

Supplementary Information: 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 Tarantino⁴,
Paolo Giudici⁴, Antonello Maruotti^{2,5,7}, Dean Everett^{2,5,6},
Gregory Joseph Fonseca^{2,3,6*}, Hao Zhou^{1,2,6*}

¹Department of Biological Sciences, College of Medicine and Health Sciences, Khalifa University, Abu Dhabi, United Arab Emirates.

²Center for Biotechnology (BTC), Khalifa University, Abu Dhabi, United Arab Emirates.

³Department of Medical Sciences, College of Medicine and Health Sciences, Khalifa University, Abu Dhabi, United Arab Emirates.

⁴Department of Economics and Management, University of Pavia, 27100, Pavia, Italy.

⁵Department of Public Health and Epidemiology, College of Medicine and Health Sciences, Khalifa University, Abu Dhabi, United Arab Emirates.

⁶Infection Research Unit, Khalifa University, Abu Dhabi, United Arab Emirates.

⁷Department GEPLI, LUMSA University.

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

Supplementary Methods

Influenza PBMC donor selection

For Influenza PBMC (GSE162632), we ranked donor and batch identifiers by SHA-256 using a prespecified salt. The first eight batches were assigned to evaluation and the other seven to training. Within each evaluation batch, three of the six donors had paired mock/influenza capture orientations matching the prespecified batch orientation. We ranked these candidates by the same rule.

In hash-ranked batch order B12, B7, B1, B2, B9, B13, B3 and B15, the eligible-candidate counts were 3, 3, 2, 3, 2, 3, 3 and 2. From the Cartesian product across batches, we retained the 520 complete combinations with four donors annotated as AFR and four as EUR in the source metadata. We represented each retained combination as an ordered tuple of donor hashes and selected the lexicographically smallest tuple, giving one donor per batch.

The salt and real donor–batch mapping are omitted from public materials. The confidential source-membership attachment contains the original metadata, donor correspondence and selected raw counts for reconstructing donor selection, preprocessing and reference means.

Model selection and training budgets

Supplementary Table 6 lists the grids, primary configurations and budget boundaries for Figs. 2–4 and Extended Data Figs. 1 and 2. Fig. 3 uses the three state-output families from these same fits. Methods defines the training groups and observation-separated validation. The expanded eight-family comparison

was specified before its new fits, with earlier external results available. All selected models were fitted before external evaluation.

The original fixed CellOT configuration used direct transport in the 2,000-gene expression space at its final checkpoint (Supplementary Table 7). The training-selected configurations instead used an autoencoder fitted within each fold, as described in Methods (Supplementary Table 6).

scGen decoder outputs were not truncated at zero. Its donor-balanced latent shift and the cellwise conversion of CPA count-space outputs are described in Methods.

Validation risk was still decreasing at the initial 1,280-epoch limit for both Influenza PBMC CPA settings, so we extended training to 2,560 epochs. Earlier test results were known when we planned this extension, but the amendment was fixed before the 2,560-epoch results and selection used only training-side validation. The learning rate of 0.0003, adversarial weight of 10, seven validation folds, physical cells, features, scales and optimization trajectories were unchanged. We compared the new checkpoint with the complete existing candidate grid using equal-donor standardized effect MSE, favouring the smaller budget for ties within 10^{-12} . Both groups selected the 2,560-epoch checkpoint.

We continued the full-training fits for seeds 17, 29 and 43 to this selected budget. The amended analysis ended at 2,560 epochs, with no higher budgets evaluated.

For scGen under the Influenza PBMC 48-cell cap, we extended both learning-rate trajectories to 640 epochs in all seven validation folds, yielding 12 candidates evaluated across 40 donors. At learning rate 0.0003, extending training from 320 to 640 epochs increased cross-validation MSE by approximately 2.4×10^{-6} . The 320-epoch checkpoint remained optimal, so the prespecified condition for extension to 1,280 epochs was not met.

The initial scGen fit used 3,200 cells from 40 training donors and the terminal 80-epoch checkpoint, without predictive checkpoint selection; loss curves were unavailable. We subsequently compared learning rates 0.001 and 0.0003 at 80, 160 and 320 epochs along continuous trajectories in seven batch-held-out folds, using only training data. Preprocessing used fold-training controls. Selection minimized equal-donor standardized treated-state prediction MSE across all 40 validation donors, choosing learning rate 0.001 at 80 epochs with an advantage of about 1.35×10^{-4} over the next candidate. Refits with seeds 17, 29 and 43 reproduced the corresponding initial weights and latent shifts. External reference comparisons used these three fits with the direct-effect, CPA and CellOT predictions fixed. The later eight-family comparison used the observation-separated validation endpoint described in Methods. This scGen training-sensitivity analysis followed inspection of the initial PBMC rankings.

Selection used unrounded scores. Source Data include all six scGen candidates, preprocessing and optimizer settings, training histories and external comparisons. The numerical records give full-precision scGen selection scores and Kang endpoint means.

Selection under alternative validation references

Configuration selection under shared references (S) and the two-block observation-separated design (O) used the same validation folds, cells, features and gene scales. For each input block, both scores used the same prediction: S defined the observed effect with that block and O with the other control block. We then swapped the blocks and scored the corresponding input-specific prediction. Scores were averaged over the two orientations within each task, tasks within each donor and validation donors equally.

Both criteria evaluated every candidate in the 320-candidate primary grid. A secondary pool added eight CPA budget extensions triggered under O, giving the same 328 candidates for both criteria. Ties within 10^{-12} favoured the smaller training budget, then the smaller representation rank and finally the lexicographically earlier candidate identifier. This sensitivity analysis was designed after the primary PBMC comparisons; configuration selection used only training donors.

For external evaluation, we crossed both selectors with S, O and D scoring in each of the four training groups. The selected candidates were evaluated with the same treated observations, genes, scales, control allocations and resource-specific task weights. We averaged scores over physical block assignments, 1,000 allocations and separately scored fitted seeds, then equally over the eight evaluation donors. The largest reference depth was primary, with the other existing depths used for sensitivity analyses. The comparison

included all 28 model-pair margins, including ties and unchanged orders. Selector effects compare S-selected and O-selected configurations under a common external scoring design; reference effects compare scoring designs within a selected model set. These descriptive comparisons use the same eight evaluation donors per resource (Supplementary Table 1).

Reference allocation and score identities

For each resource and allocation index $p \in \{0, \dots, 999\}$, we generated an allocation key by hashing the version tag, resource identifier and decimal allocation index with SHA-256. Each UTF-8 field was prefixed by its byte length as an eight-byte big-endian integer. Within each untreated-cell stratum, HMAC-SHA256 used this key to hash the stratum and cell identifiers with the same field encoding. Cells were ordered by digest, breaking ties by bitwise cell-identifier order. The original allocation analysis followed inspection of the initial PBMC rankings. We specified this revised rule after that analysis, before generating new predictions or scores.

Cells were assigned cyclically to streams B_1, B_2, B_3 . At depth n , the first n cells of each stream formed disjoint blocks, with lower depths nested within higher depths. Each label specified the complete allocation across all 88 Parse donor-cell-type strata or 40 Influenza PBMC root-task strata. Source Data include all 1,000 labels. The confidential Influenza PBMC source-membership attachment provides the information needed to reconstruct cell assignments and reference means.

For the matched-depth role patterns defined in Methods, let b_i be the three control-block means, M_i the saved state predictions generated using block i , and $r_i = M_i - Y$, where Y is the fixed treated mean. With training-derived scales and $G = 2000$, define $\langle a, b \rangle_s = G^{-1} \sum_g a_g b_g / s_g^2$ and its associated squared norm. Let

$$V = \frac{1}{6} \sum_{i \neq j} \|b_i - b_j\|_s^2, \quad K_m = \frac{1}{6} \sum_{i \neq j} \langle r_i, b_i - b_j \rangle_s. \quad (\text{S1})$$

Writing $d_{ij} = b_i - b_j$, the P and O residuals are $r_i + d_{ij}$ and $r_i - d_{ij}$. Their squared norms satisfy

$$\|r_i \pm d_{ij}\|_s^2 = \|r_i\|_s^2 + \|d_{ij}\|_s^2 \pm 2\langle r_i, d_{ij} \rangle_s. \quad (\text{S2})$$

The six ordered assignments weight every conditioning block twice, so the averaged first term equals the S squared loss. Negating the losses gives

$$U_P - U_S = -V - 2K_m, \quad U_O - U_S = -V + 2K_m. \quad (\text{S3})$$

For D, fix conditioning block i and pair the assignments of the remaining blocks j, k . The residuals $r_i + b_j - b_k$ and $r_i + b_k - b_j$ have mean squared loss $\|r_i\|_s^2 + \|b_j - b_k\|_s^2$. Averaging over input blocks gives $U_D - U_S = -V$ and $U_D = (U_P + U_O)/2$. For M, the residual $\{M(B_i) - B_j\} - (Y - B_j)$ reduces to $M(B_i) - Y$; every input block has the same marginal weight as in S, so $U_M = U_S$. Direct-effect outputs fixed across assignments have equal mean utility in all five patterns because the observation block has the same marginal weight. These identities require fixed treated means, feature weights and cached predictions, equal mapping weights and score-then-average aggregation. They apply to the matched-depth patterns, not the historical full-pool Shared construction described under ‘‘Diagnostic bounds’’.

For each configuration pair, displacement was the source-minus-target difference in utility changes from S to the comparison pattern. We averaged displacements over 1,000 labels and then eight roots equally for each resource, depth, pattern and canonically oriented pair. Adding this displacement to the S margin averaged in the same way gave the comparison-pattern margin.

For the fixed configurations in Supplementary Table 7, the maximum discrepancies in the M-S identity, D-P/O midpoint and common D-S shift across state models were 3.10×10^{-9} , 1.59×10^{-9} and 3.37×10^{-9} , respectively, all below 4×10^{-9} . The Influenza PBMC checks covered all 1,000 allocations, three depths, eight roots and eight models. For label sensitivity, we first averaged utilities over the eight roots within each label, then checked for strict opposite nonzero pairwise signs relative to S across all 28 pairs, three depths and four non-anchor patterns. Of these 336 comparisons, 45 crossed under every label and 291 never crossed. The original Parse utility tables had already been averaged over allocations, limiting this

label-level check to Influenza PBMC. Allocation-specific scores are available for all four training groups in the training-selected analyses.

The original fixed-configuration analysis highlighted five crossing pairs chosen after inspection; complete-family summaries included all 28 pairs. Source Data retain the fixed-total, depth and leave-one-composite sensitivities. The exploratory Parse RBF-scGen comparison was selected post hoc from separately constructed 84-comparison sensitivity-bound families at each depth.

Agreement between error criteria

Both error criteria were calculated within tasks before aggregation, using the same selected predictions, treated cells and control assignments. Delta-RMSE was $\{2000^{-1} \sum_g (P_g - O_g)^2\}^{1/2}$. We scored each fit separately, averaged tasks within roots and block assignments with their pattern weights, and then averaged the 1,000 labels, three stochastic seeds and eight roots equally. Deterministic configurations were fitted once. Error scores were negated for ranking. Exact score equality defined ties, and leader sets included every exact maximum. Kendall τ_b required finite scores for all eight families and a nonzero denominator; concordance was undefined for a constant complete ranking.

In Extended Data Fig. 1, the two criteria gave identical root-averaged eight-model rankings in all 80 designs. Within individual roots, full rankings agreed in 448 of 640 designs: 71 of 120 under Influenza PBMC 8-cell training, 58 of 120 under its 48-cell cap, 150 of 200 under Parse 48-cell training and 169 of 200 under its 128-cell cap. Leader sets agreed in 111, 93, 185 and 200 designs, respectively, giving 589 of 640 overall. Of the designs with different full rankings, 141 preserved the leader and 51 changed it. All concordances were defined. These descriptive comparisons reuse the same eight roots per resource; the repeated designs are not independent replicates.

For the original fixed-configuration estimability analysis, Pearson delta correlated predicted and observed effects across genes, and mean-effect cosine measured their uncentred cosine similarity in CP10K-log1p space. Pearson was undefined for a constant vector and cosine for a zero-norm vector. Centroid retrieval compared each predicted post-treatment centroid with observed targets in the same root and publisher cell-type stratum. Its score was the fraction of incorrect candidates farther away than the correct candidate, with zero credit for distance ties. Target scores were averaged within complete cell-type strata, then equally across strata within a root. Retrieval required at least two targets per complete stratum and was structurally inapplicable in Influenza PBMC, which had one. Among 160 metric-design combinations, all 40 delta-RMSE designs were estimable; 105 combinations had incomplete model families and 15 retrieval designs were structurally inapplicable. If a required score was missing, we withheld the complete-family rank summary; other defined scores are included in Source Data. This analysis followed inspection of the initial PBMC rankings.

Input sensitivity of the PBMC models

For Fig. 3, we used the primary scGen, CPA and CellOT fits from Fig. 2 under both Influenza PBMC training regimes (8 cells and a 48-cell cap) and both Parse regimes (48 cells and a 128-cell cap). Each group included all eight evaluation roots, seeds 17, 29 and 43, 1,000 allocation labels and all tested reference depths: 4/6/8 cells in Influenza PBMC and 16/24/32/40/48 in Parse. The main display uses the largest depth in each resource. We designed this unblinded analysis after inspecting the updated Fig. 2 rankings. All settings below, including comparison pairs and representation checks, were fixed before inspecting its results; the comparison set included non-reversals.

We defined $F(Z)$ as the mean native state prediction from input block Z , averaging its saved cell-level outputs with the original model adapter and gene coordinates. With observation reference A and prediction baseline B fixed, changing input C to B gives D-to-O, and changing C to A gives D-to-P. Both isolate input dependence at fixed scoring references. The primary displacement is $\|F(B) - F(C)\|_s$, using fixed training-derived gene scales and equal-gene weights. The secondary displacement, $\|(F(B) - B) - (F(C) - C)\|_s$, also changes the baseline subtracted from each prediction and is reported separately alongside $\|B - C\|_s$. Constant-state and additive-baseline controls distinguish these endpoints.

