Research Square

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

Learning reproducible cellular perturbation responses from limited independent measurements

Weijing Zhao
University of Macau
Bo Li
University of Macau
Bob Zhang
bobzhang@um.edu.mo
University of Macau
Yuqi Wang
University of Macau
Shuping Zhao
Guangdong University of Technology
Chengliang Liu
University of Macau

Article

Keywords:

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11316764/v1

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

Additional Declarations: No competing interests reported.

Learning reproducible cellular perturbation responses from limited independent measurements

Weijing Zhao[1†], Bo Li[1†], Bob Zhang[1*], Yuqi Wang[1], Shuping Zhao[2], Chengliang Liu[1]

[1]Department of Artificial Intelligence, University of Macau, Taipa, Macau SAR, China.

[2]School of Computer Science, Guangdong University of Technology, China.

*Corresponding author(s). E-mail(s): bobzhang@um.edu.mo;

[†]These authors contributed equally to this work.

Abstract

Large-scale perturbation atlases are increasingly used to train models that predict cellular responses, yet responses to individual perturbations are often measured only a few times. The resulting averages are commonly used as training targets, although they contain both reproducible perturbation effects and measurement-specific variation. Here we present RePert, a framework that uses independent measurements to construct reproducible targets for cellular-response prediction. RePert learns from separate measurements of the same perturbation, combines complementary estimates using independent calibration data, and uses a second assay to refine an existing response estimate. In two Cell Painting datasets and one gene-expression dataset, RePert estimates agreed more closely with held-out measurements than averages of the same measurements. In Cell Painting, they also better distinguished the correct compound from matched alternatives. In BBBC047, these estimates outperformed larger neural networks given identical inputs. Matched measurements from a second assay provided additional compound-specific information, with smaller gains when more Cell Painting measurements were available. The recovered responses also better preserved independently supported morphological features, response directions and patterns across doses. As training targets, they improved prediction of unseen compounds in BBBC047, whereas in sci-Plex3, with two measurements per condition, the effect depended on the prediction setting. Target construction is therefore a design dimension of perturbation modelling, distinct from model architecture and capacity.

Introduction

Large-scale perturbation atlases are becoming a foundation for predicting cellular responses and building virtual cells [1]. High-throughput assays such as Cell Painting (CP) [2] and

gene expression (GE) [3] now profile tens of thousands to more than one hundred thousand chemical and genetic perturbations [4, 5], providing models with broad coverage of how interventions alter cellular state and enabling prediction of responses to previously unseen perturbations [6–8]. This scale, however, has been achieved primarily through breadth rather than deep measurement of individual conditions. BBBC047, for example, profiled 30,616 compounds, each in quadruplicate [9], and JUMP covers about 136,000 chemical and genetic perturbations, typically with five replicates each [5]. Perturbation modelling therefore increasingly operates in a regime in which models observe many distinct interventions, but the response assigned to each intervention is estimated from only a few independent acquisitions.

Limited replication makes the learning target itself uncertain. An observed profile combines perturbation-induced changes with measurement-specific variation arising from plates, batches, acquisition procedures and cellular state [10, 11]. Averaging a small number of replicates reduces this variation but still yields an uncertain estimate, part of which may not recur in another independent experiment. Nevertheless, observed or averaged measurements are commonly used directly as training and evaluation targets for perturbation-response models [7, 8, 12]. At the same time, much of the field has focused on model architecture, scale and representation [1, 13], even as recent benchmarks have shown that simple baselines such as the training-set mean or a linear model often match deep-learning and foundation models, particularly when training data are limited [14, 15], and that apparently strong performance can partly reflect systematic variation shared across perturbations [16]. These observations motivate a more basic question that precedes model capacity: which components of an observed cellular response are reproducible enough to be worth learning?

Independent measurements provide an operational answer to this question. Replicates are usually averaged into consensus profiles [4, 17] or used to assess quality and reproducibility [10, 18], but they can also participate directly in defining the learning target. If an estimate constructed from a subset of measurements predicts a separately acquired measurement of the same perturbation that was not used to construct it, the recovered structure is supported across independent acquisitions. This principle is related to learning from independent noisy observations [19, 20], but perturbation experiments provide additional structure: repeated measurements can supply both supervision for learning response components that recur across acquisitions and independent evidence for calibrating alternative estimators. Independent evidence need not come from the same assay. CP and GE capture complementary aspects of cellular state [4, 17, 21], so once an estimate of a target-modality response is available, another assay can be treated as additional evidence. Its value should then be judged not simply by the performance of a jointly fitted multimodal model, but by whether it improves prediction of an independent target measurement beyond the information already contained in the current estimate, and whether that increment is specific to the correct perturbation. From this perspective, multimodal information is conditional and incremental evidence about the remaining uncertainty in a cellular response.

Here we introduce RePert, a framework that turns independent measurements from retrospective reproducibility checks into supervision and calibration signals for perturbation modelling (Fig. 1). RePert constructs response estimators with complementary inductive biases from limited target-modality measurements and combines them through reproducibility-calibrated aggregation (ReCA), using independent measurements that are

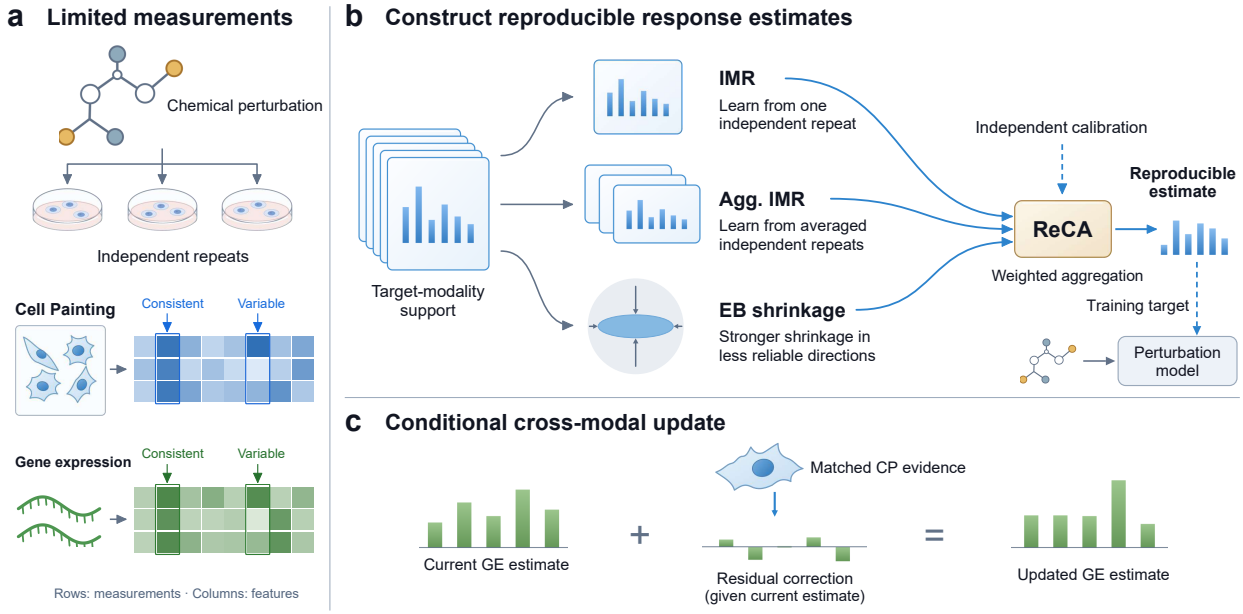


Fig. 1 | Independent measurements guide reproducible target construction and conditional multimodal updating.

separated from estimator fitting. When a second assay is available, RePert evaluates its incremental contribution beyond the current response estimate and tests the perturbation specificity of that contribution. Across BBBC047 and cpg0004-LINCS [17], this strategy recovered responses that agreed more closely with independent measurements and, in CP, retained specificity to the correct compound. Given identical inputs in BBBC047, repeat-aware estimation also outperformed larger generic neural predictors. Matched measurements from the second assay added compound-specific evidence, and added less when the target modality was measured more deeply. The recovered responses more faithfully preserved independently supported morphological features, response directions and dose organization, and improved prediction of unseen compounds when used as training targets in BBBC047. By contrast, sci-Plex3 [22], with only two independent measurements per condition, showed that the benefit depends on the available measurement structure and prediction setting. More broadly, perturbation modelling includes constructing targets that reflect the responses finite experiments can reproducibly support. RePert makes this step explicit and treats target construction as a design dimension distinct from model architecture and capacity.

Results

We used RePert (Fig. 1) to test three consequences of letting independent measurements define response targets. First, if limited measurements contain a perturbation-specific component that recurs across acquisitions, estimators supervised and calibrated by independent measurements should recover it better than averaging or generic predictors given the same inputs. Second, if a second assay provides conditional evidence, it should add compound-specific information beyond the estimate already available from the target assay, and add less as that estimate improves. Third, if the recovered responses are better targets,

they should preserve morphological structure seen in independent references and, when used for training, improve prediction for unseen compounds.

