Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Context-dependent gains from cell morphology in bioactivity prediction

Yingyi Yi[1]
The Second Affiliated Hospital of Chongqing Medical University
Xiaoqi Deng[1]
The Second Affiliated Hospital of Chongqing Medical University
Zhongxiu Xiao[1]
The Second Affiliated Hospital of Chongqing Medical University
Ya Wen[1]
The Second Affiliated Hospital of Chongqing Medical University
Ming Xia[1]
The Second Affiliated Hospital of Chongqing Medical University
Kuiqiong Liu[1]
The Second Affiliated Hospital of Chongqing Medical University
Ping Shu[1]
The Second Affiliated Hospital of Chongqing Medical University
Peilin Xiao[2]
The Second Affiliated Hospital of Chongqing Medical University
Xiaoli Liu[1]
liuxiaoli@cqmu.edu.cn

The Second Affiliated Hospital of Chongqing Medical University

Research Article

Keywords: Cell Painting, bioactivity prediction, chemical structure, cell count, cellular context, external evaluation

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11221298/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Cellular images can help prioritize compounds for further testing, but useful predictions may reflect information already captured by chemical structure or cell count. We examined whether morphology added predictive value beyond both comparators and whether that value transferred across experimental settings.

Using EU-OPENSCREEN profiles, we developed models in three HepG2 laboratories and evaluated them in a fourth laboratory. Evaluation excluded compounds and chemical scaffolds represented in our preceding analyses. Two of 12 Tox21 endpoints met prespecified sample requirements: antioxidant-response-element activity and mitochondrial-membrane-potential activity. Their test sets contained 85 and 95 compounds, respectively, representing 117 distinct compounds overall. We then compared predictions for the same compounds using HepG2 and U2OS profiles.

Adding morphology to structure and segmented-object count improved the logistic-regression model's mean Brier score by 0.0335 (95% CI: 0.0082–0.0628). The antioxidant-response estimate remained uncertain; mitochondrial-activity gains were positive in both model families. The random-forest mean gain was smaller and its interval included zero. Morphology alone and an adapted published fusion method ranked mitochondrial activity better than the selected fusion. In the paired cellular-context comparison, gains decreased by 0.0534 and 0.0357, respectively, and became slightly negative in U2OS. A laboratory-consistency selection rule made the same decisions as a simpler interval rule in the real tasks and all 2,000 simulations.

This retrospective comparison supports additional predictive value in the evaluated HepG2 setting, not a generally superior fusion method. Its empirical emphasis followed a null primary comparison between selection rules. Prospective target-context evaluation is needed before using these gains to guide preclinical screening.

Introduction

Choosing which compounds warrant further experiments is a recurring problem in drug discovery and safety assessment. Phenotypic assays address this problem by measuring cellular responses without restricting the readout to one molecular target [1, 2]. Cell Painting uses fluorescent stains and image analysis to describe several cellular compartments simultaneously [3, 4]. These measurements extend the early use of automated microscopy for multidimensional drug profiling [5]. Unlike a chemical fingerprint, which describes molecular structure, a morphology profile records what happened after cells encountered a compound. Its potential value lies in capturing responses that structure alone predicts poorly [6, 7].

There is substantial evidence that this additional information can be useful. Existing image assays have been repurposed to predict bioactivity, and multitask comparisons have identified complementary contributions from chemical and morphological representations [8, 9]. Moshkov and colleagues

combined structure, morphology and gene expression to broaden the range of predictable assay activities [10]. Fredin Haslum and colleagues subsequently demonstrated prospective screening enrichment using morphology-based prioritization [11]. Matched morphology and expression measurements also capture overlapping but distinct cellular responses [12]. Together, these studies establish the usefulness of phenotypic information; they leave open which additional measurements are worth acquiring for a particular prediction task.

A strong comparator can change that judgment. Seal and colleagues showed that cell count alone predicts many commonly used bioactivity benchmarks [13]. Rich morphology still outperformed count in selected protein-target assays, so their findings do not imply that images are generally redundant. Rather, they distinguish predicting activity from adding information beyond a simple cellular measurement. This distinction matters because broad cytotoxicity and proliferation responses can influence both reporter assays and perturbational profiles [14, 15]. An improvement over structure alone may therefore disappear once the comparator also includes cell count. Conversely, improvement beyond both inputs would provide more specific evidence for an additional predictive contribution.

Transfer introduces a second requirement. Chemical benchmarks are sensitive to how related compounds are divided between training and testing [16, 17]. Cellular measurements add variation from biological background and experimental acquisition. Transcriptomic studies have documented both shared and cell-type-specific drug responses, as well as limited reproducibility of some connectivity estimates [18, 19]. Wolff and colleagues' EU-OPENSCREEN resource provides matched compounds measured across laboratories and in HepG2 and U2OS cells [20]. Their study characterized variation in profiles and cell-count responses. These data also permit a separate question: does morphology's improvement over structure plus count survive a change in the source of the cellular measurement?

We addressed that question through an empirical comparison using two Tox21 stress-response endpoints. Models were developed across three HepG2 laboratories and evaluated in an excluded fourth laboratory, with an additional paired HepG2-U2OS comparison. The comparator included chemical structure and the measured segmented-object count. Earlier cardiomyocyte, JUMP, EveBio, Cell Health and OASIS analyses informed the comparisons and are reported as supporting analyses, not independent replications of one effect. We also evaluated whether adding laboratory-consistency requirements improved a simpler rule for selecting fusion. This design separates three questions relevant to experimental planning: whether morphology adds value, where that value persists, and whether development data can identify failures before transfer.

Methods

Study design

We reanalyzed public compound-profiling and assay data; no patient records were used. The principal evaluation linked EU-OPENSCREEN morphology to existing binary Tox21 activity labels [20]. Models were

developed with FMP-, IMTM- and MEDINA-HepG2 profiles. USC-HepG2 was excluded from model fitting and selection. FMP-U2OS provided a secondary cellular-context comparison using the same held-out compounds (Fig. 1b). Outcomes were antioxidant response element (SR-ARE) and mitochondrial membrane potential (SR-MMP) assay activity, not clinical toxicity events. Eligibility required at least 20 active compounds, 20 inactive compounds and ten chemical groups in both development and strict evaluation. All 12 candidate endpoints are listed in Supplementary Table S1.

The development pool comprised compounds represented in our preceding analyses. Strict evaluation excluded their standardized parent identities and Bemis-Murcko scaffolds [21]. New compounds with previously represented scaffolds formed a separate secondary stratum. Model specifications and selection rules were set before initial external scoring. One subsequent identity correction moved cefoxitin out of the strict set because a tautomer-dependent representation had obscured a shared scaffold. We reran the unchanged analysis; the corrected results are not a new blinded validation. The original primary comparison was between two model-selection rules. The present emphasis on prediction gains and their context dependence followed the observed results; it did not involve selecting new models from external performance.

a, Allocation by prior chemical history. Of 576 labeled parent compounds, 573 had profiles at all four HepG2 laboratories; endpoint qualification additionally required a nonmissing activity label. SR-ARE and SR-MMP strict tests overlap and are not 180 independent compounds. b, Models are developed in three HepG2 laboratories and evaluated in an excluded laboratory. The paired comparison changes the cellular input for the same strict held-out compounds; it does not isolate cell lineage from acquisition conditions. Arrows denote analysis comparisons, not chronological sample collection. Counts have no uncertainty intervals.

Data and comparators

We used version-3 aggregated EU-OPENSCREEN profiles, retaining observed wells and their plate controls. Contradictory Tox21 labels within a standardized parent structure were excluded. Chemical structure was encoded using 2,048-bit, radius-2 Morgan fingerprints. Segmented-object count was transformed as log2(count + 1) relative to plate-matched dimethyl sulfoxide controls. Morphology features excluded direct counts, object identifiers and image-position fields. Features were centered and scaled using control wells from the same plate, then aggregated by the median across replicate wells. This control-based normalization did not use test-compound labels or their feature distribution. Feature filtering, imputation and scaling within models used training data only, following established image-profiling principles [22].

