Pairwise correlation cannot tell you if a single-cell potency score adds biology: an own-baseline test and a field audit

Damir Iskakov ISKAKOV.DAM@GMAIL.COM
Independent Researcher, Astana, Kazakhstan

Abstract

Potency and stemness scores are used to nominate cancer stem-cell and drug-tolerant states from single-cell RNA sequencing (scRNA-seq) data, and they are validated by correlating the score against a pseudotime, a marker panel, or a known hierarchy. A score that restates detection depth, the number of genes detected in a cell, passes that test. A score that carries real ordering information passes it too. The criterion does not separate them, and a score that is a reparametrization of detection depth will rank a patient's cells by how many genes were detected in each. The question here is whether the pipeline built on a low-order statistic, which we call its primitive, orders cells better than that statistic alone, against a ground truth fixed before the cells were dissociated. Depth robustness is a separate question and its authors have tested it. Twelve visually staged timepoints of zebrafish development supply such a ground truth, an ordinal of twelve ranked levels over 39,505 whole-embryo cells (Farrell et al., 2018). The calibration is developmental and the target of the critique is not: the drug-tolerance datasets we work with carry binary fate labels rather than an ordering.

We audit fourteen published unsupervised scores and place nine of them on that ordinal. Skill is the Kendall τ_b between the ordinal and the score's ranks once a least-squares fit on the primitive's ranks is removed. A value counts when it lies above the 97.5th percentile of its measured null, the values this same statistic returns on simulated cells whose score carries nothing beyond its primitive. Read in the direction each score's own correlation with the stages runs, as our registered rule does, six of the nine clear, and that reading stays primary. Read post hoc in the direction each score's authors give it, as published benchmarks do, six clear as well, and four clear under both readings: CytoTRACE, NCG, whose direction we infer because its authors state none, dpath and StemFinder. The four that clear under one reading only are built on the degree-correlation primitive, which rates the later zebrafish stages more potent. In sorted haematopoietic progenitors the gene-count primitive orders cells below chance (area under the receiver operating characteristic curve, AUROC, 0.378), and the standard comparison rewards a score for beating that reversed ordering. Our own strong hypothesis, that every unsupervised potency score reduces to a fixed low-order statistic, is refuted under either reading: CytoTRACE keeps τ_b 0.288 after four primitives are removed at once.

Keywords: single-cell RNA-seq, potency scores, incremental validity, null calibration, benchmarking

Data and Code Availability. This study uses only previously published, publicly available scRNA-seq data from the Gene Expression Omnibus: GSE106474 (zebrafish embryogenesis, twelve stages), GSE113074 (Xenopus embryogenesis, ten stages), GSE117498 (sorted human haematopoiesis), GSE125970 (human intestine), GSE150949 and the olaparib and untreated arms of GSE223003 (drug-tolerant persister systems), together with STRING v12 for the interaction scaffold. Accessions, download commands and quality-control thresholds are in the appendix. All analysis code, the score implementations, the 200-seed null-calibration grid, the run records the figures read and a command-line implementation of the test are public under the MIT licence at `https://github.com/IskakovDamir/ownbaseline`; the numbers in this version are those of commit `e44b9d8`. The repository carries a single reproduction entry point with an explicit coverage ledger: the primitives, CytoTRACE v1 and SCENT's SR and CCAT regenerate end to end from a public download, and every row that cannot regenerate names its reason rather than being skipped. A test suite runs inside a clone.

Institutional Review Board (IRB). This research uses publicly available, de-identified data from

© D. Iskakov.

previously published studies, generated no new human or animal data, and accessed no identifiable human subject data. It therefore does not require IRB approval or an equivalent ethics review, and none was sought. Relevant confirmation will be provided if the paper is accepted.

1. Why this is a health problem and not a developmental one

Unsupervised potency scores were built to rank cells by differentiation potential without markers or timepoints, and they are now applied where markers are unavailable or untrustworthy: candidate cancer stem cells, pre-treatment states that survive therapy, plastic phenotypes in circulating tumour cells. The authors of SCENT (single-cell entropy) validated signalling entropy, labelled SR in their software and in our figures, on exactly this kind of claim, reporting higher entropy in relapsed acute myeloid leukaemia than in matched diagnostic samples across 19 paired bulk samples, and, at a significance they themselves call marginal, in circulating tumour cells from patients progressing under androgen receptor inhibition (Teschendorff and Enver, 2017). CytoTRACE is used the same way (Gulati et al., 2020). The question of what these scores measure is therefore a question about which cells get called resistant.

There is a specific failure mode. Cells with more detected genes tend to receive higher scores from most of these methods, and this detection depth, the number of genes detected per cell, varies with cell type, with dissociation protocol and with sorting. The same dependence is established for the neighbouring family of signature scores, where cancer cells detect more genes than normal cells across ten datasets and that imbalance biases every scoring method tested (Noureen et al., 2022). When depth happens to track the biological ordering, the score looks validated. When it opposes it, the failure is invisible to a correlation test. Neither case is diagnosed by a pairwise correlation, the score's correlation with a gold standard taken on its own, because the primitive, the low-order statistic a score is built on (here the count of detected genes), correlates with the gold standard too. Figure 1 shows both halves. In a simulation, two scores reach the same Kendall τ_b of 0.404 with a twelve-level ordering, one of them depth plus noise and the other carrying ordering that depth does not (panel a). In real tissue, depth runs against potency in sorted blood progenitors and with it in intestinal epithelium (panels b and c, section 2).

The question of whether a new measure adds anything beyond the measures already in use is old, and psychometrics settled on an answer: hierarchical regression, with the increment tested directly rather than inferred from the new measure's raw correlation with the criterion (Sechrest, 1963; Hunsley and Meyer, 2003). Genomics reached the same position independently, for prognostic signatures built on omics data (De Bin et al., 2014). Raw correlations cannot answer it, because they do not reflect the overlap between the new measure and the existing ones. The same criterion has recently done work in this field. Seven deep models for perturbation prediction fail to beat four deliberately simple baselines, and their authors read that as a failure of the evaluations to establish an increment (Ahlmann-Eltze et al., 2025). We are transferring that criterion into single-cell potency scoring rather than inventing one. Residualizing one variable on another and testing the remainder is a conditional independence test, and the modern treatment of it is the generalised covariance measure (Shah and Peters, 2020) and its assumption-lean successor (Lundborg et al., 2024). That literature carries a caveat this paper inherits: conditional independence testing admits no uniformly valid test. A verdict here is evidence against one named reduction, and weaker than a claim of independence. What is specific here is the estimator: rank residualization followed by a single Kendall τ_b, a sign-agnostic alignment step, and a decision taken against the estimator's own measured null rather than against zero, each defined at the end of this section. Calibrating a decision against an empirically measured null rather than a theoretical one is itself borrowed, from observational epidemiology (Schuemie et al., 2014).

The test needs a ground truth that is fine-grained, ordered, and fixed independently of any expression-derived score. Staged embryos supply one. The 28 Drop-seq samples of GSE106474 were collected from "20-40 visually staged" embryos each, so the stage label is assigned by morphology before dissociation and cannot be contaminated by the transcriptome it is used to judge, and it comes in twelve ordered levels rather than two. We calibrate on development and audit the methods that are deployed on patients.

Terms and the test. The *ordinal* is a ground-truth ordering of cells into a few ordered levels, assigned without reference to expression; here it is the

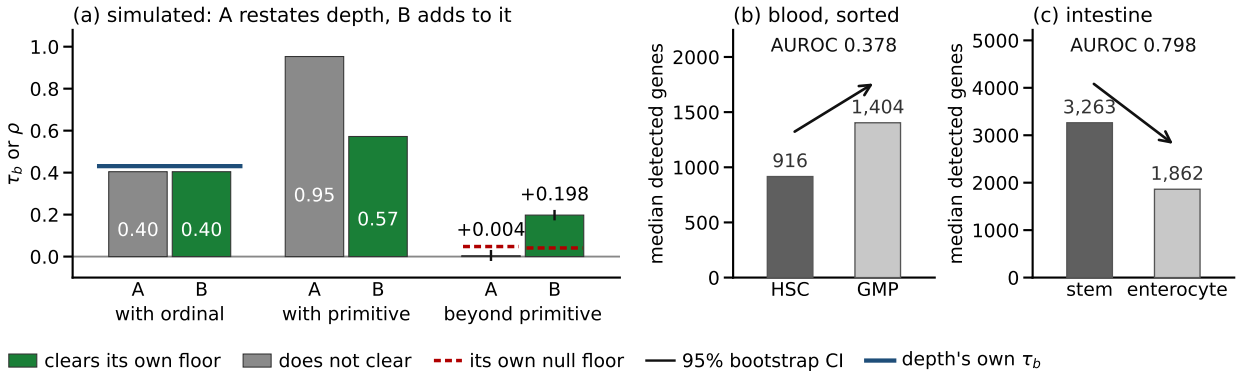


Figure 1: A correlation with the gold standard passes a score that restates its primitive, and depth decides which way the primitive points. (a) Simulated cells ($n = 3{,}000$) whose twelve-level ordinal is cut from a weighted sum of depth and a biological position independent of depth (Appendix D): A is depth plus noise, B adds part of that position, and both reach τ_b 0.404 with the ordinal. Each score has three bars, its τ_b with the ordinal, its Spearman ρ with depth and its conditional skill, Eq. (3); only the last carries a 95% bootstrap interval and, dashed, its floor, Eq. (4). The blue line is depth's own τ_b, and green marks a score that clears its floor. (b) Sorted haematopoietic progenitors (Pellin et al., 2019): haematopoietic stem cells (HSC) and granulocyte-monocyte progenitors (GMP). (c) Intestinal epithelium (Wang et al., 2020): stem cells and enterocytes. Bars are median detected genes, the arrow running from the more potent population. The AUROC, the area under the receiver operating characteristic curve, is the probability that a cell of the top potency tier has more detected genes than one of the bottom tier, a tier holding more populations than the one plotted (Appendix D).

stage of the embryo a cell came from. A score's *primitive* is the low-order statistic its own authors say it is built on or approximates. It is the score's own baseline, and the *own-baseline test* asks whether the score orders cells beyond it. Four primitives recur in this audit and form its joint set: the number of detected genes; the degree correlation PCC(x, degree), the Pearson correlation between a cell's expression and the protein-interaction degree of each gene; the Shannon entropy of the expression vector; and library size, the total count per cell. StemFinder declares a fifth, the mean expression of cell-cycle genes. *Rank residualization* removes from a score what a least-squares fit on the ranks of its primitive explains. For n cells with score s, primitives $p_1, \ldots, p_k$, ordinal o and within-sample ranks $R(\cdot)$, ties averaged,

$$\tilde{s} = \operatorname{sign}\{\rho_S(s, o)\}\, s, \tag{1}$$

$$r = R(\tilde{s}) - X\hat{\beta}, \quad X = [\mathbf{1}, R(p_1), \ldots, R(p_k)], \tag{2}$$

$$\mathrm{CS} = \tau_b(r, o) = \frac{n_c - n_d}{\sqrt{(n_0 - t_r)(n_0 - t_o)}}, \tag{3}$$

with $\operatorname{sign}(0) = 1$, $\hat{\beta}$ the least-squares coefficients, ρ_S Spearman's correlation, n_c and n_d the concordant and discordant pairs of cells, $n_0 = n(n-1)/2$ all pairs, and t_r and t_o the pairs tied in r and in o. Equation (1), the alignment step, points each score the way its marginal correlation with the ordinal runs, so the sign its authors chose does not matter. *Conditional skill* is CS, the ordering a score carries beyond its primitive. It is direction-free: Eq. (1) credits ordering in the direction the score runs on the data, which need not be the direction its authors give it, and section 3 reads every row both ways. We call the rank statistic applied to r the *kernel*. Kendall's τ_b is the primary kernel; a weighted Kendall tau that gives the top of a ranking more weight (Vigna, 2015) is the second (Appendix I). A score's *marginal skill* is $\tau_b(s, o)$ with o running toward higher potency. The *marginal gap*, a score's marginal skill minus its primitive's, is the comparison the field already makes, and the *direction check* asks whether a primitive or a score orders cells toward potency at all.