Table 1 | Datasets and profile representations

Dataset	Cell line(s)	Assay	Representation	Dimensions
BBBC047 (Rosetta)	U2OS	CP	CellProfiler features	775
		GE	L1000 profiles	977
cpg0004-LINCS	A549	CP	CellProfiler features[a]	242
sci-Plex3	A549, K562, MCF7	GE	RNA-seq pseudobulk	2,000

Limited measurements contain reproducible perturbation signals

We first examined whether a compound-specific response could be detected across independent measurements. In BBBC047 CP, we compared non-overlapping measurements from the same compound with measurements from different compounds occupying matched experimental conditions. For each compound, we calculated the Pearson correlation between same-compound measurements and subtracted the corresponding correlation obtained from the matched different-compound control. We refer to this difference as excess agreement. Across 18,082 compounds, the median excess agreement was 0.1392 when each side contained one measurement and increased to 0.2431 when each side averaged two independent measurements (Fig. 2a). The median paired increase within compounds was 0.0861, with a 95% confidence interval (CI) of 0.0845–0.0877, and 82.2% of compounds showed an increase.

Because excess agreement subtracts the agreement with matched measurements of other compounds, the increase with averaging reflects a stronger compound-specific response. Individual measurements are therefore noisy but contain a reproducible perturbation response that becomes clearer when independent measurements are combined. GE showed the same pattern in the 38 compounds with four eligible, non-overlapping measurements (Supplementary Table S1).

Independent measurements improve estimates of reproducible responses

The aggregation analysis showed that additional measurements strengthen the reproducible component of a perturbation response, but increasing experimental replication is not always practical. We therefore evaluated a different strategy: using patterns learned across compounds to improve the estimate for a new compound without collecting additional measurements. Independent-measurement regression (IMR), the simplest estimator in RePert, learns from training compounds to map the mean of the R measurements available as input (support measurements) to a separately acquired measurement of the same compound (Fig. 1b; Methods). Because this target measurement is excluded from the input, IMR can reduce its error only by learning response patterns that recur across

measurements. All estimators were evaluated on compounds not used for fitting, and agreement with the held-out measurement was measured as the Fisher z-transformed Pearson correlation.

IMR estimates agreed better with the held-out CP measurement than the mean of the same number of support measurements at every tested budget in both datasets (Fig. 2b). With one support measurement, agreement increased by 0.0805 Fisher-z units in BBBC047 (95% CI, 0.0714–0.0891; 4,023 compounds) and by 0.1328 in cpg0004-LINCS (0.1141–0.1509; 127 compounds). Gains remained positive with two and three support measurements in both datasets (Supplementary Table S1). In morphology, the gain was largest at the lowest measurement budget and decreased as more support measurements became available, consistent with simple averaging recovering more of the reproducible response when additional measurements are present.

The same principle extended beyond morphology. In BBBC047 GE, agreement with held-out measurements increased by 0.0698 Fisher-z units with one support measurement and by 0.0876 with two (Fig. 2b). Unlike in CP, the GE gain was larger with two support measurements than with one, so the value of independent-measurement learning depended on the readout as well as on the measurement budget.

We next compared three estimators with different inductive biases, each built from independent measurements. IMR learns nonlinear relationships across response features; aggregated-target IMR reduces variation in the training target by aggregating measurements excluded from the input; and covariance-based empirical-Bayes shrinkage uses cross-measurement covariance to shrink less reproducible directions. Agreement can also rise when an estimate moves towards a response shared by many compounds. We therefore evaluated these estimators using perturbation specificity, which measures whether an estimate is more similar to the held-out measurement of the correct compound than to condition-matched measurements of other compounds. With foreign references from the support plates, all three estimators improved perturbation specificity at low measurement budgets in BBBC047 (Fig. 2c). With one support measurement, the gains over the support mean were 0.0609 for IMR, 0.0783 for aggregated-target IMR and 0.0859 for empirical-Bayes shrinkage. The gains decreased as more support measurements became available (Fig. 2c). For IMR and aggregated-target IMR, this reflected a smaller gain in agreement with the correct compound. For empirical-Bayes shrinkage, agreement with other compounds also increased, from 0.044 to 0.078 Fisher-z units between one and three support measurements.

Repeat-aware estimation outperforms generic predictors at matched inputs

Empirical-Bayes shrinkage gave the largest gain at every budget. Because the three estimators rest on different assumptions, the best combination can differ across datasets and budgets. RePert therefore combines them by reproducibility-calibrated aggregation (ReCA), which uses independent calibration measurements to determine how much each candidate estimate should contribute. ReCA combines IMR, aggregated-target IMR and empirical-Bayes shrinkage using non-negative weights that sum to one, learned from independent calibration measurements.

In an independent evaluation, 60% of the original training compounds were used to fit the candidate estimators, 20% to select the ReCA weights and comparator hyperparameters, and 20% to assess the resulting procedure under four conditions, with one or three support measurements in each dataset. Relative to the support mean, ReCA increased perturbation specificity by 0.017–0.069 Fisher-z units across the four conditions (Supplementary Table S1). Relative to the strongest individual estimator, which was empirical-Bayes shrinkage in every condition, the average gain was 0.0072 (95% CI, 0.0053–0.0091). This increment was larger in cpg0004-LINCS (0.0158 and 0.0105 with one and three support measurements) than in BBBC047 (0.0006 and 0.0018; Supplementary Table S1). Empirical-Bayes shrinkage received most of the weight, aggregated-target IMR made an additional contribution, and IMR contributed appreciably only in some settings. Most of the gain over the support mean therefore came from covariance-based shrinkage, and calibration against independent measurements added a smaller, setting-dependent increment.

We next compared ReCA with standard response-estimation baselines, denoising methods and generic neural predictors on a common task. In a BBBC047 benchmark, every method received the same one, two or three 775-dimensional CP support profiles and was evaluated against the same held-out measurement and the same 32 condition-matched measurements of other compounds. The comparison included the support mean, principal component analysis (PCA) reconstruction, Noise2Self-style self-supervised denoising [20], a multilayer perceptron (MLP) control, tabular ResNet and FT-Transformer models [23], and TabM [24], alongside ReCA (Fig. 2d). The ResNet, FT-Transformer and TabM models shown in the main panel use the larger reference configurations; capacity-matched compact variants are reported in Supplementary Table S5.

ReCA achieved the highest perturbation specificity at all three measurement budgets, with values of 0.221, 0.254 and 0.282 for one, two and three support measurements, respectively. TabM was the strongest generic neural predictor, with corresponding values of 0.209, 0.238 and 0.268. The paired ReCA–TabM differences were 0.012 (95% CI, 0.008–0.017), 0.017 (0.012–0.021) and 0.014 (0.009–0.019). Larger ResNet and TabM models outperformed their compact counterparts at all three budgets, and scaling FT-Transformer had mixed effects. Empirical-Bayes shrinkage alone, the component with the largest ReCA weight, also exceeded every generic neural predictor (0.216, 0.251 and 0.280; Supplementary Table S4). The advantage of repeat-aware estimation therefore comes mainly from modelling how measurements vary, not from model capacity. Additional classical and component-level comparators, including Median, moderated z-score (MODZ), Ridge, IMR, aggregated-target IMR and empirical-Bayes shrinkage, are reported in Supplementary Tables S4–S5.

We also evaluated cpDistiller-C, a recent Cell Painting representation-learning method designed to reduce batch, row and column effects through contrastive and domain-adversarial learning [25]. cpDistiller-C produces a 50-dimensional latent representation rather than a 775-dimensional response estimate. We therefore assessed it by how well each estimate retrieved the held-out measurement of the same compound from among 32 measurements of other compounds (mean average precision, mAP@33). Across one, two and three support measurements, ReCA achieved the highest mAP@33, followed by the support mean, whereas cpDistiller-C showed the lowest retrieval performance (Fig. 2e). On this replicate-retrieval task, ReCA therefore preserved compound identity more effectively than either simple averaging or cpDistiller-C.

