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.

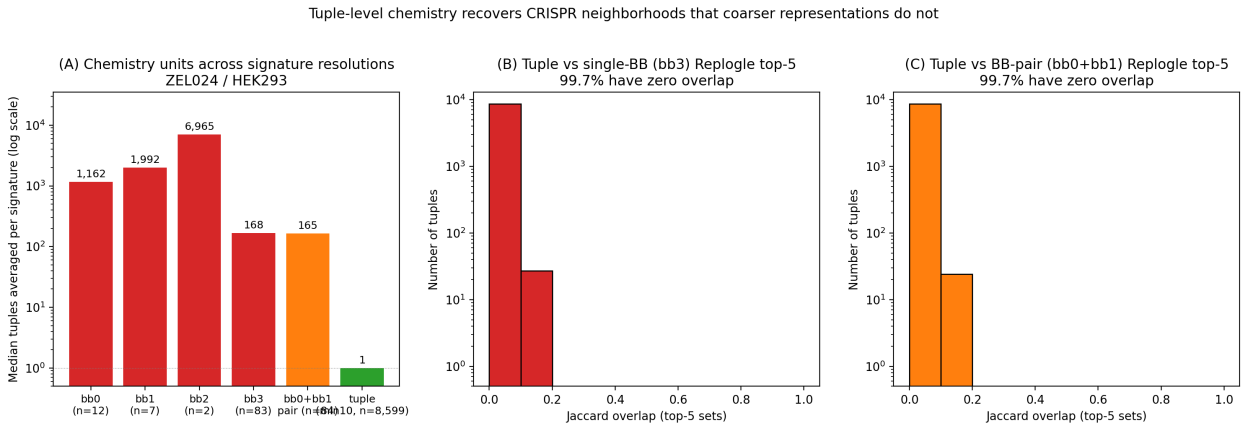


Figure 1. Aggregation diagnostics for ZEL024 / HEK293. (A) Median number of distinct chemical tuples averaged into each diagnostic signature, log-scale. Single-BB signatures average over many distinct compounds sharing one building-block position; tuple signatures average over one full molecule-level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

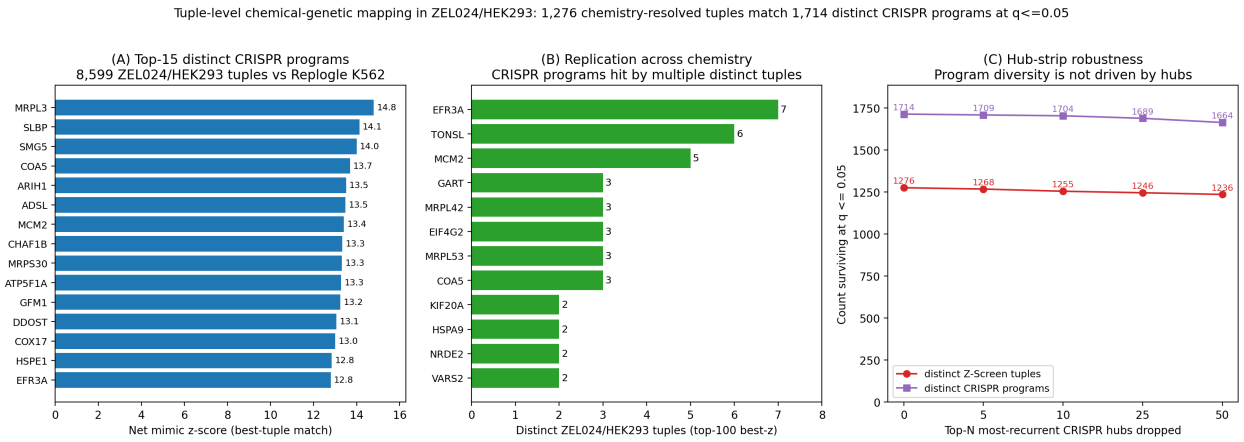


Figure 2. Tuple-level chemistry maps onto 1,714 distinct CRISPR programs in ZEL024 / HEK293. (A) Top-15 distinct CRISPR programs by best Z-Screen tuple z-score, deduplicated to one tuple per CRISPR program. (B) Replication-across-chemistry: among the top-100 best-z tuples, CRISPR programs reached by 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.

