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.

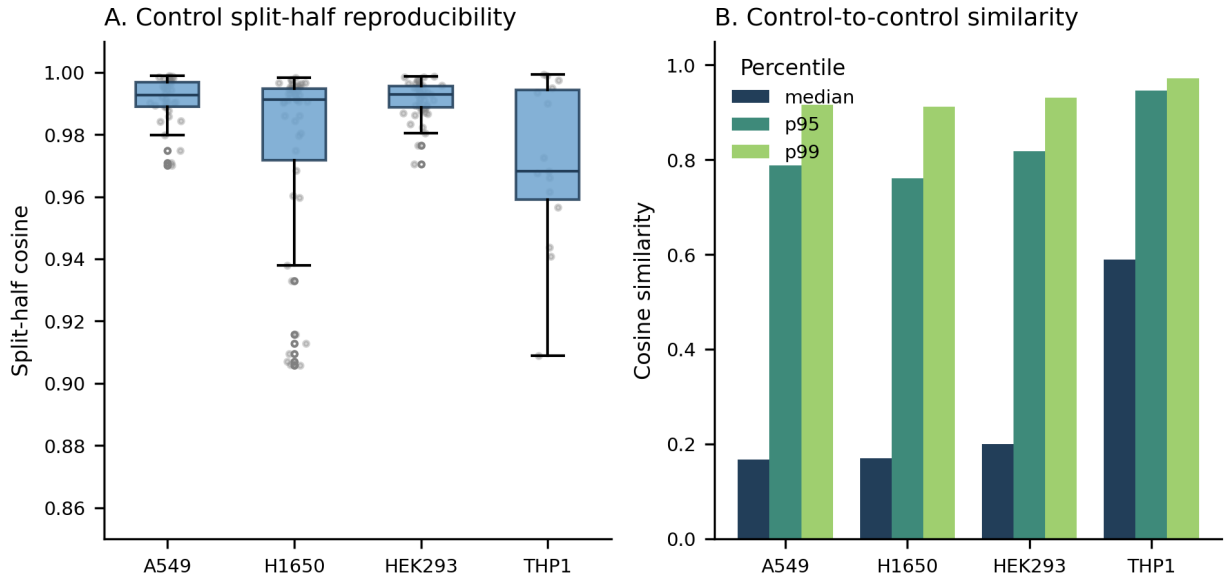


Figure 1. Named controls show high split-half reproducibility across all four cell lines, supporting downstream chemistry-to-RNA modeling.

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.