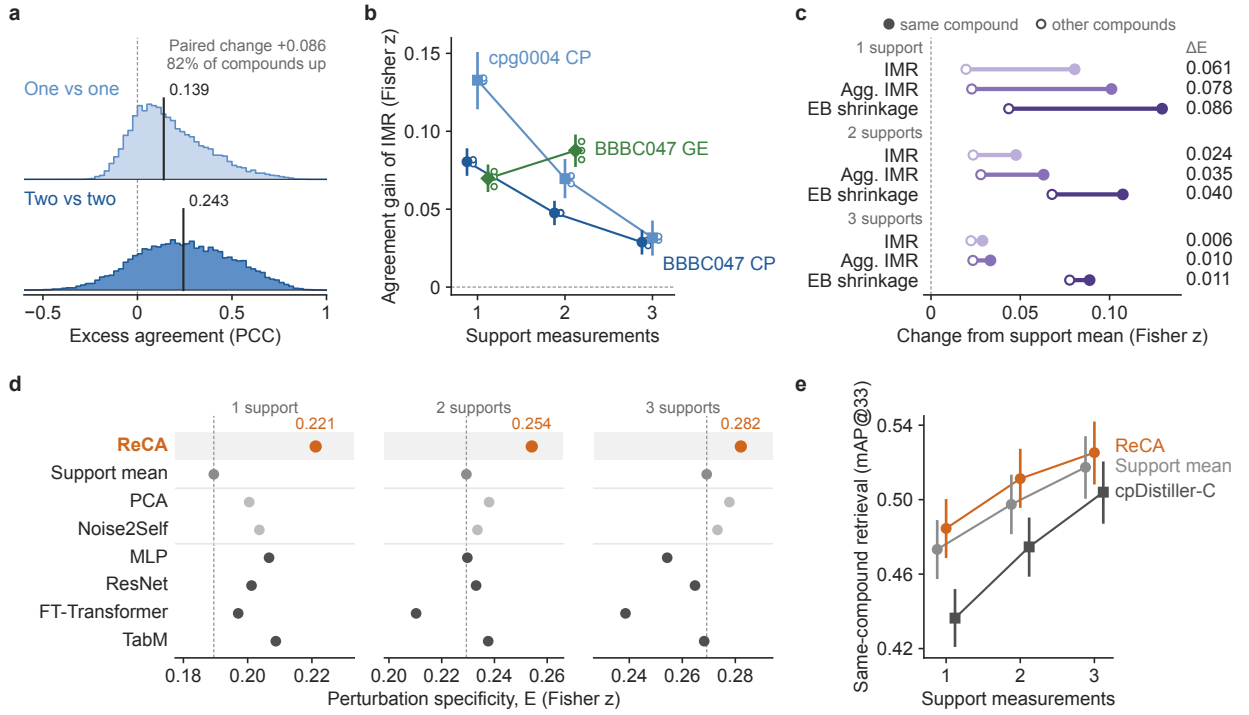


Fig. 2 | Independent measurements improve estimates of reproducible perturbation responses.

A second modality provides compound-specific, conditional evidence

RePert treats a second assay as conditional evidence that updates an existing estimate of the target-modality response (Fig. 1c). We first tested the compound specificity of this evidence by adding GE information to a CP estimate. In BBBC047, the GE evidence increased agreement between the CP estimate and a held-out CP measurement by 0.0215 Fisher-z units across 4,023 compounds (Fig. 3a, GE→CP). To test the compound specificity of this gain, we compared the correctly matched GE evidence with two controls. The improvement was 0.0374 larger than with randomly reassigned GE evidence and 0.0108 larger than with GE evidence from a different compound measured under matched conditions (3,538 compounds). The same direction was observed in a more restrictive analysis requiring exact dose matching (715 compounds; Supplementary Table S2). These controls show that the improvement required GE information from the same compound.

Using CP to update GE gave the same qualitative result. CP evidence improved agreement with an independent GE measurement by 0.0094 Fisher-z units across 3,691 compounds (Fig. 3a, CP→GE). The improvement was 0.0142 larger than with shuffled CP evidence and 0.0114 larger than with condition-matched CP evidence from another compound. The effect remained positive across all three training runs (Supplementary Fig. S1). Thus, each modality provided information about the perturbation response that was not fully captured by the other modality alone.

We next compared conditional cross-modal updating with joint prediction. For GE prediction, a model that received GE and CP together improved on the target-only estimate by 0.0049 Fisher-z units, whereas the conditional CP update improved it by 0.0094. On the same compounds and doses, the conditional update exceeded joint prediction by 0.0045 (Fig. 3b).

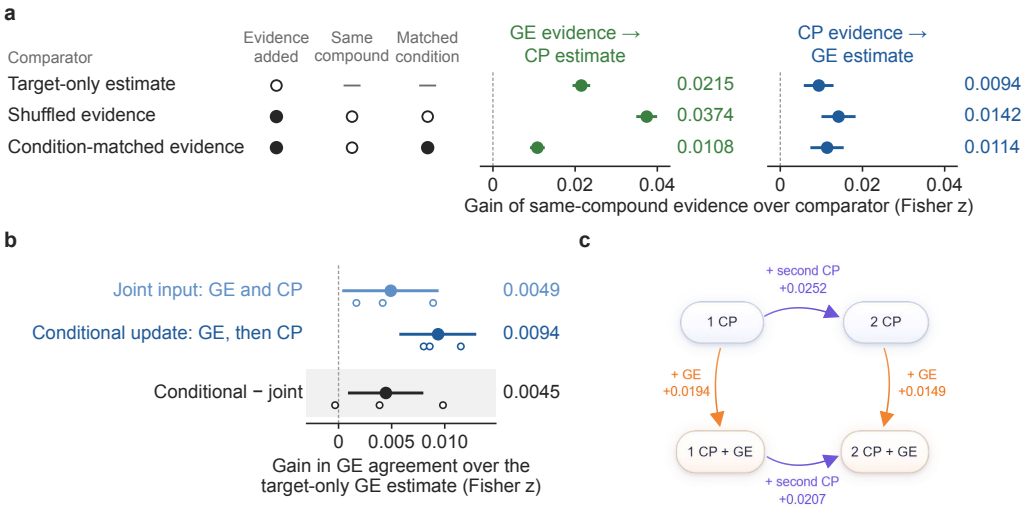


Fig. 3 | A second modality adds compound-specific, conditional evidence.

In this setting, estimating the GE response first and then using CP as additional evidence recovered more reproducible GE signal than combining the two modalities as joint inputs.

The contribution of the second modality also depended on how much information was already available from the target modality. Among 3,983 compounds evaluated against the same held-out CP measurement, adding GE after one CP measurement increased agreement by 0.0194 Fisher-z units, compared with 0.0149 after two CP measurements. Adding a second CP measurement increased agreement by 0.0252 without GE and by 0.0207 when GE evidence was already available (Fig. 3c). Thus, CP and GE provided partially overlapping information. Each source remained useful, but its incremental contribution was smaller when the other source had already supplied additional evidence.

These results show that cross-modal information acts as compound-specific evidence whose value depends on the response estimate already available. The second modality was most useful when it contributed information not already captured by the target modality, which supports using multimodal measurements conditionally.

Recovered responses preserve morphological structure supported by independent references

We next evaluated how well the recovered responses preserved the biological organization of morphological profiles. We used cpg0004-LINCS, where six doses and five measurements per condition allow independent references for feature groups and dose series. For each condition, the one-support ReCA estimate and the single measurement it was derived from (Raw) were each compared with the mean of the four other measurements. Across the seven feature groups, agreement increased from 0.5927 for Raw to 0.7339 for ReCA (difference, 0.1412 Fisher-z units; 95% CI, 0.1312–0.1511; 291 compounds; Fig. 4a; Supplementary Table S3). All seven group-level contrasts were positive, ranging from 0.1016 in mitochondria to 0.1976 in shape. The improvement therefore extended across multiple components of cellular morphology.

We then examined whether ReCA recovered the dominant morphological changes identified by the independent reference. The reference's 25 largest absolute feature changes were defined from the reference alone. ReCA increased overlap with these features from

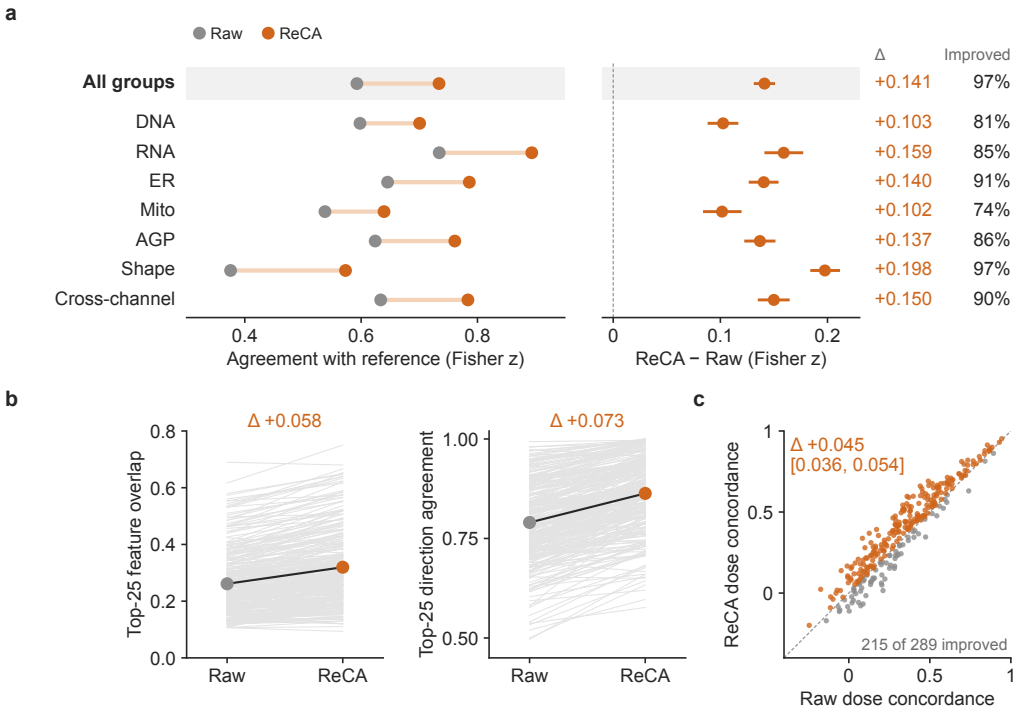


Fig. 4 | Recovered responses preserve morphological structure supported by independent references.

