Supplementary Information

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,✉]

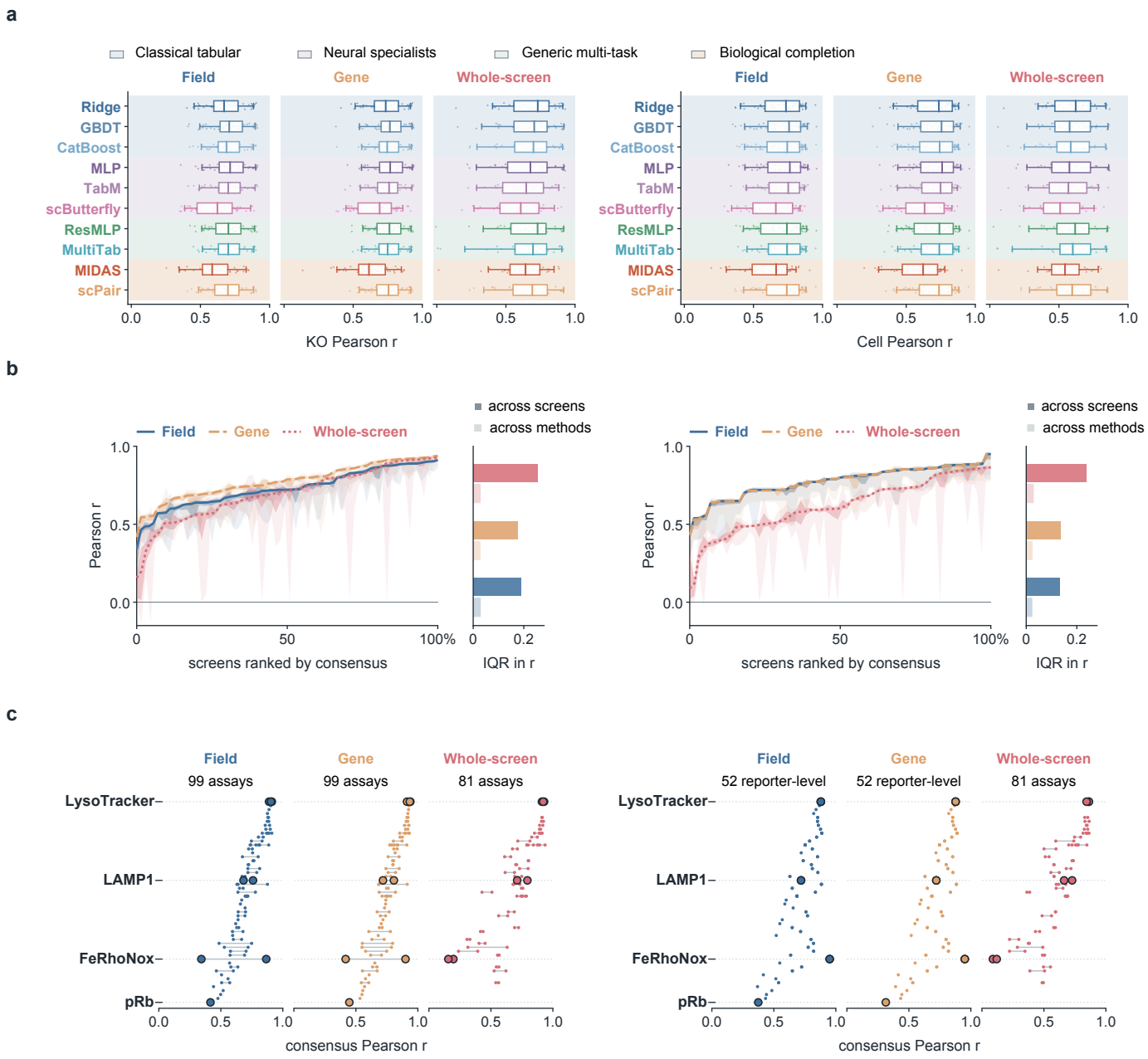
[1]School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China
[2]Shenzhen Loop Area Institute, Shenzhen, China
[3]PAMI Research Group, Department of Artificial Intelligence, University of Macau, Macau SAR
[4]Department of Computer Science, School of Computing, National University 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).

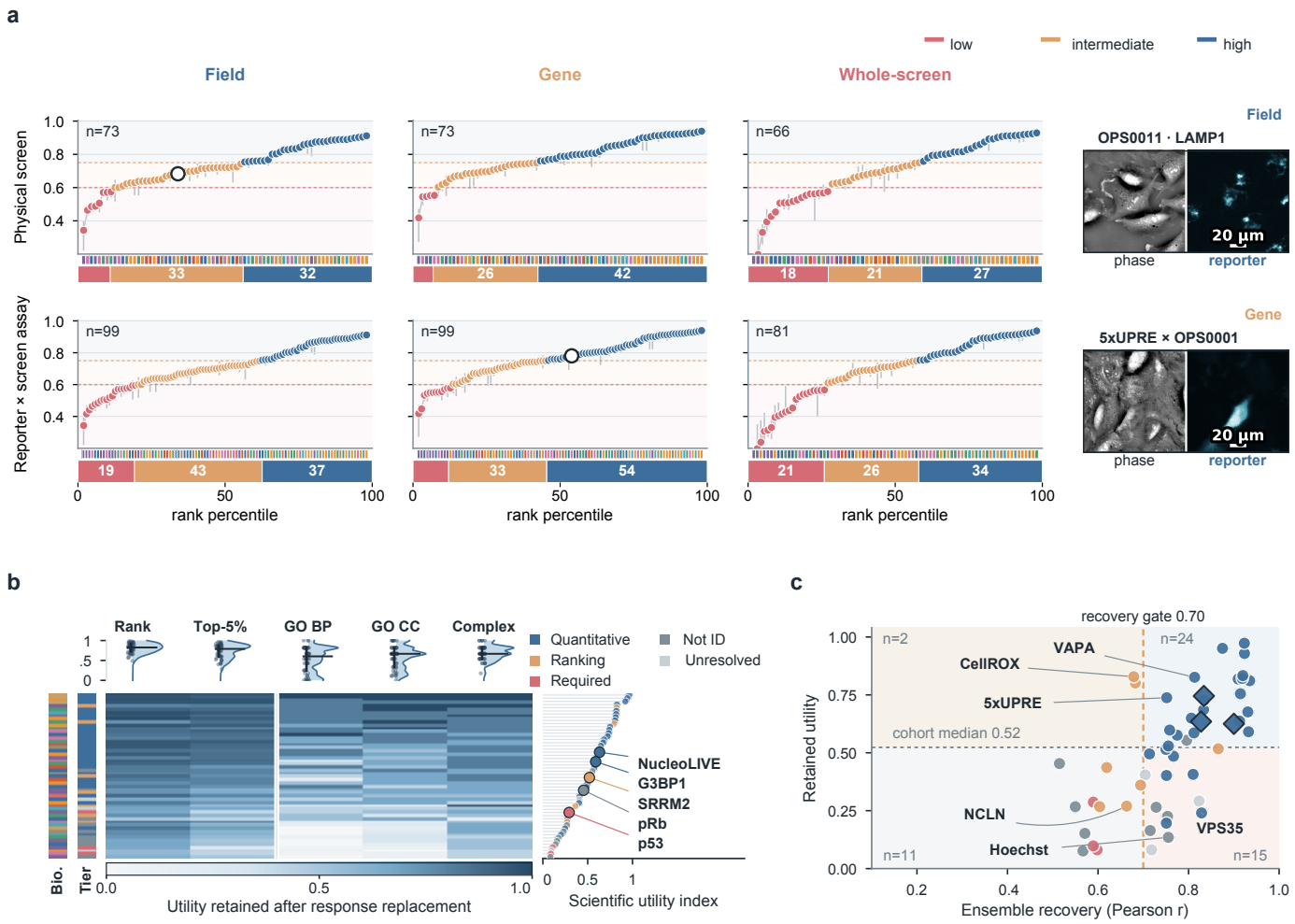
Contents

- Supplementary Figures 1 to 9
- Supplementary Tables 1 to 5
- Supplementary Data
- Supplementary Methods S1 to S17
- Supplementary Note 1
- Supplementary References

