MorphoSuff: when image-based predictions can replace targeted measurements of molecular states, organelle organization and cellular activity in perturbation screens

Mengran Li[1,2,†], Jianqing Zhu[2,†], Bo Li[3,4,†], Chengyang Zhang[5,6], Zhenchao Tang[7], Jiaying Wang[8], Wenbin Xing[1], Boyu Zhang[1], Lingbei Meng[2,9], Bob Zhang[3], Junzhou Chen[1], Quanyou Cai[10], Jinfeng Xu[11], Zelin Zang[12], Ronghui Zhang[1,✉], Lian Zhang[13,✉], Stan Z. Li[14], Jinchao Xu[2,9,✉]

[1]School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China
[2]Shenzhen Loop Area Institute, Shenzhen, China
[3]PAMI Research Group, Department of Artificial Intelligence, University of Macau, Macau SAR
[4]Department of Computer Science, School of Computing, National University of Singapore, Singapore
[5]College of Computer Science, Sichuan University, Chengdu, China
[6]Department of Pathology and Institute of Clinical Pathology, West China Hospital, Chengdu, China
[7]AI for Life Sciences Lab, Tencent, Shenzhen, China
[8]School of Life Sciences, Sun Yat-sen University, Guangzhou, China
[9]The Chinese University of Hong Kong, Shenzhen, China
[10]Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, China
[11]The University of British Columbia, Vancouver, Canada
[12]Hong Kong Institute of Science and Innovation, Hong Kong SAR
[13]Shenzhen Research Institute of Big Data, Shenzhen, China
[14]Westlake University, Hangzhou, China

[†]These authors contributed equally: Mengran Li (limr39@mail2.sysu.edu.cn), Jianqing Zhu (jianqingzhu@slai.edu.cn) and Bo Li (Boom985426@gmail.com).
[✉]Corresponding authors: Ronghui Zhang (zhangrh25@mail.sysu.edu.cn), Lian Zhang (zhanglian@sribd.cn) and Jinchao Xu (jinchaoxu@slai.edu.cn).

Abstract

Predicting molecular states, organelle organization and cellular activity from brightfield or phase images could help reveal gene function and cellular responses to compounds while reducing assay costs and experimental workload. However, replacing direct measurements requires more than accurate predictions: it also requires testing whether they support reliable biological analyses. We introduce MorphoSuff, which first assesses the stability of measured perturbation effects and then repeats the same analyses with measured and predicted values to test whether the selection of strong-effect genes or compounds, comparisons of effect sizes and functional associations are preserved. We analysed 9,996,286 same-cell phase–fluorescence pairs spanning 1,000 gene knockouts and 52 molecular and functional targets in A549 cells. Predictions often preserved perturbation rankings while compressing differences in effect magnitude. Across targets, predictions recovered a median of 79% of the 50 knockouts with the strongest measured effects. They preserved MTOR-related LysoTracker responses but underestimated p53 and p21 responses to DDX18 knockout, missing candidates and a p21 functional association with ribosomal RNA maturation. We then used MorphoSuff to decide which candidate screens could rely on predictions and which assays to retain. In one group of gene knockouts, we selected screening analyses based on how well predictions preserved the ranking of perturbation effects, then tested those choices in a separate group of knockouts. Among the selected analyses, 3.3% failed to recover enough top candidates, compared with 12.3% when predictions were used for all analyses. This approach retained 89.5% of all successful analyses. For quantitative comparisons, we used pilot measurements to choose which direct assays to retain. With the same number of direct assays and with predictions and measurements compared on the same cells, this approach reduced quantitative error by 7.7% relative to selection based on predictive correlation. In an independent primary-hepatocyte dataset, we evaluated brightfield predictions for 217 compounds. The predictions recovered 10 of the 11 compounds with the largest measured losses of metabolic activity, a recovery rate of 90.9%, 18.2 percentage points higher than that of a matched cell-count predictor. Accurate candidate screening does not guarantee reliable effect quantification or functional interpretation. MorphoSuff tests these uses separately to help researchers decide which measurements can be omitted and which should be retained.

Introduction

Understanding the function of a gene or the action of a compound requires more than detecting a change in cells: it requires identifying which molecular states, organelle structures and cellular functions have changed. Genetic perturbation screens provide a systematic way to connect these changes to individual genes[1–3]. Combining these screens with high-content imaging enables the characterization of diverse cellular and subcellular phenotypes[4–6]. Cell Painting, for example, uses multiplexed fluorescent dyes to describe the morphology and organization of several cellular components[7,8]. Measuring a particular protein state, organelle phenotype or cellular function nevertheless usually requires an appropriate label or assay. As the numbers of genes, compounds and phenotypes increase, additional reagents, imaging channels and parallel experiments add to the cost and workload[9]. Obtaining biological information from existing images that would otherwise require additional assays could therefore substantially expand screening capacity[10].

Alongside fluorescence imaging, high-content screens can also acquire brightfield and phase images. These modalities capture cellular morphology through the intrinsic optical properties of cells, without adding a separate fluorescent label for every target of interest[11]. Previous studies have predicted selected fluorescent phenotypes from unlabelled images, including the distribution and three-dimensional organization of intracellular structures[12,13]. More recent models have extended virtual staining across imaging conditions and cellular states, supporting segmentation and cell tracking[14]. Standardized brightfield imaging of single cells in nanowells has also reproduced viability dose–response curves measured with fluorescent assays[15]. These advances connect to the ambition of AI-powered virtual cells: using existing observations to predict cellular states and responses to perturbation[16]. For experimental screening, the practical opportunity is to learn predictive relationships from a subset of samples with both images and targeted measurements, then reduce repeated target-specific labelling in subsequent experiments. The same images could thereby support more biological analyses. Realizing this benefit requires determining which uses of direct measurements can be supported by predictions.

Different uses require different information to be preserved. Identifying strong-effect genes or compounds requires recovering the candidates with the strongest measured effects. Comparing perturbations quantitatively also requires preserving effect magnitudes and the differences between them. Interpreting affected biological processes further depends on which specific genes and functional associations are retained. Predictions can correctly order perturbations while compressing the differences in their effects. Similar overall rankings can also conceal changes in candidate membership that affect subsequent functional analyses. Predictive correlation, recovery of strong candidates and functional interpretation are therefore not interchangeable evaluations. Application-appropriate validation in bioimage analysis likewise emphasizes matching evaluation to the biological question being addressed[17]. Recent evaluation of dynamic three-dimensional virtual staining has also identified trade-offs between structural and phenotypic fidelity that are obscured by single-metric evaluation[18]. For measurement substitution, the direct question is: **when measured values are replaced with predictions, which results of the same biological analysis are preserved?**

To interpret differences between predictions and measurements, we first need to know whether the measured perturbation effects are reproducible across cell samples or experimental replicates[19]. Reproducibility of perturbation profiles is also an explicit criterion in the optimization of image-based assays[8]. Good predictive performance can partly reflect simple properties such as cell number; cell counts provide strong baselines in several compound-activity prediction tasks[20]. Biological context can further alter the relationship between cellular morphology and organelle states[21]. Measurement substitution must therefore be assessed for a defined target, predictor and experimental setting, with the stability of the measured effects informing the evaluation of retained biological information.

Here we introduce MorphoSuff to assess ***measurement sufficiency***: whether morphology-based predictions provide enough information to replace a targeted measurement in a specified biological analysis. MorphoSuff first assesses the stability of measured perturbation effects. It then performs the same analyses using measured and image-predicted phenotypes, comparing the strong-effect genes or compounds selected, the effect magnitudes recovered and the functional associations retained.

This evaluates the consequences of replacing a measurement for the research outcome, beyond agreement between predicted and measured values. Screening, effect quantification and functional interpretation are assessed separately within a common framework, allowing the same prediction to be evaluated according to the uses it actually supports.

We apply this framework to an optical CRISPR perturbation atlas comprising 1,000 gene knockouts, 52 molecular and functional targets and 73 screens in A549 cells[22]. Data linkage and quality control yielded 9,996,286 same-cell phase–fluorescence pairs, enabling comparisons of predictive performance and biological utility across targets under a common experimental design. We find that perturbation rankings can persist despite compression of differences in effect magnitude. Predictions recovered a median 79% of the 50 knockouts with the strongest measured effects across targets, yet retention of functional associations varied. These results make the differences between screening performance, effect quantification and functional interpretation directly testable properties of an inferred measurement.

We further test whether available evidence can identify suitable uses of predictions. One group of gene knockouts not used for model training informs use decisions, and a separate group, also excluded from training, tests those decisions. Successful analyses, accepted analyses that fail and useful analyses that are missed are counted separately. We also use pilot measurements in separately acquired screens to choose which direct assays to retain, then test whether these choices improve effect quantification for another group of genes. External evaluations extend this approach to different cellular contexts and measurement types. The OASIS primary-hepatocyte resource is used to test whether brightfield predictions recover compounds causing strong metabolic-activity losses[23]. Cell Health and LINCS extend the assessment to cell-health phenotypes and gene-expression responses, examining the support that morphology-based predictions provide for specific screening uses[24,25]. PERISCOPE provides a further experimental design for examining reference-measurement consistency[26]. By connecting predictive performance to the outcomes of biological analyses, MorphoSuff advances the question from whether a phenotype can be predicted to which biological investigations that prediction can support.

Results

We assessed measurement substitution at three levels: recovery of cellular phenotypes, preservation of knockout effects, and retention of the biological analyses performed with those effects. Comparing analyses based on measured and predicted values allowed us to identify supported uses and to test whether those uses could be selected before examining a separate set of perturbations.

A paired atlas links morphology to targeted cellular phenotypes

We evaluated MorphoSuff using the published A549 optical CRISPR perturbation atlas[22]. In the source study, pooled CRISPR perturbations targeting 1,000 genes were followed by phase and targeted-fluorescence imaging, and in situ sequencing identified the sgRNA carried by each cell. Linking these images and sgRNA identities connects each cell's morphology to its measured molecular or functional phenotype and the corresponding gene perturbation. We used these same-cell records to compare image-predicted phenotypes with direct measurements and to derive knockout responses from each (Fig. 1a).

Our summary of the analysis set[22] identified 7,344,374 distinct phase-imaged cells and 9,996,286 phase–fluorescence observations; measurements of multiple targets in the same cell contribute separate paired observations. The 52 molecular and functional targets were measured using fluorescent reporters. For each cell, we used 172 existing phase-image features, including cell area and aspect ratio, paired with 20–72 fluorescence measurements per target, such as mean and integrated intensity and spatial displacement. Together, these targets provided 1,604 measurement endpoints spanning ten biological systems, including lysosomes, mitochondria, the nucleus and signalling (Fig. 1b,c).

These measurements were distributed across 73 screens, forming 99 reporter–screen assays, with one to seven reporters per screen (Fig. 1c). Unlike resources organized around a shared morphological panel[26,27], each reporter here had its own fluorescence measurements and was observed in a subset of screens. We therefore evaluated predictions by both target and screen, separating differences in the predictability of biological phenotypes from variation across acquisition settings. The 34 reporters measured in multiple screens further allowed us to test whether recovery of the same target persisted in a separately acquired screen.

To connect this atlas structure to the biological effects being evaluated, we summarized the control-relative responses of all 1,000 knockouts for each reporter (Fig. 1d). Within-reporter response percentiles show which knockouts produce relatively strong or weak changes in each target. The FeRhoNox–ALG12, LAMP1–BORCS7 and pRb–FECH examples then link same-cell phase–fluorescence images to feature-level response maps across the knockout panel (Fig. 1e). These maps relate visible cellular morphology to the targeted response profiles used in our subsequent comparisons of perturbation ranking, effect magnitude and functional associations.

Morphology captures diverse perturbation responses across targets and experimental settings

We compared ten methods for predicting fluorescent phenotypes from phase morphology, covering models fitted separately for each target, models trained jointly across targets and methods for predicting one measurement modality from another. To compare these model families under common conditions, all received the same phase features, target measurements and predefined data partitions.

The evaluation covered three experimental settings: new imaging fields containing perturbations represented during training (Field), gene knockouts not used in training (Gene), and new screening experiments acquired separately from the training screens (Whole screen; Fig. 2a). Cell-level Pearson correlations assessed whether predictions reproduced individual-cell phenotypes; knockout-level correlations assessed whether they reproduced the changes caused by gene knockout relative to non-targeting controls. We compared predictors, individual screens and reporter–screen assays, then examined recovery across targets to determine where recovery varied within the atlas.

The ten methods differed in their recovery of both cellular phenotypes and knockout responses (Fig. 2b; Supplementary Figure 1a). In the Gene setting, median knockout-response correlations across 52 targets were 0.760–0.768 for MLP, GBDT, ResMLP and TabM, and 0.746–0.756 for scPair, MultiTab and CatBoost; the corresponding values were 0.736 for Ridge, 0.691 for scButterfly and 0.614 for MIDAS. Cell-level medians ranged from 0.622 to 0.760. MLP and GBDT had the highest observed medians for both metrics in the Field and Gene settings. In Whole screen evaluation, which included the 34 targets measured in multiple screens, the highest medians instead came from Ridge and ResMLP: 0.730 and 0.729 for knockout responses, and 0.621 and 0.617 for cellular phenotypes. Thus, the methods with the highest median performance differed between prediction within the sampled screens and transfer to separately acquired screens.