A value of CS is judged against the *measured null*: the distribution the same estimator returns on simulated cells whose ordinal is drawn from the primitive and whose score is the primitive plus independent noise, so that its true conditional skill is zero (Appendix D). A value counts when

$$\mathrm{CS} > f(n, \rho, k, L) = Q_{0.975}\big[\mathrm{CS}^*\big], \tag{4}$$

the 97.5th percentile of the null statistic CS^* over 200 draws, at the point of the measured null grid nearest the value's sample size n and score-primitive Spearman correlation $|\rho|$, for its number of primitives k and of ordinal levels L; for $k = 1$ it is the largest of three such percentiles, one at each *coupling* $c \in \{0.45, 0.55, 0.63\}$, the strength with which the null draws the ordinal from the primitive (Appendix D). We call f the value's *floor*. The interval printed beside a value is a percentile bootstrap over cells, 1,000 resamples with r refitted in each; it measures the precision of CS and takes no part in the decision. We call Eqs. (1) to (4) the *registered rule*, and a value read under them the *registered reading*: they were registered, unchanged, before the last two of the nine scores we place on the stages (section 3) met a stage label (Appendix F); for the first seven, Eq. (4) replaced an earlier decision rule after their values were known (Appendix I).

2. A primitive that orders below chance

Take haematopoietic progenitors sorted by fluorescence-activated cell sorting, whose potency ordering is the conventional one (Pellin et al., 2019), ranked into four tiers (Appendix D). Haematopoietic stem cells (HSC) carry a median of 916 detected genes. Granulocyte-monocyte progenitors (GMP), which sit below them, carry 1,404 (Figure 1b). Across that pair detection depth runs opposite to potency. The low-potency tier is itself heterogeneous: megakaryocyte-erythroid progenitors (MEP) sit at 882, below the stem compartment. Across the ranked populations the profile is non-monotone throughout, at 916 for HSC, 1,045 for multipotent progenitors (MPP), 1,018 for the broad CD34+/CD164+ gate, 1,359 for multi-lymphoid progenitors (MLP), 992 for precursor B and natural killer cells (Pre-B/NK), 882 for MEP, 1,063 for common myeloid progenitors (CMP) and 1,404 for GMP, so the confound is a property of specific population contrasts rather than a monotone reversal, which is why it is invisible to a summary correlation.

Two concordance measures agree on the sign. The AUROC, the probability that a randomly drawn cell of the top tier scores above one of the bottom tier, compares the 1,282 HSC with the 2,815 cells of the bottom tier, GMP, MEP and Pre-B/NK, and the gene-count primitive scores 0.378 on it. Over all 12,354 ranked cells its Kendall τ_b is -0.089, so the direction holds under either.

An AUROC of 0.378 is worse than a coin, and it carries as much ordering information as an equally distant value above 0.5, pointed the wrong way. Here CytoTRACE is represented by its gene-count signature, without its regression and diffusion steps (Appendix D). The field's comparison is directional, so, scored the usual way, CytoTRACE beats the primitive by a marginal gap of 0.181 and reads as evidence that the score adds biology. What it adds, in τ_b over all four tiers, is a sign correction. The gap is the distance between two orderings that are both wrong: the score sits at $+0.092$ in marginal skill and the primitive at -0.089, against a hierarchy neither of them recovers.

Rank residualization is orientation-invariant. The primitive enters as a rank covariate, and one at rank correlation $-\rho$ removes exactly as much as one at $+\rho$, so the apparent margin does not survive as a margin. The conditional skill alone does not settle this comparison either. It is positive here, 0.195 with a bootstrap interval of 0.182 to 0.210, and it clears every floor this calibration measured under τ_b below $\rho = 1$, with alignment on or off, the largest of which is 0.0841. That is a bound over all measured floors and not a matched test (Appendix E). The marginal gap and the conditional skill both read, taken alone, as skill beyond the primitive. What exposes the trap is the direction check: score and primitive sit at AUROC 0.269 and 0.378, both below 0.5, so between the stem compartment and the bottom tier the ordering they agree on is the reversed one. For the score itself the two measures part ways, τ_b $+0.092$ over all four tiers against AUROC 0.269 at the two ends. For a depth-confounded comparison the verdict rests on the conjunction of the marginal gap, the direction check and the conditional skill, and that conjunction differs from the decision rule behind the audit's verdicts (Appendix I).

We are not claiming CytoTRACE is uninformative. On the staged ordinal it is the strongest score we measured. We are claiming that this experiment, run the

standard way, provides no evidence either way, and that a reader cannot tell from the reported number.

Intestinal epithelium gives the other half (Wang et al., 2020). There the stem compartment carries 3,263 detected genes against 1,862 for enterocytes (Figure 1c), depth and hierarchy agree, and the same primitive reaches an AUROC of 0.798. Nothing about the primitive changed between the two tissues. What changed is whether dissociation and sorting happened to put the deeper cells at the top. A criterion whose verdict depends on that coincidence is measuring the coincidence.

A transfer test across cancers can mislead in a similar way. Between PC9 lung adenocarcinoma under osimertinib and Kuramochi ovarian carcinoma under olaparib, a rule fitted on gene count in one system falls 0.182 and 0.080 below the other system's own rule in AUROC on binary fate labels, while SR and CCAT, two scores whose primitive is the degree correlation, land within +0.001 to +0.008 of theirs. That agreement is arithmetic: a one-variable rule carries only the sign of its slope to the other system, so it matches the target's own rule when that sign agrees and falls below 0.5 when it flips, as gene count's does, and a degree vector permuted the same way in both systems gives CCAT no transfer at all (Appendix H).

3. The spectrum

A categorization of this field sorts these methods into average-based, entropy-based and correlation-based constructions and concludes that no benchmark exists at single-cell resolution, because published ground truths label cell groups rather than individual cells (Wang et al., 2024). That is the gap this audit works in and it is also a constraint on it, since our ordinal labels embryos. Fourteen unsupervised scores were audited and they do not all reach the same place, so the denominators matter. Five were not placed: SPIDE and scEnergy carry no measurement, StemID and cmEntropy (Grün et al., 2016; Kannan et al., 2021) equal their own Shannon-entropy primitive by construction, and StemSC (Zheng et al., 2022) was measured on the sorted atlases only (Appendix A). Nine scores were placed on the staged ordinal. Two of them, MCE and StemFinder, were run after the other seven, under a plan registered before either met a stage label (Appendix F). MCE was run from a Python port of its authors' MATLAB code; on 500 cells the port's scores differ from the original's by at most 1.8×10^{-13}. StemFinder, which scores how unevenly a cell's nearest neighbours express cell-cycle genes, was run from its authors' R code, and those authors chose among variants of the method on a benchmark that includes these embryos (Noller and Cahan, 2024). Three scores sit outside the unsupervised class and enter no count: CytoTRACE 2 is supervised, and mRNAsi and FitDevo carry gene weights fitted to labelled stem-cell samples and to timepoint labels (Malta et al., 2018; Zhang et al., 2022). Figure 2 draws the nine and Figure A1 all seventeen rows; Table A1 lists every score with its class, its primitive and the direction its authors give it.

We did not choose the primitives. The one approximate pairing is dpath's: its authors build on "metagene entropy" over a non-negative matrix factorization learned from the data (Gong et al., 2017), and we residualize dpath on the Shannon entropy of the raw expression vector.

Six of the nine carry skill beyond their primitive, each above its own measured null under τ_b: CytoTRACE at 0.653, signalling entropy at 0.430, ORIGINS (Senra et al., 2022) at 0.402, NCG at 0.402, dpath at 0.120 and StemFinder at 0.098. The two 0.402s are different measurements. NCG's is on a 3,000-cell subsample, and NCG computes its score on a gene-ontology-weighted connectome rather than STRING, so their agreement is a coincidence of rounding. StemFinder clears its floor of 0.0247 by 3.95 times. dpath is the narrowest at 2.9 times the 97.5th percentile of its own null at that sample size, 0.0411, and it is the row that fails under the weighted kernel. The three that do not clear are CCAT, SLICE and MCE, and all three fail for the same reason: their residuals are negative. CCAT (Teschendorff et al., 2021), the correlation of connectome and transcriptome, reproduces its primitive at $\rho = 0.998$ and leaves -0.221 against a floor of 0.0115 (Figure 2). SLICE (Guo et al., 2017) tracks its own entropy primitive at $\rho = 0.932$, also on the subsample, and leaves -0.113 against a floor of 0.0479. MCE tracks PCC(x, degree) at $\rho = 0.802$ and leaves -0.490 against a floor of 0.0325. A negative value does not exceed a positive floor. For CCAT and MCE the sign is the alignment step's, as the next paragraph shows; what SLICE's negative value means is a further question this design does not answer.

Direction. The alignment step lets a score earn skill in whichever direction it runs, while published benchmarks of these scores fix the direction from the method's definition and count a reversal as a failure

(Gulati et al., 2020; Teschendorff et al., 2021; Kang et al., 2025; Noller and Cahan, 2024). We therefore read every row a second way, post hoc and labelled so (Appendix G): the residual of Eq. (2) with the authors' polarity counted positive, against the floor of the same null without alignment. Four rows reverse. Their shared primitive PCC(x, degree) rises with stage on these embryos, at a Spearman of 0.516 toward the later stages, and signalling entropy, ORIGINS and CCAT rise with it, at 0.692, 0.712 and 0.496, against their authors' polarity; MCE leans the same way at 0.035, close to zero (section 6). NCG, residualized on the same primitive, runs the other way. Read in their authors' direction, signalling entropy and ORIGINS fall to −0.430 and −0.402, and CCAT and MCE rise to 0.221 and 0.490 against floors of 0.0120 and 0.0342. Six rows clear under each reading, and four clear under both: CytoTRACE, NCG, dpath and StemFinder. NCG's authors state no direction, so its polarity is inferred from its construction; read the other way it would fail, leaving three that clear under both. The registered reading stays primary.

