders. Functional drug discovery needs a second axis: whether a compound can drive a desired cellular response. Z-Screen links combinatorial building-block provenance to measured transcriptomic state, making it possible to ask what kind of chemistry generalization is being achieved. The answer is not one number. ZEL024 / HEK293 shows that dense combinatorial grids can be completed productively. ZEL031 / THP1 shows that held building blocks can be predicted beyond nearest-neighbor retrieval in a favorable system. L5 cross-library transfer specifies the prospective experiment needed to move from within-library design to scaffold-family transfer.

Introduction

Modern computational discovery has made target-centric prediction increasingly powerful. Docking, structure-aware modeling, DEL enrichment prediction, and building-block-centric DEL models all help answer whether a molecule or substructure is likely to bind a target [1-7]. Those are essential capabilities, but they leave a second design problem open: cells respond to chemistry through transport, metabolism, pathway coupling, polypharmacology, state dependence, and compensatory programs. A compound that binds is not automatically a compound that produces the desired cell state.

Z-Screen is built around that functional design problem. The platform couples one-bead-one-compound combinatorial chemistry with microwell phenotyping, preserving tuple structure and building-block provenance while measuring downstream cellular response. In the public Z-Screen bundle, each modeled compound is represented by a hashed compound identity, public chemistry embedding, library and cell-line context, building-block columns where available, and transcriptomic state summarized as a 32-dimensional scVI latent profile. This makes the central machine-learning question unusually explicit: can a model learn how chemical substructures and combinatorial tuples map to functional RNA state?

The answer depends on the split. A model that fills in an unmeasured tuple inside a dense grid is useful for prioritizing chemistry within an explored design space. A model that predicts a tuple containing building blocks withheld from training supports a stronger claim: within-library functional extrapolation from chemistry. A model that transfers from one scaffold family to another would support the hardest claim in this manuscript: prospective scaffold-family transfer of cell-state design rules. These are related but non-interchangeable capabilities.

We therefore define a five-rung generalization ladder:

Rung	Train/test regime	Design claim tested
L1	Train and test use the same building-block vocabulary; the test compound is only a held-out full combination.	Missing-tuple interpolation inside familiar chemistry.
L2	A building-block identity at one position is absent from training and appears only in test compounds.	Single-position held-building-block extrapolation.
L3	Every substantive building-block axis in the test compound contains identities absent from training.	Strict within-library out-of-vocabulary prediction.

Rung	Train/test regime	Design claim tested
L4	Test compounds are chemical-neighborhood clusters held out in projected ECFP space.	Prediction away from local analog retrieval while staying inside one library grammar.
L5	Train on one library family and test on another library family in the same cell line.	Cross-library scaffold-family transfer.

The ladder is not a ranking of good and bad results. It is an operational roadmap. L1 and dense L2 settings are the natural regimes for completing partially measured combinatorial designs. L2 and L3 ask whether the platform can move into new building-block identities while retaining the same library grammar. L4 tests whether gains survive when local chemical neighborhoods are removed. L5 marks the prospective frontier: leaving the training scaffold family while holding cellular context fixed.

All rungs are evaluated against two baselines. The first is a train-mean phenotype predictor. The second is a chemistry nearest-neighbor retrieval baseline that copies the measured phenotype of the closest training compound in the public 256-dimensional chemistry embedding. The retrieval baseline is central because it separates learned chemistry-to-cell-state structure from local analog lookup. The result is a calibrated map of current capability: ZEL031 / THP1 carries the strongest held-building-block extrapolation claim, ZEL024 / HEK293 carries the dense-grid completion claim, and L5 defines the next prospective experiment.

Results

The ladder converts generalization into a design roadmap

Question. Which kind of "new chemistry" is a chemistry-to-cell-state model being asked to handle?

Split/test. We scored the same ridge-regression model across the ladder. L1 is the missing-tuple anchor. L2 holds out building blocks at one position. L3 holds out building-block identities across all substantive positions. L4 holds out chemical neighborhoods. L5 trains on one library and tests on another library in the same cell line.

Metric. Each split is reported as row-wise cosine similarity to the measured transcriptomic state, mean squared error reduction against the train mean, cosine gain over chemistry nearest-neighbor retrieval, per-compound win rate against retrieval, and nearest-training chemistry similarity in the public projected ECFP embedding.

Interpretation. The ladder makes each model result actionable. High performance in a dense L1/L2 regime supports completion of a measured design grid. High performance in L2/L3 with lower nearest-training similarity supports held-building-block extrapolation. L4 asks whether the

signal persists away from local chemical neighborhoods. L5 defines the scaffold-family hop required for prospective cell-state design beyond the training library grammar.

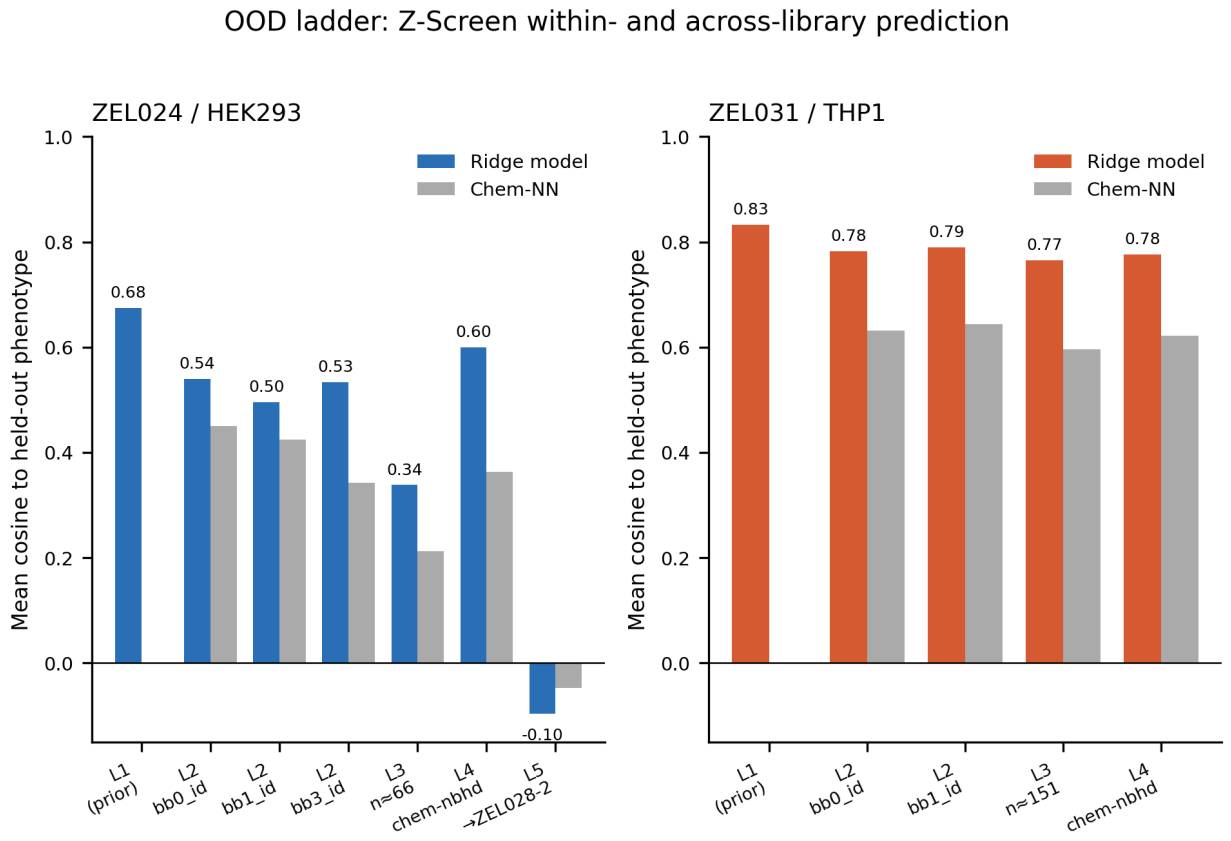


Figure 1. The generalization ladder is an experimental decision tree. L1 asks whether missing tuples can be filled in when all building blocks are familiar. L2 and L3 ask whether held building blocks can be predicted within a library. L4 asks whether gains survive chemical-neighborhood holdout within one reaction grammar. L5 asks whether a model transfers across scaffold families in the same cell line.

ZEL031 / THP1 supports genuine held-building-block extrapolation

Question. Can Z-Screen predict functional RNA states for compounds whose building-block identities were not present in training?

Split/test. The clearest test is ZEL031 in THP1 cells. In L3, every substantive building-block axis in the test compounds contained identities withheld from training. Ten random L3 draws produced an average of 151 test compounds per draw. We also ran L2 single-position holdouts for the two substantive axes, averaging about 729 test compounds per draw.

Metric. In L3, the model reached 0.765 cosine to the measured transcriptomic state, while nearest-neighbor retrieval reached 0.597. The mean gain was 0.168 cosine, the model beat retrieval on 90% of test compounds, and all 10 random draws were positive under paired bootstrap intervals. Median

nearest-training chemistry cosine was 0.642, so the test compounds were not simply near-duplicates in the public embedding space. In L2, the averaged result across bb0 and bb1 was 0.786 model cosine versus 0.638 retrieval cosine, a 0.149 gain, a 90% win rate, and median nearest-training chemistry cosine of 0.791.

Interpretation. This is the strongest functional extrapolation result in the manuscript. A conservative linear model on chemistry features predicts held-building-block transcriptomic states beyond measured-phenotype retrieval. In design terms, ZEL031 / THP1 shows that the assay can support movement into unmeasured building-block identities within a library grammar while retaining cell-state predictability.

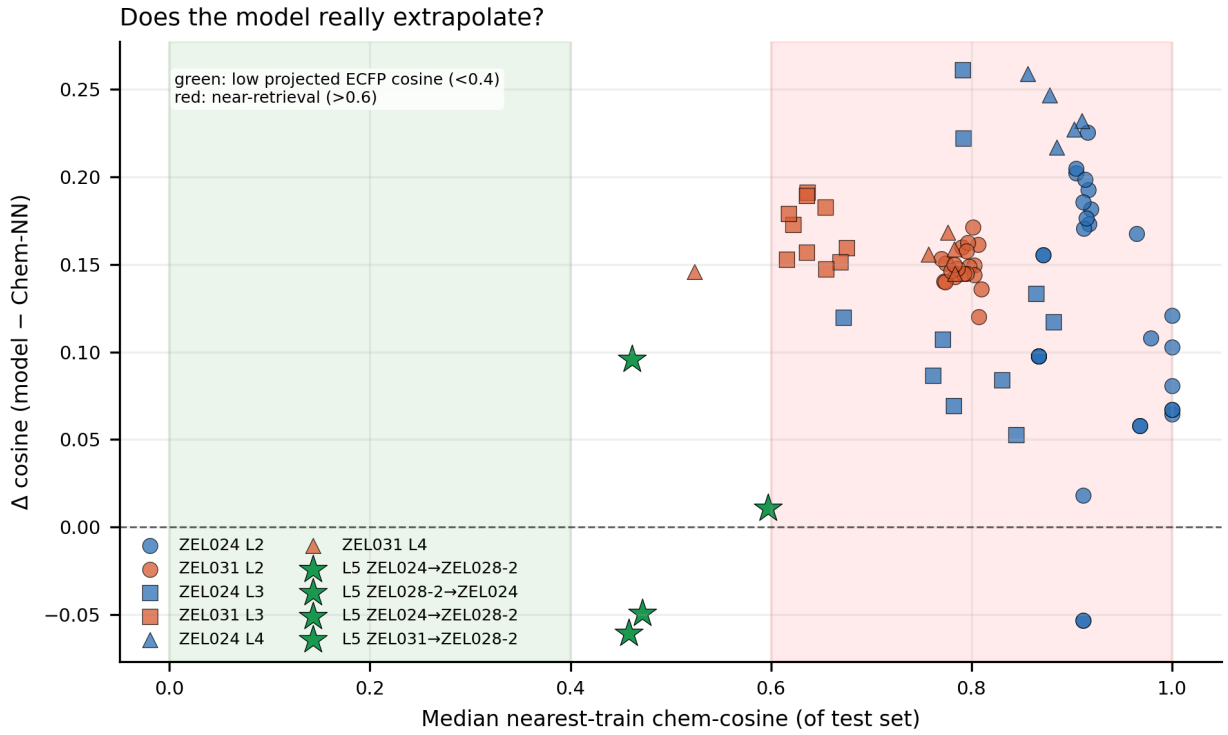


Figure 2. Retrieval-versus-extrapolation analysis plots model gain against median nearest-training similarity measured as cosine in the random-projected ECFP4 embedding, not raw-fingerprint Tanimoto. ZEL031 / THP1 sits in the lower-neighbor-similarity, high-model-gain regime, consistent with held-chemistry extrapolation rather than nearest-neighbor copying. ZEL024 / HEK293 sits in the high-neighbor-similarity regime, consistent with high-quality matrix completion within a saturated combinatorial grid.

ZEL024 / HEK293 is high-value dense-grid completion

Question. What kind of design capability is demonstrated when a dense combinatorial library predicts held-out tuples with many close training neighbors?

Split/test. ZEL024 / HEK293 was evaluated with the same L2 and L3 framework. The library covers a saturated 12 by 7 by 2 by 83 combinatorial grid, with 13,769 observed compounds out of a

13,944 possible-tuple ceiling. In L2, identities at bb0, bb1, and bb3 were held out one position at a time. In L3, identities across all substantive axes were withheld together.

Metric. L2 was positive, with 0.523 model cosine, 0.406 nearest-neighbor cosine, and a 0.117 gain. The median nearest-training chemistry cosine was 0.927, reflecting many close training neighbors inside the saturated grid. The strict L3 split was smaller and less stable: average test size fell to 67 compounds, model cosine was 0.338, nearest-neighbor cosine was 0.213, and nearest-training chemistry cosine was still 0.799. In the underlying L3 draws, mean squared error reduction against the train mean was not consistently positive, even though cosine direction remained above retrieval on average.

Interpretation. ZEL024 / HEK293 is not a weaker version of the ZEL031 / THP1 result; it is a different design mode. It shows that dense phenotypic measurement can complete a nearly saturated combinatorial grid, prioritizing compounds inside an explored chemistry space where experimental coverage is high and analog structure is rich. For chemistry campaigns, this is valuable because it turns partial measurement into a map for choosing the next tuples to synthesize or profile. The ladder keeps that value visible without relabeling it as the same out-of-vocabulary claim.

Different generalization regimes, not a head-to-head OOD comparison

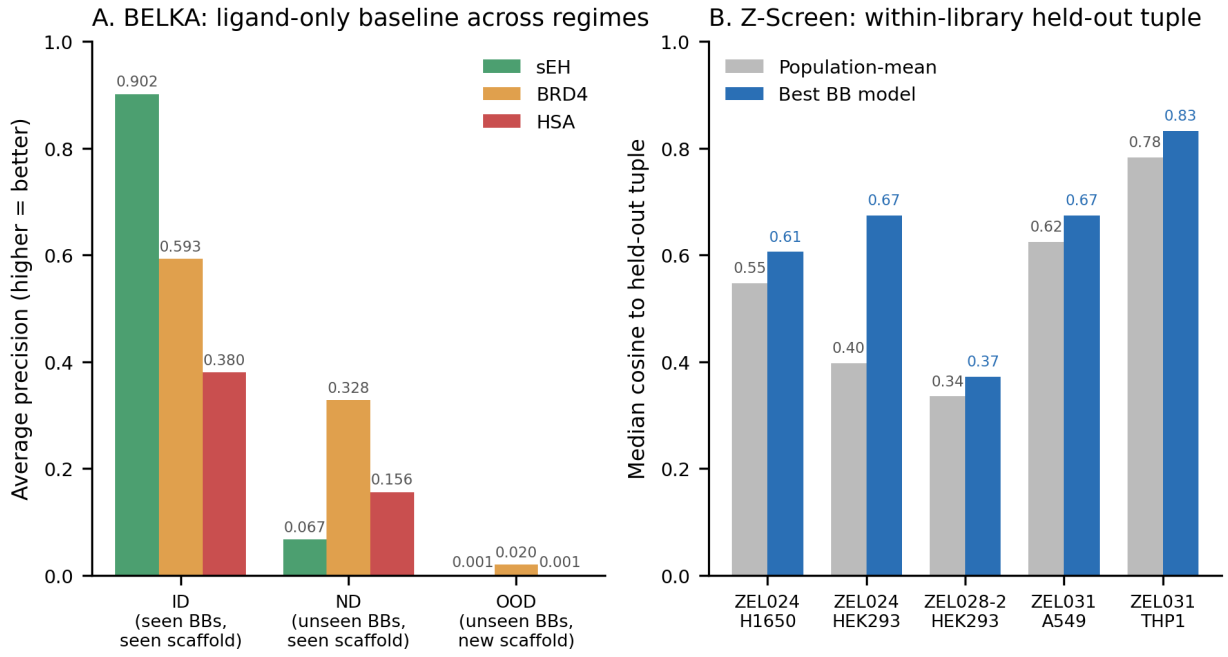


Figure 3. BELKA is used here as a vocabulary for generalization regimes, not as a head-to-head performance comparison. The Z-Screen panels separate matrix completion from harder held-chemistry regimes, with ZEL031 / THP1 carrying the strongest extrapolation signal in the present dataset.