We next asked how recovery varied across the screening experiments (Fig. 2c; Supplementary Figure 1b). For each method, we averaged the scores of the reporter–screen assays within a screen, then summarized that screen by the median across methods. For knockout responses, the interquartile range across screens was 0.175–0.256 across the three settings, whereas the median within-screen interquartile range across methods was 0.027–0.028. The screen summaries capture both the composition of the measured targets and their experimental context, prompting a closer comparison of the same target across screens.

At this finer level, each reporter–screen assay was represented by its median score across the ten methods (Fig. 2d; Supplementary Figure 1c). In the Gene setting, FeRhoNox knockout-response recovery was 0.900 in OPS0047 but 0.418 in OPS0067, whereas LysoTracker remained between 0.912 and 0.939 across three screens. Some targets therefore retained high recovery across the sampled screens, while others showed substantial differences even when reporter identity was unchanged. Recovery depended on the target–screen combination as well as on predictor choice.

We then examined which molecular and functional targets could be predicted for gene knockouts not used in training (Fig. 2e; Supplementary Figure 2a). Using the median score across methods, response recovery reached a Pearson correlation of at least 0.60 for 44 of 52 targets in gene tests and for 22 of 34 targets assessed across screens. Mean gene-test recovery ranged from 0.877 for endosomal targets to 0.620 for signalling targets. Recovery also differed between types of fluorescence measurement: intensity and intensity-range measurements were more predictable (0.822 and 0.791) than spatial displacement and sub-object distribution (0.541 and 0.525; Fig. 2f). Thus, the assessed phase features captured some aspects of the targeted phenotypes more readily than others.

To understand whether a target's behaviour in the screen could anticipate its predictability, we asked how clearly its measurements distinguished gene knockouts and how closely this pattern resembled that obtained from phase morphology (Fig. 2g). Targets that better distinguished knockouts were modestly more predictable ($\rho = 0.344$, 95% interval 0.072–0.579, $n = 52$; $\rho = 0.305$ after capping one sampled set at 300 cells per guide). However, similarity between the phase and fluorescence

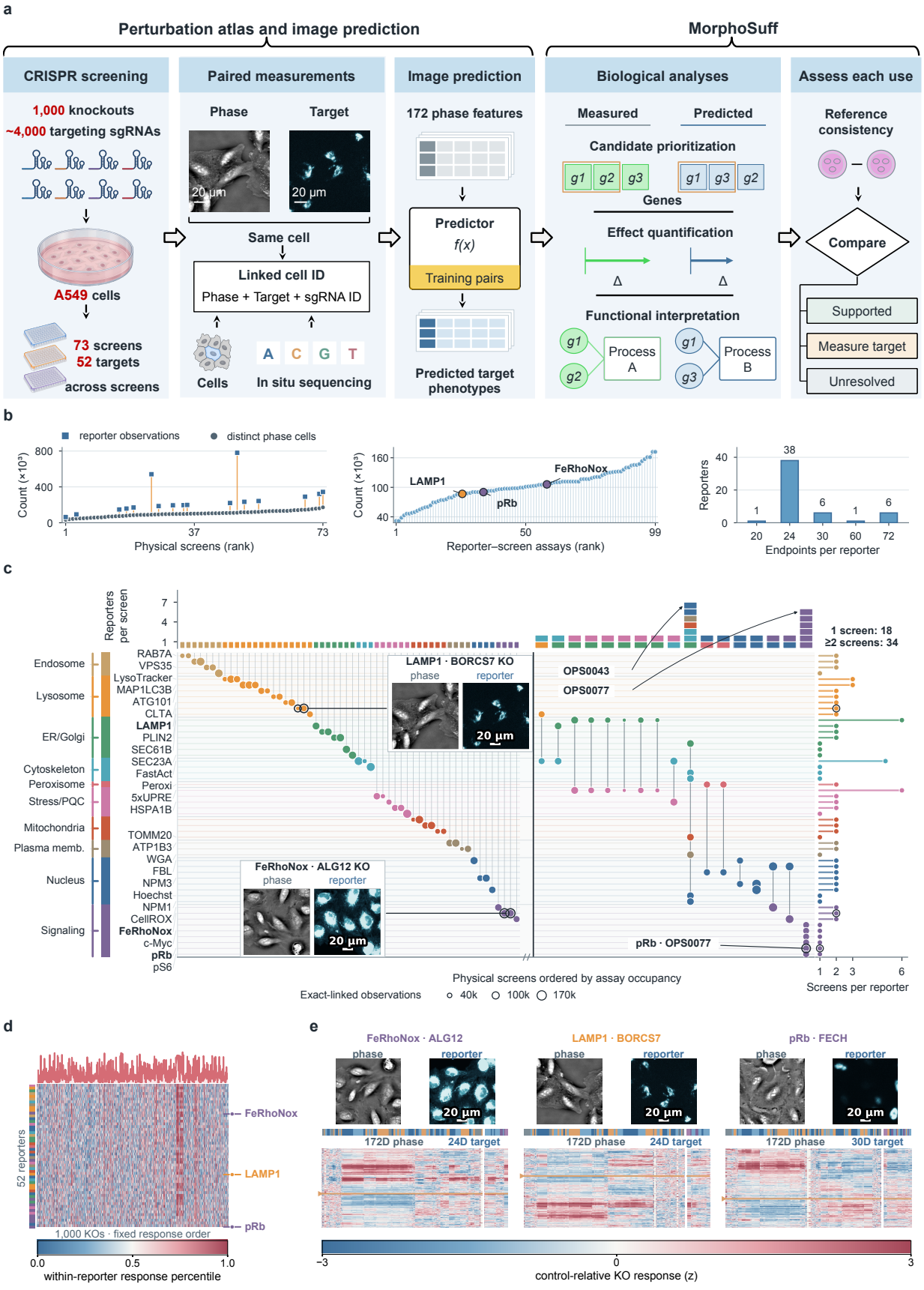

Figure 1: **A paired perturbation atlas provides the reference for measurement substitution.** **a,** Same-cell phase, targeted fluorescence and sgRNA measurements from the source atlas support prediction and the MorphoSuff comparison of biological uses. Measured and predicted phenotypes are analysed for candidate prioritization, effect quantification and functional interpretation, with reference consistency informing use assessment: $g1$–$g3$, example genes; orange outlines, selected candidates; Δ and arrow length, effect magnitude; Processes A/B, functional associations; scale bars, 20 μm. **b,** Analysis-set composition across 73 screens and 99 reporter–screen assays: 7,344,374 distinct phase cells, 9,996,286 paired reporter observations and the distribution of 1,604 endpoints among 52 reporters. **c,** Reporter coverage across screens; dot size indicates paired observations, and marginal tracks show reporters per screen and screens per reporter. Blanks denote unmeasured combinations. **d,** Control-relative response magnitudes for 52 reporters and 1,000 knockouts, expressed as within-reporter percentiles; reporters are ordered by recoverability. **e,** Same-cell phase–fluorescence images and feature-level knockout-response maps for FeRhoNox, LAMP1 and pRb. Heatmaps show signed, control-relative z-scores; image contrast is reporter-specific; scale bars, 20 μm.

patterns had no detectable association with prediction performance ($\rho = 0.090$, -0.174 to 0.345). A measurement's ability to distinguish perturbations therefore did not by itself establish how well morphology could reproduce it.

The image representation and the amount of paired training data also affected recovery. In a separate analysis of 12 targets, increasing the available paired measurements from 0.1% to 1% improved mean recovery for all ten methods; increasing the fraction to 20% gave further improvement for nine (Supplementary Figure 3a). Across six targets, engineered features reached a mean gene-test recovery of 0.739, compared with 0.594 for the pretrained Cytoland phase-image encoder[14] and 0.613 after partial fine-tuning (engineered minus fine-tuned 0.127, target-level hierarchical 95% interval 0.066–0.188; Supplementary Figure 3b).

The target-level recoverability analyses above use the median of the ten methods' scores. To examine response fidelity and measurement replacement, we instead evaluate a specific ensemble predictor: the element-wise median of the ten methods' aligned predictions for observations not used in their training. This produces one predicted response for each measured response, allowing the same biological analyses to be compared using predictions and measurements.

Predicted perturbation effects support distinct biological uses

To distinguish the response properties needed for different biological uses, we applied the atlas criteria. Combining recovery, response fidelity and reference reliability, the atlas criteria classified the ensemble predictions for 30 targets as quantitative proxies, 7 as ranking proxies, 3 as requiring direct measurement, 9 as not identifiable and 3 as unresolved (Fig. 3a; Supplementary Figure 4a). Quantitative proxies met the specified criteria for correlation, effect-magnitude spread and recovery of strong effects; ranking proxies supported ordering without meeting all quantitative criteria. Low or unavailable reference reliability prevented a substitution decision, whereas missing or intermediate fidelity evidence left a target unresolved. Across 36 threshold variations, 40 targets retained their proxy/nonproxy status and 28 retained their exact category (Supplementary Table 2).

We next asked whether recovering the order of knockout effects also preserved the differences in their sizes. Among the 52 molecular and functional targets, the median rank correlation was 0.825, whereas the median predicted-to-measured variance ratio was 0.779. Eleven targets retained less than half the measured variance in response magnitude; ten of these still had rank correlations above 0.5 (Fig. 3b). Predictions could therefore retain information about which knockouts had stronger effects while making those effects appear closer together. A useful ordering did not establish that the quantitative differences between gene effects had also been recovered.

To determine what these differences meant for biological analysis, we replaced measured knockout responses with ensemble predictions and repeated two analyses. We first selected the 50 knockouts with the strongest effects, then examined the functional associations of the selected genes using Gene Ontology biological processes and cellular components[28,29] and protein complexes from Complex Portal[30]. Predictions retained a median 79% of the strongest hits across targets. The ten leading functional terms were less uniformly retained: median Jaccard overlap was 0.603 for biological processes and 0.667 for both cellular components and protein complexes (Fig. 3c,d). Recovery of strong-effect genes and agreement in their functional interpretation were therefore related but distinct outcomes.

Because a leading term need not be statistically supported, we next asked whether predictions preserved the significant functional associations found using measurements. Significant biological-process terms were present for 33 of the 52 targets, using BH-adjusted $P \leq 0.05$ against the 1,000 assayed genes. Among the 30 quantitative proxies, 21 had such measured evidence, and all 21 retained at least one of those terms after prediction replaced measurement (Fig. 3c,d). Conversely, the six targets with no overlap between their top-ten biological-process lists had no significant terms in either the measured or predicted lists at the primary hit budget. Differences between leading annotations therefore did not always indicate the loss of an association supported by the measured screen.

VPS35 illustrates how the selection budget affects this comparison. Its top-ten biological-process lists did not overlap when the strongest 5% of knockouts were selected, but neither list contained a significant term. At the fixed 10% and 20% budgets, predictions retained four of five and five of eight measured significant terms, respectively (Supplementary Data). Ranked-list agreement and significant-term retention thus describe different aspects of functional analysis, alongside the recov-

