# REDUNDANT REGULATORY PROGRAMS CAN CHECK CELL GROUPS THEY LOCATE POORLY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Curated regulatory programs, such as pathway gene sets and regulons, are built into single-cell models, yet what they add beyond an expression encoder is unclear. We find that, used as linear features next to an encoder, programs estimate the encoder-defined state poorly but can test whether a group of cells fits the relations among programs seen in known cells. A geometric theorem explains why: the set of consistent states is wide along directions that program directions combine into only at high cost, whereas the threshold for detecting a shift of one program grows only with the cost of combining the others into it, which redundancy keeps small. Our conformally calibrated, state-invariant statistic, HellyCell, flags a group when no state fits all programs and returns a witness of at most $d+1$ programs that no state satisfies jointly ($d$: state dimension); if data are exchangeable and few programs shift, a flag whose witness misses every shifted program has probability at most the level $\alpha$. Confidence sets from programs alone are 3.7 to 8.1 times wider than the encoder's with least squares, and 19.6 to 24.7 times as the intersections of intervals that the check uses. Estimated on pilot cells, the part of a mean shift that no change of state explains predicts which of 154 real perturbations are detected in held-out groups of ten cells (rank correlation 0.89; 0.91 in the largest screen, against 0.81 for the raw shift): interferon responses are detected, most single-gene CRISPR perturbations are not. Baarda's $w$-test has more power on sparse or strong shifts, so HellyCell trades power for witnesses; invariance to the state fails under new sequencing protocols.

## 1 INTRODUCTION

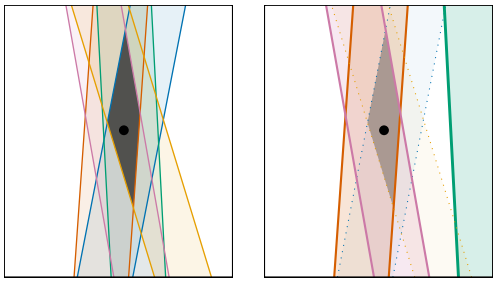
Curated regulatory knowledge, in the form of pathway gene sets and transcription-factor regulons, is increasingly built into single-cell models: as interpretable latent programs (Lotfollahi et al., 2023; Seninge et al., 2021; Kunes et al., 2024), as graphs for perturbation prediction (Roohani et al., 2024), and as activity scores (Schubert et al., 2018; Badia-i-Mompel et al., 2022; Müller-Dott et al., 2023). Yet careful evaluations find that simple baselines are hard to beat, for pretrained single-cell models (Kedzierska et al., 2025) and for perturbation prediction (Ahlmann-Eltze et al., 2025; Viñas Torné et al., 2026), and it remains unclear what regulatory structure adds to a strong expression encoder.

We find that, used as linear readings next to an encoder, programs add little to locating a cell's state but support checking whether a group of cells fits the relations among programs seen in known cells. On three datasets whose cells are thinned to a quarter of their counts, adding 380 to 540 curated programs to an expression encoder by stacking, message passing or an MLP changes the error of the full-depth state by $-5\%$ to $+4\%$, and, relative to the encoder's calibrated box, sets from the programs alone are 3.7 to 8.1 times wider with least squares and 19.6 to 24.7 times in the $\ell_\infty$ geometry used for checking (Table 1). Yet in groups of ten cells, where the set of states consistent with the programs is narrower than the population's central $90\%$ range along only 2 of 8 directions, the same programs flag 97% of held-out interferon-stimulated groups at a $5\%$ false-alarm level.

Our explanation is geometric (Figure 1). Hundreds of programs measure a state of a few dimensions, so each program's direction is re-measured by its neighbours, while some directions are measured well by none. By Theorem 2, the width of the set of states consistent with all programs along a direction $u$ lies between two multiples of the $\ell_1$ cost $\kappa(u)$ of synthesizing $u$ from program directions, whereas a shift of program $j$ is detected once it exceeds a threshold within a factor $1+\kappa_{-j}$ of the



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) a consistent cell (b) one program shifted

Figure 1: **Redundant programs check but locate poorly.** Each program constrains one projection $a_j^\top z$ of the state $z \in \mathbb{R}^2$ to a calibrated slab (strip). (a) Overlapping programs leave a feasible set (dark) that is thin across the strips but long along them, so the cell (dot) is located only coarsely. (b) A shifted program (right-most strip) empties the intersection, because its neighbours re-measure its direction; three programs, $d + 1$ for $d = 2$ (solid outlines), form a Helly witness that contains it.

one that knowing the state would give, where $\kappa_{-j}$ is the cost of synthesizing program $j$ from the others. The geometry also makes a prediction that could fail: a perturbation is detectable when the part of its mean shift that no change of state explains exceeds the calibrated threshold. Estimated on pilot cells, this signal-to-threshold ratio predicts detection in held-out groups across 154 real perturbations (rank correlation 0.89, 95% interval 0.83 to 0.94; Figure 3), and within screens better than the raw size of the shift.

The slabs are calibrated jointly by split conformal inference (Vovk et al., 2005; Lei et al., 2018), and a unit is flagged when they have no common point, which for a known unit happens with probability at most $\alpha$. Our statistic, **HellyCell**, is a Chebyshev fit that ignores the state and returns with each flag a *witness* of at most $d + 1$ programs ($d$: state dimension) whose slabs cannot hold together (Helly's theorem, Helly, 1923, made constructive by linear-programming duality). Each ingredient has precedents: reliability theory separates the precision of a least-squares estimate from the smallest detectable fault (Baarda, 1968), set-membership estimation bounds functionals over intersections of consistent sets (Milanese & Belforte, 1982), process monitoring separates residual alarms from typicality (Jackson & Mudholkar, 1979), and conformal inference calibrates sets. What is new is an $\ell_\infty$ geometry in which one atomic norm governs both tasks, with two-sided widths, exact per-program thresholds, short witnesses that are sound under sparse shifts, and a prospective test of the resulting envelope. HellyCell has limits: calibrated the same way, Baarda's $w$-test has more power on sparse or strong shifts and the global least-squares test on weak, broad ones; most single-gene CRISPR perturbations lie below its threshold; and invariance to the state fails when the sequencing protocol changes.

**Contributions.**

- **Localization and checking depend on different geometric quantities.** The width of the calibrated set along a direction is fixed, up to a factor, by the cost $\kappa(u)$ of synthesizing that direction from program directions, whereas the threshold for a shift of one program is fixed by the cost $\kappa_{-j}$ of synthesizing it from the others, and the two can be arbitrarily far apart when $d \geq 2$ (Theorem 2), although on our data the worst direction costs only about twice as much as re-measuring a program. On three datasets, widths and thresholds lie close to their upper bounds, and the thresholds carry over to shifts of a program's genes, which the theorem does not cover (Figure 2).
- **Calibrated, state-invariant checks with short witnesses.** Split-conformal calibration of residual statistics that ignore the state; HellyCell's witnesses have at most $d + 1$ programs and, with probability at least $1 - \alpha$, every witness contains a shifted program under exchangeability and sparse shifts (Theorems 1 and 3, Corollary 1; planted shifts in Q2).
- **A prospective operating envelope.** A signal-to-threshold ratio bounds the detection probability (Proposition 1); estimated on pilot cells, it predicts which real perturbations are detected in held-out cells, and invariance holds when only the state changes, not when the relation between programs and state does (Figures 3 and 4).

## 2 Related work

**Regulatory knowledge in single-cell models.** Pathway and regulon resources (Liberzon et al., 2015; Garcia-Alonso et al., 2019; Müller-Dott et al., 2023) are used to score activities (Aibar et al., 2017; Schubert et al., 2018; Badia-i-Mompel et al., 2022), also against matched control gene sets

(Zhang et al., 2022), to structure latent spaces (Lotfollahi et al., 2023; Seninge et al., 2021; Kunes et al., 2024) and to guide perturbation models (Roohani et al., 2024). Gains of foundation models (Cui et al., 2024; Theodoris et al., 2023; Hao et al., 2024; Wen et al., 2024; Levine et al., 2024) are tempered by zero-shot evaluations and linear baselines (Kedzierska et al., 2025; Ahlmann-Eltze et al., 2025; Viñas Torné et al., 2026). Perturbation screens are analysed by calling perturbed cells (Papalexi et al., 2021) and by distribution tests such as the energy distance (Peidli et al., 2024), and many perturbations have no detectable transcriptional effect (Replogle et al., 2022). We ask instead what regulatory knowledge adds as a set of measurements.

**Conformal inference for many outputs.** Split conformal prediction gives finite-sample coverage for any model (Vovk et al., 2005; Lei et al., 2018; Angelopoulos & Bates, 2023), with conditional refinements (Gibbs et al., 2025), guarantees beyond exchangeability (Tibshirani et al., 2019; Barber et al., 2023), max-type scores for simultaneous coverage (Diquigiovanni et al., 2022), outlier tests (Bates et al., 2023; Liang et al., 2024) and multi-output regions (Messoudi et al., 2021; Feldman et al., 2023; Gasparin & Ramdas, 2024), whereas intersections of per-view sets lose the guarantee (Garcia-Ceja, 2024); we calibrate all slabs jointly with a max-type score and test whether their intersection is empty.

**Redundant measurements, set membership and Helly-type certificates.** In geodetic reliability theory, the smallest detectable bias of an observation depends on its redundancy number and the precision of the estimate on the whole design (Baarda, 1968); fault detection and satellite integrity monitoring test the state-free residual of an overdetermined linear system (Chow & Willsky, 1984; Parkinson & Axelrad, 1988), with extensions to multiple faults (Blanch et al., 2010), and PCA-based process monitoring separates the squared prediction error, blind to moves within the model subspace, from Hotelling's $T^2$ (Jackson & Mudholkar, 1979). Set-membership estimation intersects measurement-consistent sets and bounds linear functionals over them by linear programs (Schweppe, 1968; Milanese & Belforte, 1982; Milanese & Vicino, 1991; Jaulin, 2009). Irreducible infeasible subsystems explain infeasible linear programs (Gleeson & Ryan, 1990; Chinneck, 2008), a discrete Chebyshev fit is fixed by $d + 1$ reference equations (Cheney, 1966), Helly's theorem has fractional versions (Katchalski & Liu, 1979; Kalai, 1984) and a colorful one, due to Lovász and reported by Bárány (1982), and optima fixed by few constraints underlie LP-type problems (Amenta, 1994; Matoušek et al., 1996), the scenario approach and optimal proper learning (Calafiore & Campi, 2006; Bousquet et al., 2020; Amenta et al., 2017). Unlike distance- and density-based out-of-distribution detectors (Liu et al., 2008; Lee et al., 2018; Sun et al., 2022; Yang et al., 2024), our statistics ignore the state; soundness under sparse shifts echoes invariant causal prediction (Peters et al., 2016) and the sparse-mechanism-shift hypothesis (Schölkopf et al., 2021).

## 3 Calibrated feasible sets from regulatory programs

**Units, states and programs.** A *unit* $y$ is a single cell or a group of $n$ cells drawn from the same stratum (for example, cell type and donor). Its state $z \in \mathbb{R}^d$ is given by an encoder fitted on training units of the known distribution; we use the top $d = 8$ principal coordinates of standardized log-expression (Appendix C varies $d$ and uses scVI and Geneformer states). A regulatory program $j \in [K]$ is a gene set $\mathcal{M}_j$ with signs $\omega_{jg} \in \{\pm 1\}$ (mode of regulation for transcription-factor regulons, $+1$ for pathways). Its *reading* $r_j(y)$ is the signed mean of standardized log-expression $\tilde{x}_g$ over $\mathcal{M}_j$, averaged over the unit's cells. A *view model* links readings to states,

$$r_j(y) = a_j^\top z + b_j + \sigma_j(y)\,\varepsilon_j, \qquad j = 1, \dots, K, \tag{1}$$

where $(a_j, b_j)$ are least-squares coefficients on training units and $\sigma_j(y)$ is a noise scale fitted on training residuals as a function of the unit size $n$ (Appendix B). For a candidate state $z$, the normalized residual is $s_j(z; y) = |r_j(y) - b_j - a_j^\top z|/\sigma_j(y)$; program $j$ defines the slab $C_j(q; y) = \{z : s_j(z; y) \le q\}$, and all programs together define the *feasible set* $F_q(y) = \bigcap_{j=1}^K C_j(q; y)$.

**Statistics, calibration and outputs.** The *consistency statistic* $T(y) = \min_z \max_j s_j(z; y)$ is the smallest multiplier at which $F_q(y)$ is non-empty: a Chebyshev fit, that is a linear program (LP) in $d+1$ variables with $2K$ constraints, whose minimizer, the Chebyshev centre, lies in every slab with margin $q - T(y)$. When a target state $z^\star$ is known (in our thinning experiments, the encoder-defined

**Algorithm 1** HellyCell: calibrated consistency test with Helly witnesses

**Require:** training and calibration units; a test unit $y$; programs $\{(\mathcal{M}_j, \omega_j)\}_{j \le K}$; level $\alpha$
Fit the encoder $z(\cdot)$ and the view model (1) $(a_j, b_j, \sigma_j)$ on training units
$\hat{q} \leftarrow$ the $\lceil (m+1)(1-\alpha) \rceil$-th smallest of $\{T(y_i)\}_{i \le m}$ on calibration units ($+\infty$ if it does not exist)
Solve the Chebyshev LP for the test unit: $T(y)$ and a basic optimal dual solution $\mu$
**if** $T(y) \le \hat{q}$ **then**
  **return** "not flagged"; optionally widths $W_u(F_{\hat{q}})$ and thresholds $\Delta_j^{\pm}$ (Theorem 2)
**else**
  $W \leftarrow \mathrm{supp}(\mu)$, so that $|W| \le d+1$ (Theorem 3)
  Deletion filter: for $k \in W$, drop $k$ if $\bigcap_{j \in W \setminus \{k\}} C_j(\hat{q}; y) = \emptyset$
  **return** "flagged" and the witness $W$
**end if**

state of the full-depth cell), the *localization statistic* is $M(y, z^\star) = \max_j s_j(z^\star; y) \ge T(y)$; the *anchored* statistic $A(y) = \max_j s_j(z(y); y)$ plugs in the unit's encoder state. Given $m$ calibration units exchangeable with the test unit, $\hat{q}$ is the $k$-th smallest calibration value of the statistic, $k = \lceil (m+1)(1-\alpha) \rceil$ ($\hat{q} = +\infty$ if $k > m$): *consistency calibration* uses $T$, *localization calibration* $M$ (Vovk et al., 2005; Lei et al., 2018). A unit is flagged if $T(y) > \hat{q}$, that is if $F_{\hat{q}}(y) = \emptyset$, and then receives a *witness*, at most $d+1$ programs whose slabs have empty intersection; passing is only the absence of evidence against consistency (Algorithm 1). The least-squares tests of fault detection, the residual sum of squares $G$ and Baarda's $w$-test $B$ (the largest residual standardized by its leverage $h_{jj}$; Baarda, 1968), are calibrated in the same way; Table 31 (Appendix E) collects the notation.

## 4 Theory: why programs check what they locate poorly

Fix a unit $y$ and drop it from the notation; all quantities below are computed at its noise scales, so the geometry depends on the unit size. Write $\tilde{a}_j = a_j / \sigma_j(y)$ for the normalized direction of program $j$, $\tilde{A} \in \mathbb{R}^{K \times d}$ for their stack and $\tilde{c}_j = (r_j(y) - b_j)/\sigma_j(y)$, so that $s_j(z) = |\tilde{c}_j - \tilde{a}_j^\top z|$. The *synthesis cost* of a direction $u \in \mathbb{R}^d$ is the $\ell_1$ cost of writing $u$ from program directions,

$$\kappa(u) = \min \left\{ \|w\|_1 : \sum_j w_j \tilde{a}_j = u \right\} \quad (+\infty \text{ if } u \notin \mathrm{span}\{\tilde{a}_j\}), \tag{2}$$

the gauge, or atomic norm (Chandrasekaran et al., 2012), of $H = \mathrm{conv}\{\pm\tilde{a}_1, \ldots, \pm\tilde{a}_K\}$: it is small when many programs jointly cover $u$. The *leave-one-out cost* $\kappa_{-j}$ is the synthesis cost of $\tilde{a}_j$ from the other programs, an $\ell_1$ cost of the kind that enters exact-recovery conditions in sparse approximation (Tropp, 2004); it is small when program $j$ is redundant with the others. $T_{-j}$ is the Chebyshev statistic without program $j$. The inradius $r(H) = \min_{\|u\|_2 = 1} \max_j |\tilde{a}_j^\top u|$ equals $1/\max_{\|u\|_2=1} \kappa(u)$ (zero if the $\tilde{a}_j$ do not span $\mathbb{R}^d$), and $W_u(F) = \max_{z, z' \in F} u^\top (z - z')$ is the width of a set $F$ along $u$. Appendix A gives the proofs and discusses how each assumption is met or tested.

**Assumption 1** (Exchangeability)**.** The view model is fitted on training units, and the calibration units and a known test unit are exchangeable (independent and identically distributed where a result compares a shifted unit with a known one).

**Assumption 2** (Sparse shift)**.** A shifted unit changes the readings of a set $S \subseteq [K]$ of programs; the programs outside $S$ follow (1) at the unit's state, with $(\varepsilon_k, \sigma_k(y))_{k \notin S}$ jointly distributed as for known units, and the unit is independent of the calibration units.

**Assumption 3** (State-respecting shift)**.** A test unit follows (1), with the coefficients fitted on training units, at some state $z_{\mathrm{new}}$ (possibly never seen in training), with $(\varepsilon, \sigma(y))$ jointly distributed as for known units, and is independent of the calibration units.

**Theorem 1** (Finite-sample validity)**.** *Under Assumption 1, localization calibration gives* $\Pr[z^\star \in F_{\hat{q}}(Y)] \ge 1 - \alpha$*, and consistency calibration gives* $\Pr[F_{\hat{q}}(Y) = \emptyset] \le \alpha$ *for a known unit. Moreover* $T(y) \le M(y, z^\star)$ *for every* $y$*, so the consistency threshold never exceeds the localization threshold computed on the same calibration units.*

Theorem 1 is split-conformal inference for a max-type score, which controls the error over all $K$ programs jointly (Appendix C).

**Theorem 2** (Localization depends on the worst direction; checking on a program's own)**.** *Fix a unit and a multiplier $q > 0$ with $q \geq T(y)$; in lower bounds, $0 \cdot \infty = 0$.*

*(a)* Localization. *For every direction $u$, $2\,(q - T(y))\,\kappa(u) \leq W_u(F_q) \leq 2q\,\kappa(u)$; hence $2(q - T(y))/r(H) \leq \operatorname{diam} F_q \leq 2q/r(H)$.*

*(b)* Checking. *Shift the reading of program $j$ by $\Delta$, in units of $\sigma_j$. The shifted feasible set is empty if and only if $\Delta > \Delta_j^+$ or $-\Delta > \Delta_j^-$, where $\Delta_j^\pm \geq 0$ are given by two LPs and satisfy*

$$q + \big(q - T_{-j}(y)\big)\kappa_{-j} \;\leq\; \tfrac{1}{2}\big(\Delta_j^+ + \Delta_j^-\big) \;\leq\; q\,(1+\kappa_{-j}), \qquad \max\{\Delta_j^+, \Delta_j^-\} \leq 2q\,(1+\kappa_{-j}).$$

*(c)* Separation. *If $d \geq 2$, then for every $0 < \epsilon \leq 1$ there are programs with $\max_j \kappa_{-j} \leq 1$ and $r(H) < \epsilon \max_j \|\tilde{a}_j\|_2$.*

In words: the feasible set is wide exactly along directions that are expensive to synthesize, whereas the threshold for program $j$ involves only $j$'s own direction and how cheaply the others reproduce it; by (c), the two can be arbitrarily far apart once $d \geq 2$. The widths in (a) are the uncertainty intervals of set-membership estimation (Milanese & Belforte, 1982), here bounded on both sides through an atomic norm. We call $\Delta_j^\pm$ the *exact detection thresholds*, $q(1 + \kappa_{-j})$ the *synthesis scale* and $2q(1 + \kappa_{-j})$ the *worst-case sufficient threshold*: a shift beyond it is always detected, and one beyond the synthesis scale in at least one sign; with the state known, the two-sign threshold of program $j$ would be $q$, so $1 + \kappa_{-j}$ is the price of not knowing it. By (a), a unit with $T(y)$ well below $\hat{q}$ cannot be pinned down along expensive directions. For point prediction, readings are functions of the measured expression, so they add information about $z^\star$ only through genes outside the encoder's input or through regularization.

