

Recursive Self-Improvement in Small-Molecule Therapeutics

Functional genomics to specify target biology, modular chemistry to navigate it, and an autonomous experimental loop that refines its own discovery procedures.

Functional genomics can define the high-dimensional cellular response required to correct a disease state. Modular combinatorial chemistry provides a structured, computable search space to reach that target by systematically assembling reusable building blocks. High-throughput cellular profiling continuously reports the transcriptomic and phenotypic consequences of each intervention, guiding subsequent synthesis cycles. This paper establishes the experimental architecture linking these three capabilities, presents retrospective benchmark evidence validating each step, and outlines a path for recursive self-improvement all the way from discovery to patient benefit.

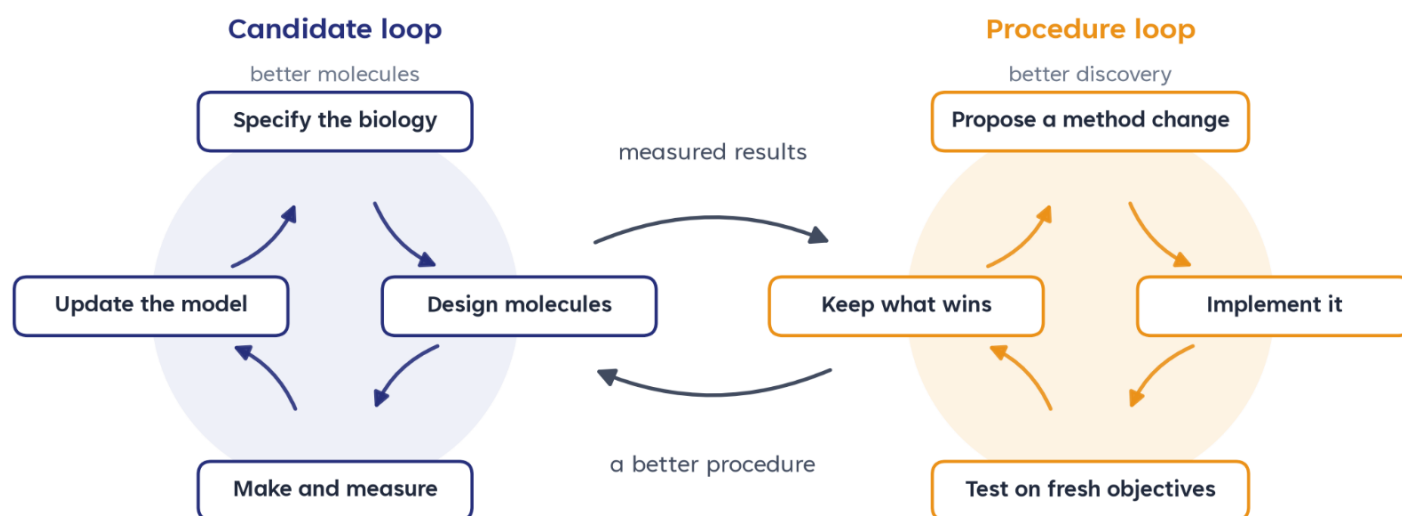


Figure 1. Two interconnected loops. The candidate loop optimizes chemical matter: specifying target biology, designing, synthesizing, measuring cellular responses, and updating molecular models. The procedure loop optimizes the experimental engine itself - refining representations, translation algorithms, and library design rules - and adopts modifications only when they outperform predecessor methods on held-out biological objectives.

162,914

molecules measured, each an assembly of modular building blocks

0.89–0.94

agreement between modeled and measured building-block effects

42.8%

recovery of top-tier chemistry under active learning, versus 14.8% at random

THE IDEA

A discovery architecture that refines its own process

Recursive self-improvement (RSI) occurs when an automated system systematically refines the mechanisms responsible for its own advancement. In software engineering and algorithmic discovery, this transition is already an empirical reality.

Autonomous coding agents iteratively rewriting their own codebase have advanced benchmark performance from 20.0% to 50.0% by retaining only revisions that survive rigorous unit testing.¹ Similarly, an autonomous system optimizing model training procedures discovered a method to multiply 4×4 complex matrices using 48 scalar multiplications - the first verified improvement over Strassen's formulation in 56 years.² Neither breakthrough was achieved through model scale in isolation; both emerged from a closed empirical loop: proposing candidate modifications, evaluating them against an honest, objective test, and committing only verifiable improvements.

Small-molecule therapeutics represents both the most consequential and the most demanding domain for such an architecture. Unlike software engineering, where candidate revisions can be evaluated against deterministic unit tests, biological discovery confronts three fundamental challenges: the evaluation substrate is a living cell rather than a software test; the candidate search space spans an intractable chemical universe; and the overarching objective - discovering an efficacious, safe drug - cannot be expressed as a simple scalar loss function.

The Z-Screen platform addresses these three obstacles through coordinated experimental principles:

- **High-dimensional biological specification.** Functional genetic perturbations (such as CRISPR interference coupled to single-cell transcriptomics) define the multi-gene cellular state a chemical intervention must reproduce. This converts an ambiguous therapeutic goal into an explicit, multi-dimensional response vector that can be continuously scored and optimized.
- **A compositional chemical action space.** Z-Screen libraries are synthesized from modular, recurring chemical building blocks. Because individual building blocks recur across thousands of distinct molecular contexts, predictive models learn transferable substructure-biology representations, enabling both predictable single-block substitutions and generative proposal of synthesis-ready candidates within a verified chemical grammar.
- **High-density, low-cost empirical feedback.** Each compound evaluation returns a whole-transcriptome profile at 50,000 compounds per experiment throughput (compared to industry standard 96-384-well format) making datasets of hundreds of thousands of deep cellular measurements routine. In the public release, 33,575 compounds carry paired high-content imaging features, as well, supported by a dedicated 11,435-well co-assay reading image and transcriptional states from the identical well.

What follows presents the retrospective benchmark evidence supporting each experimental link, clearly distinguishes verified findings from ongoing prospective claims, and defines the criteria required to establish procedural recursive self-improvement.

Genetics provides the target profile for chemical optimization

CRISPR interference suppresses target genes to reveal comprehensive downstream cellular responses.³ Coupled with high-throughput transcriptomic and phenotypic profiling, this approach provides an empirical description of the consequences of modulating a specific biological target. Our framework converts a selected genetic signature into an explicit chemical optimization objective: identifying small molecules that, within a defined cellular context, reproduce the therapeutic signature, preserve essential homeostatic pathways, and satisfy the pharmacological criteria required for an in vivo intervention.

Formulating discovery around high-dimensional transcriptomic profiles provides three decisive advantages. First, the objective is continuous and multi-dimensional rather than binary, allowing partially active molecules to provide quantitative gradients that indicate which pathway components have been engaged and where optimization is needed. Second, by mapping phenotypic states rather than isolated binding pockets, the approach uncovers multiple distinct chemical mechanisms, unlocking biology previously deemed undruggable. Third, target profiles can be benchmarked against known clinical pharmacology before chemical optimization begins.