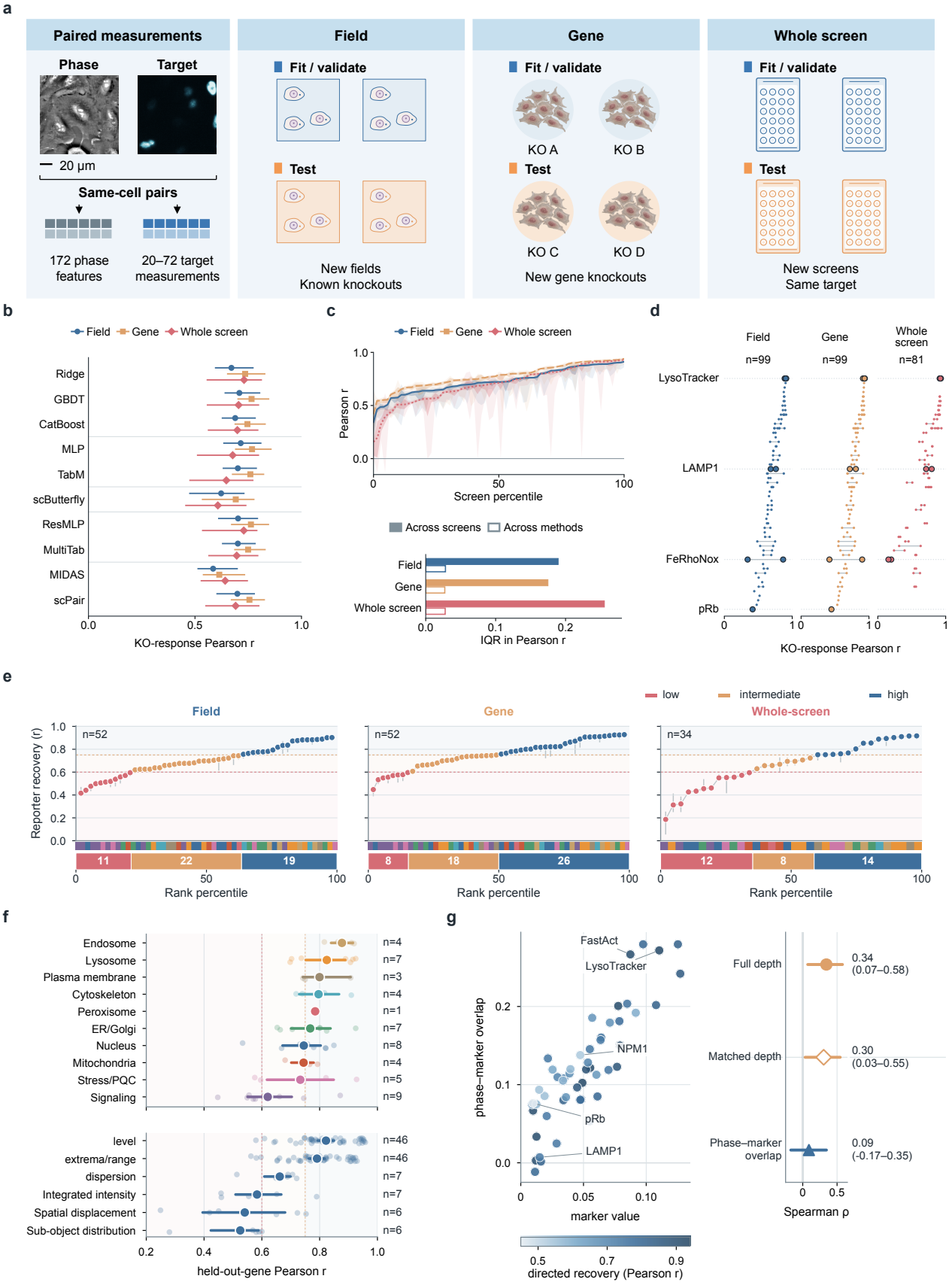

Figure 2: **Morphology captures diverse targeted responses across predictors and experimental settings. a,** Prediction on new fields, genes not used in training and separately acquired screens; blue, fitting/validation; orange, test; scale bar, 20 μm. **b,** Ten-method knockout-response medians and interquartile ranges (IQRs), using the targets in **e**. **c,** Screens ordered within each setting by median method score; bands show method IQRs and full ranges. Lower bars compare across-screen IQR with median within-screen, across-method IQR. **d,** Reporter–screen knockout scores summarized by median method scores; points retain all 99, 99 and 81 assays, with four targets highlighted. **e,** Target recoverability across field, gene and screen tests ($n = 52, 52, 34$), summarized by median method scores; bands mark descriptive strata ($r < 0.60$, $0.60 \leq r < 0.75$, $r \geq 0.75$). **f,** Recovery by biological system and endpoint family: reporter values, means and reporter-bootstrap 95% intervals (10,000 resamples). **g,** Perturbation distinctiveness and phase–reporter overlap versus gene-test recovery ($n = 52$); intervals use 3,000 reporter resamples. Matched-budget values use one draw capped at 300 cells per sgRNA. Complete knockout and cell-level comparisons accompany Supplementary Figure 1.

ery of the strong-effect genes themselves.

The supported use also depended on the predictor. With reliability and thresholds fixed, 20 of the 43 targets with reproducible measured responses supported proxy use under every predictor, 6 under none and 17 changed between proxy and nonproxy use (Supplementary Figure 5a). Fifteen of the consistently supported targets changed between quantitative and ranking categories. For example, LAMP1 supported quantitative comparison under MLP, ranking under MIDAS and neither proxy use under scButterfly (Fig. 3e). Model choice therefore affected the analyses that predictions could support, not just their correlation with measurements.

To test whether such evidence could guide screening, we used one group of gene knockouts to choose which targets' predictions to accept, then tested those choices on a separate group. Neither group had been used to train or select the predictor. We repeated this within each of the five original MLP test partitions, giving 260 target–partition analyses. For each accepted analysis, predictions selected the strongest 20% of knockout responses in the evaluation group; success required recovering at least 60% of the strongest responses identified by measurement (Fig. 3f). We compared acceptance based on prediction correlation alone, response ranking alone, or the combined evidence for recovery, ranking, magnitude spread and strong-hit retention.

When every prediction was accepted, 32 of the 260 analyses failed this screening requirement (12.3%). Acceptance based on correlation alone reduced the failure fraction to 4.2% (8 of 192 accepted analyses); ranking-based acceptance reduced it to 3.3% (7 of 211). The ranking rule retained 204 of the 228 analyses that could succeed, or 89.5%: 20 more successful analyses than correlation alone, with one fewer accepted failure. The gain in successful uses was 7.69 percentage points across all cases (reporter-cluster 95% interval 1.54–15.00), and gains were also positive for Ridge, scButterfly and the ensemble (Fig. 3g). The difference in failure fraction between the ranking and correlation rules was less certain: -0.85 percentage points (95% interval -5.14 to 3.06) at fixed thresholds and -2.05 points (-5.85 to 1.14) when both accepted 75% of cases (Supplementary Figure 6a; Supplementary Data).

Combining all response criteria yielded the most selective set: one failure among 171 accepted analyses (0.6%). This rule accepted 65.8% of cases, compared with 81.2% for ranking alone. The comparison identifies a practical choice: ranking evidence retained more successful screening uses, whereas the combined criteria accepted fewer uses and had the lowest observed failure fraction.

Same-cell morphology adds predictive information, but response ranking and magnitude can diverge

We asked whether phase morphology captured the targeted state of an individual cell, beyond the average effect of its gene knockout. For 52 reporters, we repeated the MLP analysis across five gene partitions while changing only the pairing of training measurements or the phase features supplied (Fig. 4a). Reassigning fluorescent measurements among cells with the same knockout and screen preserved the group's response distribution but removed the link to the image of the same cell; test measurements retained their original pairing. With exact pairing, median cell and knockout correlations were 0.757 and 0.769, respectively, compared with 0.291 and 0.293 after reassignment. Restricting reassignment to cells matched for area, density and eccentricity gave correlations of 0.531 and 0.502. Prediction from size–shape features alone gave 0.261 and 0.329, whereas removing those features left performance essentially unchanged. Individual-cell morphology therefore contributed information beyond average knockout responses, cell size, shape and density.

Prediction of within-perturbation variation provided complementary evidence that morphology captured differences among cells receiving the same knockout (Supplementary Figure 7a).

To determine whether this information preserved a specific biological response, we examined MTOR and RPTOR, components of the mTOR signalling system that regulates lysosome homeostasis[31]. Both knockouts increased LysoTracker mean intensity in three separately acquired screens, and predictions retained these increases (Fig. 4b). For MTOR, measured effects were 0.702, 0.939 and 1.076 control standard deviations, compared with predictions of 0.667, 0.974 and 1.110; the largest difference was 0.035. Each prediction lay within the corresponding range of acquisition-well measurements. The broader MTOR-related gene comparison also showed distinct LAMP1 responses (Fig. 4b). These results support using predictions to compare this lysosomal-dye response under the evaluated conditions.

To examine the separation between response ranking and magnitude in individual targets, we compared the predicted and measured effects of their knockouts. These are distinct comparisons:

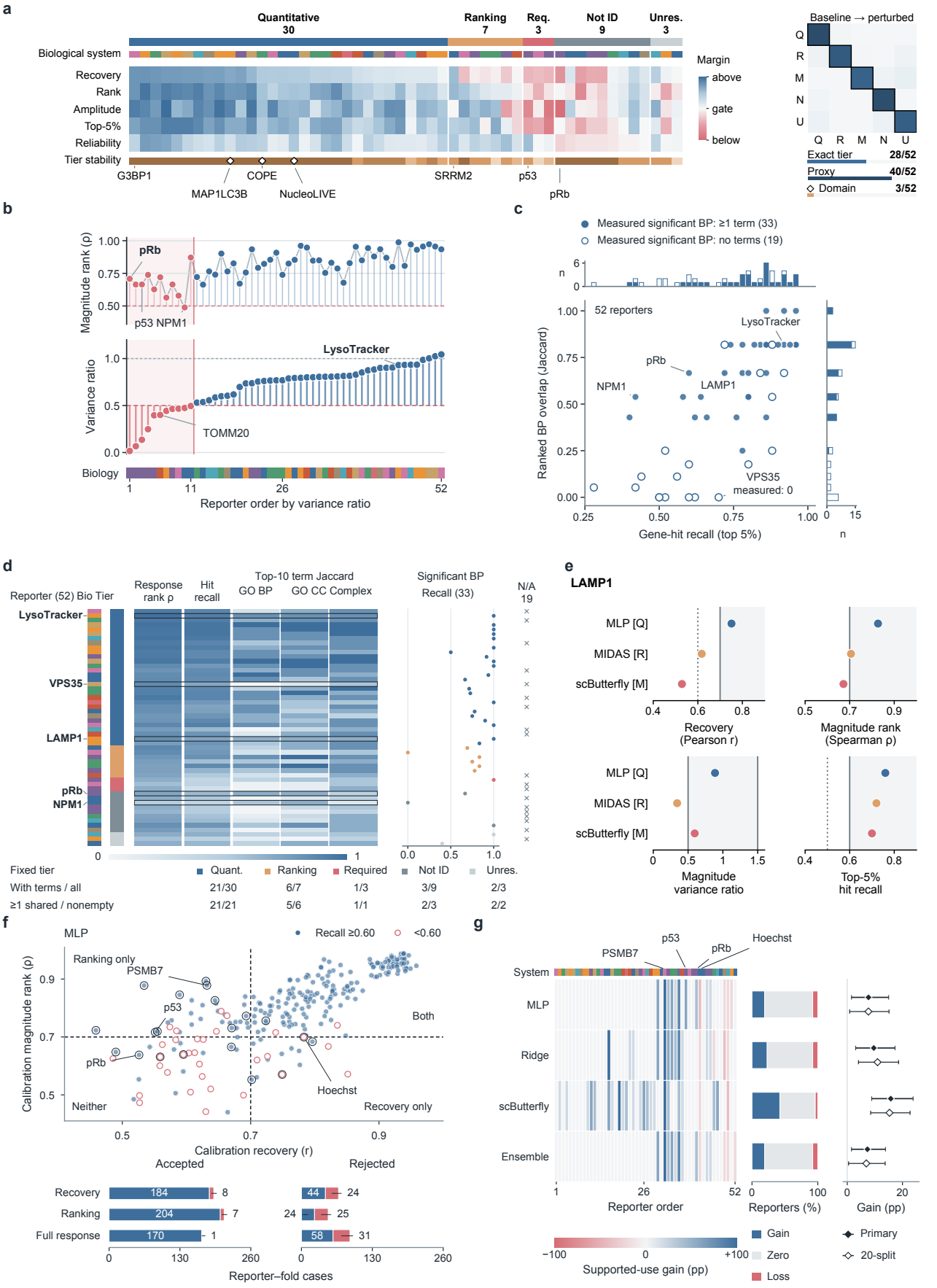

Figure 3: **Screening, effect quantification and functional interpretation require distinct evidence.** **a,** Five criterion margins and ensemble tiers for 52 targets, with threshold-sensitivity and domain summaries. **b,** Magnitude rank and variance ratio; red marks ratios below 0.5. **c,** Strong-hit recall versus top-ten biological-process overlap; filled/open points denote nonempty/empty measured significant sets. **d,** Functional retention across 52 targets; significant-term recall is undefined for 19 empty measured sets (BH-adjusted $P \leq 0.05$). **e,** LAMP1 under three predictors; shading marks quantitative criteria, dotted boundaries recovery 0.60 and hit recall 0.50. **f,** Calibration metrics and evaluation outcomes for 260 MLP target–fold cases; success requires top-20% hit recall ≥ 0.60. **g,** Rank-minus-recovery gain in supported successes: all-target matrix, primary and 20-partition mean estimates, and 95% intervals from 2,000 target-cluster bootstrap samples.

ranking asks which knockouts cause stronger responses, whereas effect-size comparison asks how far apart those responses are. We assessed the first using the rank correlation of predicted and measured response magnitudes, and the second using their variance ratio, for which a value of one indicates equal spread. LysoTracker predictions preserved both properties, with a rank correlation of 0.988 and a variance ratio of 0.930. ChromaLIVE 488 gave values of 0.828 and 0.816. By contrast, pRb retained a rank correlation of 0.708 while its predicted magnitudes occupied a much narrower range, with a variance ratio of only 0.017 (Fig. 4c). The predicted effects could therefore retain their relative order while appearing much more similar in size than the measurements showed. This distinction matters when moving from prioritizing knockouts to comparing the sizes of their effects.

We next asked whether response compression could remove a biologically relevant candidate from a screen. Disruption of ribosome biogenesis can increase p53 abundance[32]. In the A549 data, DDX18 knockout increased p53 and p21 mean intensity by 1.03 and 1.79 control standard deviations, whereas the ensemble predicted only 0.17 and 0.45 (Fig. 4d). These measurements came from the same 52 cells distributed across three wells and four sgRNAs. All three well means increased for both targets, and the guide-level comparisons supported the response, although their magnitudes varied (Fig. 4d,e; Supplementary Figure 8b). The multivariate knockout-response rank fell from 10 to 212 for p53 and from 17 to 114 for p21, excluding DDX18 from both predicted top-50 lists. All ten individual predictors also missed these two candidates (Supplementary Figure 8a; Supplementary Data).