**Proposition 1** (Operating envelope)**.** *The map $N(x) = \min_\delta \|x - \tilde{A}\delta\|_\infty$ on $\mathbb{R}^K$, which gives $T(y) = N(\tilde{c})$, is a seminorm that vanishes exactly on shifts acting through the state; a shift $\Delta$ of one program has size $|\Delta|/(1 + \kappa_{-j})$. Let a unit have normalized readings $\tilde{A}z + s + \varepsilon$, with a state $z$, a fixed shift $s$ and noise $\varepsilon$ distributed as for known units and independent of the calibration units. Conditionally on the calibration units, with $\rho = N(s)/\hat{q}$ and probabilities taken over $\varepsilon$,*

$$\Pr\big[N(\varepsilon) < (\rho - 1)\,\hat{q}\big] \;\leq\; \Pr[\textit{flag}] \;\leq\; \Pr\big[N(\varepsilon) > (1 - \rho)\,\hat{q}\big].$$

*Marginally, under Assumption 1, $\Pr[\textit{flag}] \geq 1 - \alpha - \Pr[\hat{q} \geq N(s)/2]$; in particular a unit whose shift satisfies $N(s) > 2\hat{q}$ for every calibration draw is flagged with probability at least $1 - \alpha$.*

The law of $N(\varepsilon)$ is that of $T$ for a known unit, so both bounds are estimated from calibration units, without free parameters. In least squares, the analogous quantities are the precision $(u^\top(\tilde{A}^\top\tilde{A})^{-1}u)^{1/2}$ along $u$ and the redundancy number $1 - h_{jj}$ (Baarda, 1968): the $w$-test keeps a share $\sqrt{1 - h_{jj}}$ of a shift, close to one when $K \gg d$, whereas the Chebyshev fit can absorb a share $\kappa_{-j}/(1 + \kappa_{-j})$ into the state; the $\ell_\infty$ geometry costs power, and what it buys is the next result.

**Theorem 3** (Helly witnesses are short, computable and sound)**.** *(a) If $F_q(y) = \emptyset$, there is $W \subseteq [K]$ with $|W| \leq d + 1$ and $\bigcap_{k \in W} C_k(q; y) = \emptyset$; the support of any basic optimal dual solution of the Chebyshev LP is such a set, and the deletion filter of Algorithm 1 reduces it to an irreducible one. (b) Under Assumptions 1 and 2 and either calibration, with probability at least $1 - \alpha$ every infeasible subsystem at level $\hat{q}$ contains a shifted program. Hence the probability that a unit is flagged with a witness that names no program in $S$ is at most $\alpha$, and given a flag it is at most $\min\{1, \alpha/\beta\}$, where $\beta$ is the probability of a flag.*

Part (a) is Helly's theorem (Helly, 1923) for slabs, or the $d + 1$ reference equations of a discrete Chebyshev fit (Cheney, 1966), made constructive by LP duality; the size $d+1$ is attained generically, since any $d$ slabs with linearly independent directions intersect. Part (b) holds for any statistic that cannot decrease when programs are added and is state-invariant on unshifted programs, such as $G$; specific to $T$ is that a sound witness of at most $d+1$ programs always exists and one LP finds it. It is a sparse-shift result, tested only on planted shifts: real responses, and even shifts of one program's genes, move hundreds of readings, so almost any witness meets $S$.

**Corollary 1** (Pivotality: blind to shifts that act through the state)**.** *Under Assumptions 1 and 3, $T(Y)$ has the same distribution as for a known unit, so a unit is flagged with probability at most $\alpha$, however far $z_{\mathrm{new}}$ is from the training states.*

Table 1: **Regulatory programs add little localization.** Test cells are thinned to 25% of their counts; the target is the encoder-defined state of the full-depth cell. Top: errors relative to a ridge encoder of the thinned expression (spread: mean distance of states to their mean). Bottom: median calibrated radius of calibrated sets that must contain the target, each around its own estimate with its own score, relative to the backbone box. Means over three random splits (Kang: control cells); standard deviations over splits are at most 0.05 for error ratios and 8% of each radius ratio (Appendix C, with a wider target).

| | PBMC3K | Kang (ctrl) | Pancreas |
|---|---|---|---|
| *State error relative to the backbone, state on the encoder's genes (lower is better)* | | | |
| Backbone ridge encoder (absolute error / state spread) | 1.00 (3.25/9.5) | 1.00 (2.85/10.2) | 1.00 (2.22/14.6) |
| + programs, noise-weighted fusion | 1.22 | 1.23 | 1.63 |
| + programs, cross-fitted stacking | 1.04 | 0.99 | 0.98 |
| + programs, message passing on the program graph | 0.98 | 0.99 | 0.98 |
| MLP encoder | 1.07 | 1.05 | 1.09 |
| + programs as extra MLP inputs | 1.05 | 1.04 | 1.04 |
| Programs only (weighted least squares) | 2.61 | 2.81 | 3.31 |
| *Calibrated radius at $1-\alpha = 0.9$, relative to the backbone box (coverage in parentheses)* | | | |
| Backbone box ($\ell_\infty$ over the encoder's coordinate views) | 1.0 (0.90) | 1.0 (0.90) | 1.0 (0.90) |
| Backbone ball (least squares) | 1.4 (0.91) | 1.4 (0.90) | 1.2 (0.90) |
| Programs only, least-squares ball | 3.8 (0.91) | 4.0 (0.90) | 3.7 (0.90) |
| Programs only, least-squares ellipsoid | 5.1 (0.89) | 5.6 (0.90) | 8.1 (0.90) |
| Programs only, $\ell_\infty$ feasible set $F_{\hat{q}}$ (the test's geometry) | 24.1 (0.90) | 24.7 (0.90) | 19.6 (0.90) |
| Backbone + programs, least-squares ellipsoid | 2.4 (0.90) | 2.1 (0.90) | 3.3 (0.90) |
| Backbone + programs, $\ell_\infty$ feasible set | 2.8 (0.90) | 2.5 (0.90) | 2.0 (0.90) |

Substituting $\delta = z - z_{\text{new}}$ gives $T(y) = \min_\delta \max_j |\varepsilon_j - \tilde{a}_j^\top \delta|$, free of the state, and likewise for $G$ and $B$. The corollary covers changes of the state under an unchanged view model, such as known programs in new combinations, to which typicality detectors (distances or densities in state space) respond instead; new donors, technologies or cell types may change the view model and are stress tests. If the population coefficients differ from the fitted ones by $(\delta a_j, \delta b_j)$, the residual also contains $\delta a_j^\top z_{\text{new}} + \delta b_j$, which we measure in Section 5.

## 5 Experiments

We ask four questions. **(Q1)** Do programs add localization beyond an encoder, and how much of their gap is the $\ell_\infty$ geometry? **(Q2)** How tight are the bounds of Theorem 2, do they carry over to shifts they do not cover, and are witnesses sound? **(Q3)** Can the detection of real changes be predicted before it is measured? **(Q4)** Does invariance hold under state shifts and fail under view-model shifts? The theorems hold by construction and are verified numerically on all sampled units; tests that could fail are the comparator (Q1), uncovered and multi-program shifts (Q2), the prospective test (Q3) and the shifts of Q4.

**Setup.** We use PBMC3K (Wolf et al., 2018), interferon-$\beta$ (IFN-$\beta$) stimulated and control PBMCs from eight donors (Kang et al., 2018), a pancreas atlas of nine batches from six sequencing technologies (Luecken et al., 2022), a Jurkat CRISPR screen with and without T-cell receptor (TCR) stimulation (Datlinger et al., 2017), and two CRISPR screens in K562 cells (Dixit et al., 2016; Norman et al., 2019). Programs are the 50 MSigDB hallmarks (Liberzon et al., 2015) and CollecTRI regulons (Müller-Dott et al., 2023) with at least 10 detected genes, near-duplicates removed, 380 to 540 per dataset. The state is the top $d = 8$ principal coordinates of 2,000 highly variable genes fitted on training units, with $d$ fixed in advance; units are single cells or disjoint groups of $n$ cells from one stratum (Appendix B). On Kang, detection and invariance experiments rotate donors (four for training, two each for calibration and testing); others use stratified random splits. All detectors are calibrated by the same conformal rule at $\alpha = 0.05$: the least-squares tests $G$ and $B$, the $\ell_1$ fit, typicality detectors in state space (kNN distance, Mahalanobis distance, isolation forest; Sun et al., 2022; Lee et al., 2018; Liu et al., 2008) and the largest $|z|$-score of any reading. Controls replace curated programs by size-matched random gene sets or by gene-label shuffles that keep sizes, overlaps and gene degrees.

**Q1: programs add little localization, and the $\ell_\infty$ geometry adds a further cost (Table 1).** With test cells binomially thinned to a quarter of their counts (Neufeld et al., 2024), adding the programs to an encoder of the thinned expression changes the error of the full-depth state $z^\star$ by $-5\%$ to

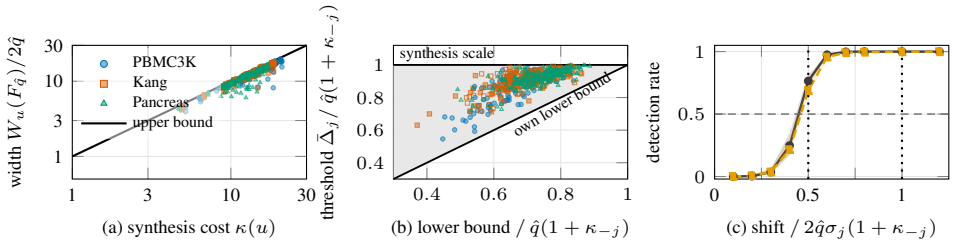



Figure 2: **Widths and thresholds lie close to the bounds of Theorem 2, which carry over to shifts they do not cover.** (a) Width of the programs-only set of thinned cells along 40 directions per cell against $\kappa(u)$ (170 random pairs per dataset). (b) Exact symmetric threshold $\bar{\Delta}_j = \frac{1}{2}(\Delta_j^+ + \Delta_j^-)$ of units that pass the test against its own lower bound, both over the synthesis scale (110 pairs per configuration; colours as in (a), open markers: groups of ten); shaded: the band of Theorem 2(b). (c) Detection of planted shifts of one program's reading (view level, solid) or of its genes (expression level, dashed); mean and range over six configurations.

$+4\%$, whether they enter a ridge encoder by cross-fitted stacking or message passing on the gene-program graph, or an MLP as extra inputs; noise-weighted fusion, the weighted least-squares state of the encoder's coordinate views and the programs together, is worse by 22% to 63%. Larger gains need information outside the encoder's input: when the state depends on every gene that the programs read, they reduce the error by up to 25%, but their $\ell_\infty$ sets stay 14 to 21 times wider than the backbone box (Appendix C). To separate information from geometry, we compare sets that must contain $z^\star$ with probability 0.9 by their *calibrated radius*, the half-diagonal of the smallest box centred at the point estimate that contains the set: the *backbone box*, the feasible set of the encoder's $d$ coordinate views with a joint multiplier; least-squares balls and ellipsoids around the weighted least-squares state; and the $\ell_\infty$ feasible set $F_{\hat{q}}$. All reach their nominal coverage. From the programs alone, least-squares sets are 3.7 to 8.1 times wider than the backbone box, and a least-squares ball is 2.7 to 3.1 times the encoder's own ball, so the programs locate the state less precisely than the encoder in either geometry (point error 2.6 to 3.3 times); the $\ell_\infty$ set is a further 2.4 to 6.4 times wider than the least-squares sets, the price of the geometry that makes witnesses short.

**Q2: the bounds are tight, carry over to spill-over shifts, and witnesses are sound (Figure 2).** *Widths.* For thinned cells, where $T$ is far below $\hat{q}$, the calibrated width along each of 40 directions is at a median 0.92 to 0.95 of its upper bound $2\hat{q}\kappa(u)$ (known units: 0.71 to 0.85), so $\kappa(u)$ sets the width. *Thresholds.* For 60 known units that pass the test and 20 random programs each, in seven configurations, every exact threshold lies in its own band, those of curated programs at a median 0.88 to 0.94 of the synthesis scale, far above their own lower bounds (median 0.67 to 0.74). *A unit-free comparison in the checking regime.* For groups of ten Kang cells, the calibrated set is narrower than the central 90% range of group states along only 2 of 8 principal directions of $\tilde{A}$ (0.7 to 4.5 times that range), whereas the synthesis scale of a program is only a factor $1 + \kappa_{-j}$ (median 1.8) above what knowing the state would allow; in natural units it is still a median of 5 across-group standard deviations of the program's reading (10 for the worst-case sufficient threshold), so single programs must change a lot, whereas broad responses add up (Appendix C). *Beyond the theorem.* Planted view-level shifts of one program obey both guarantees (in every configuration, at least 73% detected at 0.5 of the worst-case sufficient threshold and 100% at 1). Expression-level shifts move all genes of a program, and with them the readings of a median of 254 to 324 overlapping programs, which the theorem does not cover; they follow the same curve (65% to 76% detected at 0.5, at least 99.6% at 1). *Soundness.* Shifting one to five programs at once without conditioning on passing, $T$ flags a unit with a witness that misses every shifted program in at most 1.3% of cases (bound: 5%); the $w$-test detects single shifts earlier (18% against 4% at 0.3 of the worst-case threshold), and its largest residual is as often a shifted program on these shifts, though without a guarantee (Appendix C).

**Q3: real changes are detected when the threshold predicts it (Figure 3, Table 2).** With all IFN-$\beta$-stimulated Kang cells held out, state-invariant tests detect the response in groups: at $n = 10$, $T$ flags 97% and the $w$-test 100% of stimulated units, and at $n = 5$ the $w$-test is stronger (95% against 80%). Single cells need the state (anchored statistic 69%, kNN 66%, $T$ 13%). With one marginal quantile, the false-alarm rate of $T$ on control groups of a single cell type reaches 25% at $n = 10$; a quantile per cell type (Gibbs et al., 2025) brings it to at most 7%, but leaves rare types

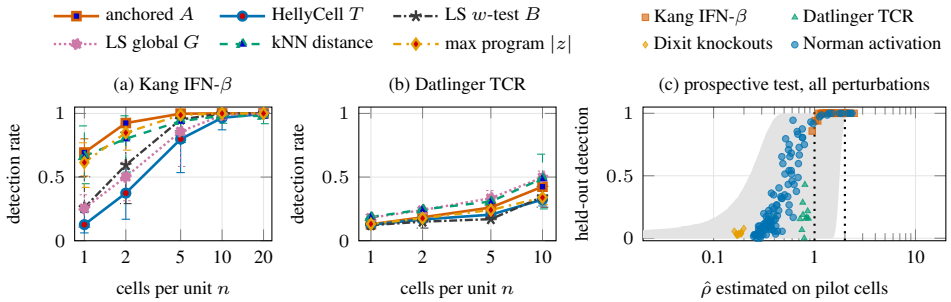



Figure 3: **Real regulatory changes**, never seen in training or calibration, at a nominal false-alarm rate of 5%. (a) IFN-$\beta$ stimulation of Kang cells and (b) TCR stimulation of Jurkat cells: means and ranges over four donor rotations or three splits. (c) Detection of held-out groups of ten cells against $\hat{\rho}$ estimated on disjoint pilot cells; shaded: the envelope of Proposition 1 from calibration units; dotted: $\hat{\rho} = 1$ and 2.

Table 2: **Witnesses certify joint infeasibility, ranked lists rank.** Explanations of detected stimulated units: share naming a reference response program (interferon or TCR programs fixed in advance; Appendix B) / share of named programs that are reference programs. A witness certifies that its programs cannot hold jointly (Theorem 3(a)); these broad responses are not covered by Theorem 3(b), so the shares are descriptive.

| | | | Kang IFN-$\beta$ | | Datlinger TCR | |
|---|---|---|---|---|---|---|
| Explanation (any / precision) | size | certificate | $n = 1$ | $n = 10$ | $n = 1$ | $n = 10$ |
| Helly witness | $\leq d + 1$ | Thm. 3(a) | 0.88 / 0.14 | 0.96 / 0.16 | 0.61 / 0.10 | 0.69 / 0.11 |
| Top-9 anchored residuals | 9 | none | 0.99 / 0.57 | 1.00 / 0.82 | 0.54 / 0.10 | 0.66 / 0.14 |
| Top-9 raw program $\|z\|$ | 9 | none | 0.98 / 0.59 | 1.00 / 0.73 | 0.47 / 0.09 | 0.67 / 0.15 |
| Top-9 $w$-test residuals | 9 | none | 0.91 / 0.28 | 1.00 / 0.55 | 0.58 / 0.12 | 0.80 / 0.18 |
| Largest $w$-test residual | 1 | none | 0.63 / 0.63 | 0.91 / 0.91 | 0.16 / 0.16 | 0.32 / 0.32 |
| 9 random programs | 9 | none | 0.20 / 0.02 | 0.20 / 0.02 | 0.44 / 0.06 | 0.44 / 0.06 |

too few calibration groups and cuts mean detection over types from 98% to 56%. TCR stimulation moves Jurkat cells in state space (an energy test is significant for 10 of 11 guide targets), but mostly along directions the view model already explains: its $\hat{\rho}$ stays below one, $T$ and the $w$-test flag only 32% and 34% of 10-cell groups, and the global test and kNN are the strongest (50% and 49%).

*Prospective test.* For each of 154 real perturbations (28 Kang, 11 Datlinger, 10 Dixit and 105 Norman), we estimate $\hat{\rho} = N(s)/\hat{q}$ on a random half of the cells and measure detection on groups of ten from the other half, cross-fitting over three splits (Appendix C). Pilot $\hat{\rho}$ predicts held-out detection (rank correlation 0.89, 95% interval 0.83 to 0.94 over perturbations): the 75 perturbations with $\hat{\rho} \leq 0.5$ are flagged in 13% of their groups on average, the 42 with $0.5 < \hat{\rho} \leq 1$ in 51%, and each of the 37 with $\hat{\rho} > 1$ in at least 94%. No detection rate leaves the envelope, which, however, binds only at its ends. Within screens, the ranking holds for Norman (0.91) and Kang (0.60), but not over the narrow range of Datlinger targets (0.14, interval $-0.48$ to 0.62). In Norman and Kang, removing the state component matters: the raw shift size predicts less well (Norman 0.81, paired difference 0.10, interval 0.03 to 0.18; Kang 0.35), a least-squares residual size as well or better (Norman 0.94; Appendix C). IFN-$\beta$ has $\hat{\rho}$ of 0.95 to 2.47 across cell types and rotations, the 10 knockouts at most 0.20; an energy test (Peidli et al., 2024) detects no effect of any knockout, so their flag rate near $\alpha$ reflects no detectable effect, whereas 102 of the 105 activations have one and their detection rises with $\hat{\rho}$: most single-gene perturbations are below the threshold rather than without effect (Appendix D). $T$ rarely flags a group that both least-squares tests miss (on average 1.1% of a perturbation's groups, at most 8.7%).

