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.

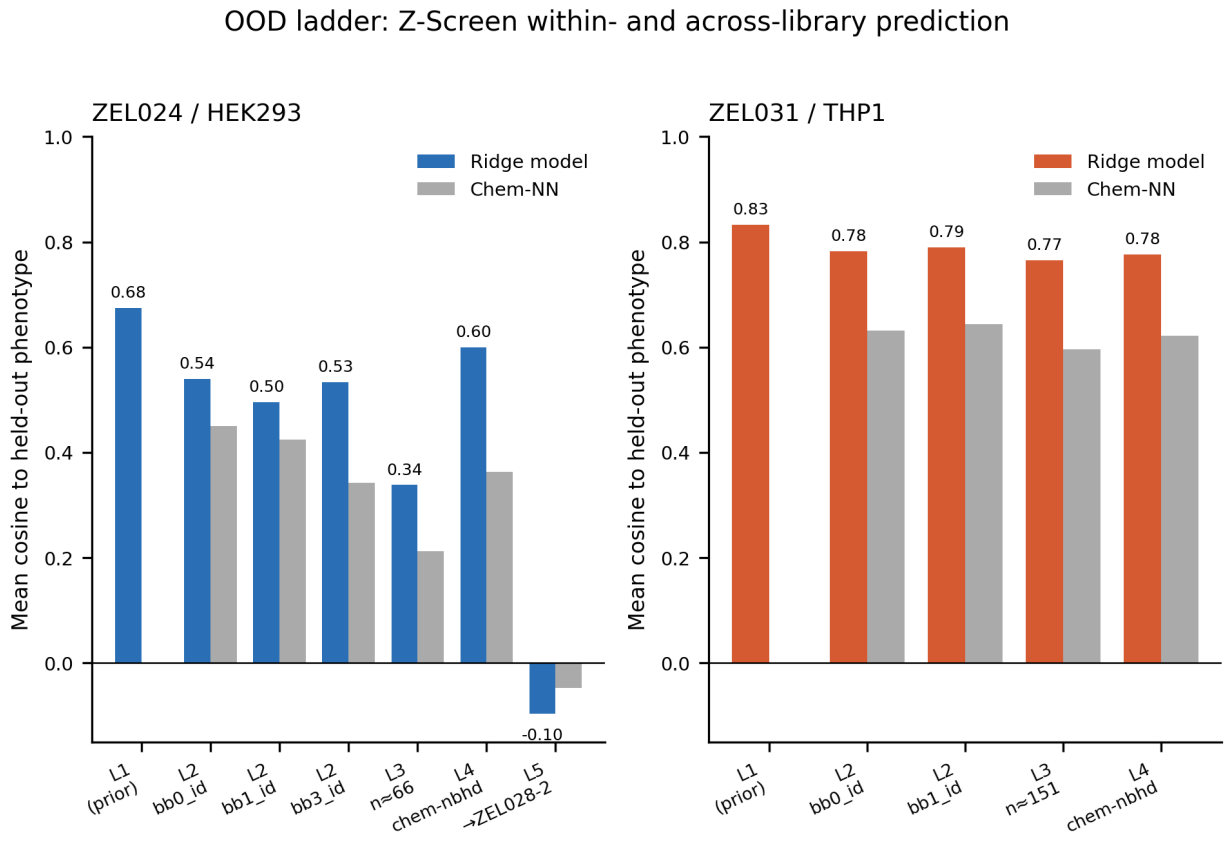

Figure 1. The generalization ladder is an experimental decision tree. L1 asks whether missing tuples can be filled in when all building blocks are familiar. L2 and L3 ask whether held building blocks can be predicted within a library. L4 asks whether gains survive chemical-neighborhood holdout within 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.