Empirical calibration precedes screening

In a HEK293 clone evaluated against a HEK293T genetic reference, 13 of 34 benchmark control drugs place their curated target within the top 10% of matching genetic perturbations, achieving a median target rank percentile of 0.144 against a label-set null distribution ($p = 0.001$). While stricter top-5% thresholds did not reach statistical significance across the full panel, specific controls demonstrated high precision. AZD7624, a selective p38 inhibitor, was profiled against 2,620 genetic knockdowns across eight independent experimental devices; its annotated target, MAPK14, ranked within the top 1% across all eight runs, with an overall median rank of 15 out of 2,620. Similarly, the METTL3 inhibitor STC-15 emerged as the highest-scoring molecule in an expanded 35-compound control panel for inducing an ER-stress response program across all five evaluated cell backgrounds.⁴

The reference library extends across diverse biological contexts. The public dataset benchmarks Zafrens chemistry against 43 distinct genetic signature sets compiled from ten published Perturb-seq studies, representing approximately 18,800 target genes across eight million perturbed cells. These references encompass primary human CD4⁺ T cells in three activation states alongside established immortalized lines, establishing that genetically derived response targets can be defined across primary and disease-relevant cellular systems.

Scope and limitations of phenomimetic scoring

Transcriptomic response resemblance constitutes an empirical hypothesis regarding a molecule's cellular mechanism of action. It does not replace direct assays for target engagement, binding affinity, selectivity, or pharmacokinetic stability, nor does it guarantee therapeutic efficacy in animal models. Its purpose is to provide a continuous, high-dimensional objective function capable of ranking every compound in a combinatorial library and calibrating prospective search strategies against validated pharmacology prior to campaign launch.

THE KEY IDEA

Modular chemistry provides a computable, transferable grammar

Conventional screening collections consist of structurally heterogeneous, disconnected chemical matter, where each measurement yields minimal transferable signal across the library. In contrast, combinatorial screening is engineered to maximize structural overlap and parameter sharing.

Every compound is constructed combinatorially: a defined building block positioned at each synthetic attachment point. Across the public pilot dataset, 627 distinct building blocks assemble 162,914 unique molecules. The median building block appears in 461 distinct chemical contexts, and even the least-frequent component of a typical molecule appears in approximately 1,100 peer molecules. Profiling a single compound therefore yields empirical evidence across an entire neighborhood of structurally related designs rather than an isolated data point.

Chemical structures withheld

Compound IDs and measured results are retained.

Figure 2. One building block, many molecular contexts. Representative building blocks and an illustrative sample of their parent molecules across two distinct libraries and four cell backgrounds. Panel labels indicate compound, cellular context, and scaffold distributions; within the public dataset, the two highlighted blocks appear in 873 and 2,110 molecules, respectively. Sixty-nine building blocks are shared between the two libraries, spanning 65% and 78% of their chemical space, with a median of 291 distinct partner blocks.

Cellular responses reflect modular chemical components

If small-molecule phenotypic responses were emergent, indivisible properties of whole structures, combinatorial design would offer little machine-learning utility. In practice, substructures exhibit strong additive predictability. Computationally removing a single building block from a predictive model recapitulates that block's empirical cellular effect with a median Spearman correlation of 0.89 to 0.94 across two cellular contexts and three validation splits, compared to an empirical label-shuffled null near zero ($z = 13.8$ to 19.2). Decomposing building blocks further into constituent fragments and summing their individual contributions predicts the complete block effect at $r = 0.97$ to 0.99 .

These substructural effects correspond directly to coherent biological programs. For striking examples - when projected onto specific gene networks, an individual building block present across 2,294 distinct molecules consistently induces a heat-shock and chaperone response signature (HSP90AA1, UBB, HSPD1, DNAJB1). A second building block positioned at the identical synthetic position across 2,267 molecules reliably activates an unfolded protein response and ER-stress signature (HSPA5, PDIA6, MANF, DDIT3, HERPUD1). Thus, individual chemical building blocks, evaluated across thousands of varied molecular scaffold contexts, cleanly map to specific cellular transcriptional states.

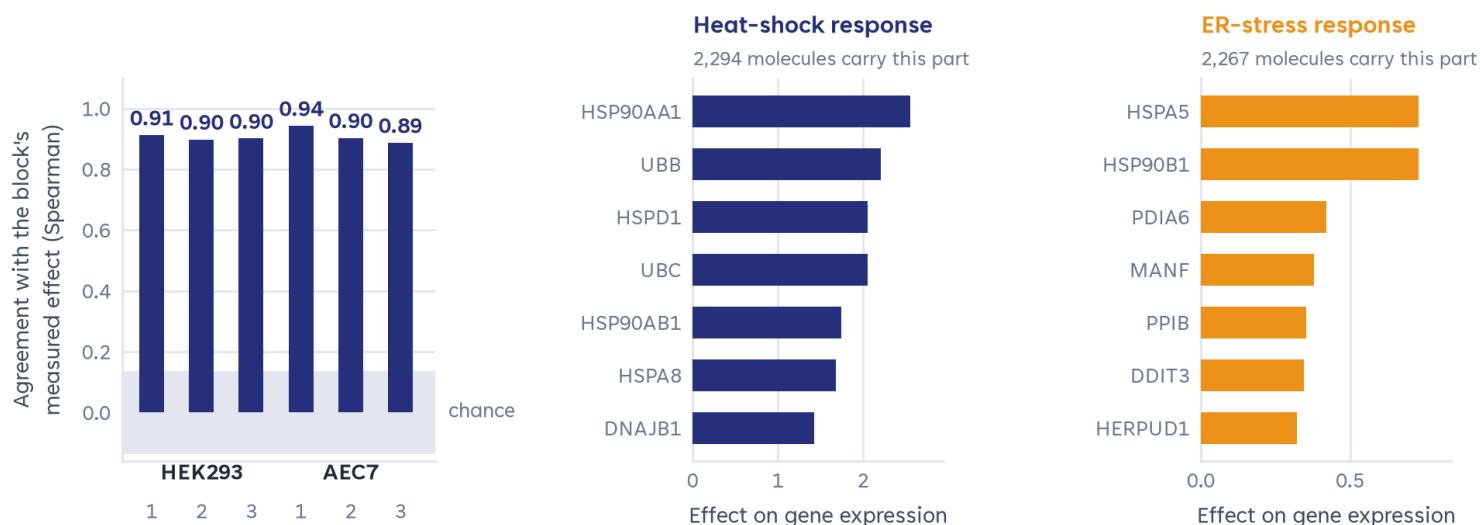


Figure 3. Computational block knockout recapitulates cellular measurements. Left: Agreement between the response decrement when a building block is computationally ablated from model inputs and the empirical response measured for that block in cells. Bars indicate median Spearman correlation across approximately 100 building blocks; the shaded envelope represents the expected distribution under chance permutations. Right: Gene-level transcriptional shifts induced by the two representative blocks described above, estimated strictly from training folds. Values reflect standardized deviations from the baseline context mean, highlighting upregulated program components across differing axis scales.

Unsupervised biological clustering independently recovers underlying chemistry

The most compelling evidence for this modular correspondence requires no initial chemical assumptions. In an unsupervised, chemistry-blind pilot study, 2,394 compounds evaluated in an initial batch were clustered solely based on their transcriptomic response vectors, without access to synthetic recipes. The identified response trajectories subsequently recurred in an independent, disjoint set of confirmation batches (mean confirmation cosine advantage: 0.581, conditional permutation $p = 0.005$). Only after response clusters were established were synthetic recipes mapped back to the clusters. Fifty-seven of 4,308 group-by-feature association tests significantly linked purely phenotypic clusters to specific building blocks or block pairs across 10 of the 12 clusters and 20 distinct blocks, exhibiting enrichments up to 8.3-fold. In one cluster, 110 of 220 members share an identical building block; in another, 50 of 155 members share a specific block pair, compared to a baseline prevalence of 7% across the broader library.