*Witnesses* have exactly $d + 1 = 9$ programs in all our flags, as general position predicts, and name a reference response program in 96% of detected 10-cell IFN-$\beta$ units (chance: 21%, Table 23); gene-shuffled or random programs still detect 74% and 89% of the units, but their witnesses name one in only 28% and 10%: detection rests largely on redundancy, what witnesses name on curated biology. But only 0.16 of their members are reference programs, against 0.82 for the nine largest anchored residuals and 0.91 for the largest $w$-test residual (Table 2), because a witness also contains the programs that pin down the violated one: it certifies but does not rank culprits.

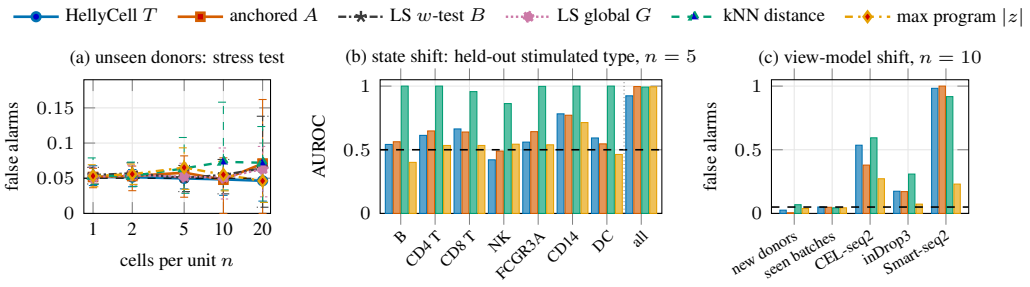



Figure 4: **Invariance holds for state shifts, not for view-model shifts.** (a) False alarms on known units of unseen Kang donors (means and ranges over four rotations; dashed: $\alpha$). (b) Stimulated cells of one Kang type are novel, those of all other types known; AUROC against control cells of the same type; right: all stimulated cells held out. (c) False alarms on known cell types under an unseen pancreas batch or protocol, and under unseen Kang donors. Bars in (b, c) show four of the six detectors (one rotation).

**Q4: invariance holds for state shifts, not for view-model shifts (Figure 4).** When only the stimulated cells of one cell type are held out, their known response in a new context acts through the state: kNN separates them from unstimulated cells of the same type (AUROC 0.86 to 1.00), whereas $T$ reaches only 0.42 to 0.66 for 6 of 7 types; the exception, CD14 monocytes (0.78), also has an elevated false-alarm rate on its own control cells (16%). The view model's estimation error at the held-out states is at most 0.03 slab half-widths (Appendix C), so the departures from 0.5 reflect changed program-state relations or noise laws, which Assumption 3 excludes; once stimulated cells enter training, $T$ no longer sees the same response in a new cell type. A new protocol or batch changes the view model itself, and $T$ flags 17% (inDrop3) to 98% (Smart-seq2) of 10-cell groups of known cell types, against 5% for seen batches (a quality signal, not novelty), and a cell type absent from training is flagged only if its programs relate to the state differently (9% to 99% of its groups). New donors may differ in more than their states: averaged over rotations, the false-alarm rate of $T$ stays between 4.6% and 5.2% at every group size, but single rotations range from 1.8% to 7.3%, and over 16 further donor partitions it reaches 15.4% at $n = 10$; on Norman control groups it is 7.6% (119 groups; binomial $p = 0.14$). Invariance is thus a property of the view model, and calibration units must match the donors, technologies and cell types in use.

## 6 Discussion, Limitations and Conclusion

**Limitations.** Our localization results concern linear program readings next to an expression encoder, with an encoder-defined target state: we did not train program-informed latent models (Lotfollahi et al., 2023), and readings are signed means rather than weighted activity scores (Badia-i-Mompel et al., 2022). The view model is linear in the state, so nonlinear relations inflate the noise scales and cost power. Reliable detection needs $\hat{\rho}$ above one, which for a single program means a change of several across-group standard deviations, and groups of five to ten cells formed within annotated strata; we did not evaluate clustering-based groups. Under sparse shifts, with probability at least $1 - \alpha$ every witness contains a shifted program, but it does not say which, and this guarantee is verified only with planted shifts: we found no real perturbation whose shifted set is sparse and known. Error control is marginal over cell types and approximate across donors, and invariance needs a fixed view model. The Chebyshev LP takes tens of milliseconds for $K \leq 800$ but would need column generation at genome scale.

**Conclusion.** Curated regulatory programs need not be accurate state estimators to be useful measurements: localization is limited by the directions they measure poorly, checking a program only by how well the others re-measure it, and the resulting threshold predicted which real perturbations are detected in held-out cells. Programs should therefore be judged also as checks: among state-invariant tests, least-squares tests give more power and HellyCell adds a short certificate to each flag; outside single-cell genomics we have tested the geometry only in a synthetic world (Appendix C).

## AI USE STATEMENT

In this work we used generative AI tools (a large-language-model assistant operating as a coding and writing agent) for the following tasks that require disclosure: implementing the methods and experiment scripts, and helping to interpret results in discussion with the authors. We also used AI tools, as recommended for disclosure, to draft and polish parts of the text, to review drafts of the paper, to search for and summarize related literature, to format references, to generate the plotting code for all figures, and to create slides for internal meetings. We did not use generative AI to produce data, to translate text, or to clean or relabel data beyond the scripted preprocessing described in Appendix B; the synthetic world of Appendix C is a scripted simulation. All AI-assisted work was reviewed by the authors: the code was tested end to end, every number in the paper is generated by script from the logged results, and each reference was checked against the publisher record or a bibliographic database. We take full responsibility for the content of this paper, including text, claims, code and figures produced with the aid of generative AI.

## REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 specify the method, and every quantity in it, including the widths, exact thresholds and synthesis costs used to test the theory, is computed by linear programs; at $d = 8$ and $K$ between 50 and 800, the consistency LP takes 12 to 59 ms per unit and a witness 99 to 153 ms on one CPU thread (Appendix C). Appendix A contains complete proofs with all assumptions stated in Section 4. Appendix B lists data sources, preprocessing, gene universes, program filtering, splits and donor rotations, the noise models, all hyper-parameters, the pre-specified reference response programs, the pilot and held-out splits of the prospective test and the compute used. All datasets are public, and every table and figure is produced by scripts from logged results with fixed seeds.

## ETHICS STATEMENT

We analyse publicly available, de-identified single-cell datasets and collect no new human data. The method flags cells or groups whose regulatory programs are jointly inconsistent. It is a research tool, and its outputs are statistical statements about consistency with a reference, not clinical assessments.

## REFERENCES

Mohammad Lotfollahi, Sergei Rybakov, Karin Hrovatin, Soroor Hediyeh-zadeh, Carlos Talavera-López, Alexander V. Misharin, and Fabian J. Theis. Biologically informed deep learning to query gene programs in single-cell atlases. *Nature Cell Biology*, 25(2):337–350, 2023.

Lucas Seninge, Ioannis Anastopoulos, Hongxu Ding, and Joshua Stuart. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nature Communications*, 12(1):5684, 2021.

Russell Z. Kunes, Thomas Walle, Max Land, Tal Nawy, and Dana Pe'er. Supervised discovery of interpretable gene programs from single-cell data. *Nature Biotechnology*, 42(7):1084–1095, 2024.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024.

Michael Schubert, Bertram Klinger, Martina Klünemann, Anja Sieber, Florian Uhlitz, Sascha Sauer, Mathew J. Garnett, Nils Blüthgen, and Julio Saez-Rodriguez. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature Communications*, 9(1):20, 2018.

Pau Badia-i-Mompel, Jesús Vélez Santiago, Jana Braunger, Celina Geiss, Daniel Dimitrov, Sophia Müller-Dott, Petr Taus, Aurelien Dugourd, Christian H. Holland, Ricardo O. Ramirez Flores, and Julio Saez-Rodriguez. decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*, 2(1):vbac016, 2022.

Sophia Müller-Dott, Eirini Tsirvouli, Miguel Vazquez, Ricardo O. Ramirez Flores, Pau Badia-i-Mompel, Robin Fallegger, Dénes Türei, Astrid Lægreid, and Julio Saez-Rodriguez. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Research*, 51(20):10934–10949, 2023.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1):101, 2025.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8): 1657–1661, 2025.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44(6):1050–1059, 2026.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Eduard Helly. Über Mengen konvexer Körper mit gemeinschaftlichen Punkten. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 32:175–176, 1923.

Willem Baarda. A testing procedure for use in geodetic networks. Publications on Geodesy, New Series, Vol. 2, No. 5, Netherlands Geodetic Commission, Delft, 1968.

Mario Milanese and Gustavo Belforte. Estimation theory and uncertainty intervals evaluation in presence of unknown but bounded errors: Linear families of models and estimators. *IEEE Transactions on Automatic Control*, 27(2):408–414, 1982.

J. Edward Jackson and Govind S. Mudholkar. Control procedures for residuals associated with principal component analysis. *Technometrics*, 21(3):341–349, 1979.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The Molecular Signatures Database hallmark gene set collection. *Cell Systems*, 1(6): 417–425, 2015.

Luz Garcia-Alonso, Christian H. Holland, Mahmoud M. Ibrahim, Denes Turei, and Julio Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, 29(8):1363–1375, 2019.

Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, 2017.

Martin Jinye Zhang, Kangcheng Hou, Kushal K. Dey, Saori Sakaue, Karthik A. Jagadeesh, Kathryn Weinand, Aris Taychameekiatchai, Poorvi Rao, Angela Oliveira Pisco, James Zou, Bruce Wang, Michael Gandal, Soumya Raychaudhuri, Bogdan Pasaniuc, and Alkes L. Price. Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nature Genetics*, 54(10):1572–1580, 2022.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, 2024.

Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. CellPLM: Pre-training of cell language model beyond single cells. In *International Conference on Learning Representations*, pp. 5649–5673, 2024.

Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, Anna Zhong, Daphne Raskin, Insu Han, Antonio Henrique de Oliveira Fonseca, Josue Ortega Caro, Amin Karbasi, Rahul Madhav Dhodapkar, and David van Dijk. Cell2Sentence: Teaching large language models the language of biology. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27299–27325, 2024.

Efthymia Papalexi, Eleni P. Mimitou, Andrew W. Butler, Samantha Foster, Bernadette Bracken, William M. Mauck, Hans-Hermann Wessels, Yuhan Hao, Bertrand Z. Yeung, Peter Smibert, and Rahul Satija. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature Genetics*, 53(3):322–331, 2021.

Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scPerturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4): 1100–1126, 2025.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, 2022.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Ziyi Liang, Matteo Sesia, and Wenguang Sun. Integrative conformal p-values for out-of-distribution testing with labelled outliers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):671–693, 2024.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, 120:108101, 2021.

Shai Feldman, Stephen Bates, and Yaniv Romano. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023.

Matteo Gasparin and Aaditya Ramdas. Merging uncertainty sets via majority vote. *arXiv preprint arXiv:2401.09379*, 2024.

Enrique Garcia-Ceja. Multi-view conformal learning for heterogeneous sensor fusion. *arXiv preprint arXiv:2402.12307*, 2024.

Edward Y. Chow and Alan S. Willsky. Analytical redundancy and the design of robust failure detection systems. *IEEE Transactions on Automatic Control*, 29(7):603–614, 1984.

Bradford W. Parkinson and Penina Axelrad. Autonomous GPS integrity monitoring using the pseudorange residual. *Navigation*, 35(2):255–274, 1988.

Juan Blanch, Todd Walter, and Per Enge. RAIM with optimal integrity and continuity allocations under multiple failures. *IEEE Transactions on Aerospace and Electronic Systems*, 46(3):1235–1247, 2010.

Fred C. Schweppe. Recursive state estimation: Unknown but bounded errors and system inputs. *IEEE Transactions on Automatic Control*, 13(1):22–28, 1968.

Mario Milanese and Antonio Vicino. Optimal estimation theory for dynamic systems with set membership uncertainty: An overview. *Automatica*, 27(6):997–1009, 1991.

Luc Jaulin. Robust set-membership state estimation; application to underwater robotics. *Automatica*, 45(1):202–206, 2009.

John Gleeson and Jennifer Ryan. Identifying minimally infeasible subsystems of inequalities. *ORSA Journal on Computing*, 2(1):61–63, 1990.

John W. Chinneck. *Feasibility and Infeasibility in Optimization: Algorithms and Computational Methods*, volume 118 of *International Series in Operations Research & Management Science*. Springer, 2008.

E. W. Cheney. *Introduction to Approximation Theory*. McGraw-Hill, New York, 1966.

Meir Katchalski and Andy Liu. A problem of geometry in $\mathbb{R}^n$. *Proceedings of the American Mathematical Society*, 75(2):284–288, 1979.

Gil Kalai. Intersection patterns of convex sets. *Israel Journal of Mathematics*, 48(2–3):161–174, 1984.

Imre Bárány. A generalization of Carathéodory's theorem. *Discrete Mathematics*, 40(2–3):141–152, 1982.

Nina Amenta. Helly-type theorems and generalized linear programming. *Discrete & Computational Geometry*, 12(3):241–261, 1994.

Jiří Matoušek, Micha Sharir, and Emo Welzl. A subexponential bound for linear programming. *Algorithmica*, 16(4–5):498–516, 1996.

Giuseppe C. Calafiore and Marco C. Campi. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51(5):742–753, 2006.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. In *Proceedings of the Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 582–609, 2020.

Nina Amenta, Jesús A. De Loera, and Pablo Soberón. Helly's theorem: New variations and applications. In *Algebraic and Geometric Methods in Discrete Mathematics*, volume 685 of *Contemporary Mathematics*, pp. 55–95. American Mathematical Society, 2017.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20827–20840, 2022.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Joel A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018.

Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M. Lanata, Rachel E. Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A. Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, 2018.

Malte D. Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F. Mueller, Daniel C. Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022.

Paul Datlinger, André F. Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C. Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–301, 2017.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.

Anna Neufeld, Ameer Dharamshi, Lucy L. Gao, and Daniela Witten. Data thinning for convolution-closed distributions. *Journal of Machine Learning Research*, 25(57):1–35, 2024.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

# A PROOFS

Throughout, $\tilde{c}_j = (r_j(y) - b_j)/\sigma_j(y)$ and $\tilde{a}_j = a_j/\sigma_j(y)$, so that $s_j(z;y) = |\tilde{c}_j - \tilde{a}_j^\top z|$. $H = \mathrm{conv}\{\pm\tilde{a}_1, \dots, \pm\tilde{a}_K\} = \{\sum_j w_j\tilde{a}_j : \|w\|_1 \le 1\}$ and $Q = \{\delta : \max_j |\tilde{a}_j^\top\delta| \le 1\}$ is its polar; $Q_{-j}$ is defined in the same way without program $j$.

**Lemma 1** (Gauge and polar)**.** *(i) The gauge of $H$ is $\kappa$: $u \in tH$ if and only if $u = \sum_j w_j\tilde{a}_j$ with $\|w\|_1 \le t$, so $\inf\{t > 0 : u \in tH\} = \kappa(u)$. (ii) $\max_{\delta\in Q} u^\top\delta = \kappa(u)$ for every $u$, with both sides infinite when $u \notin \mathrm{span}\{\tilde{a}_j\}$. (iii) $\max_{\delta\in Q}\|\delta\|_2 = \max_{\|u\|_2=1}\kappa(u) = 1/r(H)$, and $\kappa(u) \le \|u\|_2/r(H)$.*

*Proof.* (i) is the definition of $H$ as the image of the $\ell_1$ ball. (ii) is LP duality for $\max\{u^\top\delta : -\mathbf{1} \le \tilde{A}\delta \le \mathbf{1}\}$, whose dual is $\min\{\|w\|_1 : \tilde{A}^\top w = u\}$; if $u \notin \mathrm{span}\{\tilde{a}_j\}$, the primal is unbounded along the component of $u$ orthogonal to the span. (iii) The support function of $H$ is $h_H(v) = \max_j |\tilde{a}_j^\top v|$, so the Euclidean ball of radius $\rho$ lies in $H$ iff $h_H(v) \ge \rho$ for all unit $v$; the largest such $\rho$ is $r(H)$, which gives $\kappa(u) \le \|u\|_2/r(H)$ with equality for some unit $u$, and by (ii) $\max_{\delta\in Q}\|\delta\|_2 = \max_{\|u\|=1}\max_{\delta\in Q} u^\top\delta = \max_{\|u\|=1}\kappa(u)$. □

**Proof of Theorem 1.** The view model is fitted on training units only, so $T$ and $M$ are fixed functions of a unit. Let $V_1, \dots, V_m$ be the calibration values and $V_{m+1}$ the test value of either statistic, $k = \lceil (m+1)(1-\alpha)\rceil$ and $\hat{q} = V_{(k)}$, the $k$-th smallest calibration value ($\hat{q} = +\infty$ if $k > m$, in which case there is nothing to prove). If $V_{m+1} > \hat{q}$, then at least $k$ of the $m+1$ values are strictly smaller than $V_{m+1}$. For each $i \le m+1$ let $N_i$ be the number of values strictly smaller than $V_i$; at most $m+1-k$ indices have $N_i \ge k$, ties included, and by exchangeability $\Pr[N_{m+1} \ge k] = \mathbb{E}\,|\{i : N_i \ge k\}|/(m+1) \le (m+1-k)/(m+1)$. Hence $\Pr[V_{m+1} \le \hat{q}] \ge k/(m+1) \ge 1-\alpha$ without randomized tie-breaking (Vovk et al., 2005; Lei et al., 2018). For $V = M$, $M(y, z^\star) \le \hat{q}$ iff $s_j(z^\star; y) \le \hat{q}$ for all $j$ iff $z^\star \in F_{\hat{q}}(y)$. For $V = T$, the Chebyshev LP has a finite optimum that is attained, so $F_q(y) \ne \emptyset$ iff $T(y) \le q$; hence $\Pr[F_{\hat{q}}(Y) = \emptyset] = \Pr[T_{m+1} > \hat{q}] \le \alpha$. Finally $T(y) = \min_z \max_j s_j(z;y) \le \max_j s_j(z^\star;y) = M(y, z^\star)$ for every unit, so every order statistic of the $T$-values is at most the corresponding order statistic of the $M$-values. □

**Proof of Theorem 2.** (a) *Upper bound.* For $z, z' \in F_q(y)$ and every $j$, $|\tilde{a}_j^\top(z - z')| \le |\tilde{c}_j - \tilde{a}_j^\top z'| + |\tilde{c}_j - \tilde{a}_j^\top z| \le 2q$, so $z - z' \in 2qQ$ and, by Lemma 1(ii), $u^\top(z - z') \le 2q\,\kappa(u)$. *Lower bound.* Let $z_c$ be a Chebyshev centre, $\max_j s_j(z_c; y) = T(y)$, and $\eta = q - T(y) \ge 0$. For $\delta \in \eta Q$, $s_j(z_c + \delta; y) \le s_j(z_c; y) + |\tilde{a}_j^\top\delta| \le T(y) + \eta = q$, so $z_c + \eta Q \subseteq F_q(y)$ and $W_u(F_q) \ge W_u(\eta Q) = 2\eta\max_{\delta\in Q} u^\top\delta = 2\eta\,\kappa(u)$, using that $Q$ is symmetric and Lemma 1(ii). If $\eta = 0$ and $\kappa(u) = \infty$, the lower bound is 0 by the convention $0 \cdot \infty = 0$ and holds trivially; if $\eta > 0$ and $\kappa(u) = \infty$, then $Q$, hence $F_q$, is unbounded along $u$ and both sides are infinite. *Diameter.* $\mathrm{diam}\, F_q = \max_{\|u\|_2=1} W_u(F_q)$, and $\max_{\|u\|=1}\kappa(u) = 1/r(H)$ by Lemma 1(iii).