CytoTRACE's 0.653 averages a curve. Restricted to cell pairs from adjacent stages, its concordance is +0.318, and the score orders three of the eleven single-stage transitions backwards (Table A2).

CCAT sits at $\rho = 0.998$, close to $\rho = 1$, where a score's ranks equal its primitive's, the least-squares fit is exact and the residual is rounding error of order 10^{-12} that a kernel still scores (Appendix I). CCAT's largest residual, 3,937 on a rank scale of 39,505, is 4.5×10^{14} times machine precision times the cell count, so the problem does not reach this row.

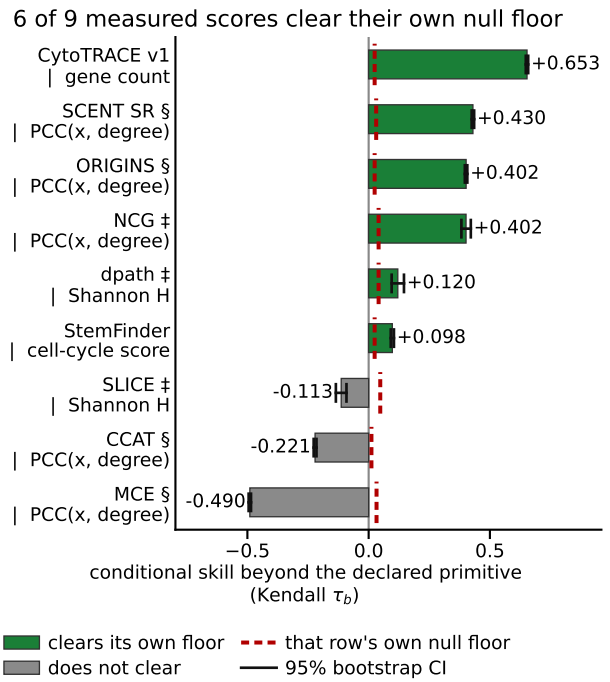


Figure 2: Field audit: six of the nine scores placed on the staged ordinal carry ordering skill beyond their own declared primitive. Each bar is a score's conditional skill, Eq. (3), on the primitive its authors declare (named under the score), against the twelve stages of GSE106474, 39,505 cells. Whiskers are 95% bootstrap intervals. The dashed mark on each bar is that row's own floor, Eq. (4). Green bars clear theirs and grey bars do not. A bar counts ordering in the direction the score's own marginal correlation with the stages runs, Eq. (1), so a negative bar is a residual running against it; § marks the four scores whose marginal ordering runs against their authors' polarity (section 3). SCENT SR is signalling entropy. NCG, dpath and SLICE (‡) were measured on a 3,000-cell subsample. The eight rows without a value are in Figure A1.

4. Surviving four primitives at once

Residualizing on one primitive at a time is a weak control. A score built from gene count and library size together can pass every single-primitive test and still carry no biology. So we residualized CytoTRACE, signalling entropy, ORIGINS and dpath on the four primitives jointly: gene count, the degree correlation, Shannon entropy and log library size.

All four survive, each above its own floor at $k = 4$: CytoTRACE falls from 0.653 to 0.288 against 0.0171, signalling entropy from 0.430 to 0.183 against 0.0227, ORIGINS from 0.402 to 0.156 against 0.0171, and dpath moves from 0.120 to 0.134 against 0.0325. NCG was own-baselined separately, from its own implementation on the gene-ontology-weighted connectome, and it also adds beyond its primitive and survives depth adjustment, a residualization on the primitive and log library size together. Residualized on the same four primitives, which exclude StemFinder's cell-cycle primitive, StemFinder keeps

0.167 against a floor of 0.0171, and MCE leaves −0.131 against 0.0233; StemFinder was not residualized on all five at once. In their authors' direction the joint residuals of signalling entropy and ORIGINS turn negative and MCE's positive (Appendix G).

Three further checks are in Appendix I: an isotonic fit in place of least squares changes no verdict; a score is called substantive only when it clears its floor and survives depth adjustment, in place of a family-wise correction; and the weighted kernel, whose null mean is far from zero, does not decide, though a rule requiring both kernels would fail dpath, 0.191 against 0.242, and give five of nine.

A second staged ordinal, in Xenopus, replicates the degree-correlation result and not the gene-count or entropy ones (Briggs et al., 2018). Residualizing on gene count may remove some biology, since total mRNA content is prognostic (Cao et al., 2022) and raised in stem cells (Kim et al., 2023); Appendix I takes up both.

The residual is real. What it consists of is open. A staged timecourse traces a developmental manifold, and a score tracking the manifold keeps skill whether or not it measures potency in any deeper sense. We mark this as the question we did not answer.

5. The authors got there first

None of these reductions are adversarial readings. CCAT is defined as the Pearson correlation between a cell's transcriptome and protein-protein interaction (PPI) node degree, and its authors define it that way in the paper introducing it (Teschendorff et al., 2021). The NCG authors build their score on network topology directly (Ni et al., 2022). Shi and colleagues put approximately 70 percent of MCE's variance down to the same correlation (Shi et al., 2020), and Teschendorff and Enver report that this correlation approximates signalling entropy at $R^2 = 0.96$ (Teschendorff and Enver, 2017). The gene-count dependence of CytoTRACE is conceded by its own successor, which reports that gene counts "are subject to dataset-specific biases, making them suboptimal for potency assessment" (Kang et al., 2025). Others treat the gene count as the quantity to improve on (Lin et al., 2024), and Kannan and colleagues showed the depth inflation of transcriptomic entropy and controlled it by keeping the top 1,000 expressed genes five years before this audit, but did not re-test the controlled score against an ordinal fixed outside the transcriptome (Kannan et al., 2021).

The R-squared report carries a caveat, and the caveat is correct: the approximation "is not a substitute for SR", and in networks where hubs are not connected to one another a positive correlation can yield a lower signalling entropy.

Our measurement is the single-cell test of that caveat, and on equivalence it comes out in the authors' favour: signalling entropy keeps conditional skill after joint residualization, so it does not restate the degree correlation. What it keeps orders these embryos toward the later stages, against the direction its authors give the score (section 3). The caveat does not extend to CCAT, which its authors define as the degree correlation itself. CCAT as its authors compute it and our realization of that statistic agree at $\rho = 0.998$, and the residual between them orders the stages at -0.221 under the registered reading and $+0.221$ in CCAT's authors' direction (section 3). Pairwise correlation cannot tell signalling entropy from CCAT, because both correlate strongly with the same primitive and with the gold standard.

6. What our own test refuted

We started from a stronger claim than the one we report. The hypothesis was that every unsupervised potency score reduces to a fixed low-order statistic, and that the whole family is tautological. The test we built to demonstrate that refuted it.

It holds for part of the family. StemID and cmEntropy equal a Shannon-entropy primitive by construction, and CCAT is the degree correlation by definition, although the residual between CCAT and our realization of it depends on the reading, as MCE's does. For the other part the hypothesis fails under either reading. CytoTRACE, a diffusion-smoothed gene count, runs in its authors' direction and keeps 0.653 against a floor of 0.0247, and 0.288 against 0.0171 after four primitives are removed at once. StemFinder, a neighbourhood statistic over cell-cycle genes, keeps 0.098 against 0.0247 on its cell-cycle primitive and 0.167 against 0.0171 on the four, two separate controls. The reduction holds by construction for StemID and cmEntropy and is false for the four scores that clear under both readings, at single-cell resolution, on this ordinal. The warrant for that refutation is that those residuals exceed the estimator's own measured null under τ_b, the criterion that replaced our earlier decision rule (Appendix I). SLICE belongs to neither part: it improves on its entropy primitive under neither reading, and its

residual, −0.113 against floors of 0.0479 (registered) and 0.0453 (authors' polarity), runs toward the later stages, which this design does not interpret.

The entropy-rate constructions split. Before MCE met a stage label we registered that MCE, an entropy rate on the interaction network plus a static entropy (Appendix F), would keep conditional skill above its floor. Under the registered rule it does not: −0.490 against 0.0325. MCE's marginal correlation with the stages is a Spearman of 0.035, so the alignment step takes its sign from almost nothing, and the large residual runs the other way, toward the earlier stages and the direction MCE's authors give it. Signalling entropy, an entropy rate over a weighted network, is the mirror case: its residual of 0.430 runs toward the later stages, against its authors' direction. Each clears its floor under one reading only, signalling entropy under the registered one at 0.430 against 0.0311 and MCE under its authors' polarity at 0.490 against 0.0342, a post hoc reading (Appendix G). Which of the two the audit credits therefore depends on how a score's direction is read, and the registered reading credits signalling entropy.

7. Limitations

The calibration is developmental and the scores are deployed on patients. That gap is deliberate and it is still a gap. We measured conditional skill where a score-independent fine ordinal exists, and we infer that a score failing there will not succeed where the ordinal is coarser and partly derived from expression. The inference is not a measurement.

The ordinal is assigned per sample rather than per cell. A six-somite embryo still contains high-potency progenitors, so a score that recovered potency perfectly would not order cells by embryo stage perfectly. The test is conservative for a failure and permissive for a success, and every positive row here is consistent with a score that tracks developmental position and no more. The coarseness is shared: score and primitive are scored on the same pairs under the same labels, so it cannot manufacture a gap between them. The order of the twelve levels is our reading of the staging series rather than the accession's listing, and Appendix C reports the audit re-run on the accession's order.

Residualization controls a realization of a primitive, not the construct behind it. Our degree primitive is computed on one scaffold, STRING v12 at a confidence of 700 on its 0 to 1,000 scale, largest connected component, with threshold and interactome fixed, so part of any surviving residual may be scaffold error rather than biology. NCG and ORIGINS are computed on networks of their own, and part of their residuals may reflect that difference.

The floor is a property of the estimator, measured on synthetic data, and a score can clear its own and still track something the nulls do not represent (Appendix D).

Five of the fourteen audited scores were never placed on the ordinal (Table A1); of the two without a measurement, scEnergy has no implementation we could run and SPIDE has one we did not run. The joint primitive set is four statistics, with StemFinder's cell-cycle score a fifth declared one, and a score reducing to a statistic outside these would pass our test and should not.

8. What to do instead

Report the own-baseline next to the score, and report the estimator's null floor at your n, kernel and ρ. Compute the low-order statistic the score is closest to, order the cells by that statistic alone, and publish both numbers. The comparison costs one line and it is the difference between a score that carries information and a score that restates detection depth.

Check the direction before reading the margin. A primitive that orders below chance turns a beaten baseline into a sign correction, and nothing in a raw margin shows this. Fix each score's direction from its definition before it meets the labels: on these embryos the degree correlation itself runs backwards (section 3).

Stop treating correlation with a pseudotime or a marker panel as validation. The primitive correlates with those too. Both scores of Figure 1a pass that test.

Cancer stem-cell and drug-tolerant-state calls made from these scores are downstream of whether the score beat its own baseline. On the zebrafish stages, six of nine did under the registered reading, six did when read post hoc in their authors' direction, and four did under both; five were never placed.