The baseline combined chemical-structure fingerprints and segmented-object count (structure plus count, labeled ‘Structure + count’ in figures). Fusion additionally incorporated morphology into the baseline through the selected fusion approach. We also evaluated each input alone, structure plus morphology, and an adapted similarity comparator based on a published fusion method[23]. The latter

combined predictions with similarity to active training compounds. We adapted it to the same scaffold and laboratory separation, retaining natural prevalence and omitting probability stretching. These changes are detailed in Supplementary Methods; this was not a reproduction of the original benchmark. Supporting analyses used L1000 and JUMP resources [24, 25], the EveBio task collection analyzed in the cell-count benchmark [13], and Cell Health and OASIS data [26, 27]. Their distinct units, endpoints and validation settings are documented in Supplementary Methods.

Model development

Logistic regression was the primary model and random forest a sensitivity analysis. Development validation combined three laboratory-held-out partitions with three scaffold-disjoint folds. Within each outer training set, three-fold scaffold validation selected hyperparameters and the fusion form. Nested selection prevented the reuse of evaluation outcomes for model tuning [28]. Each compound's rows across laboratories had total training weight one. Logistic-regression regularization was selected from three values; random forests used 200 trees and three minimum-leaf sizes. Fusion was selected between feature concatenation and a weighted average of baseline and morphology predictions. The morphology weight ranged from 0 to 1 in increments of 0.25. All candidates were selected by development Brier score, with no external-label recalibration.

The simple selection rule accepted fusion when the one-sided 95% lower bound of its development gain exceeded zero. The context rule additionally required each development laboratory's mean gain to be at least − 0.005. Margins of 0 and − 0.01 were sensitivity analyses. Otherwise, the rule selected the baseline. Simulations compared these rules under uninformative, redundant, informative, partially mismatched and shifted morphology. Twenty parameter settings each used 100 independently generated replicates. Visible shifts reversed the morphology-outcome relationship in the target setting and one development laboratory; hidden shifts introduced this reversal only in the target setting. These simulations tested rule behavior rather than reproducing the full experimental pipeline.

Evaluation

The Brier score measures the mean squared difference between a predicted activity probability and the observed binary label. We report gain as baseline Brier score minus fusion Brier score; positive values favor fusion. This is overall probability prediction error, not a measure of calibration alone. The original primary estimand was the equal-endpoint mean Brier difference between the context and simple selection rules in USC-HepG2. Supporting comparisons evaluated fusion against the baseline, average precision (AP), area under the receiver operating characteristic curve, and hits among the top 5%, 10% and 20% of ranked compounds. AP emphasizes retrieval of active compounds when they are relatively uncommon [29].

For uncertainty, we resampled chemical scaffolds while retaining paired predictions and overlapping endpoints. Brier comparisons used 5,000 resamples; AP comparisons used 2,000 attempted resamples, all valid in the reported comparisons. Percentile 95% intervals describe fixed fitted models, not the variability of repeating model development. Intervals were not multiplicity-adjusted; secondary contrasts remain exploratory. SR-ARE had 85 singleton scaffolds, making its scaffold resampling equivalent to compound resampling. Paired cellular-context analyses compared the change in gain for identical compounds and labels. Diagnostics examined repeatability, single-well inputs, random-forest seeds and morphology residualized against the baseline. Residual-model parameters were set after primary scoring. Software versions, parameter grids and supporting-study methods are provided in Supplementary Methods.

Results

Incremental prediction

The profile archive contributed 53,476 observed wells across five laboratory-cell combinations. Among 576 labeled parent structures, 353 entered development, 145 formed the strict new-entity/new-scaffold pool, and 78 entered the new-entity/seen-scaffold stratum (Fig. 1a, Supplementary Table S2). Only SR-ARE and SR-MMP met the prespecified endpoint requirements. Their strict test sets contained 85 and 95 compounds, including 23 and 21 actives, respectively (Table 1). Because compounds could contribute to both endpoints, these sets comprised 117 distinct compounds and 114 scaffolds, not 180 independent compounds.

Table 1
Endpoint populations

Endpoint	Development n / active	Development scaffolds	Strict test n / active	Strict test scaffolds
SR-ARE	263 / 53	190	85 / 23	85
SR-MMP	245 / 54	180	95 / 21	92

Unit: standardized parent compound. Endpoints overlap; strict union is 117 compounds and 114 scaffolds. Development and strict sets are not independent across endpoints.

In USC-HepG2, logistic-regression fusion improved the mean Brier score across the two endpoints by 0.0335 (95% CI: 0.0082–0.0628; Fig. 2b). SR-ARE Brier score decreased from 0.2036 to 0.1674, and SR-MMP from 0.1654 to 0.1346 (Fig. 2a). The individual SR-ARE gain remained uncertain: 0.0363 (− 0.0043 to 0.0787). The SR-MMP gain was 0.0307 (0.0120–0.0515). Random forest also showed a positive SR-MMP gain, 0.0096 (0.0014–0.0188), but its SR-ARE interval included zero. Its two-endpoint mean gain was smaller, 0.0099 (− 0.0020–0.0232). Differences were calculated before rounding.

a, Observed Brier scores from logistic-regression models using structure plus count, morphology alone or selected fusion in USC-HepG2, with circles for SR-ARE and squares for SR-MMP. Connecting lines identify an endpoint across categorical inputs, not a continuous trajectory. b, Brier gains for logistic regression and random forest. Points and bars show estimates and percentile 95% intervals from 5,000 paired scaffold bootstrap resamples; positive favors fusion. c,d, Precision-recall curves from logistic-regression predictions. Legend values are average precision; dotted lines mark the active fraction. Curves are descriptive, without confidence bands. SR-ARE has 85 compounds/85 singleton scaffolds; SR-MMP has 95 compounds/92 scaffolds; their union has 117 compounds/114 scaffolds. Joint resampling preserves endpoint overlap and does not refit models. Intervals are not multiplicity-adjusted. The adapted similarity comparator is not an exact reproduction of the original benchmark.

Ranking and screening yield did not improve uniformly. SR-ARE showed an AP increase from 0.493 to 0.575, yet the confidence interval of the difference included zero (Fig. 2c; Supplementary Table S4). For SR-MMP, AP increased from 0.421 to 0.587, with a positive difference interval (Fig. 2d; Supplementary Table S4). Morphology alone had a lower observed SR-MMP Brier score (0.128) and higher AP (0.653) than selected fusion (0.135 and 0.587; Fig. 2a,d). This ordering was reversed for SR-ARE in the same comparison (Fig. 2a,c). The adapted similarity comparator achieved an SR-MMP AP of 0.740 (Fig. 2d). Fusion retrieved more actives than the structure-plus-count baseline at the top-10% budget for both tasks, but one fewer SR-ARE active at the top-20% budget (Table 2). Improvement over the baseline therefore did not establish superiority over all input combinations or screening budgets.

Table 2
Screening yield in USC-HepG2

Endpoint	Model	Test n (active)	Top 5% hits / tested	Top 10% hits / tested	Top 20% hits / tested
SR-ARE	Baseline	85 (23)	4 / 5	6 / 9	10 / 17
SR-ARE	Fusion	85 (23)	4 / 5	8 / 9	9 / 17
SR-ARE	Similarity	85 (23)	4 / 5	6 / 9	10 / 17
SR-MMP	Baseline	95 (21)	3 / 5	6 / 10	8 / 19
SR-MMP	Fusion	95 (21)	4 / 5	7 / 10	11 / 19
SR-MMP	Similarity	95 (21)	5 / 5	9 / 10	12 / 19

All results are from logistic-regression models evaluated in USC-HepG2. Baseline: structure plus segmented-object count; Fusion: selected morphology fusion; Similarity: adapted similarity comparator. Budgets use the ceiling of 5%, 10% or 20% of compounds. Values are observed active compounds / compounds selected for testing, not confidence intervals.

In the secondary new-entity/seen-scaffold stratum, 44 and 56 compounds were evaluated for SR-ARE and SR-MMP, respectively, comprising 67 distinct compounds. The two-endpoint mean Brier gain was

smaller at 0.0056 (95% CI: −0.0180–0.0251; Supplementary Tables S3-S4). This was a different compound population, not a paired test of the effect of scaffold novelty.