Within each root, we averaged squared standardized changes over tasks, all six block permutations, allocation labels and separately fitted seeds, then took one square root. Overall summaries equally

weighted the eight root RMS values. Complete permutation averaging makes D-to-O and D-to-P displacement magnitudes identical. Utility contrasts used negative-MSE scores, averaged after scoring each seed without forming a prediction ensemble. The analysis included all three state-model pairs and all nine cross-seed combinations. S-to-M provided a balanced mean-score control despite changes in individual input predictions; NC, CM, TW, PCA and RBF provided controls that do not use model-input cells. All 48 pair-by-depth aggregate comparisons kept their order under each input contrast. Source Data include every root, seed and allocation, exact signs, the 10^{-12} tie rule and numerical residuals. Roots were the biological units; seeds and allocations were used to assess sensitivity, without new inferential intervals.

For equivalent representations, we compared $(F(Z) - B, Y - A)$ with $(F(Z), Y - A + B)$ and applied common translations to $(F(Z), Y)$, keeping model input fixed. We also recalculated displacement directly from expression vectors. The 15,552 representation checks covered the first and last tasks of all 36 prediction sets, allocations 0 and 999, every depth and all 27 role assignments. Independent double-precision calculations gave maximum residual-vector and MSE differences of 8.88×10^{-16} and 2.22×10^{-16} . Benchmark states were centred in single precision, with residuals and utilities computed in double precision; the maximum discrepancy from double-precision centring was 1.14×10^{-8} . Supplementary Fig. 1 and Supplementary Table 2 give the Kang comparisons.

Kang conditioning analysis

The Kang interferon- β panel contained paired control and stimulated profiles from eight donors. We used the official EH2259 raw counts and matched barcodes and metadata exactly to KangCrossPatient. KangCrossPatient supplied annotations only. We completed gene selection and metadata reconciliation before fitting; the same 5,000 genes were selected across the tested software versions.

Each leave-one-donor-out fold excluded the evaluation donor from preprocessing and fitting. We transformed each cell to CP10K-log1p, ranked genes by variance across training rows, broke ties by source order and restored the selected genes to source order. We fitted a centred, unscaled 100-component principal-component basis to the same training rows and used training-control sample standard deviations, floored at 0.1, as gene scales. For each held-out donor and B, CD14 Mono, CD16 Mono or CD4 T task, expression-blind SHA-256 ordering selected 12 control and 12 stimulated cells. We assigned three disjoint four-cell control blocks to roles A, B, C in all six assignments, with contrasts scored before averaging.

The fixed ridge baseline had rank 4 and penalty $\lambda = 1$, with an unpenalized intercept and a centred input basis fitted to 28 training donor-task control means. Separate, matched fits predicted stimulated-minus-control means and stimulated means, giving 16 fits across eight folds. CellFlow optimal-transport flow matching used the PBMC architecture and optimizer from cellflow-tools v0.0.9. Five initializations per fold gave 40 fits, each covering four tasks. We trained for 500,000 microsteps, accumulating gradients over 20 microsteps per update, for 25,000 optimizer updates. Predictions used the terminal fits without predictive checkpoint selection. Training histories and available checkpoints are listed in Source Data. The selected four-cell block supplied both CellFlow source cells and their mean training-fold principal-component context.

Let $\hat{y}_m(Z)$ be the terminal-state mean conditioned on block Z . We converted effect-fit ridge predictions to states by adding $\bar{Z}$; state-fit ridge and CellFlow predicted states directly. The effect and state adapters were

$$p_m^E(Z) = \hat{y}_m(Z) - \bar{Z}, \quad p_m^S(Z, P) = \hat{y}_m(Z) - \bar{P}, \quad (\text{S4})$$

so $p_m^S = p_m^E + \bar{Z} - \bar{P}$. With $\delta_{\text{obs}} = \bar{Y}_{\text{stim},12} - \bar{C}_{\text{obs}}$, utility was the negative mean squared residual over 5,000 genes, standardized by the fold's training-control scales.

The original factor-isolation analysis changed prediction centring at fixed input (Stage A), then changed input at fixed scoring references (Stage B). The contrast orientations came from earlier analyses of this resource. The factor-isolation analysis preceded the matched-state comparison below. The factor-isolation endpoints, aggregation, diagnostic bounds and interpretation criteria were fixed before calculating their outcomes. The directions drew on earlier analyses; the factor-isolation analysis was unblinded.

Stage A held $C_{\text{obs}} = A$ and $C_{\text{model}} = C$ fixed while changing C_{pred} from A to B . Define $R_k^{ab} = U_{\text{ridge effect}}^a(A, P_k, C) - U_{\text{CellFlow}}^b(A, P_k, C)$ for output adapters $a, b \in \{E, S\}$, where E subtracts the input-block mean and S subtracts the designated prediction-reference mean, with $P_0 = A$, $P_1 = B$, $M_k = R_k^{ES} - \{R_k^{EE} + R_k^{SS}\}/2$, and $H_k = U_{\text{ridge effect}}^E(A, P_k, C) - U_{\text{ridge state}}^S(A, P_k, C)$.

E1 was the donor-level mean of $M_1 - M_0$ over training realizations, tasks and block mappings; E2 was the donor-level mean of $H_1 - H_0$ over tasks and mappings. E1 compared cross-output performance with two same-output controls, whereas E2 compared ridge fits trained on effect and state targets. We numerically verified $M_k = \{R_k^{ES} - R_k^{SE}\}/2$.

Stage B held $C_{\text{obs}} = A$ and $C_{\text{pred}} = B$ fixed. With $Q(Z) = U_{\text{CellFlow}}^S(A, B, Z)$, E3 was the donor-level mean of $Q(A) - Q(C)$, and E4 was the mean of $Q(B) - Q(C)$, both over training realizations, tasks and mappings. E3 and E4 measured changes in CellFlow utility between conditioning allocations; they did not compare ridge with CellFlow.

The contrast directions $(E1, E2, E3, E4) = (+, +, -, +)$ were chosen from prior resource-specific analysis before this experiment produced its outputs. We first calculated cell means, then gene-averaged utilities and within-donor contrasts, and finally averaged the six block mappings and four tasks equally. E1, E3 and E4 were also averaged over five initializations; E2 was independent of initialization. Every endpoint required values from all eight donors; we did not reduce the donor set when a value was unavailable.

Define the fold-specific inner product $\langle a, b \rangle_W = 5000^{-1} \sum_g a_g b_g / s_g^2$. Pairing each mapping (A, B, C) with (B, A, C) holds the physical conditioning block C fixed. Expanding the Stage A contrasts within each pair cancels all prediction-dependent and treated-mean cross terms. Equal averaging gives

$$E1_d = E2_d = \text{mean}_{\text{task, mapping}} \|\bar{A} - \bar{B}\|_W^2. \quad (\text{S5})$$

E1 and E2 therefore measure the same squared reference distance and cannot serve as independent checks of model validity. We reconstructed both endpoints from the fixed block and stimulated means, fold-specific scales, ridge parameters and CellFlow output means, without refitting. All 32 donor-task values satisfied the identity to floating-point precision, with a common overall mean of 0.491. The reconstruction recovered the donor-level endpoints, conditioning margins and their midpoint relation. Anonymous derived outputs and contrasts before aggregation are provided in Source Data.

For endpoint j and donor d , we multiplied the contrast by the specified direction to obtain X_{jd} . We analysed all four endpoints jointly with all $2^8 = 256$ donor sign vectors, applying the same vector across endpoints. For $h_{jd} = (X_{jd} - \bar{X}_j)/8$, each sign-vector statistic was recentred and re-studentized, and we retained the maximum absolute statistic across endpoints. The 244th ordered maximum supplied the 0.95-reference-quantile critical value.

Equal mapping weights also give $E3 = -E4$ for the raw utility changes. The decision criteria combined the endpoint indicators, each of which required an oriented lower bound above zero. These bounds are working-model diagnostics: E1/E2 are identical, E3/E4 are opposite before orientation, and the overlapping leave-one-donor-out training sets do not establish independent donor contributions.

The matched-state ridge-CellFlow comparison fixed $C_{\text{obs}} = A$ and $C_{\text{pred}} = B$, while both models received input $Z \in \{A, B, C\}$ and used $p_m^S(Z, B)$. Ridge received the block mean; CellFlow received its source cells and their mean principal-component context, so source and context changed together. Effect scoring used residual $M(Z) - Y + A - B$; a separate treated-state calculation used $M(Z) - Y$. We reused the fixed fits and CellFlow predictions and generated ridge predictions from the fixed coefficients, without selecting checkpoints, initializations or tasks or refitting models. We specified the complete matched-state analysis after inspecting E3/E4 utilities and before calculating prediction displacement or ranking margins.

For each input change, C to A or C to B, we divided prediction changes by training-fold gene scales, squared them and averaged over genes, six mappings, four tasks and the five CellFlow initializations within each donor. We then took one square root per donor and averaged the eight donor RMS values. Utility margins were formed within donor-task-initialization-mapping cells and averaged over mappings, tasks and initializations. The donor margin $m_d(Z) = U_{\text{ridge, state, } d}(Z) - U_{\text{CellFlow, } d}(Z)$ favoured ridge when positive. Across-donor means gave all eight roots equal weight.

The matched-state family comprised $\{\overline{m}(C), \overline{m}(A), \overline{m}(B), \Delta_A, \Delta_B\}$, with $\Delta_A = \overline{m}(A) - \overline{m}(C)$ and $\Delta_B = \overline{m}(B) - \overline{m}(C)$. Equal weighting of all six mappings gives $\overline{m}(C) = \{\overline{m}(A) + \overline{m}(B)\}/2$, hence $\Delta_A = -\Delta_B$. These five quantities received two-sided working-model bounds from all $2^8 = 256$ shared donor-sign vectors, recentred and re-studentized max- $|T|$, and the 244th ordered maximum. A descriptive crossing required nonzero aggregate margins of opposite sign. The original diagnostic criterion additionally required the two margin bounds to exclude zero in opposite directions and the shift bound to exclude zero. These bounds were uncalibrated. The B margin was -0.00101 , close to zero (Supplementary Table 2).

Kang training-selection sensitivity

We compared training-selected and fixed configurations in the same eight donors. Each outer fold used seven inner-validation splits: one training donor supplied validation profiles and the other six supplied fitting data. Six-donor preprocessing used the same count transformation, 5,000-gene selection, control scales and 100-component principal-component construction. We reused identical six-donor fits across outer folds by indexing the 28 unordered pairs of excluded donors. Selection remained restricted to each outer fold's seven validation donors.

CellFlow used two learning rates, 5×10^{-5} and 10^{-4} , one inner-training seed (17), and predictive checkpoints at 125,000, 250,000 and 500,000 microsteps, with 20 microsteps per optimizer update. We compared the resulting 56 trajectories by standardized effect MSE, averaged over genes, six ordered assignments of distinct observation and prediction/input blocks, four cell types and seven validation donors. Donors had equal weight, and we selected the configuration with the lowest MSE. A learning-rate trajectory was eligible for 750,000- and 1,000,000-microstep checkpoints only when its own outer-fold risk was strictly lowest at 500,000 microsteps and improved by at least 1% relative to 250,000; no extensions were triggered. Exact risk ties favoured the shorter budget and then the lower learning rate.

Ridge used ranks 2, 4 and 8 and penalties $\lambda \in \{0.1, 1, 10\}$, with the same validation splits and endpoint. State-target validation selected the configuration; its matched effect-target fit shared the input basis and penalty. Exact risk ties favoured lower rank and then lower penalty.

All eight folds selected rank 4 and $\lambda = 10$. CellFlow selected 5×10^{-5} and 500,000 microsteps for donor 101, 10^{-4} and 250,000 for donor 1016, and 10^{-4} and 500,000 for the other six donors. Seven selections reached the tested training-budget boundary, and eight reached the ridge penalty boundary. These boundary selections do not establish training convergence.

Each selected CellFlow configuration was refitted on all seven outer-training donors with seeds 17, 29, 43, 763929762 and 85508424, yielding 40 fits at the selected budgets. Matched ridge refitting yielded 16 fits. Refitting and prediction preceded evaluation of the outer comparisons.

We fixed observation reference A and prediction reference B and supplied each of A, B and C as model input. We retained all 960 donor-task-initialization-assignment combinations and the original E1–E4 endpoint definitions. Displacement and utility summaries used the same within-donor aggregation as the fixed-configuration analysis; donor RMS values were averaged equally. Tasks, assignments and initializations were repeated measurements of the same eight donors.

For the selected fits, E1 and E2 retained the same reference-distance mean of 0.491. E3 and E4 were -0.0201 and 0.0201 . Supplementary Table 2 gives matched-state margins, utility changes and displacements for all input choices. Donor-level endpoints, utilities, selection risks and comparisons before aggregation are provided in Source Data.

Program analyses after model selection

We evaluated the two Influenza PBMC training regimes separately, using the fixed family order NC, CM, TW, PCA, RBF, scGen, CPA and CellOT. For each of the eight evaluation donors, we selected models using gene-level standardized-MSE utilities from the other seven donors, at eight controls per role across all 1,000 allocation labels. We averaged already-scored seeds, labels and selection donors equally, maximized utility, and resolved ties within 10^{-12} by the fixed family order. Selection for the 39 additional donors used all eight evaluation donors once per regime. S-selected and O-selected denote the reference pattern used for selection. We specified the full design of the updated program analysis after

the earlier analyses and updated gene-level rankings, before calculating its new program outcomes. It reused the expression data and cell memberships from the measured-response follow-up.

We recomputed each S, O and D choice from the complete candidate utility arrays. Across the two regimes, all 16 leave-one-donor-out choices and both all-eight-donor choices agreed between O and D, with no ties; all differed from S. We matched fit identities, input blocks, seeds and centring before comparing downstream outputs. Gene predictions agreed elementwise between O and D for the original and additional donors, and reconstructed program predictions exactly reproduced the stored results. Source Data include all candidate utilities, tie sets, margins, selected-fit hashes and reconstruction checks. This comparison was retrospective and reused the existing predictions and outcomes.