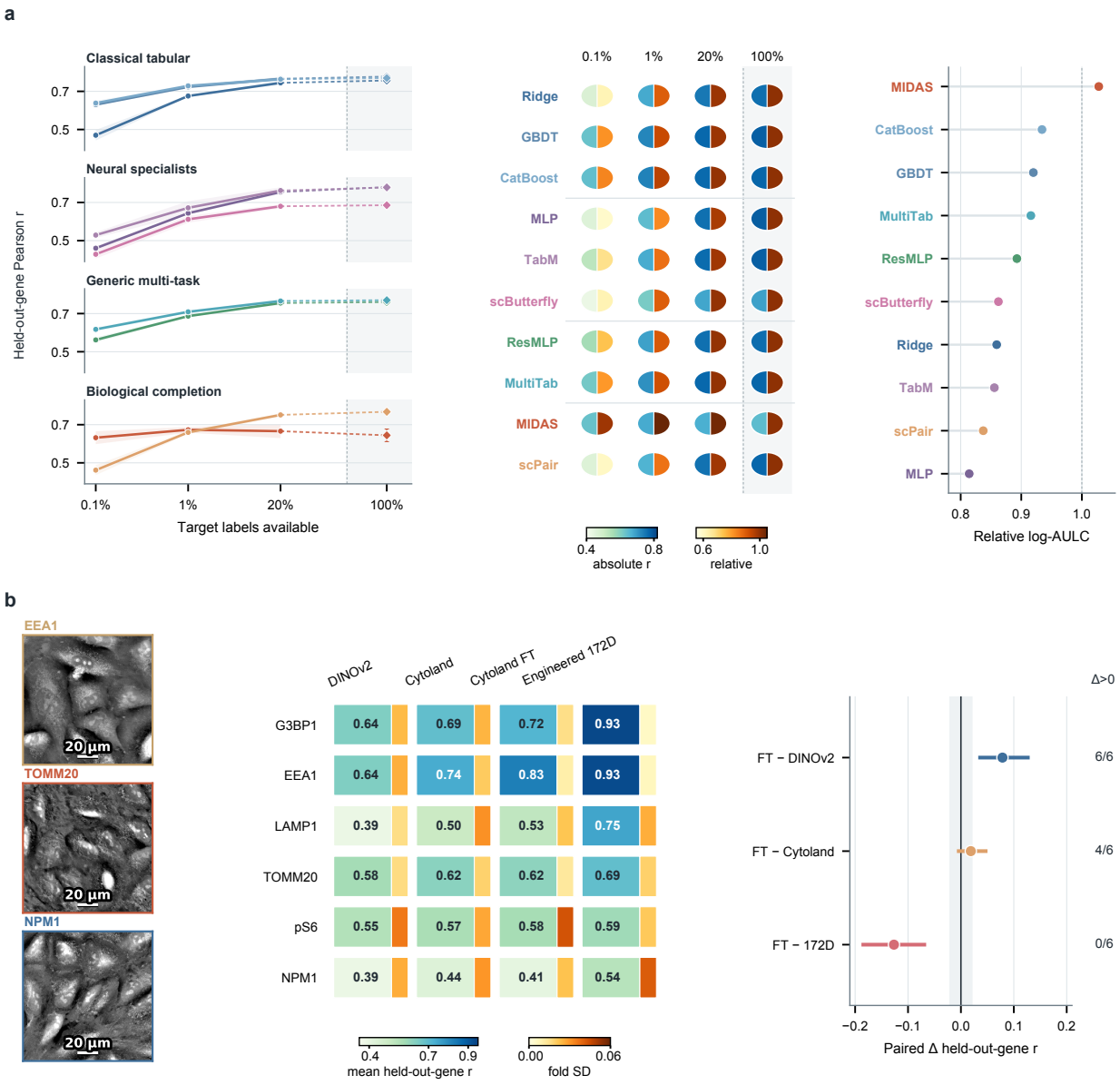
Supplementary Figures



Supplementary Figure 1: **Complete predictor comparisons across targets, screens and measurement combinations.** Panels show knockout responses on the left and cellular phenotypes on the right. **a,** Ten-method comparisons across 52 reporters (Field/Gene) or 34 repeated reporters (Whole screen). Points denote reporters; boxes show medians and interquartile ranges (IQRs), whiskers 5th–95th percentiles and pale bands method families. **b,** Screens ordered separately within each setting by the median score across methods, with bands showing the method IQR (dark) and full range (pale). Bars compare the across-screen IQR with the median within-screen, across-method IQR. **c,** Reporter–screen assay scores, summarized by the median across methods; connectors span assays of the same reporter. Field/Gene cell scores are mapped from reporter-level values to assay membership (52 distinct values), whereas Whole screen cell scores are evaluated directly in destination cells.



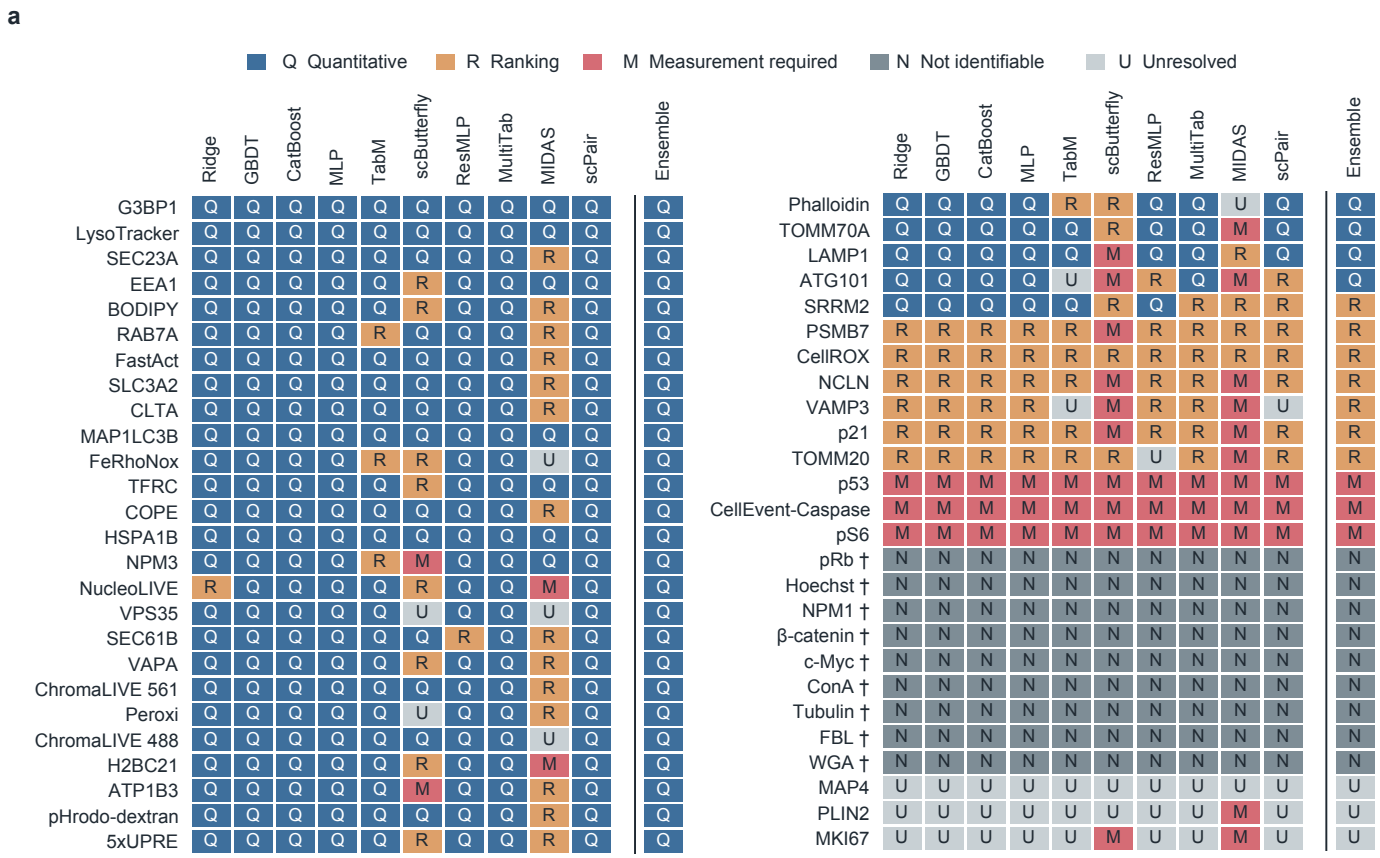
Supplementary Figure 2: **Screen-level recoverability and retained scientific utility provide complementary views of the atlas. a,** Knockout-response recoverability for physical screens and reporter–screen assays under field, gene and whole-screen holdouts. Points are unweighted medians across ten methods, ordered separately in each view; intervals span the method interquartile range, and bars count low ($r < 0.60$), intermediate ($0.60 \leq r < 0.75$) and high ($r \geq 0.75$) descriptive strata. Field/gene views contain 73 screens and 99 assays; whole-screen views contain 66 destination screens and 81 directed assays. Colour strips denote biological systems, and rings identify the exact-cell examples at right; scale bars, 20 μm. **b,** Retention of response ranking, strongest knockout hits and the ten highest-ranked functional terms after ensemble response replacement for 52 reporters, ordered by the utility index. Points, medians and interquartile ranges accompany the distributions; the utility index combines cohort-percentile hit and functional-list retention, and colours denote response tiers. **c,** Ensemble recoverability versus utility expressed relative to the leave-one-reporter-out training reference; diamonds identify domain-sensitive reporters. The recovery line at 0.70 is an operational criterion, whereas the horizontal line is the descriptive cohort median. Functional overlap in **b,c** uses the ten leading terms, irrespective of significance.



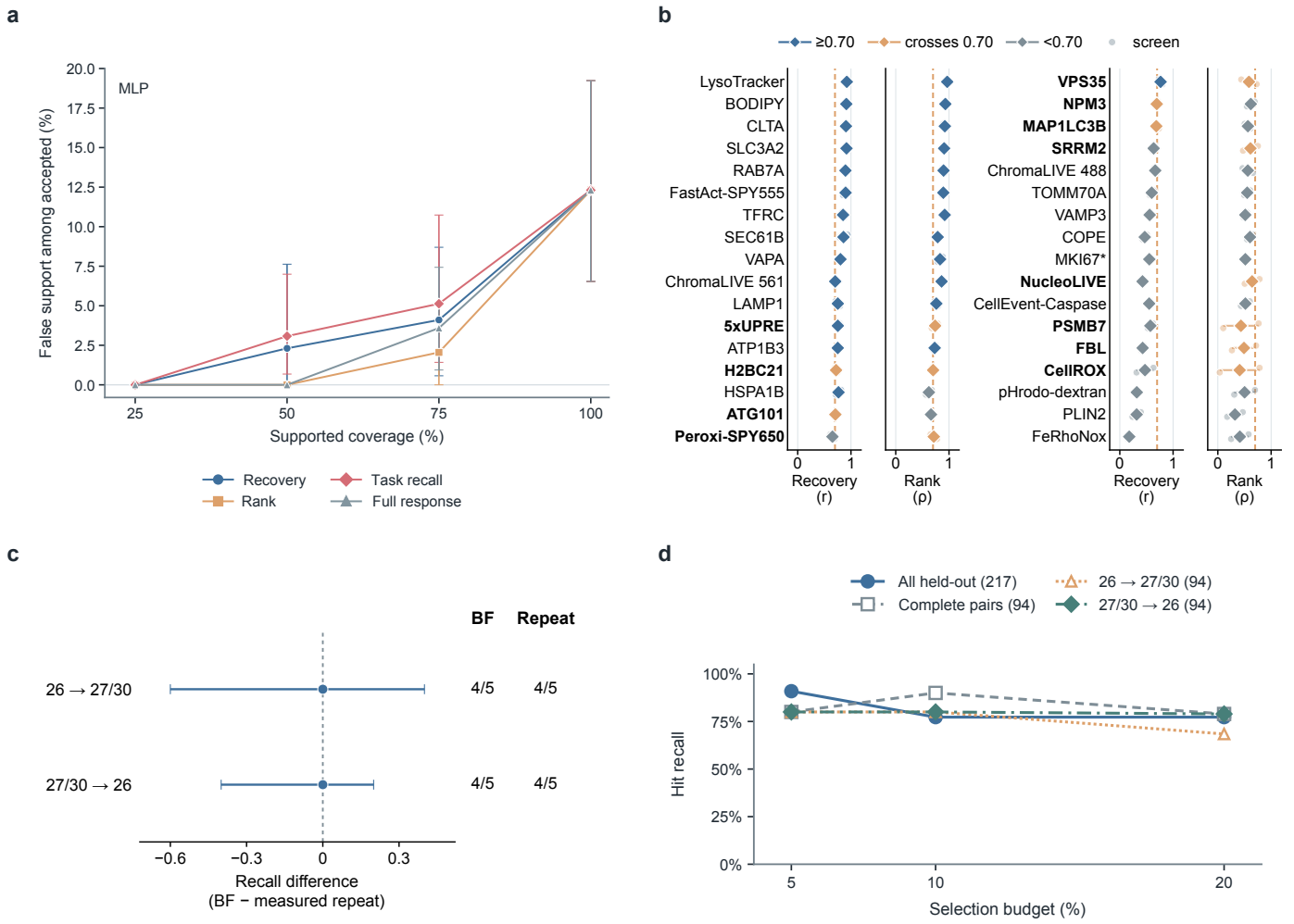
Supplementary Figure 3: **Label-efficiency calibration and representation sensitivity. a,** Held-out-gene performance on a panel of 12 reporters. Curves and bands show equal-reporter means and one standard deviation across five outer gene partitions for ten methods at 0.1%, 1% and 20% target labels; diamonds are matched 100% evaluations from a separate full-label experiment, and dashed 20–100% segments connect the experiments. In each split glyph, the left half gives Pearson correlation and the right half gives recovery relative to the method's 100% result. Relative log-AULC is the trapezoidal mean of relative recovery over $\log_{10}$ label fraction from 0.1% to 20%. **b,** Raw-phase representation comparison for six reporters. The matrix gives the five-fold mean held-out-gene Pearson correlation for each reporter and representation, and the adjacent strip gives the fold standard deviation. Paired reporter-level contrasts compare partially fine-tuned Cytoland with frozen DINOv2, frozen Cytoland and engineered 172-dimensional morphology; points are mean differences, intervals are reporter-hierarchical 95% bootstrap confidence intervals and counts denote reporters with a positive mean difference. Registered phase crops illustrate three inputs; scale bars, 20 μm.