Cellular-context transfer

The incremental gain changed when the same compounds were represented by profiles from a different cellular context (Fig. 3a). For SR-ARE, gain was 0.0362 in FMP-HepG2 and − 0.0171 in FMP-U2OS. The paired difference was − 0.0534 (95% CI: −0.1068 - −0.0018). For SR-MMP, gain changed from 0.0342 to − 0.0015, a difference of − 0.0357 (− 0.0633 to − 0.0100). Both contrasts retained identical compound identities and outcome labels. Gains decreased for 39/85 SR-ARE pairs and 41/95 SR-MMP pairs; larger declines in these pairs outweighed increases elsewhere (Fig. 3b).

a, Logistic-regression mean Brier gain in two FMP experimental contexts for identical held-out compounds. Bars show percentile 95% intervals from 5,000 scaffold resamples. The paired U2OS-minus-HepG2 changes are − 0.0534 (95% CI:−0.1068 - −0.0018) for SR-ARE and − 0.0357 (− 0.0633 - −0.0100) for SR-MMP. b, Each point shows the same compound-endpoint pair in both contexts, using (baseline probability − label)² − (fusion probability − label)². Points below the identity line have lower gains in U2OS. All 85 SR-ARE and 95 SR-MMP pairs are shown, representing 117 distinct compounds overall; endpoint overlap does not create independent replicates. Zero lines indicate equal baseline and fusion error. Biological background and acquisition differ together.

Paired comparisons among HepG2 laboratories were less conclusive: their intervals included zero (Supplementary Table S5). These comparisons did not establish equivalence between laboratories. The positive USC result demonstrated an average external gain in that setting, whereas the U2OS comparison showed a loss of gain under a larger experimental change. Cell line and acquisition conditions varied together, so the paired contrast did not isolate a biological cell-line effect.

Sources of instability

Replicate measurements contained reproducible but imperfect signal. Within the strict held-out compound pool, site-level medians across six replicate-pair correlation summaries ranged from 0.33 to 0.58, exceeding the corresponding identity-permuted controls (Supplementary Table S6). In USC, replacing replicate aggregates with single-well inputs increased fusion Brier loss by 0.0224 for SR-ARE and 0.0157 for SR-MMP. Both intervals excluded zero (Supplementary Figure S1a). This comparison changed inputs to an aggregate-trained model; it did not compare independently optimized training pipelines.

Removing baseline-predictable morphology did not preserve the external benefit. Cross-fitted residual fusion had a mean Brier score 0.0571 higher than the baseline, with a wide interval from − 0.0099 to 0.1218 (Supplementary Figure S1b; Supplementary Table S7). The estimate did not support a stable

residual increment. It also did not distinguish redundant information from useful shared signal removed by residualization.

The supporting analyses separated comparator dependence from pairing dependence (Supplementary Figures S2 - S3; Supplementary Tables S10 - S15). Against a fixed structure-plus-count-proxy baseline, JUMP's SR-ARE AP increment had a median of 0.0008 and mean of 0.0107; the other three endpoints had negative mean increments. Across 24 EveBio tasks, morphology improved on count alone, but the median of endpoint-specific median AP differences versus structure plus count was − 0.0270 for logistic regression. In Cell Health, correct perturbation pairing remained informative during cell-line holdout, while the median morphology-minus-count contrast became negative. Eight endpoints met exploratory retention criteria, with 6–9 retained under alternative thresholds. These different outcomes and controls do not constitute pooled estimates or independent replications of the EU-OPENSCREEN effect.

The cardiomyocyte and OASIS analyses addressed measurement interpretation rather than validating the principal prediction task. Cardiomyocyte response slopes were moderately reproducible after aggregation and strongly influenced by the highest dose (Supplementary Figure S4a,b). The pathway responses were dominated by a common response axis (Supplementary Table S16). Their low predictability therefore could not establish a biological prediction ceiling. In OASIS, morphology had a lower response concentration than cell count for 116 of 121 compounds detected in all four modalities (Supplementary Figure S5a). Geometric mean comparator-to-morphology ratios were 1.79 for metabolic activity, 4.28 for cell count and 8.15 for lactate dehydrogenase release (Supplementary Figure S5b). This comparison described the selected complete cases, not all screened compounds. Full results, including the negative and threshold-sensitive analyses, are retained in Supplementary Tables S10-S20.

Limits of model selection

The more elaborate context rule did not improve the original primary comparison. For logistic regression, both rules selected fusion for both endpoints. For random forest, both selected fusion for SR-ARE and the baseline for SR-MMP. Their external mean Brier difference was therefore exactly zero within each model family. The random-forest choice also missed a potential benefit: fusion subsequently performed better than the baseline on external SR-MMP data.

The two rules likewise made identical choices in all 2,000 formal simulations (Fig. 4a). Fusion was selected in 363 of 400 informative settings, 1 of 400 uninformative settings, 9 of 400 redundant settings and 140 of 400 partial-mismatch settings. Partial mismatch permuted morphology for approximately half the entities within exchange blocks, leaving some signal intact. Its mean target gain if fusion was always used was 0.0084, versus 0.0375 in the informative condition (Fig. 4b). Selection reduced the mismatch condition's mean realized gain to 0.0035 (Supplementary Tables S8–S9). In the visible-shift condition, fusion was rejected in all 200 replicates (Fig. 4a). In the hidden-shift condition, fusion was selected in 190 of 200 replicates despite a negative mean target gain (Fig. 4a,b). Neither rule reliably identified a shift absent from development data.

a, Fusion selection by either rule. Decisions are identical for every replicate. Bars give accepted fractions; labels give accepted/total counts and error bars Wilson 95% intervals. b, Distributions of target Brier gains when fusion is used in every replicate, regardless of the selection decision. Violin widths show kernel density; points and bars show mean ± 1.96 Monte Carlo standard errors, not a prediction interval. Positive favors fusion. All 2,000 independent replicates are included: 400 each for uninformative, redundant, informative and partial-mismatch morphology; 200 each for visible and hidden shifts. Results pool prespecified sample-size and noise settings. Partial mismatch can retain signal. Realized selection-rule gains are separately reported in Tables S8-S9.

Discussion

The main finding is a separation between incremental value and transferability. Morphology improved probability prediction beyond chemical structure and measured count in the evaluated external HepG2 laboratory. That improvement weakened when the same compounds were represented by U2OS profiles. Its magnitude also depended on the model, and the individual antioxidant-response estimate was uncertain. The practical contribution is a quantified boundary for using an additional cellular measurement, rather than a new algorithm that consistently outperforms existing methods.

This distinction extends the cell-count benchmark without contradicting it. Seal and colleagues showed that many apparently demanding bioactivity tasks can be predicted from count alone [13]. We included both count and structure in the comparator and still observed a positive mean external increment. The implemented baseline therefore did not capture all useful predictive information in that setting. This does not prove conditional biological independence: morphology could measure general stress more precisely than segmented-object count, and model capacity could affect the comparison. Studies comparing fingerprints and morphological descriptors for cytotoxicity support a contribution from broad stress responses [30]. The residualization result leaves this explanation open.

Our contribution differs from earlier demonstrations of broad predictive coverage or successful screening. Moshkov and colleagues asked how multiple representations expanded assay prediction [10]. Fredin Haslum and colleagues tested whether morphology-based prioritization improved prospective hit discovery [11]. We instead measured the increment over a count-inclusive comparator after chemical-history exclusion and then changed the cellular input for identical compounds. Wolff and colleagues had already characterized laboratory and cell-background differences in the source resource [20]. Our analysis connects those differences to downstream incremental prediction, rather than claiming to discover that cellular background affects morphology. Seal and colleagues' similarity-based fusion also remained competitive here, showing that this comparison concerns conditions of usefulness rather than model leadership [23].

Three explanations for the context-dependent gain remain plausible. First, the images may capture endpoint-related stress responses whose expression differs between experimental systems. Prior work combined morphology, expression and structure for mitochondrial-toxicity detection, supporting the

relevance of these complementary measurements [31]. Genetic-perturbation and cell-health studies likewise link image patterns to interpretable cellular phenotypes [32, 26]. In our Cell Health analysis, mitochondrial-channel removal most consistently reduced prediction among retained endpoints, including DNA-damage and cell-cycle readouts. This is compatible with an integrated stress response, but channel importance does not establish mitochondrial causation. Interpretable morphology representations and morphological subprofiles offer ways to examine such associations more directly [33, 34].