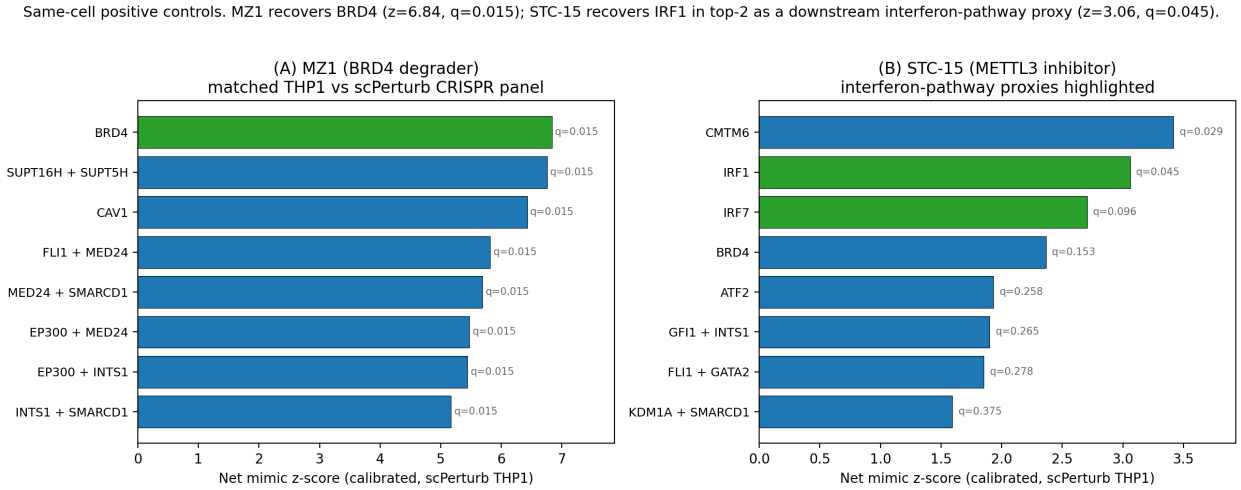


Figure 3. Same-cell positive controls in matched THP1 with calibrated empirical q-values from a permutation-null gene-set calibration. (A) MZ1 (BRD4-selective PROTAC degrader): BRD4 is the top calibrated CRISPR-like signature (z = 6.84, q = 0.015), with subsequent neighbors falling in BRD4-adjacent transcriptional 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.