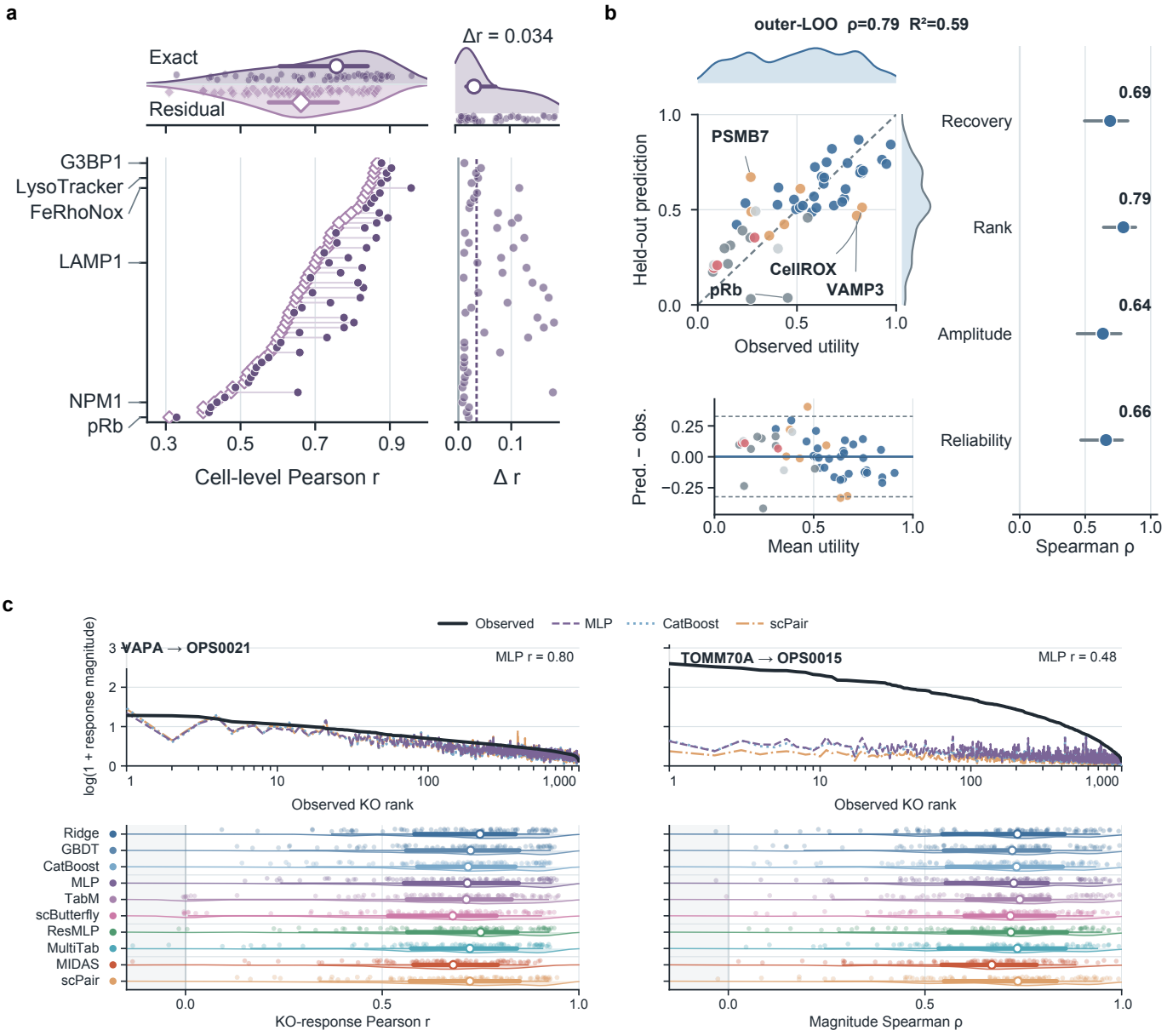
Supplementary Figure 4: **Measurement-sufficiency evidence across three experimental contexts. a,** All 52 OPS reporters evaluated against the five criteria. Columns are reporters ordered by decision tier; the upper strip denotes quantitative proxy, ranking proxy, measurement required, not identifiable and unresolved assignments (blue, orange, rose, grey and pale grey, respectively). Colour gives the signed, scaled margin to the corresponding criterion, with blue indicating that the criterion is met and red that it is not. **b,** Evidence for the two independent measurement settings and their input-permutation controls. Points show fold or screen estimates; large markers are the reported summaries and vertical lines indicate thresholds.



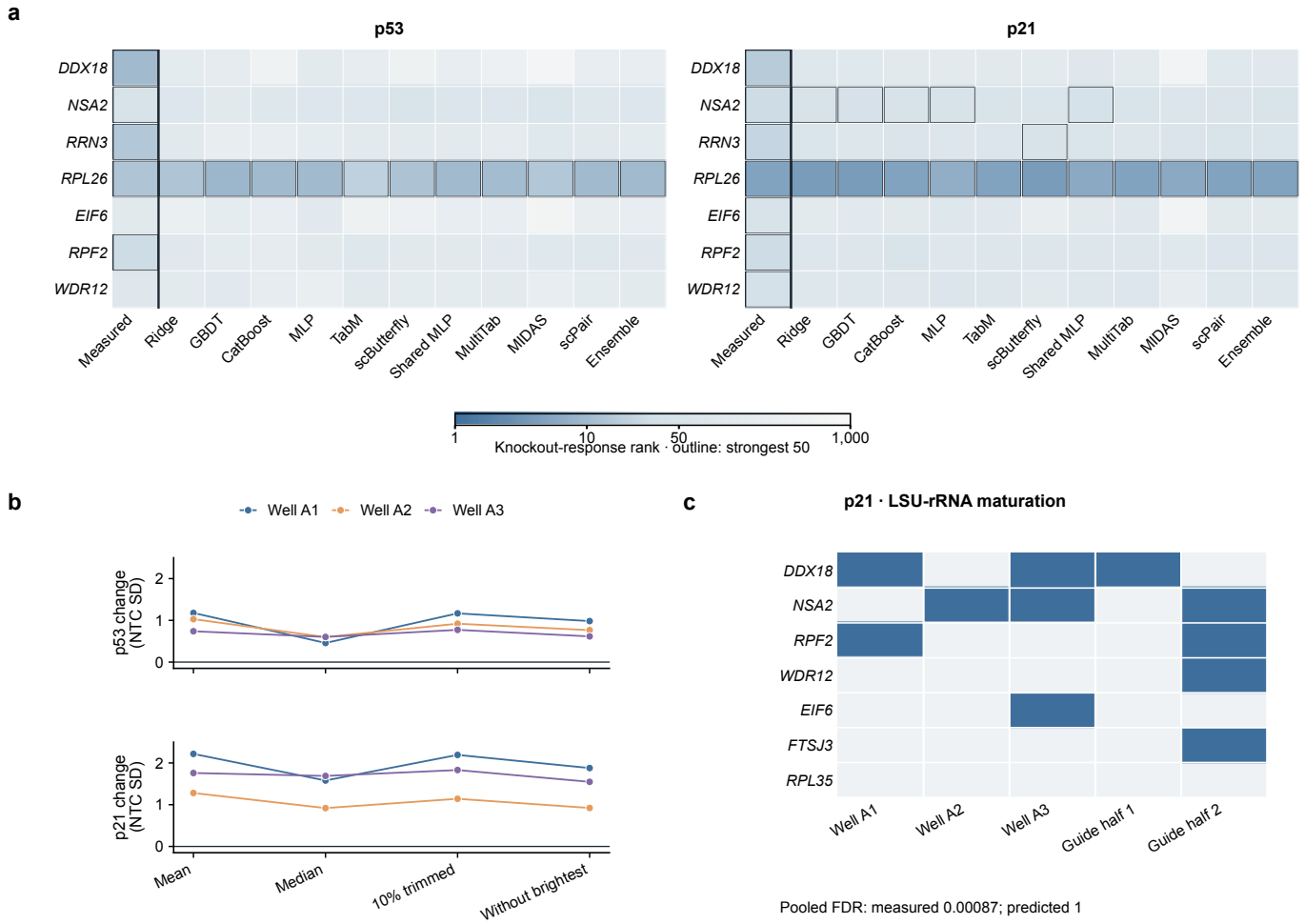
Supplementary Figure 5: **Predictor choice changes the measurement uses supported for individual targets. a,** Complete assignments for 52 reporters and ten predictors, with the fixed median ensemble as reference. The two blocks continue the reporter order used in Supplementary Figure 4a; colours and letters distinguish quantitative proxy (Q), ranking proxy (R), measurement required (M), not identifiable (N) and unresolved (U). Each method supplies its own recoverability, magnitude rank, variance ratio and strong-hit recall, using the same prediction states as the ensemble; partitions, common control scale, target reliability and thresholds are held fixed. Daggers identify the nine reliability-gated reporters.