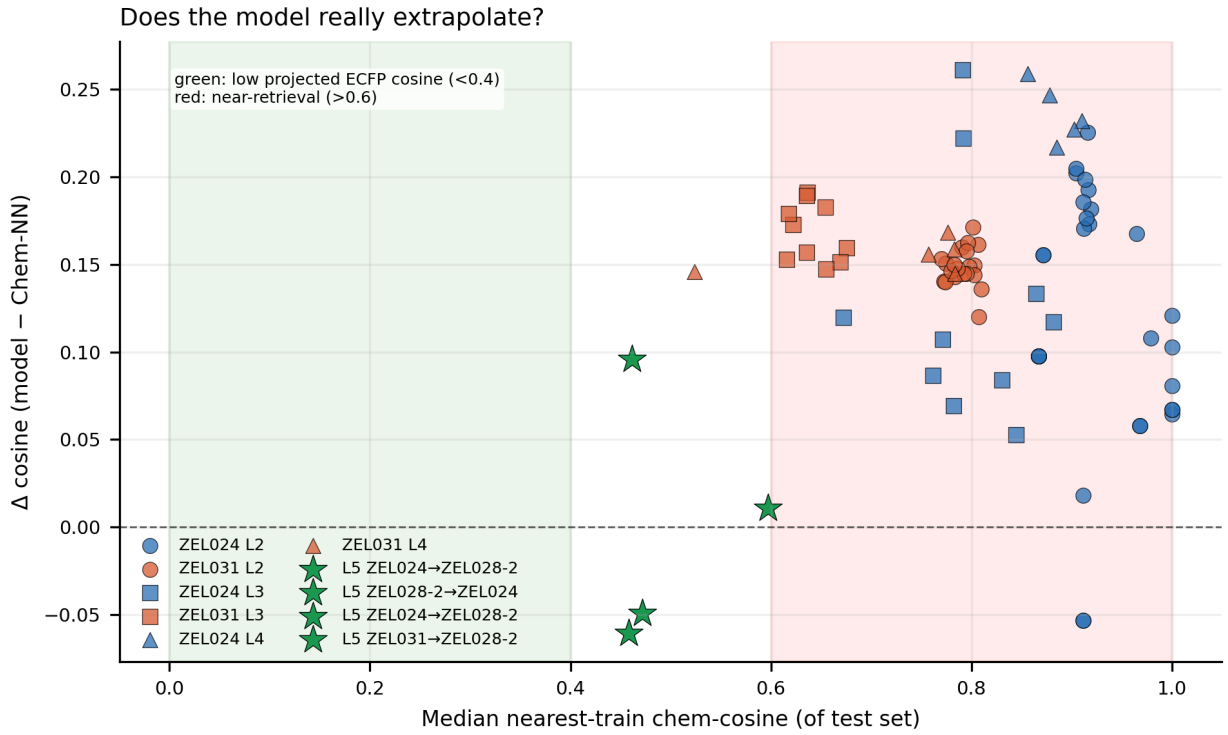

Figure 2. Retrieval-versus-extrapolation analysis plots model gain against median nearest-training similarity measured as cosine in the random-projected ECFP4 embedding, not raw-fingerprint Tanimoto. ZEL031 / THP1 sits in the lower-neighbor-similarity, high-model-gain regime, consistent with held-chemistry 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