Chemical-neighborhood holdout supports within-library signal beyond retrieval

Question. Do model gains survive when the test set is defined by chemical neighborhoods rather than by named building-block identities?

Split/test. L4 held out clusters seeded by greedy farthest-point sampling in the public projected ECFP cosine-distance space. This keeps the reaction grammar fixed but removes local chemical neighborhoods from training.

Metric. ZEL024 / HEK293 reached 0.600 model cosine versus 0.363 nearest-neighbor cosine, a 0.236 gain on average cluster sizes of 2,754 compounds. ZEL031 / THP1 reached 0.777 model cosine versus 0.622 retrieval cosine, a 0.155 gain on average cluster sizes of 711 compounds. Per-compound win rates were 84% and 92%, respectively.

Interpretation. L4 reinforces the within-library design signal. The models do more than copy the closest measured analog in the public embedding space, even when chemical neighborhoods are held out. This is an important bridge between dense-grid completion and held-building-block extrapolation because it shows that functional cell-state prediction contains recoverable chemistry structure beyond immediate local retrieval.

L5 defines the prospective scaffold-family frontier

Question. What would it take to move from within-library design to cross-library scaffold-family transfer?

Split/test. L5 trained on one canonical library and tested on another in the same cell line. Because building-block vocabularies do not overlap across libraries, L5 used the chemistry embedding alone rather than building-block one-hot features. Four train/test pairs had enough data for evaluation.

Metric. The best absolute result was ZEL024 -> ZEL028-2 in H1650: 0.194 model cosine, 0.098 nearest-neighbor cosine, a 0.096 gain, and a 70% win rate on 765 test compounds. HEK293 transfer was negative in both directions: ZEL024 -> ZEL028-2 gave -0.096 model cosine versus -0.047 retrieval, and ZEL028-2 -> ZEL024 gave -0.095 versus -0.034. A549 transfer from ZEL031 -> ZEL028-2 was near-flat at 0.098 model cosine versus 0.087 retrieval.

Interpretation. Retrospective L5 is best read as the frontier rather than as a failed version of L2-L4. It exposes the chemistry and assay conditions that a prospective scaffold-family experiment must control: same cell line, parity coverage across families, pre-registered splits, and measured held-out outcomes. A positive prospective L5 result would extend Z-Screen from within-library functional design to scaffold-family transfer. The current data specify that experiment cleanly.

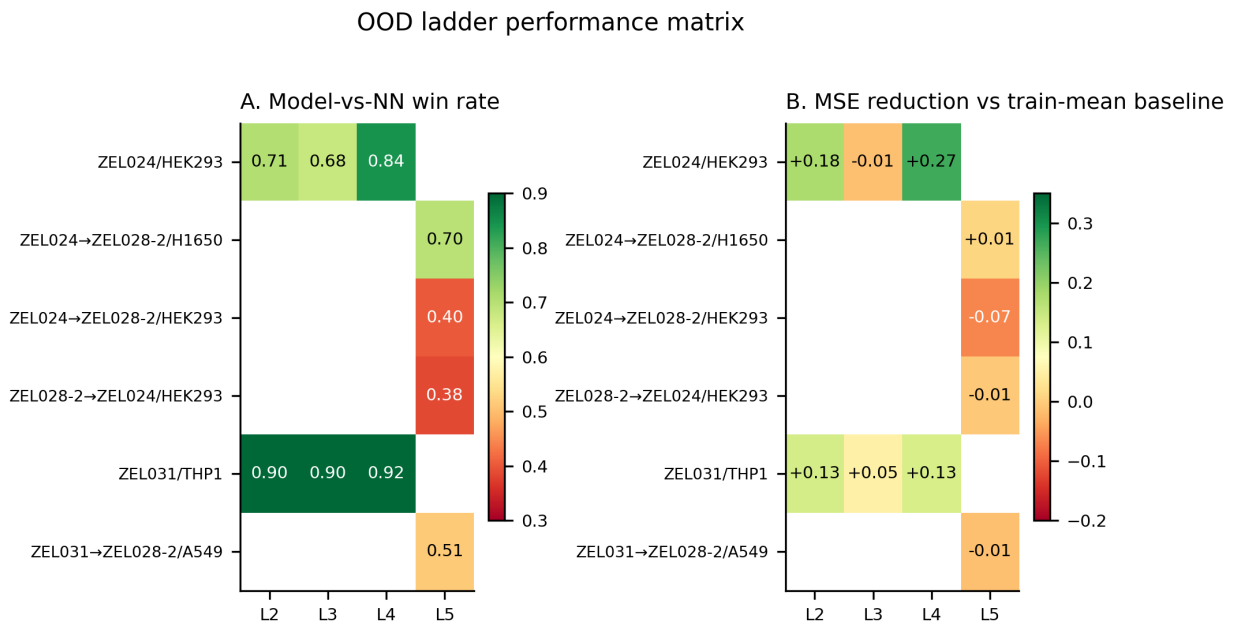


Figure 4. Per-system performance across the ladder shows strong within-library prediction in selected systems (ZEL031 / THP1 for genuine extrapolation, ZEL024 / HEK293 for matrix completion in a saturated grid), and a marked drop at retrospective cross-library scaffold-family transfer that defines the natural next experiment rather than a stable performance ceiling.

Table 1. Reader-facing summary of the ladder. Cosines are row-wise similarity to measured transcriptomic state; nearest-training similarity is cosine in the random-projected ECFP4 embedding.

Rung	System	Model	NN	Delta	Win rate	Nearest train	Interpretation
L2	ZEL024 / HEK293	0.523	0.406	0.117	0.705	0.927	dense-grid completion
L2	ZEL031 / THP1	0.786	0.638	0.149	0.900	0.791	held-building-block extrapolation
L3	ZEL024 / HEK293	0.338	0.213	0.125	0.682	0.799	small-test completion boundary

Rung	System	Model	NN	Delta	Win rate	Nearest train	Interpretation
L3	ZEL031 / THP1	0.765	0.597	0.168	0.899	0.642	strict held-chemistry extrapolation
L4	ZEL024 / HEK293	0.600	0.363	0.236	0.844	0.886	within-library signal beyond local retrieval
L4	ZEL031 / THP1	0.777	0.622	0.155	0.921	0.725	within-library signal beyond local retrieval
L5	ZEL024 -> ZEL028-2 / H1650	0.194	0.098	0.096	0.698	0.461	prospective scaffold-family lead case
L5	ZEL024 -> ZEL028-2 / HEK293	-0.096	-0.047	-0.049	0.402	0.472	retrospective transfer boundary
L5	ZEL028-2 -> ZEL024 / HEK293	-0.095	-0.034	-0.061	0.385	0.458	retrospective transfer boundary
L5	ZEL031 -> ZEL028-2 / A549	0.098	0.087	0.011	0.513	0.597	near-flat retrospective transfer

The next decisive experiment is now specified

Question. What experiment would turn L5 from a retrospective frontier into a prospective design claim?

Split/test. A clean prospective L5 experiment would hold cell line fixed, train on one or more scaffold families with adequate coverage, nominate compounds from a withheld scaffold family before measurement, and then measure those held-out compounds on a new chip. H1650 is the strongest candidate in the present tables because multiple canonical libraries were screened there at useful scale and the retrospective ZEL024 -> ZEL028-2 result is positive.

Metric. The primary endpoint should be measured-versus-predicted transcriptomic cosine on the held-out scaffold family, compared directly with chemistry nearest-neighbor retrieval. Secondary endpoints should include pathway-score correlation, train-mean improvement, replicate stability, and pre-specified per-compound win rate against retrieval.

Interpretation. The ladder turns the next experiment into an explicit design milestone. If the prospective held-out scaffold family shows a retrieval-adjusted gain near the 0.15 to 0.17 observed for ZEL031 / THP1 L2-L3, then Z-Screen would have extended from within-library cell-state design to scaffold-family transfer. If the gain is smaller, the platform still retains two present capabilities: dense-grid completion and held-building-block extrapolation within selected library/cell-line systems.

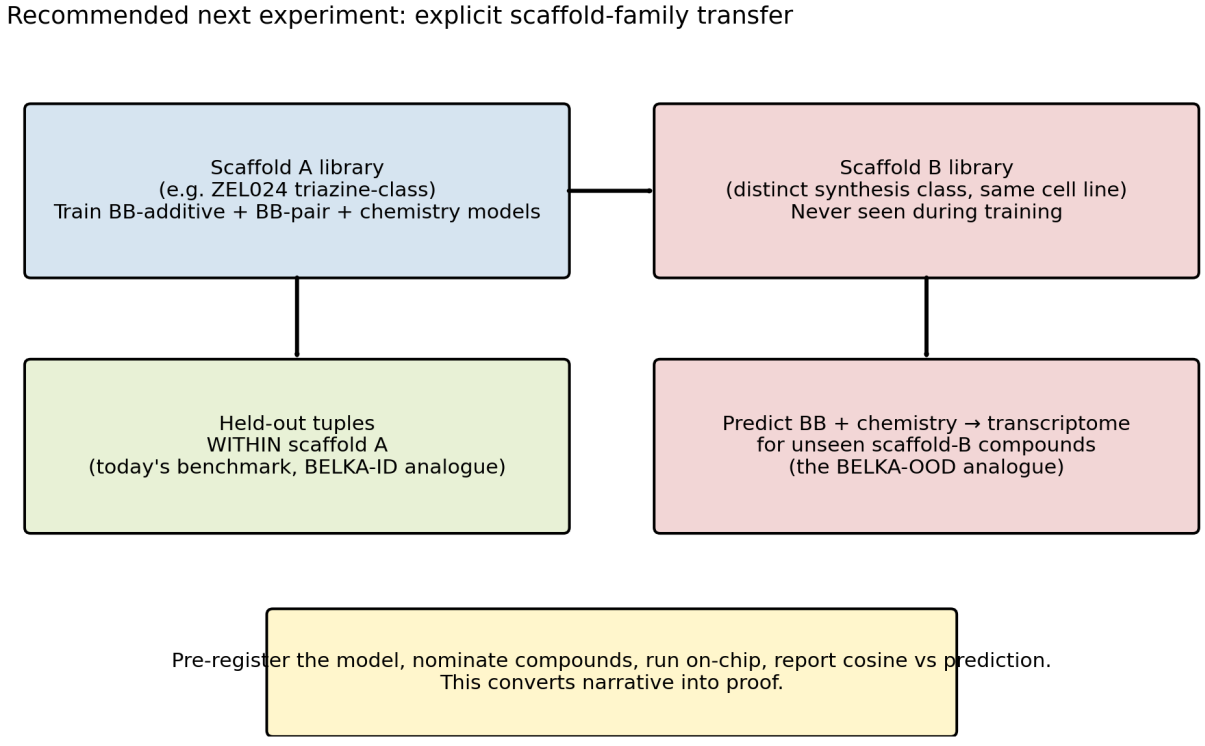


Figure 5. The next decisive experiment is a prospective same-cell scaffold-family hop with a pre-registered model, nominated compounds, and measured transcriptomic outcomes on the held-out family.

Discussion

This benchmark reframes chemistry generalization around functional design. Binding prediction, docking, and DEL modeling ask how chemistry relates to target engagement. Z-Screen adds a complementary endpoint: how chemistry relates to downstream cellular state. That shift matters because a discovery program ultimately needs compounds that create useful biology, not only molecules that score well against an isolated target.

The ladder identifies three design regimes in the present data. First, dense-grid completion is a practical capability. ZEL024 / HEK293 shows that a nearly saturated combinatorial grid can be completed with useful accuracy, supporting prioritization inside an explored chemistry design. Second, held-building-block extrapolation is achievable in a favorable system. ZEL031 / THP1 predicts multi-axis held-out building-block combinations beyond nearest-neighbor retrieval, with a large retrieval-adjusted gain and consistent positive draws. Third, cross-library scaffold-family transfer is the prospective frontier. The retrospective L5 results are informative because they specify the same-cell scaffold-family hop needed to test whether functional design rules can leave the training library grammar.

The broader message is that "generalization" should be reported as a ladder, not collapsed into a single OOD label. A matrix-completion result can be highly useful without being an extrapolation result. A held-building-block result can support functional generalization without proving scaffold transfer. A cross-library result should be tested prospectively because it changes scaffold family, chemistry coverage, and latent-space alignment at once. The ladder lets each claim keep its proper evidentiary weight.

For Z-Screen, that calibrated structure is a strength. The platform links tuple identity, building-block provenance, chemistry embeddings, and RNA state in a way that lets the field ask design-relevant questions directly. The immediate uses are to complete dense combinatorial maps and to prioritize held-building-block chemistry in systems like ZEL031 / THP1. The next use is to pre-register and execute L5 scaffold-family transfer, using the same retrieval-adjusted metrics. Together, those steps define a path from measuring combinatorial chemistry to designing cell states from chemistry.

Methods

Data and endpoints

The analyses use the public Z-Screen bundle. Per-compound chemistry is keyed by `smiles_hash`; raw Z-Screen SMILES are not included in the public package. Transcriptomic phenotypes are represented by 32-dimensional scVI latent coordinates (`D00` to `D31`) aggregated per compound within a library and cell line. Unless otherwise stated, compounds required at least three wells for L2-L4 analyses. L5 used at least three wells for training compounds and at least two wells for test compounds to retain enough cross-library test coverage.

Chemistry features

Chemistry features come from `data/ZScreen_Canonical_Dataset/RNASeqAggregate/chem_embed.parquet`. Each compound has a 256-dimensional embedding generated from a 2048-bit ECFP4 fingerprint by a fixed-seed Gaussian random projection and L2 normalization. The projection matrix is held privately, so the public embedding supports deterministic modeling and nearest-neighbor search without exposing invertible substructure information. Cosine similarity in this projected space is used for retrieval and for nearest-training similarity reports.

Model

The primary model is ridge regression with alpha equal to 10. For L2-L4, the feature matrix includes the public chemistry embedding plus split-refit one-hot encodings of available building-block columns; unseen test categories are handled as unknowns by the encoder. For L5, building-block vocabularies do not overlap across libraries, so the model uses the chemistry embedding only. Ridge regression was chosen deliberately: the benchmark asks whether chemistry features support the split-defined claim, and a conservative linear model keeps that question easier to interpret.

Baselines and metrics

The train-mean baseline predicts the average training phenotype for every test compound. The nearest-neighbor retrieval baseline finds the closest training compound by exact cosine search in the L2-normalized 256-dimensional chemistry embedding, then copies that training compound's measured phenotype. Reported metrics are row-wise cosine to the measured phenotype, mean squared error reduction against the train mean, cosine gain versus nearest-neighbor retrieval, per-compound win rate versus retrieval, and median nearest-training chemistry cosine.

Ladder implementation

L1 is the held-tuple interpolation anchor: train and test compounds share the same building-block vocabulary, and only full tuples are held out. L2 randomly holds out identities at one building-block position and removes compounds containing those identities from training. L3 holds out identities across all substantive building-block positions simultaneously, removing from training any compound containing a held identity. L4 creates chemical-neighborhood folds by greedy farthest-point seeding in projected ECFP cosine-distance space. L5 trains on one canonical library and tests on another in the same cell line.

Statistical reporting

L2 and L3 use 10 random draws per eligible system; L4 uses five chemical-neighborhood folds. Each split reports paired bootstrap confidence intervals for the model-versus-retrieval cosine difference using 2,000 paired bootstrap samples over test compounds. The summary table reports the number of positive draws where applicable, average test-set size, model cosine, nearest-neighbor cosine, retrieval-adjusted gain, win rate, and nearest-training chemistry similarity.