Acknowledgments

We thank Qingyang Wang for pointing us to the MCE code and to three methods the audit had missed, StemFinder, mRNAsi and FitDevo, and for com-

ments on the terminology, the notation and the figures of an earlier version.

References

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. doi: 10.1038/s41592-025-02772-6.

James A. Briggs, Caleb Weinreb, Daniel E. Wagner, Sean Megason, Leonid Peshkin, Marc W. Kirschner, and Allon M. Klein. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*, 360(6392):eaar5780, 2018. doi: 10.1126/science.aar5780.

Shaolong Cao, Jennifer R. Wang, Shuangxi Ji, Peng Yang, Yaoyi Dai, Shuai Guo, Matthew D. Montierth, John Paul Shen, Xiao Zhao, Jingxiao Chen, Jaewon James Lee, Paola A. Guerrero, Nicholas Spetsieris, Nikolai Engedal, Sinja Taavitsainen, Kaixian Yu, Julie Livingstone, Vinayak Bhandari, Shawna M. Hubert, Najat C. Daw, P. Andrew Futreal, Eleni Efstathiou, Bora Lim, Andrea Viale, Jianjun Zhang, Matti Nykter, Bogdan A. Czerniak, Powel H. Brown, Charles Swanton, Pavlos Msaouel, Anirban Maitra, Scott Kopetz, Peter Campbell, Terence P. Speed, Paul C. Boutros, Hongtu Zhu, Alfonso Urbanucci, Jonas Demeulemeester, Peter Van Loo, and Wenyi Wang. Estimation of tumor cell total mRNA expression in 15 cancer types predicts disease progression. *Nature Biotechnology*, 40(11):1624–1633, 2022. doi: 10.1038/s41587-022-01342-x.

Jun Dai, Shuyu Zheng, Matías M. Falco, Jie Bao, Johanna Eriksson, Sanna Pikkusaari, Sofia Forstén, Jing Jiang, Wenyu Wang, Luping Gao, Fernando Perez-Villatoro, Olli Dufva, Khalid Saeed, Yinyin Wang, Ali Amiryousefi, Anniina Färkkilä, Satu Mustjoki, Liisa Kauppi, Jing Tang, and Anna Vähärautio. Tracing back primed resistance in cancer via sister cells. *Nature Communications*, 15:1158, 2024. doi: 10.1038/s41467-024-45478-7.

Riccardo De Bin, Tobias Herold, and Anne-Laure Boulesteix. Added predictive value of omics data: specific issues related to validation illustrated by two case studies. *BMC Medical Research Methodology*, 14:117, 2014. doi: 10.1186/1471-2288-14-117.

Jeffrey A. Farrell, Yiqun Wang, Samantha J. Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F. Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392):eaar3131, 2018. doi: 10.1126/science.aar3131.

Wuming Gong, Tara L. Rasmussen, Bhairab N. Singh, Naoko Koyano-Nakagawa, Wei Pan, and Daniel J. Garry. Dpath software reveals hierarchical haemato-endothelial lineages of Etv2 progenitors based on single-cell transcriptome analysis. *Nature Communications*, 8:14362, 2017. doi: 10.1038/ncomms14362.

Dominic Grün, Mauro J. Muraro, Jean-Charles Boisset, Kay Wiebrands, Anna Lyubimova, Gitanjali Dharmadhikari, Maaike van den Born, Johan van Es, Erik Jansen, Hans Clevers, Eelco J. P. de Koning, and Alexander van Oudenaarden. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell*, 19(2):266–277, 2016. doi: 10.1016/j.stem.2016.05.010.

Gunsagar S. Gulati, Shaheen S. Sikandar, Daniel J. Wesche, Anoop Manjunath, Anjan Bharadwaj, Mark J. Berger, Francisco Ilagan, Angera H. Kuo, Robert W. Hsieh, Shang Cai, Maider Zabala, Ferenc A. Scheeren, Neethan A. Lobo, Dalong Qian, Feiqiao B. Yu, Frederick M. Dirbas, Michael F. Clarke, and Aaron M. Newman. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*, 367(6476):405–411, 2020. doi: 10.1126/science.aax0249.

Minzhe Guo, Erik L. Bao, Michael Wagner, Jeffrey A. Whitsett, and Yan Xu. SLICE: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Research*, 45(7):e54, 2017. doi: 10.1093/nar/gkw1278.

John Hunsley and Gregory J. Meyer. The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment*, 15(4):446–455, 2003. doi: 10.1037/1040-3590.15.4.446.

Suoqin Jin, Adam L. MacLean, Tao Peng, and Qing Nie. scEpath: energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data. *Bioinformatics*, 34(12):2077–2086, 2018. doi: 10.1093/bioinformatics/bty058.

Minji Kang, Gunsagar S. Gulati, Erin L. Brown, Zhen Qi, Susanna Avagyan, Jose Juan Almagro Armenteros, Rachel Gleyzer, Wubing Zhang, Chloé B. Steen, Jeremy Philip D'Silva, Janella Schwab, Michael F. Clarke, Aadel A. Chaudhuri, and Aaron M. Newman. Improved reconstruction of single-cell developmental potential with CytoTRACE 2. *Nature Methods*, 22(11):2258–2263, 2025. doi: 10.1038/s41592-025-02857-2.

Suraj Kannan, Michael Farid, Brian L. Lin, Matthew Miyamoto, and Chulan Kwon. Transcriptomic entropy benchmarks stem cell-derived cardiomyocyte maturation against endogenous tissue at single cell level. *PLOS Computational Biology*, 17 (9):e1009305, 2021. doi: 10.1371/journal.pcbi.1009305.

Yun-Kyo Kim, Brandon Cho, David P. Cook, Dan Trcka, Jeffrey L. Wrana, and Miguel Ramalho-Santos. Absolute scaling of single-cell transcriptomes identifies pervasive hypertranscription in adult stem and progenitor cells. *Cell Reports*, 42(1):111978, 2023. doi: 10.1016/j.celrep.2022.111978.

Charles B. Kimmel, William W. Ballard, Seth R. Kimmel, Bonnie Ullmann, and Thomas F. Schilling. Stages of embryonic development of the zebrafish. *Developmental Dynamics*, 203(3):253–310, 1995. doi: 10.1002/aja.1002030302.

Nikolaus Kriegeskorte, W. Kyle Simmons, Patrick S. F. Bellgowan, and Chris I. Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5):535–540, 2009. doi: 10.1038/nn.2303.

Hai Lin, Huan Hu, Zhen Feng, Fei Xu, Jie Lyu, Xiang Li, Liyu Liu, Gen Yang, and Jianwei Shuai. SCTC: inference of developmental potential from single-cell transcriptional complexity. *Nucleic Acids Research*, 52(11):6114–6128, 2024. doi: 10.1093/nar/gkae340.

Anton Rask Lundborg, Ilmun Kim, Rajen D. Shah, and Richard J. Samworth. The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52(6), 2024. doi: 10.1214/24-AOS2447.

Tathiane M. Malta, Artem Sokolov, Andrew J. Gentles, et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell*, 173(2):338–354.e15, 2018. doi: 10.1016/j.cell.2018.03.034.

Xinzhe Ni, Bohao Geng, Haoyu Zheng, Jiawei Shi, Gang Hu, and Jianzhao Gao. Accurate estimation of single-cell differentiation potency based on network topology and gene ontology information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(6):3255–3262, 2022. doi: 10.1109/TCBB.2021.3112951.

Kathleen Noller and Patrick Cahan. Cell cycle expression heterogeneity predicts degree of differentiation. *Briefings in Bioinformatics*, 25(6):bbae536, 2024. doi: 10.1093/bib/bbae536.

Nighat Noureen, Zhenqing Ye, Yidong Chen, Xiaojing Wang, and Siyuan Zheng. Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data. *eLife*, 11:e71994, 2022. doi: 10.7554/eLife.71994.

Yaara Oren, Michael Tsabar, Michael S. Cuoco, Liat Amir-Zilberstein, Heidie F. Cabanos, Jan-Christian Hütter, Bomiao Hu, Pratiksha I. Thakore, Marcin Tabaka, Charles P. Fulco, William Colgan, Brandon M. Cuevas, Sara A. Hurvitz, Dennis J. Slamon, Amy Deik, Kerry A. Pierce, Clary Clish, Aaron N. Hata, Elma Zaganjor, Galit Lahav, Katerina Politi, Joan S. Brugge, and Aviv Regev. Cycling cancer persister cells arise from lineages with distinct programs. *Nature*, 596(7873):576–582, 2021. doi: 10.1038/s41586-021-03796-6.

Danilo Pellin, Mariana Loperfido, Cristina Baricordi, Samuel L. Wolock, Annita Montepeloso, Olga K. Weinberg, Alessandra Biffi, Allon M. Klein, and Luca Biasco. A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nature Communications*, 10:2395, 2019. doi: 10.1038/s41467-019-10291-0.

Martijn J. Schuemie, Patrick B. Ryan, William DuMouchel, Marc A. Suchard, and David Madigan. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in Medicine*, 33(2):209–218, 2014. doi: 10.1002/sim.5925.

Lee Sechrest. Incremental validity: A recommendation. *Educational and Psychological Measurement*, 23(1):153–158, 1963. doi: 10.1177/001316446302300113.

Daniela Senra, Nara Guisoni, and Luis Diambra. ORIGINS: A protein network-based approach to quantify cell pluripotency from scRNA-seq data. *MethodsX*, 9:101778, 2022. doi: 10.1016/j.mex.2022.101778.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3), 2020. doi: 10.1214/19-AOS1857.

Jifan Shi, Andrew E. Teschendorff, Weiyan Chen, Luonan Chen, and Tiejun Li. Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures. *Briefings in Bioinformatics*, 21(1):248–261, 2020. doi: 10.1093/bib/bby093.

Andrew E. Teschendorff and Tariq Enver. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nature Communications*, 8:15599, 2017. doi: 10.1038/ncomms15599.

Andrew E. Teschendorff, Alok K. Maity, Xue Hu, Weiyan Chen, and Matthias Lechner. Ultra-fast scalable estimation of single-cell differentiation potency from scRNA-Seq data. *Bioinformatics*, 37(11):1528–1534, 2021. doi: 10.1093/bioinformatics/btaa987.

Sebastiano Vigna. A weighted correlation index for rankings with ties. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1166–1176, 2015. doi: 10.1145/2736277.2741088.

Qingyang Wang, Zhiqian Zhai, Qiuyu Lian, Dongyuan Song, and Jingyi Jessica Li. Categorization and analysis of 14 computational methods for estimating cell potency from single-cell RNA-seq data, 2024. arXiv:2309.13518v2; accepted at Cell Systems.

Yalong Wang, Wanlu Song, Jilian Wang, Ting Wang, Xiaochen Xiong, Zhen Qi, Wei Fu, Xuerui Yang, and Ye-Guang Chen. Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. *Journal of Experimental Medicine*, 217(2):e20191130, 2020. doi: 10.1084/jem.20191130.