0.2612 to 0.3196 (difference, 0.0584; 95% CI, 0.0533–0.0635) and increased agreement in their direction from 0.7902 to 0.8632 (difference, 0.0730; 95% CI, 0.0679–0.0784; 291 compounds; Fig. 4b). The recovered responses therefore captured both which features changed most strongly and the direction of those changes. In the independent evaluation on cpg0004-LINCS, one-support ReCA also increased perturbation specificity over the single measurement, by 0.069 Fisher-z units (Supplementary Table S1).

ReCA also better preserved the organization of responses across doses. For each compound, we represented the six-dose response by the 15 pairwise distances between doses and compared the resulting distance pattern with the independent reference across the seven feature groups. Dose concordance increased from 0.3267 for Raw to 0.3717 for ReCA (difference, 0.0450; 95% CI, 0.0356–0.0542; 289 compounds; Fig. 4c). The recovered responses therefore also preserved how responses were organized across doses.

Effects on annotation-based tasks depended on the task. ReCA improved confirmed-activity ranking, with average precision increasing from 0.1126 to 0.1791 (difference, 0.0665; 95% CI, 0.0212–0.1294; 262 compounds, of which 7 had confirmed activity; Supplementary Fig. S3). In contrast, mechanism, target and curated Reactome target-neighbour retrieval showed no detectable change (paired mAP differences, -0.0025 to -0.0002; Supplementary Table S3). Recovering reproducible morphological structure therefore improved activity ranking but not annotation retrieval.

Reproducibility-aware targets improve unseen-compound prediction in BBBC047

We next evaluated the effect of reproducibility-aware training targets on a separate model that predicts measured profiles from molecular and control features (Figs. 1b and 5). The Raw and ReCA-derived predictors used identical inputs, matched total parameter counts and the same measured evaluation targets. Performance was assessed using the Pearson

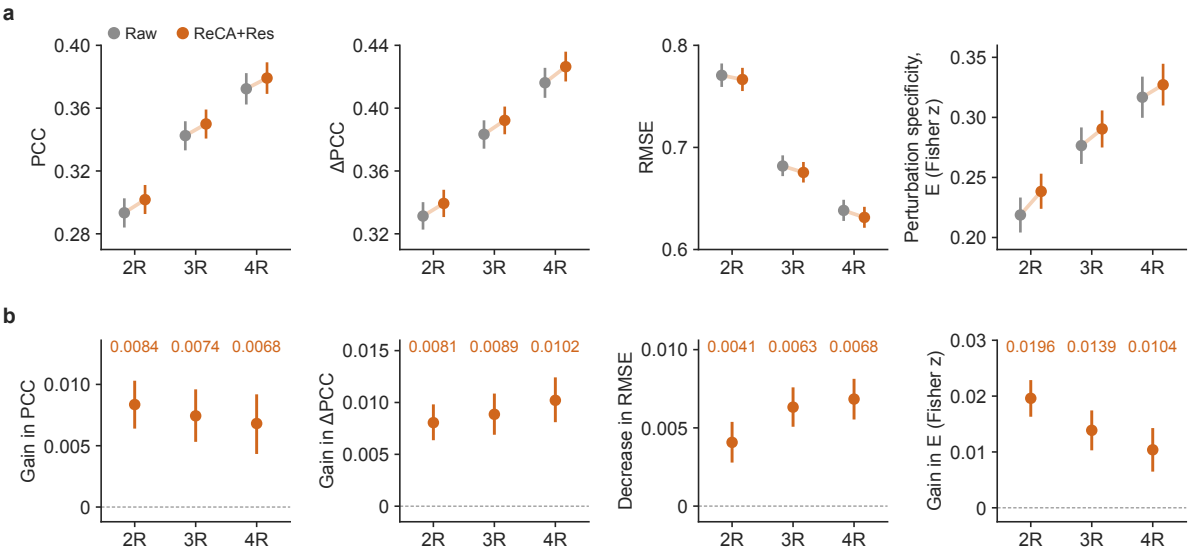


Fig. 5 | Reproducibility-aware training targets improve unseen-compound prediction in BBBC047.

correlation coefficient (PCC) on full profiles, PCC on control-subtracted responses (ΔPCC), root mean squared error (RMSE) and perturbation specificity (E).

In BBBC047, the Raw predictor was trained on the mean of exactly two, three or four selected measurements (2R, 3R and 4R). ReCA+Res split the same training target into two parts: the mean of one-support ReCA estimates computed from each selected measurement, and the remaining residual. Separate branches predicted the two parts, which were summed to reconstruct the measured profile. Across unseen compounds, ReCA+Res improved all four metrics at every measurement budget, and the 95% CI of every paired gain excluded zero (Fig. 5). PCC increased from 0.2934 to 0.3017 at 2R, from 0.3425 to 0.3500 at 3R and from 0.3723 to 0.3792 at 4R. At 3R, ΔPCC increased from 0.3834 to 0.3922, while RMSE decreased from 0.6818 to 0.6755. Perturbation specificity increased by 0.0196, 0.0139 and 0.0104 Fisher-z units at 2R, 3R and 4R (95% CIs, 0.0163–0.0229, 0.0103–0.0174 and 0.0065–0.0143), so the predictions also became more specific to the correct compound.

We also tested ReCA-derived targets in sci-Plex3, which provides only two measurements per drug–dose condition. There, the effect depended on the prediction setting and on how the targets were estimated, and no target construction improved perturbation specificity consistently (Supplementary Fig. S2 and Supplementary Tables S3 and S6).

Discussion

The responses that perturbation models can learn are bounded by the targets they are trained to reproduce. Large perturbation atlases typically estimate each condition from only a few measurements, so these targets combine perturbation effects that recur across acquisitions with measurement-specific variation. Replicate measurements are usually averaged or used to assess reproducibility; RePert uses them to supervise and calibrate response estimation. At the same measurement budget, learning from separate measurements improved agreement with held-out acquisitions in CP and GE data, and in CP the recovered responses better distinguished the correct compound from condition-matched alternatives.

Most of the specificity gain came from covariance-based empirical-Bayes shrinkage, and ReCA added a smaller, setting-dependent increment. Given identical inputs in BBBC047, shrinkage alone exceeded every generic neural predictor tested, including larger models, indicating that exploiting measurement structure and increasing model capacity address distinct aspects of the learning problem. Target construction is therefore a design dimension distinct from model architecture and capacity, which have been the main focus of current perturbation-response models [6–8, 12, 26, 27].

This perspective changes the role of repeated and multimodal measurements in perturbation modelling. In RePert, an independent acquisition provides the operational criterion for whether response structure is supported beyond the measurements used to estimate it. RePert builds on learning from independent noisy observations [19, 20], empirical-Bayes shrinkage [28, 29] and stacked generalization [30–32], and organizes them around the acquisition structure of perturbation experiments, so that separate measurements serve as evidence, calibration and evaluation. The same principle extends across assays. CP and GE each provided compound-specific information beyond an estimate obtained from the other assay, and GE added less to a CP estimate when more CP measurements were available. In the GE setting tested here, updating an existing estimate with CP also recovered more independently supported signal than joint prediction from both assays. The value of an additional assay therefore depends on what has already been measured, which connects multimodal modelling to experimental allocation: an additional replicate and an additional assay are both sources of information whose marginal value can depend on the existing measurement regime. Beyond agreement with held-out measurements, the recovered responses better preserved independently supported morphological features, response directions and dose organization, and improved confirmed-activity ranking, whereas mechanism, target and Reactome target-neighbour retrieval showed no detectable change. Reproducibility and annotation-defined biological similarity are therefore distinct criteria for target construction. When used to split training targets into a reproducible component and a residual, ReCA estimates improved prediction of unseen compounds in BBBC047 while the residual variation was retained in a separate branch.

The data analysed here set the scope of these conclusions. Independent measurements were separated at the physical-plate level within existing CP and GE resources. Measurements within the same resource can share plate effects or other systematic factors [11], and responses computed against a shared control estimate share its sampling noise, which inflates their correlation [33]. Agreement across repeats therefore measures reproducibility within an acquisition design, which can include such shared effects. Cross-modal evidence was paired by compound identity rather than by dose or acquisition. Replication depth was uneven across datasets: GE and higher-budget analyses contained fewer eligible conditions, and sci-Plex3 provided two independent measurements per condition. With two measurements, aggregated-target IMR reduces to IMR and covariance shrinkage to a diagonal rule, leaving little information for independent supervision, estimation and calibration; under this regime, no target-construction strategy consistently improved perturbation specificity. The morphological analyses used extracted profiles rather than raw images, and other biological endpoints may require different response components to be retained. Validation across sites, platforms and independently generated experiments will be important for determining how well reproducibility-defined targets transfer beyond the systems in which they are estimated.