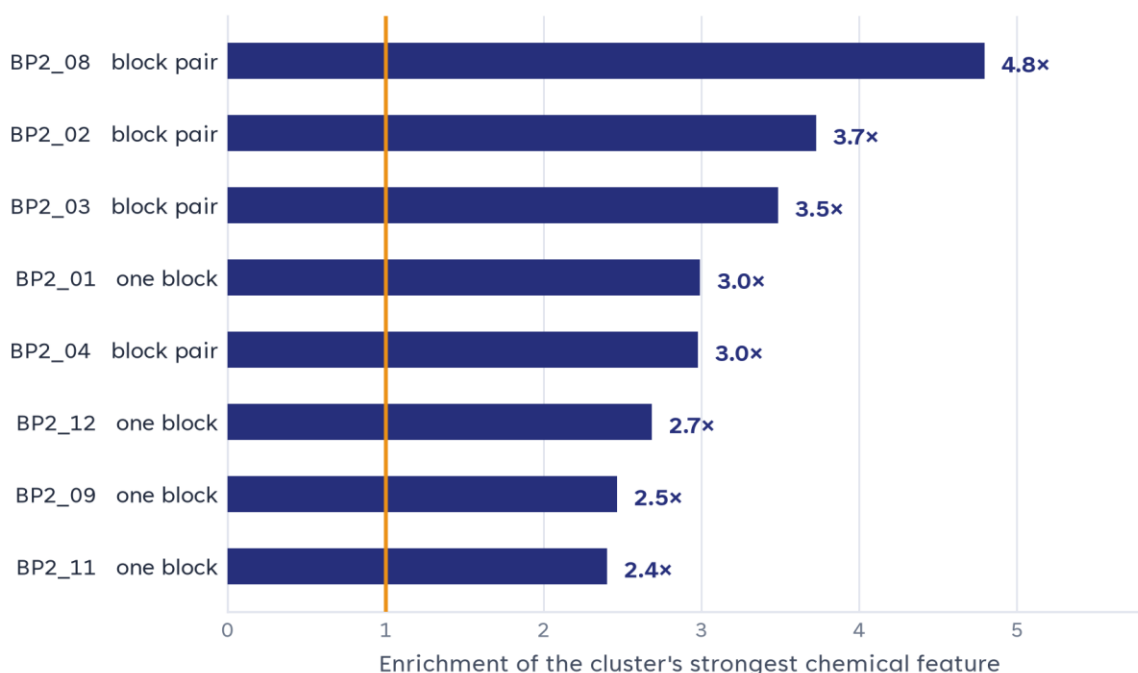


Figure 4. Biological response clusters enrich for distinct chemical building blocks. For each transcriptomically defined cluster, the fold-enrichment of its most strongly associated chemical feature relative to the library background is plotted. Cluster boundaries represent exploratory groupings: while underlying transcriptional response directions replicate reliably across batches, discrete cluster assignments remain sensitive to clustering parameters. Statistical enrichment of a building-block pair does not independently demonstrate synergistic biochemical interaction.

GENERATIVE RECOMBINATION

Predicting unmeasured molecules from recurring building blocks

Decomposing observed molecules into constituent effects is only the first requirement for an active learning platform; the second is accurately predicting the behavior of novel recombinations. The public benchmark tests this generalization capability directly: complete molecular recipes are partitioned by hash, ensuring that test molecules never appear during model training, even though each of their constituent building blocks has been observed in other contexts.

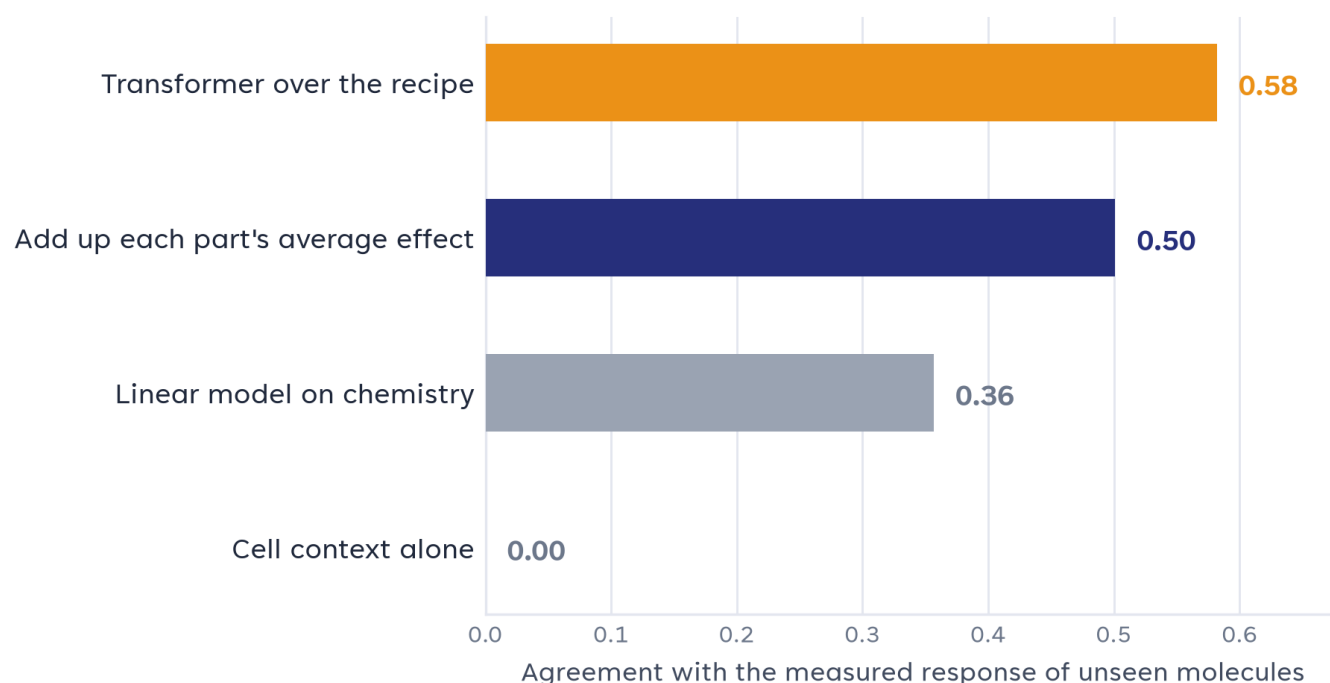


Figure 5. Generalization to unmeasured molecular combinations. Mean per-program correlation on held-out test folds within the deepest-profiled library-cell context (ZEL024 in HEK293). Baseline predictions using cell-context identity alone yield zero predictive accuracy. A transparent additive model summing marginal building-block effects estimated from training compounds achieves $r = 0.501$; a transformer model processing complete multi-position recipes reaches $r = 0.582$ (3-seed ensemble). Across eight evaluated library-cell contexts, the additive model achieves correlations ranging from 0.10 to 0.50, outperforming the transformer architecture in four of the eight settings.

That a simple additive model captures the majority of the predictive signal is a foundational result: it demonstrates that the phenotypic response of an unmade molecule can be largely approximated as the sum of its constituent substructure effects. This modularity directly informs localized structure-activity relationships (SAR): predicting how swapping a single building block redirects a molecule's biological activity.

Because the platform synthesizes molecules combinatorially on shared scaffold positions, the 190,699 profiled compounds naturally yield 5.08 million pairwise single-substitution comparisons—pairs of physically measured molecules within the same cell context that share identical building blocks at every position except one. Across these matched pairs, 207,956 represent rigorous test cases where both compounds were held out from model training. In the deepest-profiled context (HEK293 / ZEL024), the model's predicted response-shift direction aligns with the empirically measured shift at a median cosine similarity of 0.48 across 28,499 held-out pairs, rising to 0.74 for substitutions predicted to produce substantial phenotypic changes (compared to an empirical null of 0.00 from randomized pairs). In shallower, lower-coverage libraries, predictive accuracy is lower, providing an honest empirical baseline for the data density required to resolve single-block SAR.

