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.

Reference-cell design shapes model evaluation in single-cell perturbation prediction

Hao Zhou

hao.zhou@ku.ac.ae

Khalifa University https://orcid.org/0000-0001-7131-3758

Luying Su
Khalifa University

Minghao Liu
Khalifa University

Assina Abdussaitova
Khalifa University

Barbara Tarantino
University of Pavia

Paolo Giudici
University of Pavia

Antonello Maruotti
Khalifa University

Dean Everett
Khalifa University

Gregory Fonseca
Department of Medical Sciences Khalifa University

Article

Keywords:

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11119348/v1

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

Additional Declarations: There is **NO** Competing Interest.

Article

Reference-cell design shapes model evaluation in single-cell perturbation prediction

Luying Su[3,6], Minghao Liu[1,6], Assina Abdussaitova[1,2,6], Barbara Tarantino[4], Paolo Giudici[4], Antonello Maruotti[2,5,7], Dean Everett[2,5,6], Gregory Joseph Fonseca[2,3,6*], Hao Zhou[1,2,6*]

[1]Department of Biological Sciences, College of Medicine and Health Sciences, Khalifa University, Abu Dhabi, United Arab Emirates.
[2]Center for Biotechnology (BTC), Khalifa University, Abu Dhabi, United Arab Emirates.
[3]Department of Medical Sciences, College of Medicine and Health Sciences, Khalifa University, Abu Dhabi, United Arab Emirates.
[4]Department of Economics and Management, University of Pavia, 27100, Pavia, Italy.
[5]Department of Public Health and Epidemiology, College of Medicine and Health Sciences, Khalifa University, Abu Dhabi, United Arab Emirates.
[6]Infection Research Unit, Khalifa University, Abu Dhabi, United Arab Emirates.
[7]Department GEPLI, LUMSA University.

*Corresponding author(s). E-mail(s): gregoryjoseph.fonseca@ku.ac.ae; hao.zhou@ku.ac.ae;

Abstract

Single-cell perturbation benchmarks use control cells as model inputs and as baselines for predicted and observed responses. We show that reallocating these cells can reverse model rankings without changing fitted parameters or the scoring metric. Across eight model families in two peripheral blood mononuclear cell datasets, separating observation controls reversed 10 of 28 comparisons under both training regimes at the largest tested reference depths. Under balanced, equal-depth assignments scored before averaging, an exact squared-error relation links shared-to-disjoint ranking reversals between state-output and fixed direct-effect predictions to their initial margin and reference separation. Balanced input reassignment changed predictions without changing mean scores under shared scoring references. Reference-dependent selection changed secondary expression-program predictions, while observation references reversed the selected models' mean program-error ordering in 39 additional same-study donors. Refara separates input-dependent prediction changes from rescoring and assesses comparison stability across evaluated allocations. Control assignment should therefore be specified alongside models and metrics.

Single-cell perturbation models predict transcriptional responses to genetic or environmental changes. Benchmarks guide which models and predicted responses are taken forward for biological interpretation.[1–3] Predictions can also guide experiments: CellOracle used in silico perturbations to identify regulators of zebrafish axial mesoderm development, followed by experimental validation.[4]

Some models predict expression changes directly; others predict treated states that require a control baseline to define effects. For example, biolord generated infection-state counterfactuals for control hepatocytes and compared them with their observed expression to identify infection-associated genes and pathways.[5] In a prospective test, the experimental response may be estimated from controls sampled

separately from those available at inference. These samples can yield different baselines even when they come from the same untreated population.

Systema shows that subtracting a common reference from fixed predicted and observed expression leaves mean-squared error (MSE) unchanged.[6] Shared-reference correction addresses score distortion in correlation-based evaluation, while evaluation taxonomies distinguish representation, metrics and reporting choices.[7, 8] We treat model-input controls, prediction baselines and observation baselines as separate design choices. This allows us to distinguish changes in predictions from changes caused by rescoring them, and to ask when those choices reverse model comparisons with fitted parameters held fixed.

We distinguish three control roles: the observation reference defines the measured response, the prediction reference converts predicted states into effects, and the conditioning reference supplies model inputs. We call their assignment the *reference-cell design*. Our primary observation-separated evaluation shares model-input and prediction controls and uses a separate observation block. Under this design, a direct-effect prediction and its equivalent input-conditioned state receive the same score. Fully disjoint designs additionally separate the prediction baseline as a sensitivity test. Rescoring saved predictions measures scoring-reference sensitivity, whereas input audits require predictions generated from the corresponding controls.

We derive an exact squared-error condition for ranking reversals between shared and fully disjoint references, relating the initial model-performance margin to reference separation. The condition applies to fixed direct-effect predictions and balanced, equal-depth assignments scored before averaging. We test it across eight model families and two training regimes in each of two peripheral blood mononuclear cell (PBMC) datasets, then examine how reference-dependent selection changes predicted expression programs. We extend the analysis to unseen genetic combinations and ask how evaluation-unit and test choices affect model comparisons. Refara implements these audits and identifies pairwise model comparisons whose ordering persists across evaluated control allocations.

Results

Reference roles define the effect being evaluated

Control cells enter effect prediction as model inputs, as the baseline subtracted from predicted expression and as the baseline defining the observed effect (Fig. 1a,b). A state prediction $M(C_{\text{model}})$ is therefore scored as $M - C_{\text{pred}}$ against $Y - C_{\text{obs}}$; a direct-effect predictor supplies the effect directly.

Our primary observation-separated design (O) shares a control block between model input and prediction centring, while a non-overlapping block defines the observed effect. A direct-effect prediction d and its equivalent state $C_{\text{model}} + d$ therefore receive identical standardized squared-error scores for each assignment:

$$\|d - (Y - C_{\text{obs}})\|_s^2 = \|[C_{\text{model}} + d - C_{\text{model}}] - (Y - C_{\text{obs}})\|_s^2. \tag{1}$$

This equality holds before averaging, so changing the output representation alone leaves the comparison unchanged.

To compare reference designs, we formed three equally sized, non-overlapping control blocks and reassigned them among roles. The shared design (S) used one block for all roles; the observation-separated design used two. Utility was negative standardized MSE, so higher values indicated lower error. The direction of a pairwise comparison is the sign of the difference in model utilities. We scored assignments before averaging, giving each block equal weight in every role. Model parameters, treated profiles, gene weights and predictions for each input block remained fixed (Methods).

The fully disjoint design (D) used a different block for each role (Fig. 1a). Under the balanced assignments above, moving from S to D reduced the mean utility of every state-output model by exactly V, whereas fixed direct-effect predictions retained the same mean utility. Here, $V \geq 0$ is the mean squared standardized separation between control-block profiles. Thus, the direct-effect-minus-state utility margins satisfy

$$d_D = d_S + V, \qquad d_S < 0 < d_D \iff -V < d_S < 0, \tag{2}$$

A direct-effect model therefore overtakes a state-output model from S to D when its initial deficit is smaller than V. Methods defines the allocation conditions and V; Supplementary Methods gives the derivation.

Observation separation reverses comparisons across training regimes

We evaluated eight model families in paired mock and influenza-infected PBMCs (Influenza PBMC) and a multi-cytokine PBMC screen (Parse) (Fig. 2).[9–15] Training used 8 or at most 48 cells per donor–cell-type–condition stratum in Influenza PBMC, and 48 or at most 128 in Parse; the larger panels used all eligible cells up to these caps (Methods). Configurations were selected using training data only, and each regime was evaluated on the same eight held-out donors. At eight reference cells per role in Influenza PBMC and 48 in Parse, separating the observation reference reversed 10 of 28 mean-utility comparisons in each of the four dataset–training settings (Fig. 2b,c). These reversals paired each of the five direct-effect families with scGen or CellOT. The reversed pair sets were unchanged between training regimes at matched reference depths. For the focal PCA–scGen comparison, the ranking reversed in all eight evaluation donors in every training group and at every tested depth.

Using shared rather than observation-separated validation changed two of 32 configuration choices: scGen and CellOT in the 48-cell Parse training group. At the largest tested reference depths, both selectors retained the same 10 of 28 reversals between shared and fully disjoint scoring in every training group, and the leading model under each shared, observation-separated or fully disjoint design was unchanged. Both the 320-candidate grid and the expanded 328-candidate pool gave these results (Supplementary Table 1).

Increasing reference depth reduces overall shifts in pairwise margins

The same model pairs reversed under prediction-reference separation and fully disjoint references. At four reference cells per role, the no-change baseline (NC; zero predicted effect) also overtook CPA under these designs and observation separation, increasing the count from 10 to 11 of 28 pairs in both Influenza PBMC training groups. All other depths in both resources gave 10 of 28. Separating only model input produced no mean-utility ranking reversals. Increasing reference depth reduced the RMS change across all 28 pair margins for each design with separated scoring references (Fig. 2c).

Standardized MSE and unstandardized delta-RMSE gave identical donor-averaged model rankings in all 80 evaluated designs (Extended Data Fig. 1). Some donor-specific rankings differed between criteria (Supplementary Methods).

Score changes from input reassignment depend on scoring references

Changing input controls altered predictions from the selected scGen, CPA and CellOT fits used in Fig. 2. With observation and prediction references fixed, we replaced an initial, non-overlapping input block with either scoring-reference block. At the largest reference depths, equal-donor mean standardized RMS state-prediction changes ranged from 0.084 to 0.124 in Influenza PBMC and from 0.044 to 0.052 in Parse (Fig. 3a). After balanced averaging, using prediction-reference cells as input increased mean utility; using observation-reference cells decreased it. With shared scoring references, balanced input reassignment changed predictions without changing mean utility (Fig. 3b).

After subtracting each input's own baseline, displacement was close to the reference-only displacement (Fig. 3c). Re-expressing the same predictions and targets in equivalent coordinates preserved MSE to numerical precision in 15,552 checks. All three state-family mean orderings remained stable at every depth in all four training groups. In Kang, input reassignment changed the sign of the mean ridge–CellFlow margin both with the original fixed configurations and after training selection using the same eight donors; both comparisons ended near a tie (Supplementary Fig. 1 and Supplementary Table 2).[16, 17] Input sensitivity and ranking reversal are distinct outcomes.

Reference-dependent model selection changes program predictions and program-error ordering

Reference-dependent model selection changed predicted program directions, and observation controls determined which selected model had lower program error (Fig. 4). With each donor held out from selection, shared references selected scGen in both Influenza PBMC training regimes; observation-separated references selected TW, PCA or RBF across folds (Supplementary Table 3). We compared predictions for the same five cell types against a common measured response. Fully disjoint and observation-separated selection chose the same fitted models in all 16 leave-one-donor-out folds and both all-eight-donor selections, giving identical downstream predictions.

The two selection schemes predicted opposite program directions for 75 of 1,480 task–program pairs under 8-cell training and 65 under the 48-cell cap (Fig. 4a; both absolute scores >0.025). Program priorities also differed across tasks, while the overall mean profiles retained a dominant interferon-α response under both selections (Supplementary Fig. 2). Complete counts and threshold, seed and reference-assignment sensitivities are given in Extended Data Fig. 2a–d and Supplementary Table 3.