In summary, independent measurements that are usually used to check reproducibility can also define the responses that perturbation models learn. Estimators supervised and calibrated by separate measurements agreed better with held-out acquisitions and, in CP, were more specific to the correct compound. A second assay added evidence whose value depended on what had already been measured, and the recovered responses improved prediction of unseen compounds when used as training targets in BBBC047. As perturbation atlases continue to expand [5], measurements and learning objectives can be designed together so that finite experiments define responses that models can reliably learn.

Methods

Data and experimental design

We analysed CP and GE profiles from the Rosetta-paired BBBC047 resource, CP profiles from cpg0004-LINCS, and GE profiles from sci-Plex3 (Table 1). The Rosetta resource provides paired CP and L1000 GE measurements from separate acquisition plate sets [4]. Cross-modal analyses matched CP and GE profiles by compound identity alone, without pairing acquisition repeats or dose levels across assays. CP inputs were provided as CellProfiler-derived tabular profiles [9, 17, 34]; image segmentation, feature extraction and single-cell-to-well aggregation were performed by the original data providers. The fixed BBBC047 representations contained 775 CP features and 977 GE features. The cpg0004-LINCS representation contained 242 coordinates, including 241 biological features and one batch coordinate, which was excluded from feature-group analyses. The preprocessing steps are described in the Supplementary Information.

An independent response measurement was defined at the plate level rather than at the cell or technical-well level. For BBBC047, treated-cell profiles were first aggregated within each compound–plate combination and then normalized by subtracting the feature-wise median dimethyl sulfoxide (DMSO) control profile from the same plate. For cpg0004-LINCS, technical wells were median-aggregated within each compound–dose–plate condition before subtraction of the corresponding same-plate control profile. Subsequent aggregation assigned equal weight to physical plates. Compound-disjoint partitions contained 12,175, 4,059 and 4,060 BBBC047 compounds and 941, 314 and 313 cpg0004-LINCS compounds for training, validation and test, respectively. Compound identities in BBBC047 followed the supplied mapping to canonical simplified molecular-input line-entry system (SMILES) strings.

For each condition, R support measurements were averaged with equal weights to form the input profile, while a separate plate-level measurement served as the evaluation reference. Support and reference assignments were made from compound and acquisition identifiers alone. We evaluated one, two and three CP support measurements and one or two GE support measurements; with three GE supports, four compounds retained a disjoint reference. Each budget had its own set of eligible compounds, so comparisons between budgets involve different compounds. Eligible sample counts and the separate descriptive analysis of measurement aggregation are reported in Supplementary Information.

Reproducible-response formulation

For condition c, modality m and independent measurement r, we represent the observed response as

$$\boldsymbol{Y}_{c,m,r} = \boldsymbol{\mu}_{c,m} + \boldsymbol{\varepsilon}_{c,m,r}, \tag{1}$$

where $\boldsymbol{\mu}_{c,m}$ denotes the reproducible component of the response under the acquisition design and $\boldsymbol{\varepsilon}_{c,m,r}$ denotes measurement-specific variation. This formulation defines $\boldsymbol{\mu}_{c,m}$ operationally, as the response component supported across independent measurements under the acquisition design.

For clarity, we suppress the modality index below. Let S_c denote the available support measurements for condition c, and let $\boldsymbol{Y}_c^{\mathrm{held}}$ denote a separate measurement that is excluded from the support. We define the reproducibility risk of an estimator f as

$$\mathcal{R}_{\mathrm{repro}}(f) = \mathbb{E}\left[\left\| f(S_c) - \boldsymbol{Y}_c^{\mathrm{held}} \right\|_2^2\right]. \tag{2}$$

If the held-out measurement error satisfies $\mathbb{E}[\boldsymbol{\varepsilon}_c^{\mathrm{held}} \mid S_c, \boldsymbol{\mu}_c] = 0$, then

$$\mathcal{R}_{\mathrm{repro}}(f) = \mathbb{E}\left[\left\| f(S_c) - \boldsymbol{\mu}_c \right\|_2^2\right] + \mathbb{E}\left[\left\| \boldsymbol{\varepsilon}_c^{\mathrm{held}} \right\|_2^2\right]. \tag{3}$$

The second term does not depend on f. Thus, under this condition, minimizing prediction error to an independently acquired measurement has the same population optimum as minimizing error to the reproducible response component, even though the held-out measurement itself is noisy [19]. Shared acquisition effects or systematic errors can violate this condition; we therefore use matched controls and physically separated measurements where the data permit. The full derivation and assumptions are provided in Supplementary Methods.

RePert response estimation

RePert constructs three candidate estimates from the same support measurements and combines them using independent calibration measurements. The candidates differ in how information from repeated measurements is used to estimate the reproducible response: IMR predicts a separate measurement directly, aggregated-target IMR predicts an aggregate of measurements excluded from the support, and empirical-Bayes shrinkage shrinks response directions according to their estimated cross-measurement reproducibility.

Let

$$\bar{\boldsymbol{Y}}_{c,S} = \frac{1}{R} \sum_{r \in S_c} \boldsymbol{Y}_{c,r} \tag{4}$$

denote the equal-weight mean of the R support measurements for condition c.

Independent-measurement regression (IMR).

IMR learns a mapping from the support mean to a separately acquired response. Given a held-out measurement $\boldsymbol{Y}_c^{\mathrm{held}}$ that is excluded from the support set, the regression parameters are learned by minimizing

$$\hat{\theta}_{\mathrm{IMR}} = \arg\min_{\theta} \sum_{c \in \mathcal{D}_{\mathrm{train}}} \left\| f_\theta\left(\bar{\boldsymbol{Y}}_{c,S}\right) - \boldsymbol{Y}_c^{\mathrm{held}} \right\|_2^2. \tag{5}$$

The resulting estimate is $\hat{\boldsymbol{\mu}}_c^{\mathrm{IMR}} = f_{\hat{\theta}_{\mathrm{IMR}}}(\bar{\boldsymbol{Y}}_{c,S})$. Because the target measurement is disjoint from the input, the model is trained to recover response structure that generalizes across measurements rather than to reconstruct the observed support profiles themselves. We implemented f_θ as a multilayer perceptron, allowing nonlinear relationships among response features to be learned.

Aggregated-target IMR.

The aggregated-target variant of IMR uses the same regression architecture but replaces the single held-out target with an aggregate of all eligible training measurements that are excluded from the support set. For condition c, let

$$\bar{\boldsymbol{Y}}_{c,-S} = \frac{1}{|H_c|} \sum_{r \in H_c} \boldsymbol{Y}_{c,r}, \qquad H_c \cap S_c = \varnothing, \tag{6}$$

where H_c contains the eligible non-support measurements. The aggregated-target variant is trained by minimizing

$$\hat{\theta}_{\mathrm{agg}} = \arg\min_{\theta} \sum_{c \in \mathcal{D}_{\mathrm{train}}} \left\| f_\theta\left(\bar{\boldsymbol{Y}}_{c,S}\right) - \bar{\boldsymbol{Y}}_{c,-S} \right\|_2^2 . \tag{7}$$

Here, agg denotes the aggregated-target variant. Averaging the non-support measurements reduces variation in the regression target.

Covariance-based empirical-Bayes shrinkage.

This covariance-based empirical-Bayes estimator uses repeated training measurements to estimate within-condition measurement covariance and between-condition signal covariance. These covariance estimates identify response directions that are more consistently supported across measurements. After whitening by the estimated measurement covariance, directions with negative estimated signal eigenvalues are truncated to zero. A direction with non-negative signal eigenvalue λ_j receives the support-dependent shrinkage factor

$$g_j(R) = \frac{\lambda_j}{\lambda_j + 1/R}. \tag{8}$$

Directions with stronger estimated reproducible variation therefore undergo less shrinkage, whereas weakly supported directions are pulled more strongly towards the fitted population mean. The estimate is then transformed back to the original response coordinates. This covariance-based estimator provides a different inductive bias from the nonlinear IMR and aggregated-target IMR regressors [29]. Full covariance estimation and the working-model derivation are provided in Supplementary Methods.

Reproducibility-calibrated aggregation (ReCA)

ReCA combines the three candidate estimates with non-negative weights constrained to sum to one:

$$\widehat{\boldsymbol{y}}_{\mathrm{ReCA}} = \sum_{k=1}^{3} w_k \widehat{\boldsymbol{y}}_k, \qquad w_k \geq 0, \qquad \sum_{k=1}^{3} w_k = 1. \tag{9}$$

Weights are selected by prediction error against disjoint measurements from calibration compounds, separate from candidate fitting and input construction, as in stacked regression

[31, 32]. In the BBBC047 benchmark (Fig. 2d), the weights minimized the mean squared error to held-out validation measurements over the simplex. In the independent evaluation described next, the least-squares simplex weights $\hat{\boldsymbol{w}}$ fitted on n calibration compounds were shrunk towards equal weights,

$$\boldsymbol{w} = (1-\gamma)\,\tfrac{1}{3}\mathbf{1} + \gamma\,\hat{\boldsymbol{w}}, \qquad \gamma = \frac{n}{n+3}, \tag{10}$$

which corresponds to a Dirichlet(1,1,1) prior contributing one pseudo-compound per candidate, so γ is set by n. In the independent evaluation, the original training compounds were repartitioned into 60% base-fitting, 20% calibration and 20% evaluation sets. IMR, aggregated-target IMR and empirical-Bayes shrinkage were fitted on the base set. Calibration selected simplex weights and the PCA comparator's rank. Four conditions used one- and three-support CP inputs in both datasets and compared ReCA with support means, PCA reconstruction and the strongest individual estimator. The macro-average weighted conditions equally. Counts and comparator selection are specified in Supplementary Methods.