Second, a general stress response may track the assay label without identifying its intended molecular mechanism. Tox21 reporter activity can accompany nonspecific cytotoxic responses, while compound reactivity can contribute to promiscuous activity [14, 35]. Related proliferation and cell-death signals also occur in transcriptomic data [15]. Count adjustment addresses part of this explanation, but not every dimension of cell health. The retained gain could reflect sublethal changes, altered cellular organization, or a more sensitive measurement of the same underlying stress. OASIS's paired response concentrations support detection of morphological changes at lower concentrations than count changes in selected compounds [27]. Greater concentration sensitivity, however, is not evidence of target specificity or clinical harm. Target identification remains a separate inference requiring evidence linking the phenotype to molecular action [36].

Third, measurement and acquisition differences may change the learned relationship independently of molecular biology. Profile-analysis choices affect morphological similarity, and combining different imaging assays does not make their readouts interchangeable [37, 38]. The single-well sensitivity in this study shows that aggregation contributed materially to the observed performance. EU-OPENSCREEN's cell-background and acquisition differences further prevent attribution of the U2OS result to lineage alone [20]. A discriminating follow-up would profile matched compounds in both cell types at matched doses and times, with harmonized acquisition and independent viability measurements. Persistence of the difference after harmonization would strengthen a biological explanation; attenuation would implicate measurement conditions. Neither outcome has been established here.

The earlier analyses clarify why positive and negative findings should not be reduced to a vote on whether a modality works. EveBio morphology contained information beyond count, yet often did not add to structure plus count. Cell Health retained perturbation-specific information under cell-line holdout without retaining an overall advantage over count. These are different failures of incremental prediction. Likewise, our cardiomyocyte target was a noisy, high-dose-dominated summary, not an independent measurement of nine mechanisms. That setting differs from studies directly relating cardiac-cell profiles or functional measurements to cardiotoxicity [39, 40]. The originating transcriptomic study combined molecular and functional information for hazard assessment [41]; our cross-context prediction subset cannot invalidate that broader approach.

The results pose a specific test for more complex representations: improvement in the source task should be distinguished from incremental prediction in a new cellular setting. Cross-modal learning has

improved morphology-transcriptome clustering and bioactivity modeling, but those gains answer the objectives and comparisons used in that study [42]. Systema examines a related distinction in expression prediction by separating perturbation responses from systematic variation [43]. For our question, a richer representation would need to retain an advantage over structure plus count when the cellular input changes. Better alignment alone would not establish that advantage, just as model complexity alone has not assured improvement over simple linear perturbation predictors [44].

The negative selection-rule result has a specific implication. Additional conditions are useful only when they change decisions in a beneficial way. Here, laboratory-consistency requirements were redundant with the simpler rule in the evaluated tasks and simulations. The hidden-shift scenario also illustrates why favorable development evidence cannot reveal every target-specific reversal. Chemical-model evaluation guidelines and recent drug-response benchmarking similarly stress matching validation to the intended application [45, 46]. These observations favor obtaining representative target-context measurements over adding unvalidated layers of selection criteria. They do not establish that context-aware selection can never help under other designs.

For preclinical researchers, the actionable output is a conditional prioritization result. Structure-only prediction may be available before an experiment, whereas morphology requires a cellular assay. Public screening resources reduce the acquisition burden, but their chemical and phenotypic coverage remains shaped by library design [47, 48]. Repurposing collections provide another useful but selected compound space [49]. Our fixed-budget results show that a better average probability score need not retrieve more actives at every budget. The appropriate operating point therefore depends on the intended follow-up assay. Image-based environmental screening and multimodal toxicity prediction have related aims, but their endpoints differ from those evaluated here [50, 51]. Studies targeting drug-induced cardiotoxicity ranks or disease-linked perturbations address another level of biological interpretation [52, 53]. Neither a Tox21 activity label nor a favorable computational score is sufficient for a patient-level prescribing decision.

Strengths

The principal comparison combined an actual count-inclusive baseline, chemical-history exclusion and an external laboratory. The paired cellular-context test held compounds and labels constant, separating input-context changes from a change in the chemical test population. We retained a competitive published comparator, alternative models, measurement diagnostics and an unfavorable original primary result. Supporting resources offered different tests of redundancy, alignment and measurement quality rather than nominally repeating the same validation. Such diversity is useful when its limits are explicit: large perturbation resources contain multiple related measurements, not automatically independent evidence [54, 55].

Limitations

The main result concerns two qualified stress-response endpoints and one excluded HepG2 laboratory. Ten other assays lacked the required sample counts. SR-MMP gains were supported in both model families, but SR-ARE intervals included zero; no formal endpoint-by-model interaction was tested. The SR-ARE active fraction changed from 20.2% in development to 27.1% in strict evaluation, compared with 22.0% and 22.1% for SR-MMP. Although models were compared on identical test compounds, Brier gains combine probability-scale accuracy and discrimination; we did not decompose them into calibration and discrimination components.

Scaffold separation does not eliminate every chemical similarity, and benchmark composition can affect apparent performance [56, 57]. Labels were reused across contexts, intervals were not multiplicity-adjusted, and resampling did not include model-refitting uncertainty. We used extracted features rather than image encoders; deep phenotyping and few-shot learning address representations not tested here [58, 59]. Biological and technical changes were inseparable in the U2OS comparison. The identity correction and empirical reframing followed initial scoring. Prospective target-context evaluation is needed before operational use, consistent with the distinction between benchmark success and real-world toxicity prediction [60].

Conclusion

In this retrospective evaluation, morphology added predictive value beyond structure and segmented-object count in an external HepG2 laboratory, most consistently for mitochondrial activity. Gains weakened for the same compounds under U2OS experimental conditions, and adding laboratory-consistency criteria did not improve the original selection-rule comparison. These findings support target-context testing of incremental value, not a generally superior fusion algorithm or a clinical toxicity predictor.

Abbreviations

Abbreviation
Full term
AP
average precision
CI
confidence interval
DNA
deoxyribonucleic acid
FMP
Leibniz Institute for Molecular Pharmacology
IMTM
Institute of Molecular and Translational Medicine
JUMP

Joint Undertaking for Morphological Profiling
LDH
lactate dehydrogenase
MEDINA
Fundación MEDINA
SR-ARE
stress-response antioxidant response element assay
SR-MMP
stress-response mitochondrial membrane potential assay
USC
University of Santiago de Compostela

Declarations

Ethics approval and consent to participate

Not applicable. This study reanalyzed public compound and cell-profiling data and did not use patient records.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Use of AI

The authors used AI tools to improve language, grammar and spelling, translate text, and refine sentence structure, section organization and figure legends. AI also helped develop statistical analysis and plotting code for the public datasets used in this study. The authors reviewed, verified and revised the analytical procedures, code, results, figures and interpretations. These tools were not used to fabricate research data or findings, or to create or alter original research images. The authors take full responsibility for the accuracy, integrity and originality of the final work.

Funding

This work was supported by the General Program of the Chongqing Natural Science Foundation (grant no. CSTB2024NSCQ-MSX0155), the Science and Technology Bureau and the Health Committee of

Chongqing, China (grant no. 2024MSXM013), and the Kuan Ren Key-talents Program of The Second Affiliated Hospital of Chongqing Medical University.

Author Contribution

Y.Y. contributed to study design, data curation, formal analysis, visualization, writing the original draft, and manuscript review and editing. X.D. contributed to software development and manuscript review and editing. X.L. contributed to study design and supervision. Z.X.: verification of the sources of the public datasets; Y.W.: checking consistency between the figure legends and the main text; M.X.: review of the methodological descriptions for reproducibility; K.L.: verification of the definitions and footnotes in the supplementary tables; P.S.: review of the clinical interpretation and its limitations in the Discussion; P.X.: verification of the funding information and substantive review of the interpretation of the findings.

Acknowledgement

We thank the investigators and data providers responsible for the public datasets used in this study and the Gene Expression Omnibus platform for making research data accessible.