Limitations

The benchmark is retrospective, and the strongest claims are system-specific. ZEL031 / THP1 supports held-building-block extrapolation; ZEL024 / HEK293 supports dense-grid completion. The current retrospective tables do not establish broad cross-library scaffold-family transfer. L5 is also affected by uneven library coverage, sparse ZEL028-2 replication in some settings, and possible cross-library drift in the shared scVI latent space. These boundary conditions motivate the prospective L5 experiment rather than weakening the present L2-L4 design claims.

Data availability

Data tables, manuscript figures, and analysis inputs needed to reproduce this manuscript are organized in this repository under `paper3/`, `paper2/tables/`, and `data/ZScreen_Canonical_Dataset/`. A persistent archive with an assigned DOI should accompany the corresponding preprint posting.

Code availability

Analysis and figure-generation scripts are in `paper3/scripts/`. Package-level dependencies are in the repository root `requirements.txt`. The public bundle was tested with Python 3.14.3 on Windows.

References

1. Dolorfino MD, Santos Perez D, Fu Y, et al. Assessing the Generalizability of Machine Learning and Physics Methods for DNA-Encoded Libraries. bioRxiv. 2026. doi:10.64898/2026.04.18.719394v1
2. Quigley IK, Blevins A, Halverson BJ, Wilkinson N. BELKA: The Big Encoded Library for Chemical Assessment. NeurIPS 2024 Competition Track. https://neurips.cc/virtual/2024/competition/84787
3. Peterson AA, Liu DR. Small-molecule discovery through DNA-encoded libraries. Nat Rev Drug Discov. 2023;22:699-722. doi:10.1038/s41573-023-00713-6
4. McCloskey K, Sigel EA, Kearnes S, et al. Machine Learning on DNA-Encoded Libraries: A New Paradigm for Hit Finding. J Med Chem. 2020. doi:10.1021/acs.jmedchem.0c00452
5. Zhang Y, et al. Building Block-Based Binding Predictions for DNA-Encoded Libraries. J Chem Inf Model. 2023;63(16):5120-5132. doi:10.1021/acs.jcim.3c00588
6. Fitzgerald PR, Dixit A, Zhang Y, Mobley DL, Paegel BM. Building Block-Centric Approach to DNA-Encoded Library Design. J Chem Inf Model. 2024;64(12):4661-4672. doi:10.1021/acs.jcim.4c00232
7. Montoya AL, Hogendorf AS, Tingey S, Kuberan A, Yuen LH, Schuler H, Franzini RM. Widespread false negatives in DNA-encoded library data: how linker effects impair machine learning-based lead prediction. Chem Sci. 2025;16:10918-10927. doi:10.1039/D5SC00844A

Cross-cell transfer of reusable chemical-response programs in Z-Screen

Abstract

Multi-context cellular screening is powerful but scales poorly when every compound must be profiled in every cell type. Z-Screen addresses this cost by linking one-bead-one-compound chemistry, building-block provenance, tuple identity, imaging-derived features, and low-pass RNA state in a shared discovery system. Here we test whether chemical responses measured deeply in one cell context can prioritize where to measure next. In the canonical Z-Screen public dataset, two paired library/cell-line systems support primary molecule-level transfer: ZEL024 H1650/HEK293 and ZEL031 A549/THP1. Learned source-to-target maps improved held-out centered-response cosine over direct source reuse, reaching 0.279-0.371 in ZEL024 and 0.212-0.242 in ZEL031. The strongest result emerged when molecules were grouped into chemistry-resolved response programs. ZEL024 bb0+bb1 programs transferred across H1650 and HEK293 with support-weighted cosine 0.808-0.863, and ZEL031 bb0 programs transferred across A549 and THP1 with cosine 0.664-0.714. ZEL028-2 single-building-block aggregates provided useful guardrails, showing that high scores must preserve chemical specificity. Cross-library projection and gene-level decoding produced directional, hypothesis-generating signals and define the next frontier. These results establish cross-cell transfer as a practical Z-Screen scaling strategy: profile combinatorial chemistry deeply in one context, learn reusable chemical-response programs, and prioritize target-cell follow-up with explicit uncertainty.

Significance

Adding cell types is one of the fastest ways to increase both the value and the cost of a functional chemistry screen. Z-Screen creates a practical alternative to measuring every compound everywhere. The platform can learn how chemistry-linked RNA programs move between paired cell contexts, then use those maps to rank which target-cell measurements are most worth performing. The most transferable unit in the current public data is not the isolated sparse molecule; it is the reusable chemical-response program defined by shared combinatorial building blocks at a level that remains chemically specific. That is the level at which early discovery teams often make series-level decisions.

Introduction

Cell type is not a nuisance variable in chemical biology. It determines pathway wiring, target abundance, stress response, lineage state, transport, and the transcriptional consequences of

perturbation. A compound series that looks promising in one cell type may become inactive, toxic, or mechanistically different in another. The direct answer is to run every compound across every relevant cell context. That strategy is scientifically clean, but it scales poorly as chemistry, dose, timing, and cellular diversity expand.

Z-Screen is designed for a complementary operating model. It measures large one-bead-one-compound libraries while preserving the chemical provenance of each profiled molecule. A building block is a synthetic component used to assemble combinatorial molecules; a tuple is the ordered combination of building blocks that defines a compound in a library. In the public bundle, raw Z-Screen SMILES are replaced with irreversible `smiles_hash` identifiers, and chemistry-aware analyses use building-block annotations plus precomputed chemistry embeddings. The platform advantage is that compound identity, tuple structure, RNA state, and imaging-derived measurements remain linked, allowing chemical-response programs to be studied as reusable objects rather than as isolated screening hits.

This manuscript asks whether that structure can reduce the cost of multi-context biology. We define cross-cell transfer as a measured source-cell response, a paired landmark set of compounds observed in both source and target cell lines, and a learned map that estimates or prioritizes the target-cell response for held-out chemistry. The goal is not to replace direct target-cell validation. The goal is more useful and more immediate: profile chemistry deeply in one context, learn how response programs translate across biology, and decide which target-cell experiments deserve the next measurements.

Public perturbational atlases provide the broader field context. LINCS L1000 established large-scale transcriptional perturbation mapping in a reduced 978-gene representation [2]. Tahoe-100M extends single-cell drug perturbation profiling across 50 cancer cell lines for roughly 1,100 to 1,200 compounds [3]. scGeneScope contributes treatment-matched imaging and transcriptomics as a multimodal benchmark [4]. Z-Screen occupies a different axis: one of the largest public combinatorial chemistry transcriptomic datasets, with supporting imaging-derived features and explicit building-block and tuple provenance. In the full public bundle, Z-Screen contains 615,793 repaired RNA profiles, 615,721 valid scVI latent profiles, 12 combinatorial libraries, 4 cell lines, and 142,187 unique hashed compounds with chemistry embeddings. The relevant question for this paper is whether those chemistry coordinates make cross-cell response transfer useful.

We separate three levels of evidence. First, held-out molecule-level transfer tests whether a source-cell signature carries target-cell information beyond direct reuse. Second, grouped-program transfer asks whether molecules sharing defined building-block programs form more stable transferable response units. This is the primary platform result. Third, cross-library projection and gene-level decoding ask whether transfer can be extended beyond the current dense paired library settings. These frontier analyses are reported because they define where the platform can scale next.

Results

Paired library coverage defines the transfer opportunities

Question. Which library/cell-line pairs provide enough shared chemistry to test cross-cell transfer?

Comparison. The canonical Z-Screen public dataset is broad but uneven across libraries and cell lines. ZEL024 contains H1650 and HEK293 runs with 10,695 and 13,923 unique compounds, respectively; HEK293 is the denser side, with a median of 11 wells per compound. ZEL031 contains A549 and THP1 runs with 8,321 and 9,041 unique compounds, plus a much smaller H1650 run. ZEL028-2 is broad and shallow, with 40,622 A549, 25,906 H1650, and 61,396 HEK293 unique compounds, each at a median of 1 well per compound.

Metric. Eligibility was based on shared compound count after a minimum per-cell support filter. At a two-cell filter, ZEL024 H1650/HEK293 retained 6,299 shared compounds, and ZEL031 A549/THP1 retained 3,561 shared compounds. ZEL028-2 overlaps fell to 9-19 shared compounds at the same filter, and ZEL031 comparisons involving H1650 were similarly too small. At a five-cell filter, the eligible molecule-level set narrowed to 353 shared compounds for ZEL024 H1650/HEK293 and 140 for ZEL031 A549/THP1.

Interpretation. These coverage facts set the evidence hierarchy. ZEL024 and ZEL031 support primary held-out molecule-level transfer. The broader ZEL028-2 data are most informative as grouped-transfer diagnostics, aggregation guardrails, and cross-library stress tests. This structure is exactly the scaling problem Z-Screen is built to address: deep paired maps in some contexts can guide which additional cell-type measurements should be prioritized.

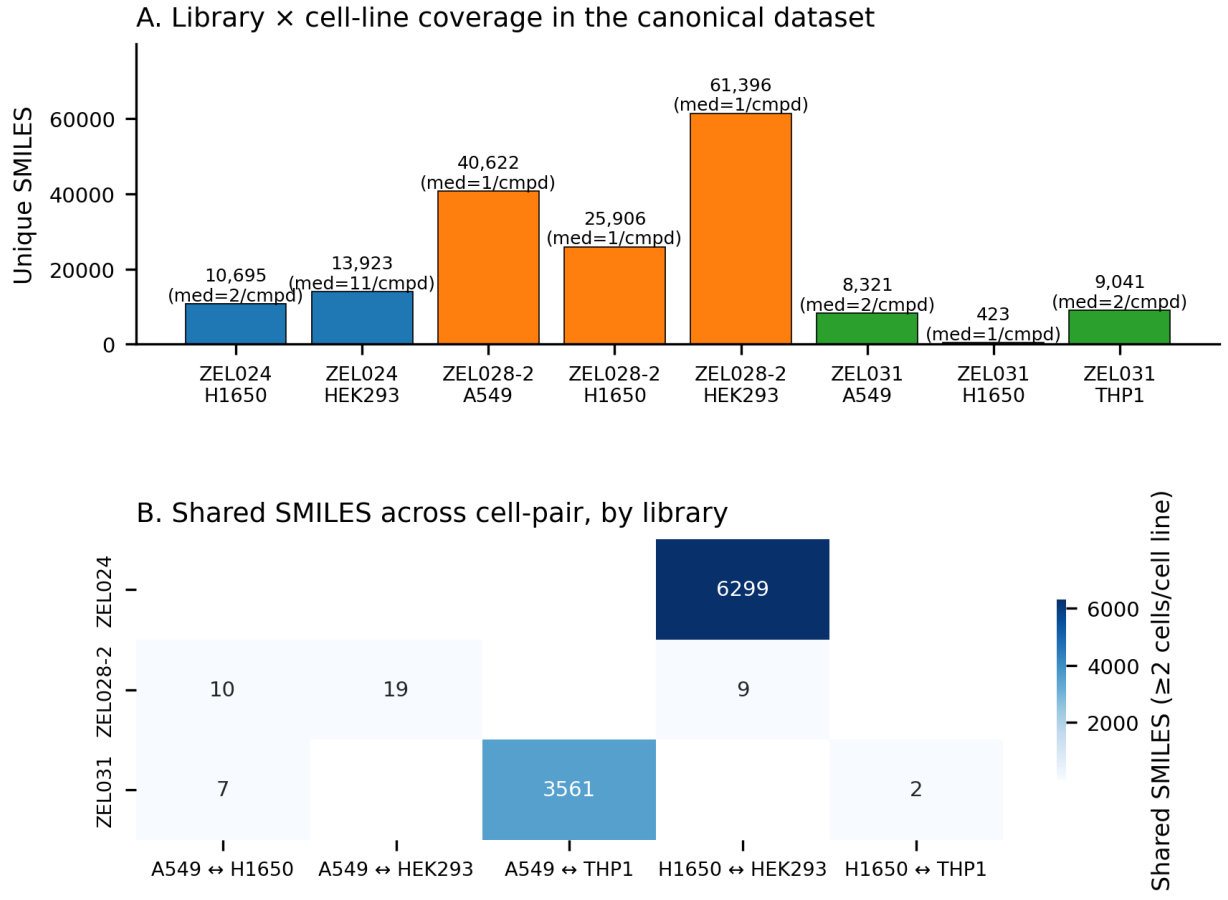


Figure 1. Library-by-cell-line coverage in the canonical Z-Screen dataset. Top: unique compounds and median wells per compound per system. Bottom: shared-compound counts across cell-pair directions, by library. The two dense paired systems used in the molecule-level transfer benchmark are ZEL024 (H1650 <-> HEK293, 6,299 shared) and ZEL031 (A549 <-> THP1, 3,561 shared).

Learned maps improve held-out molecule-level transfer

Question. For individual compounds measured in two cell lines, does a learned transfer map recover target-cell response better than simply reusing the source-cell signature?

Comparison. Each compound was represented by the mean of its 32-dimensional scVI latent coordinates within a library and cell line. The benchmark used held-out compounds from the two eligible paired systems: ZEL024 H1650 <-> HEK293 and ZEL031 A549 <-> THP1. Direct source-signature reuse, labeled identity, was compared with ridge transfer, partial least squares (PLS), and k-nearest-neighbor transfer.

Metric. The primary metric was held-out row-wise cosine similarity in centered-response space. Centering subtracts source and target training means before fitting and evaluation, focusing the benchmark on perturbation-linked variation rather than constant cell-line baseline shifts.

At the largest training sizes, learned maps consistently improved on identity:

Library	Direction	Train size	Identity	Ridge transfer	PLS transfer
ZEL024	H1650 -> HEK293	5,000	0.146	0.371	0.358
ZEL024	HEK293 -> H1650	5,000	0.146	0.279	0.275
ZEL031	A549 -> THP1	3,000	0.094	0.242	0.242
ZEL031	THP1 -> A549	3,000	0.094	0.212	0.214

Interpretation. Molecule-level transfer is detectable and useful as a prioritization signal. Direct reuse leaves substantial target-cell information uncaptured, while simple regularized maps recover additional response structure in every eligible direction. The absolute centered-response cosines also show why the molecule is not the strongest current unit of transfer: many compounds have limited well support, and low-pass transcriptomic measurements are noisy at single-molecule resolution. Z-Screen becomes more powerful when chemistry is organized at the program level.

Raw-profile benchmarks are retained as supplementary diagnostics because they capture baseline cell-state translation. In raw space, transfer scores are much higher; for example, ZEL024 HEK293 -> H1650 reaches 0.546 with ridge transfer at train size 5,000 versus 0.063 with direct reuse. The centered-response benchmark remains the main discovery readout because it better isolates perturbation-linked signal.