The ordered 2,000-gene panel, training scales and Hallmark mapping were identical across evaluation panels. We retained, in mapping order, all 37 programs with at least 15 represented genes and at least 10% of official membership. Program scores were the mean signed, training-standardized effects over represented genes. These expression summaries are not enrichment P values or measures of pathway activation.

The evaluation panel contained eight donors and five cell types (40 tasks). The additional-donor union contained 190 supported tasks. The primary cohort included B cells and combined monocytes in 39 donors (78 tasks), with complete extensions of four cell types in 39 donors (156 tasks) and five in 34 donors (170 tasks). Cell membership, acquisition metadata, treated means and the allocation algorithm stayed fixed.

Labels 0, 100, ..., 900 each assigned untreated cells to three disjoint eight-cell blocks. The primary input/prediction-reference block was B1 at label 0. The same predictions were evaluated against B1- and B2-centred responses, with B2 primary. We repeated the analysis across all ten labels, three input blocks and three observation blocks, keeping prediction centring matched to model input.

Single-cell states were averaged in recorded cell order, accumulated in float64 and stored in float32. Control means used float32 cell values with float64 accumulation and float32 storage. Each state mean was centred once by its prediction-reference mean in float32. Treated means and program calculations used float64; direct-effect predictions retained their native float32 values.

Squared and absolute program errors were calculated separately for seeds 17, 29 and 43 before equal averaging. For deterministic models, repeated seed coordinates referred to the same fit. The primary contrast was O-selected minus S-selected program MSE; both model errors were retained.

For the eight evaluation donors, aggregation averaged programs and cell types within donor, then donors equally. Each complete additional-donor cohort averaged programs and cell types within donor, donors within batch-by-capture-orientation strata, two orientations within batch and eight batches equally.

Direction and magnitude displays used mean predicted scores across seeds. Opposite directions required opposite signs with both absolute scores strictly exceeding 0, 0.01, 0.025 or 0.05; top-1/3/5/10 sets ranked absolute scores with fixed program order for ties. We calculated these summaries separately from the mean of seed-specific errors.

Fig. 4a counted donors out of eight meeting the 0.025 criterion at primary B1/label 0 for every program and cell type, without averaging tasks first. Cell types were ordered as B, CD4 T, CD8 T, combined monocytes (Mono.) and combined natural killer cells (NK). Both matrices use 0-8; white cells mean zero, with every combination represented. At the 0.025 threshold, the two selections predicted opposite directions in 75 of 1,480 task-program pairs under 8-cell training and 65 under the 48-cell cap. These spanned 20 and 23 programs and 36 and 32 of the 40 tasks, respectively. The top-five program sets differed in 33 and 34 tasks.

Supplementary Fig. 2 shows means over all 40 tasks, averaged across seeds with seed minimum-maximum ranges and the common B2 observed response. Applying the 0.025 criterion to these 37 overall mean pairs gave 0/37 disagreements in each regime.

Extended Data Fig. 2 and Supplementary Tables 3 and 4 report analyses across seeds, references, thresholds, top sets, acquisitions, weighting schemes and cohorts. The 37-program panel was primary; excluding interferon- α left 36. For deeper observations, we kept input fixed and used all eligible treated cells and untreated cells outside the input block. Because allocations reuse cells, programs overlap and donors

share acquisition batches, the reported ranges are descriptive; we did not calculate calibrated confidence intervals or significance tests.

Before scoring the additional donors, we checked the prediction procedure against stored predictions for the eight evaluation donors, using absolute tolerance 10^{-6} and relative tolerance 10^{-5} . We verified model identity, gene and task order, cell membership, block separation and finite predictions. Independent calculations checked the squared-error observation-shift identity, additive program contributions and residual invariance under a common baseline translation of stored effects. States were centred only once.

The evaluation used previously inspected expression from the same study. For paired squared errors, the squared target-noise term cancels. The remaining cross term has expectation zero when target noise has zero mean conditional on the prediction difference. Separating observation controls does not by itself remove batch effects or establish biological truth.

Measured-response follow-up

The measured-response follow-up tested a candidate from an earlier exploratory screen of 370 cell-type-by-program contrasts using the fixed model configurations in Supplementary Table 7. It was the Hallmark interferon- γ score in B cells minus combined monocytes. Before extracting expression, we fixed 39 unused donors, the represented genes, cell membership, 30 observation-reference allocations, aggregation and the decision rule. Estimates had to remain positive across allocations, batch omissions, capture orientations, alternative weighting and full-cell analysis. The follow-up failed this criterion. Neither newly selected model was run in this analysis. Supplementary Table 5 reports the original discovery predictions and the full measured-response result.

External test of reference-dependent model selection

The GSE181897 analysis tested whether observation-separated selection improved prediction on a common assessment target. We developed the protocol and adapters using earlier Parse results and recorded them locally before extracting expression or scoring this analysis. Donor-context blocks were eligible if their metadata listed at least 16 control cells and eight treated cells for each of three conditions (interferon- β , interferon- γ and TNF- α). The fixed split assigned 16 donors to selection and 46 to assessment, giving 777 primary assessment tasks. Selection and assessment donors were disjoint, but all 12 acquisition pools crossed the split. The comparison is therefore conditional on the observed pools and selection split.

We transferred seven fixed model families (NC, CM, TW, PCA, RBF, scGen and CellOT) without refitting. The adapter scored 1,968 unambiguous genes using the original training scales and filled 32 unavailable model inputs with their training-control means. CPA was excluded before scoring because its required full input axis was unavailable. For primary selection, we either shared all 16 controls across roles or split them equally: eight for input and prediction centring and eight for observation centring. Assessment used a common eight-cell input block and a separately measured effect target. We scored each fit separately, averaged losses over fits, the three conditions and eligible contexts within donor, then averaged donors equally.

Both selectors chose PCA, producing identical predictions and effect MSE of 0.0494655356. Two- and four-fold cross-fitting and the other recorded reference designs also selected PCA at total budgets of 8, 16 and 32 controls (24 selections). The primary relative improvement was 0%, below the prespecified 5% target. Source Data and the software evidence capsule provide all model losses, exclusions, cell-budget variants, treated-state checks and paired-donor calculations. This comparison assessed the transferred models against a finite-cell measured effect, not a latent biological effect.

Genetic-perturbation prediction

The Norman extension used the GEO-filtered RNA count matrix and guide annotations from GSE133344. We retained high-coverage single-cell assignments and the producer-retained non-targeting constructs, excluding NegCtrl1_NegCtrl0. The producer alias RHOXF2→RHOXF2B was retained alongside the original annotation. A two-gene task required at least 100 cells overall, five in each of

eight capture partitions and 50 cells for each constituent single-gene condition. These rules retained 110 of 131 two-gene conditions. A fixed hash assigned 55 to training and 55 to testing. All 105 single-gene conditions entered training. Hash-splitting controls within capture partition yielded 3,636 training and 3,639 evaluation controls.

Counts were normalized to 10,000 per cell across the full feature axis and log-transformed before selecting 2,000 unique Ensembl identifiers by training-control variance. Fixed scales were training-control population standard deviations, floored at 0.1. We gave the eight capture partitions equal weight when calculating targets and training-condition means. The additive state prediction was the sum of the two single-gene means minus the training-control mean. Compositional ridge regressed training-condition means on constituent-gene indicators, with equal condition weights, an unpenalized intercept and a unit penalty on the summed squared coefficients.

The 200-epoch CPA fit used a Gaussian likelihood, a 128-dimensional latent space, three 256-unit encoder and decoder layers, batch size 512 and seed 17. We used the terminal checkpoint without validation-based selection and excluded all test combinations from fitting. The split and settings were fixed before prediction, after earlier development on this dataset.

A CPA training-sensitivity analysis held out 11 of the 55 training two-gene combinations and used the remaining 60,704 cells for fitting. The Gaussian architecture, genes and scales remained fixed. Four settings crossed the native or per-epoch learning-rate schedule with adversarial weights of 1 or 10; checkpoints at 100, 200 and 400 epochs gave 12 candidates. Each validation task used non-overlapping input and observation control blocks, with 64 cells per capture in each block. Selection minimized mean standardized effect MSE across the 11 combinations; a separate 44-condition cell holdout was not used for selection.

We selected per-epoch scheduling with weight 10 at 400 epochs and refitted this configuration once to all 68,002 training cells. This refit supplied all CPA predictions in Extended Data Fig. 3; the 200-epoch checkpoint and its results were retained. Supplementary Table 8 reports this finite grid and the comparison on the same 55 held-out two-gene combinations. We had already inspected earlier results on these test combinations when planning this analysis.

The direct-effect ridge used the same constituent-gene indicators and predicted each condition’s change from the training-control mean, with capture partitions weighted equally. Targets were standardized by the fixed gene scales. We fitted without an intercept, minimizing summed squared residuals over equally weighted conditions and genes plus α times the squared coefficient norm.

We hash-ranked the 55 training combinations and assigned successive ranks cyclically to five folds; all 105 singles and controls remained in every training fold. We evaluated $\alpha \in \{0, 0.01, 0.1, 1, 10, 100\}$ using mean squared error over all 55 out-of-fold standardized effects. Values within 10^{-12} of the minimum favoured the larger α . Validation selected $\alpha = 0.1$ (MSE, 0.0184). We refitted all 161 training conditions and restored the original gene scales, yielding one fixed effect vector per test combination. We planned this extension after inspecting the state-model results and fixed its model, grid and selection rule before validation fitting or evaluation of effect predictions.

For each of 30 allocations, evaluation controls were hash-ranked within capture partition and assigned to three blocks of 135 cells. Nested prefixes gave per-partition depths of 8, 16, 32, 64, 128 and 135. All models faced the same treated target and reference allocation. We scored all 27 assignments before averaging within S, M, P, O and D, giving each block equal marginal use in every role. For all three direct-effect–state pairs, we reported effect-minus-state utility margins d_S, d_D and the reference penalty V , checking $d_D = d_S + V$. We averaged tasks and allocations equally and evaluated strict crossings both before and after aggregation. Margins within 10^{-12} of zero were treated as ties. We also checked state-output margin invariance and consistent state/effect conversion. Source Data include all models, tasks, depths and allocations, including stable comparisons.

With the training-selected CPA, the task-averaged state-output order remained compositional ridge, additive and CPA across all allocations, depths and reference patterns. CPA’s mean standardized RMS prediction displacement declined from 0.056 at 64 cells per role to 0.013 at 1,080. After averaging allocations, three of 165 state-model–task pairs reversed under P and two under O at 64 cells per role; one O comparison still reversed at 1,080. S-to-D margins between state-output predictors remained invariant. The direct-effect–state comparison still showed four of 18 overall pair-by-depth reversals, although some task-level classifications changed (Source Data).

Diagnostic bounds

The original fixed-configuration Influenza PBMC analysis partitioned 24 eligible mock cells in each evaluation root-task stratum into three depth-8 blocks. Shared assigned the 24-cell union to every active reference role; split assigned one block to each role. Rotation scored all role rotations separately and then averaged utilities. The constructions differed in overlap, per-role depth or nonlinear aggregation, so their comparisons combine these changes. The original family comprised 28 unordered configuration pairs in each of three constructions, giving 84 contrasts. The original Influenza PBMC roots, tasks, configurations, reference designs, utility and 84-contrast family were specified before inspection, after checks of data identity, cell support, training separation, reproducibility and family completeness. PCA-64 ridge versus scGen and the 56 construction-shift contrasts were selected after inspecting the 84-contrast family.

For configuration pair j , construction s and root r , d_{rjs} was the utility difference averaged equally over five tasks and $\bar{d}_{js} = 8^{-1} \sum_r d_{rjs}$. For $s \in \{\text{split}, \text{rotation}\}$, paired changes were $\Delta_{rj, \text{shared} \rightarrow s} = d_{rjs} - d_{rj, \text{shared}}$, with means over the same eight roots. These quantities describe margins among the observed roots and changes between whole constructions. The 56-member change family contained all 28 shared-to-split and 28 shared-to-rotation changes, including ten design-defined zeros.

The repeated-root working model assumes independent contributions from a common finite-variance distribution, with Rademacher sign multipliers adequately representing centred contributions. If D_{rjs} is the hypothetical random counterpart of d_{rjs} , its targets are $\theta_{js} = E(D_{rjs})$ and $\psi_{j, \text{shared} \rightarrow s} = E(D_{rjs} - D_{rj, \text{shared}})$. The within-construction and paired-change families used separate reference distributions. Bounds took the form $\bar{X}_k \pm cs_k/\sqrt{R}$, with sample-standard-deviation denominator $R - 1$.

For a family of contrasts across R roots, let X_{rk} be the contrast for root r and family member k , $\bar{X}_k = R^{-1} \sum_r X_{rk}$, $a_{rk} = (X_{rk} - \bar{X}_k)/R$, and $q_k(e) = \sum_r e_r a_{rk}$ for sign vector e . We calculated

$$T_k(e) = \frac{q_k(e)}{\left\{ \frac{R}{R-1} \left[\sum_r a_{rk}^2 - \frac{q_k(e)^2}{R} \right] \right\}^{1/2}}. \quad (\text{S6})$$

The $\lceil 0.95 \times 2^R \rceil$ th ordered value of $\max_k |T_k(e)|$ over all 2^R sign vectors supplied the critical value c .

For the ten direct-effect pairs, rotation minus Shared is exactly zero: the difference of squared losses is affine in the observation reference, and the equal-depth rotation average equals the Shared union mean. We retained these pairs in the family and set their root contributions, standard errors, bounds and multiplier statistics to zero.

For any other family member, an invalid standard error for the observed contrast or any sign vector made the entire family's bounds unbounded. Calibration simulations used a separate zero-standard-error convention: a zero numerator gave a statistic of zero, and a nonzero numerator gave positive infinity. All 25,000 calibration critical values were finite.

The 84-member within-construction family used the 244th of the $2^8 = 256$ ordered maxima and gave $c \approx 4.58$. The 56-member paired-change family used the same centring, sign schedule and re-studentization, retaining the ten design-defined zeros; its critical value was $c \approx 4.73$, and 33 nondegenerate bounds excluded zero.