The omitted candidates changed a specific functional association. Of the seven screened genes annotated to large-subunit ribosomal RNA maturation, five—DDX18, EIF6, NSA2, RPF2 and WDR12—were among the strongest 50 measured p21 responses, but none was in the predicted top 50 (Fig. 4e). The pooled measured list supported this association (BH-adjusted $P = 0.00087$), whereas the predicted list did not; repeat-specific candidate membership is shown separately in Supplementary Figure 8c. For a screen seeking these molecular responses, the predicted candidate list would therefore miss changes supported by targeted p53/p21 measurements. Cross-screen response curves and complete predictor distributions are retained in Supplementary Figure 7c.

Reference stability and experimental context explain differences in measurement use

To understand why predictions supported different uses, we examined the stability of measured knockout effects, changes between screens and variation unresolved by morphology. Gene-test recovery correlated with within-screen reliability ($\rho = 0.547$, 95% interval 0.295–0.737), cross-screen consistency (0.349, 0.014–0.625) and coverage (0.482, 0.211–0.686; Fig. 5a). Agreement between guide subsets had a similar association but an interval including zero (0.376, -0.002 to 0.668). Reliability, guide consistency and log coverage jointly accounted for 38.7% of in-sample recovery variation among the 35 targets with complete measurements (permutation $P = 0.001$).

For pRb and NPM1, the measured responses themselves showed low repeat agreement (Fig. 5b). Median agreement across well pairs, guide halves and random cell halves was -0.015, -0.020 and 0.158 for pRb, and 0.069, 0.054 and 0.197 for NPM1. Random-cell splits assessed sampling stability within the observed population, whereas well and guide comparisons also included variation between acquisitions and perturbation reagents. The lower agreement across wells and guide subsets showed why cell sampling alone was an insufficient reference for these targets. More reproducible measured responses are needed to separate prediction error from variation in the experimental effect.

Changes in control cells were associated with loss of prediction accuracy in a new screen. We assessed 81 transfer directions; correlations between control-cell shifts and transfer loss, calculated from the 34 reporter-level summaries, were 0.765 for phase and 0.837 for targeted measurements (both permutation $P = 1/3{,}001$; Fig. 5d). Control cells carry no targeted knockout, so these shifts indicate differences in acquisition or cellular context. The association with phase shifts identifies a potential warning signal that can be observed without acquiring the corresponding targeted measurement.

Even cells with similar morphology retained differences in targeted state. The discrepancy between morphology-nearest cells, relative to reference cell pairs, was 0.681 and 0.849 in two FeRhoNox contexts and 0.889 for pRb (Fig. 5e). A ratio near one means that choosing a morphologically similar neighbour removes little of the difference in targeted state. Comparisons based on raw phase thumbnails gave ratios near or above one. Such remaining variation can contribute to averaging and response compression. Recovery and magnitude ranking across separately acquired screens are sum-

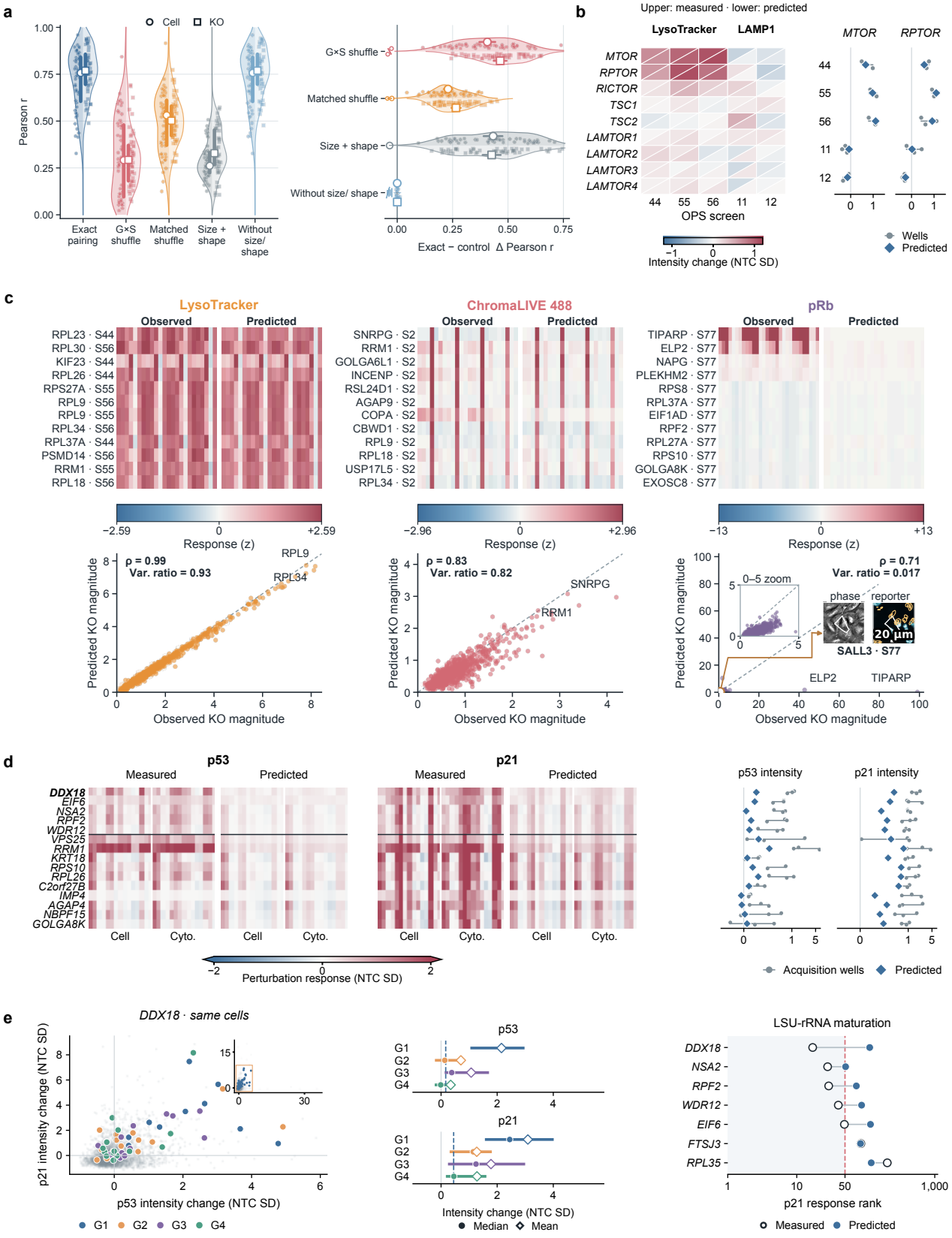

Figure 4: **Phase morphology carries same-cell information, while response ranking and magnitude can diverge. a,** Reporter-specific MLP recovery under exact pairing and four information interventions ($n = 52$, five held-out-gene partitions); G×S denotes gene–screen shuffling and matched denotes covariate-matched shuffling. Circles/squares denote cell/knockout scores, pale lines pair reporters, and open effect symbols show means with reporter-bootstrap 95% intervals. **b,** Nine MTOR-related genes across three LysoTracker and two LAMP1 screens; upper/lower triangles, measured/predicted intensity changes. Right: wells (grey) and predictions (blue). **c,** Observed versus median-of-ten-method predictions on a common reporter-wide control reference: 12 strongest held-out gene–screen profiles per heatmap and all 1,000 genes per full-range scatterplot; the pRb zoom shows 0–5 on both axes. pRb is classified as not identifiable (Fig. 3a). Screen labels use the source OPS numbering; S77 denotes OPS0077. The microscopy inset identifies an exact SALL3 cell from OPS0077 (S77); scale bar, 20 μm. **d,** Thirty-endpoint p53/p21 profiles for 15 genes; right, well intensities and predictions (symmetric-log axes, linear within ± 0.5). **e,** DDX18 paired cells (52 knockout, 4,104 NTC), four sgRNA distributions (median, mean and IQR; dashed, predicted mean), and all seven LSU-rRNA maturation genes' p21 ranks; dashed boundary, top 50.

marized in Supplementary Figure 6b. Reference instability, changes between screens and variation unresolved by the image representation therefore point to different ways of improving the inferred measurement.

We also tested screening-use decisions in 81 separately acquired target–screen contexts and in 405 original model tasks that excluded both the destination screen and evaluation genes from training and model selection (Supplementary Figure 9a). Both settings retained successful applications while exposing accepted failures and missed uses; matching the number of accepted cases reduced the difference between the ranking and correlation rules. The full decisions are provided in Supplementary Data.

Pilot measurements guide which targeted assays to retain

To test whether repeat measurements could guide experimental choices, we asked which targeted assays were worth retaining when the remainder were replaced by image predictions. We evaluated 75 target–screen combinations with three eligible wells, spanning 34 targets and 63 physical screens. The prediction and direct-measurement options used exactly the same cells and controls in two wells; the third well supplied the measured reference. On one group of genes, we compared the quantitative error of predictions with that of repeated direct measurements, then selected the 15 combinations with the largest estimated benefit of retaining the assay (Fig. 5c). Errors included all original phenotype measurements for the strongest 20% of measured gene responses. The other combinations used predictions. We compared this repeat-informed choice with retaining the same number of assays based on low predictive correlation or high prediction error alone. The complementary downstream-retention regression is retained in Supplementary Figure 7b.

On a separate group of genes, repeat-informed assay choices reduced mean normalized squared error from 0.8124 under correlation-based selection to 0.7496, a 7.7% reduction; selection by prediction error alone gave 0.7721 (Fig. 5f). The paired difference from correlation-based selection was -0.0628 (target-cluster 95% interval -0.1318 to -0.0103), and all 15 retained assays lowered evaluation error, compared with 11 of 15 selected by correlation and 13 of 15 by prediction error.

A 50-gene pilot also supported useful choices: error was 0.7598 versus 0.8195 under correlation-based selection, a 7.3% reduction, with 14 of 15 retained assays improving quantification. The small-pilot difference from prediction-error selection was less certain, with an interval including zero. Alternative gene allocations, reference-well rotations and the repeat-consistent subset are reported in Supplementary Figure 9b,c. This comparison turns repeat measurements into evidence for deciding where direct assays add the most quantitative information.

External datasets connect morphology-based prediction to specific screening uses

We evaluated four complementary public resources that differ in cellular context, measurement type and pairing design (Fig. 6). OASIS paired brightfield images with biochemical measurements in primary hepatocytes[23]. Cell Health paired fluorescent morphology with separately acquired cell-health measurements[24], and LINCS paired morphology and RNA measurements from matched compound conditions[25]. PERISCOPE supplied measurements from fluorescent channels for evaluating reference consistency[26].

In primary hepatocytes, we tested whether brightfield predictions could identify compounds causing the strongest metabolic-activity losses. Using the brightfield DINO features released with OASIS[23], Ridge was fitted on 868 development compounds and evaluated on 217 different compounds, keeping all doses and production sources of each compound together. The ranking score was the mean loss of RealTime-Glo metabolic-activity signal relative to same-plate DMSO, averaged over the assayed concentrations and retaining the direction of change. Selecting the predicted top 5% recovered 10 of the 11 largest measured losses (90.9%; compound-bootstrap interval 54.5–100%; Fig. 6a,b). Mean measured loss was 0.334 in the predicted selection and 0.338 in the measured top eleven, showing that the predicted selection retained the strongest metabolic decreases.

Brightfield predictions recovered 10/11 with both Ridge and histogram gradient boosting (HGB), compared with 8/11 and 7/11 for matched predictors using fluorescence-derived cell counts, and 0/11 and 1/11 for acquisition metadata alone (Supplementary Data). The brightfield-Ridge recovery was 18.2 percentage points higher than its matched cell-count comparator. Disrupting within-plate prediction–compound correspondence reduced mean brightfield-Ridge recall to 5.38%.