Holding selected predictions and treated cells fixed, changing only the observation reference reversed which selection scheme had lower mean program error under both training regimes (Fig. 4b). Shared observation controls favoured shared-reference selection; separated observation controls favoured observation-separated selection. The same reversal occurred in 39 additional same-study donors (Fig. 4c), evaluating all 37 programs in B cells and combined monocytes with batch- and capture-balanced averaging. Exact error contrasts and cohort, weighting and observation-depth sensitivities are given in Extended Data Fig. 2e,f and Supplementary Tables 3 and 4.

The consequences of reference design for model selection differed across the evaluated settings. Whereas shared and observation-separated selection chose different models in Influenza PBMC, both selected PCA in the external GSE181897 analysis,[18] producing identical predictions for 46 assessment donors (Supplementary Methods).

Reference design also reverses genetic-perturbation rankings

For 55 unseen two-gene combinations in the Norman K562 CRISPRa screen,[19] we compared a training-selected direct-effect ridge with three state-output predictors using matched features and scores (Extended Data Fig. 3). With direct-effect predictions fixed, task-averaged comparisons with additive and compositional ridge satisfied $-V < d_S < 0$ at 64 and 128 controls per role and reversed from S to D in all 30 allocations: four of 18 pair-by-depth comparisons. The direct-effect predictor led in task-averaged utility under both designs at greater depths and against training-selected CPA at every depth. State-output pair margins were unchanged from S to D.

Evaluation-unit and test choices change model comparisons

Reallocating control cells does not create biological replication. We therefore asked whether model comparisons also depended on which evaluation units were treated as independent and which tests were applied. In simulations, Holm-corrected exact-sign testing controlled false-direction errors but had low power with few independent units; unit-level t tests stayed near nominal error under the Gaussian null. Unstudentized bootstrapping of whole units with few units available, and treating related tasks or cells as independent, inflated errors (Extended Data Fig. 4). Acquisition links and incomplete metadata also limited testable units in plate-based perturbation data (Extended Data Fig. 5) and neural differentiation (Supplementary Methods),[20–22] with metadata gaps across 30 benchmark records (Supplementary Fig. 3).

Equal-sized evaluation subsets also yielded different decisions. In Sound Life day-7 versus day-0 B-cell and plasma-cell prediction, alternative maximum donor–batch matchings changed directional decisions for three model comparisons with predictions fixed (Fig. 5).[23, 24] Two 39-wave matchings gave RBF 28 and 30 wins against two-way ridge, with nominal full-family adjusted $P = 0.166$ and 0.0186, respectively above and below 0.05. Excluding training-linked batches left 144 maximum 12-wave matchings: NC led in 67 and RBF ridge in 77.

Refara audits reference sensitivity and comparison stability

Refara audited six prediction configurations on 55 Norman tasks. For CPA, input reassignment used predictions generated from the corresponding controls with parameters fixed; scoring-reference reassignment reused predictions. We measured sensitivity as the mean absolute change in task-averaged standardized MSE. At 64 controls per block, changing either scoring reference had about 2.7 times the sensitivity of changing the input (Table 1). At 1,080 controls per block, the ratio fell to about 1.3; all three sensitivities decreased with depth. Equivalent representations and input reassignments with fixed predictions preserved scores (Supplementary Methods).

We calibrated a reporting rule on 27 Norman tasks and assessed 28 others using standardized effect MSE, the same control pool and evaluated control allocations. At 192 controls and a 0.01 reporting gap, single-split scoring reported 323 of 420 directions; 65 failed to maintain the required margin in assessment. Calibrated reporting retained 165 directions, all exceeding the gap throughout evaluated assessment realizations. Geometry scaling gave identical decisions; other gaps yielded unsupported directions (Supplementary Methods, "Stability calibration and comparison reporting").

Discussion

Reference-cell assignment is part of the specification of a perturbation benchmark. Controls supply model inputs, define observed responses and convert predicted states into effects; changing their allocation can change model comparisons even when fitted parameters and the metric stay fixed. A score therefore describes performance under a particular reference design. Declaring that design makes the basis of a model comparison explicit.

The three control roles explain why reference changes affect model families differently. A common prediction and observation baseline cancels from the squared-error residual. Under the balanced design studied here, fully disjoint references reduce state-output utility by the mean squared standardized separation between references, while fixed direct-effect predictions retain their mean utility. The initial performance margin then determines whether a ranking reverses. Our primary observation-separated design matches model input and prediction centring, preserving equal scores for equivalent direct-effect and state representations. Fully disjoint evaluation tests sensitivity to an additional change in the prediction baseline.

Prediction sensitivity and score sensitivity require separate audits. Input changes require predictions generated from the corresponding controls; scoring-reference changes can be evaluated by rescoring saved outputs. Balanced averaging can conceal prediction changes in an unchanged mean score, so both quantities should be reported. Reference-dependent program predictions can motivate biological follow-up, but require separate validation. For example, an ordering of interferon-γ program scores between B cells and monocytes suggested by earlier fixed-model predictions did not meet its prespecified measured-response stability criterion (Supplementary Table 5). Refara records these distinctions through explicit control roles, cell memberships, prediction changes and model comparisons.

Reference stability and biological replication address different questions. Cell reallocations and training seeds measure sensitivity within an experiment; they do not add independent biological units. The 39 additional donors came from the same study and shared acquisition batches with the original cohort. Sound Life further showed that choosing different admissible evaluation subsets could change leading models and nominal test decisions even with predictions fixed. These findings link reference audits to the definition of evaluation units and the assumptions of the statistical test. Avoiding reuse of donors or batches alone does not establish independence.

The scope of these results is set by the evaluated configurations, training budgets and reference pools; some CPA fits were still improving at the budget boundary (Supplementary Table 6). The exact reversal relation requires squared-error scoring and the stated balanced design. Reporting calibration used held-out tasks from the same control pool and evaluated allocations; its ranges describe sensitivity within those allocations, not confidence bounds for new experiments. Benchmark reports should give both the comparisons retained and those left unresolved. Control assignment should therefore be treated as part of the benchmark specification. Reporting the controls used for each role, together with prediction and score sensitivity across plausible assignments, makes model comparisons more interpretable and reproducible.

Methods

Study design and experimental units

We evaluated fitted perturbation models while reassigning the control cells used as model inputs, prediction references and observation references. The primary Influenza PBMC, Parse and Kang analyses each used eight donor-level evaluation units, termed roots. Tasks combined cell type and perturbation within a donor. Control allocations, block permutations and training seeds reused the same biological samples; they did not add biological replication. The Norman tasks came from one pooled perturbation screen.

The study was not preregistered. The original 84-contrast family was specified before inspection; its focal model pair, the later 56 construction-shift contrasts and follow-up analyses were chosen after earlier results. Supplementary Methods gives the timing and available data for each analysis.

PBMC datasets and preprocessing

The Influenza PBMC dataset of Randolph et al. (GSE162632) contains paired mock and 6-h influenza-A-virus samples from 90 donors in 15 processing batches.[9, 10] A deterministic metadata- and cell-support-based rule assigned eight batches to evaluation and seven to training. Evaluation required at least 24 mock and eight treated cells in each of five tasks: B cells, CD4 T cells, CD8 T cells, combined monocytes and combined NK cells. We selected one eligible donor per evaluation batch, with four donors annotated as AFR and four as EUR, before examining response outcomes. Of 42 donors in the training batches, 40 met the training support rule and supplied 200 donor–task rows. Training and evaluation shared no donor, processing batch or capture pool. Supplementary Methods gives the donor-selection algorithm.

Parse models were fitted using four donors and evaluated on eight other donors from the same experiment; donor identity was confounded with donor-specific plate. Metadata-defined cell support across all 12 donors yielded 722 cytokine-by-cell-type tasks. Evaluation controls comprised 88 donor–cell-type pools, with 144 PBS cells per pool reused across cytokines. Primary scores weighted the 722 tasks equally within each donor, then the eight donors equally; cell types therefore received weight in proportion to their eligible task counts.

For both resources, we transformed each cell to counts per 10,000 followed by $\log(1+x)$. We selected 2,000 genes by decreasing sample variance among training untreated cells, broke ties by source-column order and restored selected genes to that order. Gene scales s_g were training-control sample standard deviations, floored at 0.1. Features, scales and fitted parameters were held fixed during evaluation.

Model training, selection and output adapters

The eight model families were no change, context mean, two-way additive ridge, PCA ridge, RBF kernel ridge, CPA, scGen and CellOT. The first five emitted direct perturbation effects; the last three emitted treated expression states. Within each donor–cell-type–condition stratum, including untreated controls, the smaller training panels used 8 cells in Influenza PBMC and 48 in Parse. Larger panels used all eligible cells up to 48 and 128, respectively: $n_{\text{used}} = \min(n_{\text{eligible}}, n_{\text{cap}})$. Smaller panels were nested within larger panels, and all models in a group used identical cells. Training-panel size and evaluation reference depth were varied separately.

Validation held out each of seven training batches in Influenza PBMC and each of four training donors in Parse. Feature learning used only the fitting donors within each fold. We selected each family's configuration by standardized effect MSE under an observation-separated, two-block design: model input and prediction centring shared one control block, while the other defined the observed effect. We scored both orientations, then averaged errors over orientations, tasks within donors and donors equally. Training was extended when the final checkpoint had the lowest validation error and improved by at least 1% over the preceding checkpoint, within the limits specified in Supplementary Table 6. Selected stochastic configurations were refitted with seeds 17, 29 and 43; deterministic models were fitted once. Supplementary Methods describes the candidate grids, budget amendments and alternative S-versus-O selection within identical candidate pools (Supplementary Table 1).

CellOT selection compared direct transport in the 2,000-gene space with transport through a 50-dimensional autoencoder. All four training groups selected the autoencoder. Each encoder and decoder was fitted within its training fold and held fixed during transport validation. Our scGen implementation used a cell-type-specific latent perturbation shift that averaged treated-minus-untreated latent mean

differences equally across training donors. Prediction encoded untreated cells by their posterior latent means, added the shift and decoded before averaging cells. Its outputs were already on the scored log-expression scale. CPA count-space outputs were normalized over the full native feature axis and log-transformed per cell before averaging. State outputs were centred once, with no further normalization or logarithm after centring. Supplementary Tables 6 and 7 and Supplementary Methods specify the fitted configurations and output adapters.

Reference assignments and scoring

For a donor–task unit, let Y_g be the treated-cell mean and $\overline{C}_{\mathrm{obs},g}$ the observation-control mean. The observed effect was $O_g = Y_g - \overline{C}_{\mathrm{obs},g}$. A state-output model generated $M_m(C_{\mathrm{model}})$ from its assigned input controls and was scored as

$$P_{mg} = M_m(C_{\mathrm{model}})_g - \overline{C}_{\mathrm{pred},g}, \qquad U_m = -\frac{1}{G}\sum_{g=1}^{G}\left(\frac{P_{mg} - O_g}{s_g}\right)^2. \tag{3}$$

Direct-effect models were scored as emitted, $P_{mg} = D_{mg}$. Model-input controls were not subtracted again during scoring. Larger utility indicated better prediction; $G = 2{,}000$ for Influenza PBMC and Parse.