For comparison, we calculated two-sided Bonferroni- t intervals for the 84-member family using $t_{1-0.05/(2 \times 84), 7} \approx 5.91$; the selected within-construction signs were unchanged. Critical values are rounded here; the full-precision values are in the numerical records.

We reconstructed pseudonymized root utilities from fixed full-sample and leave-one-root summaries as $u_r = 8U_{\text{full}} - 7U_{-r}$. Leave-one-root winners described individual-root influence.

To assess sensitivity to training realization, we used five fixed-setting CPA, scGen and CellOT fits, including the primary fit. The four additional fits were made after inspection of the primary results. Direct-effect predictions stayed fixed. Utilities were scored separately and averaged over tasks and realizations within each of eight donor-batch roots. The 25 cross-realization combinations for each stochastic pair described fitting variation within those roots. We applied the same bound calculation separately to the 84-member within-construction and 56-member paired-change families.

Calibration of the 56-member procedure used 5,000 replicates per scenario. Two of five scenarios failed the prespecified criteria (Supplementary Table 9). Under symmetric Gaussian errors, simultaneous coverage was 94.4%, with a 95% Wilson lower bound of 93.7%, below the required 94%. Under centred skew, coverage was 33.6% and the wrong-direction rate was 34.4%; its 95% Wilson upper bound of 35.7% exceeded the 5% criterion. A separate 198-scenario evaluation of the shared max- $|T|$ implementation also failed two prespecified coverage criteria. The 0.95 quantile refers to the sign-reference distribution. Enumerating every sign vector removes Monte Carlo error from that calculation, but the failed calibration precludes interpreting these working-model sensitivity bounds as calibrated 95% confidence intervals for these donors or a wider population.

Simulation design

For the root-utility model in Methods, the initial grid crossed $\delta = 0, \pm 0.25, \pm 0.5, \pm 1$, $R = 8, 12, 20, 40$ and ten nuisance states. These comprised the baseline, three common-root tie probabilities, three complete-root missingness probabilities and three heterogeneous-support cycles. A tie replaced the complete four-method root vector by its common mean; missingness removed the entire vector independently of utilities and ties. Support was balanced at 8 descendants or cycled through 6/10, 4/12 or 1/15. Each state varied one nuisance factor; interactions were not simulated. Complete-family recovery required all six true directions and no false direction; true-edge recovery was the fraction of the six true directions recovered. For simulations and original PBMC runs, analysis cells and endpoints were specified before their outputs, together with missing-data, aggregation and retention rules. The simulations used specified generating laws. Denominators included every specified cell, including zero, unavailable and no-direction outcomes.

For the hierarchical Gaussian null in Methods, $R = 2, 3, 4, 6, 8, 12, 20, 40$ roots each had eight tasks and 20 nominal cells per task. We retained the first (5, 7, 9, 11, 13, 15, 17, 19) cells in tasks 1–8, then averaged cells within tasks, tasks equally within roots and roots equally. The cell-level negative control used pseudo-rows $Z_{rtcm} = N_{\text{obs}} Y_{rtcm} / (8Rn_t)$, where n_t was the retained task-specific count and $N_{\text{obs}} = 96R$. This preserved the equal-root, equal-task point mean. Each root-count condition had 2,500 outer replicates; we recorded nonfinite calculations as failures and did not replace them.

For direction $a > b$, the exact-sign test was conditional on the observed availability pattern. It used root contrasts $D_r = U_{ra} - U_{rb}$, omitted exact zeros and calculated

$$p = 2^{-K} \sum_{j=W}^K \binom{K}{j}, \quad W = \#(D_r > 0), \quad K = \#(D_r \neq 0), \quad (\text{S7})$$

with $p = 1$ when $K = 0$. The design-averaged sign probability was used descriptively. The root CR1- t comparator used $t = \bar{D} / (s_D / \sqrt{R})$, sample-variance denominator $R - 1$, $R - 1$ degrees of freedom and a one-sided Student- t tail. Zero standard error gave $t = 0$ and $p = 1$.

Root, task and cell bootstraps resampled R , $8R$ or $96R$ paired rows, respectively. Within each procedure and outer replicate, we centred each model column and sampled one $499 \times N$ index matrix with replacement, applying it to all four models and 12 directions. For $G_m = R^{-1} \sum_r U_{rm}$ and observed contrast $D_{\text{obs}} = G_a - G_b$, the unstudentized one-sided bootstrap value was

$$p = \frac{1 + \sum_{\ell=1}^{499} I(\hat{D}_{\ell}^0 \geq D_{\text{obs}})}{500}. \quad (\text{S8})$$

Each procedure used Holm adjustment over all 12 directions at $\alpha = 0.05$. The prespecified hypothesis order resolved tied p -values, and equality at the threshold counted as rejection. Task- and cell-level resampling treated dependent descendants as independent units. Whole-root resampling preserved the unit but gave a 6.4–95.9% probability of at least one null-direction rejection across this grid (Extended Data Fig. 4b).

We used four separate PCG64 streams for data generation and the root-, task- and cell-level bootstraps. Each procedure shared its index matrix across comparisons and used a separate matrix from the other procedures. Seed inputs were the version tag, condition identifier, outer-replicate tag, six-digit zero-padded replicate index and stream purpose. These fields were separated by null bytes and hashed with SHA-256; the digest was read as an unsigned big-endian integer. After reproducing the 20 original event

counts at $R = 8, 12, 20, 40$, we extended the simulations to $R = 2, 3, 4, 6$, fixing the law, procedures, replicate counts and streams before the extension. All 10,000 added replicates were included, and Student- t tails at the added degrees of freedom agreed with SciPy at the Holm thresholds.

Each root-utility setting- δ combination in Extended Data Fig. 4a,c had 5,000 replicates. Positive and negative δ cells were generated independently and pooled only for specified aggregate endpoints by summing numerators and denominators. The smallest sign-test P value, 2^{-K} , exceeds $0.05/12$ when $K \leq 7$, so recovery is zero for $R < 8$ and Extended Data Fig. 4c displays $R \geq 8$. The minimum root counts needed for a first rejection with other family sizes are shown in Supplementary Fig. 4. Binary endpoints used pointwise two-sided 95% Clopper-Pearson intervals with $N = 2, 500, 5,000$ or $10,000$; bounded true-edge recovery used a two-sided Hoeffding interval. These quantify Monte Carlo error under the specified law.

A separate 255-setting simulation assessed simultaneous directional-error bounds with 25,000 successful replicates per setting and tail allocation $0.05/255$ (Supplementary Fig. 5). Comparator simulations covered balanced and strong-intraclass-correlation settings with 2,500 replicates per condition (Supplementary Fig. 6).

The stress analysis at $|\delta| = 0.5$ evaluated four endpoints across 40 stress-state-root-count designs, giving 160 summaries (Supplementary Fig. 7). Each design pooled 5,000 replicates at each sign of δ ; the four endpoints shared these 10,000 replicates. Structural withholding meant fewer than eight complete, non-tied roots were available; decision abstention meant at least eight were available but Holm rejected no direction. We summed these events to obtain total no-direction and calculated its interval directly from that count. Unless stated otherwise, simulation intervals were pointwise Monte Carlo intervals; candidate-component counts and matching extrema were exact graph quantities.

Acquisition and matching

We screened 30 benchmark records post hoc using fixed criteria for independent roots, complete configuration families, within-root controls and identifiable inferential procedures. Separately, we screened 17 H5AD files for perturbation, control, context and acquisition fields (Supplementary Fig. 3). Both screens were descriptive and used non-probability samples.

For GSE306429, we constructed acquisition graphs from public metadata before reading expression or the model-comparison results used here. The 58 physical plates formed the vertices. Five undirected graphs linked plates by producer/context, exact completion date, their union, or that union with completion dates within one or seven days. Connected components were candidate dependence units. The three-construction, six-configuration application contained 90 ordered directions, requiring at least $K_{\min} = \lceil \log_2(90/0.05) \rceil = 11$ non-tied roots for a first Holm rejection at familywise $\alpha = 0.05$. We could not resolve processed-cell lineage, per-unit control/perturbation support or the justification for independence; aggregate results were therefore descriptive (Source Data).

The Jerber day-52 rotenone panel contained 20 evaluated donor-linked lines in ten realized differentiation pools. Connecting lines through shared pools gave nine candidate line-pool components. Its eight-configuration family contained 56 ordered directions, requiring $K_{\min} = \lceil \log_2(56/0.05) \rceil = 11$. Even if all nine components were independent, they would fall below this threshold. Shared-pool connectivity alone did not establish independence, so we reported fixed utilities without directional intervals. RBF ridge ranked first under all three reference constructions. Acquisition-unit support for Jerber, as for GSE306429, was assessed from public records before directional testing.

For the Sound Life wave utilities defined in Methods, eligible day-0 B and plasma cells were deterministically ranked within donor-year and distributed cyclically into three disjoint blocks. Blocks contained 172–1,047 cells across waves and differed by at most one cell within a wave. The rotations assigned $(C_{\text{obs}}, C_{\text{pred}}, C_{\text{model}})$ to $(1,2,3)$, $(2,3,1)$ and $(3,1,2)$. Memberships, reference depths, features, scales and predictions stayed fixed across matchings. This matching analysis was post hoc, with donor-batch constraints held fixed.

The unrestricted graph had 93 edges, 58 donors and 40 batches, with maximum cardinality 39. Excluding batches represented among the 37 training batches gave 23 edges, 22 donors and 12 batches, with maximum cardinality 12. The two selected matchings shared 11 waves but were not nested.

Independent checks used anonymized donor, batch and wave identifiers, 93 fixed binary64 utility vectors and ten witness cases. They verified maximum matching sizes, donor–batch nonreuse, all 56 decisions per case, 112 directional win bounds across the two graphs and all 144 matchings of size 12.

Reference-role audits in Norman

We evaluated six Norman prediction configurations on 55 tasks using 30 control allocations and six depths (8, 16, 32, 64, 128 and 135 untreated cells per capture per block). The eight capture partitions gave 64–1,080 cells per block. We scored all 27 assignments of three blocks to the three reference roles before aggregation, holding fitted parameters fixed. Input-conditioned CPA used the saved prediction from the assigned input block. The other five configurations used fixed predictions: CPA, additive and compositional-ridge states, and no-change and ridge effects. Fixed CPA predictions came from the 200-epoch fit; block-specific predictions came from the validation-selected 400-epoch refit.

For each model, depth and metric, let $\bar{L}_a(\mathbf{r})$ be the loss or score in its metric direction, averaged equally over the 55 tasks at allocation a and role tuple $\mathbf{r}$. A single-role contrast was $\delta_a = \bar{L}_a(\mathbf{r}') - \bar{L}_a(\mathbf{r})$, with the other roles fixed. Table 1 reports the mean of $|\delta_a|$ over all 30 allocations and 54 directed contrasts per allocation. A two-role interaction was $I_a = \bar{L}_a(\mathbf{r}_{11}) - \bar{L}_a(\mathbf{r}_{10}) - \bar{L}_a(\mathbf{r}_{01}) + \bar{L}_a(\mathbf{r}_{00})$, holding the third role fixed; its mean absolute value used 108 directed contrasts per allocation. Thus averaging tasks preceded taking absolute values.

For input-conditioned CPA, the mean absolute observation–input score interaction decreased from 0.00442 at depth 8 to 0.000251 at depth 135. This difference-in-differences held the prediction reference fixed. Positive signed contrasts indicate higher metric values: worse for MSE/RMSE and better for Pearson/centroid accuracy. The contrasts describe one pooled screen, whose cells were reused across allocations.

We retained all models, depths, four metrics and three role pairs in Source Data, including zero input contrasts and interactions for fixed-output configurations. Repeated-primary scoring and the full reference-role audit gave identical primary scores across all 288 combinations of model, depth, metric and scoring view; the latter also measured the specified role changes. The supplied software reconstructs these summaries from the score cube and checks reference operations using small prediction arrays. Reconstruction of the complete cube requires the listed upstream prediction arrays.

Primary effect evaluation used two blocks: one for model input and prediction centring, and one for observation controls. D used three blocks as a sensitivity diagnostic.

For the first matched Norman comparison, we applied independent implementations of Systema metric formulas to six prediction configurations, 55 tasks and 2,000 genes. Expression states were restored using each configuration’s recorded baseline. We used the existing predictions without running the official package or reproducing its model fits. Replacing the training-control Pearson reference with the mean of 160 training-perturbation centroids changed one of 15 pair directions: the additive-minus-native-effect-ridge margin shifted from -0.00293 to $+0.00400$.

Repeating selected CPA inputs across 30 allocations, three blocks and six depths gave 540 input choices. All 15 task-mean pair directions remained stable under perturbation-centred Pearson and centroid accuracy despite score variation. Centroid accuracy was the fraction of 54 incorrect targets strictly farther from the prediction than the correct target, with zero credit for ties.

Fixed-reference Systema state scores had balanced role-pattern means that agreed to floating-point precision: observation and prediction centring were absent, and model-input marginals matched. This invariance is separate from the effect-target S-to-D mechanism.

Stability calibration and comparison reporting

We calibrated pairwise reporting rules on development cases and applied them to separate assessment cases. Calibration required finite scores for every model, budget and realization, finite nonnegative control geometry, and a prespecified anchor r_0 for each case. Development and full assessment tables contained at least two realizations per case, with the same realizations represented at every budget within a case. We

fixed decisions from development scores and assessment anchors before examining the remaining assessment realizations. For each comparison, we recorded the decision, reporting threshold and assessment support, together with counts of reported and held comparisons.

For model pair p , budget b , development case c and realization r , let $m_{cpb}(r)$ be the first model's utility minus the second's, with lower-is-better scores negated. With control geometry g_{cb} and geometry tolerance $\epsilon > 0$, we calculated

$$\Delta_{cpb} = \max_r |m_{cpb}(r) - m_{cpb}(r_0)|, \quad B_{pb} = \max_c \Delta_{cpb}, \quad K_{pb} = \max_{c: g_{cb} > \epsilon} \frac{\Delta_{cpb}}{g_{cb}}. \quad (\text{S9})$$

For assessment case a , the empirical range is

$$R_{apb} = \begin{cases} 0, & \text{single split,} \\ B_{pb}, & \text{repeated primary,} \\ K_{pb}g_{ab}, & \text{geometry scaled.} \end{cases} \quad (\text{S10})$$

Geometry scaling was unavailable when no development geometry exceeded ϵ , or any case had $g_{cb} \leq \epsilon$ but $\Delta_{cpb} > \tau$, where $\tau > 0$ is the numerical tolerance. Assessment geometry at or below ϵ also withheld a geometry-scaled decision. Otherwise, the anchor direction was reported only when

$$|m_{apb}(r_0)| > R_{apb} + \delta + \tau, \quad (\text{S11})$$

where $\delta \geq 0$ is the prespecified reporting gap. Numerical ties were withheld.