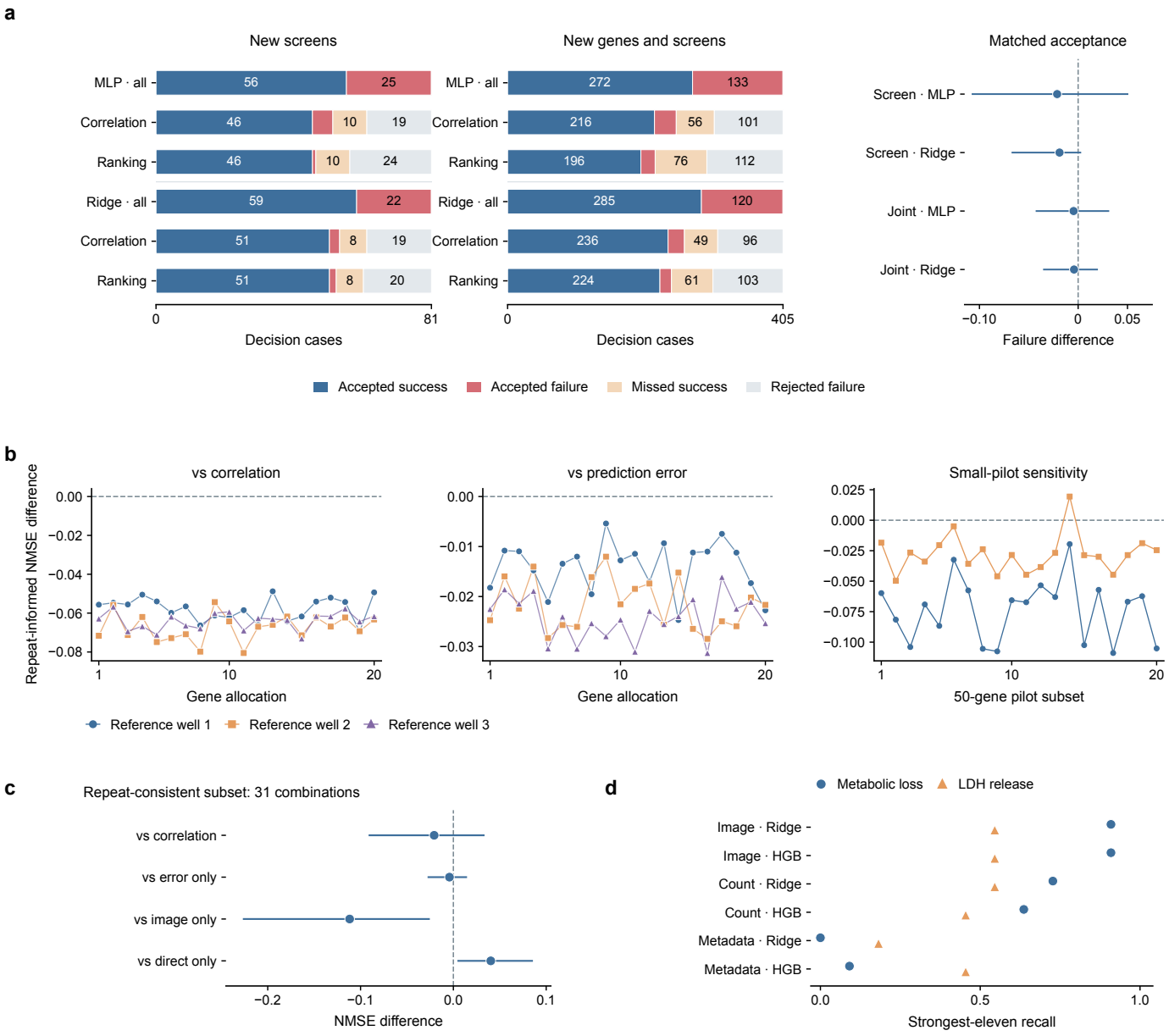
Supplementary Figure 6: **Screen context, repeat measurements and matched-coverage use decisions. a,** OPS screening failures at matched acceptance coverage. Four calibration scores selected 25%, 50%, 75% or 100% of 52 reporters within each of five folds, using frozen MLP predictions and the primary gene allocation. Success on separate evaluation genes required recall of at least 60% of the strongest 20% of measured knockout responses. Points show accepted-case failure fractions; intervals are 95% reporter-cluster bootstrap intervals (2,000 resamples), retaining all five folds per reporter. Actual coverage reflects integer-rounded selection counts. **b,** Benchmark ten-method score medians for recovery and magnitude rank across 81 directions nested in 34 reporters: points are destination evaluations, diamonds reporter medians and segments ranges, ordered down two consecutive blocks of 17 reporters. Blue remains above 0.70, grey below and gold crosses this threshold. **c,** Brightfield (BF) and measured-repeat prioritization among 94 hepatocyte-screen compounds with complete matched concentrations in source 26 and a paired source (27 or 30). Points show the difference in recall of the five strongest measured metabolic-activity losses in each cross-source direction; intervals are paired 95% compound-bootstrap intervals (2,000 resamples). Counts give retained hits for BF and the measured repeat. **d,** Recall at fixed 5%, 10% and 20% budgets, retaining the primary brightfield predictor and four evaluation populations or directions.



Supplementary Figure 7: **Complementary same-cell, downstream-retention and transfer analyses. a,** Exact-pair prediction versus an MLP fitted to cross-fitted within-gene–screen target residuals ($n = 52$), with ranked differences. Residuals use targeted measurements from peer cells in the test condition. **b,** Nested leave-one-target-out utility prediction ($n = 52$); residual lines mark mean bias and ± 1.96 standard deviations. Amplitude denotes magnitude-variance fidelity; forest intervals are 95% target-bootstrap intervals (10,000 resamples). **c,** Transfer curves show two prespecified reporter-to-destination directions and three representative predictors. Half-raincloud retain all 81 directed transfers per method, displaying density, median, interquartile range and 5th–95th percentiles for response reconstruction and magnitude ranking.



Supplementary Figure 8: **Predictor dependence and repeated-measurement support for the biological cases. a,** Measured ranks and all ten individual predictors plus the ensemble for seven fixed candidate genes in p53 and p21. Colours show log-scaled ranks among 1,000 knockouts; outlines mark the strongest 50. **b,** DDX18 intensity changes in three acquisition wells: means, medians, 10% trimmed means and means after excluding the brightest knockout cell. Each summary is centred on the corresponding well controls and divided by the common control standard deviation. **c,** Top-50 membership of all seven screened GO:0000470 members in three wells and two guide halves; blue/pale/grey, selected/not selected/missing. Pooled enrichment FDR is reported separately from repeat-specific candidate membership.



Supplementary Figure 9: **Generalization and sensitivity of prediction-use and assay choices. a,** Primary screening decisions for 81 new-screen contexts and 405 joint gene/screen-exclusion model tasks, retaining all successes, failures and omissions. Right: ranking-minus-correlation failure differences at identical acceptance counts, with 95% whole-target bootstrap intervals. Five model folds do not represent independent screens. **b,** Equal-cell quantitative-assay allocation over 20 gene allocations and three reference-well rotations; right, 20 fixed 50-gene pilots. Right-axis blue/orange compare repeat-informed choice with correlation/prediction-error choice. Allocations are sensitivity analyses, not independent experiments. **c,** The 31-combination repeat-consistent subset, with seven retained direct assays: repeat-informed-minus-comparator NMSE differences and 95% intervals from 5,000 resamples of 22 complete target clusters. **d,** All six fixed OASIS predictors evaluated on the same 217 compounds and eleven-candidate budget for metabolic loss and LDH release. Count inputs use fluorescence-derived masks.

Supplementary Tables

Supplementary Table 1: **Measurement-sufficiency criteria applied to two independent public screens.** Thresholds were retained from the OPS analysis. Values are medians over five outer folds (held-out compounds for cpg0037 and held-out genes for cpg0021), except reliability, which is estimated separately from replicate measurements. Var. ratio, predicted/measured response-magnitude variance; Recall, top-5% hit recall. Bold values fall outside the criteria for a quantitative proxy; classifications follow the hierarchy in Methods. Recoverability and reliability use the resource-specific aggregation described in Supplementary Methods.

Screen	Target (fitting unit)	r	ρ	Var. ratio	Recall	Reliab.	Tier
cpg0037	metabolic activity (well)	0.826	**0.604**	0.897	0.727	0.716	unresolved
cpg0037	LDH release (well)	**0.670**	**0.633**	0.736	0.727	0.638	unresolved
cpg0021	anti-TOMM20 (cell)	0.760	0.852	0.509	0.627	**0.005**	not identifiable
cpg0021	anti-TOMM20 (guide)	0.724	0.912	0.559	0.691	**0.047**	not identifiable

Supplementary Table 2: **Sensitivity of reporter tiers to the decision thresholds on the common control-referenced scale.** Each of eight thresholds was varied independently in stricter and less stringent directions, first by 0.05 in the metric's units and then by 10% of the threshold value. The last row varies all eight simultaneously, giving 36 configurations in total. Entries are the numbers of reporters whose tier changed among 52. The final column gives the range in the combined number of quantitative and ranking proxies, compared with 37 in the primary analysis.

	Moved by 0.05		Moved by 10%		Proxy
Threshold moved	looser	stricter	looser	stricter	total
recoverability ≥ 0.70	2	2	2	8	37
magnitude Spearman ≥ 0.70	2	3	2	6	31–39
variance ratio lower edge 0.50	1	0	1	0	37
variance ratio upper edge 1.50	0	0	0	0	37
top-5% recall ≥ 0.60 (quantitative)	0	1	0	1	37
top-5% recall ≥ 0.50 (ranking)	0	1	0	1	36–37
reliability ≥ 0.30	3	5	3	4	36–40
recoverability < 0.60 (measurement required)	3	0	3	0	37
all eight together	12	11	12	16	30–42

Supplementary Table 3: **Training and model selection for the ten full-label predictors.** All methods received the same training-standardized 172-dimensional phase representation, reporter-specific standardized target blocks and predefined train, validation and test assignments. "Observed only" means that unmeasured endpoints did not enter either the loss numerator or denominator. Specialist models were fitted independently for each reporter; shared models retained all source-reporter supervision available in a fold. The table records validation-selected benchmark states; held-out-gene response replacement also uses the archived parameter-averaged neural checkpoints described in Methods.