We evaluated 1,000 control allocations per resource at depths 4, 6 and 8 in Influenza PBMC and 16, 24, 32, 40 and 48 in Parse. Within each untreated stratum, cells were deterministically hash-ranked and distributed cyclically into three disjoint streams. At depth n, the first n cells in each stream formed blocks B_1, B_2, B_3, so lower depths were nested within higher depths for a given allocation. An allocation label specified the blocks across all strata in the resource.

Tuples followed the order $(C_{\mathrm{obs}}, C_{\mathrm{pred}}, C_{\mathrm{model}})$. We scored all 27 assignments before averaging them into five patterns. S averaged the three tuples (B_j, B_j, B_j). M, P and O averaged the six tuples (B_j, B_j, B_i), (B_j, B_i, B_j) and (B_i, B_j, B_j), respectively, for $i \neq j$. D averaged the six permutations of distinct blocks. Every block consequently had weight 1/3 in each role under each pattern. Treated profiles, feature weights and the predictions associated with each input block stayed fixed.

Each fitted seed was scored separately. We averaged task utilities within roots using the resource-specific weights, then role assignments, allocations and fitted seeds, and finally the eight roots equally. A strict pairwise reversal required opposite nonzero mean-utility margins under S and the comparison pattern, using the same pair orientation and depth. All 28 model pairs were retained. Each pairwise comparison within a root required complete data. We calculated the RMS change across the 28 pairwise margins after averaging allocations and roots. Depth-collapsed displays additionally weighted the tested depths equally. These summaries describe the evaluated donors and finite control allocations.

Quadratic identities and reversal criteria

Let b_i be the mean of block i, and define $\|a\|_s^2 = G^{-1}\sum_g a_g^2/s_g^2$. The reference-separation term was

$$V = \frac{1}{6}\sum_{i\neq j}\|b_i - b_j\|_s^2. \tag{4}$$

For a fixed model-input block, pairing the two assignments of the remaining observation and prediction blocks cancels the cross term in the squared error. Equal averaging gives $U_D - U_S = -V$ for state-output models and $U_D = (U_P + U_O)/2$. Under M and S, the common scoring reference cancels and each input block has the same marginal weight, giving $U_M = U_S$. Fixed direct-effect predictions have the same mean utility under all five patterns because observation blocks retain the same marginal weights. Thus a direct-effect-minus-state margin satisfies $d_D = d_S + V$, with an S-to-D reversal when $-V < d_S < 0$. These relations require fixed treated profiles, feature weights and input-specific predictions, equal-depth blocks, equal assignment weights and scoring before averaging. Supplementary Methods gives the P/O expansions, full derivation and allocation-level checks; the historical full-pool Shared construction is defined separately there.

Model comparisons and input-reference audits

We also ranked the selected PBMC models by unstandardized delta-RMSE, $\{G^{-1}\sum_g (P_g - O_g)^2\}^{1/2}$, calculated within each target before aggregation. Both criteria used the same predictions, treated cells and controls. Complete model rankings retained exact ties, and leader sets included every exact maximum. Rank concordance required a complete finite score vector and a defined Kendall τ_b. Supplementary Methods describes the aggregation and undefined-score rules, including the historical Pearson, cosine and retrieval comparators.

Input audits used the selected scGen, CPA and CellOT prediction banks from all four PBMC training groups. With observation block A and prediction-reference block B fixed, we changed model input from C to A or C to B using the predictions associated with those blocks. For state prediction $F(Z)$, we compared raw displacement $F(B) - F(C)$, own-input-centred displacement $[F(B) - \overline{B}] - [F(C) - \overline{C}]$, and reference-only displacement $\overline{B} - \overline{C}$. Squared changes were standardized by the training gene scales and averaged over genes, tasks, all block permutations, 1,000 allocations and three seeds within each root. We then took each root's square root and averaged the eight root RMS values. Utility contrasts used separately scored seeds. Equivalent-representation checks and numerical precision are documented in Supplementary Methods.

The supporting Kang comparison used paired control and interferon-β profiles from eight donors, with B, CD14 Mono, CD16 Mono and CD4 T tasks.[16, 25] Each donor–task supplied 12 stimulated cells and three disjoint four-cell control blocks, evaluated under all six block assignments. Ridge and CellFlow[17] received the same input block while observation A and prediction reference B stayed fixed. CellFlow input changes affected both source cells and their mean principal-component context. Leave-one-donor-out preprocessing used 5,000 training-selected genes and the same count transformation and standardized MSE as above. Alongside the original fixed fits, nested training selection fitted features, scales and models using six donors and validated on the seventh outer-training donor; selected models were refitted on all seven before outer-donor evaluation. The analyses reused the same eight donors, and folds shared training donors. Supplementary Methods gives the preprocessing, adapters, original and selected fits, and diagnostic families (Supplementary Table 2).

Program analysis after model selection

We used the primary training-selected configurations under both Influenza PBMC training regimes. For each original evaluation donor, we selected a model family under S and O using the other seven donors, 1,000 allocations and eight controls per role. Selection averaged utilities from separately scored seeds and allocations, then weighted selection donors equally. For the additional cohort, selection used all eight original donors. Model fitting and checkpoint selection used training donors only; model-family selection was a separate comparison among those fitted configurations. We also recomputed D selections and verified the selected fit identities and downstream predictions against O (Supplementary Methods).

Program scores used the fixed 2,000-gene panel, training scales and 37 eligible MSigDB Hallmark sets.[26, 27] For program H, panel G and effect Δ,

$$z_H(\Delta) = \frac{1}{|H \cap G|} \sum_{g \in H \cap G} \frac{\Delta_g}{s_g}. \tag{5}$$

State-output models used B1 for input and prediction centring, giving $M_m(B_1) - B_1$; direct-effect models retained their emitted effects. We held predictions and treated profiles fixed while defining the observed response with shared B1 or separated B2 controls. For seed s, the O-selected-minus-S-selected program-error contrast was

$$\Delta_{d,s} = \mathrm{mean}_{c,H}\left[(\hat{z}_{O,d,c,H,s} - z_{\mathrm{obs},d,c,H})^2 - (\hat{z}_{S,d,c,H,s} - z_{\mathrm{obs},d,c,H})^2\right]. \tag{6}$$

Negative values favoured O-selected predictions. Errors were calculated separately for seeds 17, 29 and 43, then averaged; deterministic models represented one fit. Original-donor summaries weighted programs and five cell types equally within donor, then the eight donors equally. Profile and direction displays used seed-mean program scores. The displayed direction comparison used an absolute-score threshold of 0.025; top-five program sets were a secondary readout. Supplementary Methods gives tie rules and secondary readouts (Supplementary Table 3).

The additional cohort comprised B-cell and combined-monocyte tasks from 39 donors in the same study. These expression data had been examined in an earlier measured-response analysis; the donors entered neither fitting nor family selection. Each task used 24 controls in three eight-cell blocks and eight fixed treated cells. We averaged programs and cell types within donor, donors within batch-by-capture-orientation strata, orientations within batch and the eight batches equally. Selected TW and PCA models emitted one effect per cell type across donors. Cohort, weighting, program-panel and observation-depth sensitivities are given in Supplementary Methods and Supplementary Table 4.

Genetic-perturbation evaluation

We evaluated 55 unseen two-gene combinations in the Norman K562 CRISPRa screen.[19] A fixed hash split the 110 supported combinations equally between training and evaluation; all 105 single-gene conditions entered training. Additive, compositional-ridge and CPA models[11] emitted absolute states. A zero-intercept ridge model emitted effects, with its penalty selected by fivefold validation within the training combinations. CPA also underwent training-side selection followed by a full-training refit (Supplementary Table 8). Non-targeting controls were separated between training and evaluation within each of eight capture partitions. Thirty allocations supplied three disjoint blocks at six depths, from 64 to 1,080 cells per role. Features and scales used training controls only. Fitted models, direct-effect predictions and treated targets stayed fixed while all S/M/P/O/D assignments were scored before task and allocation averaging. All three direct-effect–state comparisons were retained. The capture partitions came from one pooled experiment; these comparisons describe within-screen sensitivity.

Statistical comparisons and simulations

Primary reference and program comparisons summarize the evaluated donors and evaluated control allocations. For the original fixed-configuration analysis, we additionally calculated root-level contrasts and sensitivity bounds under a repeated-root working model. Calibration failed, so these bounds are reported as diagnostics, not calibrated 95% confidence intervals. Supplementary Methods, "Diagnostic bounds", defines the contrast families, working-model assumptions, multiplier construction and calibration (Supplementary Table 9).

The directional simulation included four methods and all 12 ordered comparisons. For root r, method m, descendant support n_r and $c = (1.5, 0.5, -0.5, -1.5)$, utilities were

$$U_{rm} = \delta c_m + \sqrt{0.4}\, G_r^A + \sqrt{0.4}\, Z_{rm}^A + \sqrt{0.8/n_r}\, G_r^E + \sqrt{0.8/n_r}\, Z_{rm}^E. \tag{7}$$

All G and Z draws were independent standard normals across their distinct indices. Shared G terms induced within-root correlation; descendant support affected precision, while roots remained independent testing units. We varied separation δ, root count, common-root ties, complete-root missingness and descendant support, with 5,000 replicates per setting–δ combination. The full grid and interventions are specified in Supplementary Methods, "Simulation design".

The exact-sign test concerned the conditional root-specific probability $\theta_{ij,r} = P(U_{ri} > U_{rj} \mid$ support, availability and absence of a tie). Available non-tied roots entered a one-sided test of $H_{ij} : \theta_{ij,r} \leq 1/2$ at every prespecified root position.[28] Under independent roots, the null win count is stochastically bounded by Binomial$(K, 1/2)$, where K is the number of non-tied available roots. Rejection indicates failure of this win-probability null at at least one position; mean-utility contrasts are a distinct target. Holm adjustment covered all 12 directions at familywise $\alpha = 0.05$.[29] Fewer than eight non-tied roots cannot reach the first Holm threshold. Complete-family recovery required all six true directions and no false direction; true-edge recovery was the fraction of the six true directions recovered.

To compare analysis units and tests under a global null, we generated

$$Y_{rtcm} = \sqrt{0.8} A_{rm} + \sqrt{0.1} B_{rtm} + \sqrt{0.1} E_{rtcm}, \tag{8}$$

with independent standard-normal terms across distinct indices and models, $R = 2, 3, 4, 6, 8, 12, 20, 40$ roots, eight tasks per root and 20 nominal cells per task. Unequal retained cell counts were aggregated to the same equal-task, equal-root point mean for each procedure. We compared exact-sign and root CR1-t tests with centred, paired, unstudentized root-, task- and cell-level bootstraps, using 2,500 replicates per root count. Task- and cell-level resampling treated dependent descendants as independent.

