The vcc2026 Metric Suite

2026/08/19

Contents

0 Overview **1**

1 Perturbation discrimination **3**

2 Expression error **5**

3 DGE direction fidelity **8**

4 DGE direction reach **11**

5 DGE significant set agreement **13**

6 DGE fold-change accuracy **14**

7 Submission file requirements **16**

CHAPTER 0

Overview

This document describes the six `vcc2026` metrics and how they are combined into the competition score. Parameter values are those of the packaged `configs/vcc2026.yaml`, which is the configuration the competition is scored under.

Submissions. A submission is a matrix of predicted single-cell expression whose gene axis and perturbation labels match the reference exactly, including the non-targeting control label; the number of predicted cells per perturbation is unconstrained. Values must be raw, untransformed counts: non-negative, integral, and with no cell total above 10^6. A matrix failing any of these is rejected rather than scored. Counts are required because the expression metrics compare group-summed profiles, which cannot be recovered from log-normalized data.

Evaluation data. The reference (real) measurements are pooled CRISPR-interference screens read out by single-cell RNA sequencing, in several cellular contexts, each scored independently against its own reference. Per context the panel holds 300 target constructs, one per gene, and a pooled non-targeting control drawn from 46 constructs; every perturbation is downsampled to exactly 400 cells and to a median depth of 20,000 UMI per cell, on a common axis of 18,533 genes. Cell count and sequencing depth move every metric in the suite, so fixing them removes two sources of variation that belong to the experiment rather than to the prediction.

Controls. The reference's control cells are the origin on both sides: the predicted effect is measured against them, and they form the reference group of the differential-expression test for the prediction as well as for the reference.

Gene restrictions. The perturbed gene's own row is excluded from all six metrics, and the perturbation-discrimination metric of section 2.1 removes the whole panel's 300 target genes rather than one at a time — see there for why the difference matters. Knocking a gene down and then reporting that gene as changed is the premise of the experiment rather than a prediction, and the row is worth 6–11 % of the expression metric's range on the reference contexts. The four differential-expression metrics apply two further restrictions, a low-expression filter and a significance threshold, both defined with the test in section 3.

The metrics. The competition scores the six metrics below: two computed from the expression profiles directly, and four from a differential-expression table. The `de_wilcoxon_` prefix records that the table comes from a Wilcoxon rank-sum test, and no other test is used. Five of the six are means over the perturbations of a context; `expr_mse_unbiased_capped_norm` is a single ratio of two panel-wide sums and has no per-perturbation value.

Baseline and replicate. Each metric is reported against two values measured on the same reference data and stored with it in the reference bundle, which is not released to participants. The baseline b is the context's mean perturbation response: an equal-weight average, over the constructs of that context passing a cell-count and knockdown-efficiency filter, of the per-construct mean count vector, assigned identically to every perturbation and then scored as an ordinary submission. The replicate r is the reference compared with itself: cells are split into two disjoint halves per perturbation and per control, one half is scored against the other, and five such splits (seeds derived from a pinned base seed of 0) are averaged; each half uses

metric	what it measures	section
`pds_cosine`	separability of predicted profiles	1
`expr_mse_unbiased_capped_norm`	size of the expression error	2
`de_wilcoxon_direction_fidelity_yield_raw`	correctness of predicted directions	3
`de_wilcoxon_direction_reach_raw`	depth over which directions stay correct	4
`de_wilcoxon_sig_jaccard`	agreement of responding-gene sets	5
`de_wilcoxon_lfc_nmae`	accuracy of predicted fold changes	6

its own control cells, so the two quantities whose agreement is measured are not correlated through a shared reference.

The competition score. Writing u for a metric's aggregate over the panel, each of the six is placed on the scale

$$s \;=\; \frac{u-b}{r-b}$$

so that 0 is the baseline and 1 is a repeat of the experiment, and a context's score is the equal-weight mean of the six. Values above 1 occur and are reported rather than clipped, because the replicate is estimated from half-depth data. Negative values can also occur, for a submission worse than the baseline, and are likewise reported. Four of the six members are unclamped; the expression error is clamped to $[0, 1]$ and the fold-change error is floored at -6. Each member's scaled range is stated in its own section.

Provenance of the measured values. Every number quoted in this document for a baseline b, a replicate r, a span $r - b$ or a practical value is measured on the three official reference bundles built with `cell-eval2 0.15.0` under competition `rule_version` 3, and the scaled values are read from the panel arms scored against those bundles rather than derived from b and r by hand. The baseline arm reproduces the 0 end of every member exactly, and the reference scored against itself reproduces each bundle's own stamped perfect score, so the two ends of every scale below are the ends the scorer actually applies.