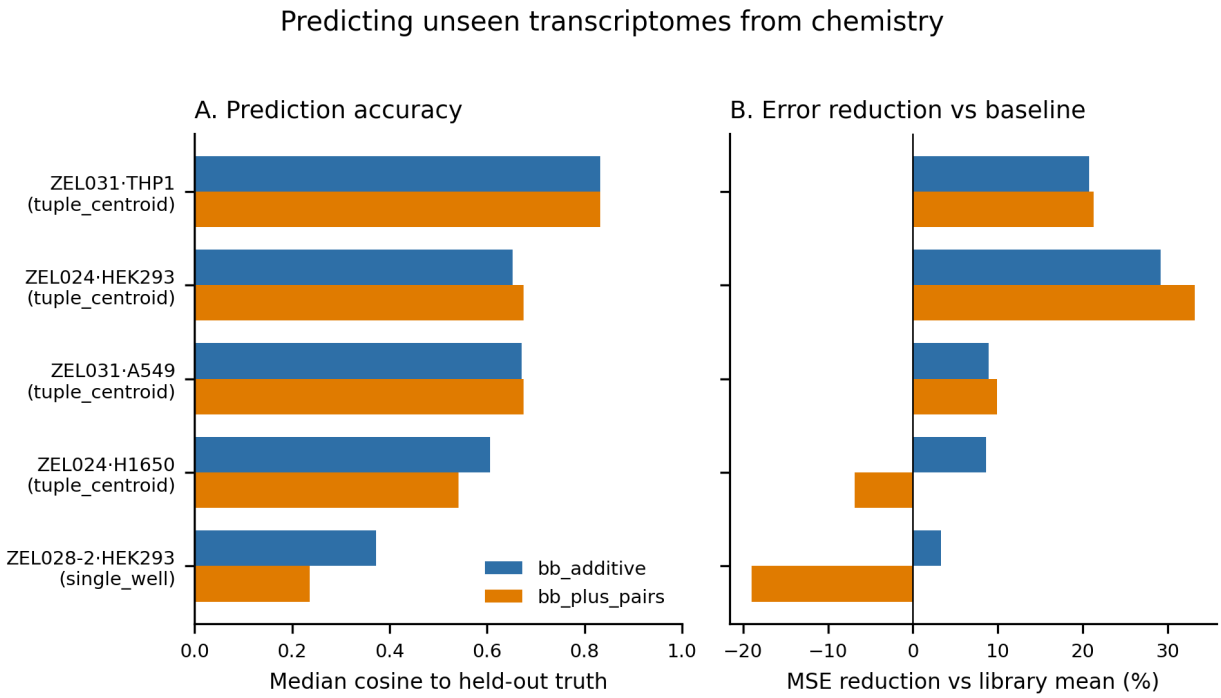


Figure 2. Building-block models improve prediction of held-out transcriptomic states across multiple library and cell-line systems, with the strongest gains in ZEL024 / HEK293 and ZEL031 / THP1.

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.

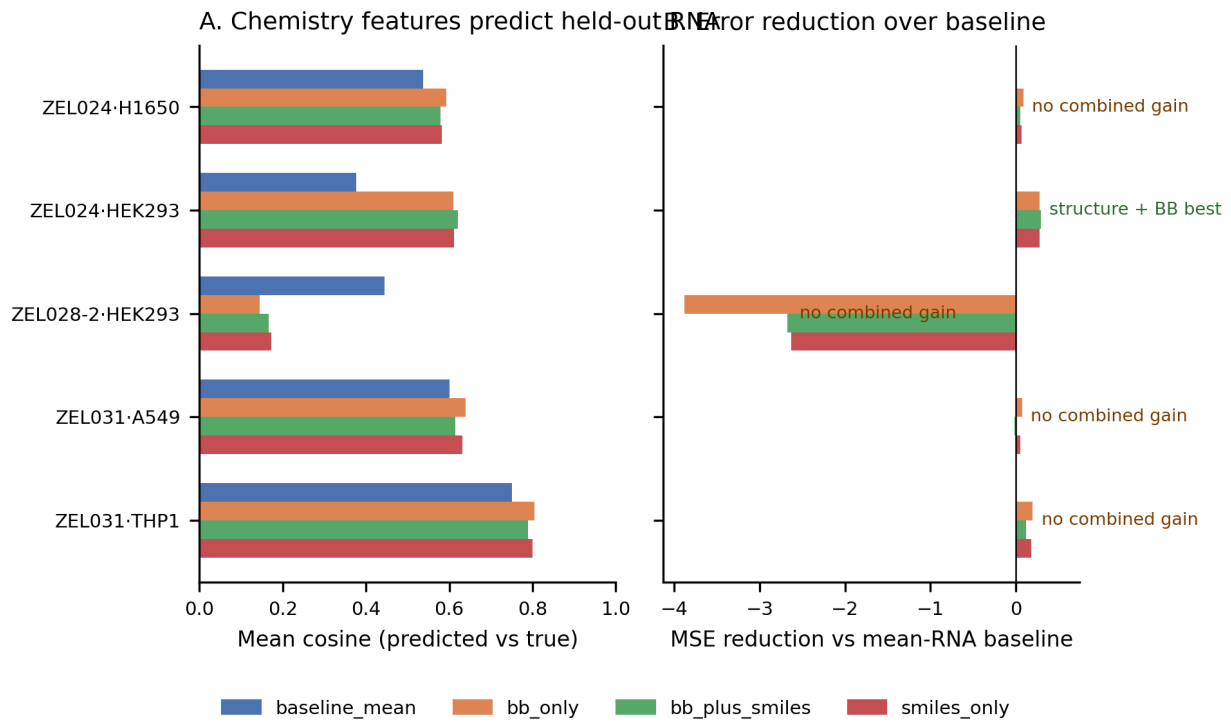


Figure 3. Chemistry-only models capture predictive signal unevenly across systems. Building-block provenance is the most reliable coordinate, and adding structure-derived descriptors improves performance most clearly in ZEL024 / HEK293 while failing to improve or hurting in weaker and sparser settings.