(b) Let $F_q^{(-j)}$ be the feasible set without program $j$ and $[l_j, u_j]$ the range of $\tilde{a}_j^\top z$ over it (two LPs; $\pm\infty$ if unbounded). After the shift, program $j$ admits exactly the $z$ with $\tilde{a}_j^\top z \in [\tilde{c}_j + \Delta - q,\ \tilde{c}_j + \Delta + q]$, so the shifted feasible set is empty iff this interval misses $[l_j, u_j]$, i.e. iff $\Delta > \Delta_j^+ := u_j + q - \tilde{c}_j$ or $-\Delta > \Delta_j^- := \tilde{c}_j - l_j + q$. Before the shift $F_q(y) \ne \emptyset$ because $q \ge T(y)$, so the two intervals intersect and $\Delta_j^\pm \ge 0$. Adding the two thresholds gives $\Delta_j^+ + \Delta_j^- = 2q + W_{\tilde{a}_j}(F_q^{(-j)})$. Part (a) applied to the programs other than $j$, for which the synthesis cost of $\tilde{a}_j$ is $\kappa_{-j}$ and the Chebyshev statistic is $T_{-j}(y) \le T(y) \le q$, gives $2(q - T_{-j}(y))\kappa_{-j} \le W_{\tilde{a}_j}(F_q^{(-j)}) \le 2q\kappa_{-j}$, which is the two-sided bound. For the maximum, take $z^\circ \in F_q(y)$; then $l_j \le \tilde{a}_j^\top z^\circ \le u_j$ and $|\tilde{c}_j - \tilde{a}_j^\top z^\circ| \le q$, so $\Delta_j^+ = (u_j - \tilde{a}_j^\top z^\circ) + (\tilde{a}_j^\top z^\circ - \tilde{c}_j) + q \le 2q\kappa_{-j} + 2q$, and likewise for $\Delta_j^-$.

(c) Let $d \ge 2$ and $0 < \epsilon \le 1$; take two copies of each standard basis vector $\iota_1, \dots, \iota_{d-1}$ of $\mathbb{R}^d$ and two copies of $(\epsilon/2)\iota_d$ as normalized directions, so that $\max_j\|\tilde{a}_j\|_2 = 1$. Every program has an exact duplicate, so $\kappa_{-j} \le 1$, while $h_H(\iota_d) = \epsilon/2$ gives $r(H) \le \epsilon/2 < \epsilon\max_j\|\tilde{a}_j\|_2$; since $\kappa$ is invariant and $r(H)$ scales linearly under a common rescaling of the programs, the construction holds at every scale. □

**Proof of Proposition 1.** Let $N(x) = \min_\delta \|x - \tilde{A}\delta\|_\infty$ for $x \in \mathbb{R}^K$; the minimum is attained because it is a finite LP. Substituting $\lambda\delta$ gives $N(\lambda x) = |\lambda| N(x)$, and $N(x + x') \le \|x + x' - \tilde{A}(\delta + \delta')\|_\infty \le N(x) + N(x')$ for minimizers $\delta, \delta'$, so $N$ is a seminorm; $N(x) = 0$ iff $x = \tilde{A}\delta$ for some $\delta$, that is iff the change acts through the state, and $N(x + \tilde{A}z) = N(x)$ for every $z$, so $T(y) = N(s + \varepsilon)$. For a shift $s = \Delta\iota_j$ of one program ($\iota_j$ the $j$-th basis vector of $\mathbb{R}^K$), every $\delta$ with $\tilde{a}_j^\top \delta = t$ has $\max_{k \ne j} |\tilde{a}_k^\top \delta| \ge |t|/\kappa_{-j}$ by Lemma 1(ii) for the programs other than $j$, with equality for some $\delta$; hence $N(s) = \min_t \max\{|\Delta - t|, |t|/\kappa_{-j}\} = |\Delta|/(1 + \kappa_{-j})$, attained at $t = \Delta\kappa_{-j}/(1 + \kappa_{-j})$. For a seminorm, $N(s) - N(\varepsilon) \le N(s + \varepsilon) \le N(s) + N(\varepsilon)$ (the left inequality applies the triangle inequality to $s = (s + \varepsilon) + (-\varepsilon)$). Fix the calibration units, hence $\hat{q}$ and $\rho = N(s)/\hat{q}$; since $\varepsilon$ is independent of them, probabilities below are over $\varepsilon$. If $N(\varepsilon) < N(s) - \hat{q}$, then $T(y) \ge N(s) - N(\varepsilon) > \hat{q}$; if $T(y) > \hat{q}$, then $N(s) + N(\varepsilon) > \hat{q}$. Hence $\{N(\varepsilon) < (\rho - 1)\hat{q}\} \subseteq \{T(y) > \hat{q}\} \subseteq \{N(\varepsilon) > (1 - \rho)\hat{q}\}$, which gives the two conditional bounds. For the marginal statement, a known unit has readings $\tilde{A}z' + \varepsilon'$ and statistic $N(\varepsilon')$, so $N(\varepsilon)$ has the law of $T$ for a known unit that is independent of the calibration units, and $\Pr[N(\varepsilon) \le \hat{q}] \ge 1 - \alpha$ by Theorem 1. On the event $\{N(\varepsilon) \le \hat{q}\} \cap \{N(s) > 2\hat{q}\}$, $T(y) \ge N(s) - N(\varepsilon) > 2\hat{q} - \hat{q} = \hat{q}$, so $\Pr[\text{flag}] \ge \Pr[N(\varepsilon) \le \hat{q}] - \Pr[N(s) \le 2\hat{q}] \ge 1 - \alpha - \Pr[\hat{q} \ge N(s)/2]$. □

**Proof of Theorem 3.** (a) Write the Chebyshev LP as $\min_{z,t} t$ subject to $\tilde{c}_j - \tilde{a}_j^\top z \le t$ and $-\tilde{c}_j + \tilde{a}_j^\top z \le t$ for all $j$. With multipliers $\mu_j^+, \mu_j^- \ge 0$ its dual is

$$\max_{\mu^\pm \ge 0} \sum_j (\mu_j^+ - \mu_j^-)\tilde{c}_j \quad \text{s.t.} \quad \sum_j (\mu_j^+ - \mu_j^-)\tilde{a}_j = 0, \qquad \sum_j (\mu_j^+ + \mu_j^-) = 1,$$

which has $d + 1$ equality constraints, so a basic optimal solution has at most $d + 1$ non-zero entries; let $W$ be the programs carrying them. Strong duality gives dual value $T(y)$. The same multipliers are feasible for the dual of the LP restricted to $W$, so by weak duality the restricted optimum satisfies $T_W(y) \ge T(y)$; removing constraints also gives $T_W(y) \le T(y)$. Hence $T_W(y) = T(y) > q$ and $\bigcap_{k \in W} C_k(q; y) = \emptyset$, which is Helly's theorem for this family (Helly, 1923; Amenta et al., 2017). The deletion filter keeps a program only if removing it makes the remaining set feasible; since removing programs can only restore feasibility, every program it keeps is necessary at the end, so its output is an irreducible infeasible subsystem (Chinneck, 2008).

(b) Let $T_{-S}(y) = \min_z \max_{k \notin S} s_k(z; y) \le T(y)$, and let $Y'$ be a known unit that, like the test unit $Y$, is independent of the calibration data. By Assumption 2, $\tilde{c}_k = \tilde{a}_k^\top z^\star + \varepsilon_k$ for $k \notin S$, where $z^\star$ is the state of $Y$; substituting $\delta = z - z^\star$ gives $T_{-S}(Y) = \min_\delta \max_{k \notin S} |\varepsilon_k - \tilde{a}_k^\top \delta|$, a function of $(\varepsilon_k, \sigma_k(Y))_{k \notin S}$ alone. Hence $T_{-S}(Y)$ and $T_{-S}(Y')$ have the same law, and $\Pr[T_{-S}(Y) \le \hat{q}] = \Pr[T_{-S}(Y') \le \hat{q}] \ge \Pr[T(Y') \le \hat{q}] \ge 1 - \alpha$ by Theorem 1. On this event $\bigcap_{k \notin S} C_k(\hat{q}; Y) \ne \emptyset$, so no subset of $[K] \setminus S$ is infeasible, and every infeasible subsystem, in particular every witness, meets $S$. A flag with a witness disjoint from $S$ therefore implies $T_{-S}(Y) > \hat{q}$, an event of probability at most $\alpha$; dividing by the probability $\beta$ of a flag, and noting that a conditional probability is at most one, gives $\min\{1, \alpha/\beta\}$. The argument for localization calibration is identical with $M_{-S}(Y, z^\star) = \max_{k \notin S} s_k(z^\star; Y) = \max_{k \notin S} |\varepsilon_k|$. □

**Proof of Corollary 1.** Under Assumption 3, $\tilde{c}_j = \tilde{a}_j^\top z_{\text{new}} + \varepsilon_j$, so $T(y) = \min_z \max_j |\varepsilon_j - \tilde{a}_j^\top (z - z_{\text{new}})| = \min_\delta \max_j |\varepsilon_j - \tilde{a}_j^\top \delta|$, a function of $(\varepsilon, \sigma(y))$ alone. A known unit $Y'$ independent of the calibration units has the same representation with its own state, so $T(Y)$ and $T(Y')$ have the same law jointly with the calibration values, and $\Pr[T(Y) > \hat{q}] = \Pr[T(Y') > \hat{q}] \le \alpha$ by Theorem 1 under Assumption 1. □

**Remarks on the assumptions.** Table 3 maps each result to its assumptions, to the experiments in which they should hold and are tested, and to those in which they are stressed.

Table 3: Results, their assumptions, the experiments in which the assumptions should hold (and are tested), and those in which they are stressed.

| Result | Assumptions | Tested where they should hold | Stressed where they may fail |
|---|---|---|---|
| Theorem 1 | exchangeability (A1) | random splits: coverage and false alarms (Table 4, Figure 6b) | donor rotations and partitions (Figure 4a, Table 21), cell types (Table 14) |
| Theorem 2 | none (any unit, $q > 0$, $q \geq T(y)$) | every sampled unit (Figure 2a,b, Table 25) | expression-level shifts, not covered by (b) (Figure 2c, Table 9) |
| Proposition 1 | fixed shift, noise as for known units (A1 for the marginal bound) | prospective test (Figure 3c, Tables 26 and 27) | responses that vary across the groups of one perturbation |
| Theorem 3(a) | none | every flag (witness size $d + 1$) | |
| Theorem 3(b) | A1 and sparse shift (A2) | planted view-level shifts of one to five programs (Tables 19 and 20) | expression-level shifts (Table 9); broad real responses (Table 2) |
| Corollary 1 | A1 and state-respecting shift (A3) | held-out stimulated cell type (Figure 4b, Tables 11 and 29) | new protocols, batches, donors and cell types (Figure 4c, Tables 15 and 18) |

Assumption 1 is the only one needed for error control and holds by construction for random splits; Proposition 1, Theorem 3(b) and Corollary 1 compare a unit with a known unit that is independent of the calibration units, which needs its stronger form, independent and identically distributed units, as random splits provide; with donor splits (Kang), calibration and test units come from different donors, so exchangeability is not guaranteed, and we measure the resulting error rates over many donor splits instead (Section 5). Groups are formed by partitioning cells into disjoint groups of $n$ within strata, identically for calibration and test, so the group is the exchangeable unit. Assumption 2 is used only for the soundness statement; in the planted experiments it holds by construction for view-level shifts of one to five programs, whereas an expression-level shift moves the readings of every program that shares a gene with the shifted one, a median of 254 to 324 programs, so $S$ is not sparse and the statement is uninformative for these shifts; the same holds for the broad real responses of Table 2. Assumption 3 is stated with the fitted coefficients; if instead the population view model holds, the estimation term $\delta a_j^\top z_{\text{new}} + \delta b_j$ enters the residual, and Section 5 measures its size at held-out states. It is the null hypothesis of the compositional-novelty experiment.

## B Implementation details

**Data.**

| Dataset | Cells | Genes | Structure | Used in |
|---|---|---|---|---|
| PBMC3K | 2629 | 13714 | healthy donor PBMCs (10x) | Q1, Q2 |
| Kang IFN-$\beta$ PBMC | 24439 | 15701 | 8 donors, 8 cell types, ctrl/stim | Q1 to Q4 |
| Pancreas atlas | 16300 | 18205 | 9 batches from 6 technologies, 14 cell types | Q1, Q2, Q4 |
| Datlinger CRISPR (Jurkat) | 5875 | 26623 | 32 targets, TCR stim/unstim | Q3 |
| Dixit Perturb-seq (K562) | 19171 | 18913 | 10 TF knockouts + controls | Q3, App. D |
| Norman Perturb-seq (K562) | 31530 | 20306 | subsample: 5960 controls, up to 250 cells per single activation | Q3, App. D |

Raw counts are used throughout. Cells with fewer than 200 detected genes, genes detected in fewer than three cells and the $0.5\%$ of cells with the largest counts are removed; for PBMC3K, cells with more than $5\%$ mitochondrial counts are also removed. For Kang we use the eight-donor batch of GEO GSE96583 with the authors' singlet calls and cell-type labels; for the pancreas atlas the counts shipped with the scvi-tools example data; for the CRISPR screens the harmonized files of scPerturb (Peidli et al., 2024).

**Genes, programs and state.** Highly variable genes (2,000, Seurat v3 flavour) are selected on training cells only. The gene universe is their union with all program genes detected in at least $1\%$ of training cells. Expression is library-normalized to $10^4$ counts, log-transformed and standardized with training means and standard deviations (clipped at $\pm 10$). Programs are the MSigDB hallmark gene sets (v2023.2) and CollecTRI regulons from OmniPath, with regulon signs from the consensus mode of regulation and complexes excluded. Programs are restricted to the gene universe, kept if they have at least 10 genes (larger ones are randomly subsampled to 200), and near-duplicates with Jaccard index above $0.8$ are removed, keeping the larger program. The state is given by the top $d = 8$ principal components of standardized highly variable genes, fitted on training units. Readings are signed means of standardized expression over program genes.

**Foundation-model state.** For the encoder ablation we also use the published Geneformer model with ten million parameters and its gene dictionaries (Theodoris et al., 2023), without fine-tuning. Each cell is rank-value encoded as in its pretraining: counts of the genes in the model's vocabulary (Ensembl identifiers from the GEO gene table) are divided by the cell's total count, multiplied by $10^4$ and divided by the gene's median over the pretraining corpus, and the top 2,048 expressed genes are ordered by this value. The cell embedding is the mean over genes of the penultimate layer, and the state is the top eight principal components of the embeddings of training cells; stimulated cells are never used to fit anything.

**View model and noise scales.** The coefficients $(a_j, b_j)$ are ordinary least squares of single-cell readings on training states. For groups, the residual variance of program $j$ at group size $n$ is modelled as $\sigma_j(n)^2 = \gamma_{0j} + \gamma_{1j}/n$ and fitted by non-negative least squares on squared residuals of training groups of sizes $n \in \{1, 2, 5, 10, 20\}$; $\gamma_{0j}$ captures between-group biological variation and $\gamma_{1j}$ within-group sampling noise. To stop a few heavy-tailed programs from dominating the maximum in $T$, we rescale each $\sigma_j$ so that the $95\%$ quantile of its normalized training residuals is the same for all programs. This rescaling uses training units only, so it does not affect validity. For the thinning experiments of Q1, single-cell noise scales are $\sigma_j(y)^2 = \gamma_{0j} + \gamma_{1j}\, 10^3/N(y) + \gamma_{2j}/(N_j(y) + 1)$, where $N$ is the cell's total count and $N_j$ the count of program $j$'s genes; the backbone's coordinate views use the same model without $N_j$, fitted on five-fold cross-fitted residuals of the ridge encoder.

**Calibration and units.** Groups are formed by randomly partitioning the cells of each stratum (cell type and donor for Kang, cell type and technology for the pancreas atlas, replicate for Datlinger) into disjoint groups of $n$ cells, identically for training, calibration and test splits. The group is the exchangeable unit. The four Kang donor rotations are listed below as training / calibration / test donors, so that every donor is tested once:

| Rotation | Training donors | Calibration donors | Test donors |
|---|---|---|---|
| 1 | 1015, 1256, 1016, 101 | 1244, 107 | 1488, 1039 |
| 2 | 1488, 1244, 1016, 107 | 1256, 1039 | 1015, 101 |
| 3 | 1015, 1488, 1039, 107 | 1016, 101 | 1256, 1244 |
| 4 | 1256, 1244, 1039, 101 | 1488, 1015 | 1016, 107 |

For Datlinger, known units are unstimulated cells of all guides, and novel units are stimulated cells. Stimulated cells with guides against the TCR-proximal kinases and adaptor LAT, ZAP70 and LCK are excluded before any analysis, because these knockouts abolish the stimulation response by design. Splits are stratified by replicate ($50/25/25\%$). All statistics are calibrated at $\alpha = 0.05$, or $\alpha = 0.1$ for the localization experiments, with the conformal quantile of Section 3.

**Linear programs.** All LPs are solved with HiGHS through SciPy. The Chebyshev LP uses the dual simplex method, so that the reported dual solution is basic and its support has at most $d + 1$ programs. Witnesses are then made irreducible by the deletion filter of Algorithm 1 (Chinneck, 2008); in all our experiments the dual support was already irreducible. Greedy peeling, used only as a diagnostic, repeatedly removes the program with the largest dual weight in the current Chebyshev solution until the remaining programs are feasible; the removal order ranks candidate culprits. Widths $W_u(F_q)$ and boxes (widths along the coordinate axes) use two LPs per direction; the exact thresholds $\Delta_j^{\pm}$ use two LPs over the feasible set without program $j$, and $T_{-j}$ one Chebyshev LP. The synthesis costs $\kappa(u)$ and $\kappa_{-j}$ are LPs in the split variables $w = w^+ - w^-$. Directions for the width tests are the $d$ right singular vectors of $\tilde{A}$, the $d$ coordinate axes and 24 random unit vectors, fixed across units. The worst-covered cost $\max_{\|u\|=1} \kappa(u) = 1/r(H)$ is a non-convex maximization; the largest $\kappa(u)$ over sampled directions is a lower bound on it, so the gap between localization and checking that we report is, if anything, understated.

**Baselines and controls.** The kNN detector scores a unit by its mean Euclidean distance to its ten nearest training units in state space. Mahalanobis distance uses the mean and covariance of training units, and isolation forest uses scikit-learn defaults. The largest program $|z|$-score standardizes each reading by the mean and standard deviation of training units. In Q1, the ridge encoder regresses the full-depth state on standardized thinned highly variable genes, with the penalty chosen by cross-validation over $[1, 10^5]$. Cross-fitted stacking regresses the state on out-of-fold encoder outputs and

program readings. Message passing augments the input with $\hat{x}_g = \sum_j |\Omega_{jg}| r_j / \sum_j |\Omega_{jg}|$, one round of gene-to-program-to-gene averaging with the unsigned weights $|\Omega_{jg}|$ of the membership matrix $\Omega$ (entries $\omega_{jg}/|\mathcal{M}_j|$ for $g \in \mathcal{M}_j$, zero otherwise), which ignores the mode of regulation. The MLP has two hidden layers of 256 units, GELU activations and dropout 0.1, and is trained with AdamW for 40 epochs. Random programs are uniformly drawn gene sets of the same sizes. Shuffled programs permute gene identities of all program genes, which preserves sizes, overlaps and gene degrees. Corrupted programs replace a fraction of each program's genes by random genes outside the program while keeping the program's name.