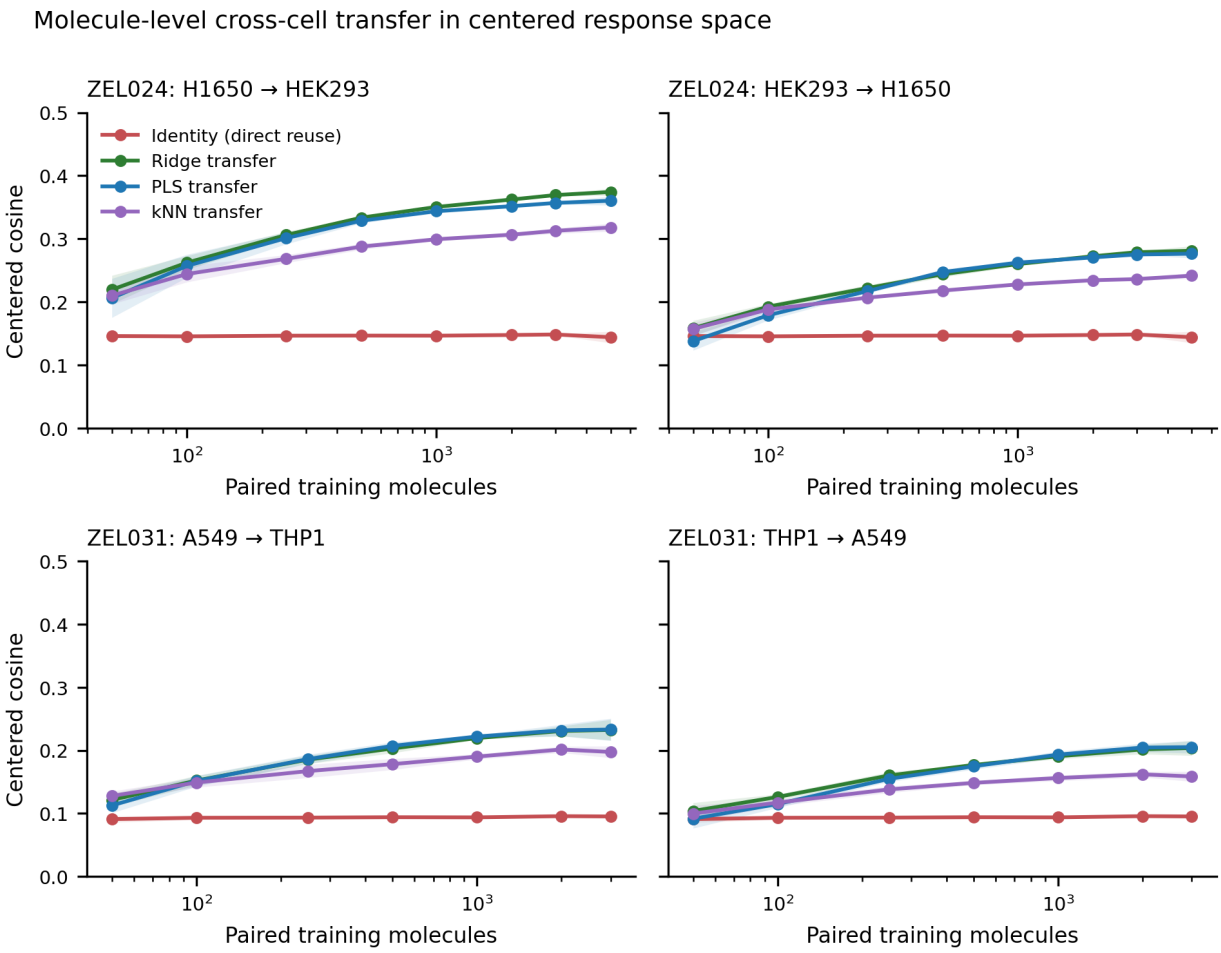


Figure 2. Molecule-level cross-cell transfer in centered-response space across four well-supported cell-pair directions in ZEL024 and ZEL031. Direct reuse of the source signature is consistently weaker than learned transfer; ridge and PLS transfer are nearly tied.

Chemistry-resolved response programs transfer strongly across cell types

Question. Are chemical-response programs defined by shared building-block provenance more reusable across cell types than individual molecules?

Comparison. We averaged molecules into grouped programs at aggregation levels that remain chemically interpretable. In ZEL024, which has four substantive building-block positions, the main program unit is a bb0+bb1 pair. In ZEL031, which has two substantive positions, the bb0 group is the closest comparable program because the other position remains the varying partner chemistry. These groupings preserve enough chemical specificity to support series-level decisions while increasing measurement support per response centroid.

Metric. The benchmark used centered ridge transfer and summarized held-out performance as support-weighted cosine similarity across shared programs. Support weighting gives greater influence to program centroids with more measured wells.

Library	Direction	Group definition	Shared groups	Median support	Identity	Ridge transfer	Gain
ZEL024	H1650 -> HEK293	bb0+bb1	84	240.5	0.471	0.863	0.392
ZEL024	HEK293 -> H1650	bb0+bb1	84	240.5	0.471	0.808	0.336
ZEL031	A549 -> THP1	bb0	129	157.0	0.419	0.714	0.296
ZEL031	THP1 -> A549	bb0	129	157.0	0.419	0.664	0.245

Interpretation. This is the central result. Chemistry-resolved grouped programs transferred with support-weighted cosine 0.664-0.863 across four cell-pair directions, improving over identity by 0.245-0.392 absolute cosine. The result turns cross-cell transfer from a molecule-by-molecule prediction exercise into a platform scaling strategy. A deep source-cell profile can reveal reusable response programs; paired landmark maps can translate those programs into target contexts; and follow-up screens can focus on the programs most likely to matter in the next cell type.

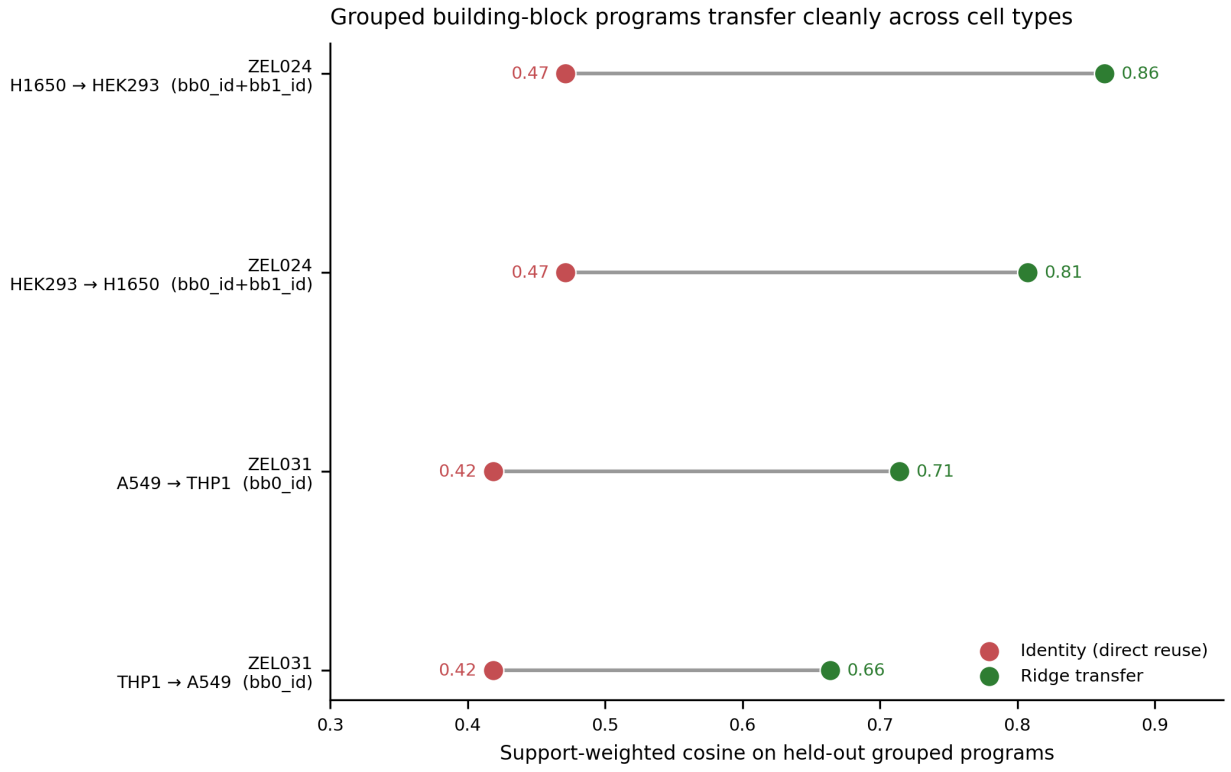


Figure 3. Grouped building-block program transfer at chemistry-resolved aggregation levels (bb0+bb1 for ZEL024, bb0 for the two-position ZEL031 library). Identity sits in the 0.42 to 0.47 range, while a regularized learned transfer map reaches support-weighted cosine 0.66 to 0.86 across four chemistry-resolved program-and-direction combinations.

The side-by-side summary highlights the scale advantage. At the molecule level, learned transfer roughly doubled centered-response cosine over identity in ZEL031 and improved ZEL024 by 0.13-0.23 absolute cosine. At the grouped-program level, learned transfer improved over identity by 0.25-0.39 absolute cosine, with all four reported directions ending in the 0.66-0.86 range.

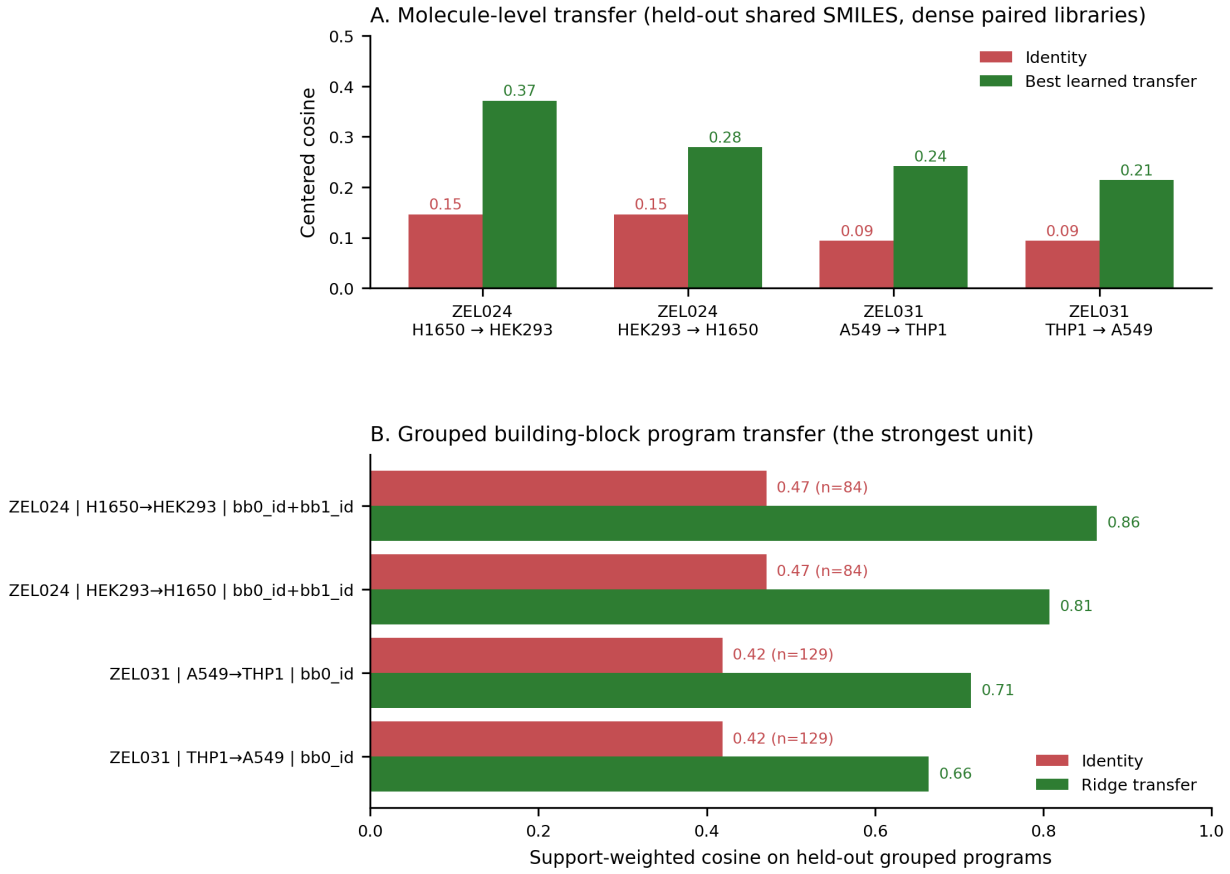


Figure 4. Best learned transfer versus identity at both molecule level (top) and grouped building-block program level (bottom). Learned transfer wins in every eligible cell-pair direction tested, with the largest gains at the grouped-program level.

Aggregation guardrails preserve chemical specificity

Question. When grouped-transfer scores are high, do they still represent chemically specific programs?

Comparison. ZEL028-2 provides an important guardrail. It has four substantive building-block positions, but some of its strongest grouped-transfer rows come from a single building-block position. A single-position centroid in this library averages over many distinct tuples and can report broad

library/cell-pair structure rather than a specific chemical-response program.

Metric. In the full grouped-transfer table, ZEL028-2 single-building-block rows reached high weighted cosines. For example, bb1 grouping reached 0.850 for A549 -> HEK293 and 0.837 for H1650 -> HEK293. More specific BB-pair rows in the same library were much shallower per group and scored lower.

Direction	Group	Class	Groups	Median support	Transfer	Gain	Interpretation
A549 -> HEK293	bb1_id	single-BB aggregate	86	526.5	0.850	0.332	high transfer but over-aggregated
H1650 -> HEK293	bb1_id	single-BB aggregate	86	333.5	0.837	0.366	high transfer but over-aggregated
HEK293 -> A549	bb1_id	single-BB aggregate	86	526.5	0.762	0.244	high transfer but over-aggregated
A549 -> HEK293	bb1_id+bb2_id	BB-pair group	6,690	5.0	0.227	0.100	more specific but shallow per group
A549 -> HEK293	bb3_id+bb4_id	BB-pair group	7,158	5.0	0.215	0.089	more specific but shallow per group
HEK293 -> A549	bb1_id+bb2_id	BB-pair group	6,690	5.0	0.213	0.086	more specific but shallow per group

Interpretation. The guardrail strengthens the grouped-program claim by defining what does and does not count as a reusable chemical-response program. High transfer is most meaningful when

the grouping remains chemically specific enough to guide synthesis or follow-up selection. ZEL024 bb0+bb1 and ZEL031 bb0 satisfy that standard in their library designs. ZEL028-2 single-building-block aggregates are useful diagnostics of transferable broad structure, but they are not the main platform evidence.

Cross-library projection marks the next scaling frontier

Question. Can a transfer map learned in one combinatorial library be applied to another library measured in the same cell-pair direction?

Comparison. A centered ridge transfer matrix was trained in one library and applied to compounds from another library. This setting is harder than within-library transfer because the held-out library can differ in building-block vocabulary, scaffold composition, sampling depth, and overlap structure.

Metric. Performance was summarized as support-weighted cosine in centered-response space, comparing identity with ridge transfer in the held-out library.

Projection	Identity	Ridge transfer	Gain
ZEL024 -> ZEL028-2, H1650 -> HEK293	0.033	0.055	0.022
ZEL024 -> ZEL028-2, HEK293 -> H1650	0.033	0.059	0.026
ZEL028-2 -> ZEL024, H1650 -> HEK293	0.119	0.121	0.002
ZEL028-2 -> ZEL024, HEK293 -> H1650	0.119	0.067	-0.052
ZEL028-2 -> ZEL031, A549 -> H1650	0.120	0.216	0.096
ZEL028-2 -> ZEL031, H1650 -> A549	0.120	0.139	0.019
ZEL031 -> ZEL028-2, A549 -> H1650	0.103	-0.009	-0.112
ZEL031 -> ZEL028-2, H1650 -> A549	0.103	0.024	-0.080

Interpretation. Cross-library projection is a frontier result. Some directions were positive, including both ZEL024 -> ZEL028-2 HEK293/H1650 projections and ZEL028-2 -> ZEL031 for A549 -> H1650. Other directions were weak or negative. The pattern shows that cross-cell maps can carry signal beyond a single library, but teacher-library depth, target-library support, and chemistry-domain match matter. The next prospective experiment should be designed explicitly for this question, with planned shared chemistry, balanced cell-line sampling, scaffold-family overlap, and predefined rejection criteria.

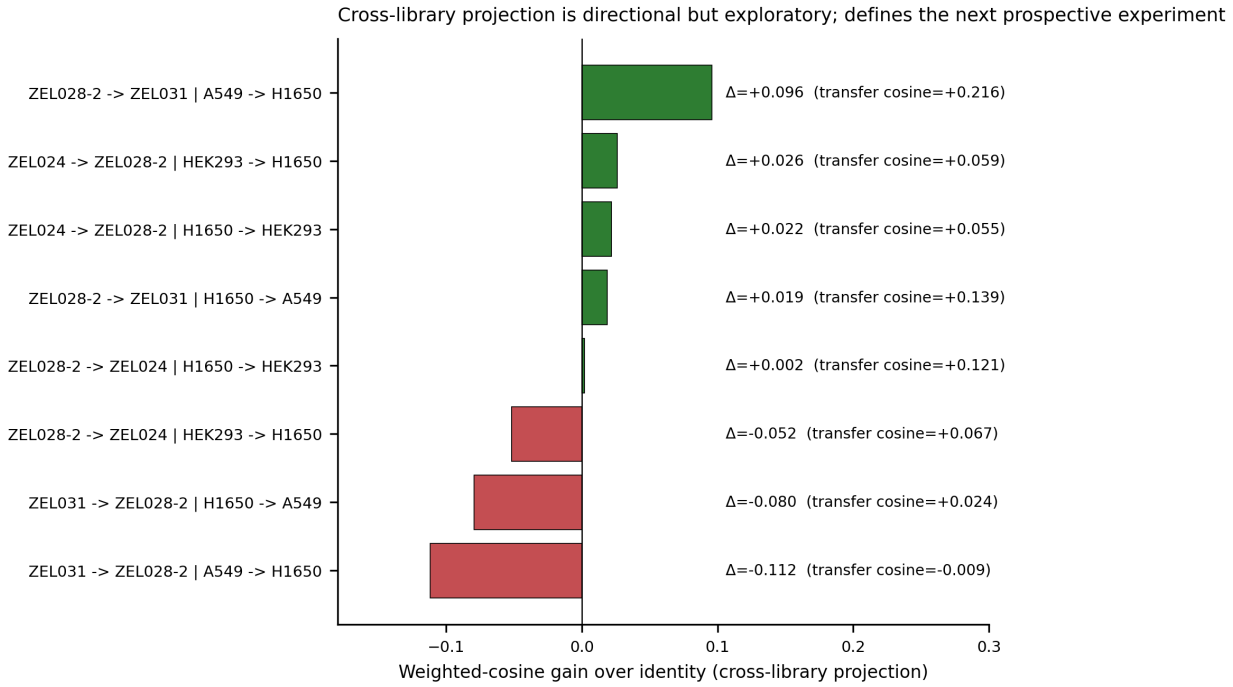


Figure 5. Cross-library projection results. Some cross-library projections are positive, but performance remains small, asymmetric, and sensitive to which library acts as teacher.