Library tuples recover calibrated phenomimic neighborhoods

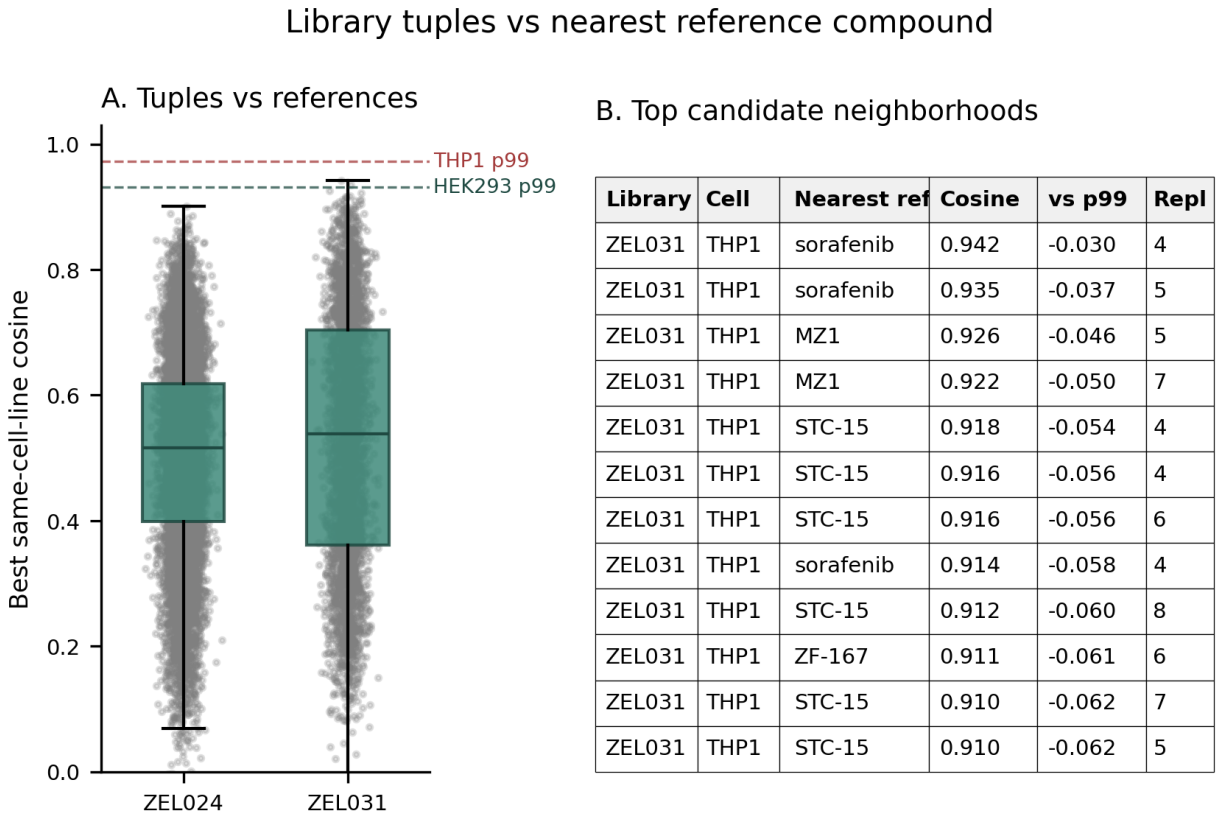
We next asked whether observed library tuples land near known compound states in a way that is useful for triage. The test compared tuple RNA centroids with named-control centroids within the same cell line, then calibrated each apparent match against the same-cell-line distribution of unrelated control-control 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.

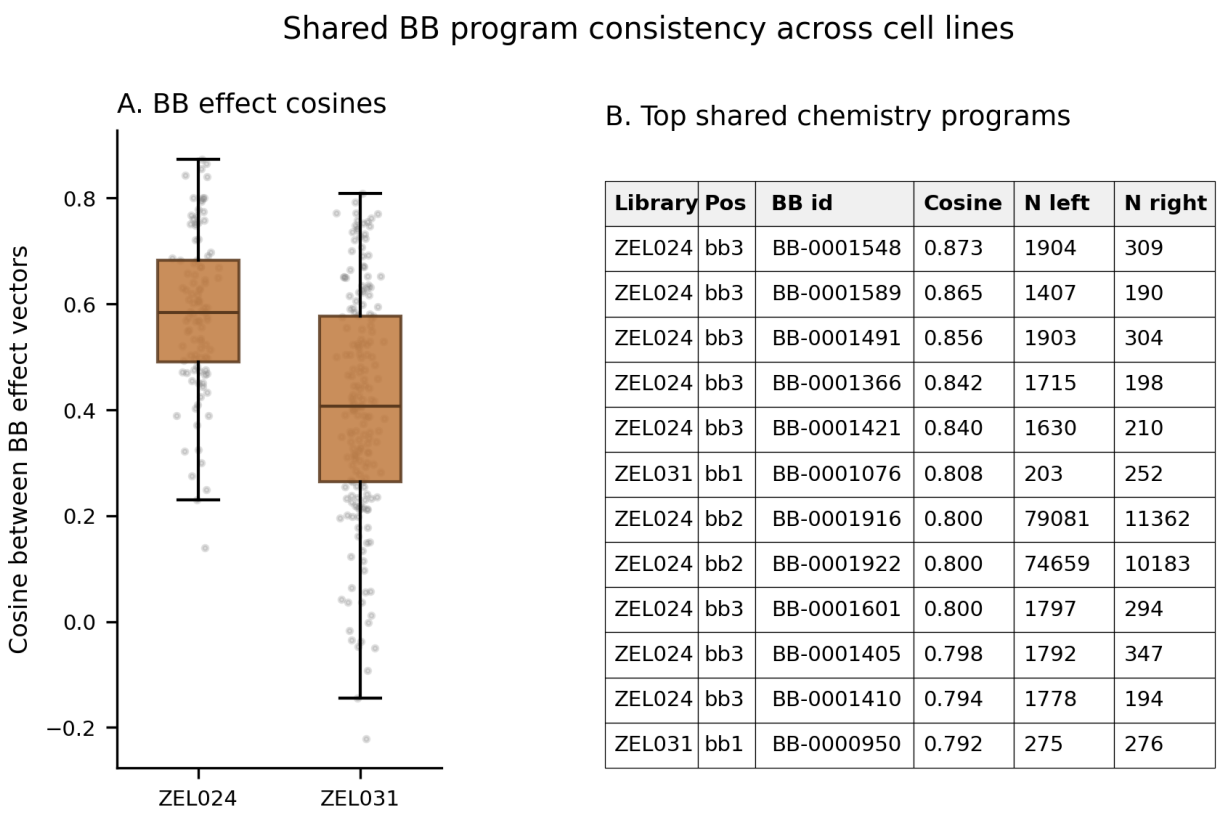
Building-block effects partly recur across cell lines