Comparisons that failed the reporting rule were held. A reported direction was supported only if its oriented margin exceeded $\delta + \tau$ in every recorded assessment realization. Otherwise, it was unsupported because of a reversal or a margin at or below the threshold. These empirical ranges describe observed variation, not confidence bounds.

For the secondary control-budget analysis, we selected the smallest tested count at which every comparison involving the unique anchor-best model was reported in its favour. Equal budgets followed their listed order; cases with no qualifying budget were held. Support required the assessment check to pass against every opponent. The rule concerns the anchor-leading model at the selected budget; a larger budget can have a different leader.

The Norman analysis held the primary target, standardized effect MSE and available control pool fixed. The 27 development tasks and 28 assessment tasks came from the same screen and used the same 30 evaluated control allocations, each evaluated for three input-block choices. Assessment tasks were held out from calibration. At 192 available controls (64 input and 128 observation controls) and reporting gap 0.01, single-split scoring reported 323 of 420 pair cases, with 65 unsupported; repeated-primary scoring reported 165, with none unsupported. The denominator comprised 15 model pairs across 28 assessment tasks. Of the 65 unsupported directions, 24 reversed in at least one assessment realization, and all 65 had at least one positive margin below the threshold; these categories overlap. Geometry scaling matched repeated-primary decisions in Norman. This retrospective analysis measured stability and reporting coverage within the supplied control pool; repeated allocations were not independent biological replicates.

At the primary gap 0.01, repeated-primary and geometry-scaled reporting each supported budget recommendations for two of 28 Norman assessment tasks, selecting 192 and 1,536 controls and holding 26. Neither rule recommended a budget in PBMC or Kang. Norman had unsupported reported directions at gaps 0, 0.005 and 0.02, and one unsupported minimum-budget recommendation at each of gaps 0 and 0.005. Geometry scaling showed no gain over repeated primary across the three resources and four prespecified gaps.

VCC scoring adapter example

For the Adamson CRISPRi example, we used scPerturb data (Zenodo record 13350497) with all 32,738 raw-count genes. We selected ATF6, ERN1 and EIF2AK3 and supplied three fixed predictions: two control-based baselines using different input pools and a truth-informed prediction. These predictions were used to check scoring behaviour, not to evaluate trained competition submissions.

Nine reference allocations produced 27 model–allocation outcomes: 15 were scored and 12 had undefined calibration. The undefined cases arose from four shared-reference allocations in which the official reach baseline attained its ceiling. We retained all 189 metric results and 189 model-pair comparisons, including unavailable values.

We calculated the six VCC metrics and their calibrated aggregate using cell-eval2 (v0.16.0). Official VCC2026 submission scoring shares held-out measured controls between prediction and observation references, with model-input controls supplied separately. Separating the scoring references defines an alternative protocol.

For calibrated metrics and their aggregate, we matched baseline and replicate estimates to the scoring-reference data and configuration. Input-only changes at fixed scoring references used the same calibration. We reported uncalibrated VCC metrics separately and restricted VCC comparison reporting to calibrated scores. The official replicate estimator uses each split half’s controls, whereas submission scoring uses a common reference.

Numerical verification

An independent implementation checked reference operations on stored Norman, PBMC and Kang predictions. The maximum absolute difference was 9.99×10^{-16} ; Source Data identify the operations and numerical results.

Independent calculations reproduced the Kang E1/E2 identity across 32 donor–task values within 3.33×10^{-16} , donor endpoints within 1.11×10^{-16} , and conditioning margins and their midpoint relation within 2.78×10^{-17} . For fixed-reference Systema state scores, balanced role-pattern means differed by at most 3.33×10^{-16} .

For the first task and allocation at depth 8, we checked 14 storage encodings (six Norman, eight PBMC) and seven common-reference translations of state outputs (four Norman, three PBMC). Storage conversion preserved decoded predictions and targets; common translation preserved residuals and scores with model input fixed. Separate checks covered the complete Norman metric panel and the independent calculation underlying Fig. 3.

For each reference operation, we specified the target, output representation and control roles. Changing the inputs of an input-conditioned model required supplied predictions generated from the corresponding controls. Input-independent models retained their predictions. Scoring-reference changes reconstructed effects or targets from saved predictions, and equivalent coordinate changes transformed predictions and targets together. Resampling observation controls from the same population changed the finite-sample target and its coupling while preserving the population estimand. A change in target type or control population required a new target definition.

For a synthetic two-block check, we supplied a direct effect e and an equivalent conditioned state $B_1 + e$. Against target $Y - B_2$, both gave residuals $e - Y + B_2$ after conversion. Independent calculation included all three models, three tasks and both allocation labels, averaging tasks within each of two units before averaging units. The equivalent-effect pair had MSE 2.5 in both pipelines; the third, fixed-state prediction had MSE 3.0.

Independent calculation reproduced VCC scores for the original and first separated allocations within 1.11×10^{-16} , including four structural nulls. It also reproduced the undefined calibration for the first shared allocation. Source Data contain these comparisons and the separate check across all allocations.

Source Data and the reconstruction guide provide the required inputs and starting points for these calculations. Source Data retain every reporting rule and gap, held and unavailable comparisons, and the earlier scGen and signed-role analyses. Supplementary Tables 2, 3 and 4 display effects, error contrasts and displacements to three significant figures; simulation probabilities and Wilson interval endpoints in Supplementary Table 9 use four decimal places.

Supplementary Table 1 | Configuration selection and comparison under alternative validation references

a, Configuration choices.

Training group	Different choices / 8	Selected under S	Selected under O
Influenza PBMC, 8 cells	0 / 8	—	—
Influenza PBMC, 48-cell cap	0 / 8	—	—
Parse, 48 cells	2 / 8	scGen: 80 epochs; CellOT: 50,000 steps	scGen: 160 epochs; CellOT: 25,000 steps
Parse, 128-cell cap	0 / 8	—	—

b, External S-to-D comparisons.

Candidate pool	Training group	Validation	Reversed pairs	Unchanged pairs	Tied pairs
Original grid	Influenza PBMC, 8 cells	S	10	18	0
Original grid	Influenza PBMC, 8 cells	O	10	18	0
Original grid	Influenza PBMC, 48-cell cap	S	10	18	0
Original grid	Influenza PBMC, 48-cell cap	O	10	18	0
Original grid	Parse, 48 cells	S	10	18	0
Original grid	Parse, 48 cells	O	10	18	0
Original grid	Parse, 128-cell cap	S	10	18	0
Original grid	Parse, 128-cell cap	O	10	18	0
Completed pool	Influenza PBMC, 8 cells	S	10	18	0
Completed pool	Influenza PBMC, 8 cells	O	10	18	0
Completed pool	Influenza PBMC, 48-cell cap	S	10	18	0
Completed pool	Influenza PBMC, 48-cell cap	O	10	18	0
Completed pool	Parse, 48 cells	S	10	18	0
Completed pool	Parse, 48 cells	O	10	18	0
Completed pool	Parse, 128-cell cap	S	10	18	0
Completed pool	Parse, 128-cell cap	O	10	18	0

Panel a lists configuration choices that differ between shared-reference (S) and observation-separated (O) validation; both candidate pools gave the same changes. The listed scGen settings use learning rate 0.0003; CellOT uses its 50-dimensional autoencoder representation. Panel b counts all 28 model-pair orders when external scoring changes from S to fully disjoint references (D), at eight reference cells per role in Influenza PBMC and 48 in Parse. A reversal has opposite nonzero mean-utility margins; a tied pair has a zero margin under either design. The original grid contains 320 candidate-group entries; the completed pool adds eight CPA budget extensions. Complete S/O/D comparisons at all depths, including reversal identities and changes between selectors, are in Source Data.

Supplementary Table 2 | Conditioning effects in the ridge–CellFlow comparison

The observation reference was block A and the prediction-centring reference was block B. Both models received the indicated conditioning block. Entries are equal-weight means across eight donor roots.

a, Utility margins and their changes. Positive ridge-minus-CellFlow margins favour ridge.

Conditioning choice or change	Mean margin or change	Diagnostic lower bound	Diagnostic upper bound
Condition on C	0.0183	0.0127	0.0239
Condition on A	0.0376	0.0261	0.049
Condition on B	−0.00101	−0.0045	0.00248
Change from C to A	0.0193	0.0129	0.0257
Change from C to B	−0.0193	−0.0257	−0.0129

b, Responses from C to B. Mean root mean square (RMS) displacement averages donor-specific RMS prediction changes in training-standardized gene units (Methods).

Model	Mean utility change	Mean RMS displacement
State-target ridge	0.00186	0.0639
CellFlow	0.0211	0.208

Bounds are working-model sensitivity diagnostics, without calibrated confidence-interval coverage (Methods; Supplementary Table 9).

c, Validation-selected configurations: utility margins and their changes. Observation A and prediction reference B remain fixed. Entries are equal-donor descriptive means for the eight donors used in panels a and b; positive margins favour state-target ridge.

Conditioning choice or change	Mean margin or change
Condition on C	0.0164
Condition on A	0.0348
Condition on B	−0.00187
Change from C to A	0.0183
Change from C to B	−0.0183

d, Validation-selected configurations: prediction responses. Displacement is the equal-donor mean of donor RMS prediction changes in training-standardized gene units. Summaries include all four cell types, five CellFlow initializations and six block assignments. Balanced assignments give equal mean displacements and opposite utility changes for C to A and C to B.

Model	Input change	Mean utility change	Mean RMS displacement
State-target ridge	C to A	−0.00183	0.0629
State-target ridge	C to B	0.00183	0.0629
CellFlow	C to A	−0.0201	0.198
CellFlow	C to B	0.0201	0.198

Panels a and b use the configurations specified in Methods and report working-model diagnostic bounds. Panels c and d use configurations selected by training-side validation on the same eight donors and report descriptive summaries without inferential intervals.

Supplementary Table 3 | Program comparisons in the eight evaluation donors

These analyses use 37 programs and five cell types per donor (40 tasks). S and O identify the reference design used for family selection. Each donor is evaluated using families selected on the other seven donors. Training cells are counted per donor–cell-type–condition stratum; ≤ 48 denotes the training cap. The program definition and selection rules are in Supplementary Methods, “Program analyses after model selection”.

a, Selection across all eight held-out-donor folds. Gaps are the best-minus-second-best mean utilities, with ranges across folds. The small O gaps show close alternatives; complete utilities, tie sets and all fold decisions are in Source Data.

Training cells	S-selected folds	O-selected folds	S gap range	O gap range
8 cells	scGen (8/8)	TW (7/8); PCA (1/8)	0.00495–0.00633	2.7×10^{-5} – 3.76×10^{-4}
≤ 48 cells	scGen (8/8)	PCA (5/8); RBF (3/8)	0.00706–0.00744	7×10^{-6} – 1.12×10^{-4}

b, Primary mean program errors. Label 0 and input/prediction-reference B1 remain fixed; B1 and B2 in the observation column give shared and separate observation controls. Errors are scored separately for each fitted seed, then averaged equally across seeds, programs, cell types and donors. Negative O–S favours O-selected predictions.

Training cells	Observation	S MSE	O MSE	O–S
8 cells	B1	0.0108	0.013	0.00224
8 cells	B2	0.0176	0.0136	-0.00399
≤ 48 cells	B1	0.00914	0.0126	0.0035
≤ 48 cells	B2	0.0154	0.0131	-0.00235

c, Direction and priority sensitivity. Directions and top-k sets use mean predicted scores across seeds. Opposite directions require both absolute scores to exceed the threshold. Top-k sets rank absolute scores, resolving ties by fixed program order. The primary prediction uses label 0/input B1; ranges cover all ten labels and three input blocks. All tested thresholds and k values are shown.

Training cells	Threshold	Opposite items /1,480	Affected tasks / 40	Task range /40
8 cells	0	301	40	39–40
8 cells	0.01	172	40	38–40
8 cells	0.025	75	36	30–38
8 cells	0.05	22	16	10–22
≤ 48 cells	0	315	39	37–40
≤ 48 cells	0.01	166	39	36–40
≤ 48 cells	0.025	65	32	27–34
≤ 48 cells	0.05	17	13	7–19

Training cells	k	Different tasks /40	Task range /40
8 cells	1	0	0–1
8 cells	3	13	11–19
8 cells	5	33	27–37
8 cells	10	40	39–40
≤ 48 cells	1	0	0–0
≤ 48 cells	3	10	8–16

Training cells	k	Different tasks /40	Task range /40
≤ 48 cells	5	34	26–37
≤ 48 cells	10	40	38–40

d, Observation-reference sensitivity. Ranges cover every shared or separated observation assignment while input and prediction-reference blocks coincide. Contrasts average separately scored seed errors, then donors equally.

Training regime	Observation relation	Assignments	Minimum O–S	Maximum O–S	Positive	Negative
8 cells	Shared	30	0.00136	0.00351	30	0
8 cells	Separated	60	-0.00605	-0.0036	0	60
≤ 48 cells	Shared	30	0.00326	0.00476	30	0
≤ 48 cells	Separated	60	-0.0041	-0.00196	0	60

Assignments reuse the same donors and cells; ranges are descriptive. Source Data, “Complete program and training records”, retain every donor error, individual-seed result and selection decision, including the all-eight-donor selection used for Supplementary Table 4. Figure 4 shows the complete primary donor comparisons; Extended Data Fig. 2 shows the main sensitivities.