To ask whether this success also supported another biochemical use of the same images, we predicted LDH release, an indicator of membrane damage, using the same development/evaluation

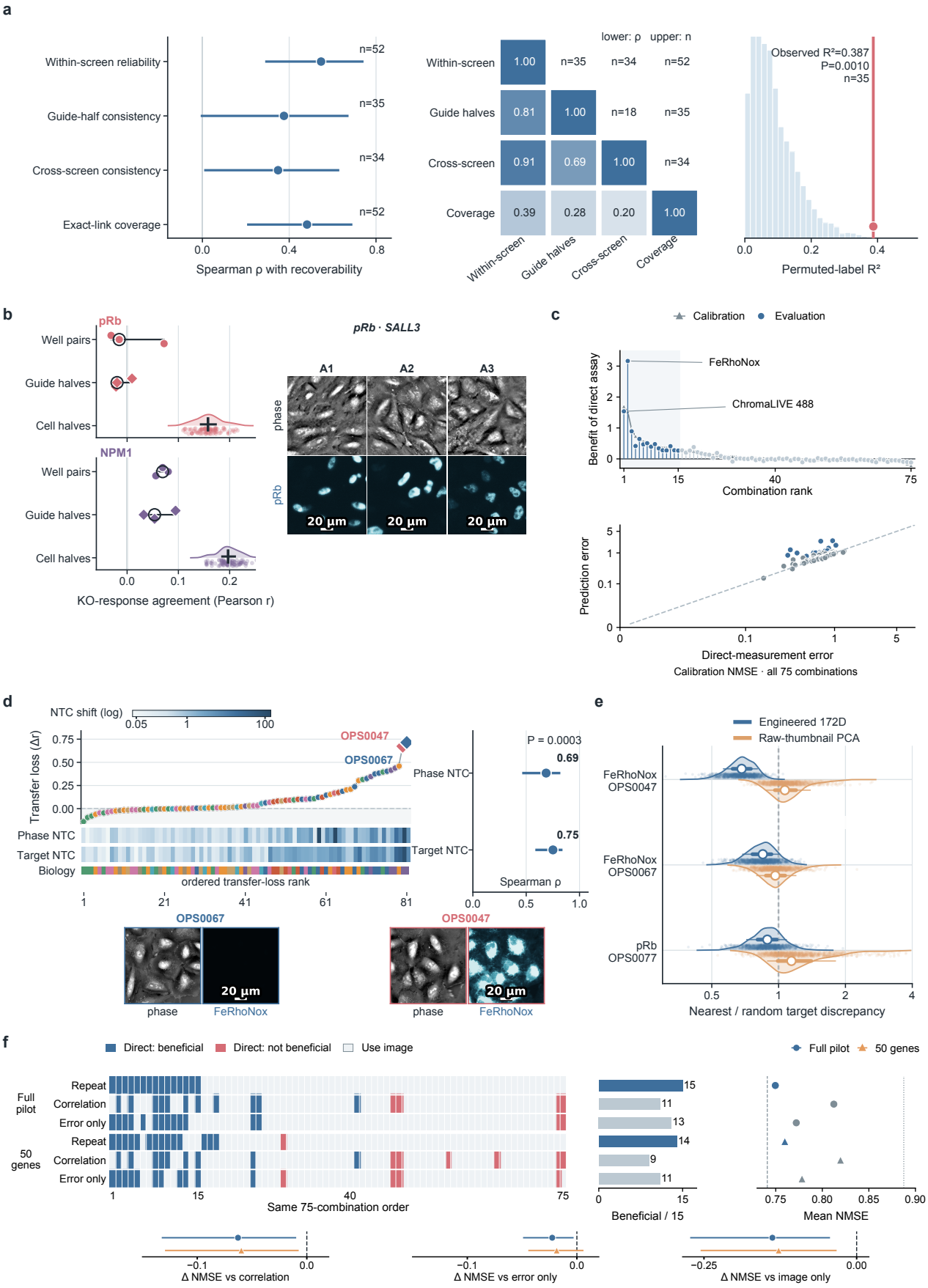

Figure 5: **Reference stability and experimental context explain differences in measurement use. a,** Recoverability versus reference consistency and coverage; reporter-bootstrap 95% intervals. **b,** pRb and NPM1 across three well pairs, three guide splits and 100 random cell splits, with paired phase–pRb examples. **c,** Calibration/evaluation direct-assay benefit across 75 equal-cell combinations. Lower: calibration prediction versus direct-measurement error (symmetric-log axes, linear within 0.05). Blue marks 15 retained assays. **d,** Control-cell shifts versus transfer loss: 81 directions (displayed correlations), 34 target medians (inference); diamonds mark FeRhoNox examples. **e,** Nearest/reference target discrepancy for engineered features and raw-phase principal components; bars, medians and interquartile ranges. Scale bars, 20 μm. **f,** Assay choices, beneficial counts and normalized mean squared error (NMSE) for full/50-gene pilots. Lower: paired differences with target-bootstrap 95% intervals (34 targets, 5,000 resamples). Dotted/dashed: all-image/all-direct error.

compounds and fixed model settings. Brightfield Ridge recovered six of the eleven strongest measured LDH responses, despite a response correlation of 0.826; the matched cell-count predictor also recovered six (Fig. 6a; Supplementary Figure 9d). Abemaciclib illustrated retained metabolic and LDH responses, whereas Fesoterodine showed preserved metabolic loss but underestimated LDH release across both production sources (Fig. 6b). Its source-averaged measured LDH changes were 0.143 and 0.095, compared with predictions of 0.073 and 0.054. Metabolic and LDH prioritization both passed the development-set recall requirement, but only the metabolic task met it in the 217 evaluation compounds. Thus, evidence for prioritizing metabolic injury did not establish that the membrane-damage assay could also be replaced.

Repeated measurements provided an additional comparison for the same selection task. Among 94 compounds with complete dose-matched measurements in two production sources, predictions from one source recovered four of the five strongest measured losses in the other, in both directions; selection based on the repeated biochemical measurement also recovered four of five (Supplementary Figure 6c). At fixed selection budgets of 10% and 20%, predictions retained 17/22 and 34/44 hits, respectively (Supplementary Figure 6d). The hepatocyte analysis thus supported prioritization of strong metabolic losses across several selection budgets and repeated measurements.

A complementary assessment used the CellProfiler brightfield features supplied by OASIS[23] to evaluate reconstruction of the broader biochemical response. Metabolic-activity and lactate dehydrogenase (LDH) release predictions reached recovery of 0.83 and 0.67 and top-5% recall of 0.727, but magnitude ranks of 0.604 and 0.633 left both endpoints unresolved under the full response tiers (Supplementary Table 1). The focused prioritization analysis used a different image representation, signed response score and aggregation. Evaluating the intended selection established a concrete use without requiring every property of the biochemical response to be recovered.

Cell Health tested prediction of ten cell-health measures in ES2 and HCC44 cells, giving 20 endpoint–context cases with 25 training, 12 calibration and 12 evaluation genes. In ES2, S-phase prediction retained three of the four strongest responses at $r = 0.604$, whereas late-apoptosis prediction retained two of four at $r = 0.950$ (Fig. 6c). Random selection at this four-of-twelve budget has expected recall of 1/3. The calibration task rule accepted five cases: three succeeded on evaluation genes and two failed, while seven other successful cases were missed and eight unsuccessful cases were rejected (Fig. 6d). Adding plate-repeat agreement left the accepted set unchanged. For HCC44 live-cell count, prediction retained three of four responses against the plate consensus but two of four against the second plate. These comparisons show how the selected genes and the measured reference, rather than correlation alone, determine the supported screening use.

LINCS extended the assessment to RNA, evaluating all 978 landmark probes on 260 compounds excluded from model training. Across the full RNA panel, calibration evidence enriched successful uses relative to accepting every prediction (Fig. 6e). Without selection, 24 of 978 Ridge and 20 of 978 MLP endpoints succeeded (2.5% and 2.0%). Task-only acceptance raised these success fractions to 35.3% and 35.7%: six successes and eleven failures among seventeen accepted Ridge endpoints, and five successes and nine failures among fourteen accepted MLP endpoints. Coverage was 1.7% and 1.4%, with 18 and 15 successful endpoints missed. Recovery-only accepted seven Ridge endpoints (five successes, two failures) and twelve MLP endpoints (four successes, eight failures). Adding repeat agreement left the primary acceptance sets unchanged, while the rank criterion accepted no endpoint. Across ten Ridge partitions, the task-plus-repeat success fraction ranged from 33.3% to 60.0%. Calibration therefore concentrated successful RNA applications in a small subset, but the balance of successes, failures and missed uses differed between predictors and selection rules.

For BRCA1, predictions retained 17 of the 26 strongest absolute responses against the consensus and against each of two separately acquired RNA references, at signed $r = 0.515$ (Fig. 6f). Sixteen of the seventeen consensus-recovered hits agreed in response direction, and thirteen recovered hits were common to both reference comparisons. This supplied a successful prioritization example at moderate overall correlation, complementing the high-correlation Cell Health example in which fewer strong responses were recovered.

PERISCOPE demonstrated the role of reference reproducibility in interpreting otherwise favourable prediction metrics. Anti-TOMM20 predictions from four fluorescent channels met the response-fidelity criteria for a quantitative proxy, including cell-level recovery of 0.760 and magnitude rank of 0.852. However, guide-partition reliability was 0.005 at the cell level and 0.047 for guide profiles, leading to a not-identifiable assignment (Supplementary Table 1). Input-permutation

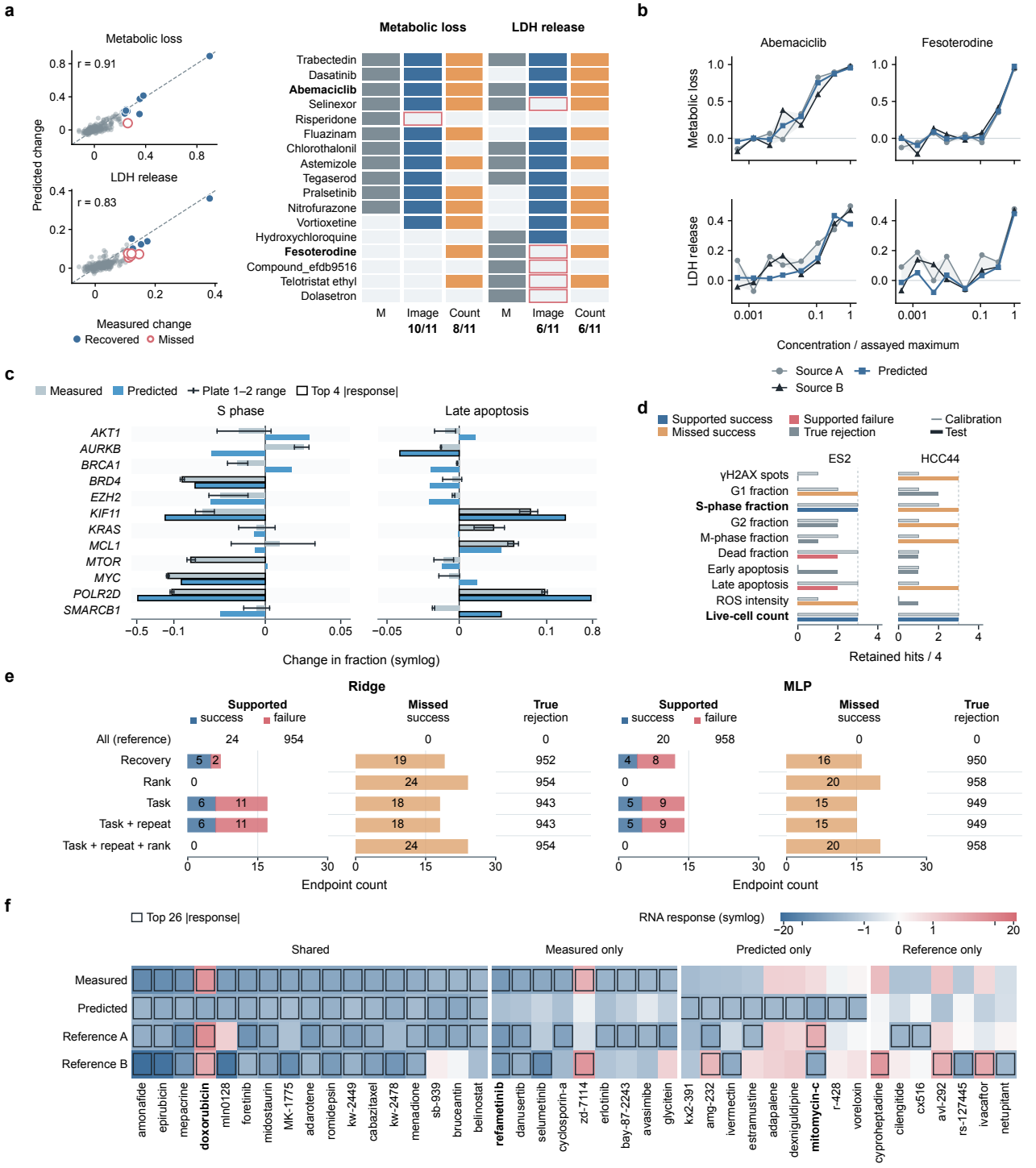

Figure 6: **Independent resources connect response prediction to specific measurement uses. a,** Left: measured versus brightfield-Ridge responses for all 217 evaluation compounds; dashed lines, identity; blue/rose points, recovered/missed measured top-eleven candidates. Right: complete 17-compound union of measured/brightfield top-eleven sets for metabolic loss and LDH release. M, measured; Image, brightfield; Count, matched cell-count predictor. Filled cells mark selection; rose outlines mark missed measured candidates; bottom counts give recovered hits. **b,** Source-resolved metabolic and LDH dose responses for Abemaciclib and Fesoterodine. Grey/black curves, production sources A/B; blue, mean prediction; shading, measured source range, not a confidence interval. Concentration is divided by each compound's assayed maximum. **c,** Cell Health responses for twelve held-out genes in ES2: measured consensus (grey bars), predictions (blue bars), two-plate ranges (whiskers) and the four largest absolute responses in each set (black outlines). S-phase and late-apoptosis fractions use symmetric-log axes, linear within ± 0.01. **d,** Cell Health task outcomes across twenty prespecified endpoint–context cases. Calibration and test retained-hit counts are shown separately out of four; colours distinguish supported success, supported failure, missed success and true rejection. **e,** LINCS endpoint decisions for Ridge and MLP across five acceptance rules plus an accept-all reference. Supported successes and failures share a 0–30 count scale with missed successes; true rejections are labelled directly. **f,** LINCS BRCA1 response map for all 42 compounds in the union of measured, predicted and two separately acquired reference top-26 sets. Colours show signed, plate-DMSO-centred RNA responses on one symmetric-log scale; outlines mark top-26 membership.

analyses provided complementary checks of predictive information (Supplementary Figure 4b). Across these resources, MorphoSuff connected prediction to a defined biological use, with direct comparisons of selected perturbations and repeated measurements determining what that use could support.