Supplementary Methods gives the row scaling, 499-resample algorithms, zero-standard-error and failure conventions, and supporting simulations. Binary endpoints used pointwise two-sided 95% Clopper–Pearson intervals;[30] bounded true-edge recovery used Hoeffding intervals. These intervals quantify Monte Carlo error under the simulated law.

Acquisition-unit assessment

For GSE306429, we constructed plate-linkage graphs from public metadata before examining expression or method-comparison results. Connected components defined candidate acquisition units. The 90-direction comparison family required at least 11 non-tied independent roots for the first Holm rejection at $\alpha = 0.05$; processed-cell lineage, control/perturbation support and the justification of independence remained unresolved, so utilities were reported descriptively. In the Jerber day-52 rotenone resource,[22] shared differentiation pools connected the evaluated lines into nine candidate components, below the 11-root requirement for its 56-direction family even if those components qualified as independent. Supplementary Methods, "Acquisition and matching", specifies the graph rules and the separate benchmark-record screen.

Sound Life matching sensitivity

We compared maximum donor–batch matchings using fixed predictions and utilities for 93 same-pool donor–year waves in B-cell and plasma-cell data. Each eligible wave formed an edge between its donor and recorded day-0/day-7 batch; labels had to be nonempty and consistent across visits. A maximum-cardinality matching selected waves without reusing a donor or batch. The unrestricted graph admitted 39-wave matchings; excluding batches represented among the 37 training batches gave a separate graph admitting 12-wave matchings. Supplementary Table 7 gives the eight fitted predictors.

For each wave, day-0 controls were split into three disjoint blocks and rotated across input, prediction-reference and observation-reference roles. The observed effect used all eligible day-7 cells. State outputs used their assigned input block and were centred once; direct-effect outputs stayed fixed. Rotation accuracy was $1 - G^{-1} \sum_g \min(z_g^2, 25)/25$, with $G = 2,000$ and residuals standardized by training day-0 standard deviations floored at 0.1. Wave utilities averaged the three rotations equally, and every selected wave had equal weight.

Every matching used the same eight-model, 56-direction exact-sign/Holm procedure at familywise $\alpha = 0.05$. We enumerated all 144 maximum 12-wave matchings. For the 39-wave graph, exact-rational optimization of additive objectives provided attainable bounds and witness matchings. A direction was classified as rejected under every, no or some maximum matchings. It remained unresolved when a non-rejecting matching was found but a rejecting matching could neither be found nor ruled out. Analogous checks identified possible leading models. These are finite-graph classifications; sign-test validity additionally requires the independence and win-probability assumptions. Supplementary Methods gives the control allocation, graph construction and reconstruction checks.

Software audits and comparison reporting

Refara uses predictions from the corresponding controls for input audits and rescores saved outputs for scoring-reference audits, with model parameters fixed (Table 1; Supplementary Methods).

We calibrated reporting on 27 Norman tasks and assessed 28 held-out tasks from the same screen. Both sets used standardized effect MSE and the same control pool, 30 evaluated control allocations and three input-block choices. For first-minus-second utility $m_{cpb}(r)$, indexed by development case c, model pair p, control budget b and realization r, the repeated-primary range relative to anchor r_0 was

$$B_{pb} = \max_{c,r} |m_{cpb}(r) - m_{cpb}(r_0)| . \tag{9}$$

Assessment anchor directions were reported only when their absolute margins exceeded $B_{pb} + \delta + \tau$, with reporting gap δ and numerical tolerance τ; otherwise they were withheld. Decisions were fixed from development scores and assessment anchors before checking remaining assessment realizations. A reported direction was unsupported if any oriented assessment margin was at most $\delta + \tau$. We recorded reporting coverage and unsupported directions. Supplementary Methods describes the full rule, variants and VCC adapter.

Use of generative AI

OpenAI tools were used to proofread and reorganize the manuscript and code. The authors take responsibility for the work.

Data availability

The study used public datasets from GEO (GSE162632, GSE96583, GSE306429 and GSE133344), ExperimentHub (EH2259), Parse Biosciences, the Jerber study and the Allen Institute.[10, 15, 19, 21, 22, 24, 25] Source accessions and provider links are listed in the references. The external model-selection test used GSE181897.[31]

Source Data and supporting reconstruction materials accompany the submission.

Code availability

Refara source code, a Python package and examples accompany the submission. Author-written reference-design code is distributed under the BSD 3-Clause licence; data and third-party software retain their respective licences. The package will be made publicly available at https://github.com/hz424/Refara during peer review.

References

[1] Peidli, S. et al. scPerturb: harmonized single-cell perturbation data. *Nat. Methods* **21**, 531–540 (2024). https://doi.org/10.1038/s41592-023-02144-y.

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

[3] Wu, Y. et al. PerturBench: benchmarking machine learning models for cellular perturbation analysis. *Adv. Neural Inf. Process. Syst.* **38**, 106937–106977 (2025). https://doi.org/10.52202/085713-3225.

[4] Kamimoto, K. et al. Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**, 742–751 (2023). https://doi.org/10.1038/s41586-022-05688-9.

[5] Piran, Z., Cohen, N., Hoshen, Y. & Nitzan, M. Disentanglement of single-cell data with biolord. *Nat. Biotechnol.* **42**, 1678–1683 (2024). https://doi.org/10.1038/s41587-023-02079-x.

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

[7] Nicol, P. B., Shivakumar, S. & Irizarry, R. A. Spurious correlation inflates performance in single-cell perturbation prediction. Preprint at https://www.biorxiv.org/content/10.64898/2026.05.07.723486v2 (2026).

[8] Schäfer, P. S. L. et al. Towards principled evaluation of single-cell perturbation prediction models. Preprint at https://www.biorxiv.org/content/10.64898/2026.07.23.740433v1 (2026).

[9] Randolph, H. E. et al. Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. *Science* **374**, 1127–1133 (2021). https://doi.org/10.1126/science.abg0928.

[10] Randolph, H. E. et al. Single-cell RNA-sequencing reveals pervasive but highly cell type-specific genetic ancestry effects on the response to viral infection. Gene Expression Omnibus GSE162632 (2021). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162632.

[11] Lotfollahi, M. et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.* **19**, e11517 (2023). https://doi.org/10.15252/msb.202211517.

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

[13] Bunne, C. et al. Learning single-cell perturbation responses using neural optimal transport. *Nat. Methods* **20**, 1759–1768 (2023). https://doi.org/10.1038/s41592-023-01969-x.

[14] Oesinghaus, L. et al. A single-cell cytokine dictionary of human peripheral blood. Preprint at https://www.biorxiv.org/content/10.64898/2025.12.12.693897v1 (2025).

[15] Parse Biosciences. *10 million human PBMCs in a single experiment.* https://www.parsebiosciences.com/datasets/10-million-human-pbmcs-in-a-single-experiment/ (accessed 24 August 2026).

[16] Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018). https://doi.org/10.1038/nbt.4042.

[17] Klein, D. et al. CellFlow enables generative single-cell phenotype modeling with flow matching. Preprint at https://doi.org/10.1101/2025.04.11.648220 (2025).

[18] Hartoularos, G. C. et al. Reference-free multiplexed single-cell sequencing identifies genetic modifiers of the human immune response. Preprint at https://www.biorxiv.org/content/10.1101/2023.05.29.542756v2 (2023).

[19] Norman, T. M. et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365**, 786–793 (2019). https://doi.org/10.1126/science.aax4438.

[20] DeMeo, B. et al. Active learning framework leveraging transcriptomics identifies modulators of disease phenotypes. *Science* **390**, eadi8577 (2025). https://doi.org/10.1126/science.adi8577.

[21] Treacy, D. et al. Perturbational single-cell RNA sequencing of 104 compounds across six immortalized and four primary cell lines. Gene Expression Omnibus GSE306429 (2025). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE306429.

[22] Jerber, J. et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.* **53**, 304–312 (2021). https://doi.org/10.1038/s41588-021-00801-6.

[23] Gong, Q. et al. Multi-omic profiling reveals age-related immune dynamics in healthy adults. *Nature* **648**, 696–706 (2025). https://doi.org/10.1038/s41586-025-09686-5.

[24] Gustafson, C. E. et al. *Longitudinal Dynamics of Health and Age.* Human Immune System Explorer https://doi.org/10.57785/bn17-3h16 (2024).

[25] Crowell, H. L. muscData: Multi-sample multi-group scRNA-seq data. *Bioconductor*, version 1.26.0. https://doi.org/10.18129/B9.bioc.muscData (2026).

[26] Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005). https://doi.org/10.1073/pnas.0506580102.

[27] Liberzon, A. et al. The Molecular Signatures Database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015). https://doi.org/10.1016/j.cels.2015.12.004.

[28] Dixon, W. J. & Mood, A. M. The statistical sign test. *J. Am. Stat. Assoc.* **41**, 557–566 (1946). https://doi.org/10.1080/01621459.1946.10501898.

[29] Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979). https://www.jstor.org/stable/4615733.

[30] Clopper, C. J. & Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404–413 (1934). https://doi.org/10.1093/biomet/26.4.404.

[31] Hartoularos, G. C. et al. Optimal pooled single-cell RNA-sequencing elucidates gene by environment interactions in circulating immune cells. Gene Expression Omnibus GSE181897 (2021). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE181897.

Author contributions

H.Z. conceived the study, developed the methods and software, carried out the investigation and analyses, curated the data, prepared the visualizations and wrote the original draft. L.S. contributed to the analyses, software and visualization. M.L. contributed to the analyses. G.J.F. contributed to study design. All authors reviewed and edited the manuscript.

Acknowledgements

The authors thank the high-performance computing (HPC) team at Khalifa University, especially Sanyo Illikkal and Wadud Miah, for their support.

Funding

H.Z. is supported by Khalifa University FSU Grant 8474000820 and the Center for Biotechnology (BTC), Khalifa University.

Competing interests

The authors declare no competing interests.

Ethics and consent

This study analysed publicly available datasets and involved no new participant recruitment or human-data collection. Ethics approvals and consent procedures are reported in the source studies and provider records.