Implementation details.

IMR and aggregated-target IMR used multilayer perceptrons with Gaussian error linear unit (GELU) activations, a 32-dimensional bottleneck and hidden widths of 256 for BBBC047 CP and GE or 128 for cpg0004 CP. Inputs and targets were standardized on fitting data. Models used AdamW [35] with learning rate 3×10^{-4}, weight decay 10^{-5}, batch size 256 and at most 80 epochs. Checkpoints were selected within each experiment's development compounds; in the independent ReCA evaluation, epochs were selected within the base set and the models were refitted on that set. Further numerical settings are in Supplementary Methods.

Conditional cross-modal evidence

We tested whether a second modality could improve an already available estimate of the target response. Let $\boldsymbol{P}_0$ denote the target-only estimate and let e denote additional evidence from the other modality. We express the updated estimate as

$$\boldsymbol{P}_1 = \boldsymbol{P}_0 + \beta\widehat{\boldsymbol{R}}(e \mid \boldsymbol{P}_0), \tag{11}$$

where $\widehat{\boldsymbol{R}}(e \mid \boldsymbol{P}_0)$ represents the component of the remaining response that can be predicted from the additional evidence conditional on the current estimate. The update strength β was selected using validation compounds. In the CP-to-GE experiment, the candidate grid included $\beta = 0$, allowing the target-only estimate to be retained when the additional modality did not improve the validation criterion. This formulation evaluates cross-modal information by its incremental contribution beyond an existing response estimate.

GE-to-CP implementation.

For GE-to-CP updating, $\boldsymbol{P}_0$ was a fixed CP estimate combining IMR with a virtual CP prior. The candidate residual was defined as the difference between a GE-informed CP prediction and the corresponding virtual CP prior. A single update strength was selected on validation compounds. To determine whether any gain depended on compound-specific GE information, the correctly matched residual was compared with two controls: residual vectors shuffled across compounds, and residual vectors obtained from metadata-matched

foreign compounds. These controls altered the compound identity of the cross-modal evidence while leaving the fitted predictors unchanged.

CP-to-GE implementation.

For CP-to-GE updating, the residual mapper was conditioned jointly on the target-only GE estimate and the observed CP profile,

$$\widehat{\boldsymbol{R}} = f(\boldsymbol{P}_0, \boldsymbol{Y}_{\mathrm{CP}}). \tag{12}$$

The training target was the residual GE response, $\boldsymbol{Y}_{\mathrm{GE}} - \boldsymbol{P}_0$. To prevent a compound from contributing to its own residual target through $\boldsymbol{P}_0$, target-only estimates for the training compounds were generated by five-fold compound-level cross-fitting. Model checkpoints were selected using training compounds, and validation compounds were used to select β from a grid that included zero. Correct, shuffled and metadata-matched foreign CP inputs were then evaluated using the same target-only estimate, fitted residual mapper and selected update weight.

As a direct-fusion comparator, we trained a model to predict GE from the concatenation of the measured GE support profile and CP profile. Conditional updating and direct fusion were compared on the same compounds and against the same target-only starting estimate. In a separate sequential experiment, we measured the incremental contribution of GE after one or two CP support measurements using a common disjoint CP reference. Network architectures, evidence-matching procedures and validation grids are described in Supplementary Methods.

Evaluation metrics

Repeat agreement.

For an estimated response $\widehat{\boldsymbol{y}}$ and an independent reference $\boldsymbol{y}$, repeat agreement was defined as the Fisher z-transformed Pearson correlation:

$$A(\widehat{\boldsymbol{y}}, \boldsymbol{y}) = \operatorname{arctanh}\left[\operatorname{corr}(\widehat{\boldsymbol{y}}, \boldsymbol{y})\right]. \tag{13}$$

For numerical stability, correlations were clipped to $[-1+\epsilon, 1-\epsilon]$, with $\epsilon = 10^{-6}$, before transformation. In the repeat-budget, cross-modal and morphological analyses, each condition i was assigned a fixed matched-foreign measurement-pair baseline n_i that was shared across methods. The corrected agreement was therefore $A_i - n_i$. Because the same n_i was used for all methods, this term cancels in paired method comparisons.

Perturbation specificity.

The estimator-decomposition and independent ReCA analyses used a prediction-dependent foreign correction. For condition i, perturbation specificity was defined as

$$E_i = A(\widehat{\boldsymbol{y}}_i, \boldsymbol{y}_i) - \frac{1}{K_i}\sum_{j=1}^{K_i} A(\widehat{\boldsymbol{y}}_i, \boldsymbol{y}_{ij}^{\mathrm{foreign}}), \tag{14}$$

where $\boldsymbol{y}_i$ is the correct independent reference and $\boldsymbol{y}_{ij}^{\mathrm{foreign}}$ denotes a matched reference from a different compound. The same foreign-compound identifiers were used across

methods, but the resulting correction depends on the prediction and therefore differs from the fixed measurement-pair baseline used above. Foreign references were matched using the available acquisition and dose information; the exact matching rules are given in Supplementary Methods.

Response-estimation and representation benchmarks.

The BBBC047 same-space response-estimation benchmark compared the support mean, median, MODZ, PCA reconstruction, ridge regression, Noise2Self-style denoising, IMR, aggregated-target IMR, empirical-Bayes shrinkage and ReCA using identical physical support and reference assignments in the 775-dimensional CP space. Perturbation specificity E was the primary endpoint. Neural architecture comparisons used the same support/reference assignments and evaluated MLP, ResNet, FT-Transformer and TabM under capacity-matched and larger reference configurations [23, 24].

cpDistiller-C, which produces a 50-dimensional representation, was evaluated by same-compound retrieval. Each query was compared with one held-out profile from the same compound and matched foreign-compound profiles, with mAP@33 as the endpoint and one positive profile per query. Support assignments, parameter counts, fitting procedures and retrieval construction are detailed in Supplementary Methods.

Prediction metrics.

For profile i, let $\hat{\Delta \boldsymbol{y}}_i$ and $\Delta \boldsymbol{y}_i$ denote the predicted and observed control-subtracted responses of dimension d, and let $\boldsymbol{b}_i$ denote the matched control baseline. We computed

$$\mathrm{PCC}_i = \mathrm{corr}\left(\hat{\Delta \boldsymbol{y}}_i + \boldsymbol{b}_i, \Delta \boldsymbol{y}_i + \boldsymbol{b}_i\right), \tag{15}$$

$$\Delta\mathrm{PCC}_i = \mathrm{corr}\left(\hat{\Delta \boldsymbol{y}}_i, \Delta \boldsymbol{y}_i\right), \tag{16}$$

$$\mathrm{RMSE}_i = \sqrt{\frac{1}{d}\sum_{j=1}^{d}\left(\hat{\Delta y}_{ij} - \Delta y_{ij}\right)^2}. \tag{17}$$

For predictors, perturbation specificity E was computed as in equation (14), comparing each prediction with the measured profile of the correct compound and with matched measured profiles of other compounds. Metrics were first computed at the profile level and then aggregated at the compound or drug level according to the analysis-specific evaluation procedure. PCC and ΔPCC in Fig. 5 are untransformed correlations. RMSE is the same for full and control-subtracted profiles, because the prediction and observation share the control baseline.

Biological organization analyses

Morphological analyses used cpg0004-LINCS and a one-support ReCA model. For each of two support rotations, the estimate from one physical profile was compared with the mean of four other physical profiles of the same compound–dose condition. The 241 biological features were grouped as DNA, RNA, endoplasmic reticulum, mitochondria, actin/Golgi/plasma membrane, shape and cross-channel features. Figure 4a displays the six structural groups, the cross-channel group and the macro-average over all seven. Feature-group agreement and dominant-feature recovery were scored per support view. Top-25

overlap measured shared dominant standardized features, while direction agreement used the reference-defined top 25. Dose concordance correlated predicted and reference vectors of all 15 pairwise $1 - \mathrm{PCC}$ distances among six doses using Spearman correlation, with equal weight across feature groups. Supplementary Methods give the weights and eligible condition counts.

Activity ranking used labels requiring reproducible and sufficiently large effects in disjoint confirmation measurements, with thresholds derived from validation data. Mechanism and target retrieval ranked dose-matched training/validation consensus profiles by PCC against Raw or ReCA test queries. Pathway retrieval used the same Reactome enrichment of curated target genes from the top 25 neighbours, with Benjamini–Hochberg correction within query [36, 37]. Activity average precision (AP) and the three retrieval mAP endpoints retained separate eligible sets. Exact feature counts, support eligibility, retrieval cohorts and scoring details are in Supplementary Methods.