Discussion

The value of morphological prediction lies not only in its accuracy, but also in the biological analyses it can support. In the A549 perturbation atlas, the same predictions could preserve the selection of strong-effect knockouts while compressing differences in effect magnitude, and their retention of functional associations varied among targets. Screening, quantitative comparison and functional interpretation are therefore distinct capabilities, rather than automatic consequences of improved predictive accuracy. MorphoSuff brings these uses into a common framework, repeating the same analyses with measured and predicted molecular and functional phenotypes to test which experimental conclusions are preserved. This extends evaluation beyond numerical agreement to the scientific questions that predictions can help answer.

Preserving the order of perturbation effects does not necessarily preserve the differences between them. For eleven targets, the variance of predicted effect magnitudes was less than half the measured variance, yet ten retained rank correlations above 0.5. A predictor can therefore place strong-effect genes near the top of a ranking while underestimating the differences between strong and weak perturbations. The same-cell experiments showed that individual-cell morphology provided information beyond the average response to a gene knockout. However, morphologically similar cells could still differ in their molecular or functional state. Averaging over these unresolved states could narrow the range of predicted values, providing one possible explanation for the coexistence of preserved rankings and compressed magnitudes. For experimental use, this distinction separates prioritizing candidates from comparing their effect sizes: evidence for the first use does not establish the second.

Recovering most strong-effect genes does not guarantee that their functional interpretation is fully preserved. Across targets, predictions recovered a median 79% of the strongest knockouts, but functional associations also depend on which genes enter the candidate set and which biological processes they jointly support. Changes in a small number of candidates can affect whether a process has sufficient statistical support. Among the 30 quantitative proxies, 21 had significant biological-process associations in the measured data, and all 21 retained at least one of these associations in the predictions. Predictions thus preserved some functional evidence supported by direct measurements. Changes in the ordering of leading annotations nevertheless need to be distinguished from the loss of significant associations. For VPS35, neither measured nor predicted data yielded significant associations at the primary candidate-set size, whereas larger prespecified sets revealed shared significant associations. The p21 analysis showed a concrete loss: omission of five ribosomal-RNA maturation candidates removed an association supported by the pooled measured list, while the DDX18 example linked candidate loss to p53/p21 changes observed across acquisition wells. When the purpose of an experiment is to explain the biological processes affected, the functional results themselves should therefore be compared, rather than inferred to be reliable from candidate recovery alone.

Judging whether predictions can replace measurements first requires knowing whether the measured perturbation effects are reproducible. When repeat measurements differ substantially, disagreement between prediction and measurement can reflect either prediction error or variation in the experimental reference. The PERISCOPE anti-TOMM20 analysis illustrated this distinction: predictions met the response-fidelity criteria, but measured responses from different sgRNA subsets were inconsistent. Repeat measurements should therefore reflect the experimental conditions across which predictions will be used. Randomly splitting cells from the same population mainly assesses sampling stability, whereas comparisons across sgRNAs or wells also include variation between perturbation reagents and data-acquisition units. These estimates answer different questions. Choosing a repeat design that matches the intended use helps determine whether the next priority is to improve the predictor or the consistency of the measured perturbation effects.

Improvements in models and image representations should be judged by the biological analyses they make possible. In this study, predictor choice changed not only correlation scores but also the uses supported for the same target. For example, LAMP1 predictions supported quantitative

comparison, ranking alone or no proxy use, depending on the predictor. Whether a target can be predicted usefully is therefore not a fixed property independent of the model and experimental setting. A recent comparison of self-supervised learning methods on fluorescent Cell Painting images similarly found that representations learned by different mechanisms had distinct strengths across downstream tasks[33]. Model comparisons should consequently ask whether an increase in correlation makes a previously unsupported candidate-selection, effect-size or functional analysis possible. Assessing such changes connects computational improvements directly to experimental needs.

Choosing which prediction uses to accept involves balancing the retention of useful analyses against the avoidance of failures. In the gene-knockout validation, selection based on effect ranking retained 204 successful prioritization analyses, compared with 184 when selection used predictive correlation alone; the two rules accepted seven and eight failed analyses, respectively. Requiring all response criteria produced a smaller usable set, with one failure among 171 accepted analyses. These choices serve different experimental priorities. Broad initial screening may favour retaining more potentially useful predictions, whereas costly follow-up may favour a smaller set of uses supported by stricter requirements. Reporting successes, failures and missed successful uses together makes these benefits and costs explicit, rather than reducing all experiments to a single accuracy threshold.

Pilot measurements can guide which targeted assays to retain. Comparing prediction error with the variation between repeated measurements identifies where a direct assay adds quantitative information. In the equal-cell analysis, this evidence selected 15 assays that all improved quantification on a separate group of genes, whereas four of the 15 assays selected by correlation did not. A 50-gene pilot also supported useful choices. This provides a practical role for MorphoSuff in planning a subsequent screen: retain direct measurements where they add information for the intended analysis and use predictions where the pilot evidence supports substitution. The retrospective comparison establishes an effect-quantification benefit at a fixed assay allowance; prospective screens can now test the corresponding reagent and workload savings.

Morphological predictions can provide concrete screening value, but this value must be tested for the intended use. In primary hepatocytes, brightfield predictions recovered ten of the eleven compounds causing the strongest measured metabolic-activity losses among 217 compounds not used for model training. A matched predictor using cell counts derived from fluorescence images recovered eight. This provides direct evidence for prioritizing compounds with strong metabolic effects using brightfield images. LDH-release prediction recovered six of eleven, showing why evidence for the metabolic task did not establish that membrane-damage measurements could also be omitted. Cell Health and LINCS illustrated the distinction between correlation and screening performance in both directions: a highly correlated cell-health prediction could still miss the required strong-effect candidates, whereas BRCA1 prediction retained strong compound responses at moderate correlation. External evaluation can thus move beyond comparing correlation coefficients across datasets to identifying the experimental tasks that predictions support in new cellular settings and measurement types.

Future work should expand the range of cellular backgrounds and test whether learning directly from raw images enables more biological uses. Paired imaging and targeted measurements in additional cell lines would establish which prediction uses persist across cellular backgrounds and which require further training or reassessment. A complementary direction is to learn directly from raw phase or brightfield images, capturing cellular variation that the current engineered features may not adequately describe. Self-supervised learning offers several approaches to cellular image representation[34–36]. Applying these approaches to phase and brightfield images should be followed by testing whether they improve the preservation of effect sizes and functional associations, as well as predictive correlation. Ultimately, these advances should inform experimental design by identifying which targeted assays can be reduced, which should be retained, and how these choices balance assay costs, experimental workload and biological information.

Methods

Study design and paired measurements

We developed MorphoSuff and analysed the public A549 optical perturbation atlas[22]. The resource links pooled CRISPR perturbations to phase images, targeted fluorescence measurements and sgRNA identities in the same cells. The analysis covered 1,000 knockout genes, 52 molecular and

functional targets, 73 screens and 99 target–screen combinations. The linked data contained 7,344,374 distinct phase-imaged cells and 9,996,286 phase–fluorescence observations, because a cell could contribute measurements for more than one target. Each cell had 172 engineered phase features describing intensity, texture, geometry and spatial organization, paired with 20–72 fluorescence measurements per target, such as intensity and intracellular distribution. The 52 target-specific blocks contained 1,604 measurement endpoints in total.

We used the source study's cell linkage and feature definitions. Inclusion required valid same-cell linkage, an observed target block and assignment to an evaluation partition; unmeasured target–screen combinations remained missing. Feature filtering, imputation and model-input and target standardization were fitted on training cells and applied unchanged to validation and test cells. The separately defined control reference used to compare response magnitudes across partitions is described below. Targets were the primary units for atlas-wide inference, with cells, genes, screens and data partitions retained as nested observations. Preprocessing criteria and linkage records are provided in Supplementary Methods S1 and Supplementary Data.

Prediction methods and generalization tests

We compared ten methods spanning four modelling approaches. Linear and tree-based methods comprised Ridge regression[37], gradient-boosted decision trees implemented with XGBoost[38], and CatBoost[39]. Target-specific neural models comprised an MLP implemented for this study and TabM[40]. Shared neural models comprised our residual MLP (ResMLP) and MultiTab implementations, which learned across targets. Single-cell multimodal methods comprised scButterfly[41], MIDAS[42] and scPair[43]; MIDAS used shared training, whereas scButterfly and scPair were fitted separately by target. The study-specific architectures and all method adaptations, objectives and selection procedures are specified in Supplementary Table 3.

Every method was trained using the same paired phase features and targeted fluorescence measurements, with identical data partitions. At prediction time, the fitted models received phase features without the corresponding target-channel measurements. The three evaluation settings were new imaging fields containing perturbations represented during training (Field), gene knockouts absent from training (Gene), and separately acquired screens excluded from training (Whole screen). Field and Gene evaluations used five folds: fold k was tested, the next cyclic fold was used for model selection, and the other three for training. Fields were grouped by physical screen–well–tile identity and genes by knockout identity; linked observations retained the same assignment across targets. Whole-screen evaluation required a target measured in both source and destination screens and excluded the complete destination screen from training and model selection. It comprised 81 directed evaluations for 34 targets across 66 destination screens.

Benchmark scores used validation-selected model states and were averaged over folds or eligible destination screens within each target. The recoverability landscape summarized the median of ten method scores. Response replacement instead used a fixed ensemble: for each target, fold and gene–screen combination, we took the element-wise median of ten aligned response vectors before averaging across screens. Its recovery was calculated from those combined predictions. Checkpoint averaging and the prediction states used for benchmarking and response replacement are documented in Supplementary Methods S1 and Supplementary Data.

Perturbation responses, fidelity and reference reliability

For target r, screen s, knockout gene g and endpoint e, we defined the measured response relative to non-targeting controls from the same screen:

$$\Delta y_{rsge} = \overline{y}_{rsge} - \overline{y}_{rs0e}. \tag{1}$$

Here $\overline{y}_{rsge}$ is the endpoint mean over eligible cells with knockout g, $\overline{y}_{rs0e}$ is the corresponding non-targeting-control mean, and 0 denotes the control group. Hats denote predictions. Predicted responses were calculated in the same way from predictions for cells not used to train the corresponding model. Cell-level recovery was Pearson correlation across non-control test cells, calculated separately for each endpoint and averaged equally across endpoints. Knockout-response recovery used the same endpoint-wise calculation across eligible gene–screen responses:

$$R_r^{\mathrm{KO}} = \frac{1}{|E_r|} \sum_{e \in E_r} \mathrm{corr}_{\mathrm{P}} \left(\Delta y_{r \cdot \cdot e}, \Delta \widehat{y}_{r \cdot \cdot e} \right), \tag{2}$$

Here R_r^{KO} is knockout-response recovery, $\mathrm{corr_P}$ denotes Pearson correlation, and the dots index eligible gene–screen pairs for a given endpoint. The set E_r contains endpoints with an estimable correlation, and $|E_r|$ is their number. Constant or insufficient paired values yielded undefined correlations and were excluded from this average. Field and Gene recovery scores were calculated within each fold before averaging across folds. Screen- and assay-level aggregation is specified in Supplementary Methods S1.

To compare multivariate responses across folds, we converted measured and predicted endpoint changes to one target-wide control scale: the pooled standard deviation of unique non-targeting-control cells after subtracting their screen-specific means. This was an evaluation reference, separate from training-set standardization (Supplementary Methods S2, Eq. S1). Responses for each gene were then averaged equally across screens. Their measured and predicted magnitudes were

$$M_{rg} = \|\Delta\vec{y}_{rg}\|_2, \qquad \widehat{M}_{rg} = \|\Delta\widehat{\vec{y}}_{rg}\|_2. \tag{3}$$

Here $\Delta\vec{y}_{rg}$ and $\Delta\widehat{\vec{y}}_{rg}$ contain the measured and predicted endpoint changes for gene g after common-control scaling and equal-weight averaging across screens; $\|\cdot\|_2$ denotes the Euclidean norm. We measured agreement in magnitude ranking by Spearman correlation and preservation of magnitude spread by

$$Q_r^{\mathrm{var}} = \frac{\mathrm{Var}_g(\widehat{M}_{rg})}{\mathrm{Var}_g(M_{rg})}. \tag{4}$$

Here Var_g denotes variance across knockout genes, so $Q_r^{\mathrm{var}} = 1$ indicates equal measured and predicted magnitude variance. Strong-hit recall was the proportion of the 50 genes with the largest measured magnitudes recovered among the 50 largest predicted magnitudes. Within-gene profile fidelity was Pearson correlation across at least three variable, finite endpoint coordinates.