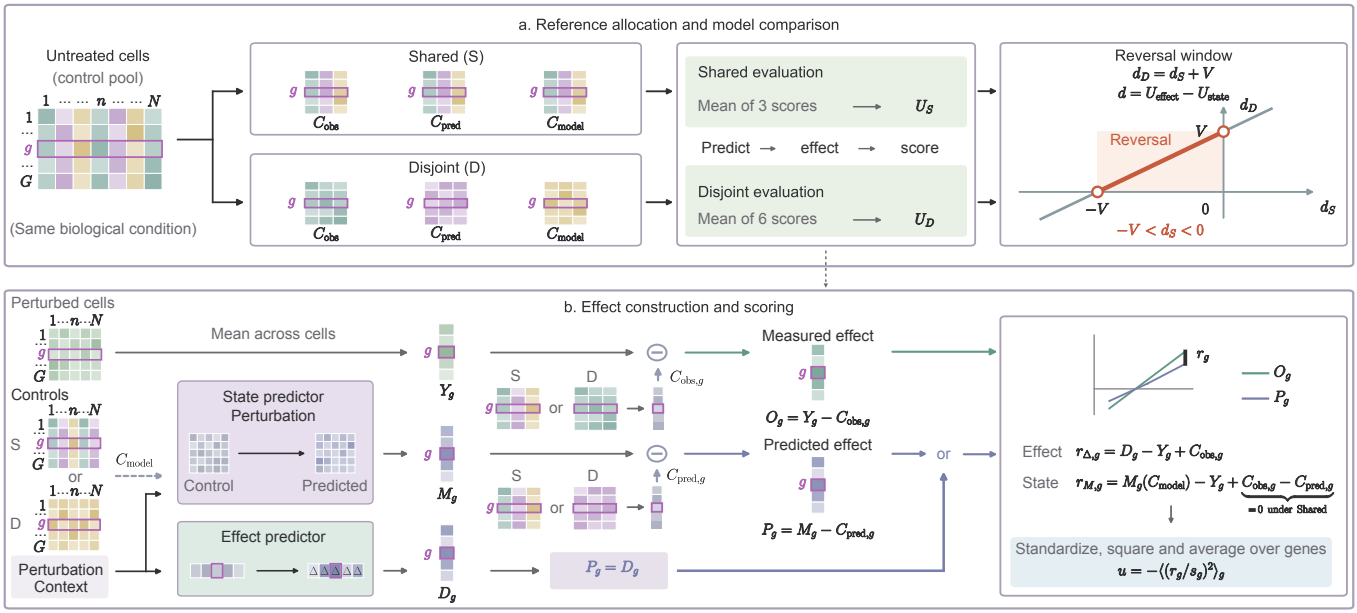


Figure 1 | Reference-cell allocation determines effect scores and ranking reversals
a, Three equal-depth control blocks from one biological condition supply observation reference C_{obs}, prediction baseline C_{pred} and model input C_{model}. Shared (S) assignments use one block for all roles; disjoint (D) assignments use a different block for each. Negative standardized squared-error utilities are scored before averaging three shared assignments or six disjoint permutations. The direct-effect-minus-state margin obeys $d_D = d_S + V$, where $V \geq 0$ is the mean squared standardized distance between block means. The interval $-V < d_S < 0$ gives strict reversals; open endpoints denote ties. Assignment weights are equal; treated expression, gene scales, fits and direct-effect predictions are fixed. **b,** Treated expression Y_g becomes measured effect $O_g = Y_g - C_{\mathrm{obs},g}$. A state output $M_g(C_{\mathrm{model}})$ becomes $P_g = M_g - C_{\mathrm{pred},g}$; a direct-effect output gives $P_g = D_g$. Utility negates the gene-mean squared residual $r_g = P_g - O_g$, standardized by training-derived scales s_g. Shared scoring references cancel explicit baselines in state residuals but can retain model-input dependence. D probes combined reference sensitivity; primary protocol O shares input and prediction centring but separates observation controls. Matrices and lines are schematic; highlighted rows represent one gene.

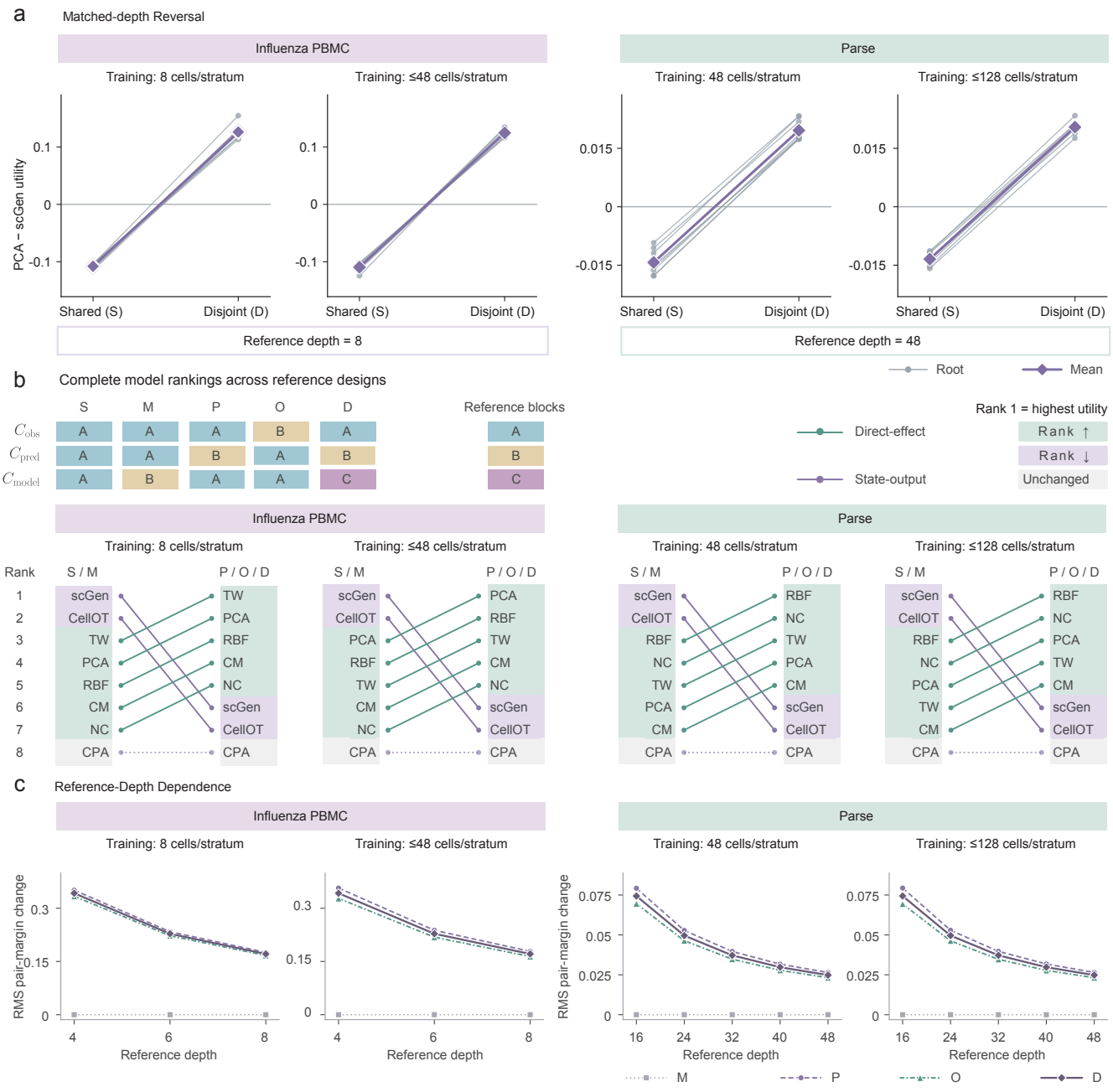


Figure 2 | Reference sharing changes model comparisons
a, PCA-family minus scGen utility under shared (S) and disjoint (D) references at matched depth. Lines pair eight evaluation donors per resource; diamonds are equal-donor means. Training counts are per donor–cell-type–condition stratum, including controls, not reference depths. Caps ≤48 and ≤128 use all eligible cells when fewer are available. **b,** Eight-model ranks at the depths in a; rank 1 is best. S/M and P/O/D have identical ranks here, not necessarily utilities. Teal/purple lines identify direct-effect/state-output families; shading shows improvement, worsening or no rank change from S. Dotted lines mark unchanged ranks. Key letters identify shared blocks: S shares all roles; M/P/O separate input/prediction/observation references; D separates all three. **c,** RMS changes from S in equal-donor-mean utility margins over all 28 pairs, including unchanged pairs. Utility is negative equal-gene standardized MSE, averaged over block permutations, 1,000 allocations and three separately scored scGen, CPA and CellOT fits. Regimes reuse donors; allocations repartition controls and seeds vary initialization. Panels b/c are descriptive. Models use training-side selection (Supplementary Table 6). NC, no change; CM, context mean; TW, two-way ridge; PCA, PCA ridge; RBF, RBF ridge.