**Pre-specified reference response programs.** For Kang, reference programs are those whose name contains INTERFERON, IFN, ISG or ANTIVIRAL, together with the STAT1, STAT2, IRF1, IRF2, IRF3, IRF7, IRF8 and IRF9 regulons. For Datlinger, they are those whose name contains TNFA, NFKB, IL2, TCR, T CELL RECEPTOR, ZAP-70, CD28, INFLAMMATORY or ALLOGRAFT, together with the NFAT, NF-$\kappa$B/REL, AP-1 (JUN, FOS), EGR, NR4A, IRF4, BATF and ATF3 regulons. These lists were fixed before the final experiments; they are a proxy for the unknown set of shifted programs, not its ground truth.

**Operating envelope.** For every real perturbation, the mean shift $s$ of the readings is the difference between the mean reading of its perturbed cells and that of its reference cells, divided by the noise scales $\sigma_j(10)$ of groups of ten cells; $\hat{\rho} = N(s)/\hat{q}$ uses the Chebyshev LP and the quantile $\hat{q}$ of $T$ at $n = 10$, and the envelope of Proposition 1 uses the empirical distribution of $T$ on the calibration units of the same dataset. In the prospective test, $s$ is estimated on a random half of the perturbed and reference cells only, and detection is measured on groups formed from the other half (Appendix C). The Kang perturbations are the stimulated cells of one cell type in the test donors of one rotation, the Datlinger perturbations the stimulated cells of one guide target (TCR-proximal knockouts excluded, as above), and the screens' perturbations the cells of one target.

**Compute.** All experiments ran on a laptop CPU (Intel i9, 8 cores, 16 GB RAM); no GPU was needed. The complete experimental suite takes about eight hours of wall-clock time on this machine (the linear programs run in parallel on eight worker processes), and runtimes per unit are reported in Appendix C.

## C Additional results

**Calibration across levels.** Table 4 reports, for every nominal level, the empirical coverage of the full-depth target state by the jointly calibrated backbone-plus-program set (localization calibration) and the empirical false-alarm rate of the consistency test on held-out known single cells (consistency calibration). Both match their targets, as Theorem 1 guarantees.

Table 4: Empirical coverage and false-alarm rate versus the nominal level (thinned cells, 25% of counts; mean over splits).

| Dataset | Quantity | $\alpha = 0.01$ | $\alpha = 0.025$ | $\alpha = 0.05$ | $\alpha = 0.1$ | $\alpha = 0.2$ |
|---|---|---|---|---|---|---|
| PBMC3K | coverage (target $1-\alpha$) | 0.993 | 0.971 | 0.948 | 0.900 | 0.794 |
| | false-alarm rate (target $\alpha$) | 0.008 | 0.027 | 0.049 | 0.102 | 0.217 |
| Kang (ctrl) | coverage (target $1-\alpha$) | 0.988 | 0.976 | 0.953 | 0.904 | 0.796 |
| | false-alarm rate (target $\alpha$) | 0.009 | 0.024 | 0.050 | 0.097 | 0.198 |
| Pancreas | coverage (target $1-\alpha$) | 0.991 | 0.973 | 0.949 | 0.902 | 0.799 |
| | false-alarm rate (target $\alpha$) | 0.012 | 0.027 | 0.051 | 0.102 | 0.198 |

**Thinning level.** Table 5 repeats Q1 at three thinning levels (split 0). The conclusions do not depend on the depth: programs never reduce the error of a ridge encoder by more than a few percent, and they always inflate the calibrated radius.

Table 5: Q1 at different thinning levels (fraction of counts kept), split 0. Errors and radii relative to the backbone.

| Dataset, kept | stacking | message passing | fusion | radius: backbone+programs | radius: programs only |
|---|---|---|---|---|---|
| Kang (ctrl), 0.1 | 0.99 | 0.99 | 1.22 | 2.8 | 25 |
| Kang (ctrl), 0.25 | 1.00 | 0.99 | 1.23 | 2.4 | 24 |
| Kang (ctrl), 0.5 | 1.00 | 0.99 | 1.25 | 2.3 | 31 |
| Pancreas, 0.1 | 0.96 | 0.98 | 1.56 | 1.7 | 13 |
| Pancreas, 0.25 | 0.98 | 0.99 | 1.65 | 2.0 | 21 |
| Pancreas, 0.5 | 0.98 | 0.99 | 1.78 | 2.1 | 31 |
| PBMC3K, 0.1 | 0.98 | 0.98 | 1.25 | 3.0 | 23 |
| PBMC3K, 0.25 | 1.03 | 0.98 | 1.22 | 2.8 | 25 |
| PBMC3K, 0.5 | 1.06 | 0.99 | 1.26 | 2.7 | 30 |

**All detectors.** Tables 6 and 7 list every detector.

Table 6: All detectors on Kang IFN-$\beta$: detection rate (TPR) and false-alarm rate (FPR) at $\alpha = 0.05$, and AUROC, by cells per unit $n$; mean over four donor rotations $\times$ two grouping seeds; the largest standard error of a TPR or FPR entry is 0.113.

| | $n = 1$ | | | $n = 2$ | | | $n = 5$ | | | $n = 10$ | | | $n = 20$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | TPR | FPR | AUC | TPR | FPR | AUC | TPR | FPR | AUC | TPR | FPR | AUC | TPR | FPR | AUC |
| HellyCell $T$ | 0.13 | 0.05 | 0.69 | 0.37 | 0.05 | 0.83 | 0.80 | 0.05 | 0.96 | 0.97 | 0.05 | 1.00 | 1.00 | 0.05 | 1.00 |
| anchored $A$ | 0.69 | 0.05 | 0.91 | 0.92 | 0.05 | 0.98 | 1.00 | 0.06 | 1.00 | 1.00 | 0.05 | 1.00 | 1.00 | 0.07 | 1.00 |
| LS global $G$ | 0.26 | 0.05 | 0.76 | 0.50 | 0.05 | 0.87 | 0.86 | 0.05 | 0.98 | 0.99 | 0.05 | 1.00 | 1.00 | 0.06 | 1.00 |
| LS $w$-test $B$ | 0.26 | 0.05 | 0.75 | 0.59 | 0.06 | 0.90 | 0.95 | 0.05 | 0.99 | 1.00 | 0.05 | 1.00 | 1.00 | 0.06 | 1.00 |
| kNN distance | 0.66 | 0.05 | 0.93 | 0.80 | 0.05 | 0.96 | 0.94 | 0.06 | 0.98 | 0.98 | 0.07 | 0.99 | 0.98 | 0.07 | 0.98 |
| Mahalanobis | 0.33 | 0.05 | 0.81 | 0.44 | 0.05 | 0.86 | 0.58 | 0.07 | 0.92 | 0.70 | 0.07 | 0.94 | 0.75 | 0.06 | 0.96 |
| Isolation forest | 0.38 | 0.06 | 0.84 | 0.46 | 0.06 | 0.89 | 0.57 | 0.06 | 0.93 | 0.65 | 0.07 | 0.95 | 0.67 | 0.06 | 0.96 |
| max program $\lvert z \rvert$ | 0.61 | 0.05 | 0.88 | 0.85 | 0.06 | 0.96 | 0.99 | 0.07 | 1.00 | 1.00 | 0.06 | 1.00 | 1.00 | 0.05 | 1.00 |

Table 7: All detectors on Datlinger TCR stimulation: detection rate (TPR) and false-alarm rate (FPR) at $\alpha = 0.05$, and AUROC, by cells per unit $n$; mean over three splits; the largest standard error of a TPR or FPR entry is 0.126.

| | $n = 1$ | | | $n = 2$ | | | $n = 5$ | | | $n = 10$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | TPR | FPR | AUC | TPR | FPR | AUC | TPR | FPR | AUC | TPR | FPR | AUC |
| HellyCell $T$ | 0.13 | 0.04 | 0.63 | 0.17 | 0.05 | 0.67 | 0.20 | 0.02 | 0.75 | 0.32 | 0.04 | 0.81 |
| anchored $A$ | 0.13 | 0.05 | 0.62 | 0.18 | 0.07 | 0.67 | 0.26 | 0.04 | 0.76 | 0.42 | 0.09 | 0.80 |
| LS global $G$ | 0.18 | 0.05 | 0.65 | 0.24 | 0.06 | 0.71 | 0.33 | 0.06 | 0.81 | 0.50 | 0.05 | 0.88 |
| LS $w$-test $B$ | 0.12 | 0.05 | 0.62 | 0.15 | 0.06 | 0.65 | 0.17 | 0.03 | 0.72 | 0.34 | 0.08 | 0.76 |
| kNN distance | 0.19 | 0.05 | 0.70 | 0.25 | 0.06 | 0.74 | 0.31 | 0.04 | 0.83 | 0.49 | 0.04 | 0.89 |
| Mahalanobis | 0.13 | 0.05 | 0.60 | 0.17 | 0.05 | 0.64 | 0.20 | 0.02 | 0.71 | 0.41 | 0.06 | 0.82 |
| Isolation forest | 0.14 | 0.05 | 0.61 | 0.16 | 0.04 | 0.66 | 0.25 | 0.04 | 0.75 | 0.48 | 0.03 | 0.84 |
| max program $\lvert z \rvert$ | 0.13 | 0.04 | 0.63 | 0.18 | 0.05 | 0.66 | 0.24 | 0.04 | 0.72 | 0.34 | 0.08 | 0.76 |

**Corrupted priors and culprit ranking.** Table 8 replaces a fraction of every program's genes by random genes. False alarms stay at $\alpha$ because calibration adapts to any fixed set of programs, while detection and explanation degrade gradually. The peel top-1 column shows that the first program removed by greedy peeling (the largest dual weight) is rarely a reference response program: peeling does not rank culprits for this broad response.

Table 8: Corrupted programs ($n = 10$; four rotations with one grouping seed each, so the uncorrupted row differs slightly from Table 6). Witness any: share of witnesses naming a reference response program; anchored precision: share of flagged programs that are reference programs; peel top-1: share of units whose first peeled program is a reference program (base: share of reference programs).

| Data | Genes replaced | FPR | TPR | witness any | anchored precision | peel top-1 | base |
|---|---|---|---|---|---|---|---|
| Kang | 0% | 0.04 | 0.97 | 0.96 | 0.71 | 0.20 | 0.025 |
| Kang | 20% | 0.03 | 0.97 | 0.98 | 0.69 | 0.30 | 0.025 |
| Kang | 50% | 0.05 | 0.92 | 0.97 | 0.68 | 0.32 | 0.025 |
| Kang | 80% | 0.04 | 0.92 | 0.65 | 0.31 | 0.21 | 0.025 |
| Datlinger | 0% | 0.04 | 0.32 | 0.72 | 0.43 | 0.19 | 0.060 |
| Datlinger | 20% | 0.07 | 0.36 | 0.61 | 0.32 | 0.11 | 0.060 |
| Datlinger | 50% | 0.02 | 0.26 | 0.53 | 0.12 | 0.10 | 0.060 |
| Datlinger | 80% | 0.03 | 0.17 | 0.45 | 0.06 | 0.03 | 0.060 |

**Threshold law, all configurations.** Table 9 lists detection at four normalized shifts, witness soundness and peeling for every configuration, including random and shuffled programs. The collapse holds for every program family, since the threshold involves only the geometry of the views; what biology changes is which programs are named and how large a gene-level change is needed. Detection is measured on units that pass the test before the shift, and $\kappa_{-j}$ differs slightly from its value for the thinned cells of Q1 (Table 1), which use the noise scales of thinned cells.

Table 9: Planted shifts on held-out known units that pass the test before the shift. Detection at normalized shift $x$; soundness: share of witnesses containing the shifted program (view) or a moved program (expression; every program sharing a gene with the shifted one moves, so this share is uninformative there); peel: share of detected shifts whose first peeled program is the shifted one.

| Data | $n$ | Programs | shift | median $\kappa_{-j}$ | $x = 0.2$ | $x = 0.3$ | $x = 0.5$ | $x = 1$ | soundness | peel |
|---|---|---|---|---|---|---|---|---|---|---|
| Kang | 1 | bio | expression | 0.79 | 0.007 | 0.04 | 0.66 | 1.00 | 1.000 | 0.96 |
| Kang | 1 | bio | view | 0.79 | 0.001 | 0.04 | 0.75 | 1.00 | 1.000 | 0.97 |
| Kang | 5 | bio | expression | 0.80 | 0.006 | 0.02 | 0.68 | 1.00 | 1.000 | 0.98 |
| Kang | 5 | bio | view | 0.80 | 0.003 | 0.03 | 0.75 | 1.00 | 1.000 | 0.99 |
| Kang | 10 | bio | expression | 0.83 | 0.001 | 0.02 | 0.70 | 1.00 | 1.000 | 0.97 |
| Kang | 10 | bio | view | 0.83 | 0.001 | 0.03 | 0.77 | 1.00 | 1.000 | 0.98 |
| Kang | 10 | random | expression | 1.00 | 0.023 | 0.15 | 0.87 | 1.00 | 1.000 | 0.97 |
| Kang | 10 | random | view | 1.00 | 0.021 | 0.15 | 0.88 | 1.00 | 1.000 | 0.97 |
| Kang | 10 | shuffled | expression | 0.96 | 0.018 | 0.11 | 0.83 | 1.00 | 1.000 | 0.97 |
| Kang | 10 | shuffled | view | 0.96 | 0.016 | 0.11 | 0.86 | 1.00 | 1.000 | 0.98 |
| Pancreas | 1 | bio | expression | 0.80 | 0.001 | 0.02 | 0.65 | 1.00 | 1.000 | 0.98 |
| Pancreas | 1 | bio | view | 0.80 | 0.000 | 0.02 | 0.73 | 1.00 | 1.000 | 0.99 |
| Pancreas | 10 | bio | expression | 0.82 | 0.007 | 0.04 | 0.68 | 1.00 | 1.000 | 0.98 |
| Pancreas | 10 | bio | view | 0.82 | 0.007 | 0.03 | 0.78 | 1.00 | 1.000 | 0.99 |
| PBMC3K | 1 | bio | expression | 0.82 | 0.017 | 0.06 | 0.76 | 1.00 | 1.000 | 0.96 |
| PBMC3K | 1 | bio | view | 0.82 | 0.009 | 0.06 | 0.81 | 1.00 | 1.000 | 0.97 |
| PBMC3K | 1 | random | expression | 0.98 | 0.007 | 0.09 | 0.86 | 1.00 | 1.000 | 0.98 |
| PBMC3K | 1 | random | view | 0.98 | 0.009 | 0.09 | 0.87 | 1.00 | 1.000 | 0.98 |
| PBMC3K | 1 | shuffled | expression | 0.99 | 0.024 | 0.08 | 0.81 | 1.00 | 1.000 | 0.96 |
| PBMC3K | 1 | shuffled | view | 0.99 | 0.013 | 0.09 | 0.84 | 1.00 | 1.000 | 0.97 |

**State encoder.** Table 10 repeats the Kang IFN-$\beta$ experiment (rotation 1) with principal-component states of dimension 4, 8 and 16, with an scVI latent (Lopez et al., 2018) of dimension 8 trained on the training units only, and with the top 8 principal components of cell embeddings from a pretrained single-cell foundation model, Geneformer (Theodoris et al., 2023) (Appendix B). False alarms stay near $\alpha$ and witnesses keep naming an interferon program for every encoder. This table uses one rotation and one grouping seed, so its detection rates differ from the four-rotation means of Table 6 by the seed variation reported in the caption of Table 6. The scVI latent makes $T$ more powerful than principal components, and a larger state dimension ($d = 16$) makes it less powerful, plausibly because more directions are then available to absorb a shift. With states from the pretrained foundation model, $T$ flags 8% ($n = 1$), 59% ($n = 5$) and 98% ($n = 10$) of stimulated units at false-alarm rates of 3% to 6%.

Table 10: State encoder ablation (Kang IFN-$\beta$, rotation 1): TPR / FPR at $\alpha = 0.05$ of HellyCell's statistic $T$, and share of witnesses naming an interferon program.

| Encoder | TPR/FPR $n = 1$ | TPR/FPR $n = 5$ | TPR/FPR $n = 10$ | witness $n = 1$ | witness $n = 5$ | witness $n = 10$ |
|---|---|---|---|---|---|---|
| PCA d=4 | 0.08 / 0.05 | 0.49 / 0.04 | 0.94 / 0.07 | 0.88 | 0.97 | 0.98 |
| PCA d=8 | 0.06 / 0.05 | 0.47 / 0.04 | 0.94 / 0.05 | 0.82 | 0.98 | 1.00 |
| PCA d=16 | 0.06 / 0.05 | 0.26 / 0.04 | 0.60 / 0.06 | 0.91 | 0.97 | 0.99 |
| scVI d=8 | 0.13 / 0.04 | 0.87 / 0.03 | 1.00 / 0.03 | 0.94 | 1.00 | 1.00 |
| Geneformer d=8 | 0.08 / 0.04 | 0.59 / 0.03 | 0.98 / 0.06 | 0.90 | 0.99 | 1.00 |

**Compositional novelty, all cell types.** Table 11 gives detection and false-alarm rates for the seven cell types with enough cells (megakaryocytes are too few); false alarms are measured on the control cells of the same type.

Table 11: Kang, one stimulated cell type held out (rotation 1, one grouping seed, so rates differ from other single runs by seed variation): TPR / FPR at $\alpha = 0.05$.

| Held-out type | $n$ | HellyCell $T$ | anchored $A$ | kNN | max program $\lvert z \rvert$ |
|---|---|---|---|---|---|
| B cells | 1 | 0.10 / 0.09 | 0.07 / 0.09 | 0.28 / 0.05 | 0.08 / 0.09 |
| B cells | 5 | 0.08 / 0.04 | 0.02 / 0.02 | 1.00 / 0.00 | 0.03 / 0.08 |
| CD14+ Monocytes | 1 | 0.05 / 0.10 | 0.08 / 0.10 | 0.63 / 0.03 | 0.07 / 0.11 |
| CD14+ Monocytes | 5 | 0.45 / 0.16 | 0.44 / 0.18 | 1.00 / 0.01 | 0.28 / 0.15 |
| CD4 T cells | 1 | 0.05 / 0.04 | 0.04 / 0.03 | 0.08 / 0.00 | 0.03 / 0.02 |
| CD4 T cells | 5 | 0.06 / 0.03 | 0.03 / 0.01 | 0.97 / 0.00 | 0.02 / 0.01 |
| CD8 T cells | 1 | 0.12 / 0.04 | 0.12 / 0.02 | 0.06 / 0.04 | 0.10 / 0.04 |
| CD8 T cells | 5 | 0.15 / 0.00 | 0.00 / 0.00 | 0.62 / 0.00 | 0.00 / 0.00 |
| Dendritic cells | 1 | 0.20 / 0.10 | 0.20 / 0.10 | 0.91 / 0.41 | 0.15 / 0.08 |
| Dendritic cells | 5 | 0.58 / 0.22 | 0.50 / 0.44 | 1.00 / 1.00 | 0.33 / 0.33 |
| FCGR3A+ Monocytes | 1 | 0.08 / 0.09 | 0.05 / 0.05 | 0.70 / 0.16 | 0.06 / 0.05 |
| FCGR3A+ Monocytes | 5 | 0.06 / 0.08 | 0.06 / 0.00 | 1.00 / 0.20 | 0.06 / 0.00 |
| NK cells | 1 | 0.08 / 0.12 | 0.04 / 0.06 | 0.08 / 0.03 | 0.04 / 0.06 |
| NK cells | 5 | 0.03 / 0.04 | 0.00 / 0.00 | 0.14 / 0.04 | 0.08 / 0.00 |
| ALL | 1 | 0.06 / 0.05 | 0.51 / 0.04 | 0.66 / 0.04 | 0.42 / 0.04 |
| ALL | 5 | 0.60 / 0.06 | 0.99 / 0.03 | 1.00 / 0.05 | 0.97 / 0.04 |

**Runtime.** Table 12 reports single-threaded wall-clock time per unit for the consistency LP and for witness extraction.