Two other important results must be highlighted. When a test molecule contains an entirely novel building block unseen during training, predictive accuracy remains substantially above the baseline floor ($r = 0.34$ versus 0.18 for random; $\Delta = +0.155$, 95% CI: +0.124 to +0.188, positive across all 15 validation folds). However, explicit chemical-structure descriptors and learned block embeddings perform equivalently in this regime, indicating that predictions rely primarily on the familiar portions of the remaining recipe. Furthermore,

individual whole-compound phenotypic profiles do not reliably transfer across divergent cell types, even though building-block functional signatures transfer consistently (median cosine similarity of 0.24 to 0.67 across lineages, positive for 88% to 94% of blocks, against a null near zero). The transferable unit of information is the substructure effect, not the compound-level profile.

The design space accessible from known building blocks

Because candidate molecules are generated directly within the combinatorial grammar of the platform, proposed compounds are synthesizable by construction. The building-block vocabularies profiled to date define 2,079,661 nominal compound-by-cellular-context states, of which the 190,699 measured profiles represent approximately 9% of this space. Enumerated purely as distinct chemical recipes, the existing blocks specify 722,397 unique molecules, compared to the 162,914 physically synthesized and screened. In the largest library, introducing a single new building block at one position creates 7,744 novel synthesizable combinations. While this combinatorial lattice represents theoretical enumerations rather than a universal guarantee of synthetic yield, it defines the vast, accessible design space unlocked by a calibrated substructure vocabulary.

THE LOOP

Iterative feedback converts screening into active optimization

A predictive biological model is a valuable asset; an active learning loop that continuously refines that model while directing experimental synthesis compounds in value.

To evaluate this capability, retrospective screening replays were conducted across measured libraries. Each search strategy began with an identical initial pool of 200 measured compounds and subsequently selected six sequential batches of 200 compounds (1,400 compounds total). Three acquisition strategies were evaluated: an adaptive active-learning policy that refits its model after each batch, a static policy that ranks the entire library based on the initial model, and uniform random selection. Compound measurements were revealed only upon acquisition, and a pre-designated reserved set of high-performing compounds remained permanently inaccessible to all policies.

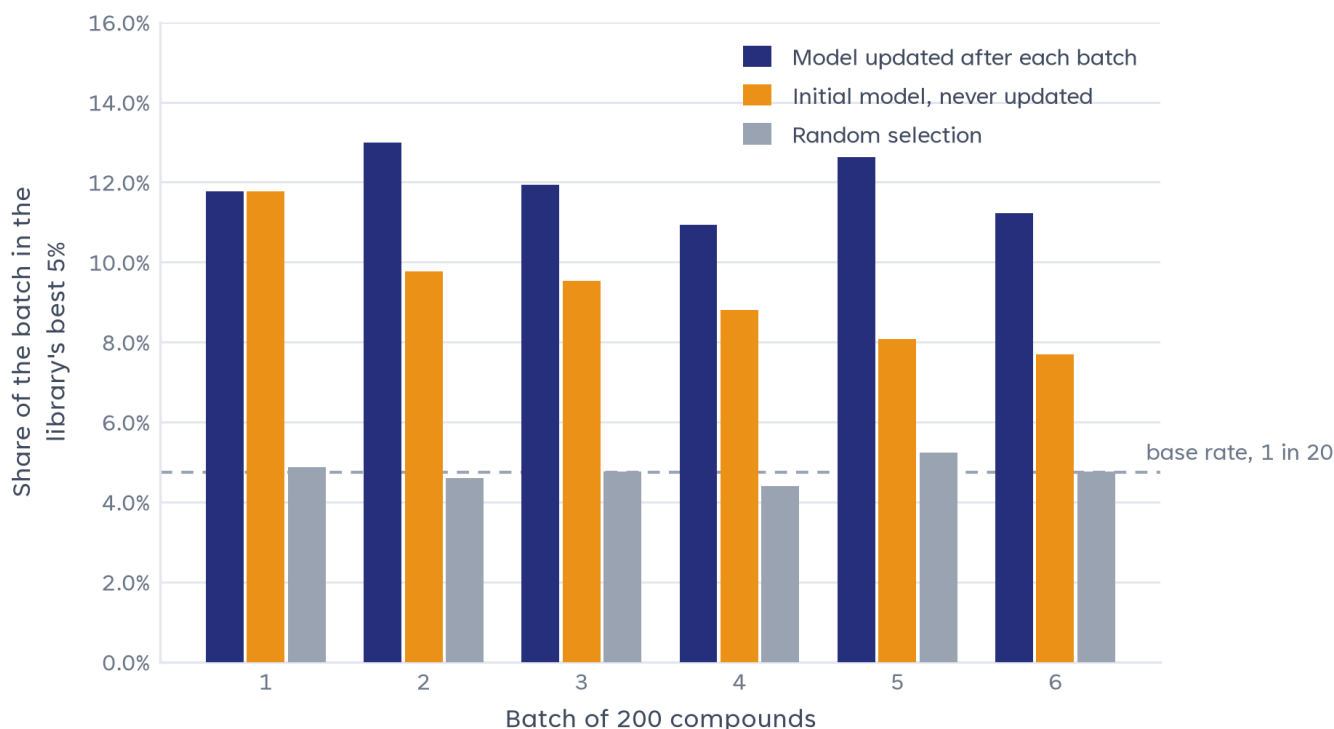


Figure 6. Batch yield trajectories under active learning. Proportion of each 200-compound batch scoring in the library's top 5% response tier, pooled across five responsive target objectives and three initialization seeds. The first informed batch increases discovery yield from approximately 5% (random baseline) to ~12.5%. While the static model's yield subsequently decays as its initial top candidates are exhausted, the adaptive updating policy sustains high hit rates across successive rounds. (Initial informed batches are identical by construction, so performance divergence emerges in batch two.)

At the cumulative 1,400-compound budget, adaptive updating recovers 42.8% of the top-performing chemistry for the strongest objective, compared to 30.7% for the static policy and 14.8% for random acquisition. Adaptive updating outperformed both comparative strategies across all initialization starts for five of the six evaluated objectives. Accounting for the inaccessible reserved test split (which holds 51 to 75 of the top compounds per objective, placing the theoretical ceiling near ~89%), recovering 42.8% represents roughly half of all discoverable high-affinity chemical matter.