Supplementary Table 4 | Program evaluation in 39 additional donors

This within-study evaluation uses 39 additional donors, eight acquisition batches and 78 B-cell or combined-monocyte tasks. Model families were selected once using the eight evaluation donors before prediction in the additional cohort: S selected scGen in both training regimes; O selected TW under 8-cell training and PCA under the 48-cell cap. Additional donors did not enter fitting or selection. $\Delta\text{MSE} = \text{MSE}_\text{O} - \text{MSE}_\text{S}$; negative values favour O-selected predictions.

a, Primary program errors. Label 0, input/prediction-reference B1 and treated cells are fixed; B1/B2 denote the observation controls. Errors are scored separately for each seed and averaged over seeds, the 37 programs and cell types within each donor. We then average donors within each batch –capture-orientation stratum, the two orientations within each batch and the eight batches, assigning equal weight at each step.

Training regime	Observation	S MSE	O MSE	ΔMSE
8 cells	B1	0.0159	0.0202	0.00423
8 cells	B2	0.0238	0.0205	−0.00335
≤48 cells	B1	0.0153	0.0185	0.00323
≤48 cells	B2	0.0224	0.0189	−0.00351

b, Reference sensitivity. Rows summarize all indicated assignments, with scores averaged over separately scored seeds. Ten labels and three input/prediction-reference blocks give 30 shared, 60 separate and 30 deeper-observation settings. The same primary direction occurs for all three individually scored seeds (Source Data).

Training regime	Observation	Settings	Minimum ΔMSE	Maximum ΔMSE	Positive	Negative
8 cells	Shared	30	0.00336	0.00541	30	0
8 cells	Separate	60	−0.00449	−0.00103	0	60
8 cells	Deeper	30	−0.00499	−0.00386	0	30
≤48 cells	Shared	30	0.0022	0.00413	30	0
≤48 cells	Separate	60	−0.00452	−0.00258	0	60
≤48 cells	Deeper	30	−0.00423	−0.00328	0	30

c, Acquisition sensitivity. NI capture identifies which capture supplied untreated cells. Orientation-specific estimates average donors within orientation and batch, then batches equally. Individual-batch estimates average the two orientations; omissions weight the remaining seven batches equally. Ranges span all batches or batch omissions.

Training regime	Summary	Estimates	ΔMSE , B1	ΔMSE , B2
8 cells	NI capture 1	1	0.00369	−0.00382
8 cells	NI capture 2	1	0.00478	−0.00288
8 cells	Individual batches	8	−0.000873 to 0.01	−0.00679 to 0.00263
8 cells	Leave one batch out	8	0.00341 to 0.00497	−0.0042 to −0.00286
≤48 cells	NI capture 1	1	0.00309	−0.00407
≤48 cells	NI capture 2	1	0.00337	−0.00294
≤48 cells	Individual batches	8	0.00127 to 0.00705	−0.00525 to −0.00127
≤48 cells	Leave one batch out	8	0.00268 to 0.00351	−0.00383 to −0.00326

d, Cohort, weighting and target sensitivity. Label 0/input B1 are fixed. Equal-donor and equal-batch weighting omit the primary capture-orientation standardization. Cohort and program variants retain the primary hierarchy. The deeper target uses all treated cells and all untreated cells outside input B1.

Training regime	Specification	Donors	Tasks	Programs	Δ MSE, B1	Δ MSE, B2/deeper
8 cells	Four cell types	39	156	37	0.00335	−0.00351
8 cells	Five cell types	34	170	37	0.00422	−0.00288
8 cells	Equal donors	39	78	37	0.00306	−0.00467
8 cells	Equal batches	39	78	37	0.00289	−0.00482
8 cells	Without interferon- α	39	78	36	0.00432	−0.00339
8 cells	Deeper target	39	78	37	—	−0.00436
≤ 48 cells	Four cell types	39	156	37	0.00302	−0.00339
≤ 48 cells	Five cell types	34	170	37	0.00343	−0.00325
≤ 48 cells	Equal donors	39	78	37	0.00325	−0.00351
≤ 48 cells	Equal batches	39	78	37	0.00319	−0.00356
≤ 48 cells	Without interferon- α	39	78	36	0.00319	−0.00369
≤ 48 cells	Deeper target	39	78	37	—	−0.00369

e, Prediction interpretation. Direction counts use mean predicted scores across seeds, primary input B1/label 0, opposite signs and both absolute scores above 0.025. Top-five sets rank absolute scores. The last two columns count programs with negative mean error contrasts.

Training regime	Opposite task—program pairs	Affected tasks	Affected donors	Different top-five sets	O lower, B1	O lower, B2
8 cells	140/2,886	64/78	37/39	66/78	2/37	35/37
≤ 48 cells	77/2,886	45/78	30/39	68/78	4/37	35/37

Figure 4c displays all 39 donor comparisons. Exact donor errors, seed results and program contributions are in Source Data, “Complete program and training records”. These donors share study batches; reference settings reuse cells and programs overlap. The ranges describe sensitivity, without confidence intervals or significance tests.

Supplementary Table 5 | Within-study validation of the exploratory interferon- γ contrast

This table reports measured responses for the interferon- γ contrast nominated by the earlier fixed-configuration discovery screen.

The fixed contrast was the Hallmark interferon- γ response in B cells minus combined monocytes. It was selected after screening 370 contrasts in eight discovery donors and tested in 39 unused donors from the same study and eight batches. Estimates are signed 59-gene scores in training-standard-deviation units; positive values match the discovery direction. Allocation and batch-omission ranges are dependent sensitivity summaries, not confidence intervals.

Analysis	Estimate or result
Discovery observed contrast	0.0454
Discovery S-selected contrast	0.0892
Discovery D-selected contrast	-0.0338
Primary validation estimate, batch/capture-orientation weighting	0.0538
Primary validation estimate, equal donor	0.0242
Full-cohort range over 30 allocations	0.0397 to 0.0758
Range over full cohort and eight batch omissions	0.0198 to 0.0957
Primary allocation, capture orientation 1	0.0950
Primary allocation, capture orientation 2	0.0125
Capture orientation 1, range over 30 allocations	0.0662 to 0.1235
Capture orientation 2, range over 30 allocations	-0.0201 to 0.0671
Full-cell estimate, capture orientation 1	0.0288
Full-cell estimate, capture orientation 2	-0.0059
Fixed-primary-allocation donor directions	22 positive; 17 negative
Descriptive stability criterion	Not met

This validation assessed the measured-response direction; neither selected model shown in Fig. 4 and Extended Data Fig. 2 was run as part of this test.

Supplementary Table 6 | Training-selected configurations for the PBMC analyses

Figs. 2–4 and Extended Data Figs. 1 and 2 use these settings. Training panels contain 8 or at most 48 cells per donor–cell-type–condition stratum in Influenza PBMC, and 48 or at most 128 in Parse, including controls. Models within a group use the same physical cells. Configuration selection holds out seven training batches in Influenza PBMC (40 donors) or four training donors in Parse and minimizes equal-donor standardized effect MSE with independent observation controls. Reference-cell depth at evaluation is specified separately.

a, Candidate grids.

Family	Initial grid
NC / CM	No hyperparameter grid; no change / context mean
TW	$\alpha \in \{0.01, 0.1, 1, 10, 100\}$
PCA	The same α grid; maximum rank $\in \{8, 32, 64, 128\}$
RBF	The same α grid; $\gamma \in \{0.1, 0.5, 2\}$
scGen	Learning rate $\in \{0.001, 0.0003\}$; epochs $\in \{20, 40, 80, 160, 320\}$
CPA	Learning rate $\in \{0.0001, 0.0003\}$; adversarial weight $\in \{1, 10\}$; epochs $\in \{24, 80, 160, 320\}$
CellOT	Direct genes / AE50; outer steps $\in \{1,000, 5,000, 10,000, 25,000, 50,000, 100,000\}$

A terminal optimum with at least 1% improvement over the preceding same-setting checkpoint initially triggered budget doubling across all folds, up to 1,280 epochs or 400,000 CellOT outer steps. Ties within 10^{-12} favoured smaller budgets, then smaller representation rank, then lexicographic candidate identifier. One structurally distinct setting within 1% of the minimum was retained when available, giving 15 additional configurations. Complete selection records are in Source Data, “Complete program and training records”. Stochastic configurations used refit seeds 17, 29 and 43; deterministic configurations were fitted once.

CPA validation was subsequently extended to 2,560 epochs in both Influenza PBMC groups (Supplementary Methods).

b, Primary settings. α is ridge regularization, γ the RBF coefficient and lr the learning rate. AE50 denotes the native 50-dimensional autoencoder representation.

Family	Influenza PBMC: 8	Influenza PBMC: ≤48	Parse: 48	Parse: ≤128
NC	No fit	No fit	No fit	No fit
CM	Context mean	Context mean	Context mean	Context mean
TW	$\alpha=10$	$\alpha=1$	$\alpha=100$	$\alpha=1$
PCA	$\alpha=1$, rank=8	$\alpha=1$, rank=8	$\alpha=100$, rank=128	$\alpha=1$, rank=128
RBF	$\alpha=1$, $\gamma=0.1$	$\alpha=1$, $\gamma=0.5$	$\alpha=10$, $\gamma=2$	$\alpha=1$, $\gamma=2$
CPA	lr=0.0003, adversarial weight=10; 2,560 epochs	lr=0.0003, adversarial weight=10; 2,560 epochs	lr=0.0003, adversarial weight=10; 1,280 epochs	lr=0.0003, adversarial weight=10; 320 epochs
scGen	lr=0.0003; 160 epochs	lr=0.0003; 320 epochs	lr=0.0003; 160 epochs	lr=0.001; 40 epochs
CellOT	AE50; 100,000 outer steps	AE50; 100,000 outer steps	AE50; 25,000 outer steps	AE50; 100,000 outer steps

c, Budget diagnostics. Terminal gain is the relative validation-risk reduction from the preceding same-setting checkpoint.

Training group	Family	Selection position	Terminal gain
Influenza PBMC: 8	scGen	Interior optimum	Not applicable
Influenza PBMC: 8	CPA	Budget limit; still improving	2.63%
Influenza PBMC: 8	CellOT	Terminal gain below 1%	0.26%
Influenza PBMC: ≤ 48	scGen	Terminal gain below 1%	0.74%
Influenza PBMC: ≤ 48	CPA	Budget limit; still improving	1.07%
Influenza PBMC: ≤ 48	CellOT	Terminal gain below 1%	0.068%
Parse: 48	scGen	Interior optimum	Not applicable
Parse: 48	CPA	Terminal gain below 1%	0.905%
Parse: 48	CellOT	Interior optimum	Not applicable
Parse: ≤ 128	scGen	Interior optimum	Not applicable
Parse: ≤ 128	CPA	Terminal gain below 1%	0.706%
Parse: ≤ 128	CellOT	Terminal gain below 1%	0.053%

CPA validation error was still decreasing at the budget limit in both Influenza PBMC groups, so convergence was not established. Extending scGen to 640 epochs under the 48-cell cap left the 320-epoch checkpoint best among those evaluated.

Supplementary Table 7 | Model definitions and training settings

Sound Life uses the configurations below, with dataset-specific adaptations listed after the tables. These configurations also underlie the fixed-configuration PBMC comparisons; Supplementary Table 6 gives the validation-selected PBMC settings used in Figs. 2–4 and Extended Data Figs. 1 and 2. Source Data contain the scGen checkpoint comparisons and full performance records for these configurations.

a, Direct-effect configurations.

Model	Model definition	Fitting settings
No change	Zero effect for every gene.	Deterministic.
Context mean	Mean training-donor effect.	Deterministic.
Two-way additive ridge	Additive effects of one-hot context and perturbation indicators.	Regularization strength 1.
PCA-64 ridge	Rank-64 principal components of training effects, followed by additive ridge regression.	Regularization strength 1.
RBF kernel ridge	Radial basis function kernel on fixed one-hot task indicators.	Kernel coefficient 0.5; regularization strength 1.

All three ridge models used closed-form solutions in 64-bit floating-point arithmetic. The RBF coefficient scales squared distance in the negative exponent.

b, Absolute-state configurations.

Model	Model definition	Training settings
CPA	Nonvariational latent state of 128 dimensions; three 256-unit encoder/decoder layers. Negative-binomial reconstruction on the full count axis.	CPA 0.8.8; batch size 1,024; logarithmic-sigmoid dose response; 24 epochs; terminal checkpoint without validation-based selection.
scGen	Global variational autoencoder on 2,000 genes, with 512 hidden units, 100 latent dimensions, two layers and dropout 0.2. Latent shifts are averaged equally across donors.	scGen 2.1.1; batch size 512; 80 epochs.
CellOT	Dual input-convex transport networks on 2,000-gene CP10K-log1p expression, with four 64-unit hidden layers. Separate maps for 90 Parse cytokines or one influenza-A-virus condition in Influenza PBMC.	Batch size 64; learning rate 10^{-4} ; 100,000 outer updates, each with 10 inner updates; terminal state without validation-based selection.

Sound Life settings. Direct-effect fits weighted donors equally. Context means were estimated within influenza year; additive ridge used vaccine and influenza-year indicators. PCA retained at most 64 components, limited by the rank of the training data. CPA and CellOT used deterministically selected, donor-balanced panels of 15,360 cells. CPA used the full raw-count gene set, and scGen used transformed expression for the selected 2,000 genes.

Supplementary Table 8 | Norman CPA training sensitivity

a, Training-side selection. Mean standardized effect MSE across 11 held-out training combinations. The lowest MSE selected the setting and checkpoint. All 12 candidates are shown.

Learning-rate schedule	Adversarial weight	100 epochs	200 epochs	400 epochs
Native	1	0.0812	0.0824	0.0879
Native	10	0.0766	0.0729	0.0874
Per epoch	1	0.0814	0.0817	0.0802
Per epoch	10	0.0755	0.0724	0.0717

Native scheduling retained CPA 0.8.8 stepping; per-epoch scheduling advanced the same scheduler once each epoch. The selected checkpoint was at the grid boundary. This comparison assesses the specified training procedures; it does not establish convergence or isolate the effect of epoch count.

b, Common-target evaluation. The 55 held-out two-gene combinations were evaluated with the same treated targets, genes and scales. B1 supplied CPA input and centred state predictions; non-overlapping B2 cells supplied observation centring. No change predicted zero effect. Each block contained 64 controls per capture. Mean squared error (MSE) and mean absolute error (MAE) use standardized gene-effect errors, averaged over genes within each task and then equally across tasks. Amplitude ratio is the task-mean ratio of standardized predicted to measured effect norms.