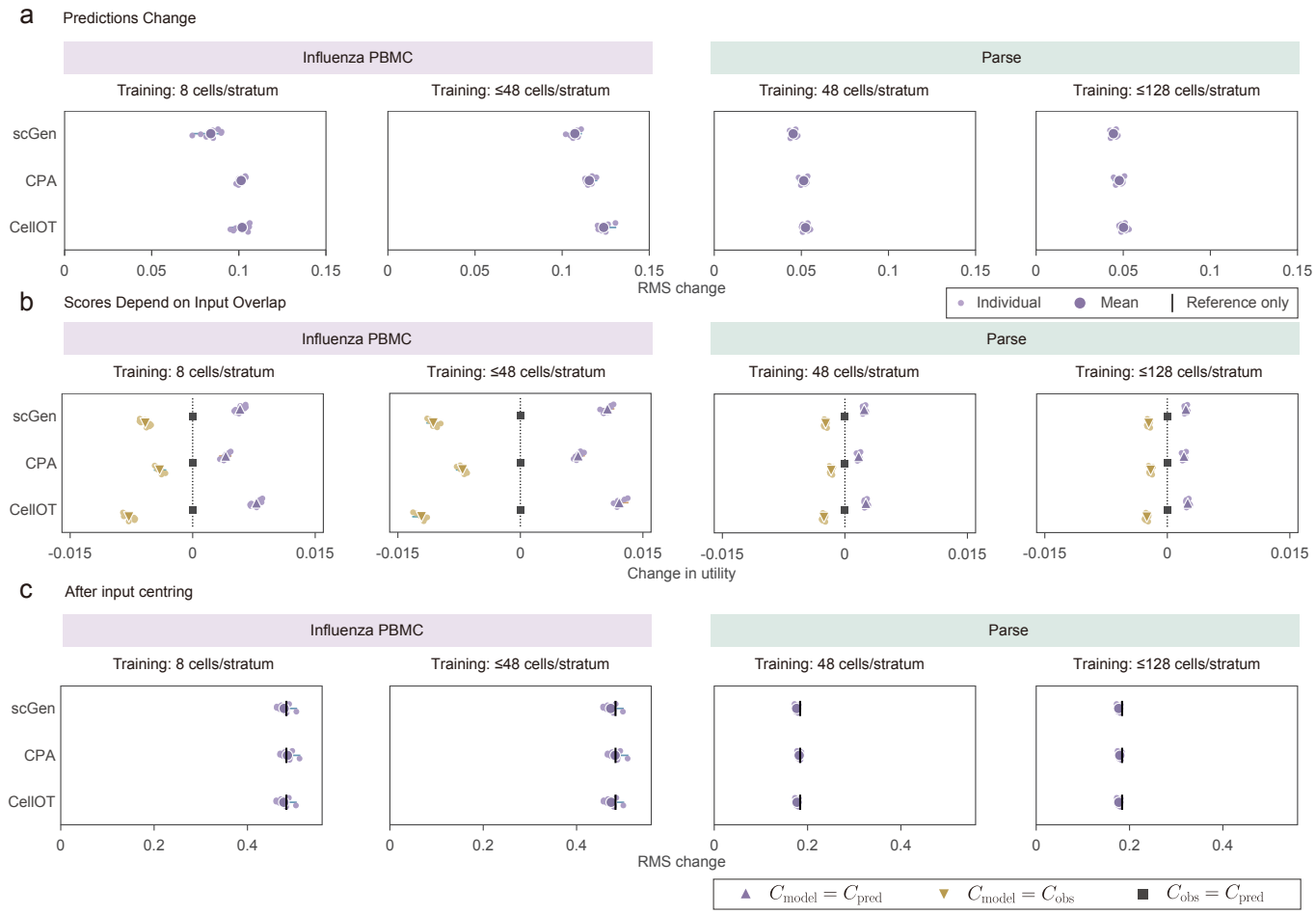


Figure 3 | Model-input changes and their score responses
The Fig. 2 scGen, CPA and CellOT fits receive reassigned inputs with scoring references fixed (cells per role: Influenza PBMC, 8; Parse, 48). **a,** Standardized RMS prediction displacement. **b,** Utility (negative standardized MSE) changes when input matches the prediction or observation reference; squares show balanced input reassignment under shared scoring references. **c,** Displacement after subtracting each input's own mean; black ticks show equal-donor mean reference-only displacement. Small dots are eight donors, large symbols their equal-weight means, and lines their minimum–maximum ranges. RMS follows averaging squared changes within donor; utility averages separately scored errors. Both include three seeds and 1,000 allocations, reusing donor cells. Balanced allocations give identical RMS summaries for the two input changes.

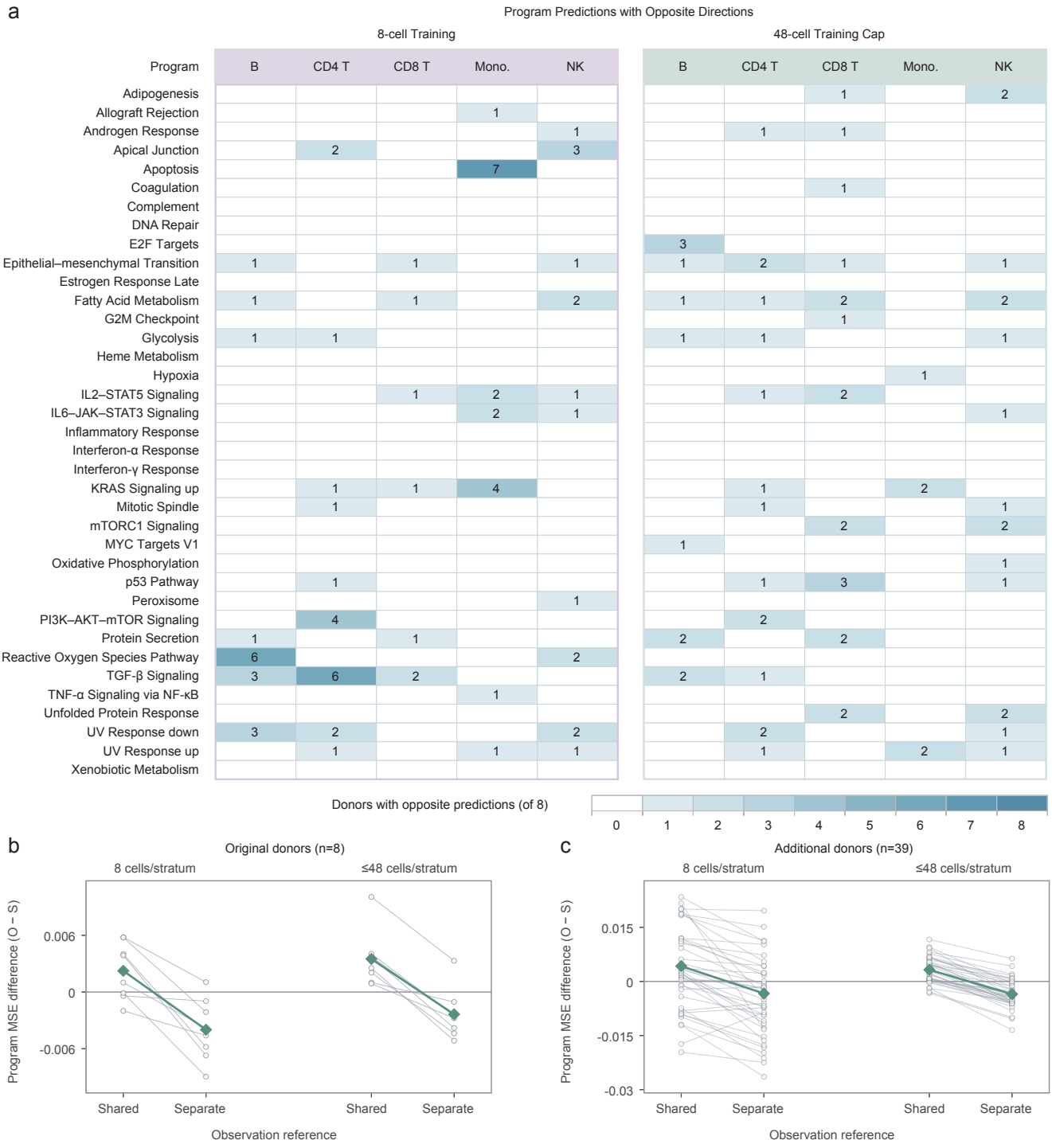


Figure 4 | Reference choices change program predictions and measured model accuracy
Models were selected with shared (S) or observation-separated (O) references (Supplementary Tables 3 and 4). **a,** Donors (of eight) with opposite seed-mean predictions and both absolute scores >0.025, across all 37 Hallmark programs and five cell types. Scores average training-standardized gene effects; white denotes zero. **b,c,** Program-MSE differences (O−S) for eight original (b) and 39 additional same-study donors (c); positive values favour S-selected models. Predictions and treated observations are fixed while observation controls change from shared B1 to separated B2. Lines pair donors; diamonds are equal-donor (b) or batch/capture-balanced (c) means. Errors are scored before averaging three stochastic-model seeds or one deterministic fit. Donors recur across regimes; y-axis ranges differ. O and D selected identical fits and predictions in all 18 settings. Mean profiles and sensitivities are in Supplementary Fig. 2 and Extended Data Fig. 2.