Prediction with RePert-derived supervision

We tested whether RePert-derived response estimates improve the training targets of separate downstream predictors. In BBBC047, we evaluated prediction of the raw response averaged across a fixed number of selected physical repeats, using molecular and control inputs. In sci-Plex3, we evaluated transfer of drug responses from two source cell lines to a held-out target cell line and, separately, prediction of unseen drugs within a cell line. In both settings, the compared predictors used matched input information and were evaluated against the same target profiles.

BBBC047 molecular-input prediction.

For each compound and repeat budget $k \in \{2, 3, 4\}$, exactly k distinct physical plates were selected. ReCA was applied independently to each selected repeat. We defined

$$\bar{Y} = \frac{1}{k}\sum_{r=1}^{k} Y_r, \qquad \bar{C} = \frac{1}{k}\sum_{r=1}^{k} \mathrm{ReCA}_{1R}(Y_r), \qquad \bar{R} = \bar{Y} - \bar{C}. \tag{18}$$

The Raw predictor used $\bar{Y}$ as its training target. The ReCA+Res predictor instead used the ReCA-derived component $\bar{C}$ and the remaining residual $\bar{R}$ as separate branch targets. The two predicted components were summed to reconstruct $\bar{Y}$. Both predictors took molecular fingerprints and a control profile as their only inputs and were matched in model capacity. Evaluation was performed on unseen compounds with $\bar{Y}$ as the common target, so the benchmark assessed prediction of the selected-measurement mean. ReCA targets for training compounds were generated out of fold at the compound level. Repeat selection, model architectures and training procedures are detailed in Supplementary Methods.

sci-Plex3 cross-cell-line prediction.

We used the 24-hour sci-Plex3 dataset, comprising 183 drugs and 613 drug–dose conditions with two replicate measurements [22]. Responses were represented as pseudobulk log-counts per million after subtraction of matched Vehicle profiles, with 2,000 features selected from source-cell Vehicle measurements. For each held-out target cell line, a downstream predictor estimated drug responses using the other two cell lines together with

extended-connectivity fingerprints (ECFP4) [38], log-dose and the target-cell Vehicle baseline. Source-cell ReCA targets were averaged across the two replicates and then equally across the two source cell lines. Raw, ReCA and ReCA+Res predictors used identical inputs and matched capacity and differed only in their source-side training targets (Supplementary Table S6). Seen-drug and unseen-drug transfer were evaluated separately. Within-cell-line prediction of the same 36 unseen drugs used the target cell line's own training drugs as the source. Treated profiles from the held-out target cell line were excluded from model fitting and selection. Source-only calibration, model architecture and exact drug partitions are detailed in Supplementary Methods.

Statistical Analyses

Comparisons were paired by compound. Doses and support rotations were grouped within compound, and training seeds were model realizations rather than biological replicates. Uncertainty used bootstrap resampling [39], generally with 10,000 paired compound samples and two-sided 95% percentile intervals. Aggregation was analysis-specific: repeat-budget, estimator-decomposition and cross-modal scores were averaged across three seeds within compound before resampling; independent ReCA evaluation used one fitted candidate set per condition. Predictor analyses used five seeds, with BBBC047 scalar scores averaged within compound and sci-Plex3 prediction vectors averaged before condition-level scoring. The latter resampled drugs within each of three fixed target cell lines. Analysis-specific resampling procedures are given in Supplementary Methods.

The BBBC047 benchmark, morphology analyses and predictor comparisons used the predefined test partition, with models, analysis settings and endpoints prespecified before final evaluation. A separate 60/20/20 ReCA analysis used compounds held out from all development-stage evaluations.

Data availability

Paired BBBC047 CP and L1000 GE profiles were obtained from the Rosetta resource CDRP-BBBC047-Bray-CP-GE [4]. The resource is distributed through Cell Painting Gallery [40] under accession `cpg0003-rosetta`. Dataset descriptions and download instructions are available at https://github.com/carpenter-singh-lab/2022_Haghighi_NatureMethods. The study also used cpg0004-LINCS [17] and sci-Plex3 [22]. Source data are provided in Supplementary Data 1 and Supplementary Data 2. No new experimental measurements were generated.

Code availability

The custom code used to implement RePert, perform the analyses and generate the figures and tables is publicly available at https://github.com/Zhao-weijing/RePert-code under the MIT license. Numerical source data, a minimal synthetic example and instructions for environment setup and code execution are provided in the repository.

Acknowledgements and funding

This work was supported by the Science and Technology Development Fund, Macao S.A.R. (FDCT), grant 0028/2023/RIA1.

Author contributions

W.Z. and B.L. conceived the study and developed the methodology. W.Z., B.L. and C.L. developed the software, performed the experiments and analysed the data. W.Z., B.L. and C.L. wrote the original draft. W.Z., B.L., B.Z., Y.W., S.Z. and C.L. reviewed and edited the manuscript. B.Z. supervised the study, administered the project and acquired funding.

Competing interests

The authors declare no competing interests.

References

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

[2] Bray, M.-A. *et al.* Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols* **11**, 1757–1774 (2016). URL https://doi.org/10.1038/nprot.2016.105.

[3] Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437–1452.e17 (2017). URL https://doi.org/10.1016/j.cell.2017.10.049.

[4] Haghighi, M., Caicedo, J. C., Cimini, B. A., Carpenter, A. E. & Singh, S. High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. *Nature Methods* **19**, 1550–1557 (2022). URL https://doi.org/10.1038/s41592-022-01667-0.

[5] Chandrasekaran, S. N. *et al.* JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv* (2023). URL https://www.biorxiv.org/content/10.1101/2023.03.23.534023v2. Preprint.

[6] Hetzel, L. *et al.* Predicting Cellular Responses to Novel Drug Perturbations at a Single-Cell Resolution. *Advances in Neural Information Processing Systems* **35**, 26711–26722 (2022). URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/aa933b5abc1be30baece1d230ec575a7-Abstract.html.

[7] Roohani, Y., Huang, K. & Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology* **42**, 927–935 (2024). URL https://doi.org/10.1038/s41587-023-01905-6.

[8] Palma, A., Theis, F. J. & Lotfollahi, M. Predicting cell morphological responses to perturbations using generative modeling. *Nature Communications* **16**, 505 (2025). URL https://doi.org/10.1038/s41467-024-55707-8.

[9] Bray, M.-A. *et al.* A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *GigaScience* **6**, giw014 (2017). URL https://doi.org/10.1093/gigascience/giw014.

[10] Caicedo, J. C. *et al.* Data-analysis strategies for image-based cell profiling. *Nature Methods* **14**, 849–863 (2017). URL https://doi.org/10.1038/nmeth.4397.

[11] Arevalo, J. *et al.* Evaluating batch correction methods for image-based cell profiling. *Nature Communications* **15**, 6516 (2024). URL https://doi.org/10.1038/s41467-024-50613-5.

[12] Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nature Methods* **16**, 715–721 (2019). URL https://doi.org/10.1038/s41592-019-0494-8.

[13] Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods* **21**, 1470–1480 (2024). URL https://doi.org/10.1038/s41592-024-02201-0.

[14] Ahlmann-Eltze, C., Huber, W. & Anders, S. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods* **22**, 1657–1661 (2025). URL https://doi.org/10.1038/s41592-025-02772-6.

[15] Wei, Z. *et al.* Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods* **23**, 451–464 (2026). URL https://doi.org/10.1038/s41592-025-02980-0.

[16] Viñas Torné, R. *et al.* Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology* **44**, 1050–1059 (2026). URL https://doi.org/10.1038/s41587-025-02777-8.

[17] Way, G. P. *et al.* Morphology and gene expression profiling provide complementary information for mapping cell state. *Cell Systems* **13**, 911–923.e9 (2022). URL https://doi.org/10.1016/j.cels.2022.10.001.

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

[19] Lehtinen, J. *et al.* Noise2Noise: Learning Image Restoration without Clean Data. In Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research 80, 2965–2974 (2018). URL https://proceedings.mlr.press/v80/lehtinen18a.html.

[20] Batson, J. & Royer, L. Noise2Self: Blind Denoising by Self-Supervision. In Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research 97, 524–533 (2019). URL https://proceedings.mlr.press/v97/batson19a.html.

[21] Moshkov, N. *et al.* Predicting compound activity from phenotypic profiles and chemical structures. *Nature Communications* **14**, 1967 (2023). URL https://doi.org/10.1038/s41467-023-37570-1.

[22] Srivatsan, S. R. *et al.* Massively multiplex chemical transcriptomics at single-cell resolution. *Science* **367**, 45–51 (2020). URL https://doi.org/10.1126/science.aax6234.

[23] Gorishniy, Y., Rubachev, I., Khrulkov, V. & Babenko, A. Revisiting Deep Learning Models for Tabular Data. *Advances in Neural Information Processing Systems* **34**, 18932–18943 (2021). URL https://proceedings.neurips.cc/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html.

[24] Gorishniy, Y., Kotelnikov, A. & Babenko, A. TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling. *The Thirteenth International Conference on Learning Representations* (2025). URL https://openreview.net/forum?id=Sd4wYYOhmY.