Method	Training scope	Implementation class	Target and loss semantics	Validation and model selection	Implementation and provenance
Ridge	Reporter-specific; one estimator per observed endpoint	Float64 closed-form ridge regression without an additional intercept[1]	Each endpoint used its observed rows; validation and test inputs were clipped to the training feature range	$\alpha \in \{0.1, 1, 10, 100, 1000\}$ selected by validation MSE	Released study implementation; centred inputs and targets make the intercept redundant
GBDT	Reporter-specific; one estimator per observed endpoint	XGBoost regression trees[2]: 300 trees, depth 6, learning rate 0.05, row and column subsampling 0.8	Squared-error objective on observed training rows	Validation RMSE with early-stopping patience 20	Official XGBoost implementation; complete software environment and configuration released
CatBoost	Reporter-specific; one joint multi-output estimator per reporter	CatBoost[3]: 300 iterations, depth 6, learning rate 0.05, L_2 leaf penalty 3 and 128 borders	MultiRMSE over the complete observed reporter block; Plain boosting	Validation early stopping with patience 20 and restoration of the best iteration	Official CatBoost implementation; complete software environment and configuration released
MLP	Reporter-specific; one multi-output network per reporter	Hidden widths 256 and 128; Linear–GELU–LayerNorm blocks	Mean squared error over the standardized observed reporter block	AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch 1,024; up to 100 epochs, patience 10; lowest validation MSE restored	Released study implementation and machine-readable configuration
TabM	Reporter-specific; one multi-output model per reporter	Three 256-wide blocks, parameter-efficient ensemble size 16 and dropout 0.05[4]	Observed-only squared error over the reporter endpoint block; ensemble predictions averaged before export	AdamW, learning rate 10^{-3}, weight decay 3×10^{-4}; lowest validation MSE restored	Official TabM implementation; source revision and configuration released
scButterfly	Reporter-specific paired phase–reporter model	Phase encoder 256–128, target encoder 128–128, latent dimension 128, dropout 0.1 and phase masking 0.5[5]	Paired reconstruction, translation, KL and adversarial objectives; unavailable endpoints masked	100 phase-pretraining, 100 target-pretraining and 200 joint epochs; 50-epoch KL warm-up; best validation state restored	Public scButterfly implementation with released sparse-target adapter
ResMLP	Shared across source reporters with reporter-specific output heads	Width 512, expansion 1,024, four residual blocks, head width 128 and dropout 0.05	Reporter-balanced observed-only MSE, equalized over endpoints and active reporters	AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}; 100 epochs, two-epoch warm-up, cosine decay and EMA 0.999	Released study implementation and machine-readable configuration
MultiTab	Shared across the source-reporter panel with reporter-specific tokens and heads	Token dimension 32, four attention heads, two blocks, feed-forward width 128, head width 64 and dropout 0.05	Reporter-balanced observed-only MSE as for ResMLP	AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}; 100 epochs, two-epoch warm-up, cosine decay and EMA 0.999	Released study implementation and machine-readable configuration
MIDAS	Shared across all source reporters represented as sparse modalities	Biological latent 128, technical latent 32, modality and shared widths 768, two-layer decoder, dropout 0.1[6]	Phase and observed-only reporter reconstruction, biological and technical KL terms and modality alignment (weight 50)	AdamW, learning rate 10^{-4}, weight decay 0.01; 100 joint epochs; lowest mean reporter validation MSE restored	Public MIDAS implementation with released sparse-modality adapter
scPair	Reporter-specific paired phase–reporter model	Encoder widths 256–64, cross-modal width 128 and dropout 0.1[7]	Sum of phase reconstruction, reporter reconstruction and bidirectional translation MSE on complete reporter rows	AdamW, learning rate 10^{-4}, weight decay 10^{-4}, batch 16,384; 40 epochs of 64 steps; best validation state restored	Public scPair modules with released paired-modality adapter; EMA and top-three checkpoint averages archived

Supplementary Table 4: **Calibration rules for task-use decisions.** All gates use calibration perturbations; evaluation perturbations supply the independent task outcome. Here r is the resource-specific recovery correlation, ρ is rank of response magnitudes, v is predicted/measured magnitude variance, h is top-budget recall, and q is reference-repeat signed correlation. OPS full-response does not include q.

Rule	Calibration gate	Application
Accept all	No gate	OPS, Cell Health, LINCS
Recovery	$r \geq 0.70$	All three
Rank	$\rho \geq 0.70$	All three
Rank + hit	$\rho \geq 0.70$ and $h \geq 0.50$	OPS
Full response	$r, \rho \geq 0.70$, $0.50 \leq v \leq 1.50$, $h \geq 0.60$	OPS; task-budget hit gate
Original-budget full	Same, with h at 5%	OPS sensitivity
Task	$h \geq 0.60$	Cell Health, LINCS
Task + repeat	$h \geq 0.60$, $q \geq 0.30$	Cell Health, LINCS
Task + repeat + rank	$h \geq 0.60$, $q \geq 0.30$, $\rho \geq 0.70$	LINCS

Supplementary Table 5: **Decision units, partitions and random-selection references.** Candidate counts and budgets refer to the primary evaluation, not pooled observations across assays. Random expected recall is k/N for selecting k of N candidates.

	OPS	Cell Health	LINCS
Input / target	Phase features / fluorescent reporter profiles	Fluorescent morphology / cell-health measurements	Fluorescent morphology / landmark RNA
Paired unit	Same cell, aggregated by gene–screen	Guide within cell context; separate assay wells	Chemical condition; separate assay plates
Acquisition context	A549, matched phase/reporter images	ES2 and HCC44, separate health panels	A549, images at 48 h and RNA at 24 h
Decision case	Reporter–outer fold	Endpoint–cell context	RNA probe–predictor
Primary cases	260 (52×5)	20 (10×2)	978 per predictor
Predictor	Fixed MLP predictions; three sensitivities	Fixed Ridge per case	Ridge and MLP
Training	Three of five original gene folds	25 genes	775 compounds
Model validation	Next cyclic fold	Fixed penalty	Training-only splits
Decision calibration	91–111 genes per fold	12 genes	257 compounds
Evaluation candidates	92–111 genes per fold	12 genes	260 compounds
Evaluation hits	$\lceil 0.20N \rceil$	4	26
Required retained hits	$\lceil 0.60k \rceil$	3	16
Random expected recall	$k/N \approx 0.20$	1/3	0.10
Repeat evidence	Within-screen cell halves; sensitivity cohort	Two endpoint plates	Odd/even X-plate groups
Input requirement	Paired fluorescent targets for training	Stained images	Stained images
Allocation sensitivity	20 allocations without refitting	20 gene partitions	10 Ridge chemical-group partitions

Supplementary Data

Numerical evidence for measurement sufficiency and biological use. One data archive contains OPS predictor-specific metrics and assignments, calibration–evaluation decisions, gene-response magnitudes, hit selections and functional-enrichment records, and OASIS, Cell Health and LINCS task evidence. The external data include OASIS well predictions and compound losses, Cell Health guide/gene predictions and reference plates, and complete primary LINCS predictions and RNA references. Metrics, decisions and

partition records are retained for every reported allocation. The biological-case extension adds matched-cell and acquisition summaries, complete functional-member ranks, equal-cell assay-choice metrics, new-screen and joint gene/screen decision records, and the LDH compound/dose results. The archive includes an index, field definitions, source attribution and checksums.

Supplementary Methods

S1. OPS preprocessing, prediction states and aggregation

Preprocessing used the source atlas's same-cell features[8]. Phase variables finite in fewer than 80% of training cells were removed. Remaining non-finite entries were replaced by the training median, variables with training standard deviation at most 10^{-8} were removed, and retained variables were centred and scaled using training means and standard deviations. A cell was excluded for a target if any endpoint in its observed target block was non-finite. Target endpoints were standardized within target using training means and standard deviations, with invariant endpoints removed. Shared models masked unobserved endpoints from both the loss numerator and denominator.

Field assignments used a global physical screen–well–tile hash, including cells linked to more than one target. Target-level scores averaged five folds or, for Whole screen, eligible destination directions. Knockout-response scores for target–screen assays were averaged over folds; screen summaries then averaged member assays equally. The Field/Gene cell-level screen and assay views reweighted target-level cell scores by assay membership: 99 memberships represented 52 distinct target scores. Whole-screen cell scores were evaluated directly in destination cells.

The benchmark used validation-selected single-model states. Response replacement used parameter-averaged checkpoints for TabM, scButterfly, ResMLP, MultiTab, MIDAS and scPair, together with the benchmark's fitted Ridge, GBDT, CatBoost and MLP states. The element-wise median combined the ten aligned response vectors within target, fold and gene–screen before equal-weight averaging across screens. Method-score medians summarized the benchmark; ensemble recoverability was calculated from ensemble predictions. Prediction states are recorded in Supplementary Data and implementations in Supplementary Table 3.

Method comparisons paired targets against the prespecified target-specific MLP reference. Two-sided Wilcoxon signed-rank tests used 52 targets for Field/Gene and 34 for Whole screen. Distinct differences required Holm-adjusted $P < 0.05$ across 54 comparisons[9] and an absolute median paired difference of at least 0.02. Screen and assay summaries were descriptive because they share targets and acquisition structure. The descriptive recovery strata ($r < 0.60$, $0.60 \leq r < 0.75$, $r \geq 0.75$) were separate from the measurement-sufficiency tiers.

S2. Sensitivity to the cross-fold response scale

The prediction models standardized each target endpoint using the training cells of its outer partition. Endpoint-wise correlations are invariant to this scaling, but combining response vectors across partitions can change relative endpoint weights because training-set standard deviations differ. The primary response-profile analyses used a common control reference:

$$c_{re} = \left\{ \frac{1}{N_{r0}} \sum_{s} \sum_{i \in I_{rs0}} (y^{\mathrm{raw}}_{ire} - \overline{y}^{\mathrm{raw}}_{rs0e})^2 \right\}^{1/2}, \tag{S1}$$

where I_{rs0} contains unique NTC cells for target r in screen s, $N_{r0} = \sum_s |I_{rs0}|$, and the screen-specific control mean is subtracted before pooling. The scale uses control measurements for evaluation, independently of model fitting. All 1,604 scales were finite and exceeded 10^{-8}. A response standardized with training-fold standard deviation $\sigma^{(f)}_{re}$ was converted as $\Delta y^{\mathrm{common}} = \Delta y^{(f)} \sigma^{(f)}_{re} / c_{re}$, with the same conversion applied to predictions.

Here c_{re} is the common control scale for target r and endpoint e; s indexes screens and i indexes unique NTC cells. The subscript 0 denotes NTCs, y^{raw}_{ire} is the unstandardized measurement and $\overline{y}^{\mathrm{raw}}_{rs0e}$ is its screen-specific NTC mean. The denominator N_{r0} is the total control-cell count, so the scale uses the population

mean square. In the conversion, f indexes the outer fold, $\sigma_{re}^{(f)}$ is the training-cell standard deviation, and $\Delta y^{(f)}$ and Δy^{common} denote responses in fold-standardized and common-control coordinates, respectively.

Response-profile, replacement and tier calculations were repeated in the original fold-standardized coordinates. ER NCLN was the only target to change tier: unresolved on the fold-standardized scale and a ranking proxy on the primary common scale. Recovery and reliability were unchanged. Paired values accompany the source data.

S3. Reference consistency, screen transfer and local phenotype discrepancy

Within-screen reliability followed the NTC-normalized split-half procedure in Methods, with a deterministic phase-row hash and 1,000 gene-bootstrap resamples. Guide-half consistency used three fixed partitions for genes with at least four guides. Endpoints were standardized within well, then averaged equally over each guide subset and wells. One hundred random cell-half splits within gene and well supplied a sampling reference. Cross-screen consistency was the median correlation of knockout-magnitude vectors across screen pairs sharing a target. For pRb and NPM1, three well pairs, three guide splits and 100 cell splits were retained separately.