Reference reliability assessed whether measured knockout effects were stable when calculated from separate cell groups. Within each screen, a deterministic split divided cells into two halves. Endpoints were standardized against non-targeting controls using the median and median absolute deviation within each half; knockout means were summarized by their root-mean-square magnitude. Pearson agreement between halves required at least 20 common knockout genes, and target reliability was the median across screens. Separate well-pair, guide-partition and random-cell-split comparisons assessed agreement across acquisition wells, perturbation reagents and cell samples, respectively. These repeat designs and the comparison of morphologically similar cells are detailed in Supplementary Methods S3.

Measurement replacement and sufficiency classification

For each target, we performed the same hit-selection and functional analyses using measured responses and fixed-ensemble predictions. Gene Ontology biological-process and cellular-component terms[29] and Complex Portal complexes[30] were tested by hypergeometric enrichment of the top 50 genes, using the 1,000 assayed knockout genes as background and retaining terms with 3–250 assayed members. Ranked-term retention was the Jaccard overlap of the ten terms with the smallest unadjusted P values, with ties resolved by full term name. Significant-term retention was evaluated separately after Benjamini–Hochberg adjustment[44] across all eligible terms within each target, annotation family, selection budget and measured or predicted list. Significant sets used adjusted $P \leq 0.05$; recall was undefined for an empty measured set and Jaccard overlap was undefined when both sets were empty. Functional-annotation sensitivity analyses are described in Supplementary Methods S4.

Each sufficiency assessment specified the target, predictor, image representation, cellular context and intended use. For each target, the atlas classification used knockout-response recovery r (Eq. 2, averaged over five Gene folds), magnitude Spearman correlation ρ, magnitude-variance ratio $v = Q_r^{\mathrm{var}}$, top-5% hit recall h_5 and reference reliability q, applying the following rules in order.

Category	Condition, evaluated in the order shown
Not identifiable	q missing or $q < 0.30$; missing and low reliability are recorded separately.
Unresolved	Reliability passes, but required response-fidelity evidence is missing.
Quantitative proxy	$r \geq 0.70$, $\rho \geq 0.70$, $0.50 \leq v \leq 1.50$, and $h_5 \geq 0.60$.
Ranking proxy	Quantitative criteria are not all met, but $\rho \geq 0.70$ and $h_5 \geq 0.50$.
Measurement required	Neither proxy criterion is met, and $r < 0.60$.
Unresolved	All remaining cases with complete evidence.

These numerical criteria were selected for the atlas and fixed before final assignment. They describe response fidelity and candidate retention; functional-term retention was evaluated separately. A domain-sensitivity flag recorded transfer loss without changing the five-tier classification. We assessed threshold sensitivity under 36 fixed variations and predictor dependence by recomputing the same metrics and rules for each method, keeping reference reliability and data partitions unchanged (Supplementary Methods S5 and S6). The supplementary utility index and its prediction analysis were not used for tier assignment; their definitions and sensitivity analyses are given in Supplementary Methods S7.

Testing use decisions on separate knockout genes

We tested whether evidence from one group of knockout genes could identify prediction uses that remained successful on another group. Within each original Gene test fold, complete genes were divided into calibration and evaluation groups; neither group had been used to train or select that fold's predictor. We reused the fixed target-specific MLP as the primary predictor and Ridge, scButterfly and the ensemble for sensitivity analyses. Each target–fold pair formed one decision case, giving 260 cases across 52 targets and five folds.

The task was to recover the strongest 20% of measured knockout responses, rounding the candidate count up. Evaluation success required recall of at least 0.60. We compared five rules, applying all acceptance criteria to calibration genes only; h_{20} denotes recall at this 20% budget.

Rule	Calibration requirement
Accept all	No selection of uses.
Recovery only	Pearson $r \geq 0.70$.
Rank only	Magnitude Spearman $\rho \geq 0.70$.
Rank plus hit retention	$\rho \geq 0.70$ and $h_{20} \geq 0.50$.
Full response	$r \geq 0.70$, $\rho \geq 0.70$, $0.50 \leq v \leq 1.50$, and $h_{20} \geq 0.60$.

We recorded accepted successes, accepted failures, missed successes and correct rejections. Acceptance was reported as a proportion of all cases, failure as a proportion of accepted cases, and retained successes as a proportion of all cases that met the evaluation task criterion. A zero denominator yielded an undefined proportion. Missing calibration metrics could not satisfy a criterion, and missing evaluation outcomes were excluded from performance denominators and reported separately.

The primary analysis included all 52 targets without a reliability gate; the 43 targets passing the atlas reliability criterion formed a sensitivity subset. Twenty fixed gene allocations assessed partition sensitivity without retraining or changing the rules. Additional analyses compared rules at equal acceptance proportions and varied selection budgets and success requirements. Confidence intervals used 2,000 paired target-cluster bootstrap resamples, retaining all folds and allocations for each sampled target. Exact allocations, normalized rule scores for matched coverage and the complete task definitions are provided in Supplementary Methods S8 (OPS) and Supplementary Tables 4 and 5.

Controls for same-cell information and screen transfer

To determine what same-cell pairing contributed, we kept the MLP and Gene partitions fixed and compared exact pairing with shuffled training and validation targets within each gene–screen group. A second shuffle matched cell area, local cell density and eccentricity before exchanging targets. Test cells retained their correct pairing. Size–shape-only and phase-minus-size–shape inputs tested the contribution of gross geometry. A separate residual analysis subtracted a cross-fitted mean from other cells with the same gene and screen, then predicted the remaining single-cell variation. Because

the test residual requires measured phenotypes from peer cells, this arm assessed conditional same-cell information rather than prediction of a new perturbation mean.

For screen transfer, we related performance changes to shifts in source and destination non-targeting controls, analysing the 81 transfer directions descriptively and the 34 target-level summaries for inference. We also compared fluorescence differences between morphology-nearest cells and other cells with the same knockout. Full matching and residual procedures are given in Supplementary Methods S9; control-shift and nearest-neighbour analyses in Supplementary Methods S3; and the source-atlas perturbation-distinctiveness analysis[22] in Supplementary Methods S10.

Biological cases and pilot-guided assay selection

The MTOR/RPTOR and DDX18 cases used the original ensemble predictions and the same endpoint scales as the response-replacement analysis. We compared mean-intensity changes with measurements from individual acquisition wells and sgRNAs. DDX18 p53/p21 observations were linked by exact phase-cell identity. For the p21 functional example, we retained all seven assayed members of GO:0000470 and compared their ranks, top-50 membership and enrichment using the existing 1,000-gene background. Case selection, eligibility and repeat calculations are given in Supplementary Methods S14.

Assay selection used 75 target–screen combinations with three eligible wells. The destination screen was absent from predictor fitting and model selection. We analytically reconstructed the original Ridge coefficients using the original source cells, preprocessing and selected penalty, checking agreement with the deposited predictions before calculating matched-well predictions. Prediction and direct options used the same knockout cells and non-targeting controls in two wells, with equal well weights; the third well was the measured reference. Calibration genes supplied assay choices and disjoint evaluation genes supplied outcomes.

For each combination and gene group, quantitative error was

$$E(\vec{z}) = \frac{\sum_{g \in H} \|\vec{z}_g - \vec{y}_g\|_2^2}{\sum_{g \in H} \|\vec{y}_g\|_2^2}. \tag{5}$$

Here $\vec{y}_g$ is the third-well measured response vector for gene g on the common control scale, $\vec{z}_g$ is either the two-well predicted vector or the two-well direct-measurement vector, and H contains the strongest 20% of reference responses in that calibration or evaluation group, ranked by $\|\vec{y}_g\|_2$ and rounding up. All retained endpoints enter the vectors. We ranked assays by their calibration error reduction, $E(\vec{z}^{\text{pred}}) - E(\vec{z}^{\text{direct}})$, retaining up to 15 positive-benefit combinations and using predictions for the remainder. Equal-count comparators selected assays by lowest calibration correlation or highest prediction error; random, all-prediction and all-direct choices supplied additional references. Final errors were averaged equally across the 75 combinations. We repeated choices using 50 calibration genes, with 40–50 eligible genes per combination. Target- and screen-cluster intervals, repeat-consistent subsets and allocation sensitivities are specified in Supplementary Methods S15. The additional frozen-model screen and joint gene/screen use-decision evaluations are specified there separately.

Validation in independent data resources

OASIS: biochemical measurements in primary hepatocytes. OASIS provides well-paired brightfield and biochemical measurements for 1,085 compounds in primary human hepatocytes[23]. The response-fidelity analysis used 808 brightfield CellProfiler features and Ridge regression, with entire compounds assigned to five folds. Recovery of metabolic activity and lactate dehydrogenase release used the deposited plate-normalized values; response-fidelity metrics additionally used plate-control-centred responses. Reliability used disjoint replicate-well halves. Feature extraction depended on the deposited fluorescence-derived segmentation masks (Supplementary Methods S11).

The focused metabolic-injury task used the released 768-dimensional brightfield DINO features[23] and Ridge regression with L2-penalty strength $\alpha = 100$. The fixed partition contained 868 development compounds and 217 test compounds; all doses and sources of each test compound were excluded from fitting. All 868 development compounds contributed to training, while compound-level development scoring excluded Hycanthone because it lacked one of the eight concentration positions, leaving 867 scored compounds. Metabolic-activity loss was the negative change from same-plate DMSO. For each compound, responses were averaged within source at matched actual

concentrations, equally across sources, and then across its eight assayed concentration positions. We compared the eleven largest measured and predicted losses, with matched cell-count and acquisition-metadata predictors and measured-repeat comparisons in the 94 compounds with complete concentrations in two sources. The cell counts came from fluorescence-derived masks. Compound bootstrap resampling preserved doses and source pairing; additional predictors, budgets and correspondence controls are specified in Supplementary Methods S12.

The LDH task used the same 768-dimensional features, compound partition, train-only preprocessing and fixed Ridge/HGB settings. LDH response was the positive change from same-plate DMSO, averaged over sources and assayed concentrations as above. We compared the strongest eleven measured and predicted increases. Source-resolved examples retained every available actual concentration and production source; their shaded ranges describe source variation. The complete predictor comparison, development-use assessment and LDH fit details are given in Supplementary Methods S16.

Cell Health: health phenotypes in two cellular backgrounds. Fluorescent Cell Painting profiles were linked to separately acquired cell-health measurements by cell context and sgRNA in ES2 and HCC44[24]. We analysed ten endpoints in each context, covering cell-cycle fractions, cell death, live-cell count, oxidative stress and DNA damage. Whole genes, including their guides and both contexts, were assigned to 25 training, 12 calibration and 12 evaluation genes. Separate Ridge models used training-only preprocessing and a fixed penalty. The task selected the four largest absolute responses in each twelve-gene group; success required recovering at least three of the four measured candidates. We compared correlation-, rank- and task-based acceptance, with or without a repeat-agreement criterion from the two reference plates. Exact inclusion criteria, rules and allocation sensitivity are given in Supplementary Methods S8 (Cell Health) and Supplementary Tables 4 and 5.

LINCS: compound-induced RNA responses. We linked A549 Cell Painting at 48 h to separately acquired L1000 RNA measurements at 24 h for matched 10 μM compound treatments[25]. Inclusion required at least two image plates and two RNA plates. The fixed partition comprised 775 training, 257 calibration and 260 evaluation compounds, grouping chemical identities and all their observations together. Ridge and MLP selection used training compounds only. The task selected the 26 largest absolute RNA responses per endpoint, with success requiring at least 16 recovered candidates. Acceptance rules used recovery, rank, task recall and repeat agreement, separately or in combination. Two RNA references were formed from disjoint acquisition groups; repeat-supported success required recovery against both, in addition to the consensus. Their construction, complete rules and chemical-group uncertainty analyses are described in Supplementary Methods S8 (LINCS).

PERISCOPE: prediction between fluorescence channels. We predicted anti-TOMM20 measurements from DAPI, ConA, phalloidin and WGA features in the PERISCOPE resource[26], using both single-cell and guide-level profiles. Cross-channel features involving anti-TOMM20 were excluded from predictors, and gene-grouped fivefold evaluation used validation-selected Ridge regression. The deposited segmentation included the target channel. Reliability was agreement between disjoint guide halves within each plate, summarized across plates. Resource-specific response aggregation and repeat definitions are specified in Supplementary Methods S11. Permuted-input controls are described in Supplementary Methods S13.

Representation analyses and statistical inference

Label-efficiency analyses used twelve targets with 0.1%, 1% or 20% of paired training and validation observations, while leaving test data unchanged. We distinguished target-only predictors from shared predictors that additionally used complete measurements for the remaining forty targets. On a separate six-target panel, we compared engineered phase features[22], a frozen DINOv2 encoder[45], a frozen Cytoland encoder[14], and partially fine-tuned Cytoland[14], using the same five Gene partitions. Sampling, crop preprocessing, fine-tuning and evaluation details are provided in Supplementary Methods S17.