Emily C. Willroth and Olivia E. Atherton. Best laid plans: A guide to reporting preregistration deviations. *Advances in Methods and Practices in Psychological Science*, 7(1):25152459231213802, 2024. doi: 10.1177/25152459231213802.

Feng Zhang, Chen Yang, Yihao Wang, Huiyuan Jiao, Zhiming Wang, Jianfeng Shen, and Lingjie Li. FitDevo: accurate inference of single-cell developmental potential using sample-specific gene weight. *Briefings in Bioinformatics*, 23(5):bbac293, 2022. doi: 10.1093/bib/bbac293.

Hailong Zheng, Jiajing Xie, Kai Song, Jing Yang, Huiting Xiao, Jiashuai Zhang, Keru Li, Rongqiang Yuan, Yuting Zhao, Yunyan Gu, and Wenyuan Zhao. StemSC: a cross-dataset human stemness index for single-cell samples. *Stem Cell Research & Therapy*, 13(1):115, 2022. doi: 10.1186/s13287-022-02803-5.

Ruiqing Zheng, Ziwei Xu, Yanping Zeng, Edwin Wang, and Min Li. SPIDE: A single cell potency inference method based on the local cell-specific network entropy. *Methods*, 220:90–97, 2023. doi: 10.1016/j.ymeth.2023.11.006.

Appendix A. The full audit

Table A1 lists the seventeen scores this audit considered. Fourteen are unsupervised and enter its counts; CytoTRACE 2 is supervised, and mRNAsi and FitDevo score cells with gene weights fitted to labelled samples. The categories are those of the categorization this audit works from (Wang et al., 2024), which places three of the scores in none.

CytoTRACE 2. CytoTRACE 2 escapes the fixed-statistic class on a pre-registered probe of three datasets, at a median margin of 0.175 over raw gene counts (Kang et al., 2025). We did not replicate that comparison at atlas scale. Across twenty-three datasets only published per-cell scores were available, so the baseline there is a proxy rather than raw gene counts and the two comparisons are not the same quantity. On the four datasets where the raw-gene-count baseline is obtainable the median margin is larger, 0.389 and 0.440 under the two kernels. The margin against raw gene counts at atlas scale is unmeasured, and we report the probe as a boundary illustration rather than a replicated result.

Figure 2 draws the nine rows that carry a value. A bar chart cannot show the absence of one without inventing a number for it, so the rows that carry no value are collected in Figure A1, with the reason printed on each row rather than gathered into a legend.

The eight divide four ways. SPIDE and scEnergy were blocked and produced no run artefact. SPIDE's source paper is paywalled with no open-access or repository copy, though its title declares a local cell-specific network entropy and the categorization places it with SCENT (Zheng et al., 2023; Wang et al., 2024). A first-party Python implementation is public and we did not run it; the omission is ours and the code is public, so a reader can close it. scEnergy, the per-cell energy of the scEpath pipeline, is distributed only inside a MATLAB pipeline whose network construction uses the Statistics toolbox and MATLAB graph objects that GNU Octave does not provide, and we had no licence (Jin et al., 2018). StemSC was measured on the sorted atlases and not on this ordinal. CytoTRACE 2 is supervised, and mRNAsi and FitDevo score cells with gene weights fitted to labelled samples, so the three sit outside the unsupervised class this audit covers; the two trained rows print their boundary results on the row, and Appendix F tabulates them. StemID and cmEntropy equal a Shannon-entropy primitive by construction, and they appear in a band at the foot of the plot that carries no scale. The band carries none on purpose: plotting them at a skill of zero would report a number we did not measure.

Appendix B. Resolution of the zebrafish result

The audit reports one Kendall τ_b per score against the twelve-stage ordinal. That statistic pools every cross-stage pair, so a score can reach a high value by separating early embryos from late ones while ordering neighbouring stages poorly. Table A2 decomposes each published value by stage gap. It covers seven rows; MCE and StemFinder, added after it was computed, were not decomposed. Stage ranks map to hours post fertilisation as 1:3.3, 2:3.7, 3:4.3, 4:4.7, 5:5.3, 6:6, 7:6.7, 8:8, 9:9, 10:10, 11:11, 12:12, taken from the standard staging series for this organism (Kimmel et al., 1995) rather than from the accession, which records stage names only.

CytoTRACE's 0.653 averages a curve. Concordance restricted to a single stage gap is +0.318, and it rises monotonically with separation to +0.896 four stages apart and +0.997 or above from eight. At one stage of resolution the score is unreliable: it orders three of the eleven single-stage transitions backwards. CytoTRACE, signalling entropy and ORIGINS share one shape: concordance rises with separation and saturates. CCAT runs the other way and deepens from −0.094 to −0.363. NCG reaches about +0.6 by gap five and then varies between +0.42 and +0.64 without further gain. At single-stage resolution CytoTRACE reverses three of the eleven transitions, at 3.3 to 3.7 hours, 3.7 to 4.3, and 5.3 to 6; signalling entropy reverses five and ORIGINS three. Over all cross-stage pairs the concordance is +0.700 for CytoTRACE and +0.461 for signalling entropy, above the reported 0.653 and 0.430 because the denominator of τ_b also counts the pairs tied within a stage. Restricted to adjacent stages, CytoTRACE retains 0.455 of its +0.700 and signalling entropy 0.523 of its +0.461, on 1,000 cell-level bootstrap resamples with 95% intervals of $[0.448, 0.461]$ and $[0.511, 0.535]$.

Table A1: The seventeen scores. Category follows Wang et al. (2024), with our class added where a score is trained or supervised: average-based scores weight important genes, entropy-based scores measure how evenly expression spreads, with or without an interaction network, and correlation-based scores correlate expression with a gene weight vector. The primitive is the statistic each score's own authors say it is built on or approximates, the one it is residualized on; the two trained rows declare none and are read on the joint set, rows marked – carry no residual, and StemSC's is the fraction of its reference gene pairs whose within-cell order matches the stem-cell reference, used on the sorted atlases only. Polarity is the direction each score's authors give it, "potent" when a higher value means a more potent cell; for NCG (*) it is inferred from the construction, since its authors state none (Appendix G). Placement: on the zebrafish ordinal, on a 3,000-cell subsample of it (‡), as a boundary row, or the reason it carries no value there. A boundary row is a trained score, outside the unsupervised class and the audit's counts, residualized on the joint set and reported in Table A3.

score	category	primitive	polarity	placement	source
CytoTRACE v1	average-based	gene count	potent	ordinal	Gulati et al. (2020)
ORIGINS	average-based	PCC(x, degree)	potent	ordinal	Senra et al. (2022)
NCG	average-based	PCC(x, degree)	potent*	ordinal ‡	Ni et al. (2022)
SR (signalling entropy)	entropy, network	PCC(x, degree)	potent	ordinal	Teschendorff and Enver (2017)
MCE	entropy, network	PCC(x, degree)	potent	ordinal	Shi et al. (2020)
SPIDE	entropy, network	–	–	no measurement	Zheng et al. (2023)
dpath	entropy	Shannon entropy	potent	ordinal ‡	Gong et al. (2017)
SLICE	entropy	Shannon entropy	potent	ordinal ‡	Guo et al. (2017)
scEnergy	entropy	–	–	no measurement	Jin et al. (2018)
StemID	entropy	Shannon entropy	–	equal by construction	Grün et al. (2016)
cmEntropy	entropy	Shannon entropy, top 1,000 genes	–	equal by construction	Kannan et al. (2021)
CCAT	correlation-based	PCC(x, degree)	potent	ordinal	Teschendorff et al. (2021)
StemFinder	none	cell-cycle score	differentiated	ordinal	Noller and Cahan (2024)
StemSC	none	gene-pair order fraction	–	sorted atlases only	Zheng et al. (2022)
mRNAsi	correlation-based, trained	joint set	potent	boundary row	Malta et al. (2018)
FitDevo	correlation-based, trained	joint set	potent	boundary row	Zhang et al. (2022)
CytoTRACE 2	supervised	–	–	outside the class	Kang et al. (2025)

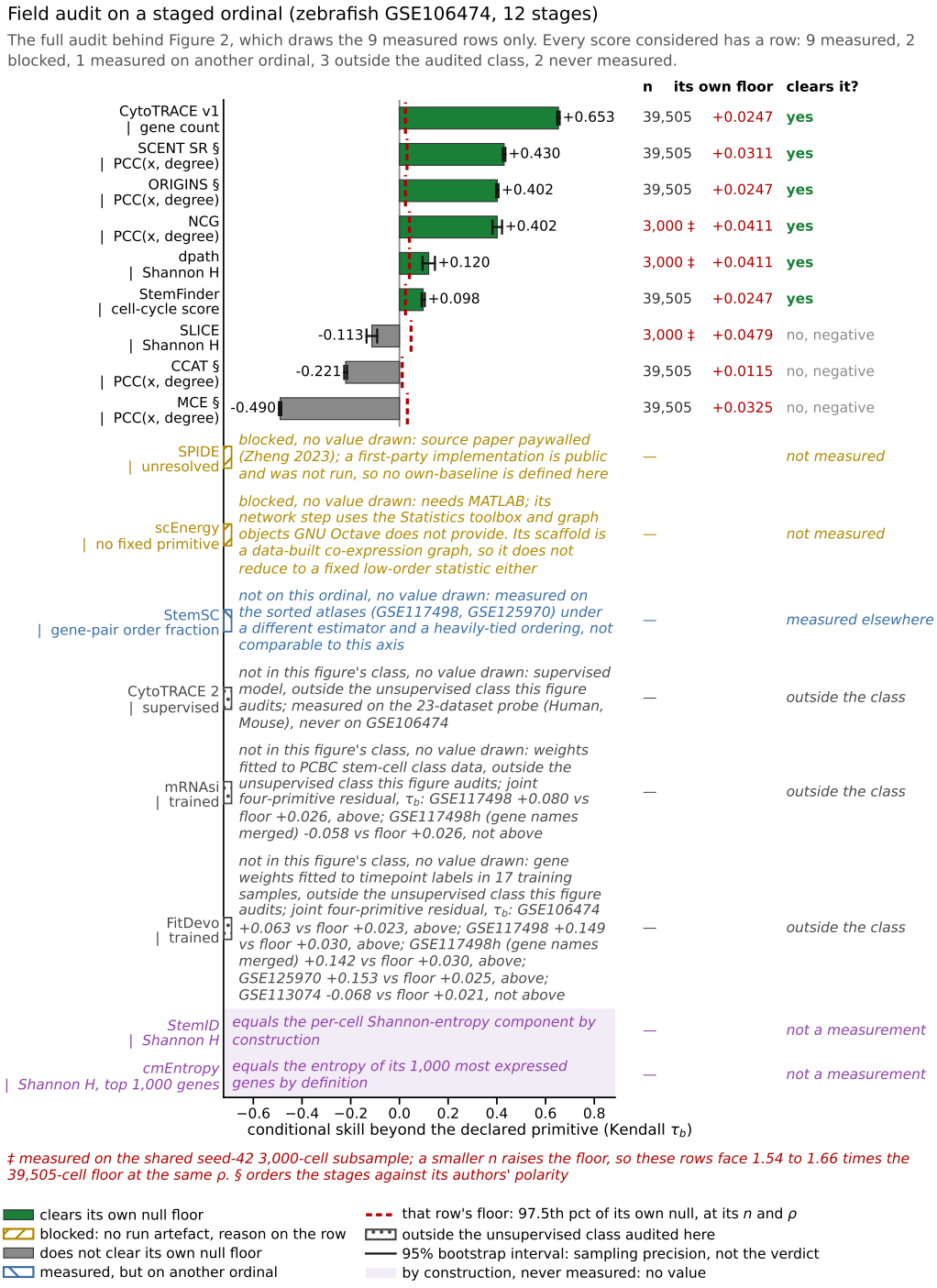


Figure A1: The full 17-row audit behind Figure 2, which draws the nine measured rows only. Bars, whiskers and dashed floors read as there; the side table gives each row's n, its floor and whether the value clears it. The eight rows without a value print their reason on the row, the two trained rows with their boundary results (Table A3). StemID and cmEntropy equal a Shannon-entropy primitive by construction and sit in the shaded band at the foot, which carries no scale.