Table 12: Runtime per unit (ms, single thread): consistency LP / witness extraction.

| $d$ | $K = 50$ | $K = 100$ | $K = 200$ | $K = 400$ | $K = 800$ |
|---|---|---|---|---|---|
| 4 | 12 / 71 | 12 / 66 | 16 / 70 | 22 / 81 | 38 / 92 |
| 8 | 12 / 99 | 14 / 102 | 19 / 130 | 30 / 130 | 59 / 153 |
| 16 | 13 / 164 | 20 / 225 | 20 / 141 | 39 / 193 | 98 / 194 |
| 32 | 19 / 480 | 23 / 448 | 52 / 485 | 81 / 530 | 130 / 552 |

**Geometry of the feasible set for groups of cells.** Table 13 repeats the geometric analysis of Q1 for groups of $n = 10$ Kang control cells, the setting in which $T$ detects most interferon-stimulated groups. We compare the extent of the calibrated feasible set along each right-singular direction $v_k$ of the normalized program matrix with the central $90\%$ range of group states along that direction (3.29 standard deviations for a Gaussian spread). Along 2 of the 8 directions the set is narrower than that range, so the programs do locate a group coarsely there; along the others it is up to 4.5 times wider. The guaranteed checking threshold of a program is a median of 9.8 standard deviations of its reading across known groups (interquartile range 8.0 to 11.4); averaged over the two signs, the exact threshold is at most half of it (Theorem 2(b), Figure 2). The checks therefore detect large regulatory changes, such as the interferon response, rather than subtle ones.

Table 13: Kang control groups ($n = 10$): median extent of the calibrated feasible set along each right-singular direction $v_k$ of $\tilde{A}$, the standard deviation of group states along $v_k$, and the extent relative to one standard deviation and to the central 90% range (3.29 s.d.).

| Direction $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Singular value of $\tilde{A}$ | 5.76 | 3.03 | 2.48 | 2.19 | 1.93 | 1.64 | 1.52 | 1.29 |
| Extent of $F_{\tilde{q}}$ along $v_k$ | 7.9 | 11.0 | 15.9 | 17.5 | 20.1 | 23.6 | 25.5 | 27.7 |
| s.d. of group states along $v_k$ | 3.6 | 2.7 | 5.2 | 2.2 | 3.2 | 1.8 | 2.3 | 1.9 |
| Extent / s.d. | 2.2 | 4.1 | 3.1 | 8.1 | 6.3 | 13.3 | 11.3 | 14.8 |
| Extent / 90% range | 0.66 | 1.23 | 0.93 | 2.45 | 1.91 | 4.05 | 3.43 | 4.50 |

**A controlled test of the dichotomy.** Figure 5 isolates Theorem 2 in a synthetic world with known states. There are $K = 150$ views of a $d = 8$ dimensional state, with unit noise; each view direction lies in a fixed 4-dimensional subspace up to a leakage component of relative size $\epsilon$ in the other four directions, and states have standard deviation 3; calibration uses $\alpha = 0.05$. As $\epsilon$ shrinks the views become redundant: the smallest singular value of $\tilde{A}$ falls and the calibrated localization radius grows from 11.6 to 186, roughly like $1/\epsilon$ (the box holding the central 90% of states in every coordinate has half-diagonal 14), while joint coverage stays at 0.95. Over the same range the checking cost stays between 1.22 and 1.38, a shift of half the predicted threshold is detected in 93% to 95% of units, every shift at the threshold is detected, and witnesses always contain the shifted view. Means over three seeds.

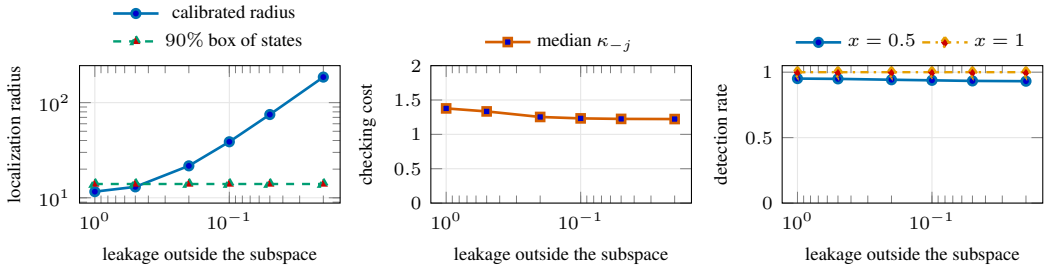



Figure 5: Synthetic test of Theorem 2: as views collapse towards a subspace (right), localization fails (left) while the checking cost and detection stay flat (middle, right).

**Per-type error control.** Table 14 breaks the false-alarm rate on known units down by cell type. With one marginal quantile, typicality detectors flag most units of the rare types (megakaryocytes and dendritic cells), whereas the state-invariant tests stay much closer to $\alpha$ but not at it, because the noise law also differs between types. Per-type calibration restores type-level control for single cells, but groups of rare types are then too few to calibrate, which costs power.

Table 14: Kang IFN-$\beta$, per-type error control (four donor rotations $\times$ two grouping seeds). Largest false-alarm rate over the eight cell types on control units of unseen donors, with one marginal conformal quantile or with a separate quantile per cell type (Mondrian); last columns: unweighted mean detection over cell types at $n = 5$ and $n = 10$ (Table 6 pools units instead). A type with fewer than 19 calibration units gets an infinite quantile and is never flagged.

| | largest false-alarm rate over cell types | | | | | | mean detection over cell types | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $n = 1$ | | $n = 5$ | | $n = 10$ | | $n = 5$ | | $n = 10$ | |
| Method | marginal | per-type | marginal | per-type | marginal | per-type | marginal | per-type | marginal | per-type |
| HellyCell $T$ | 0.15 | 0.06 | 0.13 | 0.07 | 0.25 | 0.07 | 0.87 | 0.56 | 0.98 | 0.56 |
| LS global $G$ | 0.15 | 0.06 | 0.31 | 0.06 | 0.33 | 0.08 | 0.93 | 0.64 | 0.99 | 0.56 |
| LS $w$-test $B$ | 0.12 | 0.06 | 0.21 | 0.07 | 0.25 | 0.08 | 0.95 | 0.63 | 0.99 | 0.56 |
| kNN distance | 0.72 | 0.06 | 1.00 | 0.08 | 1.00 | 0.12 | 0.96 | 0.63 | 0.99 | 0.56 |
| max program $\lvert z \rvert$ | 0.19 | 0.06 | 0.63 | 0.08 | 0.95 | 0.10 | 0.99 | 0.65 | 1.00 | 0.56 |

**Never-seen cell types.** Table 15 removes one cell type at a time from training and calibration. Corollary 1 predicts that a state-invariant statistic flags units of the removed type at rate $\alpha$ if the

type follows the same view model, whereas typicality detectors flag them because their states are far from the training states. B cells and CD14 monocytes are flagged by the state-invariant tests and by kNN, so their programs relate to the state differently from those of the other types. CD8 T cells and FCGR3A monocytes are flagged by kNN in 41% to 42% of groups but by $T$ in only 9% to 11%: their states are atypical, but their programs largely follow the view model learned from the other types.

Table 15: Never-seen cell types (Kang control cells only). Each type is removed from training and calibration; the encoder, the view model and all detectors are fitted without it. Flag rate on units of the removed type and false-alarm rate on units of the known types of the test donors, at $\alpha = 0.05$; mean over the donor rotations in which the removed type has at least 200 control cells in the test donors (other types are not removed). Right: flag rate per removed type at $n = 5$.

| | $n = 1$ | | $n = 5$ | | removed type, $n = 5$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | new type | known | new type | known | B cells | CD14 monocytes | CD4 T cells | CD8 T cells | FCGR3A monocytes | NK cells |
| HellyCell $T$ | 0.17 | 0.05 | 0.49 | 0.05 | 0.99 | 0.83 | 0.16 | 0.09 | 0.11 | 0.31 |
| anchored $A$ | 0.17 | 0.05 | 0.45 | 0.05 | 0.99 | 0.75 | 0.10 | 0.11 | 0.18 | 0.19 |
| LS global $G$ | 0.20 | 0.05 | 0.52 | 0.05 | 0.99 | 0.95 | 0.14 | 0.05 | 0.35 | 0.27 |
| LS $w$-test $B$ | 0.16 | 0.05 | 0.46 | 0.06 | 0.96 | 0.78 | 0.13 | 0.12 | 0.12 | 0.29 |
| kNN distance | 0.22 | 0.05 | 0.56 | 0.06 | 0.82 | 1.00 | 0.14 | 0.41 | 0.42 | 0.33 |
| Mahalanobis | 0.19 | 0.05 | 0.33 | 0.06 | 0.01 | 1.00 | 0.01 | 0.01 | 0.41 | 0.39 |
| Isolation forest | 0.20 | 0.05 | 0.34 | 0.05 | 0.04 | 0.99 | 0.01 | 0.01 | 0.12 | 0.56 |
| max program $\|z\|$ | 0.20 | 0.05 | 0.47 | 0.06 | 0.92 | 0.95 | 0.07 | 0.06 | 0.27 | 0.17 |

**Least-squares residual tests on planted shifts.** Table 16 scores the planted shifts of Q2 with all three state-invariant statistics. A view-level shift $\Delta$ changes program $j$'s standardized least-squares residual by $\sqrt{1 - h_{jj}}\,\Delta$, with $h_{jj}$ the leverage (median $\sqrt{1 - h_{jj}} = 0.99$), whereas the Chebyshev fit absorbs up to a share $\kappa_{-j}/(1 + \kappa_{-j})$ of it, so $B$ detects smaller shifts than $T$. The global test $G$ spreads its sensitivity over all $K$ programs and detects a single shifted program late, but it is the strongest state-invariant test for the broad TCR response (Table 7).

Table 16: Planted shifts scored by the Chebyshev statistic $T$ and by the least-squares residual tests $B$ and $G$ on the same held-out known units (each test calibrated at $\alpha = 0.05$ on the same calibration units; detection among units that pass the respective test before the shift). Shifts are fractions $x$ of $T$'s worst-case bound $2\hat{q}\sigma_j(1 + \kappa_{-j})$. Last column: share of shifts detected by $B$ at $x \geq 0.5$ whose largest standardized residual is the shifted program (view) or a moved program (expression).

| | | | $T$ | | | | $B$ | | | | $G$ | | | | $B$ top |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data | $n$ | shift | $x = 0.2$ | $x = 0.3$ | $x = 0.5$ | $x = 1$ | $x = 0.2$ | $x = 0.3$ | $x = 0.5$ | $x = 1$ | $x = 0.2$ | $x = 0.3$ | $x = 0.5$ | $x = 1$ | correct |
| Kang | 1 | view | 0.00 | 0.04 | 0.75 | 1.00 | 0.01 | 0.21 | 0.92 | 1.00 | 0.01 | 0.02 | 0.03 | 0.23 | 1.00 |
| Kang | 1 | expression | 0.01 | 0.04 | 0.66 | 1.00 | 0.01 | 0.16 | 0.89 | 1.00 | 0.01 | 0.02 | 0.06 | 0.58 | 1.00 |
| Kang | 10 | view | 0.00 | 0.03 | 0.77 | 1.00 | 0.01 | 0.17 | 0.91 | 1.00 | 0.00 | 0.00 | 0.00 | 0.16 | 1.00 |
| Kang | 10 | expression | 0.00 | 0.02 | 0.70 | 1.00 | 0.01 | 0.11 | 0.87 | 1.00 | 0.00 | 0.00 | 0.02 | 0.48 | 1.00 |
| Pancreas | 1 | view | 0.00 | 0.02 | 0.73 | 1.00 | 0.01 | 0.13 | 0.93 | 1.00 | 0.00 | 0.00 | 0.02 | 0.12 | 1.00 |
| Pancreas | 1 | expression | 0.00 | 0.02 | 0.65 | 1.00 | 0.01 | 0.09 | 0.92 | 1.00 | 0.00 | 0.01 | 0.03 | 0.25 | 1.00 |
| PBMC3K | 1 | view | 0.01 | 0.06 | 0.81 | 1.00 | 0.01 | 0.19 | 0.88 | 1.00 | 0.01 | 0.01 | 0.04 | 0.40 | 1.00 |
| PBMC3K | 1 | expression | 0.02 | 0.06 | 0.76 | 1.00 | 0.02 | 0.15 | 0.85 | 1.00 | 0.01 | 0.02 | 0.08 | 0.77 | 1.00 |

**Anchored flagged sets.** Table 17 gives the composition of the anchored flagged sets of detected stimulated units.

Table 17: Anchored flagged sets of detected IFN-$\beta$-stimulated Kang units: mean and median size, mean number of interferon programs, share of sets with at least one other program, and precision averaged over units. This is a separate run with its own groups, so its precision differs slightly from Table 23.

| $n$ | rotation | units | mean size | median size | interferon | other present | precision |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1581 | 2.2 | 2 | 2.1 | 0.11 | 0.94 |
| 1 | 2 | 2610 | 2.8 | 2 | 2.4 | 0.25 | 0.90 |
| 1 | 3 | 2478 | 3.2 | 3 | 3.0 | 0.12 | 0.95 |
| 1 | 4 | 1722 | 4.0 | 4 | 3.9 | 0.12 | 0.96 |
| 1 | all | 8391 | 3.1 | 3 | 2.8 | 0.16 | 0.93 |
| 5 | 1 | 610 | 6.1 | 5 | 5.2 | 0.28 | 0.92 |
| 5 | 2 | 692 | 10.9 | 8 | 6.2 | 0.76 | 0.70 |
| 5 | 3 | 686 | 10.0 | 7 | 6.6 | 0.54 | 0.81 |
| 5 | 4 | 424 | 10.2 | 8 | 7.1 | 0.61 | 0.80 |
| 5 | all | 2412 | 9.3 | 7 | 6.2 | 0.55 | 0.80 |
| 10 | 1 | 303 | 7.4 | 6 | 6.1 | 0.25 | 0.91 |
| 10 | 2 | 343 | 20.8 | 12 | 6.9 | 0.97 | 0.52 |
| 10 | 3 | 341 | 16.6 | 10 | 7.8 | 0.88 | 0.68 |
| 10 | 4 | 207 | 14.4 | 9 | 7.7 | 0.74 | 0.71 |
| 10 | all | 1194 | 15.1 | 9 | 7.1 | 0.72 | 0.69 |

**Batch and protocol shifts.** Table 18 lists the flag rates behind Figure 4c, including single cells.

Table 18: Flag rates on known cell types under donor, batch and protocol shifts at $\alpha = 0.05$ (Figure 4c), for single cells and 10-cell groups; Kang uses rotation 1 with one grouping seed, whereas Figure 4a averages eight runs. inDrop3 is a new batch of a technology whose other batches are in training; CEL-seq2 and Smart-seq2 are protocols related to, but distinct from, CEL-seq and SMARTer.

| | HellyCell $T$ | | anchored $A$ | | kNN distance | | max program $\lvert z\rvert$ | |
|---|---|---|---|---|---|---|---|---|
| Shift | $n = 1$ | $n = 10$ | $n = 1$ | $n = 10$ | $n = 1$ | $n = 10$ | $n = 1$ | $n = 10$ |
| Kang, new donors | 0.05 | 0.03 | 0.04 | 0.00 | 0.04 | 0.07 | 0.04 | 0.04 |
| Pancreas, seen batches | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.04 | 0.05 | 0.04 |
| Pancreas, Smart-seq2 held out | 0.06 | 0.98 | 0.06 | 1.00 | 0.25 | 0.92 | 0.08 | 0.23 |
| Pancreas, CEL-seq2 held out | 0.05 | 0.54 | 0.05 | 0.38 | 0.21 | 0.59 | 0.07 | 0.27 |
| Pancreas, inDrop3 held out | 0.06 | 0.17 | 0.07 | 0.17 | 0.05 | 0.31 | 0.04 | 0.07 |

**Simultaneous shifts without conditioning.** In Table 9 shifts are applied to units that pass the test first, so soundness and detection beyond the bound hold by construction there. Tables 19 and 20 remove that condition and shift up to five programs at once, so that a unit that is already inconsistent, or a witness made only of unshifted programs, counts against the method. The $\ell_1$ fit barely detects these shifts because it is built to ignore a few large residuals, and its residual sum grows little. On these planted shifts the largest residual of the $w$-test is as often a shifted program as the witness contains one (soundness columns), but it carries no guarantee when several programs shift.