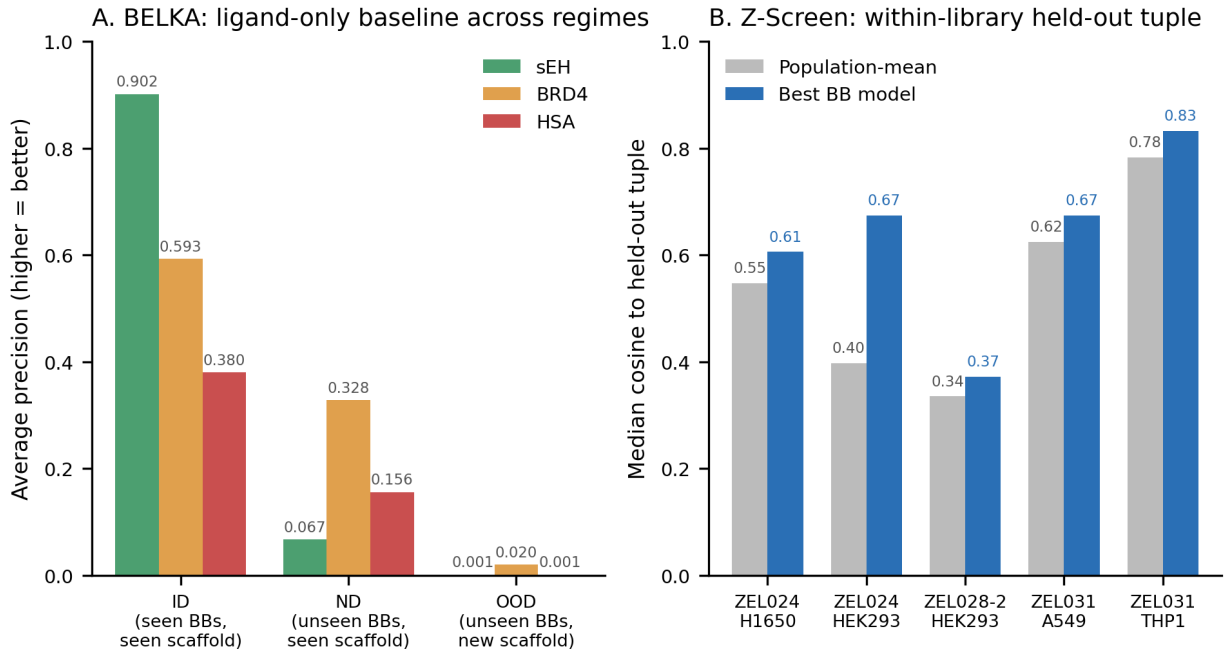

Figure 3. BELKA is used here as a vocabulary for generalization regimes, not as a head-to-head performance comparison. The Z-Screen panels separate matrix completion from harder held-chemistry regimes, with ZEL031 / THP1 carrying the strongest extrapolation signal in the present dataset.

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.