Table A2: Concordance at each stage gap, zebrafish ordinal, same residuals as the main audit. Column g_k restricts the concordant-minus-discordant count to cell pairs whose stage ranks differ by exactly k, normalised by the tie-corrected pair count within that restriction, so each column is a concordance and not an average of per-stratum correlations. Each row is its score's residual, Eq. (2), of the aligned score, Eq. (1), so a positive entry is concordance in the direction the score's own marginal correlation runs: toward the earlier stages for CytoTRACE, NCG, dpath and SLICE, toward the later stages for signalling entropy, ORIGINS and CCAT (Appendix G). With the stage-gap restriction lifted and the pairs tied within a stage counted, the same routine returns the reported τ_b of each row, as `scipy.stats.kendalltau` does. NCG, dpath and SLICE run on the shared seed-42 3,000-cell subsample, whose earliest stage holds 28 cells; dpath's g_{10} and g_{11} rest on that block alone and are reported here for completeness rather than as a finding.

row	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
CytoTRACE v1	+0.318	+0.575	+0.755	+0.896	+0.964	+0.982	+0.990	+0.997	+0.998	+0.999	+0.999
SCENT SR	+0.241	+0.331	+0.415	+0.544	+0.700	+0.694	+0.802	+0.796	+0.841	+0.866	+0.853
ORIGINS	+0.248	+0.318	+0.397	+0.449	+0.614	+0.659	+0.764	+0.822	+0.893	+0.964	+0.953
NCG	+0.224	+0.373	+0.499	+0.510	+0.604	+0.446	+0.638	+0.439	+0.417	+0.454	+0.537
dpath	+0.191	+0.276	+0.061	−0.014	+0.075	−0.058	+0.322	+0.015	−0.255	−0.973	−0.975
SLICE	+0.048	−0.001	−0.108	−0.149	−0.261	−0.474	−0.427	−0.577	−0.391	−0.079	+0.013
CCAT	−0.094	−0.141	−0.288	−0.252	−0.406	−0.320	−0.463	−0.320	−0.343	−0.351	−0.363

Appendix C. Stage order, and the audit re-run on the other one

The accession lists the shield stage between bud and the three-somite stage. Shield is about six hours post fertilisation and sits developmentally between 50 and 60 percent epiboly (Kimmel et al., 1995), so we rank it sixth, and we pre-registered the accession's own order as a sensitivity check before running the audit.

Re-running on that ordinal moves every value and changes no verdict. CytoTRACE reads 0.593 against 0.653, signalling entropy 0.402 against 0.430, ORIGINS 0.398 against 0.402 and dpath 0.159 against 0.120, each still above the floor for its own n and ρ; the two negative rows stay negative, CCAT at −0.208 against −0.221 and SLICE at −0.098 against −0.113. NCG runs on a separate implementation and was not re-run, and MCE and StemFinder were added after this sensitivity and were not re-run on it. The point estimates are in the released decider output under `conditional_skill_shield_sensitivity_pointest`; no intervals were computed for them.

Appendix D. Reporting details

Composition of the ordinal. Quality control keeps cells with at least 200 detected genes, leaving 39,505 cells over 23,974 genes, of which 8,899 intersect the STRING scaffold. Per stage, as rank, hours post fertilisation, cells and median detected genes: 1, 3.3, 311, 3,121; 2, 3.7, 200, 2,613; 3, 4.3, 1,174, 1,931; 4, 4.7, 1,506, 1,610.5; 5, 5.3, 5,864, 1,937.5; 6, 6.0, 1,046, 1,291; 7, 6.7, 4,201, 1,772; 8, 8.0, 6,324, 1,492; 9, 9.0, 5,626, 1,377; 10, 10.0, 7,162, 1,195; 11, 11.0, 1,631, 1,047; 12, 12.0, 4,460, 1,040.5. Stage sizes range over a factor of 36, and the median detected gene count falls with developmental stage, which is why the gene-count primitive carries skill on this ordinal at all.

Implementations. Four scores were reimplemented and checked against the reference implementation on a shared 500-cell subsample before use: CytoTRACE v1 as a port of the published pipeline (most variable genes, gene-counts signature, non-negative least squares, Markov diffusion) over the authors' own core functions; signalling entropy and CCAT as vectorizations of the SCENT algorithm, agreeing with the R implementation at Spearman 1.000 and a maximum absolute deviation below 10^{-13}; ORIGINS as $x^{\mathsf{T}}Ax$ on its native differentiation network, agreeing with `ORIGINS::activity` at Spearman 1.000. Three were run from the authors' own code: SLICE through `SLICE::getEntropy` with gene-ontology kappa similarity and bootstrap entropy, dpath through metagene entropy over its weighted-Poisson non-negative matrix factorization at $K = 5$, and NCG against its published gene-

ontology-weighted connectome. Two were added later. MCE was ported from its authors' MATLAB code and StemFinder was run from its authors' R package; Appendix F gives both. In section 2 CytoTRACE is represented by its gene-count signature, the geometric mean of the 200 genes most correlated with gene count, without its regression and diffusion steps. No score's parameters were tuned; each runs as its authors distribute it, apart from StemFinder's number of components, which its authors choose by eye and we fixed by a registered rule.

Tiers and AUROC in the sorted atlases. GSE117498 ranks 12,354 cells in four tiers: HSC (1,282 cells) at the top; MPP (215) and the broad CD34+/CD164+ gate (6,343); CMP (1,576) and MLP (123); and GMP (1,012), MEP (1,211) and Pre-B/NK (592) at the bottom. The 9,058 cells of its three CD34-negative or CD34-low gates carry no rank. GSE125970 ranks all 14,537 cells: stem cells (660) at the top, transit-amplifying cells (1,096) and progenitors (4,156) in the middle, and enterocytes (5,719), goblet cells (2,019), Paneth-like cells (759) and enteroendocrine cells (128) at the bottom. The AUROC of section 2 is the probability that a randomly drawn cell of the top tier scores above one of the bottom tier, ties counted half, on the raw score with the middle tiers left out; Kendall τ_b uses every ranked cell. Depth is not monotone in the intestine either: transit-amplifying cells carry a median of 4,111.5 detected genes, above the stem cells' 3,263.

The simulation of Figure 1. Depth d and a biological position b are independent standard normal over 3,000 cells, and the ordinal cuts $0.6\,d + 0.8\,b$ into twelve equally occupied levels. Score A is $d + 0.3\,\varepsilon_A$ and score B is $d + 0.5\,b + s\,\varepsilon_B$, with $s = 1.27$ found by bisection so that both reach the same τ_b with the ordinal. The seed is fixed in `figures/make_fig_failure.py`, and the floors are read as for every other value.

The null for one primitive. Cells are drawn as $p \sim N(0, I_n)$, $s = ap + \sqrt{1-a^2}\,\varepsilon$ and $o = \mathrm{bin}_L(\gamma p + \sqrt{1-\gamma^2}\,\eta)$, with ε and η independent standard normal. $a = 2\sin(\pi\rho/6)$ gives s and p the Spearman correlation ρ, $\gamma = 2\sin(\pi c/6)$ gives p the Spearman correlation c with the variable the ordinal is cut from, and bin_L cuts that variable into L equally occupied levels; at $\rho = 1$ the score is $\exp(p)$. Given p, the score is independent of the ordinal. We call this the *coupled null*; every floor in the paper is read from it or from its form for several primitives below. The floor is the largest of the three 97.5th percentiles at $c \in \{0.45, 0.55, 0.63\}$, a range set by the absolute Spearman correlations of gene count, PCC(x, degree) and Shannon entropy with the twelve stages.

The null grid. The null is measured rather than assumed. The grid crosses five sample sizes, 500, 3,000, 10,000, 39,505 and 127,607, six score-primitive rank correlations, 0.0, 0.5, 0.9, 0.93, 0.998 and 1.0, and alignment on or off, under an uncoupled null, in which the ordinal is independent of both score and primitive, and under the coupled null at its three coupling strengths, all on a twelve-level ordinal, 240 grid points; a sweep at $n = 3{,}000$ over 2, 4 and 50 levels adds 72, for 312 grid points at 200 seeds each. A reported value is placed against the grid point nearest it in both axes; no interpolation is performed between grid points, so a value whose n or ρ falls between grid points is judged against a neighbour rather than against its own coordinates. For residuals on more than one covariate the grid is coarser still, covering $n \in \{3{,}000,\ 39{,}505\}$ and excluding $\rho = 1.0$. Pooling is over seeds within a grid point, and the percentile is recomputed on the pooled sample rather than averaged across seeds.

The null for several primitives and other ordinals. For a residual on $k > 1$ primitives the covariates are drawn as $p_j = 0.6\,u + 0.8\,v_j$ with $u, v_1, \ldots, v_k$ independent standard normal, so they correlate with one another at 0.36; the ordinal is cut from their standardized mean at coupling $c = 0.55$, and the score tracks p_1 alone at Spearman ρ, so its true conditional skill given all k is zero. The grid measures $k \in \{1, 2, 4\}$ at $n \in \{3{,}000,\ 39{,}505\}$. For the rows of Appendix F, a value whose n, ρ or number of ordinal levels lies far from every grid point gets a design point of its own: the same generators and seeding, run at its own n, ρ, k and L before the value is computed. The boundary rows of Appendix F are read this way, on ordinals of three, four and ten levels. To check that design points and grid points agree, four grid points were rerun as design points; the 97.5th percentiles of τ_b differ by at most 0.004.

On the threshold itself. A percentile of a measured null is a bright line, and bright lines invite the objection that a decision has been substituted for an inference. Two things limit that here. The floor is not a p value and it is not compared against a fixed α across rows: it is the estimator's own null at that value's own n and ρ, so it differs by row on purpose and is not comparable between rows. And a placement is never reported alone: the value, its bootstrap interval and its floor are all tabulated, so a reader who

prefers a different rule can apply it to the same numbers. A value below its floor is a failure to establish an increment, and we do not read it as evidence that no increment exists.