Table 19: Several programs shifted at once on held-out known units, with no condition on passing the test before the shift. $|S|$ programs, chosen at random or coupled (one program and its most aligned neighbours), move by $x$ times their own worst-case bound. Detection rate and soundness (share of flagged units whose explanation contains a shifted program: the Helly witness for $T$, the top residual for $B$ and for the $\ell_1$ fit); last column: share of all units that $T$ flags with a witness naming no shifted program, which Theorem 3(b) bounds by $\alpha = 0.05$. Null rows ($x = 0$): false-alarm rates.

| | | | | $T$ | | $B$ | | $\ell_1$ fit | | $G$ | $T$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Data | $n$ | shifted set | $x$ | detect | sound | detect | sound | detect | sound | detect | unsound flags |
| Kang | 1 | none | 0 | 0.05 | | 0.05 | | 0.05 | | 0.06 | |
| | | random, $\|S\| = 1$ | 0.5 | 0.79 | 0.99 | 0.93 | 0.99 | 0.07 | 1.00 | 0.09 | 0.007 |
| | | random, $\|S\| = 1$ | 1 | 1.00 | 1.00 | 1.00 | 0.99 | 0.09 | 1.00 | 0.33 | 0.000 |
| | | random, $\|S\| = 2$ | 0.5 | 0.96 | 0.99 | 0.95 | 0.98 | 0.07 | 0.82 | 0.10 | 0.007 |
| | | random, $\|S\| = 2$ | 1 | 1.00 | 1.00 | 1.00 | 0.99 | 0.11 | 1.00 | 0.77 | 0.000 |
| | | random, $\|S\| = 3$ | 0.5 | 1.00 | 0.99 | 1.00 | 0.99 | 0.09 | 1.00 | 0.16 | 0.007 |
| | | random, $\|S\| = 3$ | 1 | 1.00 | 1.00 | 1.00 | 0.99 | 0.12 | 1.00 | 0.99 | 0.000 |
| | | random, $\|S\| = 5$ | 0.5 | 1.00 | 1.00 | 1.00 | 0.99 | 0.11 | 0.94 | 0.38 | 0.000 |
| | | random, $\|S\| = 5$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.24 | 1.00 | 1.00 | 0.000 |
| | | coupled, $\|S\| = 1$ | 0.5 | 0.71 | 0.99 | 0.92 | 0.98 | 0.07 | 0.70 | 0.09 | 0.007 |
| | | coupled, $\|S\| = 1$ | 1 | 1.00 | 1.00 | 1.00 | 0.99 | 0.08 | 1.00 | 0.26 | 0.000 |
| | | coupled, $\|S\| = 2$ | 0.5 | 0.98 | 0.99 | 0.99 | 0.99 | 0.07 | 1.00 | 0.11 | 0.007 |
| | | coupled, $\|S\| = 2$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.10 | 1.00 | 0.79 | 0.000 |
| | | coupled, $\|S\| = 3$ | 0.5 | 0.99 | 0.99 | 1.00 | 0.99 | 0.09 | 1.00 | 0.18 | 0.007 |
| | | coupled, $\|S\| = 3$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.11 | 1.00 | 0.99 | 0.000 |
| | | coupled, $\|S\| = 5$ | 0.5 | 1.00 | 0.99 | 1.00 | 0.99 | 0.11 | 1.00 | 0.39 | 0.007 |
| | | coupled, $\|S\| = 5$ | 1 | 1.00 | 1.00 | 1.00 | 0.99 | 0.19 | 1.00 | 1.00 | 0.000 |
| Kang | 10 | none | 0 | 0.02 | | 0.01 | | 0.03 | | 0.02 | |
| | | random, $\|S\| = 1$ | 0.5 | 0.78 | 1.00 | 0.93 | 1.00 | 0.03 | 1.00 | 0.03 | 0.000 |
| | | random, $\|S\| = 1$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.03 | 1.00 | 0.15 | 0.000 |
| | | random, $\|S\| = 2$ | 0.5 | 0.97 | 1.00 | 0.99 | 0.99 | 0.03 | 1.00 | 0.05 | 0.000 |
| | | random, $\|S\| = 2$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.06 | 1.00 | 0.67 | 0.000 |
| | | random, $\|S\| = 3$ | 0.5 | 1.00 | 1.00 | 1.00 | 1.00 | 0.05 | 1.00 | 0.10 | 0.000 |
| | | random, $\|S\| = 3$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.10 | 1.00 | 0.99 | 0.000 |
| | | random, $\|S\| = 5$ | 0.5 | 1.00 | 1.00 | 1.00 | 1.00 | 0.07 | 1.00 | 0.19 | 0.000 |
| | | random, $\|S\| = 5$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.15 | 1.00 | 1.00 | 0.000 |
| | | coupled, $\|S\| = 1$ | 0.5 | 0.72 | 1.00 | 0.89 | 1.00 | 0.03 | 1.00 | 0.03 | 0.000 |
| | | coupled, $\|S\| = 1$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.04 | 1.00 | 0.18 | 0.000 |
| | | coupled, $\|S\| = 2$ | 0.5 | 0.95 | 1.00 | 0.99 | 1.00 | 0.03 | 1.00 | 0.06 | 0.000 |
| | | coupled, $\|S\| = 2$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.06 | 1.00 | 0.69 | 0.000 |
| | | coupled, $\|S\| = 3$ | 0.5 | 0.97 | 1.00 | 0.98 | 1.00 | 0.04 | 1.00 | 0.07 | 0.000 |
| | | coupled, $\|S\| = 3$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.09 | 1.00 | 0.95 | 0.000 |
| | | coupled, $\|S\| = 5$ | 0.5 | 1.00 | 1.00 | 1.00 | 1.00 | 0.08 | 1.00 | 0.23 | 0.000 |
| | | coupled, $\|S\| = 5$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.18 | 1.00 | 1.00 | 0.000 |

Table 20: Continued from Table 19: several programs shifted at once on held-out known units, with no condition on passing the test before the shift. $|S|$ programs, chosen at random or coupled (one program and its most aligned neighbours), move by $x$ times their own worst-case bound. Detection rate and soundness (share of flagged units whose explanation contains a shifted program: the Helly witness for $T$, the top residual for $B$ and for the $\ell_1$ fit); last column: share of all units that $T$ flags with a witness naming no shifted program, which Theorem 3(b) bounds by $\alpha = 0.05$. Null rows ($x = 0$): false-alarm rates.

| | | | | $T$ | | $B$ | | $\ell_1$ fit | | $G$ | $T$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Data | $n$ | shifted set | $x$ | detect | sound | detect | sound | detect | sound | detect | unsound flags |
| Pancreas | 1 | none | 0 | 0.07 | | 0.05 | | 0.07 | | 0.08 | |
| | | random, $\|S\| = 1$ | 0.5 | 0.75 | 1.00 | 0.93 | 0.98 | 0.07 | 0.82 | 0.09 | 0.000 |
| | | random, $\|S\| = 1$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.08 | 1.00 | 0.13 | 0.000 |
| | | random, $\|S\| = 2$ | 0.5 | 0.99 | 1.00 | 1.00 | 0.99 | 0.09 | 0.93 | 0.12 | 0.000 |
| | | random, $\|S\| = 2$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.11 | 1.00 | 0.36 | 0.000 |
| | | random, $\|S\| = 3$ | 0.5 | 1.00 | 1.00 | 1.00 | 0.98 | 0.10 | 0.80 | 0.13 | 0.000 |
| | | random, $\|S\| = 3$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.11 | 1.00 | 0.86 | 0.000 |
| | | random, $\|S\| = 5$ | 0.5 | 1.00 | 1.00 | 1.00 | 1.00 | 0.11 | 0.94 | 0.20 | 0.000 |
| | | random, $\|S\| = 5$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.15 | 1.00 | 1.00 | 0.000 |
| | | coupled, $\|S\| = 1$ | 0.5 | 0.76 | 0.98 | 0.96 | 0.98 | 0.07 | 0.73 | 0.09 | 0.013 |
| | | coupled, $\|S\| = 1$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.09 | 1.00 | 0.17 | 0.000 |
| | | coupled, $\|S\| = 2$ | 0.5 | 0.95 | 1.00 | 0.99 | 0.99 | 0.09 | 0.85 | 0.11 | 0.000 |
| | | coupled, $\|S\| = 2$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.10 | 1.00 | 0.36 | 0.000 |
| | | coupled, $\|S\| = 3$ | 0.5 | 1.00 | 1.00 | 1.00 | 1.00 | 0.10 | 1.00 | 0.14 | 0.000 |
| | | coupled, $\|S\| = 3$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.11 | 1.00 | 0.81 | 0.000 |
| | | coupled, $\|S\| = 5$ | 0.5 | 1.00 | 1.00 | 1.00 | 0.99 | 0.11 | 0.88 | 0.19 | 0.000 |
| | | coupled, $\|S\| = 5$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.15 | 1.00 | 1.00 | 0.000 |
| PBMC3K | 1 | none | 0 | 0.05 | | 0.07 | | 0.02 | | 0.02 | |
| | | random, $\|S\| = 1$ | 0.5 | 0.75 | 1.00 | 0.88 | 0.98 | 0.03 | 0.80 | 0.07 | 0.000 |
| | | random, $\|S\| = 1$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.05 | 1.00 | 0.35 | 0.000 |
| | | random, $\|S\| = 2$ | 0.5 | 0.99 | 0.99 | 0.97 | 0.99 | 0.05 | 1.00 | 0.12 | 0.007 |
| | | random, $\|S\| = 2$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.09 | 1.00 | 0.92 | 0.000 |
| | | random, $\|S\| = 3$ | 0.5 | 1.00 | 1.00 | 0.99 | 0.99 | 0.05 | 1.00 | 0.21 | 0.000 |
| | | random, $\|S\| = 3$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.12 | 1.00 | 1.00 | 0.000 |
| | | random, $\|S\| = 5$ | 0.5 | 1.00 | 1.00 | 1.00 | 1.00 | 0.11 | 1.00 | 0.65 | 0.000 |
| | | random, $\|S\| = 5$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.31 | 1.00 | 1.00 | 0.000 |
| | | coupled, $\|S\| = 1$ | 0.5 | 0.87 | 1.00 | 0.90 | 0.99 | 0.03 | 0.80 | 0.05 | 0.000 |
| | | coupled, $\|S\| = 1$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.05 | 1.00 | 0.38 | 0.000 |
| | | coupled, $\|S\| = 2$ | 0.5 | 0.93 | 1.00 | 0.99 | 1.00 | 0.03 | 1.00 | 0.14 | 0.000 |
| | | coupled, $\|S\| = 2$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.09 | 1.00 | 0.91 | 0.000 |
| | | coupled, $\|S\| = 3$ | 0.5 | 1.00 | 1.00 | 1.00 | 1.00 | 0.05 | 1.00 | 0.21 | 0.000 |
| | | coupled, $\|S\| = 3$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.13 | 1.00 | 0.99 | 0.000 |
| | | coupled, $\|S\| = 5$ | 0.5 | 1.00 | 1.00 | 1.00 | 1.00 | 0.11 | 1.00 | 0.62 | 0.000 |
| | | coupled, $\|S\| = 5$ | 1 | 1.00 | 1.00 | 1.00 | 1.00 | 0.32 | 1.00 | 1.00 | 0.000 |

**Many donor splits.** Table 21 repeats the Kang experiment on 16 further random donor partitions to estimate how often each detector exceeds its nominal false-alarm rate when the test donors are new; the $\ell_1$ fit (least absolute deviations, a state-invariant statistic suited to sparse faults) is included.

Table 21: False alarms on control units of unseen donors over 16 random partitions of the eight Kang donors (four training, two calibration, two test), at $\alpha = 0.05$: mean / 90th percentile / maximum over splits, and the share of split and group-size combinations above 7.5%.

| Method | $n = 1$ | $n = 5$ | $n = 10$ | share $> 7.5\%$ |
|---|---|---|---|---|
| HellyCell $T$ | 0.050 / 0.064 / 0.079 | 0.048 / 0.067 / 0.112 | 0.050 / 0.081 / 0.154 | 0.08 |
| anchored $A$ | 0.051 / 0.065 / 0.077 | 0.056 / 0.087 / 0.146 | 0.055 / 0.093 / 0.115 | 0.19 |
| LS global $G$ | 0.051 / 0.064 / 0.087 | 0.059 / 0.098 / 0.180 | 0.056 / 0.101 / 0.192 | 0.17 |
| LS $w$-test $B$ | 0.051 / 0.065 / 0.076 | 0.051 / 0.074 / 0.090 | 0.047 / 0.066 / 0.115 | 0.08 |
| $\ell_1$ fit | 0.051 / 0.065 / 0.090 | 0.058 / 0.092 / 0.172 | 0.055 / 0.087 / 0.231 | 0.15 |
| kNN distance | 0.051 / 0.063 / 0.065 | 0.056 / 0.080 / 0.100 | 0.062 / 0.117 / 0.123 | 0.17 |
| Mahalanobis | 0.052 / 0.067 / 0.068 | 0.052 / 0.075 / 0.097 | 0.049 / 0.082 / 0.103 | 0.10 |
| Isolation forest | 0.051 / 0.066 / 0.074 | 0.050 / 0.067 / 0.082 | 0.049 / 0.062 / 0.109 | 0.04 |
| max program $\lvert z \rvert$ | 0.053 / 0.073 / 0.097 | 0.058 / 0.101 / 0.120 | 0.058 / 0.103 / 0.123 | 0.21 |

**A target state the encoder cannot see.** Table 22 repeats Q1 with the state defined by all genes of the universe, so that programs read genes that shape the target but that the encoder does not see. Programs then improve the point estimate by up to 25% (best program-informed estimator per dataset: PBMC3K 6%, Kang 13%, Pancreas 25%), as Section 4 predicts for information outside the encoder's input, whereas calibrated sets built from the programs alone stay 14 to 21 times wider than the encoder's own.

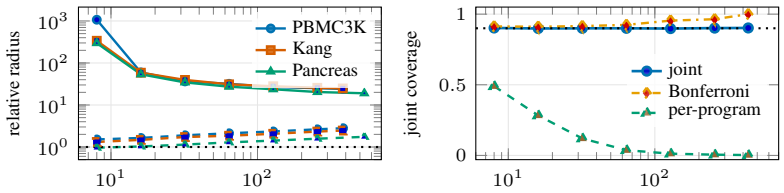
Table 22: Q1 with the target state defined on the whole gene universe (highly variable genes and every detected program gene) while the encoder reads only the thinned highly variable genes: error relative to the ridge encoder (MLP rows relative to the MLP without programs) and calibrated radius relative to the encoder's box. Means over three splits (MLP rows: one split).

| Dataset | fusion | stacking | message passing | MLP + programs | radius: backbone + programs | radius: programs only |
|---|---|---|---|---|---|---|
| PBMC3K | 1.26 | 1.06 | 0.95 | 0.94 | 3.2 | 21.5 |
| Kang | 1.09 | 0.87 | 0.94 | 0.92 | 2.7 | 19.3 |
| Pancreas | 1.01 | 0.75 | 0.82 | 0.86 | 2.5 | 14.0 |

Table 23: Explanations of detected novelty with attribution controls. *Any*: share of explanations naming a reference response program; *precision*: share of named programs that are reference programs, averaged over units. Witnesses have $d + 1 = 9$ programs; ranked lists are compared in Table 2, which comes from a separate run that regroups the cells with other seeds and scores at most 250 novel units per setting, so its shares differ by a few points. The chance row is the probability that nine programs drawn at random include a reference program; Table 2 draws them. With random or gene-shuffled programs, $T$ still flags 89% and 74% of the 10-cell IFN-$\beta$ units (curated programs: 97%).

| | | Kang IFN-$\beta$ | | Datlinger TCR | |
|---|---|---|---|---|---|
| Explanation | Programs | $n = 1$ | $n = 10$ | $n = 1$ | $n = 10$ |
| Helly witness, any | curated | 0.88 | 0.96 | 0.67 | 0.72 |
| Helly witness, any | gene-shuffled | 0.23 | 0.28 | 0.40 | 0.44 |
| Helly witness, any | random sets | 0.14 | 0.10 | 0.45 | 0.39 |
| chance for 9 random programs | | 0.21 | 0.21 | 0.43 | 0.43 |
| Anchored flagged set, precision | curated | 0.94 | 0.71 | 0.27 | 0.43 |
| Anchored flagged set, precision | gene-shuffled | 0.03 | 0.02 | 0.04 | 0.07 |
| Anchored flagged set, precision | random sets | 0.02 | 0.01 | 0.04 | 0.06 |
| base rate of reference programs | | 0.025 | 0.025 | 0.060 | 0.060 |

**Adding programs and joint calibration.** Figure 6 repeats Q1 as programs are added: the programs-only radius levels off far above the backbone, and the joint radius grows because $\hat{q}$ increases with $K$ while new slabs cut no new direction. Per-program calibration collapses the joint coverage of all slabs, Bonferroni over-covers, and joint calibration stays near nominal, as Theorem 1 guarantees.




(a) number of programs $K$ (b) number of programs $K$

Figure 6: (a) Median calibrated radius at 90% coverage over 150 thinned test cells per split, relative to the backbone's box (dotted), as programs are added; solid: programs only, dashed: backbone plus programs. (b) Joint coverage of all program slabs under per-program, Bonferroni and joint calibration (dotted: 0.9).

**Variation over splits and the localization comparator.** Table 24 repeats the main rows of Table 1 as mean ± standard deviation over the three random splits (the MLP row of the wider target uses one split). The least-squares comparator uses the same thinned cells, the same programs and the same noise scales as the feasible set. For the programs alone (and for the backbone and programs together), it takes the weighted least-squares state $\hat{z}$ and calibrates either the Euclidean error $\|z^\star - \hat{z}\|_2$ (a ball) or the Mahalanobis error $\|\tilde{A}(z^\star - \hat{z})\|_2$ (an ellipsoid with shape $\tilde{A}^\top \tilde{A}$), each by its own conformal quantile at $\alpha = 0.1$; the calibrated radius of a ball of radius $q$ is $q\sqrt{d}$, and that of the ellipsoid is $q\,(\mathrm{tr}(\tilde{A}^\top \tilde{A})^{-1})^{1/2}$, the half-diagonal of its bounding box. All sets reach their nominal coverage. The ellipsoid is wider than the ball because the Mahalanobis residual of the programs is heavy-tailed (its calibrated quantile is two to four times the Gaussian value), so a better-shaped set pays for a larger multiplier.

Table 24: Localization (Table 1) as mean ± standard deviation over three random splits: state error and calibrated radius relative to the backbone; last block: the wider target, the state of the full-depth cell on every gene that the programs read.

| | PBMC3K | Kang (ctrl) | Pancreas |
|---|---|---|---|
| *State error relative to the backbone ridge encoder* | | | |
| + programs, noise-weighted fusion | 1.22 ± 0.02 | 1.23 ± 0.01 | 1.63 ± 0.02 |
| + programs, cross-fitted stacking | 1.04 ± 0.02 | 0.99 ± 0.00 | 0.98 ± 0.00 |
| + programs, message passing | 0.98 ± 0.01 | 0.99 ± 0.00 | 0.98 ± 0.00 |
| MLP encoder | 1.07 ± 0.02 | 1.05 ± 0.00 | 1.09 ± 0.01 |
| + programs as extra MLP inputs | 1.05 ± 0.02 | 1.04 ± 0.01 | 1.04 ± 0.01 |
| Programs only (weighted least squares) | 2.61 ± 0.05 | 2.81 ± 0.02 | 3.31 ± 0.04 |
| *Calibrated radius relative to the backbone box* | | | |
| Backbone ball (least squares) | 1.40 ± 0.03 | 1.41 ± 0.01 | 1.20 ± 0.08 |
| Programs only, least-squares ball | 3.76 ± 0.06 | 4.01 ± 0.04 | 3.71 ± 0.28 |
| Programs only, least-squares ellipsoid | 5.07 ± 0.06 | 5.58 ± 0.09 | 8.15 ± 0.59 |
| Programs only, $\ell_\infty$ feasible set | 24.06 ± 0.49 | 24.74 ± 0.62 | 19.63 ± 1.54 |
| Backbone + programs, least-squares ellipsoid | 2.40 ± 0.07 | 2.14 ± 0.01 | 3.30 ± 0.20 |
| Backbone + programs, $\ell_\infty$ feasible set | 2.80 ± 0.04 | 2.47 ± 0.05 | 1.96 ± 0.13 |
| *State on all genes that the programs read: error relative to the backbone* | | | |
| + programs, cross-fitted stacking | 1.06 ± 0.03 | 0.87 ± 0.00 | 0.75 ± 0.00 |
| + programs, message passing | 0.95 ± 0.01 | 0.94 ± 0.01 | 0.82 ± 0.01 |
| + programs as extra MLP inputs (vs. MLP; one split) | 0.94 | 0.92 | 0.86 |
| Programs only, $\ell_\infty$ feasible set: calibrated radius relative to the backbone box | 21.5 ± 0.8 | 19.3 ± 0.6 | 14.0 ± 0.2 |

**Mechanism tests, all configurations.** Table 25 lists, for every configuration of the planted-shift experiments, the exact thresholds of Theorem 2(b) and the widths of Theorem 2(a), verified numerically, on the same units (60 units that pass the test and 20 random programs each; widths along 40 directions for 30 of the units). $R^2$ is the share of the variance of the log threshold across programs explained by a predictor with one free constant; the least-squares redundancy predictor $1/\sqrt{1-h_{jj}}$ orders programs almost as well as $1+\kappa_{-j}$ (rank correlations within 0.05), but varies too little across programs to predict the size of the thresholds. Random and shuffled programs have $\kappa_{-j} \approx 1$ for most programs, so there is less variation to explain. In reading units, the noise level $\hat{q}\sigma_j$ alone explains 0.941 of the variance of the thresholds across curated programs, and the leave-one-out cost raises this to 0.985: the noise level carries most of the variation, and $1 + \kappa_{-j}$ a fourfold smaller remainder.