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.

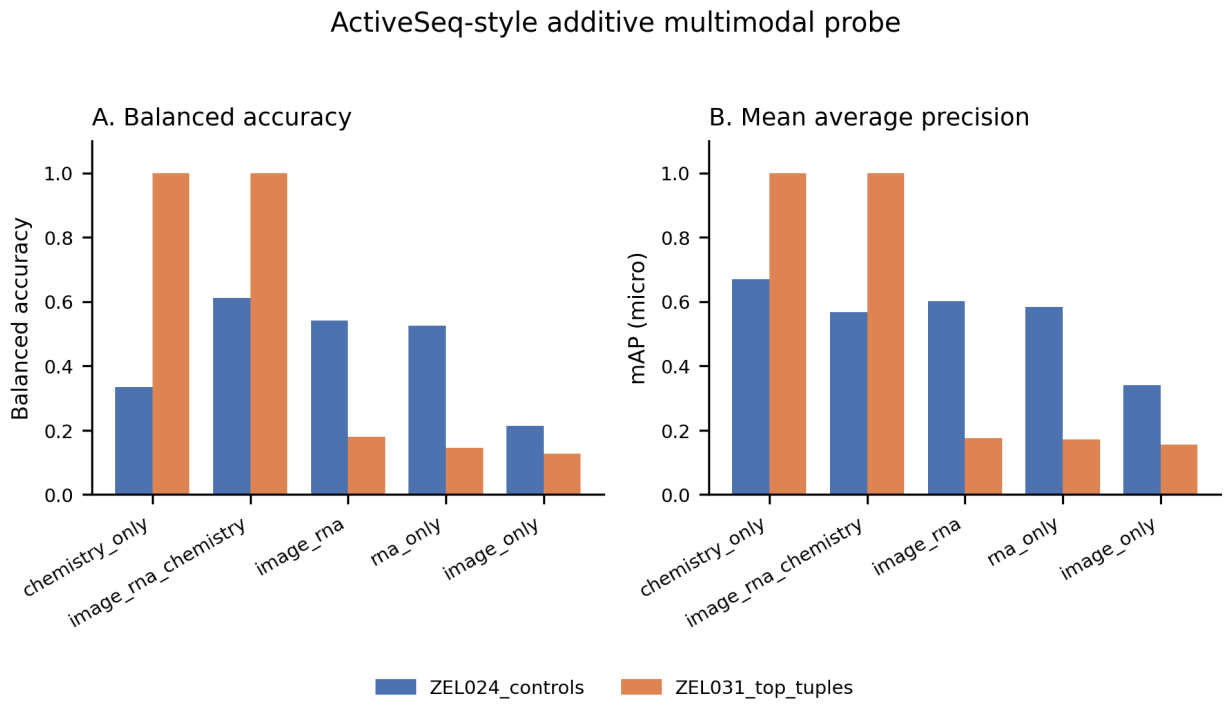


Figure 6. In a simple additive benchmark, image plus RNA modestly outperforms either single modality alone in ZEL024 controls, while RNA and chemistry provenance remain the dominant signals in the current dataset.

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