The floor a score has to clear is a property of the estimator. It is positive and close to flat in n, while a bootstrap interval narrows as one over the square root of n, so at atlas scale an estimate whose interval excludes zero is not evidence of anything: at $n = 39{,}505$ the τ_b floors sit two to seven interval half-widths above zero. Those floors were measured on synthetic data under the nulls for one primitive and for several given above, so a score can clear its own floor and still track structure in the atlases that these nulls do not represent.

Compute. The null grid is the only expensive step: 312 grid points at 200 seeds each, run across nine processes. The block at $n = 3{,}000$, 120 grid points and 24,000 replicates, completed in 35 seconds. The 48 grid points at $n = 127{,}607$ took about 3,541 seconds for the first 36 on nine processes and 279 seconds for the other 12 on six. The two sorted-atlas runs in section 2 took 87 and 39 seconds. The estimator-defect check of Appendix I was run on two platforms, as reported there. No per-machine hardware specification was recorded, so we state wall-clock and parallelism only.

Assets and licences. All expression data are previously published and public in NCBI GEO: GSE106474, GSE113074, GSE117498, GSE125970, GSE150949 and GSE223003. GEO attaches no licence to submitted data. Its disclaimer states that "NCBI places no restrictions on the use or distribution of the GEO data", while noting that "some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted". The scaffold is STRING v12.0 at confidence 700, distributed under a Creative Commons BY 4.0 licence. Among the implementations we ran, ORIGINS is MIT-licensed, SLICE is GPL-3, SCENT and dpath state GPL-3 and GPL (≥ 2) in their package metadata without carrying a licence file, and NCG states no licence at all. CytoTRACE v1 carries conflicting terms, GPL-3 in its package metadata and a non-commercial agreement in its licence file, and we do not assert which governs. Of the implementations we could not run, scEpath likewise states no licence. Of those added later, the stemFinder package declares its licence as PRIVATE in its metadata, FitDevo's repository carries no licence file, MCE.m states no licence, and the mRNAsi weights ship in TCGAbiolinks, which is GPL (≥ 3).

Appendix E. Methods note

Cell counts. Briggs et al. report 136,966 cells profiled by inDrop across ten timepoints; our prepared Xenopus matrix carries 127,607 after our own quality control, and every Xenopus number in this paper is computed on that set. GSE106474 as published reports 38,731 cells (Farrell et al., 2018). Our prepared matrix carries 39,505, the difference being our own quality control applied to the raw GEO UMI count table rather than the authors' filtered set; the thresholds are in the released `prep` stage and every number in this paper is computed on the 39,505.

Gene names. The GSE117498 files spell 4,137 genes two ways: the seven sorted-population files in HGNC form (HLA-A), the four broad-gate files in the form R's `make.names` gives (HLA.A). We merge each pair before taking the union of genes across files. Without the merge each such gene splits into two columns, one of them zero in every broad-gate cell, and the mitochondrial and HLA genes of the scaffold are among them; the section 2 values then read 0.229 for the marginal gap and 0.232 for the conditional skill, and no verdict changes.

Estimators. The section 2 value of 0.195 comes from the single-covariate estimator on a four-tier heavily tied ordering, and the null grid simulates the twelve-level ordinal under the multi-covariate estimator, so no matched null grid point exists for it. Two rank-residual implementations exist in the released code. The audit rows in sections 3 and 4 use the multi-covariate least-squares estimator with an intercept. The depth-confound value in section 2, 0.195, was produced by the single-covariate implementation, whose slope carries an n/(n-1) inflation and which fits no intercept; the measured divergence between the two reaches 3.7×10^{-2} at n = 106 and falls to 6.6×10^{-7} at n = 39,505. Both are orientation-invariant and the section 2 comparison is already polarity-aligned.

Appendix F. The four rows added last

MCE. Its authors' MATLAB function fits, for each cell, a Markov chain to the interaction network and the cell's normalized expression by iterative scaling, adds the chain's entropy rate to the static entropy

of the normalized expression, and divides by the logarithm of the network's number of non-zero entries (Shi et al., 2020). We ported it to Python with its start, its stopping rule (largest change below 10^{-2}) and its iteration cap, ran the unmodified MATLAB file in GNU Octave 11.3.0 on the 500-cell subsample used to validate SCENT, and matched it at Spearman 1.000 with a largest absolute difference of 1.8×10^{-13} and identical step counts in every cell. The 10^{-2} rule leaves the score within 8.3×10^{-8} of the fixed point solved to a residual of 10^{-6}. MCE.m prescribes no input transform, so MCE received signalling entropy's input, $\log_2(c + 1.1)$ of counts c scaled to the median library size, over the 8,468 genes of its network, as registered in advance. All 39,505 cells were scored, 109 to 659 steps each, none undefined. The null does not calibrate MCE's alignment: at its grid point the simulated scores correlate with the ordinal at 0.40 to 0.57, never near the 0.035 that set MCE's sign.

The prediction for MCE. Before MCE met a stage label we registered that, as an entropy-rate construction, it would keep conditional skill above its floor on PCC(x, degree). Section 6 reports that it does not, and why the alignment step decides it. The registered rule is one-sided and does not ask whether a residual as large as MCE's in the other direction fits the null.

StemFinder. StemFinder was run from its authors' R functions, unmodified, with the processing steps of the package and its vignette: 2,500 variable genes with the cell-cycle genes removed, 50 principal components, round($\sqrt{n}$) = 199 nearest neighbours, and the package's human S-phase and G2/M lists, of which 81 distinct genes are present in the zebrafish matrix after matching symbols in upper case. The authors choose the number of components by eye from an elbow plot. We registered a rule in its place, the point of the standard-deviation curve farthest from the chord joining its ends, and it kept seven. Its declared primitive is the package's own gene-set score on the same genes. Nothing is fitted, but its authors chose among variants of the method on a benchmark that includes GSE106587, the SuperSeries that contains GSE106474.

The trained rows. mRNAsi and FitDevo were run with their published weights, untouched, and enter no count. Neither declares a primitive, so each is residualized on all four at once. Table A3 gives the values on the sorted progenitors, the intestinal epithelium and the two embryo series. Three caveats bear on them. FitDevo's authors evaluated their weights once on the zebrafish stage labels, and zebrafish and Xenopus symbols were matched to human ones in upper case, although the authors left zebrafish out of training for too few homologous genes. In GSE117498 the broad-gate files spell 4,137 genes in R's `make.names` form and the sorted files in HGNC form; the code path we inherit unions them by exact name, so those genes read zero in broad-gate cells. GSE117498h is the registered sensitivity with the names merged. The plan made the unmerged GSE117498 rows decisive because the inherited code path loads the data that way; the rest of this paper merges the names (Appendix E), and on GSE117498h mRNAsi does not clear. Both null generators cut the ordinal into equally occupied levels, which the real tiers are not: most GSE125970 cells sit in its lowest tier, and the second-highest of the four GSE117498 tiers holds over half of its ranked cells. FitDevo clears on GSE117498 and on GSE117498h, with a residual that runs against its authors' polarity on both; on the zebrafish stages, the intestinal epithelium and the Xenopus timepoints it runs with that polarity. On the Xenopus timepoints FitDevo shows MCE's pattern: a marginal correlation of 0.116 sets the sign, and the residual runs the other way. mRNAsi does not keep its verdict: its residual runs against the authors' polarity on both GSE117498 and GSE117498h, at -0.080 and -0.058, while its marginal correlation with the tiers moves from -0.021 to $+0.010$, which flips the alignment and with it the verdict.

Appendix G. The direction of each score

The registered estimator takes each score's sign from the score's own Spearman correlation with the ordinal, Eq. (1). Published benchmarks of these scores fix the sign from the method's definition before scoring and report a reversed ordering as a failure (Gulati et al., 2020; Teschendorff et al., 2021; Kang et al., 2025; Noller and Cahan, 2024; Zhang et al., 2022). Taking the sign from the same labels the skill is then scored against is a selection on those labels, the circularity that Kriegeskorte et al. (2009) describe. We therefore read every placed row a second time with the sign fixed by its authors' polarity, after all registered values were known, and report the two readings side by side for every row, with the registered one primary, as guidance on reporting deviations from a registered plan asks (Willroth and Atherton, 2024). Table A4 gives both.

Table A3: The trained boundary rows. Each value is the Kendall τ_b of the aligned score's rank residual on the four primitives at once (detected genes, PCC(x, degree), Shannon entropy, $\log_{10}$ library size), with its 95% bootstrap interval over 1,000 resamples. L is the number of ordinal levels: four ranked progenitor tiers in GSE117498, three in the intestinal epithelium of GSE125970, twelve zebrafish stages in GSE106474 and ten Xenopus timepoints in GSE113074. ρ is the largest Spearman correlation between the score and any one primitive, and the floor is the null's 97.5th percentile at the grid point or design point nearest that ρ, at $k = 4$ and the row's own L, with its value at $L = 12$ in brackets. ρ_S is the raw score's Spearman correlation with the ordinal as stored, whose sign the alignment step takes; GSE106474 and GSE113074 are stored from early to late, GSE117498 and GSE125970 from low to high potency. The last column is the same residual with the authors' polarity counted positive. GSE117498h is GSE117498 with the gene names merged. None of these rows enters the audit's count.

				CS, four primitives		floor			authors'
score	data	L	n	ρ	[95% CI]	(at L=12)	verdict	ρ_S	polarity
mRNAsi	GSE117498	4	12,354	+0.634	+0.0800 [+0.0671, +0.0915]	+0.0263 (+0.0245)	above	−0.021	−0.0800
mRNAsi	GSE117498h	4	12,354	+0.648	−0.0578 [−0.0693, −0.0446]	+0.0256 (+0.0254)	not above	+0.010	−0.0578
FitDevo	GSE106474	12	39,505	+0.942	+0.0631 [+0.0580, +0.0679]	+0.0227	above	−0.377	+0.0631
FitDevo	GSE117498	4	12,354	+0.782	+0.1494 [+0.1359, +0.1617]	+0.0301 (+0.0266)	above	−0.087	−0.1494
FitDevo	GSE117498h	4	12,354	+0.784	+0.1425 [+0.1290, +0.1550]	+0.0303 (+0.0274)	above	−0.088	−0.1425
FitDevo	GSE125970	3	14,537	+0.944	+0.1527 [+0.1425, +0.1630]	+0.0246 (+0.0235)	above	+0.708	+0.1527
FitDevo	GSE113074	10	127,607	+0.814	−0.0681 [−0.0711, −0.0651]	+0.0215 (+0.0216)	not above	+0.116	+0.0681

The second reading uses the same residual, Eq. (2), with the authors' polarity counted positive and the ordinal oriented toward the earlier, more potent stages. Its floor is the 97.5th percentile of the same grid point with alignment off, since a sign fixed in advance is not chosen from the data; the two floors differ by at most 0.0026 across the nine rows. Because the rank residual and τ_b both change sign with the score, the second value is the registered one with its sign reversed wherever a score's marginal ordering runs against its authors' polarity, and equal to it elsewhere.