Availability of data and materials

EU-OPENSCREEN aggregated profiles were obtained from version 3.0.0 of the public Zenodo dataset (https://doi.org/10.5281/zenodo.19347244). Tox21 activity labels were obtained from the complete training-data archive of the NCATS Tox21 Data Challenge (https://tripod.nih.gov/tox21/challenge/data.jsp). The source archive used was tox21_10k_data_all.sdf.zip.

JUMP profiles and metadata are available through https://github.com/jump-cellpainting/datasets/tree/v0.11.0; the analysis used assembled COMPOUND/v1.0 profiles, with Harmony profiles for morphology and profiles_var_mad_int.parquet for count proxies. The EveBio benchmark inputs, including CP_all_EveBio.csv and evebio_assay_data_smiles.csv, are available at https://github.com/srijitseal/The_Seal_Files/tree/main/The_EveBio_dataset.

Cell Health consensus profiles and readouts are available from https://github.com/broadinstitute/cell-health/tree/v2.0. OASIS point-of-departure summaries are available in 2_downstream_analysis/compiled_results/SI_tables of https://github.com/broadinstitute/2024_09_09_Axiom_OASIS. The OASIS analysis reused these tabulated summaries rather than reprocessing raw images.

Cardiomyocyte transcriptomic data are available in GEO under GSE262419 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE262419). L1000 data are available under

GSE92742 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742) and GSE70138 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70138).

Analysis subsets and processing are described in Additional file 1. Additional file 2 contains machine-readable result tables and figure source data. These files do not replace the original profile archives or constitute the complete analysis software. No restricted patient-level data are included.

Code availability

All the codes of this research have been uploaded to GitHub: https://github.com/Ethan1990-code/Morphology-Code.

References

1. 1. Moffat JG, Vincent F, Lee JA, et al (2017) Opportunities and challenges in phenotypic drug discovery: an industry perspective. Nature Reviews Drug Discovery 16:531–543. https://doi.org/10.1038/nrd.2017.111
2. 2. Warchal SJ, Unciti-Broceta A, Carragher NO (2016) Next-Generation Phenotypic Screening. Future Medicinal Chemistry 8:1331–1347. https://doi.org/10.4155/fmc-2016-0025
3. 3. Bray M-A, Singh S, Han H, et al (2016) Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nature Protocols 11:1757–1774. https://doi.org/10.1038/nprot.2016.105
4. 4. McQuin C, Goodman A, Chernyshev V, et al (2018) CellProfiler 3.0: Next-generation image processing for biology. PLOS Biology 16:e2005970. https://doi.org/10.1371/journal.pbio.2005970
5. 5. Perlman ZE, Slack MD, Feng Y, et al (2004) Multidimensional Drug Profiling By Automated Microscopy. Science 306:1194–1198. https://doi.org/10.1126/science.1100709
6. 6. Caicedo JC, Singh S, Carpenter AE (2016) Applications in image-based profiling of perturbations. Current Opinion in Biotechnology 39:134–142. https://doi.org/10.1016/j.copbio.2016.04.003
7. 7. Chandrasekaran SN, Ceulemans H, Boyd JD, Carpenter AE (2021) Image-based profiling for drug discovery: due for a machine-learning upgrade? Nature Reviews Drug Discovery 20:145–159. https://doi.org/10.1038/s41573-020-00117-w
8. 8. Simm J, Klambauer G, Arany A, et al (2018) Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. Cell Chemical Biology 25:611–618.e3. https://doi.org/10.1016/j.chembiol.2018.01.015
9. 9. Trapotsi M-A, Mervin LH, Afzal AM, et al (2021) Comparison of Chemical Structure and Cell Morphology Information for Multitask Bioactivity Predictions. Journal of Chemical Information and Modeling 61:1444–1456. https://doi.org/10.1021/acs.jcim.0c00864
10. 10. Moshkov N, Becker T, Yang K, et al (2023) Predicting compound activity from phenotypic profiles and chemical structures. Nature Communications 14:1967. https://doi.org/10.1038/s41467-023-

37570-1

11. 11. Fredin Haslum J, Lardeau C-H, Karlsson J, et al (2024) Cell Painting-based bioactivity prediction boosts high-throughput screening hit-rates and compound diversity. Nature Communications 15:3470. https://doi.org/10.1038/s41467-024-47171-1
12. 12. Way GP, Natoli T, Adeboye A, et al (2022) Morphology and gene expression profiling provide complementary information for mapping cell state. Cell Systems 13:911–923.e9. https://doi.org/10.1016/j.cels.2022.10.001
13. 13. Seal S, Dee W, Shah A, et al (2026) Counting cells can accurately predict small-molecule bioactivity benchmarks. Nature Communications 17:2436. https://doi.org/10.1038/s41467-026-68725-5
14. 14. Escher BI, Henneberger L, König M, et al (2020) Cytotoxicity Burst? Differentiating Specific from Nonspecific Effects in Tox21 in Vitro Reporter Gene Assays. Environmental Health Perspectives 128:077007. https://doi.org/10.1289/ehp6664
15. 15. Szalai B, Subramanian V, Holland CH, et al (2019) Signatures of cell death and proliferation in perturbation transcriptomics data—from confounding factor to effective prediction. Nucleic Acids Research 47:10010–10026. https://doi.org/10.1093/nar/gkz805
16. 16. Yang K, Swanson K, Jin W, et al (2019) Analyzing Learned Molecular Representations for Property Prediction. Journal of Chemical Information and Modeling 59:3370–3388. https://doi.org/10.1021/acs.jcim.9b00237
17. 17. Wu Z, Ramsundar B, Feinberg EN, et al (2018) MoleculeNet: a benchmark for molecular machine learning. Chemical Science 9:513–530. https://doi.org/10.1039/c7sc02664a
18. 18. Niepel M, Hafner M, Duan Q, et al (2017) Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. Nature Communications 8:1186. https://doi.org/10.1038/s41467-017-01383-w
19. 19. Lim N, Pavlidis P (2021) Evaluation of connectivity map shows limited reproducibility in drug repositioning. Scientific Reports 11:17624. https://doi.org/10.1038/s41598-021-97005-z
20. 20. Wolff C, Neuenschwander M, Beese CJ, et al (2025) Morphological profiling data resource enables prediction of chemical compound properties. iScience 28:112445. https://doi.org/10.1016/j.isci.2025.112445
21. 21. Bemis GW, Murcko MA (1996) The Properties of Known Drugs. 1. Molecular Frameworks. Journal of Medicinal Chemistry 39:2887–2893. https://doi.org/10.1021/jm9602928
22. 22. Caicedo JC, Cooper S, Heigwer F, et al (2017) Data-analysis strategies for image-based cell profiling. Nature Methods 14:849–863. https://doi.org/10.1038/nmeth.4397
23. 23. Seal S, Yang H, Trapotsi M-A, et al (2023) Merging bioactivity predictions from cell morphology and chemical fingerprint models using similarity to training data. Journal of Cheminformatics 15:56. https://doi.org/10.1186/s13321-023-00723-x
24. 24. Subramanian A, Narayan R, Corsello SM, et al (2017) A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell 171:1437–1452.e17.