Atlas-wide comparisons used targets as the inferential unit. Method comparisons used two-sided paired Wilcoxon signed-rank tests, with Holm correction across the 54 planned comparisons[46]; a

difference required Holm-adjusted $P < 0.05$ and an absolute median paired effect of at least 0.02. The MLP was the prespecified statistical reference, while all ten methods were reported. Bootstrap analyses preserved the grouping relevant to each claim: targets for atlas-wide comparisons, whole genes for response-repeat estimates and whole compounds for compound-selection tasks. Unless otherwise specified, intervals were percentile 95% intervals from 10,000 target-level resamples. Exact sample sizes and resampling settings are given in the corresponding legends and source tables; paired method comparisons are detailed in Supplementary Methods S1.

Software and hardware

Analyses used Python 3.10.20, NumPy 1.26.4, pandas 2.1.4, SciPy 1.11.4 and scikit-learn 1.7.2; neural models used PyTorch 2.3.1, gradient-boosted trees used XGBoost 2.1.4 and CatBoost used version 1.2.10. Computations used NVIDIA H100 GPUs (80 GB) and multi-core CPUs. Independent reporter- and fold-level jobs were run in parallel, with up to four concurrent jobs per GPU in the same-cell analyses. Figures were rendered with Matplotlib 3.7.5.

Data availability

The optical perturbation atlas was generated by the OPS study[22]. We used revision 6 of its public Zenodo record (https://doi.org/10.5281/zenodo.20495192), released 1 June 2026 under CC BY 4.0[47]. Full imaging and molecular data remain available through the Biohub OPS Explorer (https://biohub.ai/ops-explorer) under the source resource's terms. The external resources cpg0037-oasis, cpg0021-periscope, cpg0004-lincs (Cell Painting) and cpg0003-rosetta (LINCS-Pilot1 L1000) are available through the Cell Painting Gallery[23,25,26,48]. Cell Health inputs and targets were obtained from the public version 2.0 deposit (https://github.com/broadinstitute/cell-health/tree/v2.0)[24]. Source-file hashes, LINCS RNA S3 version identifiers, image-feature manifests and chemical/guide matching records accompany Supplementary Data.

OASIS DINO profiles and metadata for metabolic-activity prioritization were obtained from the source study's Axiom OASIS deposit (https://doi.org/10.5281/zenodo.17067683, Jessica Ewald, CC BY 4.0). Derived tables retain source attribution and the deposited numerical values; predictions, aggregation and selection records are outputs of the present analysis.

Processed OPS phase features, exact-linked reporter targets, metadata and fixed field- and gene-holdout assignments are available in the MorphoSuff dataset repository (https://huggingface.co/datasets/Amanda1998/MorphoSuff). One Supplementary Data archive accompanies this manuscript. It contains OPS predictor-specific evidence, calibration–evaluation decisions and functional analyses; and OASIS, Cell Health and LINCS prediction and selection evidence. Primary LINCS predictions and references are stored without duplicating shared arrays; metrics and decision records cover all reported allocations. Full raw microscopy remains with the source resources; provenance and source-specific licences accompany the data.

Gene Ontology[28,29] and Complex Portal[30] were reused as public third-party annotations. The releases, download dates and identifiers used in the analysis are recorded in the source-data provenance.

Code availability

MorphoSuff software for data preparation, model fitting and measurement-sufficiency assessment is available at https://github.com/limengran98/MorphoSuff under the MIT License. The accompanying Supplementary Data contain the numerical records underlying the reported decision, functional and external-use analyses. The Python package provides dataset adapters and shared response-fidelity, reliability and measurement-decision analyses for use with other paired datasets or user-supplied predictions.

Author contributions

Mengran Li, Jianqing Zhu and Bo Li designed the study with Ronghui Zhang, Lian Zhang and Jinchao Xu. Mengran Li developed the assessment framework and software, curated the data, performed the computational analyses and prepared the figures. Jianqing Zhu and Bo Li contributed to

methodological development, validation and interpretation of the results. Chengyang Zhang, Zhenchao Tang, Jiaying Wang, Wenbin Xing, Boyu Zhang, Lingbei Meng, Bob Zhang, Junzhou Chen, Quanyou Cai, Jinfeng Xu, Zelin Zang and Stan Z. Li contributed to validation, interpretation of the results and manuscript revision. Ronghui Zhang, Lian Zhang and Jinchao Xu supervised the study and coordinated the project. Mengran Li, Jianqing Zhu and Bo Li drafted the manuscript. All authors discussed the results, revised the manuscript and approved the final version.

Funding

This work was supported by the National Natural Science Foundation of China (grant no. 52672548).

Competing interests

The authors declare no competing interests.

References

[1] Christoph Bock et al. High-content CRISPR screening. *Nature Reviews Methods Primers*, 2:8, 2022. doi: 10.1038/s43586-021-00093-4.

[2] Atray Dixit et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167:1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038.

[3] Joseph M. Replogle 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] Michael Boutros, Florian Heigwer, and Christina Laufer. Microscopy-based high-content screening. *Cell*, 163:1314–1325, 2015. doi: 10.1016/j.cell.2015.11.007.

[5] David Feldman, Avtar Singh, Jonathan L. Schmid-Burgk, Rebecca J. Carlson, Anja Mezger, Anthony J. Garrity, Feng Zhang, and Paul C. Blainey. Optical pooled screens in human cells. *Cell*, 179(3):787–799.e17, 2019. doi: 10.1016/j.cell.2019.09.016.

[6] Luke Funk, Kuan-Chung Su, Jimmy Ly, David Feldman, Avtar Singh, Brittania Moodie, Paul C. Blainey, and Iain M. Cheeseman. The phenotypic landscape of essential human genes. *Cell*, 185 (24):4634–4653.e22, 2022. doi: 10.1016/j.cell.2022.10.017.

[7] Mark-Anthony Bray et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11:1757–1774, 2016. doi: 10.1038/nprot.2016.105.

[8] Beth A. Cimini et al. Optimizing the Cell Painting assay for image-based profiling. *Nature Protocols*, 18(7):1981–2013, 2023. doi: 10.1038/s41596-023-00840-9.

[9] Jan Oscar Cross-Zamirski et al. Label-free prediction of Cell Painting from brightfield images. *Scientific Reports*, 12:10001, 2022. doi: 10.1038/s41598-022-12914-x.

[10] Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D. Boyd, and Anne E. Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, 20:145–159, 2021. doi: 10.1038/s41573-020-00117-w.

[11] YongKeun Park, Christian Depeursinge, and Gabriel Popescu. Quantitative phase imaging in biomedicine. *Nature Photonics*, 12:578–589, 2018. doi: 10.1038/s41566-018-0253-x.

[12] Eric M. Christiansen et al. In silico labeling: predicting fluorescent labels in unlabeled images. *Cell*, 173:792–803.e19, 2018. doi: 10.1016/j.cell.2018.03.040.

[13] Chawin Ounkomol, Sharmishtaa Seshamani, Mary M. Maleckar, Forrest Collman, and Gregory R. Johnson. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature Methods*, 15:917–920, 2018. doi: 10.1038/s41592-018-0111-2.

[14] Ziwen Liu et al. Robust virtual staining of landmark organelles with Cytoland. *Nature Machine Intelligence*, 7:901–915, 2025. doi: 10.1038/s42256-025-01046-2.

[15] Pan Deng, Deasung Jang, Samuel G. Berryman, Simon P. Duffy, and Hongshen Ma. Regularized single-cell imaging enables generalizable AI models for stain-free cell viability screening. *Small Methods*, 10(4):e01369, 2026. doi: 10.1002/smtd.202501369.

[16] Charlotte Bunne et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024. doi: 10.1016/j.cell.2024.11.015.

[17] Jianxu Chen, Matheus P. Viana, and Susanne M. Rafelski. When seeing is not believing: application-appropriate validation matters for quantitative bioimage analysis. *Nature Methods*, 20:968–970, 2023. doi: 10.1038/s41592-023-01881-4.

[18] Alexandr A. Kalinin et al. The DynaCell evaluation suite for dynamic 3D virtual staining of live cells. *arXiv preprint arXiv:2611.02324*, 2026. URL https://arxiv.org/abs/2611.02324.

[19] Juan C. Caicedo et al. Data-analysis strategies for image-based cell profiling. *Nature Methods*, 14:849–863, 2017. doi: 10.1038/nmeth.4397.

[20] Srijit Seal et al. Counting cells can accurately predict small-molecule bioactivity benchmarks. *Nature Communications*, 17:2436, 2026. doi: 10.1038/s41467-026-68725-5.

[21] Nitsan Elmalam and Assaf Zaritsky. Cell context-dependent in silico organelle localization in label-free microscopy images. *Nature Methods*, 23:405–416, 2026. doi: 10.1038/s41592-025-02960-4.

[22] Chad Liu et al. A multimodal perturbation atlas defines the phenotypic resolution of cellular morphology. *bioRxiv*, 2026. doi: 10.64898/2026.06.01.728087.

[23] Jessica D. Ewald et al. Cell Painting for cytotoxicity and mode-of-action analysis in primary human hepatocytes. *Cell Systems*, 17(5):101566, 2026. doi: 10.1016/j.cels.2026.101566.

[24] Gregory P. Way et al. Predicting cell health phenotypes using image-based morphology profiling. *Molecular Biology of the Cell*, 32:995–1005, 2021. doi: 10.1091/mbc.E20-12-0784.

[25] Marzieh Haghighi, Juan C. Caicedo, Beth A. Cimini, Anne E. Carpenter, and Shantanu Singh. High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. *Nature Methods*, 19:1550–1557, 2022. doi: 10.1038/s41592-022-01667-0.

[26] Meraj Ramezani et al. A genome-wide atlas of human cell morphology. *Nature Methods*, 22: 621–633, 2025. doi: 10.1038/s41592-024-02537-7.

[27] Srinivas Niranj Chandrasekaran et al. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, 2023. doi: 10.1101/2023.03.23.534023.

[28] Michael Ashburner et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25: 25–29, 2000. doi: 10.1038/75556.

[29] Gene Ontology Consortium. The Gene Ontology knowledgebase in 2023. *Genetics*, 224:iyad031, 2023. doi: 10.1093/genetics/iyad031.

[30] Birgit H M Meldal et al. Complex Portal 2022: new curation frontiers. *Nucleic Acids Research*, 50: D578–D586, 2022. doi: 10.1093/nar/gkab991.

[31] Agnes Roczniak-Ferguson, Constance S. Petit, Florian Froehlich, Sharon Qian, Jennifer Ky, Brittany Angarola, Tobias C. Walther, and Shawn M. Ferguson. The transcription factor TFEB links mTORC1 signaling to transcriptional control of lysosome homeostasis. *Science Signaling*, 5(228): ra42, 2012. doi: 10.1126/scisignal.2002790.

[32] Pavel Moudry, Katarina Chroma, Sladana Bursac, Sinisa Volarevic, and Jiri Bartek. RNA-interference screen for p53 regulators unveils a role of WDR75 in ribosome biogenesis. *Cell Death & Differentiation*, 29(3):687–696, 2022. doi: 10.1038/s41418-021-00882-0.

[33] Walid M. Abdelmoula et al. Advancing morphological profiling with self-supervised learning through comparative insights into representation learning of cellular perturbations. *Nature Methods*, 23:2181–2194, 2026. doi: 10.1038/s41592-026-03270-z.

[34] Michael Doron et al. Unbiased single-cell morphology with self-supervised vision transformers. *bioRxiv*, 2023. doi: 10.1101/2023.06.16.545359.

[35] Oren Kraus et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11757–11768, 2024. doi: 10.1109/cvpr52733.2024.01117.

[36] Kian Kenyon-Dean et al. ViTally consistent: Scaling biological representation learning for cell microscopy. *Proceedings of Machine Learning Research*, 267:29735–29752, 2025. URL https://proceedings.mlr.press/v267/kenyon-dean25a.html.

[37] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

[38] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[39] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, pages 6638–6648, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html.

[40] Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Sd4wYYOhmY.

[41] Yichuan Cao et al. scButterfly: a versatile single-cell cross-modality translation method via dual-aligned variational autoencoders. *Nature Communications*, 15:2973, 2024. doi: 10.1038/s41467-024-47418-x.

[42] Zhen He et al. Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. *Nature Biotechnology*, 42:1594–1605, 2024. doi: 10.1038/s41587-023-02040-y.

[43] Hongru Hu and Gerald Quon. scPair: Boosting single cell multimodal analysis by leveraging implicit feature selection and single cell atlases. *Nature Communications*, 15:9932, 2024. doi: 10.1038/s41467-024-53971-2.

[44] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[45] Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

[46] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

[47] Chad Liu et al. A multimodal perturbation atlas defines the phenotypic resolution of cellular morphology. Zenodo, 2026.

[48] Erin Weisbart et al. Cell Painting Gallery: an open resource for image-based profiling. *Nature Methods*, 21(10):1775–1777, 2024. doi: 10.1038/s41592-024-02399-z.