Four rows reverse, and they share a primitive. The degree correlation PCC(x, degree) rises with stage on these embryos, at a Spearman of 0.516 toward the later stages, and signalling entropy, ORIGINS, CCAT and MCE rise with it, against the polarity their authors give them. NCG, residualized on the same primitive on the subsample, runs toward the earlier stages. Under the second reading signalling entropy and ORIGINS fall below their floors and CCAT and MCE clear theirs, so the count stays at six of nine, and four rows clear under both readings: CytoTRACE, NCG, dpath and StemFinder. The joint four-primitive residuals follow the same signs. No null with alignment off was measured for four covariates, so the joint values under the second reading are set beside the registered k = 4 floors only.

The table also gives the null's 2.5th percentile at each row's registered grid point, for description. CCAT, SLICE and MCE lie below it, so their residuals are not values the null produces; the registered rule is one-sided and draws no verdict from that tail, and neither do we.

Appendix H. Transfer across cancers adds nothing beyond the target's own fit

Scores of this family are also claimed to capture structure holding across tumour types. We tested gene count, and SR and CCAT, two scores whose primitive is the degree correlation, for such transfer, between PC9 lung adenocarcinoma under osimertinib (Oren et al., 2021) and Kuramochi ovarian carcinoma under olaparib (Dai et al., 2024), fitting on one system and evaluating on the other in both directions. Each system gives its pre-treatment cells a binary fate label: the cycling-persister fate in PC9, positive for 150 of its 1,108 profiles, and the pre-resistant fate in Kuramochi, positive for 904 of 9,394. For a statistic we fit a class-balanced logistic regression on it alone in the source system S, score the target system T with it, and report $\Delta = \mathrm{AUROC}_{S\to T} - \mathrm{AUROC}_T$, its AUROC on the target labels minus the five-fold

Table A4: Each placed score read two ways. ρ_S is the raw score's Spearman correlation with the stage ordinal stored from early (1) to late (12), whose sign the alignment step takes; polarity is the direction the score's authors give it (sources in Table A1; NCG* inferred), and "marginal" says whether the score's ordering of the stages runs with that polarity, potency falling with stage. Registered: conditional τ_b on the declared primitive after alignment, against its registered floor, from the coupled null of Appendix D with alignment on. Authors' polarity: the same residual with the authors' direction counted positive, against the alignment-off floor of the same grid point. Null 2.5%: the registered grid point's 2.5th percentile, the smallest over its coupling strengths. Joint: the four-primitive residual under the two readings. NCG, dpath and SLICE use the 3,000-cell subsample. Computed after the registered values were known; no registered verdict changes.

score	n	ρ_S	polarity	marginal	registered / floor	authors' / floor	null 2.5%	joint, reg. / auth.
CytoTRACE	39,505	−0.933	potent	with	+0.653 / 0.0247	+0.653 / 0.0249	+0.0092	+0.288 / +0.288
signalling entropy	39,505	+0.692	potent	against	+0.430 / 0.0311	−0.430 / 0.0326	+0.0148	+0.183 / −0.183
ORIGINS	39,505	+0.712	potent	against	+0.402 / 0.0247	−0.402 / 0.0249	+0.0092	+0.156 / −0.156
NCG	3,000	−0.311	potent*	with	+0.402 / 0.0411	+0.402 / 0.0414	−0.0065	–
dpath	3,000	−0.444	potent	with	+0.120 / 0.0411	+0.120 / 0.0414	−0.0065	+0.134 / +0.134
StemFinder	39,505	+0.445	differentiated	with	+0.098 / 0.0247	+0.098 / 0.0249	+0.0092	+0.167 / +0.167
SLICE	3,000	−0.525	potent	with	−0.113 / 0.0479	−0.113 / 0.0453	+0.0001	–
CCAT	39,505	+0.496	potent	against	−0.221 / 0.0115	+0.221 / 0.0120	−0.0011	–
MCE	39,505	+0.035	potent	against	−0.490 / 0.0325	+0.490 / 0.0342	+0.0159	−0.131 / +0.131

Table A5: What the second reading changes against the registered plan, in the form Willroth and Atherton (2024) recommend. Both rows were added on 2026-10-12, after every registered value was known.

item	registered	added	effect
sign of a score	sign of its Spearman correlation with the ordinal, Eq. (1)	a second reading with the sign fixed by the authors' polarity, reported beside the first	four of nine verdicts differ: signalling entropy and ORIGINS fail, CCAT and MCE clear; the count stays six; four rows clear under both
floor for that reading	coupled null, alignment on	the same grid point with alignment off	floors move by at most 0.0026; no verdict depends on the change

cross-validated AUROC of the same regression fitted within T. A one-variable fit is monotone, so a transferred AUROC below 0.5 means the statistic relates to fate with opposite signs in the two systems.

Gene count and the two scores behave oppositely. Gene count anti-transfers, at $\Delta = -0.182$ and -0.080 in the two directions, with transferred AUROCs of 0.409 and 0.458, so a rule learned in one cancer is worse than nothing in the other. SR and CCAT sit near zero throughout, Δ spanning +0.001 to +0.008 across the two scores and both directions.

Read quickly, the second looks like stability. It is arithmetic. A one-variable rule is monotone in its statistic, so on the target it reaches that statistic's own AUROC or one minus it, and Δ sits near zero whenever the fitted sign agrees between the two systems, with or without a network; gene count's sign flips, and its transferred and own AUROCs sum to within 0.004 of one. Nor does the shared STRING network carry the agreement: recomputing CCAT with one random permutation of the degree vector applied to both systems returns a transferred AUROC of 0.500 averaged over 200 permutations. What transfers on the real network is a low-order axis whose relation to fate happens to run the same way in both cancers, and the target's own fit already reaches it, so a claim that such a score reveals conserved cross-cancer structure has to beat the target's own baseline first.

Appendix I. Robustness checks and the decision rule

Residualization here is by least squares on ranks, which removes the linear part of a rank-rank relation and leaves any nonlinear part to be scored as skill.

We replaced it with a non-parametric monotone (isotonic) fit. Over the seven rows it was run on, MCE and StemFinder not among them, no value moves by more than 0.034 in Kendall τ_b, no verdict changes including CCAT's and SLICE's, and the largest gain in explained variance from dropping linearity is 0.052, on NCG. CytoTRACE, the most nonlinear pipeline over its primitive, moves least of the three largest rows, at 0.007 (pre-registered).

We report no family-wise correction. In its place we require two conditions together before a score is called substantive: the value exceeds the 97.5th percentile of the estimator's own measured null, taken at that value's own sample size and its own score-primitive rank correlation, and the effect survives depth adjustment. A bootstrap interval excluding zero is not evidence here. We measured the rule we previously used on synthetic data whose true conditional skill is exactly zero, and the full conjunction of both-kernel agreement, an interval excluding zero and survival of the depth control fired in 40 of 40 replicates at $n = 39{,}505$ and $\rho = 0.5$, the grid point of CytoTRACE and ORIGINS, in 40 of 40 at $\rho = 0.9$, the simulated point nearest signalling entropy's 0.93, and in 35.5 percent of 200 replicates at $n = 3{,}000$ and $\rho = 0.5$, where dpath and NCG sit. A criterion a true zero satisfies at those rates does not distinguish a score that adds something from one that does not. We state the replacement so that a reader can apply their own correction to the tabulated values.

We removed the weighted Kendall tau. Under a null in which the ordinal is drawn independently of both the score and the primitive, its mean is not zero: it rises from 0.188 at n = 500 to 0.309 at n = 127,607, and stands at 0.241 at n = 3,000 and 0.291 at n = 39,505. It returned a positive verdict on every null replicate we ran, 720 of 720, and a kernel pinned at one answer cannot disagree with anything, so the agreement condition it was there to enforce was inert by construction. Dropping it has a cost. Under a rule requiring both kernels to clear their own floors, dpath does not clear its weighted-tau null, 0.191 against a 97.5th percentile of 0.242, and the count in section 3 would be five of nine rather than six. ORIGINS carries the same cost one stage later: its joint-4 residual under weighted tau, 0.155, sits below that grid point's null mean of 0.175.

On removing depth as though it were noise. Total mRNA content differs systematically between haematopoietic compartments and is prognostic across fifteen tumour types in its own right (Cao et al., 2022). In stem and progenitor compartments the same quantity has been read as hypertranscription, a phenotype rather than a technical artefact (Kim et al., 2023). Residualizing a score on gene count therefore removes something partly biological, and a reader is entitled to ask whether we strip out the signal. Our claim is narrower than the objection assumes. A score whose ordering is recoverable from gene count alone has not been shown to add anything to gene count, whatever gene count turns out to mean. That holds however much biology the primitive carries, because it is a claim about what this experiment established. A paper that publishes the score without the baseline leaves its reader unable to separate the two.

Xenopus. A second staged ordinal in Xenopus, ten timepoints from zygotic genome activation to early organogenesis, reproduces the degree-correlation result (Briggs et al., 2018). Its shallower chemistry nulls the gene-count and entropy primitives, so only the degree-correlation result replicates there.

The weighted kernel. Its floor is largest exactly where the score-primitive correlation is highest, which is where a reduction claim most needs it to be small. The second kernel is the weighted Kendall tau of Vigna (2015) in its additive hyperbolic form, as `scipy.stats.weightedtau` computes it with `rank=True`. Ranking the cells by decreasing (r, o) and writing $\pi(i)$ for cell i's place in that ranking, counted from zero, each pair carries the weight $w_{ij} = 1/(\pi(i)+1) + 1/(\pi(j)+1)$, and

$$\tau_w = \frac{\sum_{i<j} w_{ij}\ \mathrm{sgn}(r_i - r_j)\,\mathrm{sgn}(o_i - o_j)}{\sqrt{\sum_{i<j} w_{ij}\,[r_i \neq r_j]}\,\sqrt{\sum_{i<j} w_{ij}\,[o_i \neq o_j]}}; \quad (5)$$

the reported value averages this with the same quantity computed from the ranking by decreasing (o, r). Pairs at the top of the ranking weigh most.

The estimator defect. CCAT sits near the rank correlation where our estimator has a recorded defect. At exact rank identity the least-squares fit is exact and the residual is the rounding error of the subtraction, of order 10^{-12}. That debris is monotone in the primitive rank, so where the primitive predicts the ordinal it inherits that association and both kernels score it, reaching 0.87 in magnitude with a sign that changes with the seed. How large it is also depends on the machine, because the debris is the rounding error of a least-squares fit: across sixteen runs of the defect test one platform put all of them between 0.71

and 0.87, and another put fourteen there and collapsed two toward zero, the two where the residual itself landed two to three orders smaller. We first recorded the defect on a fixture whose ordinal is independent of the primitive, where τ_b stays near 0.02 and only the weighted kernel misbehaves, and that is the narrower case. What separates a residual from debris is its magnitude. CCAT's reaches 3,937 on a rank scale of 39,505, which is 4.5×10^{14} times machine precision times the cell count, so the defect does not reach this row.