https://doi.org/10.1016/j.cell.2017.10.049

25. 25. Chandrasekaran SN, Ackerman J, Alix E, et al (2023) JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations
26. 26. Way GP, Kost-Alimova M, Shibue T, et al (2021) Predicting cell health phenotypes using image-based morphology profiling. Molecular Biology of the Cell 32:995–1005. https://doi.org/10.1091/mbc.e20-12-0784
27. 27. Ewald JD, Titterton KL, Bäuerle A, et al (2026) Cell Painting for cytotoxicity and mode-of-action analysis in primary human hepatocytes. Cell Systems 17:101566. https://doi.org/10.1016/j.cels.2026.101566
28. 28. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7:91. https://doi.org/10.1186/1471-2105-7-91
29. 29. Saito T, Rehmsmeier M (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE 10:e0118432. https://doi.org/10.1371/journal.pone.0118432
30. 30. Seal S, Yang H, Vollmers L, Bender A (2021) Comparison of Cellular Morphological Descriptors and Molecular Fingerprints for the Prediction of Cytotoxicity- and Proliferation-Related Assays. Chemical Research in Toxicology 34:422–437. https://doi.org/10.1021/acs.chemrestox.0c00303
31. 31. Seal S, Carreras-Puigvert J, Trapotsi M-A, et al (2022) Integrating cell morphology with gene expression and chemical structure to aid mitochondrial toxicity detection. Communications Biology 5:858. https://doi.org/10.1038/s42003-022-03763-5
32. 32. Rohban MH, Singh S, Wu X, et al (2017) Systematic morphological profiling of human gene and allele function via Cell Painting. eLife 6:e24060. https://doi.org/10.7554/elife.24060
33. 33. Seal S, Carreras-Puigvert J, Singh S, et al (2024) From pixels to phenotypes: Integrating image-based profiling with cell health data as BioMorph features improves interpretability. Molecular Biology of the Cell 35:mr2. https://doi.org/10.1091/mbc.e23-08-0298
34. 34. Pahl A, Schölermann B, Lampe P, et al (2023) Morphological subprofile analysis for bioactivity annotation of small molecules. Cell Chemical Biology 30:839–853.e7. https://doi.org/10.1016/j.chembiol.2023.06.003
35. 35. Matlock MK, Hughes TB, Dahlin JL, Swamidass SJ (2018) Modeling Small-Molecule Reactivity Identifies Promiscuous Bioactive Compounds. Journal of Chemical Information and Modeling 58:1483–1500. https://doi.org/10.1021/acs.jcim.8b00104
36. 36. Schenone M, Dančík V, Wagner BK, Clemons PA (2013) Target identification and mechanism of action in chemical biology and drug discovery. Nature Chemical Biology 9:232–240. https://doi.org/10.1038/nchembio.1199
37. 37. Ljosa V, Caie PD, ter Horst R, et al (2013) Comparison of Methods for Image-Based Profiling of Cellular Morphological Responses to Small-Molecule Treatment. SLAS Discovery 18:1321–1329. https://doi.org/10.1177/1087057113503553

38. 38. Rose F, Basu S, Rexhepaj E, et al (2018) Compound Functional Prediction Using Multiple Unrelated Morphological Profiling Assays. SLAS Technology 23:243–251. https://doi.org/10.1177/2472630317740831
39. 39. van Hasselt JGC, Rahman R, Hansen J, et al (2020) Transcriptomic profiling of human cardiac cells predicts protein kinase inhibitor-associated cardiotoxicity. Nature Communications 11:4809. https://doi.org/10.1038/s41467-020-18396-7
40. 40. Blanchette AD, Burnett SD, Grimm FA, et al (2020) A Bayesian Method for Population-wide Cardiotoxicity Hazard and Risk Characterization Using an In Vitro Human Model. Toxicological Sciences 178:391–403. https://doi.org/10.1093/toxsci/kfaa151
41. 41. Tsai H-HD, Ford LC, Burnett SD, et al (2024) Informing Hazard Identification and Risk Characterization of Environmental Chemicals by Combining Transcriptomic and Functional Data from Human-Induced Pluripotent Stem-Cell-Derived Cardiomyocytes. Chemical Research in Toxicology 37:1428–1444. https://doi.org/10.1021/acs.chemrestox.4c00193
42. 42. Ha SV, Jaensch S, Kańduła MM, et al (2025) Cross modality learning of cell painting and transcriptomics data improves mechanism of action clustering and bioactivity modelling. Scientific Reports 15:23010. https://doi.org/10.1038/s41598-025-05914-0
43. 43. Viñas Torné R, Wiatrak M, Piran Z, et al (2026) Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. Nature Biotechnology 44:1050–1059. https://doi.org/10.1038/s41587-025-02777-8
44. 44. Ahlmann-Eltze C, Huber W, Anders S (2025) Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Nature Methods 22:1657–1661. https://doi.org/10.1038/s41592-025-02772-6
45. 45. Bender A, Schneider N, Segler M, et al (2022) Evaluation guidelines for machine learning tools in the chemical sciences. Nature Reviews Chemistry 6:428–442. https://doi.org/10.1038/s41570-022-00391-9
46. 46. Bernett J, Iversen P, Picciani M, et al (2026) Critical evaluation of drug response prediction models with DrEval. Nature Communications 17:4238. https://doi.org/10.1038/s41467-026-72903-w
47. 47. Bray M-A, Gustafsdottir SM, Rohban MH, et al (2017) A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. GigaScience 6:giw014. https://doi.org/10.1093/gigascience/giw014
48. 48. Wawer MJ, Li K, Gustafsdottir SM, et al (2014) Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. Proceedings of the National Academy of Sciences 111:10911–10916. https://doi.org/10.1073/pnas.1410933111
49. 49. Corsello SM, Bittker JA, Liu Z, et al (2017) The Drug Repurposing Hub: a next-generation drug library and information resource. Nature Medicine 23:405–408. https://doi.org/10.1038/nm.4306

50. 50. Nyffeler J, Willis C, Lougee R, et al (2020) Bioactivity screening of environmental chemicals using imaging-based high-throughput phenotypic profiling. Toxicology and Applied Pharmacology 389:114876. https://doi.org/10.1016/j.taap.2019.114876
51. 51. Liu A, Seal S, Yang H, Bender A (2023) Using chemical and biological data to predict drug toxicity. SLAS Discovery 28:53–64. https://doi.org/10.1016/j.slasd.2022.12.003
52. 52. Seal S, Spjuth O, Hosseini-Gerami L, et al (2024) Insights into Drug Cardiotoxicity from Biological and Chemical Data: The First Public Classifiers for FDA Drug-Induced Cardiotoxicity Rank. Journal of Chemical Information and Modeling 64:1172–1186. https://doi.org/10.1021/acs.jcim.3c01834
53. 53. Cerisier N, Dafniet B, Badel A, Taboureau O (2023) Linking chemicals, genes and morphological perturbations to diseases. Toxicology and Applied Pharmacology 461:116407. https://doi.org/10.1016/j.taap.2023.116407
54. 54. Haghighi M, Caicedo JC, Cimini BA, et al (2022) High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. Nature Methods 19:1550–1557. https://doi.org/10.1038/s41592-022-01667-0
55. 55. Keenan AB, Jenkins SL, Jagodnik KM, et al (2018) The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. Cell Systems 6:13–24. https://doi.org/10.1016/j.cels.2017.11.001
56. 56. Rohrer SG, Baumann K (2009) Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. Journal of Chemical Information and Modeling 49:169–184. https://doi.org/10.1021/ci8002649
57. 57. Chaput L, Martinez-Sanz J, Saettel N, Mouawad L (2016) Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance. Journal of Cheminformatics 8:56. https://doi.org/10.1186/s13321-016-0167-x
58. 58. Pratapa A, Doron M, Caicedo JC (2021) Image-based cell phenotyping with deep learning. Current Opinion in Chemical Biology 65:9–17. https://doi.org/10.1016/j.cbpa.2021.04.001
59. 59. Ha SV, Leuschner L, Czodrowski P (2024) FSL-CP: a benchmark for small molecule activity few-shot prediction using cell microscopy images. Digital Discovery 3:719–727. https://doi.org/10.1039/d3dd00205e
60. 60. Seal S, Mahale M, García-Ortegón M, et al (2025) Machine Learning for Toxicity Prediction Using Chemical Structures: Pillars for Success in the Real World. Chemical Research in Toxicology 38:759–807. https://doi.org/10.1021/acs.chemrestox.5c00033