A complete-case linear model related recovery to within-screen reliability, guide-half consistency and $\log_{10}$ linked-cell coverage for 35 targets. Its R^2 was compared with 3,000 target-level permutations of recovery. Cross-screen consistency was analysed separately because it was available for 34 repeated targets.

For transfer, source controls pooled all other screens measuring the target; source and destination each had a deterministic cap of 20,000 cells. The control shift for d valid phase features or target endpoints was

$$D_{r,s\to t}^{\text{NTC}} = \left\{ \frac{1}{d} \sum_{e=1}^{d} \left(\frac{\mu_{rt0e} - \mu_{rs0e}}{\sigma_{rs0e}} \right)^2 \right\}^{1/2}, \tag{S2}$$

with s denoting pooled sources and t the destination. Non-finite coordinates and source standard deviations at most 10^{-8} were excluded. Transfer loss was the median-across-methods Gene recovery minus Whole-screen recovery. The 81 directions were displayed individually; inference used medians within 34 targets and 3,000 target-level permutations. Phase-control shift can be measured before acquiring the destination target, whereas target-control shift requires that measurement. The domain-sensitive flag identified targets with cross-screen target consistency at or above the cohort median and a Whole-screen-minus-Gene recovery difference in the lowest quartile. Destination-specific recovery and rank were separately classified as always above 0.70, always below 0.70 or crossing 0.70.

Here $D_{r,s\to t}^{\text{NTC}}$ is the root-mean-square standardized NTC mean shift for target r; e indexes the d retained phase features or target endpoints. The quantities μ_{rs0e} and μ_{rt0e} are pooled-source and destination NTC means, respectively, σ_{rs0e} is the source NTC standard deviation, and the subscript 0 denotes NTCs.

Local phenotype discrepancy used FeRhoNox in two screens and pRb in one, with 80,975 linked non-control cells. Groups required at least six cells. Each cell's nearest neighbour was selected using root-mean-square distance in 172 phase features standardized within gene–screen, requiring a different guide when available. Target endpoints were centred on the well-specific NTC median, scaled by its median absolute deviation (MAD) and clipped to $[-10, 10]$. For each gene, median root-mean-square target discrepancy among directed nearest pairs was divided by the median over all eligible unordered pairs. A 32-dimensional principal-component representation fitted to raw phase thumbnails within target and screen supplied a second morphology representation. Microscopy examples used outcome-independent medoid selection.

S4. Significant and ranked functional evidence

Functional sensitivity analyses used fixed ensemble predictions on the common response scale, with the same Gene Ontology[10] and Complex Portal[11] annotations for measured and predicted lists. Enrichment tests, Benjamini–Hochberg adjustment[12] and empty-set definitions followed Methods. Budgets were 5%, 10% and 20%; ranked-list lengths were 5, 10 and 20 terms.

The cohort-relative downstream-retention index used 1,001 source-atlas labels, including NTC, and source-order resolution of ranking ties. The main functional panels use the 1,000-gene background and de-

terministic full-term-name ties. Changing only the background left all 156 primary ranked-list comparisons unchanged; full-term-name tie resolution changed twenty. Complete scores, both sensitivity comparisons and redundancy-aware and member-matched summaries accompany Supplementary Data.

S5. Predictor-specific measurement criteria and prediction-state sensitivity

Predictor dependence used the ten methods' predictions for 52 targets and five Gene partitions, aligned by target, endpoint and gene–screen identity on the common control scale (Eq. S1). Recoverability was the equal-endpoint Pearson correlation within fold, averaged over five folds. Response vectors were averaged equally across screens within each gene, and their Euclidean magnitudes supplied rank fidelity, the predicted-to-observed variance ratio and recall of the strongest 50 of 1,000 genes. No top-50 boundary had tied magnitudes. Measured target reliability, partitions, thresholds and rule precedence were held fixed. The primary comparison used the same prediction states as the response-replacement ensemble; stability summaries used the 43 targets with reliability at least 0.30.

Sensitivity analyses repeated the response-level criteria with single-model benchmark states and compared recovery calculated from ensemble predictions with the median of method scores. The latter substitution changed ATG101 from quantitative to ranking: ensemble recovery was 0.714 and the median method score was 0.697. Changing the ensemble from checkpoint-averaged to single-model states changed only SRRM2: its magnitude-variance ratio was 0.495 versus 0.504, crossing the fixed 0.50 boundary. Both complete predictor matrices accompany Supplementary Data. Functional-term retention used the fixed ensemble.

S6. Sensitivity of the reporter tiers to the threshold values

Each of the eight thresholds was varied independently in both directions under two schemes: an absolute change of 0.05 in the metric's units and a relative change of 10% of the threshold value. All eight thresholds were also varied together, yielding 36 scenarios with fixed target evidence. For the two-sided variance-ratio interval, greater stringency moved both limits towards 1.0 and lower stringency moved them apart.

Rule precedence remained fixed: ranking-proxy eligibility was evaluated before the measurement-required rule, including when recovery was below 0.60.

S7. Cohort-relative downstream-retention modelling

Let $P_{52}(z_r)$ be the average-tie percentile rank of a quantity among the 52 targets. The downstream-retention index was

$$U_r = \frac{1}{4}\left[P_{52}(H_r) + P_{52}(J_r^{\mathrm{BP}}) + P_{52}(J_r^{\mathrm{CC}}) + P_{52}(J_r^{\mathrm{CP}})\right], \tag{S3}$$

where H is top-5% hit recall and J denotes the source-order top-ten overlaps for biological process, cellular component and Complex Portal described above. Magnitude rank was evaluated separately as a predictor of this downstream-retention index. Association with recovery used Spearman correlation; utility values were rounded to 12 decimal places before ranking to preserve percentile ties.

In this notation, r indexes targets, z_r denotes a component value and $P_{52}(z_r)$ is its ascending average-tie rank divided by 52. U_r is the cohort-relative downstream-retention index, H_r is top-5% hit recall, and J_r^{BP}, J_r^{CC} and J_r^{CP} are top-ten Jaccard overlaps for Gene Ontology biological-process terms, Gene Ontology cellular-component terms and Complex Portal complexes, respectively.

Ridge regression predicted retention from recovery, magnitude rank, within-screen reliability and magnitude-variance fidelity, defined as $\exp\{-|\log Q_r^{\mathrm{var}}|\}$ using Eq. 4. Strict nested leave-one-target-out evaluation constructed every target component from training-target empirical distributions. In each outer split, the 51 training targets supplied average-tie percentile ranks; a test component was mapped to that training distribution, with values outside its range mapped to 0 or 1 and ties assigned the mean training percentile. Inner leave-one-target-out validation repeated this training-only construction and selected $\alpha \in \{0.01, 0.1, 1, 10, 100\}$ by squared error. The excluded target's measured downstream outcomes supplied only its evaluation value. Prediction was assessed by outer-loop Spearman correlation and R^2; univariate comparisons used the same outer targets.

The quantity Q_r^{var} is the predicted-to-measured magnitude-variance ratio in Eq. 4; log denotes the natural logarithm, α is the Ridge L2 penalty and R^2 is the coefficient of determination.

In this complementary analysis, response fidelity and reference reliability predicted variation in a downstream-retention index combining strong-hit recall and ranked-term overlap (Supplementary Figure 2b,c). Nested leave-one-reporter-out prediction with Ridge reached $\rho = 0.786$ and $R^2 = 0.587$ (Supplementary Figure 7b). This related the response-level evidence to the extent of biological-analysis retention across targets.

S8. Retrospective task-decision validation

Complete rules and resource-specific budgets are given in Supplementary Tables 4 and 5. Outcome denominators follow Methods: missing calibration metrics cannot satisfy a criterion, missing evaluation outcomes are reported separately, and zero acceptance gives an undefined false-support fraction.

OPS. The primary calibration/evaluation gene counts were 105/106, 93/94, 91/92, 98/99 and 111/111 across the five original outer test folds. All guides and screens for a gene stayed together. Each case used predictions from its original fold model, converted to the common control scale before equal-weight averaging across screens. Seeds 20261008–20261027 changed only calibration/evaluation membership. Secondary analyses used 5% or 10% selection budgets, success cut-offs of 0.50, 0.70 or 0.80, and a full-response rule using top-5% calibration recall with the primary top-20% evaluation task. Supported-use gain was the difference in accepted successes divided by all cases.

For matched coverage, single-criterion rules ranked targets by calibration recovery, rank or task recall. Combined rules used the minimum normalized criterion ratio:

$$m_{\mathrm{full}} = \min\left\{\frac{r}{0.70}, \frac{\rho}{0.70}, \frac{v}{0.50}, \frac{1.50}{v}, \frac{h_b}{0.60}\right\},$$
$$m_{\mathrm{rank+hit}} = \min\left\{\frac{\rho}{0.70}, \frac{h_b}{0.50}\right\}.$$

Here r is recovery, ρ magnitude rank, v the predicted/measured magnitude-variance ratio and h_b hit recall at selection budget b. A ratio of at least one satisfies its criterion. Within each predictor–fold–seed, the highest-scoring $\lceil cn \rceil$ targets were accepted for $c = 0.25, 0.50, 0.75, 1.00$, with target identifier breaking ties. Two thousand paired target-cluster bootstrap draws kept all folds and repeated allocations together (Supplementary Figure 6a).

Here ρ is Spearman correlation of response magnitudes, and the primary selection budget is $b = 20\%$. In the matched-coverage selection, c is the acceptance fraction, n the number of eligible targets and $\lceil cn \rceil$ the accepted count rounded upwards.

The calibration scores m_{full} and $m_{\mathrm{rank+hit}}$ correspond to the full-response and rank-plus-hit rules, respectively; min selects the smallest normalized criterion ratio.

Cell Health. The Cell Health resource[13] supplied ten endpoints in each of ES2 and HCC44: G1, S, G2 and mitotic fractions; dead, early-apoptotic and late-apoptotic fractions; live-cell count; ROS; and H2AX. Each context had 49 genes and 100 targeting guides; 100 guides had finite targets on both reference plates for nine endpoints and 94 for ROS. A549 was excluded because duplicate mappings were unresolved. Wells were averaged by guide and plate, centred on the median of ten infected non-targeting guides and averaged equally across plates; empty-well controls were excluded. The two assays were paired by guide and context, not by cell or well.