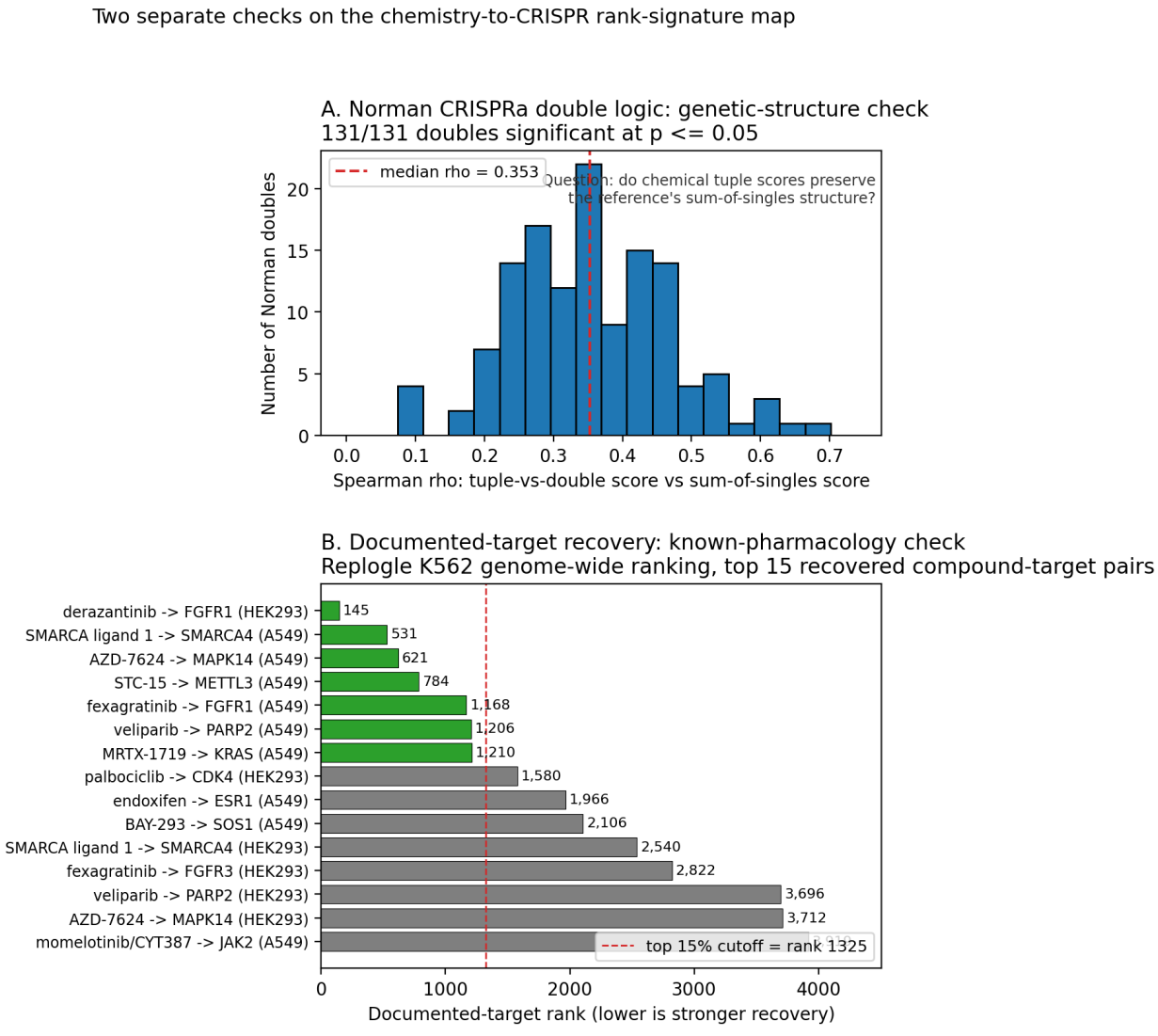


Figure 5. Two separate validation checks. (A) Norman combinatorial logic for pleiotropy-aware

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.

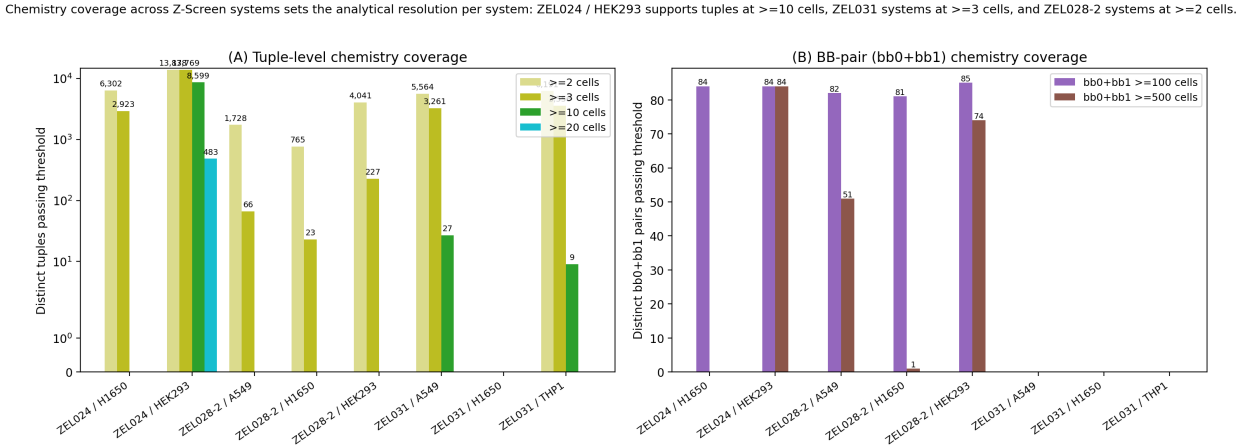
The calibrated min2 full-tuple scans still show structured biology. The deduplicated calibrated table contains 943 HEK293 tuple calls, 760 A549 calls, and 448 H1650 calls, for 2,151 total best full-tuple CRISPR calls used in the hierarchy. The strongest full-tuple matches reached high net z values, and 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.