Gene-level decoding turns transferred latents into ranked hypotheses

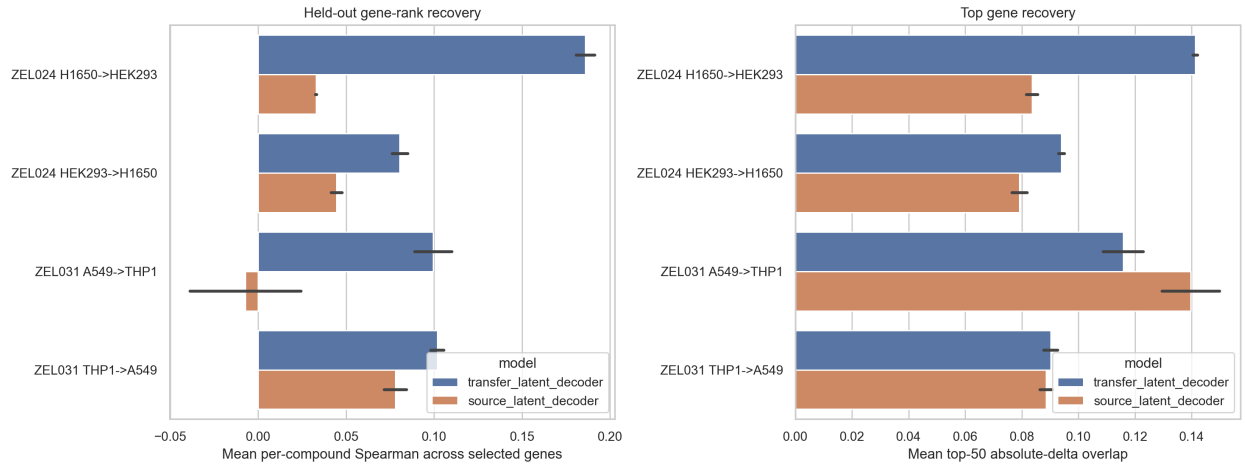
Question. Can transferred latent profiles be decoded into ranked gene-level hypotheses in the target cell type?

Comparison. This experiment composed two models: a source-to-target latent transfer map fit on shared compounds, and a target-cell ridge decoder fit from target-cell scVI centroids to log-CPM pseudobulk deltas over 800 target-training variable genes. A rank signature is the ordered list of genes most increased or decreased by a perturbation. The goal here was ranked gene-hypothesis generation, not full transcriptome reconstruction.

Metric. Held-out predictions were scored by per-compound Pearson and Spearman correlations over gene deltas and by top absolute-delta gene overlap. The clearest gain occurred in ZEL024 H1650 -> HEK293: the direct source-latent decoder reached mean gene-delta Spearman 0.033 and top-50 absolute-delta overlap 0.084, while the transferred-latent decoder reached Spearman 0.186 and top-50 overlap 0.141. An oracle decoder using the measured target latent reached Spearman 0.353 and top-50 overlap 0.187. The reverse ZEL024 direction and both ZEL031 directions produced smaller transferred-latent Spearman values of 0.081-0.102.

Interpretation. Gene-level decoding extends cross-cell transfer from latent prioritization toward interpretable biology. The current signal is sufficient to nominate ranked target-cell genes or pathways for inspection, especially in ZEL024 H1650 -> HEK293, while the oracle gap shows room for better target-cell decoders and richer training data. This is a natural next layer for Z-Screen:

transfer the program, decode the target-cell hypothesis, then measure the most decision-relevant cell type directly.



Supplementary Figure 2. Exploratory gene-level transfer. A target-cell decoder composed with the learned source-to-target latent map partially recovers held-out gene-rank structure, with the clearest gain in ZEL024 H1650 -> HEK293 and weaker gains in the remaining directions.

Discussion

Cross-cell transfer is a platform scaling advantage for Z-Screen. The strongest evidence is that chemistry-resolved response programs, not just individual compounds, can be translated across cell types with high support-weighted cosine. In ZEL024, bb0+bb1 programs transferred between H1650 and HEK293 at 0.808-0.863. In ZEL031, bb0 programs transferred between A549 and THP1 at 0.664-0.714. These are reusable chemical-response programs: they retain building-block specificity, average over enough measured chemistry to stabilize the RNA signal, and map onto the way discovery teams reason about series-level follow-up.

Molecule-level transfer provides a complementary result. Learned maps improved over direct source reuse in every eligible dense direction, with centered-response cosine reaching 0.279-0.371 in ZEL024 and 0.212-0.242 in ZEL031. These values are not the headline because sparse single-molecule profiles are noisy, but they matter operationally. They show that paired landmark data capture compound-specific target-cell information and can help rank molecules within or around a transferable program.

The ZEL028-2 guardrail clarifies the chemical standard for this claim. Single-building-block aggregates in a four-position library can score highly because they average broad library structure. Those rows are valuable diagnostics, but they are too coarse to act as the main chemical-program evidence. The stronger platform interpretation comes from groupings that preserve chemistry specificity while increasing support: bb0+bb1 in ZEL024 and bb0 in the two-position ZEL031 design.

Cross-library and gene-level analyses define the next build-out. Cross-library projection already shows directional gains, but performance depends on which library acts as teacher and how well the

chemistry domains overlap. Gene-level decoding shows that transferred latents can be converted into ranked target-cell hypotheses, with a clear ZEL024 H1650 -> HEK293 gain and measurable headroom relative to oracle target-latent decoding. Together, these analyses point to a prospective architecture in which Z-Screen builds a shared chemical-response latent space, learns cell-type-specific adapters, rejects out-of-domain projections, and allocates new target-cell measurements where uncertainty and expected information gain are highest.

The practical workflow is immediate. Profile a combinatorial library deeply in a source cell type. Use a paired landmark panel to learn how source responses translate into a target cell type. Identify chemical-response programs that are likely to retain, lose, invert, or shift activity across contexts. Then measure the most informative programs and representative molecules directly in the target cell line. This turns cross-cell biology from an exhaustive grid into an adaptive measurement problem.

Methods

Data inputs

All analyses used the canonical Z-Screen RNA aggregate dataset distributed with this repository:

- `data/ZScreen_Canonical_Dataset/RNASeqAggregate/canonical_obs.parquet`
- `data/ZScreen_Canonical_Dataset/RNASeqAggregate/canonical_scvi_latents.parquet`
- `data/ZScreen_Canonical_Dataset/RNASeqAggregate/canonical_chemistry.parquet`
- `data/ZScreen_Canonical_Dataset/RNASeqAggregate/chem_embed.parquet`

Rows were retained when they had a non-null `smiles_hash` and `chemistry_grain == "building_block_annotation"`. In the public bundle, `smiles_hash` is the chemistry-resolved compound identifier used for joins; raw Z-Screen SMILES are not included. Compound-control-only rows were excluded from the primary transfer benchmark because their annotations are mechanism-of-action labels rather than combinatorial tuple coordinates.

Latent representation

Transfer benchmarks used 32-dimensional scVI latent coordinates [1]. Each well in the RNA aggregate is a low-pass pseudobulk measurement from a small number of cells, so individual genes are noisy at the single-well level. The scVI latent provides a denoised representation for comparing chemistry-linked states across cell types. Gene-level recovery was evaluated separately with the exploratory decoder analysis.

Centered-response benchmark

Cell lines differ in baseline transcriptional state. A raw-space model can therefore score well by learning constant cell-line offsets. For the primary benchmark, source and target training means were subtracted before model fitting and evaluation. This centered-response formulation focuses performance on perturbation-linked variation and is the closest available proxy for response in the absence of matched vehicle controls in the canonical dataset.

Molecule-level transfer

For each library, cell line, and compound identifier, scVI latent coordinates were averaged across wells. Shared compounds between a source and target cell line were split into training and held-out test sets. Identity, global-shift where applicable, ridge transfer, PLS transfer, and k-nearest-neighbor transfer were compared. Performance was measured as row-wise cosine similarity on held-out compounds and averaged across repeated splits. The manuscript reports centered-response results as the primary molecule-level analysis and raw-profile results as supplementary diagnostics.

Grouped building-block transfer

Within each library and cell-pair direction, compounds were grouped by candidate building-block keys such as `bb0_id`, `bb1_id`, and multi-column combinations such as `bb0_id+bb1_id`. For each shared group, source and target centered latent signatures were averaged. A centered ridge transfer model was fit across repeated train-test splits, and performance was summarized as support-weighted cosine similarity across held-out groups. The main manuscript reports aggregation levels that remain chemically interpretable for the corresponding library design.

ZEL028-2 aggregation guardrail

ZEL028-2 grouped-transfer outputs were reviewed separately because single-building-block centroids in a four-position library collapse many distinct tuples. The guardrail table reports both high-scoring single-building-block rows and more specific BB-pair rows. Single-building-block rows are treated as broad diagnostics rather than chemistry-resolved program claims.

Cross-library projection

For a fixed source-target cell-line direction, a centered ridge transfer matrix was trained in one library and applied to another library. Because sparse libraries are singleton-heavy, adaptive minimum-well thresholds were used to retain feasible exploratory comparisons while preserving stricter filters in denser settings. Directional asymmetry was reported directly because teacher-library depth, target-library support, and chemistry-domain match are part of the result.

Exploratory gene-level decoder

For the four dense paired directions, a source-to-target latent transfer map was fit on shared training compounds. A separate target-cell ridge decoder was trained to map target-cell scVI centroids to log-CPM pseudobulk deltas over 800 variable genes selected from target training cells. Held-out predictions were scored by per-compound Pearson and Spearman correlation over gene deltas and by top-50 and top-100 absolute-delta gene overlap. This analysis was used for ranked gene-hypothesis generation.

Limitations

- Matched vehicle controls are unavailable in the canonical dataset, so centered response is a proxy for perturbational response rather than a fully DMSO-normalized effect.
- The primary molecule-level benchmark is limited to the dense paired libraries ZEL024 and ZEL031; ZEL028-2 becomes too sparse under strict per-molecule support filters.
- The strongest current transfer unit is the chemistry-resolved grouped program, with molecule-level transfer acting as supporting prioritization evidence.
- ZEL028-2 single-building-block results are intentionally treated as aggregation guardrails because they over-aggregate a four-position library.
- Cross-library projection depends on teacher-library depth, target-library support, and chemistry-domain match; the prospective version should include planned shared chemistry and balanced sampling.
- Gene-level decoding should be read as ranked hypothesis generation, not full transcriptome reconstruction.
- Current transfer models use averaged 32-dimensional scVI latent signatures and do not model dose, time, metabolism, tissue architecture, or organism-level context.

Data availability

Data tables, manuscript figures, and analysis inputs needed to reproduce this manuscript are organized under `paper4/` and `data/ZScreen_Canonical_Dataset/` in this repository. The public bundle replaces raw Z-Screen SMILES with `smiles_hash` identifiers and precomputed irreversible chemistry embeddings as described in the repository README. A persistent-archive deposition with an assigned DOI will accompany the corresponding preprint posting.

Code availability

Analysis and figure-generation scripts are in `paper4/scripts/`. Package-level dependencies are in the repository root `requirements.txt`. The package was tested with Python 3.14.3 on Windows.

References

1. Lopez R, Regier J, Cole MB, et al. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15:1053-1058. doi:10.1038/s41592-018-0229-2
2. Subramanian A, Narayan R, Corsello SM, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171(6):1437-1452.e17. doi:10.1016/j.cell.2017.10.049
3. Tahoe Therapeutics. Tahoe-100M: A 100-million single-cell perturbational atlas across 50 cancer cell lines. Dataset card and February 2025 release. https://huggingface.co/datasets/tahoe-bio/Tahoe-100M
4. Dapello J, Nassar M, Eksi R, et al. scGeneScope: A treatment-matched single-cell imaging and transcriptomics dataset and benchmark for treatment response

modeling. NeurIPS 2025 Datasets and Benchmarks. OpenReview. https://openreview.net/pdf/f7d541dae38bcf88a79789a4c6440aadfec123c7

Z-Screen maps combinatorial small-molecule tuple states into CRISPR perturbation spaces

Abstract

Z-Screen is a chemical-genetic mechanism-mapping framework that connects combinatorial small-molecule tuple states to CRISPR perturbation spaces, including paired CRISPR references for pleiotropic state interpretation. Each Z-Screen tuple is the full ordered set of building blocks that forms one molecule, and each chemical or CRISPR perturbation was represented as a signed rank signature and calibrated with a size-matched permutation null. The best-supported system was ZEL024 / HEK293: 8,599 four-position tuples were compared with 8,603 Replogle K562 knockout signatures, yielding 1,276 tuples matched to 1,714 CRISPR programs at empirical q <= 0.05; after removing the 50 most recurrent CRISPR neighbors, 1,236 tuples and 1,664 programs remained. Collapsing the same wells to single-building-block or pair averages produced almost disjoint CRISPR neighborhoods, showing why the actionable unit is the full tuple. Same-cell THP1 controls recovered BRD4 for MZ1 and interferon-pathway proxies for STC-15. Norman paired CRISPR activation states showed that tuple scores preserve genetic sum-of-singles structure, supporting combinatorial references for deconvolving mechanism hypotheses behind small-molecule pleiotropy. BAY-293/V-ATPase and ZEL028-2 ATP6V1A/FKBP9 nominate concrete follow-up hypotheses for validation. Together, these results establish a chemistry-resolved map from combinatorial small-molecule states to genetic perturbation state space.

Significance

Functional genomics and medicinal chemistry usually enter a discovery campaign through different doors. CRISPR screens nominate perturbation states; chemistry screens nominate molecules. Z-Screen links the two by mapping exact small-molecule tuple states into single-gene, paired-gene, and pathway-level CRISPR perturbation spaces.

The key output is a calibrated, chemistry-addressable hypothesis table. A match means that a compound-level transcriptional state overlaps a CRISPR perturbation state more than expected under a matched-size null. Because small molecules can engage primary targets, collateral targets, pathway feedback, and compensatory states at the same time, each CRISPR neighbor is treated as a candidate mechanism axis for validation. The value of tuple resolution is that every hypothesis points back to exact building-block coordinates that can be resynthesized, varied, and tested.

Introduction

Single-cell CRISPR perturbation atlases have made genetic states searchable at large scale. Perturb-seq and related methods measure the transcriptional consequence of perturbing genes in pooled single-cell experiments [1,2]. The Replogle K562 atlas provides genome-scale knockout signatures [4]. The Norman K562 CRISPR-activation atlas includes both single-gene perturbations and paired perturbations, making it possible to ask how a double perturbation relates to its component singles [3]. scPerturb harmonizes many public single-cell perturbation datasets across cell types and technologies [9].

For chemical biology, the corresponding question is richer than "what is the target of this molecule?" Small molecules can produce composite states: partial modulation of a primary target, collateral target engagement, complex disruption, feedback, stress, adaptation, or convergence onto a downstream pathway. A single CRISPR neighbor can nominate one axis of a molecule-induced state, but it should not be read as a direct biochemical mechanism without orthogonal validation. Paired CRISPR references add a second layer: they let us ask whether a chemical state resembles a combination of genetic axes more than either axis alone, which is the relevant setting for deconvolving mechanism hypotheses behind pleiotropy.