Figures

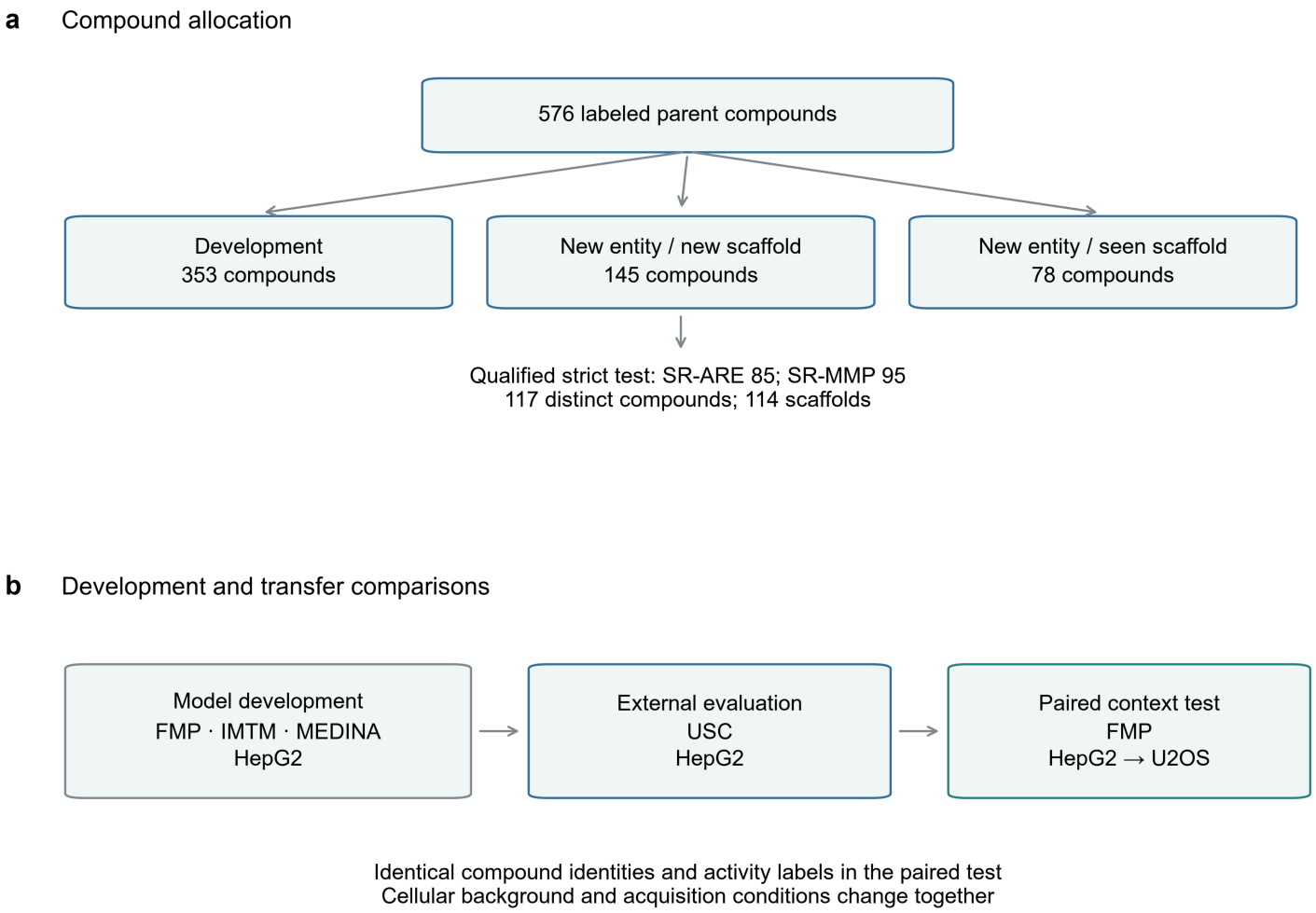


Figure 1

Study design and evaluation populations

a, Allocation by prior chemical history. Of 576 labeled parent compounds, 573 had profiles at all four HepG2 laboratories; endpoint qualification additionally required a nonmissing activity label. SR-ARE and SR-MMP strict tests overlap and are not 180 independent compounds. b, Models are developed in three HepG2 laboratories and evaluated in an excluded laboratory. The paired comparison changes the cellular input for the same strict held-out compounds; it does not isolate cell lineage from acquisition conditions. Arrows denote analysis comparisons, not chronological sample collection. Counts have no uncertainty intervals.

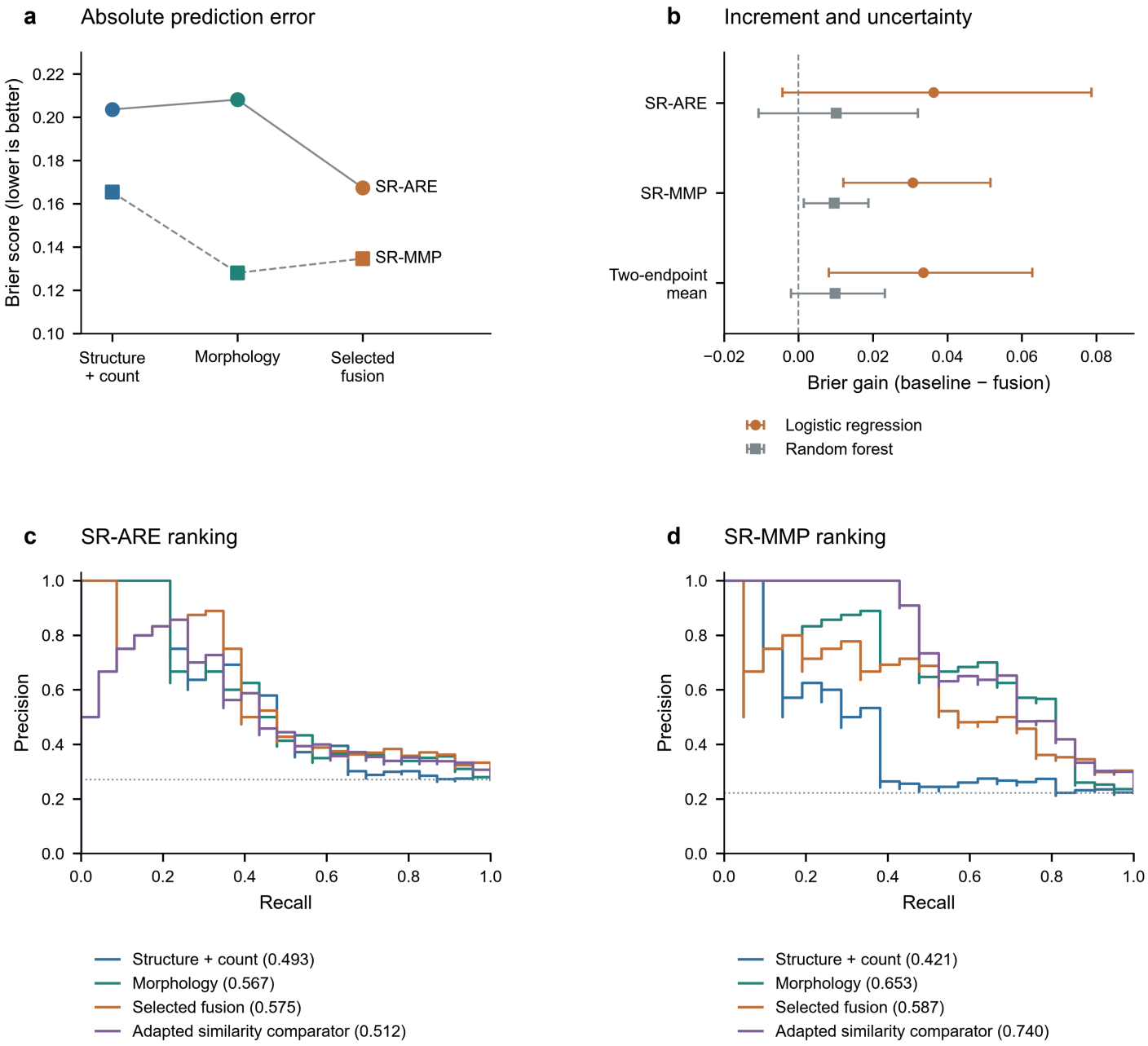


Figure 2

Prediction gains and screening trade-offs

a, Observed Brier scores from logistic-regression models using structure plus count, morphology alone or selected fusion in USC-HepG2, with circles for SR-ARE and squares for SR-MMP. Connecting lines identify an endpoint across categorical inputs, not a continuous trajectory. b, Brier gains for logistic regression and random forest. Points and bars show estimates and percentile 95% intervals from 5,000 paired scaffold bootstrap resamples; positive favors fusion. c,d, Precision-recall curves from logistic-regression predictions. Legend values are average precision; dotted lines mark the active fraction.