[25] Yan, C. *et al.* Triple-effect correction for Cell Painting data with contrastive and domain-adversarial learning. *Nature Communications* **16**, 6886 (2025). URL https://doi.org/10.1038/s41467-025-62193-z.

[26] Lotfollahi, M. *et al.* Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology* **19**, e11517 (2023). URL https://doi.org/10.15252/msb.202211517.

[27] Wang, X. *et al.* Prediction of cellular morphology changes under perturbations with a transcriptome-guided diffusion model. *Nature Communications* **16**, 8210 (2025). URL https://doi.org/10.1038/s41467-025-63478-z.

[28] Efron, B. Tweedie's Formula and Selection Bias. *Journal of the American Statistical Association* **106**, 1602–1614 (2011). URL https://doi.org/10.1198/jasa.2011.tm11181.

[29] Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics* **51**, 187–195 (2019). URL https://doi.org/10.1038/s41588-018-0268-8.

[30] Wolpert, D. H. Stacked generalization. *Neural Networks* **5**, 241–259 (1992). URL https://doi.org/10.1016/s0893-6080(05)80023-1.

[31] Breiman, L. Stacked regressions. *Machine Learning* **24**, 49–64 (1996). URL https://doi.org/10.1007/bf00117832.

[32] van der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super Learner. *Statistical Applications in Genetics and Molecular Biology* **6**, Article 25 (2007). URL https://doi.org/10.2202/1544-6115.1309.

[33] Nicol, P. B., Shivakumar, S. & Irizarry, R. A. Spurious correlation inflates performance in single-cell perturbation prediction. *bioRxiv* (2026). URL https://doi.org/10.64898/2026.05.07.723486. Preprint.

[34] Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology* **7**, R100 (2006). URL https://doi.org/10.1186/gb-2006-7-10-r100.

[35] Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. International Conference on Learning Representations (2019). URL https://openreview.net/forum?id=Bkg6RiCqY7.

[36] Gillespie, M. *et al.* The Reactome pathway knowledgebase 2022. *Nucleic Acids Research* **50**, D687–D692 (2022). URL https://doi.org/10.1093/nar/gkab1028.

[37] Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995). URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[38] Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50**, 742–754 (2010). URL https://doi.org/10.1021/ci100050t.

[39] Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7**, 1–26 (1979). URL https://doi.org/10.1214/aos/1176344552.

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

Display item legends

Table 1 | Datasets and profile representations

Dimensions refer to the model representations used here; analysis-specific compound counts are reported in Methods. CP, Cell Painting; GE, gene expression. [a]The 242-coordinate representation includes 241 biological features and one batch coordinate, excluded from biological analyses.

Figure 1 | Independent measurements guide reproducible target construction and conditional multimodal updating. a, Few independent Cell Painting (CP) and gene-expression (GE) measurements contain patterns that recur across measurements and patterns that vary. **b,** Support profiles yield candidate estimates from independent-measurement regression (IMR), its aggregated-target variant (Agg. IMR) and covariance-based empirical-Bayes shrinkage (EB shrinkage). Reproducibility-calibrated aggregation (ReCA) combines these estimates using weights selected on independent calibration measurements. The resulting estimate can serve as the training target of a perturbation model with molecular inputs. **c,** Matched CP evidence conditions a residual correction to the current GE estimate. The update strength is selected on validation compounds; a weight of zero retains the current estimate. CP-to-GE updating is shown; GE-to-CP updating has the same form with a CP starting estimate. Profile colours are schematic.

Figure 2 | Independent measurements improve estimates of reproducible perturbation responses. a, Per-compound excess agreement in BBBC047 CP, the same-compound Pearson correlation coefficient (PCC) minus that of matched different-compound controls, when each side is one measurement or the average of two independent measurements (18,082 compounds). Vertical lines mark medians; the paired change is the median within-compound difference. **b,** Agreement gain of IMR over the mean of the same number of support measurements. Filled symbols and bars show means and 95% CIs over compounds; open circles show the three training runs. Compounds with one to three support measurements: BBBC047 CP, 4,023, 3,982 and 3,588; cpg0004 CP, 127, 120 and 120;

BBBC047 GE, 3,691 and 3,121. **c,** Change from the support mean in agreement with the held-out measurement of the same compound (filled) and with measurements of other compounds from the support plates (open) for the three ReCA components in BBBC047 CP. The gap is the specificity gain ΔE (printed; 95% CIs in Supplementary Table S1). **d,** Perturbation specificity (E) of all methods on the same BBBC047 CP task (3,149, 3,124 and 2,832 compounds with one to three support measurements). All methods receive the same support profiles; the dashed line marks the support mean. ResNet, FT-Transformer and TabM use the larger reference configurations, and MLP is the capacity-matched feed-forward control. Paired contrasts with 95% CIs are given in Supplementary Tables S4 and S5. **e,** Retrieval of the held-out measurement of the same compound among 33 candidates (mAP@33) for the support mean, ReCA and cpDistiller-C on the compounds in **d**. Points and bars show means and 95% CIs. Agg. IMR, aggregated-target IMR; CI, confidence interval; CP, Cell Painting; EB, covariance-based empirical-Bayes shrinkage; GE, gene expression; IMR, independent-measurement regression; mAP@33, mean average precision over 33 candidates; MLP, multilayer perceptron; PCA, principal component analysis; ReCA, reproducibility-calibrated aggregation.

Figure 3 | A second modality adds compound-specific, conditional evidence. a, Gain from same-compound cross-modal evidence over three comparators: the target-only estimate, randomly reassigned (shuffled) evidence and evidence from a different compound measured under matched conditions. Glyphs show whether a comparator adds evidence and whether that evidence comes from the same compound and a matched condition (filled, yes; open, no; dash, not applicable). GE$\rightarrow$CP denotes GE evidence added to a CP estimate; CP$\rightarrow$GE denotes CP evidence added to a GE estimate. Points and bars show mean paired differences and 95% CIs over compounds (GE to CP, 4,023 compounds and 3,538 for the condition-matched comparator; CP to GE, 3,691 compounds). **b,** Gain in GE agreement over the same target-only GE estimate when GE and CP are joint inputs to one model and when CP updates the GE estimate conditionally, and the difference between the two (3,691 compounds). Filled symbols and bars show means and 95% CIs over compounds, and open circles show the three training runs. **c,** Sequential additions of CP and GE evidence. Nodes represent CP response estimates constructed from one or two CP measurements, with or without GE evidence. Purple horizontal arrows add a second CP measurement; orange vertical arrows add GE evidence. Arrow labels show mean gains in agreement with the same held-out CP measurement across 3,983 compounds. All gains are in Fisher-z units. CI, confidence interval; CP, Cell Painting; GE, gene expression.

Figure 4 | Recovered responses preserve morphological structure supported by independent references. In cpg0004-LINCS, Raw is a single support measurement and ReCA the one-support estimate derived from it; the reference is the mean of four other measurements of the same compound–dose condition. **a,** Agreement with the reference by feature group (Fisher z). Left, Raw and ReCA means; right, paired difference with 95% CI and the percentage of compounds that improved. All groups is the mean over the seven groups. ER, endoplasmic reticulum; Mito, mitochondria; AGP, actin, Golgi and plasma membrane. **b,** Overlap with the reference's 25 largest feature changes and agreement in their direction. Grey lines show compounds; points and bars show means and 95% compound-bootstrap intervals; Δ is the paired mean difference. **c,** Dose concordance of each compound for Raw and ReCA; orange marks compounds that improved. The dashed diagonal marks equal concordance, and Δ is the paired mean difference with its 95%

CI. Analyses include 291 compounds in **a** and **b** and 289 in **c**. Scores were aggregated within compound across doses, support rotations and seeds, and intervals use 10,000 compound-bootstrap resamples. CI, confidence interval; ReCA, reproducibility-calibrated aggregation.

Figure 5 | Reproducibility-aware training targets improve unseen-compound prediction in BBBC047. The Raw predictor was trained on the mean of two, three or four selected measurements per compound (2R, 3R and 4R). ReCA+Res was trained on the same measurements, with separate branches predicting the mean one-support ReCA estimate and the residual (Res). Both were evaluated against the same measured profiles of unseen compounds. **a,** Mean score of each predictor: Pearson correlation coefficient on full profiles (PCC), PCC on control-subtracted responses ($\triangle$PCC), root mean squared error (RMSE) and perturbation specificity E. Lower is better for RMSE and higher for the other metrics. **b,** Paired gain of ReCA+Res over Raw, signed so that positive values favour ReCA+Res (for RMSE, the decrease); printed values are the mean gains. Points and bars show means and pointwise 95% CIs from 10,000 compound bootstrap resamples. PCC, $\triangle$PCC and RMSE were evaluated on 4,040, 4,000 and 3,606 compounds at 2R, 3R and 4R, and E, which requires matched measurements of other compounds, on 3,915, 3,874 and 3,478. CI, confidence interval; ReCA, reproducibility-calibrated aggregation.

Supplementary Files

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

- SupplementaryData2.xlsx
- SupplementaryInformation.pdf
- SupplementaryData1.xlsx