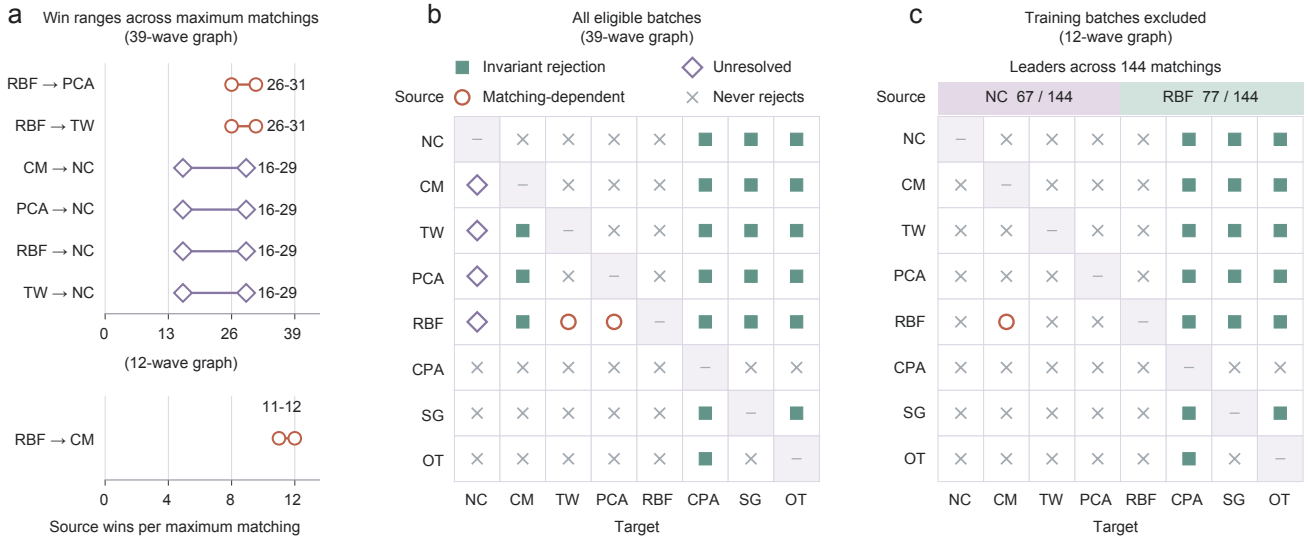
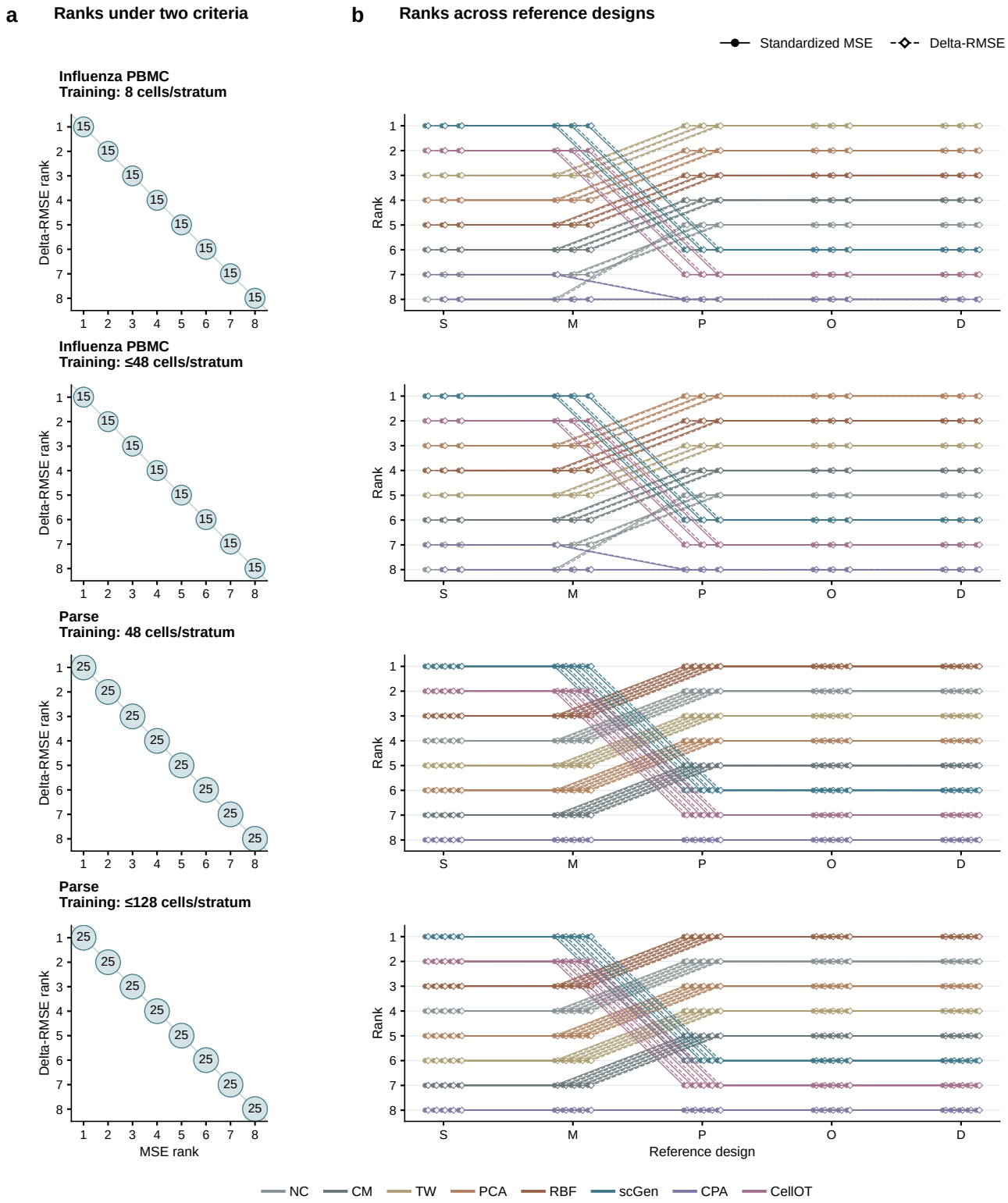
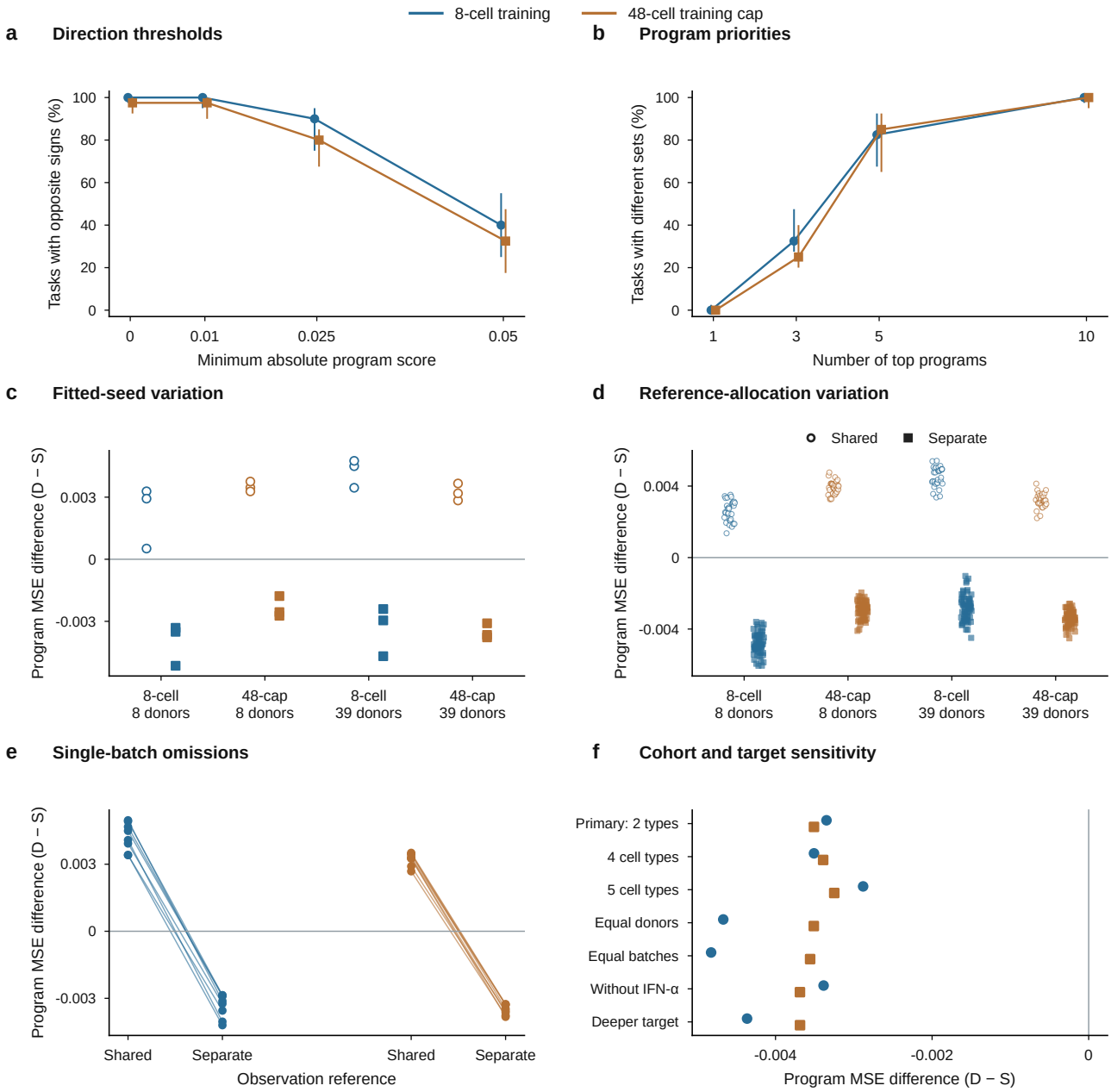


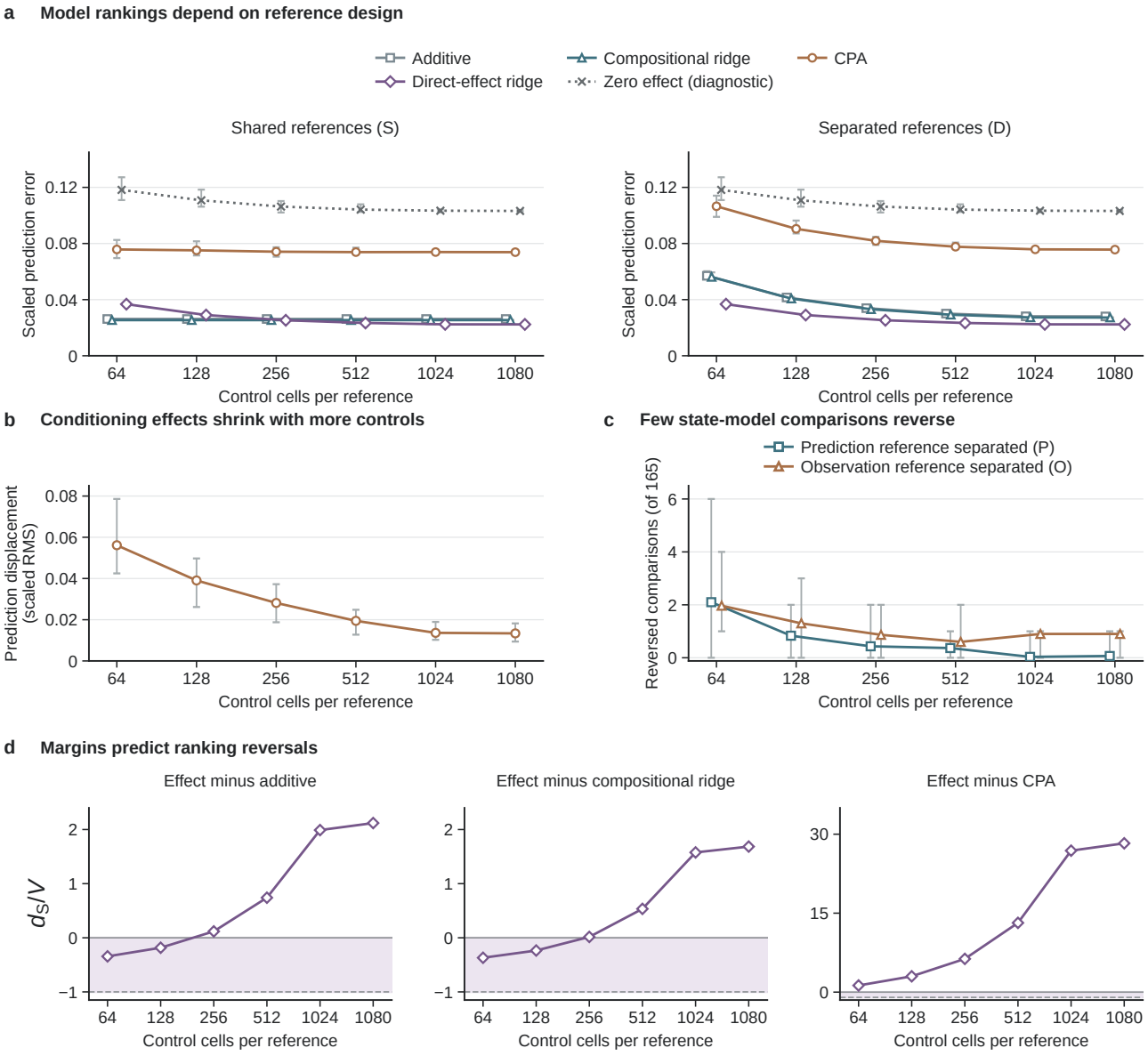
Figure 5 | Maximum matchings can support different directional decisions
Predictions, reference rotations and per-wave utilities are fixed across donor–batch matchings (models: Supplementary Table 7). **a,** Source-model win ranges across admissible maximum matchings for seven directions. Circles mark rejecting and nonrejecting witnesses; diamonds mark unresolved rejection possibilities. **b,c,** All 56 directions for the 39-wave graph and the 12-wave graph excluding training-linked batches. Rows are sources, columns targets. Squares denote invariant rejection, circles matching dependence, diamonds unresolved certification and crosses no rejection. The strip counts leaders across all 144 maximum 12-wave matchings. One-sided exact-sign tests use Holm familywise α=0.05. Inference requires independent units and the declared win-probability null; graph extrema concern recorded admissibility rules. NC, no change; CM, context mean; TW, two-way ridge; PCA, PCA-64 ridge; RBF, RBF ridge; SG, scGen; OT, CellOT.