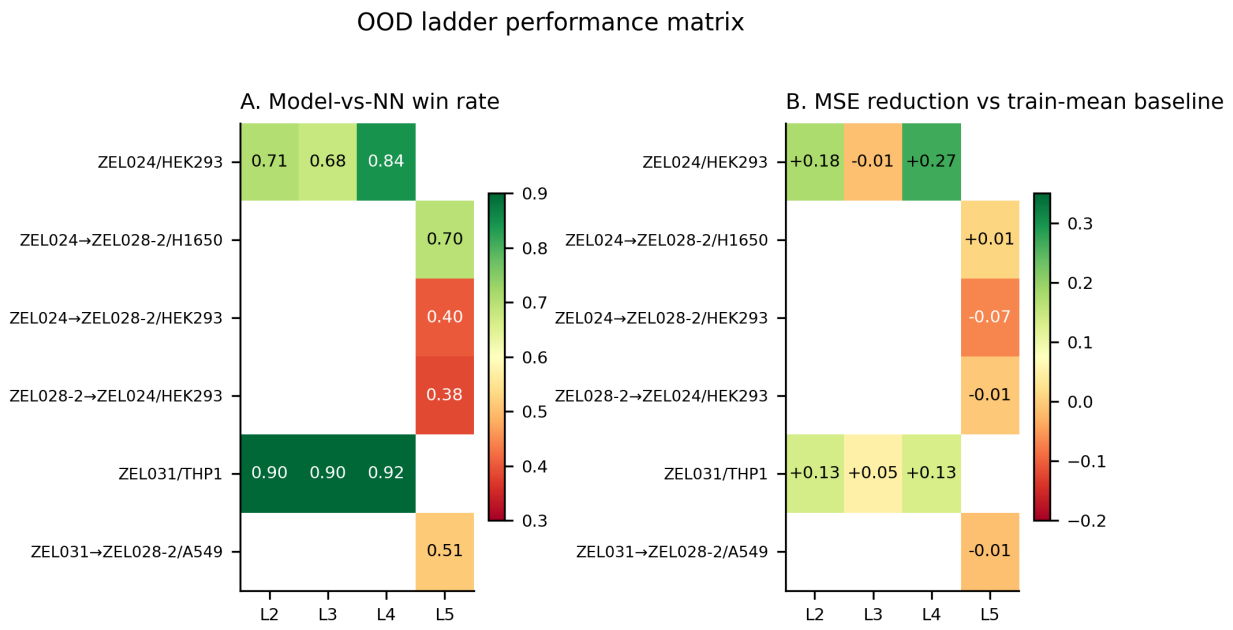

Figure 4. Per-system performance across the ladder shows strong within-library prediction in selected systems (ZEL031 / THP1 for genuine extrapolation, ZEL024 / HEK293 for matrix completion in a saturated grid), and a marked drop at retrospective cross-library scaffold-family transfer that defines the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

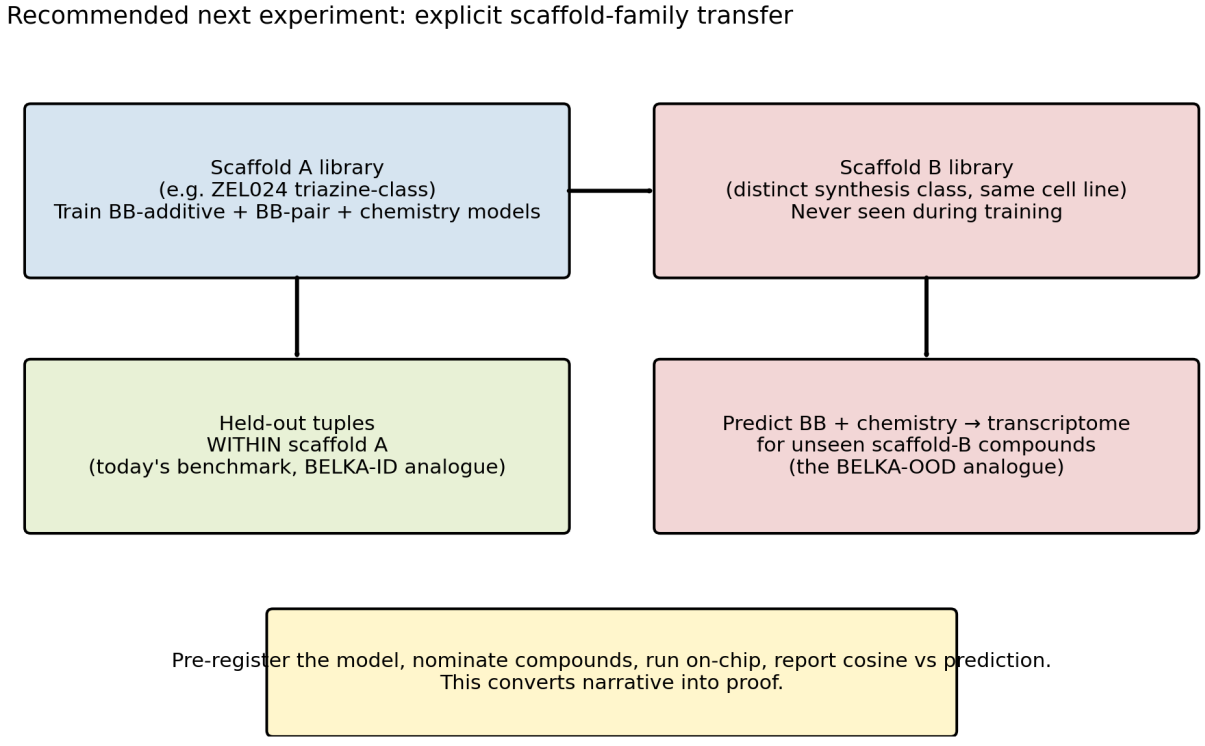

Figure 5. The next decisive experiment is a prospective same-cell scaffold-family hop with a pre-registered model, nominated compounds, and measured transcriptomic outcomes on the held-out family.

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