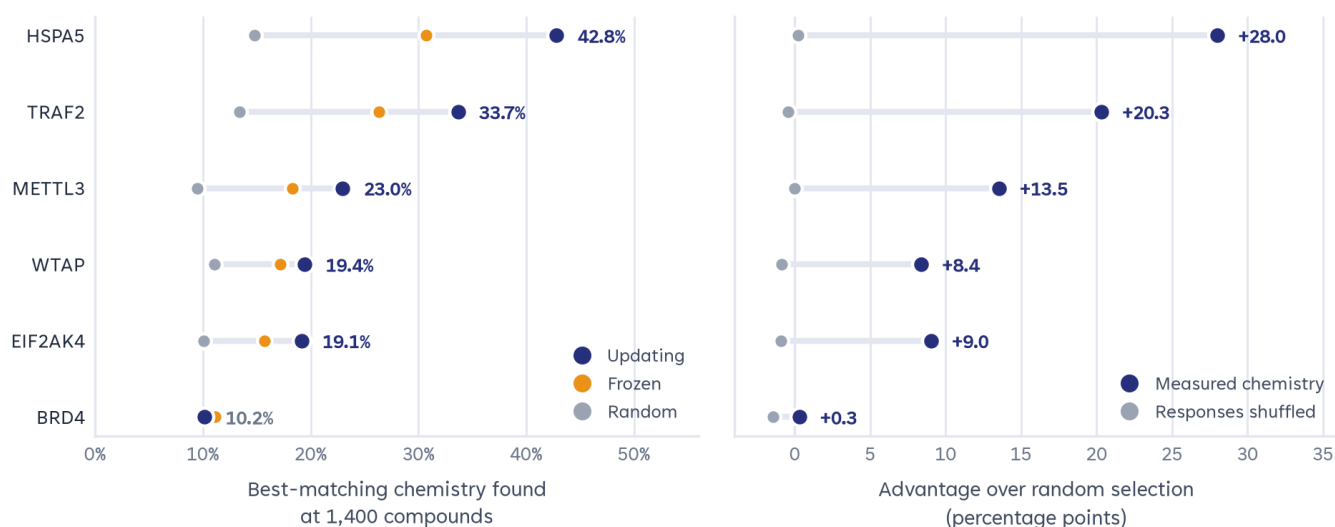


Figure 7. Recovery performance across six response objectives under matched experimental budgets. Left: Cumulative proportion of top-5% compounds recovered at 1,400 total acquisitions. Right: Recovery advantage over random selection using empirical chemistry-response pairings versus label-permuted controls. Performance gains depend strictly on meaningful chemical-phenotypic structure rather than procedural search bias. (Single permutations per context serve as a diagnostic check rather than an asymptotic null distribution.)

The BRD4 objective provides an instructive negative control: it is the sole target where adaptive updating underperformed the static model across all three runs. High-scoring compounds for BRD4 exhibited the lowest chemical coherence among the six objectives: the most enriched building block in the recipe increased top-tier prevalence by only 2.13-fold (and 1.23-fold at the first position), whereas the strongest objective achieved 3.24-fold and 1.89-fold enrichments at the same positions. The active learning framework successfully exploits genuine structure-activity relationships, while transparently signaling when chemical-phenotypic coupling is absent.

The selections made by the active loop are mechanistically interpretable. Guided solely by an unsupervised phenotypic target, the adaptive policy progressively focuses its acquisitions on a compact set of key building blocks: 67% of the compounds selected for the top objective contain one of its three primary building blocks, compared to a 39% prevalence across the broader library. Shuffling response labels completely eliminates this selective convergence.

The trade-off between target optimization and global parameter estimation

Concentrating an experimental budget on identifying top-tier candidate molecules is fundamentally distinct from calibrating an accurate global model of the entire chemical library. In these empirical replays, random compound acquisition yielded superior predictive correlation on reserved out-of-distribution molecules compared to greedy adaptive updating across every evaluated objective. Candidate discovery yield and global model accuracy must therefore be tracked as separate performance metrics. Notably, introducing a naive diversity allocation (reserving 40 of each 200 selections for random exploration) reduced overall hit recovery across five of six objectives without systematically broadening chemical diversity. Optimizing exploration-exploitation trade-offs remains an active engineering priority, and prospective discovery campaigns must deliberately evaluate both endpoints.

BREADTH

High-dimensional readouts support multiplexed discovery campaigns

Because whole-transcriptome profiling captures comprehensive cellular states, an experimental campaign optimized for an initial objective simultaneously generates actionable data across dozens of secondary targets. In pilot benchmarks, profiled libraries resolve into 1,007 chemistry-linked response families across eight experimental settings; 855 families exhibit statistically significant pathway signatures, and in 762 families, every member shares a common building block that provides an immediate synthetic handle for that biological axis. Within a single library-cell context, 102 distinct families point along 37 mutually orthogonal transcriptional program directions.

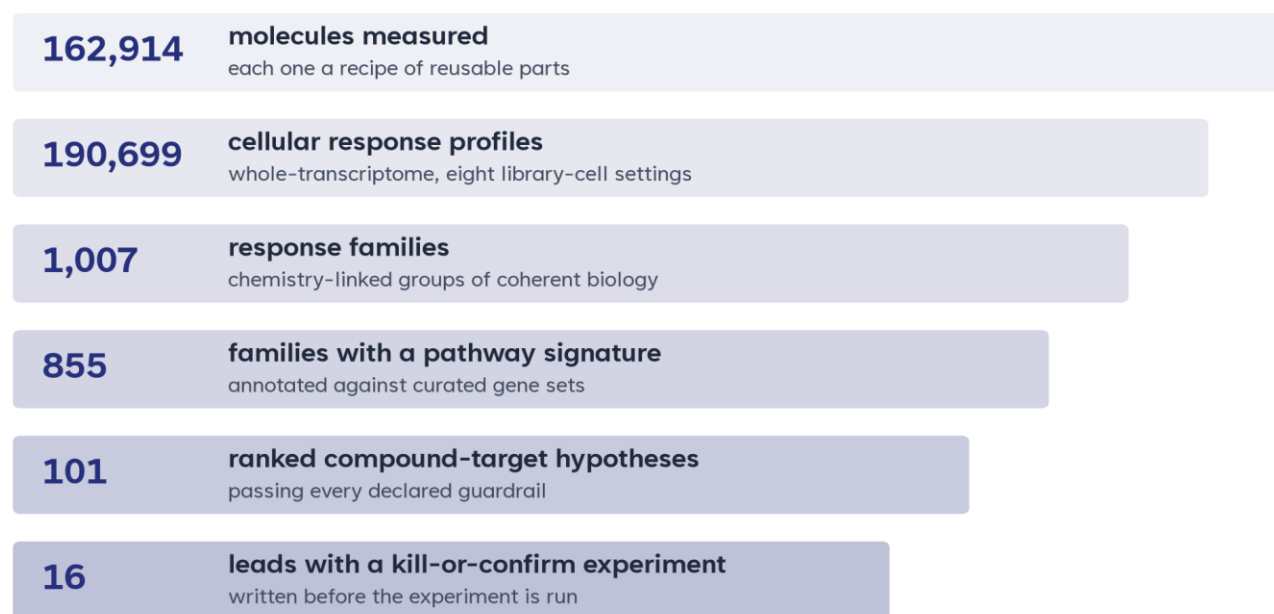


Figure 8. Funneling combinatorial libraries into qualified experimental leads. Progression from broad combinatorial synthesis to verified biological leads carrying structured follow-up hypotheses. Operationally, the final tier establishes falsifiable confirmation assays; designating a compound as a lead does not substitute for downstream target validation. Response families represent tight structural clusters rather than exhaustive partitions, with a median family size of four molecules.

This multiplexed visibility provides the operational basis for program branching: a library synthesized to address objective A inherently yields the empirical data required to qualify objective B, allowing subsequent synthesis cycles to advance both objectives simultaneously while initiating preliminary exploration of objective C. Generating this secondary intelligence incurs zero additional screening cost.