Model	MSE	MAE	Amplitude ratio
Validation-selected CPA	0.0766	0.2018	0.408
CPA, 200 epochs	0.0952	0.2244	0.239
Additive	0.0296	0.1287	1.139
Compositional ridge	0.0286	0.1236	0.812
No change	0.1034	0.2315	0.000

The validation-selected CPA had lower MSE than the 200-epoch fit in 50 of 55 tasks, but lower MSE than additive and compositional ridge in only 5 and 4 tasks, respectively. The comparison uses previously evaluated test combinations from one pooled screen; captures and tasks are descriptive units. Source Data include both CPA configurations and every task-level result.

Supplementary Table 9 | Calibration of the diagnostic bounds

Each scenario comprised 5,000 replicates with eight experimental units and 56 contrasts. True contrasts followed the repeating sequence 0, +0.35 and −0.35. The scenarios and assessment criteria were chosen after empirical results were known, but before the calibration results were examined.

Simulation scenario	Simultaneous coverage (lower 95% bound)	Probability of any wrong direction (upper 95% bound)	Both criteria met
Symmetric Gaussian errors	0.9436 (0.9369)	0.0216 (0.0260)	No
Heavy-tailed symmetric errors	0.9568 (0.9508)	0.0154 (0.0192)	Yes
Centred skewed errors	0.3356 (0.3226)	0.3442 (0.3575)	No
Unequal error scales across units	0.9690 (0.9638)	0.0110 (0.0143)	Yes
Errors correlated across contrasts	0.9620 (0.9563)	0.0214 (0.0258)	Yes

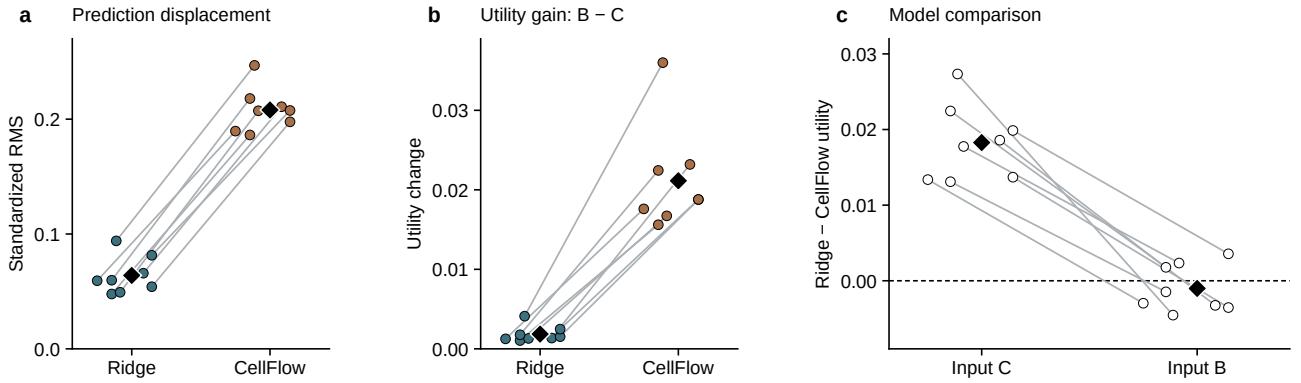
Parentheses give the indicated lower or upper endpoints of 95% Wilson intervals for the simulation probabilities. A scenario passed if the lower endpoint for simultaneous coverage was at least 0.94 and the upper endpoint for any wrong direction was at most 0.05. A wrong direction occurred when a bound lay entirely below zero for a nonnegative truth, or entirely above zero for a nonpositive truth.

The five error models included independent standard normal draws, independent Student's t draws with three degrees of freedom, and centred log-normal draws obtained by exponentiating 0.8 times a standard normal draw and subtracting the population mean. We also considered independent normal draws with unit-specific scales of 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2 and 2.25, and Gaussian errors with a shared factor of loading 0.9 whose sign alternated across contrasts. The bound construction is defined in Supplementary Methods, "Diagnostic bounds".

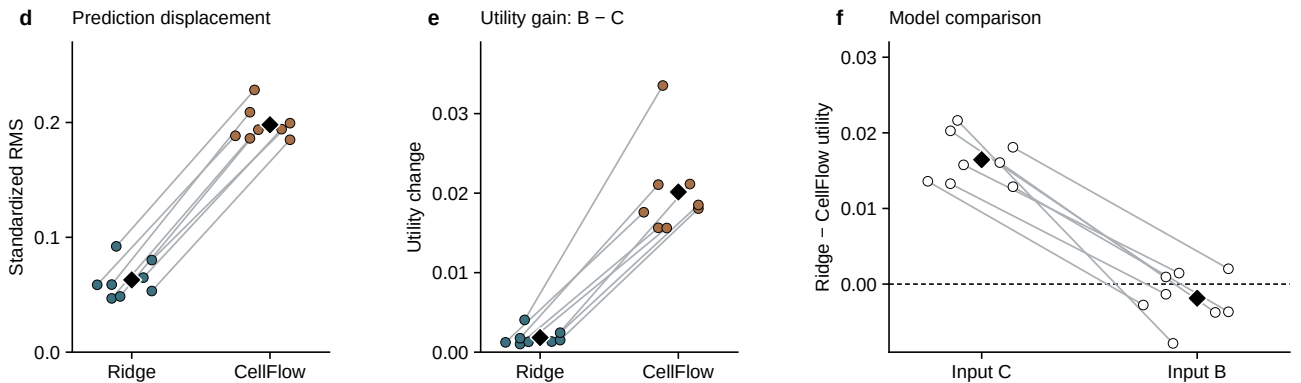
Model-input references across training configurations

Model input: C → B | Fixed observation controls: A; prediction baseline: B

Original fixed configurations



Selected training configurations



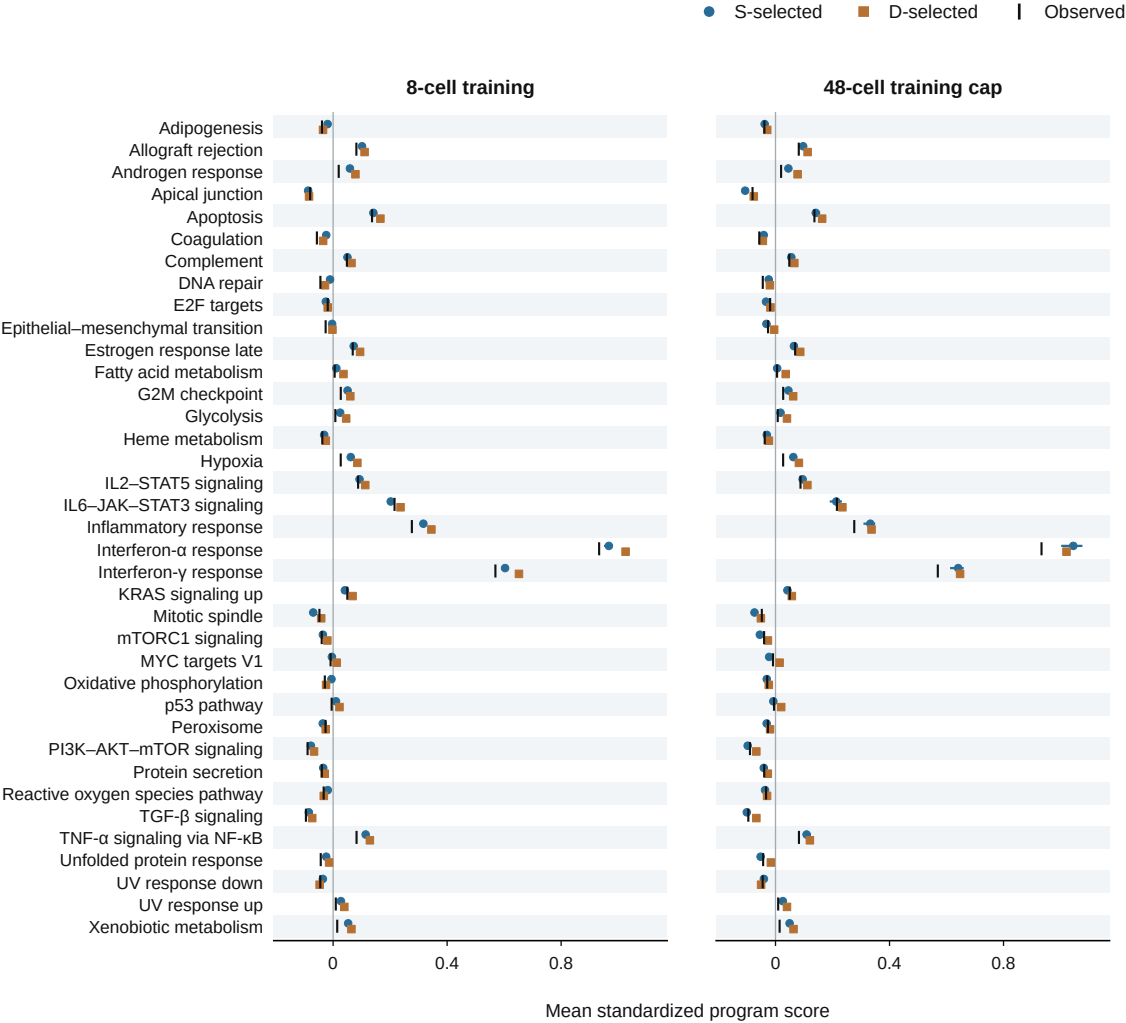
○ Paired donor (n = 8) ◆ Equal-donor mean

Supplementary Figure 1 | Model-input references in the Kang ridge–CellFlow comparison

Original fixed (a–c) and inner-donor-selected (d–f) configurations predict interferon- β responses in the same eight Kang donors. A/B/C are disjoint four-cell control blocks. Input changes from C to B while observation A and prediction reference B stay fixed; CellFlow source cells and mean context change together. **a,d**, Standardized RMS prediction displacement. **b,e**, Utility change $U(B) - U(C)$, using negative standardized MSE. **c,f**, Ridge-minus-CellFlow utility; positive favours ridge. Lines pair donors; black diamonds are equal-donor means. Each donor averages four cell types and six assignments; CellFlow uses five separately scored fits. RMS follows within-donor averaging of squared changes. Offsets separate marks. Assignments and initializations reuse donor cells; no inferential intervals are shown (Supplementary Table 2).

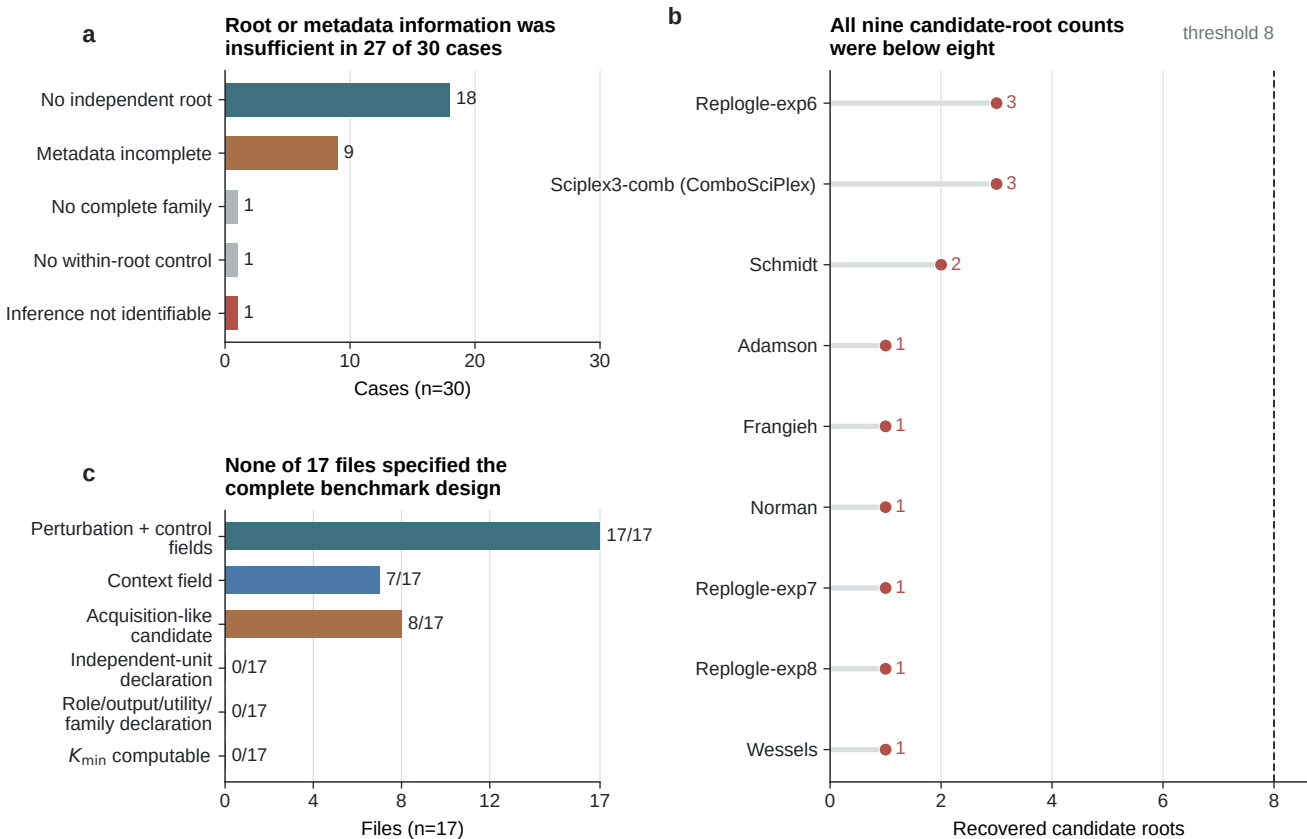
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750