Curves are descriptive, without confidence bands. SR-ARE has 85 compounds/85 singleton scaffolds; SR-MMP has 95 compounds/92 scaffolds; their union has 117 compounds/114 scaffolds. Joint resampling preserves endpoint overlap and does not refit models. Intervals are not multiplicity-adjusted. The adapted similarity comparator is not an exact reproduction of the original benchmark.

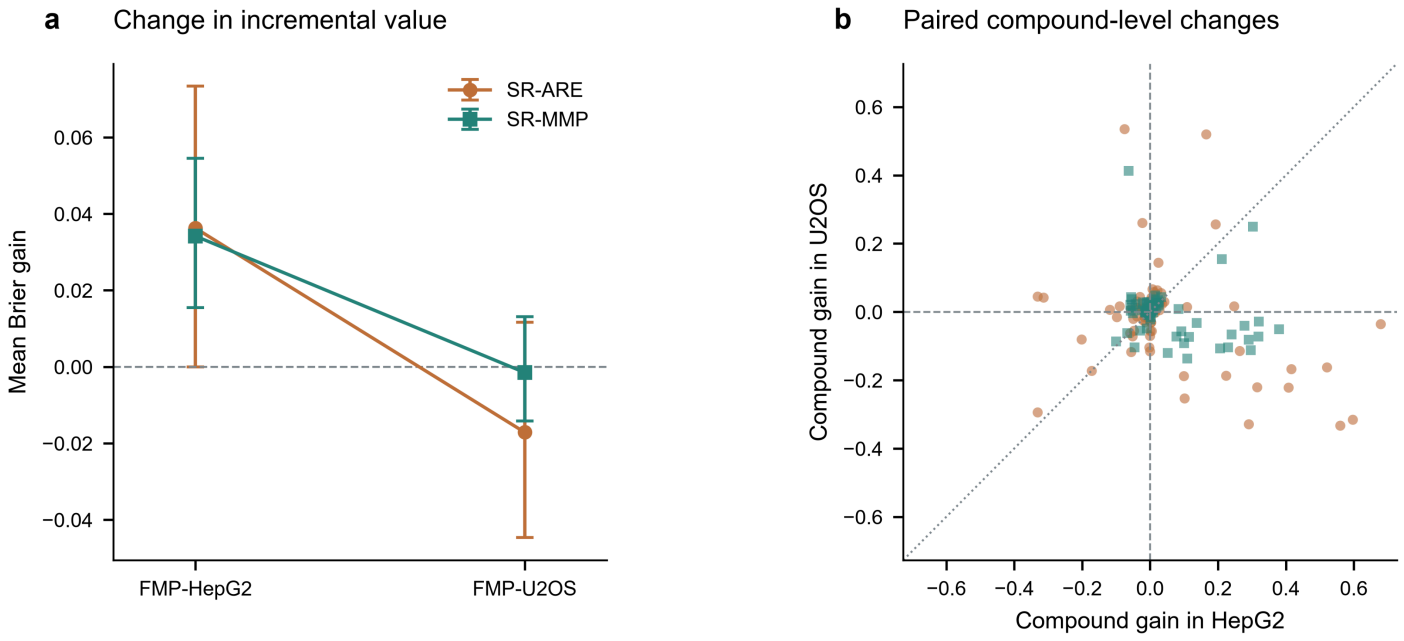


Figure 3

Changes in gain under paired context transfer

a, Logistic-regression mean Brier gain in two FMP experimental contexts for identical held-out compounds. Bars show percentile 95% intervals from 5,000 scaffold resamples. The paired U2OS-minus-HepG2 changes are −0.0534 (95% CI:−0.1068 - −0.0018) for SR-ARE and −0.0357 (−0.0633 - −0.0100) for SR-MMP. b, Each point shows the same compound-endpoint pair in both contexts, using (baseline probability − label)² − (fusion probability − label)². Points below the identity line have lower gains in U2OS. All 85 SR-ARE and 95 SR-MMP pairs are shown, representing 117 distinct compounds overall; endpoint overlap does not create independent replicates. Zero lines indicate equal baseline and fusion error. Biological background and acquisition differ together.

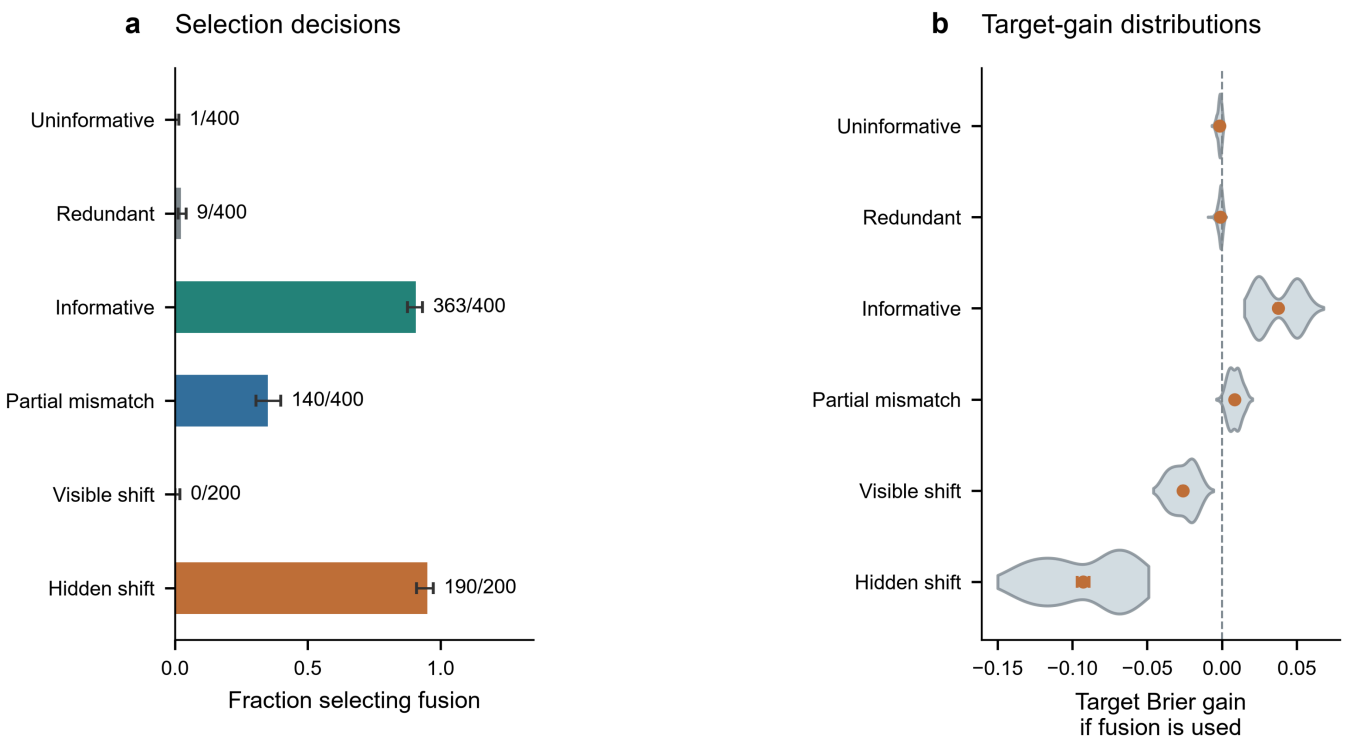


Figure 4

Selection decisions and target performance

a, Fusion selection by either rule. Decisions are identical for every replicate. Bars give accepted fractions; labels give accepted/total counts and error bars Wilson 95% intervals. b, Distributions of target Brier gains when fusion is used in every replicate, regardless of the selection decision. Violin widths show kernel density; points and bars show mean ±1.96 Monte Carlo standard errors, not a prediction interval. Positive favors fusion. All 2,000 independent replicates are included: 400 each for uninformative, redundant, informative and partial-mismatch morphology; 200 each for visible and hidden shifts. Results pool prespecified sample-size and noise settings. Partial mismatch can retain signal. Realized selection-rule gains are separately reported in Tables S8-S9.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.zip
- SupplementaryTables.zip
- SupplementaryFigures.zip