Predictors used 1,783 numeric Cell Painting features, excluding identifiers. Imputation, constant-feature filtering and input/target standardization used training genes only; targets were not imputed. A separate Ridge model with $\alpha = 100$ and equal total weight per gene was fitted for each endpoint–context pair. Predictions were returned to control-relative units. Twenty seeds (20261008–20261027) defined allocation sensitivity, keeping all guides and both contexts for each gene together. Family-balanced summaries retained the five

prespecified endpoint families. Figure 6c uses symmetric-log axes, linear within ± 0.01; metrics and selection use untransformed fraction changes.

LINCS. The LINCS-Pilot1 resource[14] matched A549 Cell Painting at 48 h and L1000 at 24 h by chemical identity and actual 10 μM condition, with at least two plates per assay. DMSO, MG132 and bortezomib were excluded as prediction targets. The cohort contained 1,292 compounds in 1,286 chemical-identity groups, defined by the first 14 InChIKey characters, falling back to compound identifiers. All doses and replicates of a group belonged to one partition. RNA reference A combined odd X-acquisition groups and B even groups. The consensus included A and B, so these three comparisons share reference observations. Compound–plate means were averaged equally across plates, without cross-assay well pairing.

Image features used plate-DMSO median/MAD normalization and RNA used plate-DMSO centring. There were 1,663 shared numeric morphology features after exclusion of coordinates, identifiers and non-shared columns. Training compounds alone determined median imputation, 1st/99th-percentile clipping, constant-feature filtering and input/target standardization. Ridge selected $\alpha \in \{10, 100, 1000, 10000\}$ by three-fold chemical-group cross-validation within training. The MLP used 256/128 ReLU units, dropout 0.1, AdamW with learning rate 0.001 and weight decay 0.01, batch size 128, at most 200 epochs and stopping patience 25. A training-only 20% validation partition selected the epoch before refitting on all training compounds.

All 978 landmark probes were retained: 957 uniquely annotated and 21 ambiguously annotated. Task-plus-repeat used calibration A/B signed Pearson ≥ 0.30; repeat-supported evaluation required success against each reference separately. Allocation sensitivity used ten fixed Ridge partitions. Two hundred joint compound-group bootstrap draws preserved the entire probe panel and recomputed evaluation hit sets, with calibration acceptance fixed.

S9. Same-cell information controls

For the pairing-derangement controls, training and validation target vectors were reassigned separately within gene–screen groups; test pairs remained intact. Covariate-matched derangement greedily paired cells without replacement by distance in standardized log cell area, log field–tile cell density and eccentricity. Density counted linked cells in the same screen, well and tile. Matching covariates were scaled over the target cohort using morphology alone. Pairs exchanged target vectors; an unmatched odd cell and the closest existing pair formed a randomly oriented three-cycle, and singletons were unchanged. Feature-schema annotations defined the size–shape subset.

For residual prediction, each gene–screen group was divided into $\min(5, n)$ near-equal blocks within training, validation and test separately. The target for cell i in block b was

$$\widetilde{y}_{ire} = y_{ire} - \frac{1}{|I^{(-b)}_{r,g(i),s(i)}|} \sum_{j \in I^{(-b)}_{r,g(i),s(i)}} y_{jre}, \tag{S4}$$

where the reference contained peer cells outside block b from the same group; singleton residuals were zero. The MLP was fitted directly to these residuals. Test residuals used measured targets from peer test cells, assessing cell-level association conditional on the measured group response. Scores averaged five folds within target, and exact-pair advantages used paired target-level differences and bootstrap intervals.

Here n is the number of cells in the gene–screen group within the relevant training, validation or test partition; i and j index cells, r the target and e the endpoint. The value y_{ire} is the measured endpoint, $\widetilde{y}_{ire}$ its residual, and $g(i)$ and $s(i)$ are the gene and screen of cell i. The set $I^{(-b)}_{r,g(i),s(i)}$ contains peer cells outside block b within that same partition and group, and vertical bars denote its size.

We also predicted each cell's deviation from its gene–screen mean, estimating that mean from other cells and excluding the cell's own evaluation block (Supplementary Figure 7a). Correlations for these residual phenotypes were positive for all 52 reporters, with a median of 0.662, compared with 0.757 for exact-cell

prediction. This complementary analysis showed that morphology was informative about differences in targeted state among cells receiving the same perturbation.

We also examined response recovery and ranking when predictions were applied to separately acquired screens. Across 810 predictor–transfer combinations, median response Pearson correlation was 0.715 and magnitude-rank correlation was 0.723, with variation among destination screens (Supplementary Figure 7c). Thus, the same-cell information captured by morphology could support response prediction, but its preservation across screens and its suitability for quantitative comparison required separate assessment.

S10. Biological summaries and phase–target overlap

Source annotations assigned targets to ten biological systems[8]. Endpoint families described intensity level, extrema or range, dispersion, integrated intensity, sub-object distribution and spatial displacement. Family summaries first averaged endpoints within target and family, then summarized these target-level values. Biological-system summaries likewise used targets as the unit.

Perturbation distinctiveness used the source study's copairs mean average precision (mAP), averaging 1,001 perturbation labels, including non-targeting control (NTC). Profiles from different guides targeting the same gene were positives and other perturbations were negatives, with cosine distance. Phase–target overlap used a $53 \times 1{,}001$ matrix M of phase and 52 target distinctiveness profiles. For profile r and perturbation g, $\mathrm{TF}_{rg} = M_{rg} / \sum_h M_{rh}$ and $\mathrm{IDF}_g = \log\{54/(\mathrm{df}_g + 1)\}$, where df_g counted profiles exceeding the across-profile mean for perturbation g. Overlap was Pearson correlation between the weighted phase and target profiles. A source-study subsampling draw capped each guide at 300 cells and repeated guide aggregation and mAP calculation. Associations with directed prediction recovery used Spearman correlation over 52 targets and 3,000 target-bootstrap resamples.

Here M_{rg} is the mAP for profile r and perturbation label g, with r indexing the 53 profiles and the summation index h running over all 1,001 labels. TF denotes term frequency and IDF inverse document frequency; the weighted coordinate is $\mathrm{TF}_{rg}\mathrm{IDF}_g$. The count df_g is taken over the 53 profiles, log is the natural logarithm, and 54 is the profile count plus one.

S11. Endpoint aggregation in the independent applications

The CellProfiler-based OASIS analysis used 21,695 wells from 66 plates and 1,085 compounds in a commercial five-donor primary-hepatocyte pool[15]. Inputs retained 808 brightfield-attributable features, excluding 120 brightfield–fluorescence cross-channel, 4,512 fluorescence-attributable and 200 channel-free segmentation features. The deposited object masks were derived from fluorescence images. Each metabolic-activity or lactate-dehydrogenase-release target supplied one endpoint. Compound-level fivefold partitions kept all wells and concentrations together. Training-only median imputation and standardization preceded Ridge with $\alpha = 1$. Recovery was Pearson correlation of observed and predicted compound–plate means of the deposited plate-normalized endpoint, excluding controls. Response metrics first subtracted the within-fold plate-control mean and then averaged plates equally within compound. Each fidelity value was the median of five folds. Reliability correlated compound responses from seeded disjoint halves of replicate wells pooled across plates and concentrations.

PERISCOPE[16] included nine A549-TetR-Cas9 plates, with 1,080,000 cells selected by a fixed cap of the first 120,000 streamed rows per plate and all 704,771 deposited gene–guide–plate median profiles. Predictors retained 2,508 features exclusive to DAPI, ConA, phalloidin and WGA; targets were 823 anti-TOMM20 features. Sixty predictor–target cross-channel and 375 channel-free features were excluded. Segmentation used images including anti-TOMM20. Non-control genes were divided into five folds, the next cyclic fold supplied validation, and NTCs remained in training. Validation endpoint-standardized MSE selected the Ridge penalty before refitting on the outer training set. Test responses averaged cells or profiles by gene and plate and subtracted the same-plate NTC mean.

PERISCOPE recovery pooled gene–plate–endpoint response coordinates for Pearson correlation after dividing observed and predicted responses by each endpoint's observed response standard deviation. An endpoint required positive spread exceeding 10^{-3} times the median endpoint spread. This differs from OPS

equal-endpoint mean correlation (Eq. 2). Magnitudes first averaged responses equally across plates and standardized endpoints by observed across-gene standard deviations within fold. Both evaluation scales were shared by observations and predictions and did not enter fitting. Reliability split each eligible gene's guides into disjoint halves within plate, centred on plate controls, standardized by the full-guide response spread and took the median of nine plate correlations; genes with fewer than two guides and negligible-spread endpoints were excluded. Numerical tier thresholds were retained from OPS, with aggregation and repeat definitions matched to each resource.

S12. Metabolic-activity prioritization: repeated measurements, controls and uncertainty

The focused OASIS analysis used all 768 released brightfield DINO features[15], derived from stained, fixed-cell acquisitions. The 868 development compounds supplied 13,157 wells and the 217 test compounds supplied 3,300 wells, grouping all doses and production sources by compound. The production sources used for cross-source comparisons were represented in both development and test sets. All 868 development compounds entered fitting; Hycanthone's seven rather than eight assayed concentrations excluded it only from compound-level development scoring, leaving 867 scores. Features were centred on same-plate DMSO means before training-only median imputation and standardization. Ridge used $\alpha = 100$ with DINO features alone. Measured and predicted responses were separately centred on their same-plate DMSO means.

The target was the normalized RealTime-Glo MT signal in `Metadata_mtt_normalized`. Loss was the negative signed change relative to plate controls, in normalized-signal units rather than percentages of viable or dead cells. Within-compound aggregation matched actual concentrations before averaging sources and then the eight concentration positions. All 217 test compounds had eight positions across the union of sources. Cross-source comparisons used the 94 compounds with all eight matched concentrations in both sources: 38 between sources 26 and 27, and 56 between sources 26 and 30. The other 123 comprised 31 source-25/26 and 92 source-25/27 pairs with incomplete source-25 dose coverage. Source roles were assigned by lexical order within compound; both pooled and actual source-pair directions accompany Supplementary Data.

The histogram gradient-boosting comparator used the same 768 features, with 100 iterations, learning rate 0.1, at most 15 leaves, L2 regularization 10 and no early stopping. Matched Ridge and gradient-boosting models used either same-plate-control-centred cell counts from fluorescence-derived segmentation or acquisition metadata. Metadata comprised numeric row, column and dose position, plus one-hot-encoded plate and production source, with training-only category fitting and unknown-category handling. DMSO dose position was set to zero.

At each budget, measured and predicted top sets had the same size, $k = \lceil fN \rceil$; overlap divided by k was both recall and precision. Compound identifiers resolved ties. Bootstrap resampling drew compounds 2,000 times, retaining doses and source roles, distinguishing repeated draws by unique instance identifiers and recomputing the top sets. Paired comparisons used the same draws for prediction-to-other-source and measured-source-to-other-source recall. Intervals quantify compound-sampling uncertainty conditional on the fitted models and observed production sources.