A detailed case study illustrates both the power of this approach and the rigorous controls it requires. Two chemical subseries sharing an identical core scaffold but differing at an adjacent partner position exhibited reciprocal genetic target resemblance: one series achieved cosine similarities of 0.112 against HSPA5 and 0.050 against TRAF2, whereas the partner series scored 0.052 against HSPA5 and 0.082 against TRAF2, with individual members of each series replicating the shift. When the analysis was restricted strictly to chemistry-matched pairs, the TRAF2 association remained robust; however, when further constrained to pairs evaluated

within the identical experimental batch, the distinction attenuated. This underscores that target resemblance is governed by complete combinatorial pairings rather than isolated cores, demonstrating that prospective campaigns must rigorously balance chemical variants across experimental batches by design.

THE RECURSION

Procedural optimization: the outer recursive loop

The candidate loop described above operates within a fixed experimental procedure, learning solely from molecular measurements. True recursive self-improvement in therapeutics introduces an autonomous outer loop that refines the discovery procedure itself. The procedure loop iteratively modifies how cellular states are represented, how genetic and chemical profiles are aligned, how acquisition functions balance risk, and how new combinatorial libraries are designed. Research agents propose these algorithmic and methodological refinements, and the system deploys them into production only when they empirically outperform earlier standards.

To maintain scientific rigor, procedural advancement must be evaluated against strict falsification criteria. A modified discovery procedure earns adoption only by outperforming a frozen predecessor on entirely novel biological objectives that were withheld during its development—tested from identical starting datasets, under identical synthesis and screening budgets, and with all failed syntheses and invalid assays explicitly counted in performance denominators. This rigorous comparison separates genuine recursive self-improvement from standard model retraining and forms the core empirical hypothesis of the next platform phase.

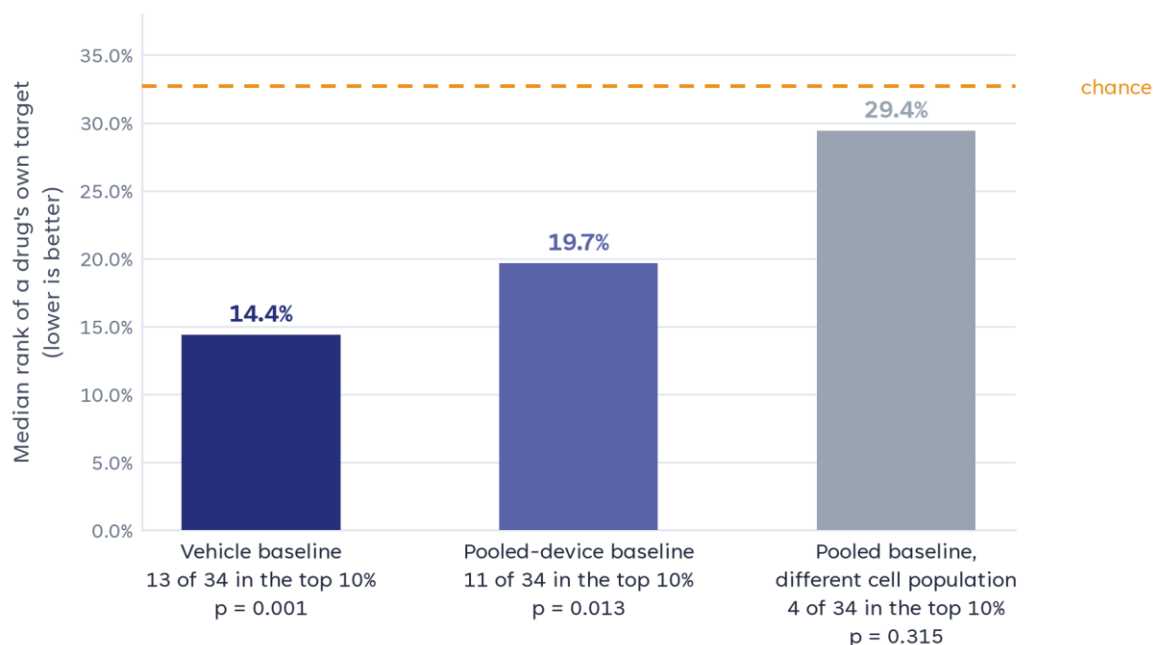


Figure 9. Evaluating discovery procedures as empirical objects. Thirty-four benchmark control drugs evaluated against 18,311 candidate genetic references across three alternative analytical procedures. Shifting the baseline normalization from a pooled reference to a vehicle control improves the median target percentile rank from 19.7% to 14.4% and increases top-10% target identifications from 11 to 13; evaluating an alternative cell population using pooled baselines degrades performance to chance expectations. This retrospective comparison illustrates how procedural modifications measurably alter discovery efficiency, providing the prototype for prospective procedure-level optimization.

This analysis demonstrates the outer loop in miniature. Across identical chemical compounds and genetic references, modifying the analytical procedure altered discovery yield dramatically: shifting target ranking from chance (33%) to 19.7%, and ultimately to 14.4%. Because neither the chemical matter nor the underlying biological measurements varied, the performance gain is attributable solely to procedural refinement, demonstrating that discovery methodology can be treated as a measurable, optimizable variable.

The modularity of combinatorial chemistry is what makes systematic procedural evaluation practical. Holding a core structure fixed while systematically varying partner groups yields built-in controls across thousands of data points. Hypotheses regarding chemical rules can be tested at scale, and substitutions that eliminate or redirect biological activity define the boundaries of the design grammar. Identifying these boundaries ensures that successive library designs grow progressively more efficient rather than simply larger.

SCALE

Information scaling and compounding returns in high-throughput profiling

System capacity enhances discovery performance across three distinct dimensions, each of which must demonstrate clear empirical gains on withheld evaluation data.

First, expanding chemical diversity systematically enhances model accuracy. Iteratively training the additive marginal model on expanding subsets of measured compounds and evaluating performance on identical held-out test sets yields a strictly monotonic scaling curve: correlation increases from 0.036 at 250 compounds to 0.190 at 49,000 compounds, maintaining an improvement rate of ~ 0.013 per doubling at the current frontier and ~ 0.022 per doubling across two orders of magnitude.

Second, increasing cellular profiling depth substantially improves signal-to-noise. Across 1,562 chemical families, concordance between two disjoint subsets of compounds within the same structural family rises from $r = 0.37$ when pooling 30 to 99 single cells to $r = 0.99$ when pooling over 5,000 cells (Spearman correlation of 0.882 across families). While statistical averaging accounts for part of this gain, this scaling illustrates why large combinatorial libraries yield compounding value: every profiled molecule contributes to the aggregate estimate of multiple overlapping structural families simultaneously.

Third, experimental replication refines reference stability. Aggregating independent experimental batches increases correlation with held-out validation batches from 0.420 (single batch) to 0.690 (eight batches) in one context, and from 0.424 to 0.668 across five batches in another, improving across 20 of 20 resampled partitions in both settings.

Finally, the performance advantage of active learning expands over successive cycles. Pooled across the five responsive target objectives, the recovery advantage of the updating policy over the static model widens from 1.2 percentage points following the second batch to 6.0 percentage points after the sixth, and expands to a 15.9 percentage-point advantage over random acquisition. Performance trajectories show no evidence of plateauing within the tested budget.