Table 25: Mechanism tests of Theorem 2 on the units of the planted-shift experiments: exact symmetric threshold $\bar{\Delta}_j = \frac{1}{2}(\Delta_j^+ + \Delta_j^-)$ relative to the synthesis scale $\hat{q}(1+\kappa_{-j})$ (median and 10th to 90th percentile), share inside the band of Theorem 2(b), $R^2$ of $\log \bar{\Delta}_j$ for the predictors $1+\kappa_{-j}$ and $1/\sqrt{1-h_{jj}}$, and median width $W_u(F_{\hat{q}})$ relative to $2\hat{q}\kappa(u)$.

| Data | $n$ | Programs | $K$ | $\hat{q}$ | median $\kappa_{-j}$ | $\bar{\Delta}_j/\hat{q}(1+\kappa_{-j})$ | in band | $R^2$: $\kappa_{-j}$ | $R^2$: $h_{jj}$ | $W_u/2\hat{q}\kappa(u)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| PBMC3K | 1 | curated | 389 | 3.73 | 0.81 | 0.88 [0.75, 0.94] | 100% | 0.62 | 0.09 | 0.71 |
| Kang | 1 | curated | 384 | 4.03 | 0.77 | 0.93 [0.80, 0.98] | 100% | 0.74 | 0.09 | 0.83 |
| Kang | 10 | curated | 384 | 3.80 | 0.81 | 0.91 [0.83, 0.96] | 100% | 0.84 | 0.10 | 0.80 |
| Pancreas | 1 | curated | 536 | 4.43 | 0.79 | 0.94 [0.86, 0.98] | 100% | 0.83 | 0.07 | 0.85 |
| Pancreas | 10 | curated | 536 | 3.68 | 0.82 | 0.92 [0.84, 0.96] | 100% | 0.82 | 0.07 | 0.83 |
| Kang | 10 | random | 384 | 3.03 | 1.03 | 0.80 [0.69, 0.87] | 100% | 0.46 | 0.07 | 0.61 |
| Kang | 10 | shuffled | 384 | 3.11 | 0.98 | 0.82 [0.70, 0.90] | 100% | 0.51 | 0.07 | 0.60 |
| *Thinned single cells of Q1 (programs only, localization calibration at $\alpha = 0.1$)* | | | | | | | | | | |
| PBMC3K | 1 | curated | 389 | 5.83 | | | | | | 0.92 |
| Kang | 1 | curated | 380 | 6.10 | | | | | | 0.95 |
| Pancreas | 1 | curated | 536 | 5.48 | | | | | | 0.93 |

**A unit-free comparison.** For a unit direction $u$, $\kappa(u)$ carries the units of the state, whereas $\kappa_{-j}$ is scale-free. Scaled to the length of a typical program, $\kappa(u) \cdot \mathrm{median}_j \|\tilde{a}_j\|_2$ is scale-free: along the worst of the sampled directions it is a median of 1.5 to 1.8 over the units of the threshold experiments, against a median $\kappa_{-j}$ of 0.77 to 0.82. The worst-covered direction thus costs about twice as much as re-measuring a program, and the calibrated set along it is about as wide as the slab of one typical program along that program's own direction: hundreds of programs locate the state no better than one program measures its own projection. In natural units, for groups of ten cells the calibrated set is wider than the central $90\%$ range of group states along 94% (Kang) and 83% (Pancreas) of the sampled directions (median 2.4 and 1.7 times that range). Of the 384 Kang programs, 49 are hallmarks (median $\kappa_{-j}$ 1.02 at $n = 10$) and 335 regulons (0.77); the 10 interferon programs are not unusually redundant ($\kappa_{-j}$ 0.77 to 1.60), so detection of the interferon response does not rest on near-duplicate IRF or STAT regulons.

**Prospective test of the operating envelope.** For every real perturbation, the perturbed cells and their reference cells (the same cell type and donors for Kang, unstimulated cells with the same guide for Datlinger, control cells for the screens; perturbations with fewer than 20 cells on either side are skipped, and Datlinger cells without an assigned guide form one perturbation) are split at random into halves. The mean shift $s$ of the normalized readings, hence $\hat{\rho} = N(s)/\hat{q}$, is estimated on one half (pilot); groups of $n = 10$ cells are formed from the other half (held out) and scored by $T$; the roles of the halves are swapped, and this is repeated for three random splits, so that $\hat{\rho}$ never sees the cells on which detection is measured. $\hat{q}$ and the envelope of Proposition 1 come from calibration units only, and the envelope is evaluated at the mean pilot $\hat{\rho}$, from the calibration units of the perturbation's own dataset (Figure 3c shades the envelope of the calibration units of all four datasets pooled). Confidence intervals resample perturbations (4,000 bootstrap draws), because groups of one perturbation are not independent. Table 26 also reports an energy-distance permutation test (Peidli et al., 2024) between perturbed and reference cells (199 permutations, up to 300 cells per side, Benjamini-Hochberg at $0.05$ within each screen), in the state space and in the space of program readings, and the share of groups flagged by $T$ but by neither least-squares test, each calibrated by its own conformal quantile. The noise of a pilot mean inflates $\hat{\rho}$ on average, because $N$ is a seminorm, hence convex (Jensen's inequality): the median inflation relative to the estimate from all cells is 1% (Kang), 11% (Norman), 36% (Datlinger), 38% (Dixit), largest where reference cells are few or the true shift is small.

Table 26: Prospective test ($n = 10$, $\alpha = 0.05$): number of perturbations, median pilot $\hat{\rho}$, rank correlation of pilot $\hat{\rho}$ with held-out detection [95% bootstrap interval], mean held-out detection, detection rates outside the envelope, perturbations with a significant energy test (state / readings) and the mean share of groups flagged by $T$ alone.

| Perturbations | number | median $\hat{\rho}$ | rank correlation | detection | outside envelope | energy test | $T$ only |
|---|---|---|---|---|---|---|---|
| IFN-$\beta$, per Kang cell type and rotation | 28 | 1.51 | 0.60 [0.32, 0.78] | 0.99 | 0 | 28 / 28 | 0.0% |
| TCR stimulation, per Datlinger guide target | 11 | 0.79 | 0.14 [−0.48, 0.62] | 0.18 | 0 | 10 / 2 | 0.0% |
| Dixit knockouts, per target | 10 | 0.18 | 0.41 [−0.45, 0.96] | 0.04 | 0 | 0 / 0 | 0.5% |
| Norman activations, per target | 105 | 0.42 | 0.91 [0.85, 0.94] | 0.36 | 0 | 102 / 102 | 1.5% |
| All | 154 | 0.51 | 0.89 [0.83, 0.94] | 0.44 | 0 | | 1.1% |

**Which pilot quantity predicts detection?** Table 27 compares, on the same pilot and held-out halves, the ratio $\hat{\rho}$ of Proposition 1 with quantities that ignore the geometry: the raw size of the mean shift, which keeps the part that a change of state explains; least-squares residual sizes of the shift, which remove that part in the Euclidean geometry; and the flag rate of $T$ measured directly on groups of pilot cells. Within screens, removing the state component matters: $\hat{\rho}$ predicts better than the raw size (paired difference of rank correlations, Norman 0.10, Kang 0.24), whereas the Euclidean residual size predicts as well, so what matters is that the state is removed, not the norm; pooled over screens all shift sizes do equally well, because the screens differ more than the perturbations within them. The direct pilot flag rate predicts best, as a measurement of the same quantity should; $\hat{\rho}$ needs only the mean shift and comes with the bounds of Proposition 1.

Table 27: Rank correlation with held-out detection by $T$ ($n = 10$) of quantities computed on the pilot halves of the prospective test [95% bootstrap interval over perturbations].

| Pilot quantity | all (154) | Kang (28) | Datlinger (11) | Dixit (10) | Norman (105) |
|---|---|---|---|---|---|
| $\hat{\rho} = N(s)/\hat{q}$ (state removed, $\ell_\infty$) | 0.89 [0.83, 0.94] | 0.60 [0.32, 0.78] | 0.14 [−0.48, 0.62] | 0.41 [−0.45, 0.96] | 0.91 [0.85, 0.94] |
| $\max_j \lvert s_j \rvert/\hat{q}$ (state not removed) | 0.88 [0.83, 0.91] | 0.35 [0.02, 0.61] | 0.40 [−0.40, 0.93] | 0.76 [0.30, 0.96] | 0.81 [0.74, 0.86] |
| $(G(s)/\hat{q}_G)^{1/2}$ (state removed, least squares) | 0.89 [0.83, 0.94] | 0.57 [0.32, 0.76] | 0.41 [−0.21, 0.80] | 0.60 [−0.04, 1.00] | 0.94 [0.90, 0.97] |
| $B(s)/\hat{q}_B$ (largest standardized residual) | 0.89 [0.82, 0.93] | 0.51 [0.26, 0.72] | 0.19 [−0.45, 0.63] | 0.28 [−0.51, 0.91] | 0.88 [0.81, 0.93] |
| flag rate of $T$ on pilot groups (direct measurement) | 0.97 [0.95, 0.98] | 0.64 [0.10, 0.99] | 0.68 [0.01, 0.98] | 0.41 [−0.43, 0.96] | 0.97 [0.94, 0.98] |

**Preprocessing.** Table 28 repeats the Kang IFN-$\beta$ experiment (rotation 1) with the program-size cap, the near-duplicate threshold and the noise scales changed. The geometry and the conclusions do not depend on these choices: the leave-one-out costs stay small, the worst sampled synthesis cost stays large, false alarms stay near $\alpha$, and groups of ten stimulated cells are detected with witnesses that name an interferon program. On this dataset no pair of programs has a Jaccard index between 0.8 and 0.9, so this variant reproduces the default.

Table 28: Sensitivity to program preprocessing and noise scales (Kang, rotation 1, one grouping seed): number of programs, leave-one-out cost $\kappa_{-j}$ (median, 10th to 90th percentile) and largest sampled $\kappa(u)$ at $n = 10$, and TPR / FPR of $T$ at $\alpha = 0.05$ and the share of witnesses naming an interferon program.

| Variant | $K$ | $\kappa_{-j}$, $n = 10$ | max $\kappa(u)$ | TPR / FPR, $n = 1$ | TPR / FPR, $n = 10$ | witness |
|---|---|---|---|---|---|---|
| default (cap 200, Jaccard 0.8, tail equalizer) | 384 | 0.80 [0.49, 1.28] | 4.6 | 0.06 / 0.05 | 0.93 / 0.04 | 0.99 |
| no tail equalizer | 384 | 0.81 [0.48, 1.27] | 4.6 | 0.05 / 0.05 | 0.91 / 0.05 | 1.00 |
| robust scale (MAD), no equalizer | 384 | 0.78 [0.47, 1.25] | 4.3 | 0.05 / 0.05 | 0.87 / 0.02 | 1.00 |
| programs capped at 100 genes | 384 | 0.82 [0.50, 1.33] | 4.6 | 0.06 / 0.05 | 0.93 / 0.04 | 1.00 |
| no cap on program size | 384 | 0.80 [0.48, 1.30] | 4.6 | 0.06 / 0.05 | 0.93 / 0.04 | 0.99 |
| near-duplicate threshold 0.7 | 382 | 0.80 [0.49, 1.29] | 4.6 | 0.06 / 0.05 | 0.94 / 0.04 | 1.00 |
| near-duplicate threshold 0.9 | 384 | 0.80 [0.49, 1.28] | 4.6 | 0.06 / 0.05 | 0.93 / 0.04 | 0.99 |

**Size of the estimation term of Corollary 1.** Table 29 bounds the part of the residual at a held-out state that comes from estimating the view model: the standard deviation of $\delta a_j^\top z_{\mathrm{new}} + \delta b_j$ is $s_j \, (x^\top (X^\top X)^{-1} x)^{1/2}$ for the training design $X$, $x = (z_{\mathrm{new}}, 1)$ and the single-cell residual standard deviation $s_j$, which we report in units of the slab half-width $\hat{q}\sigma_j(n)$. It stays below a few hundredths of a half-width even for states several times farther from the training states than calibration units are, and the AUROC of $T$ is not ordered by that distance (Spearman correlation 0.00 over the seven types), so the departures from chance in Figure 4b reflect changed program-state relations rather than estimation error.

Table 29: Held-out stimulated cell types (Kang, rotation 1, $n = 5$): largest and median standard deviation of the estimation term over held-out groups and programs, in slab half-widths; distance of held-out groups to the ten nearest training groups relative to that of calibration groups; AUROC of $T$ (Figure 4b).

| Held-out type | largest term | median term | distance ratio | AUROC of $T$ |
|---|---|---|---|---|
| B cells | 0.024 | 0.015 | 6.2 | 0.54 |
| CD14+ Monocytes | 0.032 | 0.021 | 6.0 | 0.78 |
| CD4 T cells | 0.018 | 0.009 | 3.4 | 0.61 |
| CD8 T cells | 0.017 | 0.010 | 2.8 | 0.66 |
| Dendritic cells | 0.024 | 0.017 | 8.4 | 0.59 |
| FCGR3A+ Monocytes | 0.033 | 0.024 | 7.8 | 0.56 |
| NK cells | 0.024 | 0.015 | 1.9 | 0.42 |

## D PERTURBATION SCREENS: A NEGATIVE RESULT

We asked whether HellyCell recovers the perturbed transcription factor's own CollecTRI regulon in two CRISPR screens: knockouts of ten transcription factors in K562 cells (Dixit et al., 2016) and CRISPR activation of single genes in K562 cells (Norman et al., 2019). The model is fitted and calibrated on control cells, units are disjoint groups of $n = 10$ cells, $T$ is calibrated by its own conformal quantile at $\alpha = 0.05$ as in Algorithm 1, and Table 30 ranks the target's regulon among all programs by witness frequency, peeling order, anchored residual and the standard differential-activity $|t|$-statistic between perturbed and control cells.

In the knockout screen, $T$ flags 2.7% of perturbed groups on average, at a false-alarm rate of 4.3% on control groups, and even differential activity places the target's regulon at median rank 266 of 419, no better than chance: regulon activity carries no detectable on-target signal in this screen, every knockout has $\hat{\rho} < 1$ (Figure 3c), and an energy test detects no effect of any knockout (Table 26). In the activation screen, 23 of 105 targets have a regulon, and only 3 of them move their own regulon near the top of the differential-activity ranking (IRF1 rank 5, TP73 rank 6, CEBPA rank 11 of 428); $T$ flags 100%, 100%, 100% of their groups, respectively, and 35% of groups over all 105 targets, although an energy test finds an effect in the state space for 102 of the 105 activations analysed in Table 26. For these targets the witness-frequency ranks of the target's own regulon are 37, 5, 66. A witness only has to contain *some* shifted program, and an activated transcription factor shifts many downstream programs, so the target's own regulon is a narrow reference for witnesses. Because standard analysis fails on most targets, these screens can neither validate nor refute witnesses; what they do show is that most single-gene perturbations lie below the detection threshold (Proposition 1), whether or not they have an effect.

Table 30: Perturbation screens ($n = 10$). Detection: share of perturbed groups flagged by $T$, averaged over targets. Other columns: median rank of the target's own regulon among all programs (share of targets with rank $\leq 10$), over the targets that have a regulon; witness and peeling ranks only over targets with at least one flagged group (n/a: none).

| Screen | targets (with regulon) | detection | false alarms | witness | peeling | anchored | differential activity |
|---|---|---|---|---|---|---|---|
| Dixit (knockout) | 10 (10) | 0.03 | 0.043 | 221 (0.00) | 221 (0.00) | 246 (0.00) | 266 (0.00) |
| Norman (activation) | 105 (23) | 0.35 | 0.076 | 232 (0.04) | 232 (0.00) | 132 (0.09) | 231 (0.09) |

# E NOTATION

Table 31: Notation.

| Symbol | Meaning |
|---|---|
| $y$, $n$ | a unit (single cell or group of $n$ cells) |
| $z \in \mathbb{R}^d$, $z^\star$ | cell state (top $d$ principal coordinates; scVI in Table 10); target state of a unit |
| $z(y)$, $z_{\mathrm{new}}$ | encoder state of unit $y$; state of a test unit under Assumption 3 |
| $\tilde{x}_g(y)$ | standardized log-expression of gene $g$, averaged over the unit's cells |
| $K$, $\mathcal{M}_j$, $\omega_{jg}$ | number of programs; gene set of program $j$; sign of gene $g$ in program $j$ |
| $r_j(y)$ | reading of program $j$: signed mean of $\tilde{x}_g$ over $\mathcal{M}_j$ |
| $a_j, b_j, \sigma_j(y), \varepsilon_j$ | direction, offset, noise scale and standardized noise in (1) |
| $s_j(z; y)$ | normalized residual $\lvert r_j - b_j - a_j^\top z\rvert/\sigma_j$ |
| $C_j(q; y)$, $F_q(y)$ | slab of program $j$ at multiplier $q$; feasible set, the intersection of all slabs |
| $T(y)$, $M(y, z^\star)$, $A(y)$ | consistency, localization and anchored statistics |
| $T_{-j}(y)$, $z_c$ | $T$ without program $j$; Chebyshev centre (the minimizer in $T$) |
| $m$, $k$, $\hat{q}$, $\alpha$ | calibration units; rank $\lceil (m+1)(1-\alpha)\rceil$; conformal quantile; level |
| $\tilde{a}_j$, $\tilde{A}$, $\tilde{c}_j$ | normalized direction $a_j/\sigma_j$; their stack; normalized centred reading |
| $H$, $Q$, $r(H)$ | $\mathrm{conv}\{\pm\tilde{a}_k\}$; polar of $H$; its inradius |
| $\kappa(u)$, $\kappa_{-j}$ | $\ell_1$ synthesis cost of direction $u$; leave-one-out cost of program $j$ |
| $W_u(F)$ | width of a set $F$ along direction $u$ |
| $\Delta$, $\Delta_j^{\pm}$, $\bar{\Delta}_j$ | shift of a reading in noise units; exact one-sided detection thresholds; their mean |
| $q(1+\kappa_{-j})$, $2q(1+\kappa_{-j})$ | synthesis scale; worst-case sufficient threshold of program $j$ |
| $N(x)$ | seminorm $\min_\delta \lVert x - \tilde{A}\delta\rVert_\infty$ of normalized readings; $T(y) = N(\tilde{c})$ |
| $s$, $\rho$, $\hat{\rho}$ | mean shift of normalized readings; ratio $N(s)/\hat{q}$; its estimate on pilot cells |
| $x$ | planted shift over the worst-case sufficient threshold $2\hat{q}\sigma_j(1+\kappa_{-j})$ |
| $\hat{z}$ | weighted least-squares state, $\arg\min_z \sum_j s_j(z; y)^2$ |
| $W$ | witness: programs whose slabs have empty intersection, $\lvert W\rvert \le d+1$ |
| $S$, $\beta$ | shifted programs (Assumption 2); probability of a flag |
| $G$, $h_{jj}$ | least-squares residual sum of squares; leverage of program $j$ |
| $B$ | Baarda's $w$-statistic, the largest residual over $\sqrt{1-h_{jj}}$ |
| $\delta a_j$, $\delta b_j$ | population minus fitted view-model coefficients |
| $\iota_j$ | $j$-th standard basis vector |