Connectivity Map and LINCS L1000 established the idea of matching compound and genetic expression signatures [5,6], and newer resources such as L2S2 support direct chemical perturbation to CRISPR signature search [7]. Z-Screen adds chemistry resolution. Each Z-Screen well contains a one-bead-one-compound combinatorial molecule and produces a per-well pseudobulk transcriptional readout from the 5 to 10 cell colony in that micro-well. The chemical identity is not only "a compound"; it is an ordered tuple of building blocks. A building block is a chemical subunit installed at a defined library position, such as bb0, bb1, bb2, bb3, or bb4. A tuple is the full ordered set of populated building-block identities for one combinatorial compound. In the ZEL024 headline library, the tuple is bb0+bb1+bb2+bb3; in ZEL028-2, bb1, bb3, and bb4 are variable while bb2 is fixed; in ZEL031, the substantive tuple is bb0+bb1.

This distinction is central because Z-Screen has multiple possible grains. A full tuple is one molecule-level chemistry. A building-block pair or a single-building-block average pools many distinct molecules that share a component. Those aggregated signatures are useful diagnostics, but they are not equivalent chemistry units. In this manuscript, "tuple resolution" means the full molecule-level state, while single-BB and BB-pair signatures are used to test what is lost when chemistry is over-collapsed.

All comparisons are made through rank signatures rather than shared normalized expression vectors. A rank signature is the ordered set of genes most increased and most decreased relative to an appropriate background, with direction preserved. A Z-Screen signature compares a target tuple, subset, or control compound against the matching chemical background; a CRISPR signature compares a perturbation against controls in its source dataset. A chemistry-CRISPR score combines mimic overlap, reverse overlap, net mimic overlap, and rank cosine over shared top genes. A calibrated match is a retained chemistry-CRISPR pair whose observed score is evaluated against a

size-matched permutation null and assigned a Benjamini-Hochberg q-value. This calibration answers a narrow, reproducible question: do the strongest up/down genes agree more than expected by chance for a reference signature of the same size?

Two recurrent-neighbor terms are used throughout. A recurrent neighbor is a CRISPR perturbation that appears as a top match for many chemical queries. A hub is an especially recurrent CRISPR neighbor, operationally flagged by its recurrence rank. Hubs can represent broadly responsive biological states, so the headline table is accompanied by hub-strip diagnostics that remove the most recurrent CRISPR programs and recount surviving calibrated matches.

Finally, ZEL028-2 requires a hierarchy because it has many observed tuples but shallow per-tuple support. The full-tuple layer gives the most chemistry-specific hypotheses, usually at only two wells per tuple. Variable-pair subset signatures, such as bb1+bb3, bb1+bb4, and bb3+bb4, aggregate more wells and test whether part of the same chemistry supports the same CRISPR neighbor. The strictest hierarchy tier requires a non-hub full-tuple call plus calibrated subset support at q <= 0.05. Top-50-only subset support is reported separately as a broader, hypothesis-generating tier.

The goal of this manuscript is to test Z-Screen as a chemical-genetic mechanism map. We ask whether tuple-resolved small-molecule states can be placed in the same rank-signature frame as public CRISPR states, whether the resulting matches are calibrated and robust to recurrent-neighbor artifacts, and whether the matches remain attached to actionable chemistry coordinates.

Results

A calibrated rank-signature framework compares chemical and CRISPR perturbation states

The first question is whether Z-Screen chemistry and public CRISPR atlases can be compared without forcing incompatible datasets into a single expression normalization. Z-Screen rows are per-well pseudobulks from one-bead-one-compound micro-wells; Replogle and Norman are K562 single-cell CRISPR datasets; scPerturb is a harmonized collection of public perturbation datasets. We therefore represented each perturbation as a signed top-gene rank signature and compared rank-level agreement.

The comparison uses four related quantities: mimic overlap, reverse overlap, net mimic overlap, and rank cosine over shared top genes. Mimic overlap counts genes moving in the same direction in chemistry and CRISPR; reverse overlap counts opposite-direction agreement; net mimic overlap subtracts reverse from mimic. Rank cosine adds rank-order information when enough top genes are shared. For each chemical query, the top 50 CRISPR neighbors were retained, and the top 10,000 rows by net mimic and total mimic overlap were calibrated with 1,000 size-matched random gene sets per CRISPR signature. Empirical p-values were converted to q-values across calibrated rows. The full reproduction order and table outputs are in `paper5/README.md`.

The output is a calibrated state hypothesis, not a direct biochemical mechanism. A strong calibrated match says that the chemical state resembles a CRISPR state more than expected under the

null. The match may reflect direct target engagement, an upstream or downstream pathway, a compensatory program, or a convergent transcriptional response. The Results therefore report the evidence level for each class of claim: ZEL024 / HEK293 as the best-supported tuple system, same-cell THP1 controls as validation, Norman doubles as combinatorial reference geometry, and BAY-293 plus ZEL028-2 as follow-up hypotheses.

Full tuples are the actionable chemistry unit (Figure 1)

The next question is whether building-block aggregation preserves the same CRISPR neighborhood as full-tuple chemistry. ZEL024 / HEK293 provides the cleanest test because it has 13,931 observed four-position tuples and 8,599 tuples with at least 10 wells. The same wells can also be collapsed into 84 bb0+bb1 pair signatures or 83 single-bb3 signatures. Those aggregates pool many molecules: the bb0+bb1 pair signatures average a median of 165 distinct tuples, and the bb3 single-position signatures average a median of 168 distinct tuples. Other single-position aggregates in ZEL024 / HEK293 can average thousands of tuples.

For each of the 8,599 ZEL024 / HEK293 tuples, we compared its top five Replogle CRISPR neighbors with the top five neighbors from the containing bb0+bb1 pair and from the containing bb3 single-building-block diagnostic. The top-five overlap was almost always zero. Tuple versus bb3 had zero overlap for 99.686 percent of tuples, with median Jaccard 0.000; tuple versus bb0+bb1 pair had zero overlap for 99.721 percent, also with median Jaccard 0.000 (`paper5/tables/ZEL024_HEK293_resolution_summary.csv`).

The conclusion is practical: coarser averages can produce nearly independent CRISPR neighborhoods because they mix many distinct compounds. If a discovery team wants to act on a CRISPR-like state, the state needs to be attached to the molecule-level tuple that produced it. Figure 1 therefore justifies making ZEL024 / HEK293 full tuples, not single-building-block aggregates, the headline analysis grain.

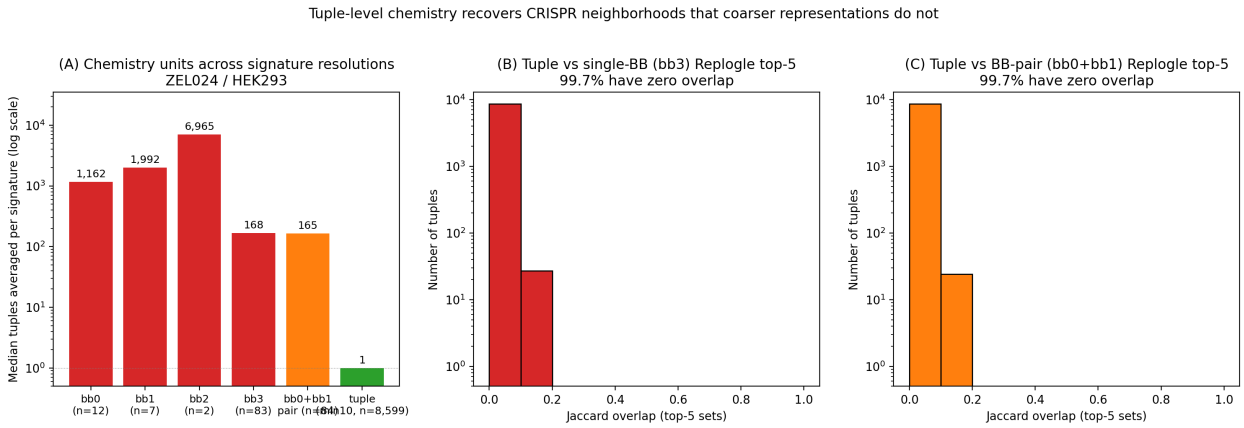


Figure 1. Aggregation diagnostics for ZEL024 / HEK293. (A) Median number of distinct chemical tuples averaged into each diagnostic signature, log-scale. Single-BB signatures average over many distinct compounds sharing one building-block position; tuple signatures average over one full molecule-level tuple. (B) Top-5 Replogle CRISPR matches at tuple resolution share zero genes

with the containing single-BB bb3 diagnostic for 99.686 percent of 8,599 tuples. (C) The same comparison at tuple-versus-bb0+bb1 pair diagnostic resolution, with 99.721 percent zero overlap. The finding argues against chemical over-collapse, not for treating single-BB averages as chemistry units.

ZEL024 / HEK293 maps tuple chemistry onto a broad CRISPR-state space (Figure 2)

The headline result is that full-tuple chemistry produces calibrated CRISPR-state matches at scale. We compared 8,599 ZEL024 / HEK293 tuple signatures with 8,603 Replogle K562 knockout signatures, a search space of approximately 74 million chemistry-CRISPR scores. The top 10,000 rows were calibrated by the size-matched permutation null.

All 10,000 calibrated rows passed q <= 0.05, spanning 1,276 distinct chemical tuples and 1,714 distinct CRISPR programs (`paper5/tables/ZEL024_HEK293_tuples_min10_vs_replogle_calibrated.csv`). The strongest deduplicated CRISPR programs reached net z from 12.8 to 14.8 and concentrated in interpretable transcriptional state neighborhoods, including DNA replication and chromatin maintenance (MCM2, SLBP, CHAF1B, ARIH1, ADSL), mitochondrial function (MRPL3, MRPS30, COX17, COA5, ATP5F1A, GFM1, HSPE1), and trafficking or nonsense-mediated decay (EFR3A, DDOST, SMG5).

The table also contains chemical convergence across independent tuples. Among the top-100 best-z tuples, 12 CRISPR programs were reached by two or more chemically distinct tuples. EFR3A appeared with seven distinct tuples, TONSL with six, MCM2 with five, and GART, MRPL42, EIF4G2, MRPL53, and COA5 each recurred three times (`paper5/tables/ZEL024_HEK293_tuples_min10_vs_replogle_dedup_top_per_crispr.csv`). These recurrences do not prove direct mechanisms, but they are the expected pattern if independent chemistry coordinates can converge on similar CRISPR-like transcriptional states.

We then asked whether the map was dominated by recurrent CRISPR hubs. After progressively removing the 5, 10, 25, and 50 most recurrent CRISPR perturbations, the q <= 0.05 map remained broad. Dropping the top 50 recurrent neighbors still left 1,236 distinct chemical tuples and 1,664 distinct CRISPR programs (`paper5/tables/ZEL024_HEK293_tuples_min10_vs_replogle_hub_strip_summary.csv`). The calibrated ZEL024 / HEK293 tuple map is therefore not explained by a small set of broad CRISPR neighbors.

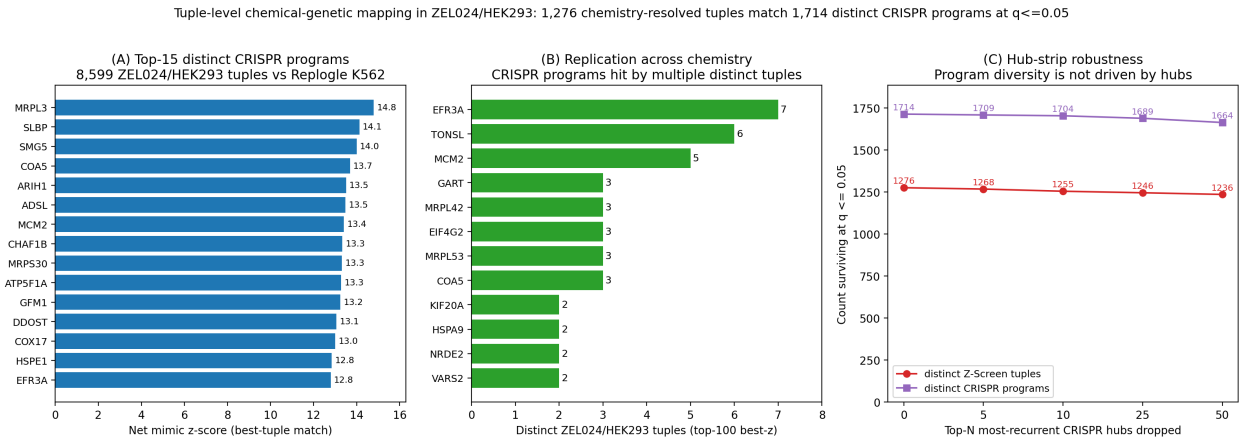


Figure 2. Tuple-level chemistry maps onto 1,714 distinct CRISPR programs in ZEL024 / HEK293. (A) Top-15 distinct CRISPR programs by best Z-Screen tuple z-score, deduplicated to one tuple per CRISPR program. (B) Replication-across-chemistry: among the top-100 best-z tuples, CRISPR programs reached by multiple chemically distinct tuples. (C) Hub-strip robustness: the calibrated top-10,000 table after dropping the top-N most-recurrent CRISPR perturbations. Tuple and program diversity at q <= 0.05 are essentially unchanged at N = 50.

Same-cell controls validate the state-matching framework (Figure 3)

The same-cell THP1 control panel anchors the interpretation of the larger cross-cell maps. It provides two complementary cases: one in which the direct genetic analog is represented, and one in which the useful CRISPR neighbor is a downstream pathway proxy.

MZ1 is a BRD4-targeting PROTAC degrader [10]. The THP1 MZ1 control library was sparse, so its compound signature used a same-cell compound-pool fallback background rather than a same-library vehicle control. Under that broader background, MZ1 recovered BRD4 as the top calibrated CRISPR-like signature in the THP1 scPerturb panel (net mimic = 20, rank cosine = 0.539, net z = 6.84, q = 0.015; `paper5/tables/internal_control_summary_combined.csv`). Neighboring matches fell in BRD4-adjacent transcriptional and chromatin programs and are treated as state-neighborhood matches, not additional direct target calls.

STC-15 is a METTL3 inhibitor, but METTL3 is absent from the local THP1 CRISPR panel. The expected same-cell question is therefore whether STC-15 recovers a downstream pathway proxy rather than the missing direct genetic analog. It did: IRF1, an interferon-pathway proxy for METTL3-inhibition biology [11], was the second calibrated match (net z = 3.06, q = 0.045), with IRF7 third at a weaker q-value. This validates the intended interpretation rule. When the direct analog is represented, the framework can recover it; when it is absent, the top interpretable hit may be a downstream state.

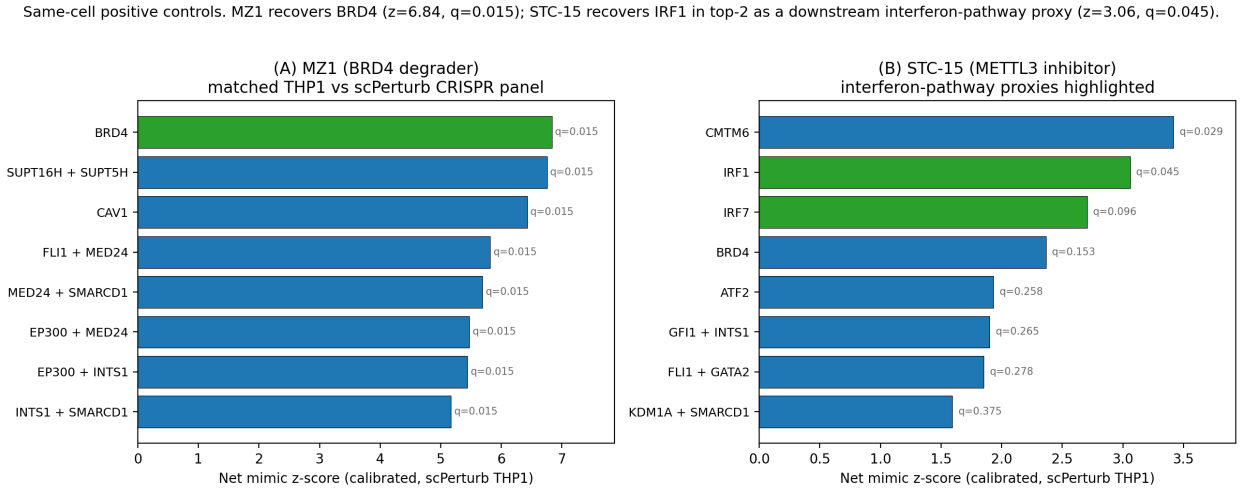


Figure 3. Same-cell positive controls in matched THP1 with calibrated empirical q-values from a permutation-null gene-set calibration. (A) MZ1 (BRD4-selective PROTAC degrader): BRD4 is the top calibrated CRISPR-like signature (z = 6.84, q = 0.015), with subsequent neighbors falling in BRD4-adjacent transcriptional and chromatin programs. (B) STC-15 (METTL3 inhibitor): METTL3 itself is absent from the local CRISPR panel; the downstream interferon-pathway proxy IRF1 is recovered in the top two calibrated matches (z = 3.06, q = 0.045).