In $k = \lceil fN \rceil$, f is the selection fraction, N the number of compounds in the evaluation population and k the selected count, with the ceiling rounding upwards.

For the full 217-compound analysis, 2,000 permutations reassigned test predictions among wells within plate while retaining the fitted models. Dose position was constant within plate, giving 61 plate–dose groups, and all test wells were eligible for permutation. Exact hypergeometric distributions described random-selection overlap separately for each population and budget. These random-selection references are distinct from the compound-bootstrap intervals for model performance.

S13. Permuted-input negative control

Input profiles were permuted within plate at well level for CellProfiler-based OASIS and guide-profile level for PERISCOPE, retaining target measurements, perturbation labels and control status. OASIS retained Ridge $\alpha = 1$. PERISCOPE used a fixed penalty-to-training-row ratio of $\alpha/n = 0.01$, selected in four of the five main-analysis folds. Supplementary Table 1 reports the nested-selected main analysis and Supplementary Note 1

the fixed-penalty permutation. Reference reliability used the unchanged target measurements.

Here α is the Ridge L2 penalty and n the number of training rows, so α/n is the penalty per training row.

S14. Candidate effects and acquisition support in the biological cases

The biological cases are descriptive follow-ups of the original response-replacement results. The MTOR-related comparison retained MTOR, RPTOR, RICTOR, TSC1, TSC2 and LAMTOR1–4 in LysoTracker screens OPS0044/0055/0056 and LAMP1 screens OPS0011/0012. We compared the mean-intensity endpoint of each assay, using its original common-control scale. Acquisition-well means required at least five finite knockout cells and twenty non-targeting-control cells; each well's own control mean was subtracted. All eligible wells were retained, including screens with only two eligible MTOR wells. LysoTracker and LAMP1 measurements came from different screens and describe different assay signals.

For DDX18, the exact phase-row identifiers linked p53 and p21 measurements in OPS0077. The paired extract contained 52 knockout cells and 4,104 NTC cells; the knockout cells occupied three wells (18/17/17 cells) and four sgRNA sequences (10/14/14/14 cells). We retained all cells and guides, including the weakest guide response. Cell distributions, mean-intensity shifts and sensitivity to cell extremes were compared with the fixed gene-level ensemble prediction. The blue prediction is a gene mean, not a prediction specific to each acquisition well or displayed cell. Multivariate ranks used all original endpoints and the same 1,000 knockout genes as the primary top-50 analysis. The individual-predictor comparison retained all ten original prediction states.

The p21 functional comparison used all seven assayed members of GO:0000470 (maturation of LSU-rRNA), including the two members absent from the measured top 50. We retained measured/predicted ranks and candidate membership separately. Three acquisition-well lists and two lexicographically alternated sgRNA halves supplied repeat-specific membership; each eligible guide required at least five cells. The pooled functional association used the existing hypergeometric test, 1,000-gene universe and BH adjustment across all eligible biological-process terms. It was not assumed that pooled enrichment significance also held in every well or guide half.

S15. Equal-cell assay choices and use decisions in new screens

The equal-cell analysis retained all 75 previously eligible three-well target–screen combinations (34 targets, 63 physical screens), using the first three eligible wells in lexicographic order. Original Whole-screen models excluded the entire destination from fitting and validation, while knockout genes could have appeared in other source screens. We reconstructed the original source-only Ridge operator from its saved preprocessing, original fitting-cell assignments, source-only feature-clipping limits and previously selected penalty. No new penalty search was performed. All 75 deposited response profiles were reproduced; the maximum absolute difference was 1.9×10^{-6} common-control standard deviations. Raw-cell checks independently rebuilt the matched well means and retained cell-identity hashes for all 225 wells.

For each reference-well rotation, eligible genes had at least five knockout cells in every required well, with at least twenty NTC cells per well. The prediction option averaged cell predictions within each of the other two wells, subtracted the predicted NTC mean of that well and averaged the two responses equally. The direct option applied identical cell membership, NTC subtraction and well weights to measured fluorescence. The third well supplied the reference response. Endpoint scales were the original target-wide NTC standard deviations. Calibration and evaluation genes followed the existing 20 gene-role allocations; the primary allocation used 402–498 eligible calibration and 401–502 evaluation genes across combinations.

Quantitative error followed Eq. 5. The strongest 20% of the measured reference magnitudes were selected independently within calibration and evaluation groups, rounding up. Undefined zero-energy errors were recorded as missing. The repeat-informed priority was calibration prediction error minus calibration direct-measurement error. At the primary 20% allowance, the 15 largest positive priorities received direct measurements and the rest predictions. Correlation and prediction-error comparators received exactly the same number of direct choices, prioritizing the lowest flattened calibration response correlation or the highest calibration prediction error, respectively; ties were resolved by target and physical-screen identifiers. The

random reference used the expected error under uniform allocation of the same direct allowance.

The 50-gene pilot sampled only from calibration genes; source subset 0 was primary and all twenty fixed subsets were retained. Each combination had 40–50 eligible pilot genes, and evaluation genes remained unchanged. Primary paired uncertainty used 5,000 bootstrap resamples of all 34 complete target clusters; physical-screen resampling supplied a separate sensitivity. We also retained twenty full-pilot gene allocations, all three reference-well rotations and a calibration-repeat-consistent subset ($q \geq 0.30$, 31 combinations, 22 targets, 30 screens). Its direct allowance was seven combinations. Large-intensity direction errors and candidate recall were recorded separately from multivariate quantitative error.

The screening-use extensions reused original frozen MLP/Ridge predictions and the original response rules. The new-screen analysis contained 81 target–screen contexts in 34 targets and 66 destination screens. The joint analysis contained the same contexts evaluated under five original gene folds, giving 405 model tasks; each model excluded its evaluation genes and destination screen from fitting and validation. Gene roles were separated within each corresponding original test fold, without pooling calibration labels across fitted models. Calibration/evaluation sizes in joint tasks were 91–111/92–111 genes. Original fold-specific NTC references and partition hashes were checked against raw cells. The strongest-20% screening task required evaluation recall of at least 0.60. We retained all five rules, repeat-gated sensitivities and twenty gene-role allocations; the primary summaries in Supplementary Figure 9a use accept-all, correlation and ranking without a repeat gate. Matched-acceptance comparisons preserved each original fold's accepted-case count. Confidence intervals used 5,000 whole-target or whole-screen resamples, not cells or repeated allocations.

S16. LDH prioritization using the matched OASIS image representation

The LDH extension used the same 868 development compounds, 217 evaluation compounds and 768 brightfield DINO features as the focused metabolic task[15]. Development and evaluation contained 13,157 and 3,300 wells, respectively. Ridge used $\alpha = 100$; histogram gradient boosting used 100 iterations, 15 maximum leaf nodes, learning rate 0.1, L2 regularization 10, no early stopping and random seed 20260911. Training-only preprocessing and all source/dose observations remained grouped by compound. Brightfield, fluorescence-derived cell-count and acquisition-metadata inputs were each evaluated with both fixed predictors. Independent normal-equation reconstruction reproduced Ridge predictions to a maximum difference of 1.7×10^{-14}.

The LDH response was the positive plate-DMSO-centred normalized assay signal, whereas metabolic loss was the negative change in metabolic signal. Within compound, we averaged measurements within actual source/concentration, then equally across sources and the eight assayed concentration positions. The evaluation selected the strongest eleven compounds. Development-use evidence came from the existing out-of-fold predictions; scoring required all eight concentration positions, leaving 867 development compounds. At the top-5% development budget, both brightfield-Ridge tasks recovered 34 of 44 measured candidates and passed the fixed 0.60 requirement. Evaluation recovered 10 of 11 metabolic and 6 of 11 LDH candidates; the development acceptance did not remove that evaluation failure.

The complete main-panel candidate display uses the union of measured and brightfield-predicted top-eleven sets across both assays, giving 17 compounds, with the matched cell-count selections displayed on those same rows. Candidate ordering first follows the measured metabolic-loss rank. Source-resolved curves show every available actual concentration for Abemaciclib and Fesoterodine, selected retrospectively as a retained response and an LDH omission; all other compounds, including the missed metabolic candidate and all five missed LDH candidates, remain in Supplementary Data. Source ranges describe acquisition variation and do not estimate between-donor variation. The LDH signal is not interpreted as a percentage of dead cells.

S17. Label efficiency and raw-phase representations

The label-efficiency panel comprised pS6, NPM1, SEC23A, LAMP1, 5xUPRE, TOMM20, FastAct, EEA1, WGA, Peroxi, Hoechst and G3BP1, selected before learning-curve analysis to span biological systems and full-label recovery. Hamilton largest-remainder allocation sampled 0.1%, 1% or 20% of observations within control-status-by-gene strata, independently of target values. The same fractions and sampling indices were

used across methods in training and validation; all test cells remained. ResMLP, MultiTab and MIDAS additionally used the fully labelled 40-target donor atlas, while the other seven methods used target-only supervision. Performance summaries used equal-target means and standard deviations across five Gene folds. Matched 100% values came from a separate full-label experiment on the same panel and partitions.

Raw-image comparisons used EEA1, LAMP1, TOMM20, NPM1, pS6 and G3BP1 in five Gene partitions, selected before comparison to span six biological systems and the recovery range with registered crops. Representations were the 172 engineered phase features[8], frozen DINOv2 ViT-S/14[17], frozen Cytoland VS-Cyto2D[18] and partially fine-tuned Cytoland[18]. Registered 384×384 level-0 crops were centred by the valid-pixel median and scaled by the interquartile range, with padding set to that median and no augmentation. Partial fine-tuning fixed the stem and first two stages and trained the last two stages with a 256–128 endpoint head. AdamW used batch size 48, backbone/head learning rates $10^{-5}/3 \times 10^{-4}$, two warm-up epochs followed by cosine decay, at most 25 epochs and validation stopping patience 6. Predictions were paired within target and fold, with targets as the bootstrap unit (Supplementary Figure 3b).

Supplementary Note 1: Input permutation distinguishes prediction failure from inadequate target reproducibility

Within-plate permutation of OASIS CellProfiler-derived brightfield profiles reduced recoverability from 0.83 to 0.09 for metabolic activity and from 0.67 to 0.05 for LDH release. Both targets became measurement required: reference reliability was unchanged, but the permuted profiles no longer recovered the measured responses.

For PERISCOPE, guide-profile permutation gave recoverability of -0.0004, magnitude-rank correlation of 0.022, variance ratio of 0.025 and strong-hit recall of 0.064 under the fixed-penalty control. Anti-TOMM20 remained not identifiable because guide-partition reliability was unchanged at 0.047. These controls distinguish loss of predictive information from inadequate reference reproducibility (Supplementary Fig. 4b).

Supplementary References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.