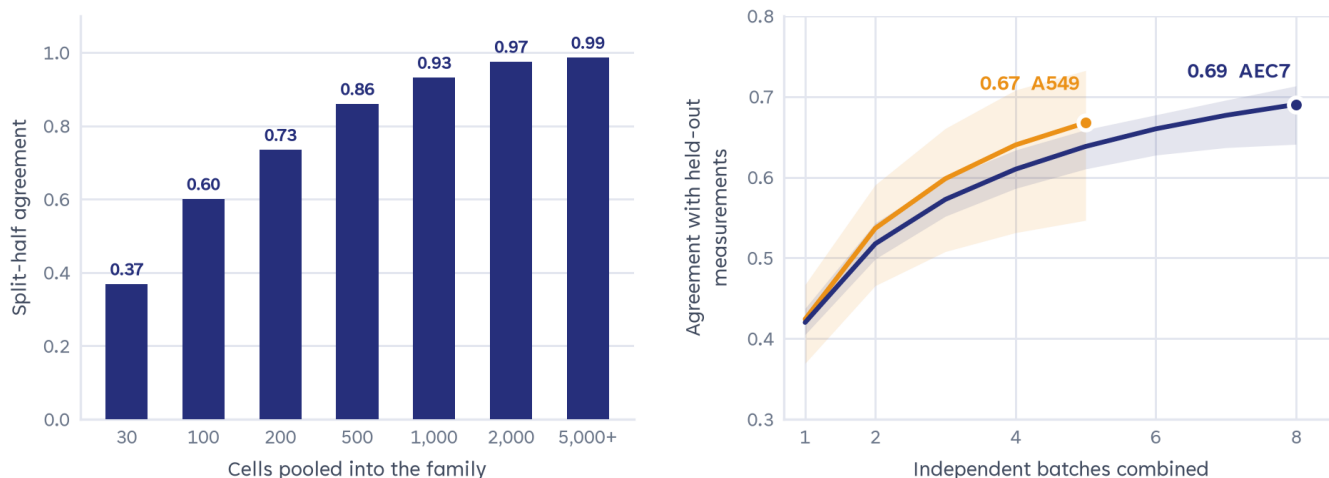


Figure 10. Compounding returns of cellular depth and experimental replication. Left: Biological measurement reproducibility as a function of profiled cell count, evaluated across 1,562 endothelial structural families (larger families incorporate structural diversity, representing empirical library properties rather than an idealized scaling law). Right: Correlation with held-out validation measurements of identical benchmark controls as independent experimental batches are cumulatively pooled.

TRANSFER

Generalizing discovery methodologies to primary and disease-relevant substrates

The discovery framework maintains its mathematical structure across biological substrates: specifying a target phenotype, mapping genetic references, establishing an intended functional modulation, and executing a targeted chemical search. When migrating to a new biological context, the platform transfers substructure representations, validated synthetic chemistry, and optimized procedural heuristics, while fresh measurements rapidly quantify the extent of biological transfer. Context transfer thus becomes an empirical measurement rather than an assumption. As shown in earlier benchmarks, substructure-level program signatures transfer robustly across lineages, whereas individual whole-compound phenotypic profiles do not.

Patient-derived primary samples represent the critical translational substrate, where assay miniaturization provides a decisive operational advantage: minimizing sample consumption per condition enables extensive multiplexed profiling from scarce clinical specimens. The Z-Screen platform has demonstrated this capability by profiling whole-transcriptome responses in primary patient and donor T cells across a kinase-directed library and 31 diverse control compounds. In this example known control compounds are used as references and the objective is to find novel small molecules that carry desired biology from these controls while avoiding undesirable biology relevant to the particular patient. In the example below novel molecules that phenomimic AZD-3463, BAY-293, BGT226, GCN2-IN-7, IRAK-inhibitor 1, n-deshydroxyethyl Dastanib, XL019, aloperine, etc., are found directly in a patient's T cell screen, suggesting patient-specific, iterative tuning of therapeutic biology is feasible. Translating ex vivo cellular responses into predictive models of clinical exposure, tissue physiology, and therapeutic outcome defines the prospective development path.

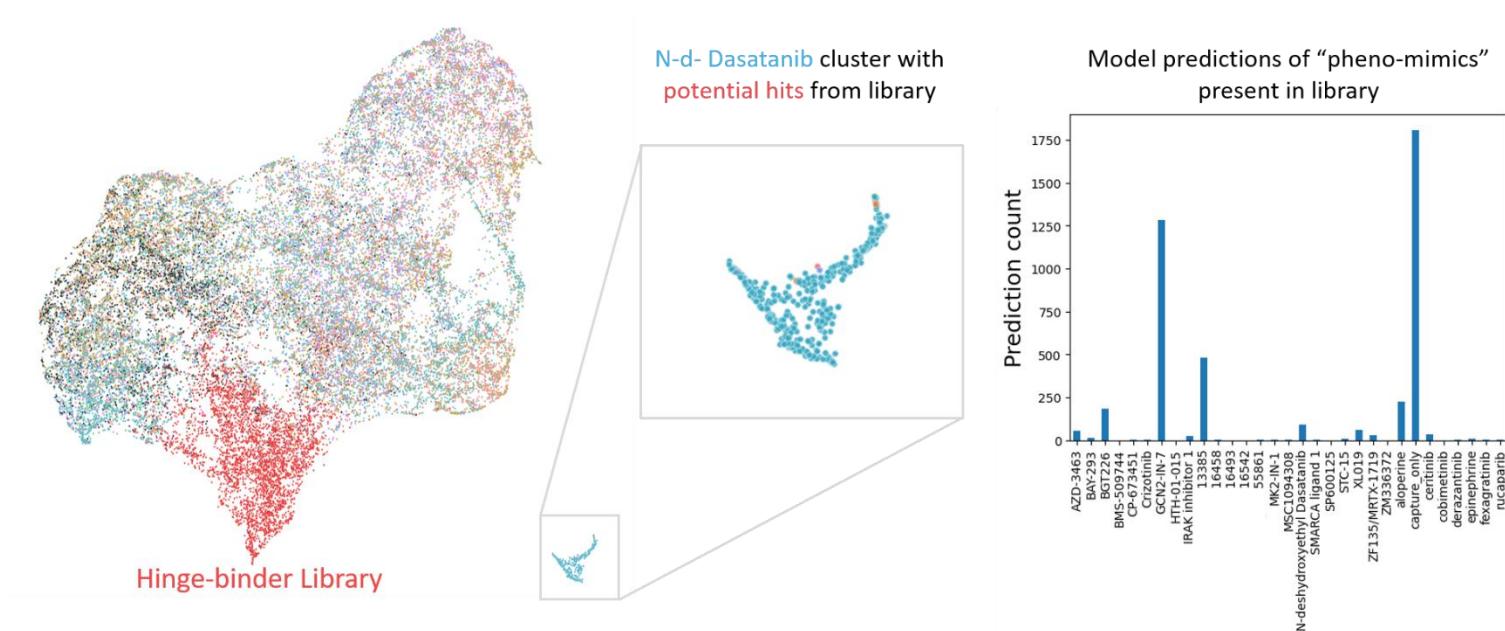


Figure 11. Potential for RSI directly on patient samples. Left: Transcriptome UMAP of 24,000 compounds (colored red) profiled on a cancer patient's CD4+ T cells. The library was generated around diverse hinge-binder scaffolds - the privileged pharmacophores of kinase inhibitors. 31 control compounds were spiked in alongside, represented by various colors in the UMAP. Inset: N-deshydroxyethyl Dasatinib shows a very distinct transcriptional signature in this patient a few molecules from the library induce a similar phenotype. Right: Only a small fraction of the library overlaps with the signatures of the spiked-in compounds. ~1,750 molecules map to no-compound controls, representing the null-set; ~1,250 overlap with GCN2-IN-7, with a number of other internal and external control compounds showing strong mapping to some of the library molecules.