BAY-293 nominates a V-ATPase and trafficking state hypothesis (Figure 4)

We then asked how the framework behaves when a compound has a known primary pharmacology but the top transcriptional neighborhood is not the annotated target. BAY-293 is an SOS1 inhibitor profiled in A549 in a non-combinatorial control library. In the documented-target recovery table, SOS1 is present but not dominant in the A549 ranking (rank 2,106 within the ranking universe used by `paper5/tables/known_biology_recovery.csv`). The calibrated Replogle neighborhood instead concentrates on V-ATPase and vesicular-trafficking perturbation states.

The top 12 calibrated BAY-293 A549 matches all had empirical q = 0.003, with net z from 9.5 to 11.9. They included V-ATPase complex genes ATP6V1A, ATP6V1H, ATP6V1B2, ATP6V1C1, and ATP6V1E1, plus trafficking-associated genes ZW10, SACM1L, TMED2, ZNHIT1, WDR7, CCDC115, and CLTC (`paper5/tables/zscreen_compounds_vs_replogle_calibrated.csv`). This is a bridge example, not a target claim. The result says that the BAY-293-induced A549 state resembles V-ATPase and trafficking CRISPR states more strongly than its SOS1 knockout-like state in this cross-cell rank comparison.

That distinction creates a useful follow-up axis. RAS signaling and endosomal trafficking intersect biologically, so the V-ATPase/trafficking neighborhood is plausible enough to test. It also provides context for ZEL028-2, where an unrelated combinatorial tuple later nominates ATP6V1A with calibrated subset support. The shared gene label is best read as a recurrent state hypothesis that links two independent observations.

Figure 4. BAY-293 bridge example. The SOS1 inhibitor BAY-293 in A549 matches the V-ATPase complex (red) and vesicular-trafficking CRISPR knockouts (blue) at calibrated empirical q = 0.003. Top hit ZW10 reaches net z = 11.86; nine V-ATPase or trafficking components reach net z above 10.

Paired CRISPR states help parse pleiotropy while documented-target recovery stays conditional (Figure 5)

Norman provides the pleiotropy-aware test. For each of 131 Norman K562 paired CRISPR-activation perturbations, we compared the vector of ZEL024 / HEK293 tuple scores against the double with the sum of the tuple scores against the two constituent singles. All 131 doubles produced positive Spearman correlations with p <= 0.05; the median rho was 0.353. For 98.5 percent of doubles, the best tuple's score against the double exceeded its score against either single (`paper5/tables/norman_tuple_combinatorial_logic_summary.csv`).

This result does not say that a molecule is literally a two-gene perturbation. It says that tuple-level chemical scores respect the combinatorial structure of a paired genetic atlas. That is the rationale for using paired CRISPR references to interpret pleiotropic molecule-induced states: a strong A+B state match can prioritize candidate axes, interactions, or downstream programs for validation without assigning two direct biochemical targets.

The same-cell controls test clean examples. A broader documented-target scan asks a harder question: when a compound has a known primary target, does a cross-cell CRISPR rank list place that target unusually high? In `paper5/tables/known_biology_recovery.csv`, 34 compound-cell-line comparisons cover 21 Z-Screen compounds. Among the 27 comparisons evaluated in the

8,839-entry ranking universe used for the Replogle-based target-recovery scan, seven placed the documented target in the top 15 percent: FGFR1 for derazantinib in HEK293, SMARCA4 for SMARCA ligand 1 in A549, MAPK14 for AZD-7624 in A549, METTL3 for STC-15 in A549, FGFR1 for fexagratinib in A549, PARP2 for veliparib in A549, and KRAS for MRTX-1719 in A549. STC-15 in HEK293 was scored against a larger merged Replogle+Norman universe and is reported separately in the table.

The misses are informative rather than incidental. Cobimetinib, GSK126, BAY-293, and AZD-7624 illustrate why chemical-CRISPR similarity is not target deconvolution. A molecule can inhibit a protein without producing a knockout-like transcriptional profile in the tested cell context, and cross-cell ranking can move an annotated target up or down depending on pharmacology, cell state, timing, and pathway feedback. The appropriate conclusion is enrichment with important conditions, not universal target recovery.

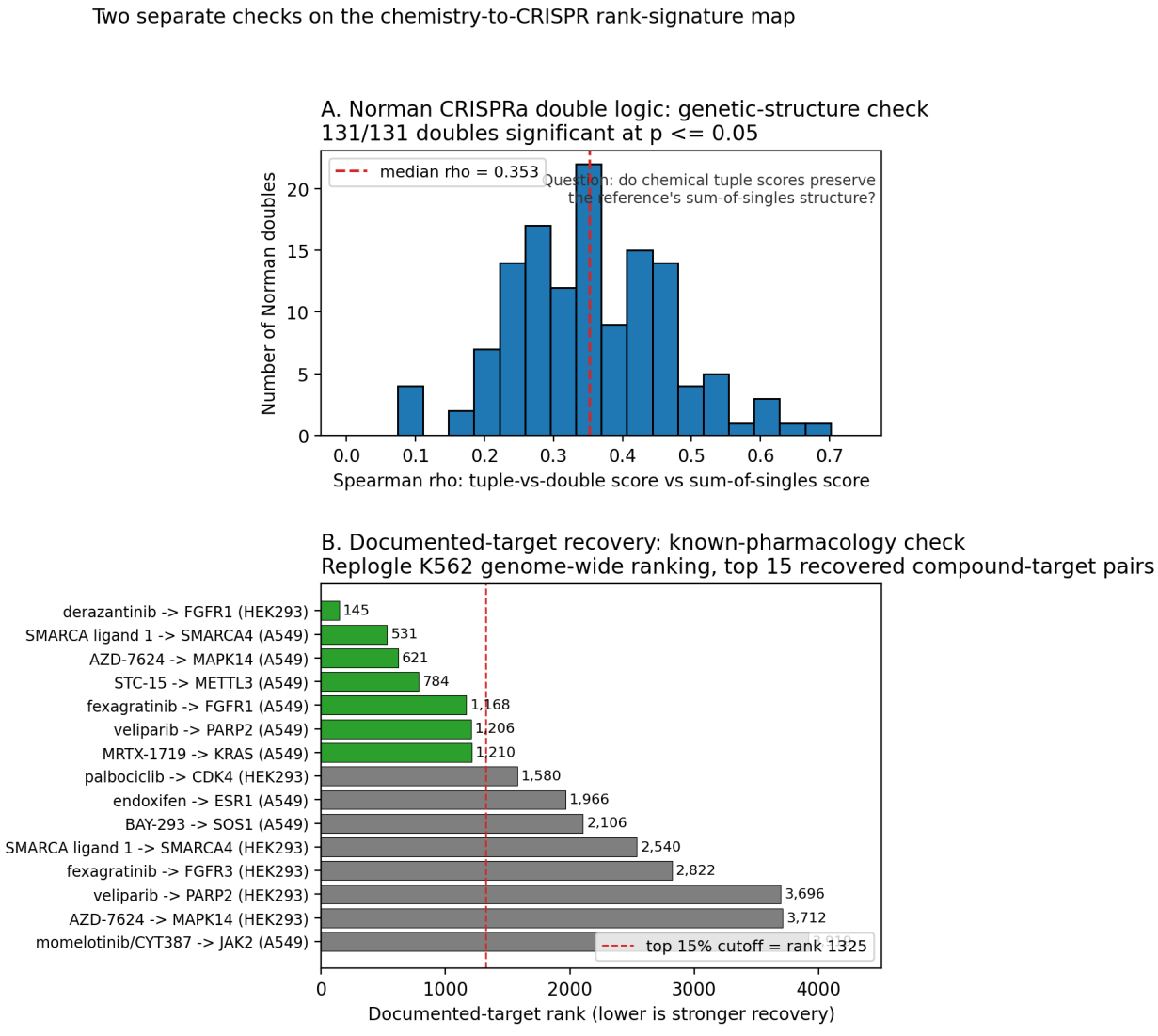


Figure 5. Two separate validation checks. (A) Norman combinatorial logic for pleiotropy-aware

interpretation: distribution of Spearman correlations across 131 Norman K562 doubles between tuple-vs-double match scores and sum-of-singles match scores; all 131 doubles significant at p <= 0.05; median rho = 0.353. The result tests whether tuple scores respect known single-versus-paired genetic structure, a requirement for using paired CRISPR states to parse composite chemical phenotypes. (B) Documented-target recovery: ranks within the Replogle K562 cross-cell ranking for the top-15 best-recovered compound-target pairs; the dashed line marks the top 15 percent of the 8,839-perturbation reference.

ZEL028-2 uses a hierarchy to separate shallow tuple hypotheses from corroborated calls (Supplementary Figures 2 and 3)

ZEL028-2 asks whether the framework can still be useful when molecule-level coverage is much shallower than ZEL024 / HEK293. The raw observed universe is large: 61,396 full tuples in HEK293, 40,622 in A549, and 25,906 in H1650. However, only 4,041, 1,728, and 765 full-tuple signatures, respectively, qualify at the relaxed >=2-well threshold, and the median qualifying tuple has only two wells (`paper5/tables/resolution_feasibility.csv`). Those full-tuple calls are chemistry-specific but noisy, so they are treated as a hypothesis leaderboard.

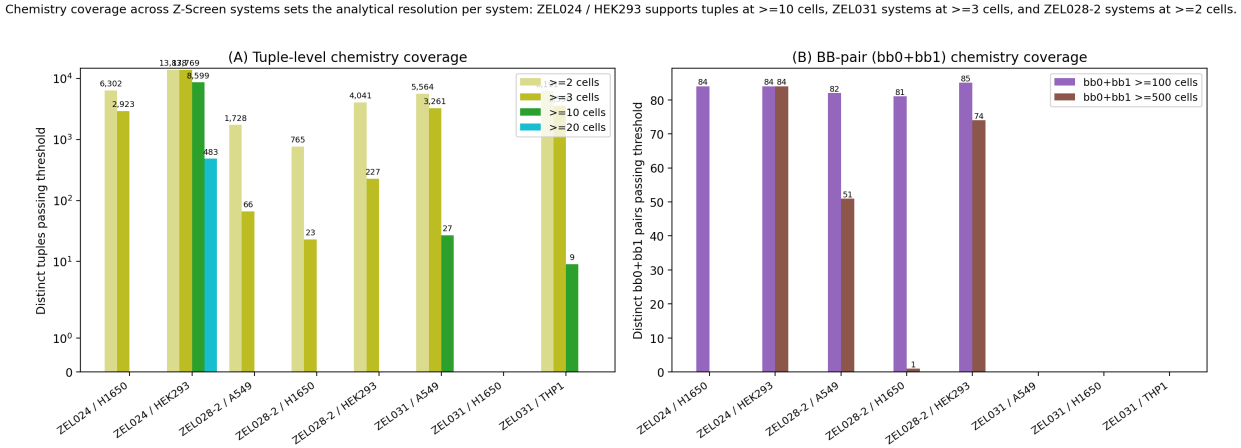
The calibrated min2 full-tuple scans still show structured biology. The deduplicated calibrated table contains 943 HEK293 tuple calls, 760 A549 calls, and 448 H1650 calls, for 2,151 total best full-tuple CRISPR calls used in the hierarchy. The strongest full-tuple matches reached high net z values, and the recurring CRISPR neighborhoods included RNA processing, mitochondrial translation, helicase activity, and translation initiation. However, 1,967 of the 2,151 calls were assigned to top-50 recurrent hub programs, reinforcing the need for a stricter support model.

The hierarchy therefore adds variable-pair subset evidence. In ZEL028-2, bb2 is fixed and bb0 is absent, so the meaningful variable pairs are bb1+bb3, bb1+bb4, and bb3+bb4. At >=10 wells per subset signature, these panels provide 3,051 to 3,103 signatures per pair in HEK293, 909 to 1,260 in A549, and 141 to 257 in H1650 (`paper5/tables/ZEL028-2_subset_coverage.csv`). Each pair panel was scanned against Replogle, and the variable-pair scans were permutation-calibrated with the same size-matched null used for the headline ZEL024 / HEK293 analysis.

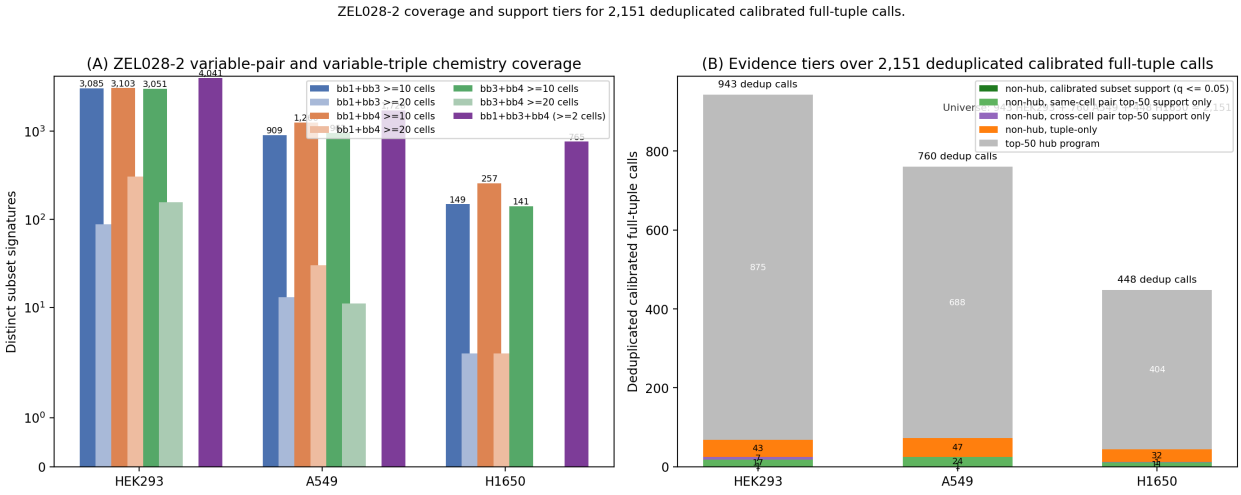
Two non-hub candidates pass the strictest FDR-controlled subset tier: HEK293 ATP6V1A and A549 FKBP9 (`paper5/tables/ZEL028-2_hierarchical_mechanism_support_top_candidates.csv`). The ATP6V1A tuple has full-tuple net z = 15.47 and q = 0.001, with bb3+bb4 subset support at z = 7.94 and q = 0.001. The FKBP9 tuple has full-tuple net z = 11.87 and q = 0.001, with bb1+bb4 subset support at z = 8.96 and q = 0.001. These are exciting because they turn a shallow library into specific chemistry-plus-CRISPR follow-up packages. Eleven additional non-hub candidates have top-50 subset support without calibrated subset support. The calibrated tier and top-50-only tier are kept separate because only the former is FDR-controlled.

ATP6V1A is the clearest cross-observation link in the current data. BAY-293 in A549 produced a calibrated V-ATPase/trafficking state neighborhood that included ATP6V1A, and an unrelated ZEL028-2 HEK293 tuple independently points to the ATP6V1A CRISPR state with same-cell

variable-pair calibrated support. This does not establish V-ATPase target engagement by either chemistry. It does create a concrete follow-up package: a specific ZEL028-2 tuple, its supporting bb3+bb4 subset, and a V-ATPase transcriptional-state hypothesis.



Supplementary Figure 2. Z-Screen chemistry coverage across systems. (A) Distinct tuples passing >=2, >=3, >=10, and >=20 well thresholds in each system. ZEL024 / HEK293 supports tuple-level signatures at >=10 wells, ZEL031 systems at >=3 wells, and ZEL028-2 systems at >=2 wells. (B) Distinct bb0+bb1 pairs passing 100 and 500 well thresholds in libraries that populate bb0; ZEL028-2 has no substantive bb0 position, so its bb0+bb1 pair counts collapse to the same single-bb1 aggregation regardless of bb2/bb3/bb4 and are not used as a chemistry unit anywhere in this manuscript. The chemistry-resolved ZEL028-2 analysis uses variable-pair (bb1+bb3, bb1+bb4, bb3+bb4) and variable-triple (bb1+bb3+bb4) signatures (Supplementary Figure 3). ZEL024 supports both tuple and bb0+bb1 pair resolution; ZEL031 has too many distinct pairs for any single pair to carry meaningful well counts.