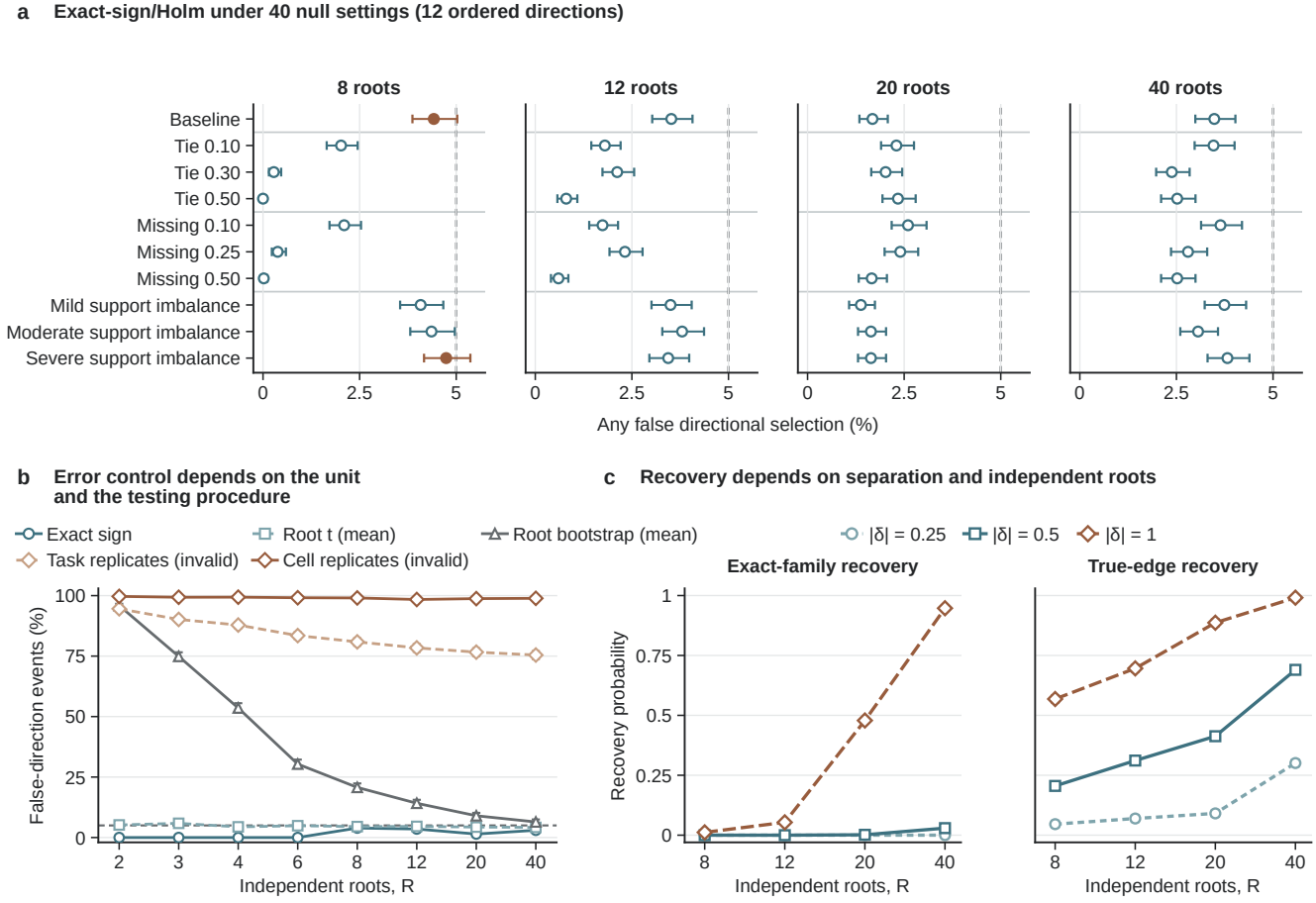
Extended Data Figure 1 | Model rankings under standardized MSE and delta-RMSE
a, Paired ranks of eight models across 80 reference designs (640 model–design pairs), by resource and training regime. Labels give counts; circle area increases with count. **b,** Five role designs at reference depths of 4/6/8 cells (Influenza PBMC) or 16/24/32/40/48 (Parse); offsets indicate increasing depth. Solid lines/filled circles show standardized MSE; dashed lines/open diamonds show unstandardized delta-RMSE. S shares all roles; M/P/O separate input/prediction/observation references; D separates all three. Both criteria use identical predictions and allocations, averaging assignments, 1,000 allocations, separately scored seeds and eight equally weighted roots. Training counts are per donor–cell-type–condition stratum, not reference depths; $\leq$48/$\leq$128 are caps, using all eligible cells when fewer are available. Models use training-side selection (Supplementary Table 6). Ties receive average ranks and retain co-leaders. Seeds and allocations reuse biological roots; root-specific concordance is reported in Supplementary Methods.



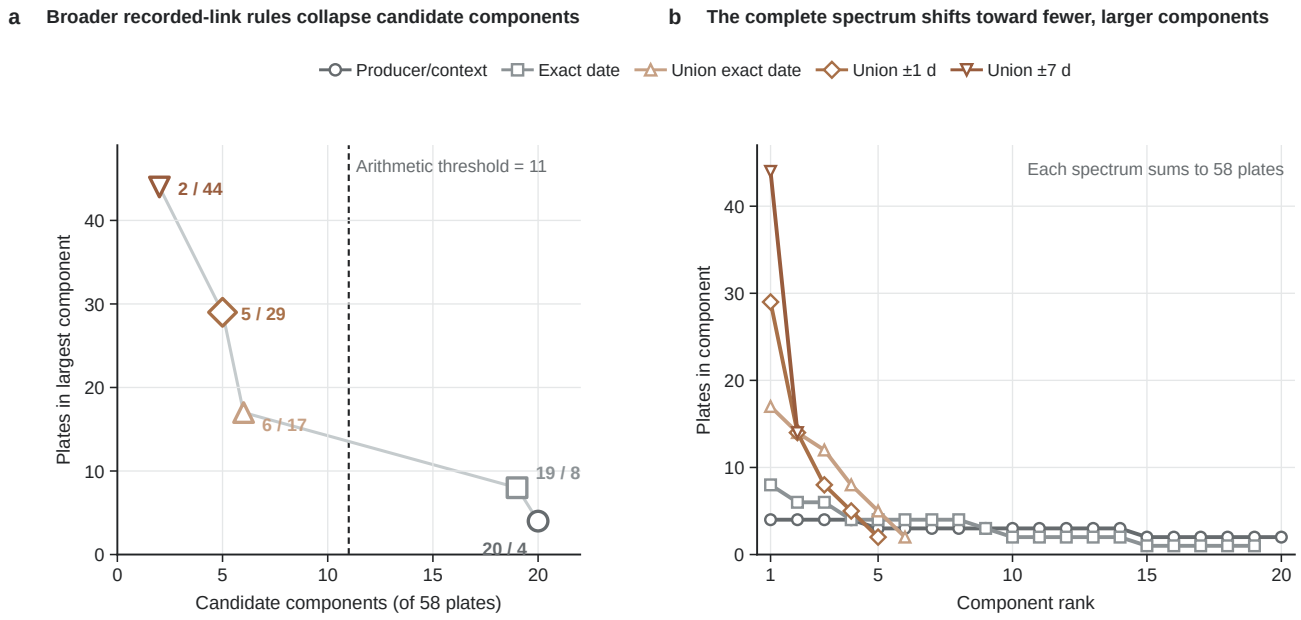
Extended Data Figure 2 | Program-interpretation and accuracy sensitivities under updated training
Blue/orange indicate 8-cell training/48-cell cap (Supplementary Table 6); S/D denote shared/disjoint-reference selection. O selected identical fits and predictions to D in every setting. **a,b,** Percentages of 40 original donor–cell-type tasks with opposite program directions (a; both absolute seed-mean scores exceed the threshold) or different top-k sets (b; ranked by absolute seed-mean score). Points use label 0/input B1; ranges span ten labels and three input blocks. **c,** Program-MSE contrasts for separately scored seeds 17, 29 and 43 in eight original and 39 additional donors. Input/prediction reference is B1; open circles use observation B1, filled squares B2. **d,** All 30 shared and 60 separate assignments, averaging scored seed errors; symbols follow c. **e,** Additional-donor estimates omitting each batch; lines pair observation choices. **f,** Cohort, weighting, program-panel and deeper-target sensitivities (Supplementary Table 4), with label 0/input B1 and observation B2. Positive MSE_D−MSE_S favours S-selected. Errors average 37 programs unless indicated. Original donors receive equal weight; additional donors balance batches and capture orientations (eight batches; seven in e). Deterministic fits repeat across seeds. Ranges describe same-study sensitivity, not inferential intervals; deeper targets are defined in Supplementary Methods.



Extended Data Figure 3 | Reference effects in genetic-perturbation prediction
CPA uses training-side selection (Supplementary Table 8). **a,** Standardized MSE for 55 held-out combinations under shared (S) and disjoint (D) references; zero effect is a diagnostic baseline. **b,** Standardized RMS displacement after changing CPA inputs. **c,** Reversals from S among 165 state-model–task pairs under separate prediction (P) or observation (O) references. Points and bars in a–c show means and full ranges over 30 allocations. **d,** Effect-minus-state margin divided by reference penalty, each averaged over tasks and allocations. Shading marks $-1 < d_S/V < 0$, the exact reversal interval; CPA uses a separate vertical scale. Eight capture partitions belong to one pooled screen.



Extended Data Figure 4 | Simulation performance depends on independent replication and test choice
a, Probability of any false directional selection across 40 global-null conditions (5,000 replicates each); orange marks have upper limits above 5%. **b,** Hierarchical-Gaussian-null rejection proportions (2,500 replicates per root count). Task/cell procedures treat descendants as independent; Holm correction covers 12 directions. **c,** Complete-family and true-edge recovery by adjacent-method separation $|\delta|$, pooling 10,000 signed-signal replicates per setting. Panels a/c use root-level exact-sign/Holm at familywise $\alpha = 0.05$; c starts at eight roots, the first-rejection minimum. Pointwise 95% Monte Carlo intervals use Clopper–Pearson, or Hoeffding for true-edge recovery. Sign and mean procedures test different hypotheses.



Extended Data Figure 5 | Recorded links alter candidate acquisition geometry
a, Candidate-component count versus largest-component size for 58 plates. Five rules link plates by producer/context, exact completion date, their union, or unions allowing dates within one or seven days. The dashed line marks the 11 non-tied roots required for a first Holm rejection among 90 ordered directions at familywise $\alpha = 0.05$. **b,** Component-size spectra, each summing to 58 plates. These exact graph counts do not establish independence.

Mean absolute score change

Table 1 Score sensitivity to reference-cell reassignment in CPA

Reassigned role	**Prediction or score update**	**64 cells/block ($d = 8$)**	**1,080 cells/block ($d = 135$)**
Model input	Select the prediction generated from reassigned input cells.	0.00854	0.00176
Prediction reference	Re-centre the same state prediction.	0.0235	0.00223
Observation reference	Recalculate the observed effect using the reassigned controls.	0.0234	0.00222

Scores are standardized MSE for input-conditioned CPA, averaged equally over 55 Norman tasks. Entries average the absolute change in this mean score when one reference role is reassigned and the other two remain fixed. Larger values indicate greater score sensitivity. Each block contains d untreated cells from each of eight capture partitions. The 30 allocations reuse cells from the same screen. Full allocation details, additional depths and fixed-prediction controls are given in Supplementary Methods and Source Data.

Supplementary Files

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

- SupplementaryInformation.pdf