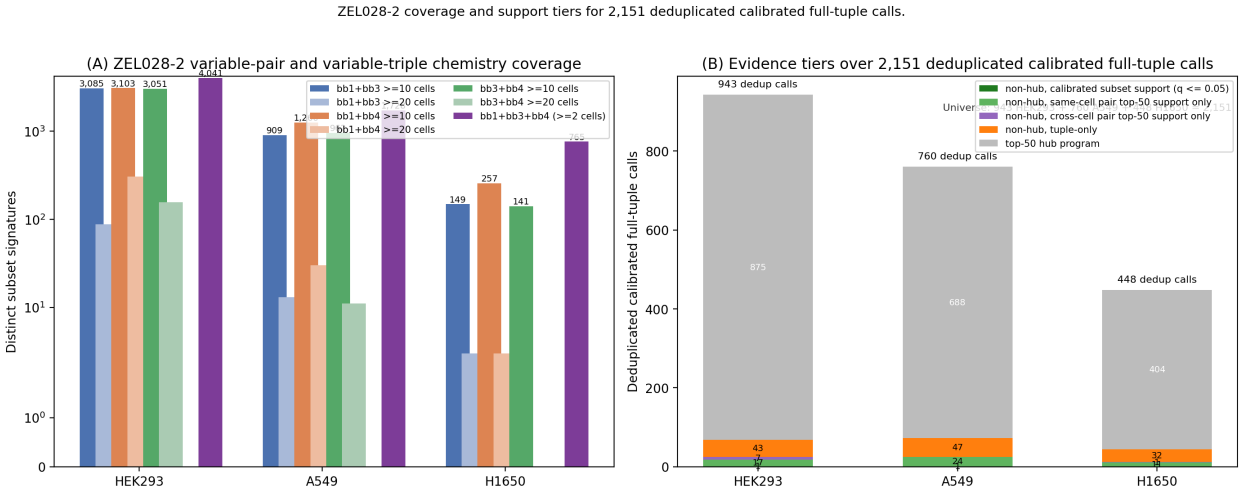
Two non-hub candidates pass the strictest FDR-controlled subset tier: HEK293 ATP6V1A and A549 FKBP9 (`paper5/tables/ZEL028-2_hierarchical_mechanism_support_top_candidates.csv`). The ATP6V1A tuple has full-tuple net z = 15.47 and q = 0.001, with bb3+bb4 subset support at z = 7.94 and q = 0.001. The FKBP9 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.

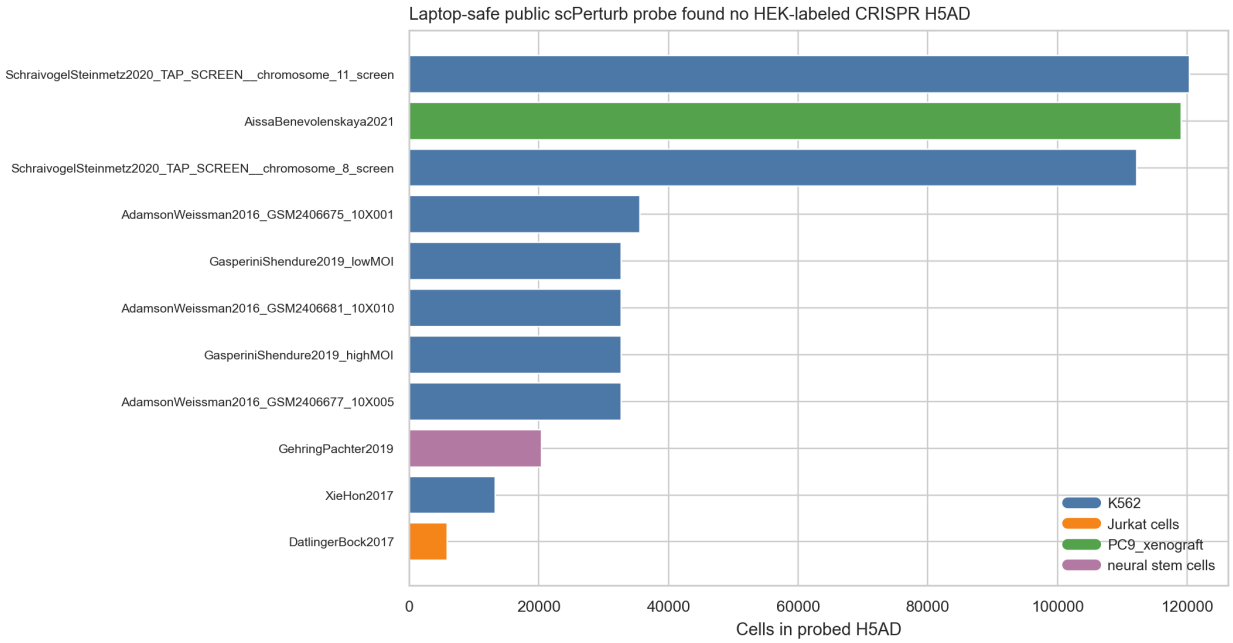


Supplementary Figure 3. ZEL028-2 multi-resolution chemistry coverage and hierarchical mechanism support. (A) Variable-pair (bb1+bb3, bb1+bb4, bb3+bb4) subset signatures qualifying at >=10 wells per signature and at >=20 wells per signature, plus the variable-triple bb1+bb3+bb4

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.