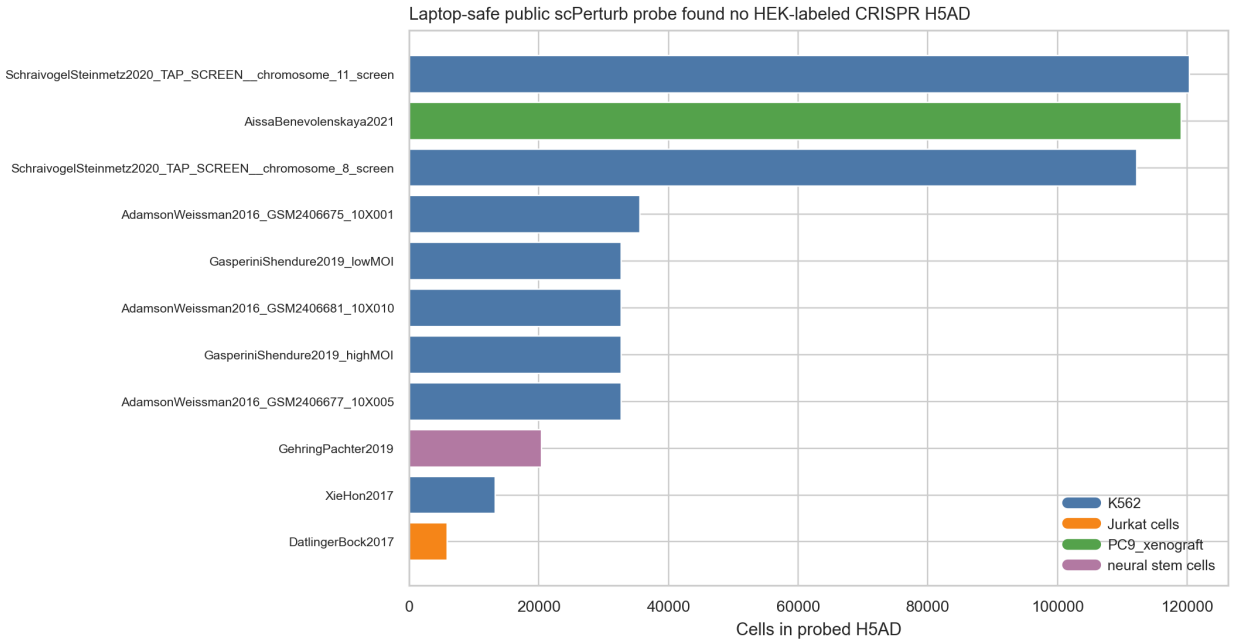


Supplementary Figure 3. ZEL028-2 multi-resolution chemistry coverage and hierarchical mechanism support. (A) Variable-pair (bb1+bb3, bb1+bb4, bb3+bb4) subset signatures qualifying at >=10 wells per signature and at >=20 wells per signature, plus the variable-triple bb1+bb3+bb4

at >=2 wells, across HEK293, A549, and H1650. (B) Hierarchical mechanism-support breakdown of the 2,151 deduplicated calibrated full-tuple calls. Bars are stacked from bottom to top: non-hub with calibrated subset support at q <= 0.05 (FDR-controlled tier; 2 candidates total: HEK293 ATP6V1A and A549 FKBP9), non-hub with same-cell variable-pair top-50 support only, non-hub with cross-cell variable-pair top-50 support only, non-hub tuple-only, and top-50 hub programs. Total counts above each bar give the number of deduplicated calibrated full-tuple calls per cell line.

A targeted public probe did not identify a matched HEK293 single-cell CRISPR reference (Supplementary Figure 1)

The HEK293-rich Z-Screen systems would ideally be benchmarked against a matched HEK293 single-cell CRISPR atlas. A laptop-scale probe of 11 scPerturb H5AD files under 500 MB read observation metadata only and searched for HEK293 or HEK293T CRISPR datasets. The probed files covered K562, Jurkat or T-cell, PC9 xenograft, and neural stem cell perturbation panels; none surfaced a HEK-labeled CRISPR dataset (`paper5/tables/scperturb_h5ad_probe_summary.csv`). HEK293-to-Replogle comparisons are therefore reported as cross-cell rank-signature state hypotheses. This probe is not a comprehensive public-data survey.



Supplementary Figure 1. Targeted probe of 11 small (<500 MB) scPerturb H5AD files for matched HEK293 single-cell CRISPR data. Each row reports cell-line label, perturbation type, perturbation count, and dataset dimensions; no HEK-labeled CRISPR dataset was identified locally.

Discussion

The central result is that Z-Screen turns combinatorial chemistry into a searchable genetic-state map. In ZEL024 / HEK293, 8,599 molecule-level tuples produce calibrated matches to a broad Replogle CRISPR-state space, and those matches remain broad after recurrent-neighbor removal.

The result matters because the unit of action is exact chemistry: each hit points back to a bb0 / bb1 / bb2 / bb3 tuple that can be resynthesized, varied, and retested.

The resolution result is as important as the scale result. Collapsing the same wells to single-building-block or pair averages produced almost disjoint CRISPR neighborhoods, so the biological hypothesis changes when the chemistry is over-collapsed. For mechanism mapping, the full tuple is not a convenience; it is the coordinate system that keeps a CRISPR-like state experimentally actionable.

The validation results set useful interpretation boundaries. Same-cell THP1 controls show that the method can recover a direct genetic analog when represented (MZ1 to BRD4) and a downstream pathway proxy when the direct analog is absent (STC-15 to IRF1). The broader documented-target scan shows enrichment but not universality. This is the expected behavior for transcriptional state matching: powerful for nominating mechanism axes, inappropriate as an unvalidated one-target answer.

Paired CRISPR references make the framework more appropriate for small-molecule biology. Norman doubles show that tuple scores preserve sum-of-singles logic across the ZEL024 / HEK293 library. That finding supports paired CRISPR states as reference axes for composite chemical phenotypes, while preserving the caveat that a paired-state match nominates mechanistic axes and interactions rather than direct dual target engagement.

BAY-293 and ZEL028-2 show how the map becomes an experimental queue. BAY-293 nominates a V-ATPase/trafficking state despite known SOS1 pharmacology, while the ZEL028-2 hierarchy identifies ATP6V1A in HEK293 and FKBP9 in A549 as non-hub, subset-supported candidates. The ATP6V1A call is strengthened as a follow-up hypothesis by the independent BAY-293 V-ATPase/trafficking state, but neither observation should be read as biochemical target identification.

The practical use case is chemistry-resolved state matching against genetic perturbation atlases. A team can begin with a desired CRISPR state, a paired CRISPR state, or a pathway neighborhood and prioritize calibrated Z-Screen tuples that move cells toward similar transcriptional states. The next step is experimental: matched-cell CRISPR validation, biochemical target engagement, resynthesis of prioritized tuples, and follow-up libraries around the implicated building-block coordinates.

Methods

Data inputs

The Z-Screen RNA aggregate is the repaired canonical dataset at `data/ZScreen_Canonical_Dataset/RNASeqAggreg`
Each Z-Screen measurement used here is a per-well pseudobulk from a one-bead-one-compound micro-well. Public CRISPR rank signatures are bundled in `paper5/external_data/`: Replogle K562 genome-scale Perturb-seq [4], Norman K562 CRISPR activation [3], and scPerturb THP1 signatures [9]. The paper5 reproduction order is documented in `paper5/README.md`.

Z-Screen tuple and subset signatures

Tuple signatures were built from wells with building-block annotations. For each library and cell line, wells were grouped by the full populated tuple identifier. ZEL024 / HEK293 retained 8,599 four-position tuples at >=10 wells for the headline analysis, with a stricter >=20-well sensitivity set of 483 tuples. ZEL031 used a >=3-well threshold in A549 and THP1. ZEL028-2 used a >=2-well threshold because deeper tuple thresholds were not supported; the qualifying full tuples were 4,041 in HEK293, 1,728 in A549, and 765 in H1650. For each signature, the background was the same library and cell line excluding the target wells. The effect vector was logCPM(target) minus logCPM(background), and the top 250 up and top 250 down genes were retained.

BB-pair and single-position signatures were built with the same target-versus-background logic but pooled many full tuples sharing one building block or one building-block pair. These signatures are diagnostic aggregations, not molecule-level chemistry units. ZEL028-2 variable-pair and variable-triple signatures were built directly for bb1+bb3, bb1+bb4, bb3+bb4, and bb1+bb3+bb4 because bb2 is fixed and bb0 is absent in that library.

CRISPR rank signatures

The Replogle, Norman, and scPerturb source datasets were converted upstream into the same signed up/down rank format. Replogle contributes 8,603 K562 knockout signatures. Norman contributes 236 K562 CRISPR-activation signatures, including 105 singles and 131 doubles. scPerturb contributes 78 THP1 perturbation signatures from harmonized public perturbation data. Each CRISPR signature uses controls from its source dataset as background.

Chemistry-to-CRISPR comparison and calibration

Each chemistry and CRISPR rank signature was converted into a signed rank vector. Comparisons recorded mimic up-overlap, mimic down-overlap, reverse up-down overlap, reverse down-up overlap, total mimic overlap, total reverse overlap, net mimic overlap, rank cosine, and shared top-gene counts. Only the top 50 CRISPR matches per chemical query were retained for large scans.

Permutation calibration was applied to the top 10,000 rows by net mimic and total mimic overlap. For each row, 1,000 random gene sets matched to the CRISPR up-set and down-set sizes were drawn from the dataset-specific gene universe. Empirical p-values were computed as `(1 + nulls >= observed) / (1 + n permutations)` and converted to Benjamini-Hochberg q-values across calibrated rows. The headline ZEL024 / HEK293 scan used seed 13. This non-parametric calibration was used because the rank-overlap null depends on the signature size and gene universe.

Recurrent-neighbor and resolution diagnostics

Recurrent-neighbor tables count how many distinct chemical queries nominate each CRISPR perturbation in the retained top-K and calibrated tables. Hub-strip summaries remove the top N most recurrent CRISPR perturbations and recount calibrated rows, distinct chemical queries, and distinct CRISPR programs at q <= 0.05.

The ZEL024 / HEK293 resolution comparison used each tuple's top five Replogle neighbors and compared them with the top five neighbors from the containing bb0+bb1 pair and bb3 single-building-block diagnostic. Pairwise Jaccard overlaps were summarized in `paper5/tables/ZEL024_HEK293_resolution_summary.csv`.

Same-cell controls and documented-target recovery

MZ1 and STC-15 were evaluated against matched-cell THP1 scPerturb references. Their source control libraries were sparse, so compound rank signatures used a same-cell compound-pool fallback background: all THP1 compound-treated wells excluding the target compound. This broader background is noted in the interpretation, especially for absolute z-scores. The integrated control summary is `paper5/tables/internal_control_summary_combined.csv`.

The documented-target recovery scan compares Z-Screen compound signatures with CRISPR rank lists and records whether documented primary targets appear high in the resulting ranking. The output is `paper5/tables/known_biology_recovery.csv`. Because many comparisons are cross-cell and some compounds act through catalytic inhibition, degradation, allostery, or pathway feedback, the table is interpreted as enrichment and diagnostic context rather than a direct target-deconvolution benchmark.

Norman combinatorial logic

For each of the 131 Norman double-gene perturbations, the corresponding single-gene perturbations were identified in the Norman metadata. Across ZEL024 / HEK293 tuples, the Spearman correlation was computed between the tuple-vs-double net mimic score and the sum of tuple-vs-single-A plus tuple-vs-single-B scores. The same logic was rerun on the ZEL028-2 variable-pair and variable-triple panels. This test evaluates whether paired CRISPR references preserve useful combinatorial state geometry for chemical signatures; it does not assign two direct targets to a molecule.

ZEL028-2 hierarchy

The ZEL028-2 hierarchy starts from deduplicated calibrated full-tuple calls, one best CRISPR call per represented tuple query, across HEK293, A549, and H1650. Each row is annotated for same-cell variable-pair top-50 support, same-cell variable-pair calibrated support at q <= 0.05, cross-cell variable-pair support, variable-triple support, and whether the CRISPR perturbation is a top-50 recurrent hub in the full-tuple calibrated table. The strict tier requires a non-hub full-tuple call and at least one calibrated variable-pair subset support at q <= 0.05. The broader tier requires top-50 subset support without calibrated subset support. Outputs are `paper5/tables/ZEL028-2_hierarchical_mechanism_support.csv`, `paper5/tables/ZEL028-2_hierarchical_mechanism_support_summary.csv`, and `paper5/tables/ZEL028-2_hierarchical_mechanism_`

Limitations

The framework compares transcriptional states, not biochemical binding. A chemistry-to-CRISPR match remains a state-level hypothesis until validated by orthogonal target engagement, genetic rescue, matched-cell perturbation, or independent profiling. This caveat is strongest for cross-cell comparisons such as HEK293 Z-Screen tuples versus K562 Replogle knockouts.

Same-cell CRISPR coverage is currently THP1-dominant. The targeted scPerturb probe did not identify a local HEK293 single-cell CRISPR reference, but it was intentionally laptop-scale and not exhaustive. The HEK293 results should therefore be read as cross-cell state hypotheses, not cell-of-origin target calls.

Tuple resolution depends on coverage. ZEL024 / HEK293 is the strongest tuple-level system because it supports 8,599 tuples at >=10 wells. ZEL028-2 full-tuple calls are much shallower, usually two wells per tuple, and are therefore treated as hypothesis-generating unless supported by the hierarchy. ZEL031 systems are lower-coverage two-position tuple analyses. All output tables carry the well-count and aggregation metadata needed to evaluate any individual match.

Recurrent CRISPR perturbations are real structural features of large perturbation atlases. The hub-strip diagnostic shows that the headline ZEL024 / HEK293 result remains broad after top-hub removal, but individual hub-associated calls should still be interpreted cautiously.

Data availability

The reproducibility package, including the canonical Z-Screen RNA-seq dataset, bundled public CRISPR rank signatures, derived match tables, calibrated outputs, recurrence diagnostics, resolution diagnostics, same-cell controls, hierarchy tables, and figure scripts, is contained in this repository. The repaired canonical RNA-seq aggregate is at `data/ZScreen_Canonical_Dataset/RNASeqAggregate/`. Paper 5 analysis outputs are in `paper5/tables/`, figures in `paper5/figures/`, and public CRISPR rank signatures in `paper5/external_data/`.

Code availability

All paper 5 analysis and figure-generation scripts are in `paper5/scripts/`. The reproduction order is in `paper5/README.md`, and package-level dependencies are listed in the repository root `requirements.txt`. The codebase was tested with Python 3.14.3 on Windows.

References

1. Dixit A, Parnas O, Li B, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell. 2016;167(7):1853-1866.e17.
2. Datlinger P, Rendeiro AF, Schmidl C, et al. Pooled CRISPR screening with single-cell transcriptome readout. Nat Methods. 2017;14:297-301.

3. Norman TM, Horlbeck MA, Replogle JM, et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science. 2019;365(6455):786-793.
4. Replogle JM, Saunders RA, Pogson AN, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell. 2022;185(14):2559-2575.e28.
5. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929-1935.
6. Subramanian A, Narayan R, Corsello SM, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171(6):1437-1452.e17.
7. Torcato M, Niepel M, Kuleshov MV, et al. L2S2: chemical perturbation and CRISPR KO LINCS L1000 signature search engine. Nucleic Acids Res. 2025;53(W1):W338-W350.
8. Lopez R, Regier J, Cole MB, et al. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15:1053-1058.
9. Peidli S, Durakis Green T, Shen C, et al. scPerturb: harmonized single-cell perturbation data. Nat Methods. 2024;21:531-540.
10. Zengerle M, Chan KH, Ciulli A. Selective Small Molecule Induced Degradation of the BET Bromodomain Protein BRD4. ACS Chem Biol. 2015;10(8):1770-1777.
11. Yankova E, Blackaby W, Albertella M, et al. Small-molecule inhibition of METTL3 as a therapeutic strategy against acute myeloid leukaemia. Nature. 2021;593(7860):597-601.