Mean predicted and observed program responses



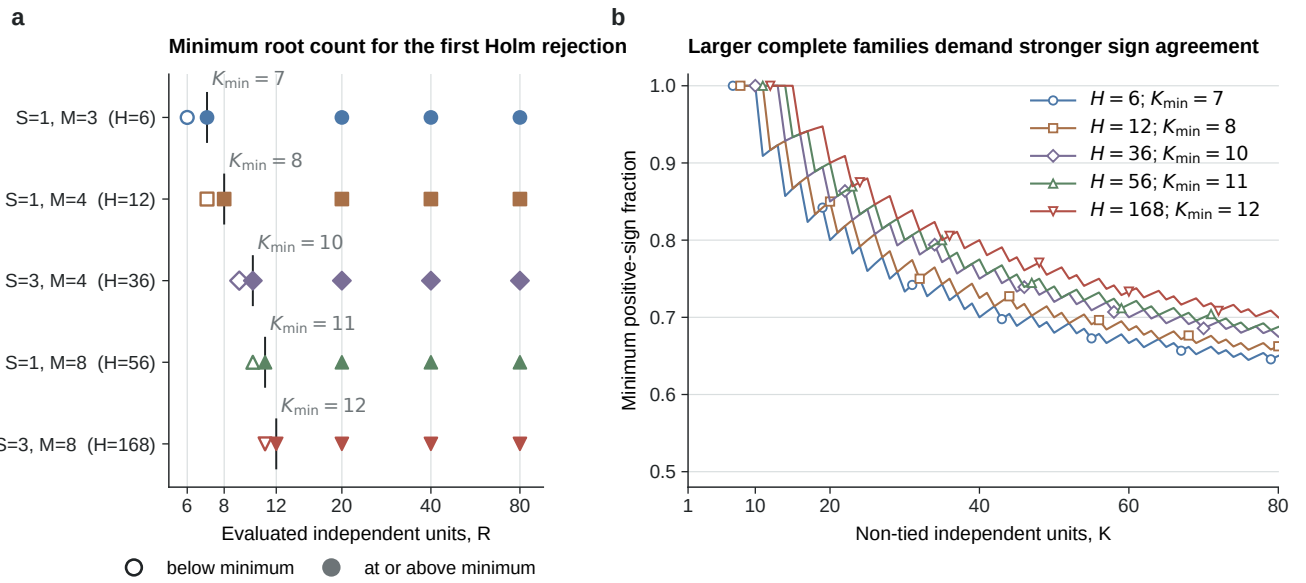
Supplementary Figure 2 | Mean predicted and observed program responses after reference-dependent selection

All 37 Hallmark programs under 8-cell training (left) and a 48-cell cap (right). S/D denote shared/disjoint-reference selection, excluding each donor from its own selection; O selected identical fits and predictions to D in every fold. Points equally average five cell types and eight donors, then seeds 17, 29 and 43; ranges span minimum-maximum seed-specific task means. Deterministic fits repeat across seeds. Black marks show the common B2-centred observed response. States use B1 for input and prediction centring; direct effects need no additional centring. Scores average training-standardized effects over represented genes in fixed program order. Fig. 4a shows disagreements before task averaging.



Supplementary Figure 3 | Experimental design information in screened public records

a, Missing design information in 30 benchmark records: 18 lacked identifiable independent roots; nine had incomplete metadata; one each lacked a complete model family, within-unit control or identifiable statistical procedure. **b**, Candidate-root counts in the nine incomplete records, all below the screening threshold of eight; inference also requires independence and adequate family-specific support. **c**, Annotations in 17 H5AD files: all identify perturbation/control, seven context and eight possible acquisition units. None specifies both independent units and the reference cells, outputs, utility and comparisons needed for inference or calculation of $K_{\min}$, the first-Holm-rejection minimum of non-tied independent units. The screens used separate non-probability samples.



Supplementary Figure 4 | Independent-unit requirements for directional testing

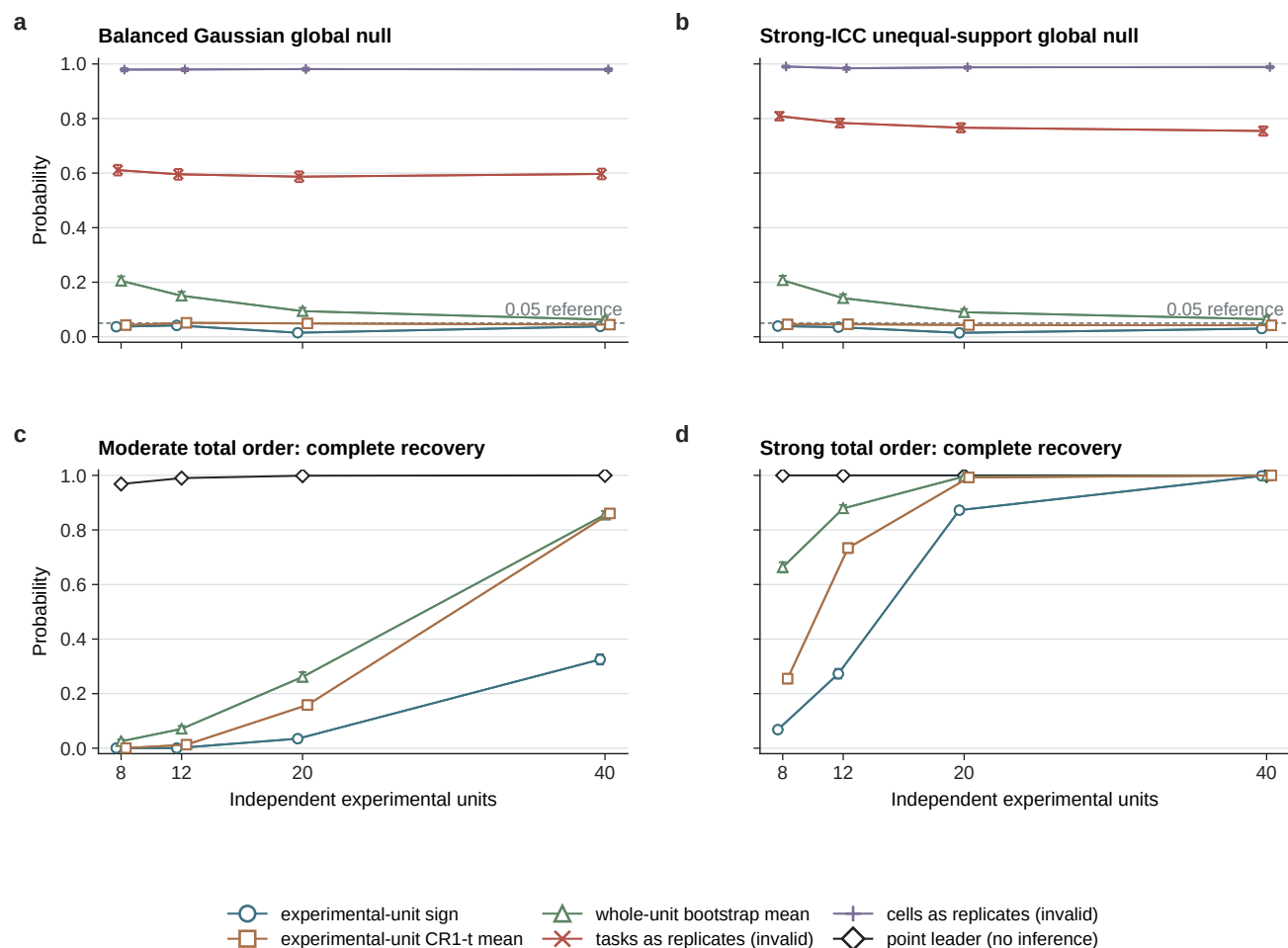
a, Root counts R in 25 designs with S reference constructions and M configurations, giving $H = SM(M-1)$ ordered comparisons. The five families contain 6/12/36/56/168 directions. At familywise $\alpha = 0.05$, black ticks mark $K_{\min}$, the first-Holm-rejection minimum of non-tied independent units. Open/filled markers indicate designs below/at or above this threshold; ties reduce usable units. **b**, Minimum positive-difference fraction for first rejection across $K = 1, \dots, 80$ non-tied units; positive favours the tested configuration. Below $K_{\min}$, even unanimity cannot reject. All 425 values follow the testing rule without outcome data.



Supplementary Figure 5 | Directional error across 255 simulated settings

Marks are simultaneous one-sided exact Clopper-Pearson upper bounds with tail allocation 0.05/255, based on 25,000 successful Monte Carlo replicates per setting. Rows show five geometries with S reference constructions, M configurations and H ordered directions; offsets separate settings. Filled circles represent 205 independent-root common-sign settings; open diamonds represent 50 independent heterogeneous rootwise-intersection settings. The vertical line marks the prespecified 0.06 limit. Generating laws are defined in Supplementary Methods.

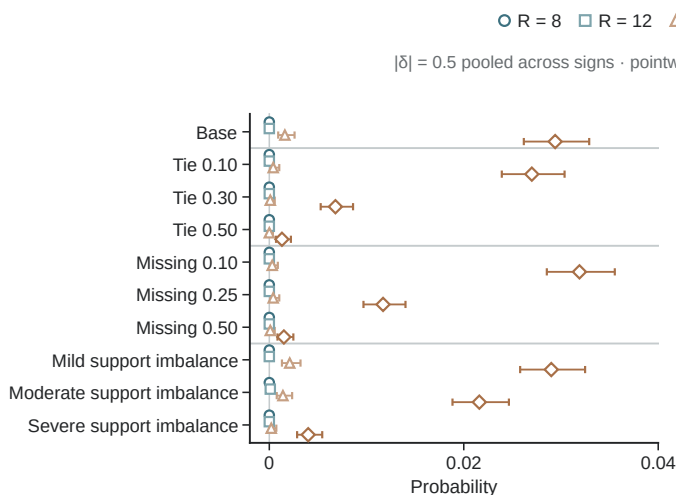
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002



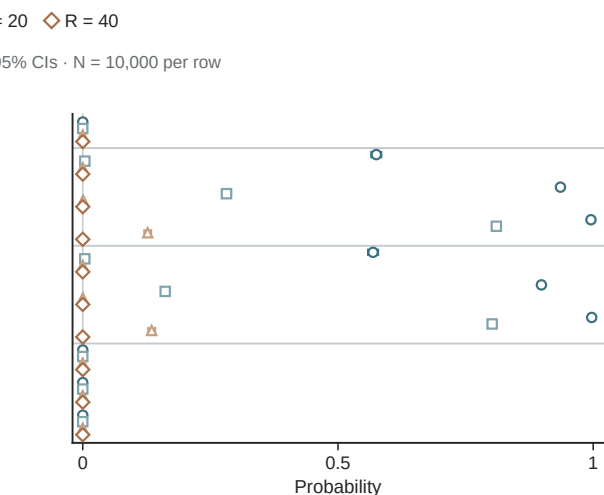
Supplementary Figure 6 | Experimental-unit and descendant-level inference

a,b, Any-direction rejection under balanced and strong-intraclass-correlation global nulls. Root-preserving procedures retain experimental units; task/cell procedures treat descendants as independent. The dashed line marks the nominal 0.05 level for root-preserving procedures. **c,d,** Complete target-relation recovery under moderate and strong ordered alternatives. Points summarize 2,500 Monte Carlo replicates; bars show pointwise two-sided 95% Clopper–Pearson intervals where defined. Black diamonds mark descriptive point leaders; lines connect procedures across root counts. Sign and equal-unit mean procedures target different estimands.

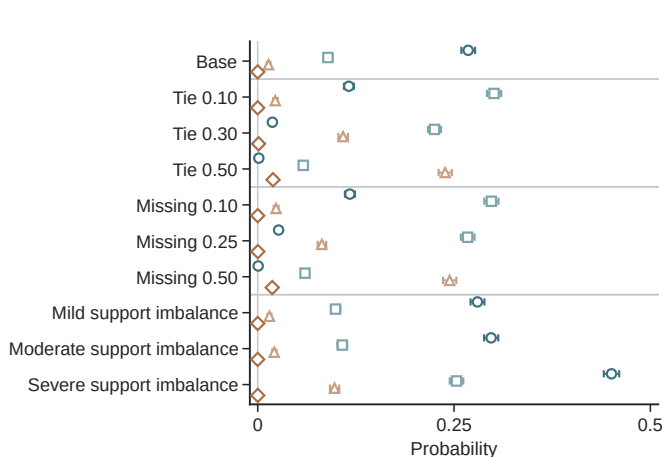
a Exact-family recovery



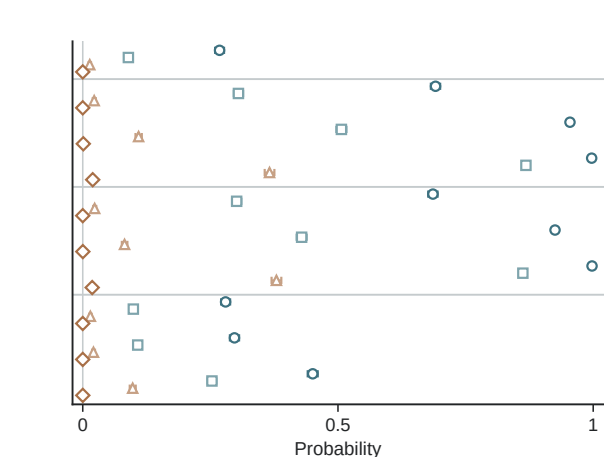
b Structural withholding



c Decision abstention



d Total no-direction rate



Supplementary Figure 7 | Structural withholding and decision abstention under stress

Ten stress states and four root counts ($R = 8, 12, 20, 40$) give 40 designs at $|\delta| = 0.5$. Each pools 5,000 independently generated replicates per sign of δ ; all endpoints share these 10,000 replicates. **a**, Exact-family recovery. **b**, Structural withholding. **c**, Decision abstention. **d**, Total no-direction rate. Points are pooled estimates; bars are pointwise two-sided 95% Clopper–Pearson intervals. Withholding and abstention partition the total; its interval is calculated directly. Axes differ by endpoint. Stress states vary ties, missing complete roots and heterogeneous support.