WHAT COMES NEXT

Empirical validation gates for therapeutic recursive self-improvement

- 1. Prospective biological recapitulation.** A frozen predictive model nominates candidate compounds; nominated molecules are physically synthesized and evaluated in fresh cellular assays; and an independent functional or biophysical assay confirms whether the specified biological phenotype was successfully induced. This establishes the physical therapeutic quine: an actionable description of target biology paired with an executable recipe that reliably recreates it.
- 2. Chemical generalization to unmeasured combinations.** Pre-specified novel combinations outperform matched chemical controls in independent experimental batches, including verified validation of substitutions predicted to fail. Falsifying specific negative predictions provides a more rigorous test of chemical grammar than evaluating top-scoring candidates alone.
- 3. Procedural self-improvement.** A revised discovery procedure measurably outperforms its predecessor on biological objectives withheld during its optimization, evaluated under equal computational and experimental budgets, with all synthetic and assay failures fully accounted for in performance metrics.

Each validation gate is designed to be strictly falsifiable, and comprehensive operational records preserve both successes and informative failures.

Platform Capability	Required Empirical Deliverable	Strategic Value Created
Genetically Specified Chemical Reproduction	Novel synthesized molecules reproduce target functional phenotypes in defined biological contexts	Validated therapeutic starting series and an empirical, quantifiable discovery benchmark
Autonomous Procedural Improvement	Successive discovery procedures optimize novel biological targets more efficiently under matched initial conditions	Compounding, program-agnostic discovery capability transferred into subsequent campaigns
Cross-Context Transfer and Branching	Functional activity confirmed across secondary biological objectives or complex primary substrates	Multiplexed pipeline opportunities and an expanding empirical intervention atlas
Patient-Derived Ex Vivo Validation	Chemical response models evaluated in primary clinical material and prospectively correlated with outcomes	A verified, translational bridge connecting screening models directly to human clinical response

Expanding experimental capacity directly accelerates each of these capabilities: providing sharper chemical contrast per cycle, generating higher-fidelity response profiles, and increasing the diversity of biological contexts in which design principles are tested. The empirical evidence presented above identifies precisely where additional throughput yields compounding returns, establishing a rational, evidence-based foundation for platform scaling.

DEFINITIONS

Methodological definitions and benchmark conventions

- **Recovery.** The proportion of a library's pre-identified top-5% responding compounds identified by a given search strategy. Total universe denominators are 464 and 660 compounds and explicitly include pre-reserved compounds that no policy is permitted to acquire. This metric quantifies the retrieval of high-scoring measured phenotypic profiles; it does not denote an assay hit rate, biochemical affinity, or target engagement.
- **Screening replay.** An empirical simulation using previously measured compound datasets, in which phenotypic profiles are hidden from models until selected by an acquisition policy. Three initialization seeds evaluate sensitivity to initial training sets within the same underlying dataset; they do not represent biological replicate experiments. In certain contexts, response baselines were normalized across the full collection, meaning that not every preprocessing step was strictly prospective.
- **Held-out molecular combinations.** Cross-validation splits established by hashing unique compound identifiers. No test recipe is present during model training, while all individual constituent building blocks are observed in alternate combinations. This represents an unseen-combination generalization benchmark using familiar components, rather than an evaluation on entirely novel chemical scaffolds.
- **In silico block knockout.** The computational ablation of a building-block feature from model inputs, rather than physical chemical deconstruction in the laboratory. The evaluation assesses concordance between the model's loss of predicted effect and the empirical response attributable to that building block in cells.

- **Response families.** Coherent, structurally interpretable clusters within the response landscape, identified where pairwise chemical similarity and transcriptomic response similarity both exceed context-specific thresholds. Families overlap and do not form an orthogonal or independent partition of the library.
- **Permuted controls.** Control evaluations where compound-response pairings are randomized once per context, matched on cross-validation splits, budgets, and random seeds. This serves as a diagnostic indicator of genuine association rather than an asymptotic empirical null distribution or formal p-value.
- **Batch agreement.** Concordance between independent experimental measurements of identical control compounds across separate runs. This metric quantifies technical reproducibility and assay precision, rather than model generalization or biological novelty.

Scope and data sources

All quantitative analyses presented in this paper reflect retrospective evaluations of experimental measurements conducted by Zafrens, compiled in ZScreen_Pilot_Package_v2.0.0 and ZScreen_RSI_v0.2.0. No prospective chemical synthesis, de novo functional screening, or active prospective campaigns are reported herein. Platform operational throughput and cost figures reflect company operational specifications rather than package-verified benchmarks.

REFERENCES

Sources and Citations

1. Zhang J, Hu S, Lu C, Lange R, Clune J. Darwin Gödel Machine: open-ended evolution of self-improving agents. arXiv:2505.22954 (2025; revised 2026). Software-agent evidence, not a biological result.
2. Novikov A, Vū N, Eisenberger M, et al. AlphaEvolve: a coding agent for scientific and algorithmic discovery. arXiv:2506.13131 (2025).
3. Replogle JM, Saunders RA, Pogson AN, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell 185, 2559–2575.e28 (2022). doi:10.1016/j.cell.2022.05.013.
4. Calibration, the eight-device AZD7624 result and the three procedural settings. Zafrens. ZScreen_RSI_v0.2.0, evidence/bridge: control calibration, label-set nulls and family statistics. HEK293-CLONE chemical controls against a HEK293T genetic reference; AEC7 chemistry against a teloHAEC reference.
5. Block knockout, parts-sum, recipe counts, the held-out benchmark and the design space. Zafrens. ZScreen_Pilot_Package_v2.0.0, annex_chemistry/attribution_certificate.csv (block knockout and parts-sum certificates), core/recipes.parquet, core/benchmark/fold0_baseline_comparison.csv, core/splits/fold_assignments.parquet, figures/measured_and_nominal_design_space.csv.
6. The chemistry-blind grouping and its enrichments. Zafrens. ZScreen_RSI_v0.2.0, biology_first: chemistry-blind response clustering with disjoint discovery and confirmation batch sets, and the conditional maximum-statistic chemical enrichment test.
7. The replay, the batch yields and the shuffled control. Zafrens. ZScreen_RSI_v0.2.0, replays/extension: six response objectives, four policies, three matched starts, real and permuted rewards; results/campaign_results.csv and the selection tables.
8. The response-family atlas and the funnel. Zafrens. ZScreen_Pilot_Package_v2.0.0, annex_clusters, annex_hypotheses and annex_phenomimicry: response-family census, hypothesis ledger and ranked genetic-correspondence list.
9. Scale, depth and the partner-substitution example. Zafrens. ZScreen_RSI_v0.2.0: measurement_design and atlas/recipe_groups for batch accumulation, paired imaging and RNA studies and family split-half statistics; atlas/matched_partner and case_studies/02_partner_responses for the partner-substitution comparison and its chemistry-matched and stratum-matched sensitivity analyses.
10. DeMeo B, Nesbitt C, Miller SA, et al. Active learning framework leveraging transcriptomics identifies modulators of disease phenotypes. Science 390, eadi8577 (2025). doi:10.1126/science.adi8577. Model-nominated compounds gave a 19.6% hit rate against 1.1% for random selection from the same inventory.
11. Roohani YH, Hua TJ, Tung PY, et al. Virtual Cell Challenge: toward a Turing test for the virtual cell. Cell 188, 3370–3374 (2025). doi:10.1016/j.cell.2025.06.008. The 2025 challenge closed with perturbation-prediction models not consistently beating naive baselines, which is the case for measured perturbation data rather than for more modelling alone.
12. Zafrens. PatientTcellDrugSeq, internal record, November 2025: transcriptional responses in patient and donor T cells to a kinase library and 31 diverse controls, and predicted phenomimics within that library.