CHAPTER 1

Perturbation discrimination

`pds_cosine` measures whether a predicted profile is identifiable: is the prediction for a given perturbation closer to that perturbation's measured profile than to any other perturbation's?

Definition. A group's expression profile is formed by summing counts over its cells, normalizing the result to a fixed total $\mathrm{TS} = 5 \times 10^4$, and log-transforming:

$$b_{p,g} \;=\; \log\!\left(1 + \mathrm{TS}\,\frac{P_{p,g}}{\sum_{g'} P_{p,g'}}\right), \qquad P_{p,g} \;=\; \sum_{c \in \mathcal{C}_p} y_{c,g},$$

where $\mathcal{C}_p$ is the set of cells labelled p and $y_{c,g}$ the raw count of gene g in cell c. The effect of a perturbation is its profile minus the reference control profile, on both sides: $\delta_q = b_q - b_{\mathrm{ctrl}}$ for the reference and $\hat{\delta}_p = \hat{b}_p - b_{\mathrm{ctrl}}$ for the prediction. For each perturbation p the predicted effect is compared with every one of the $n = 300$ measured effects by cosine distance,

$$d_{p,q} \;=\; 1 - \frac{\langle \hat{\delta}_p,\ \delta_q \rangle}{\|\hat{\delta}_p\|\ \|\delta_q\|}.$$

Both vectors are evaluated with the coordinates of **all 300 panel target genes** removed — not just p's own — so every one of the n^2 distances is computed in the same fixed feature space, and $d_{p,q} = 1$ whenever either has zero norm. Removing only p's own gene would leave q's knockdown visible in δ_q for every $q \neq p$: a submission could then spike the other 299 target genes, land at cosine distance > 1 from every competitor while its own comparison — where its gene is removed — sat at $d_{p,p} \approx 1$, and win every rank on no information beyond the published target list. Measured on the three reference contexts, that submission scored 0.798 / 0.757 / 0.761 against the 0.530 / 0.528 / 0.510 baselines the row-scope design itself produced; with the whole panel removed it scores 0.500, which is also where the baseline now sits, and a perfect prediction still scores 1.

The score is one minus the normalized rank of the correct match,

$$\mathrm{PDS}_p \;=\; 1 - \frac{k_p}{n-1}, \qquad k_p \;=\; \#\{q : d_{p,q} < d_{p,p}\} \;+\; \tfrac{1}{2}\left(\#\{q : d_{p,q} = d_{p,p}\} - 1\right),$$

so that a tied block of candidates shares the average of the positions it spans. The metric's value for a context is the mean of PDS_p over its 300 perturbations.

Theoretical limits. k_p ranges over $[0, n-1]$, so $\mathrm{PDS}_p \in [0, 1]$: it is 1 when the prediction is strictly closest to its own measured effect and 0 when it is farther from its own than from all 299 others. A prediction carrying no information about which perturbation is which ranks uniformly and scores 0.5 in expectation. Two degenerate predictions land there exactly rather than in expectation. A prediction that emits the same profile for every perturbation produces one shared distance row, so the ranks are a permutation of $\{0, \ldots, n-1\}$ and the mean is exactly 0.5. A prediction whose effect is zero, the reference control pasted onto every perturbation, has zero norm, so its entire row ties at distance 1 and every perturbation takes the mid-rank

$(n-1)/2$, i.e. 0.5 individually and not merely on average. The rank is taken over the 300 constructs actually scored, so the same submission evaluated against a different panel is a different quantity.

Practical values. Over the three reference contexts:

quantity	range
baseline b	0.500
replicate r	0.927 – 0.984

The baseline is exactly 0.500 on all three contexts, the no-information point itself: with all 300 panel target genes removed, a predictor handed the context's mean response carries nothing that distinguishes one perturbation from another, every distance row ties, and every perturbation takes the mid-rank. That is the property this metric exists to detect, and here it is detected exactly rather than approximately. The replicate shows that half of a context's cells recover the identity of nearly every perturbation. Its standard deviation over the five splits is 0.5–1.3 % of the span $r - b$ that the score divides by, and because the replicate is already near the ceiling there is little headroom above it: a submission reproducing the reference exactly scores 1.03–1.17. Together with the direction reach of section 4 that is the least headroom of any unclamped member: the set agreement of section 5 reaches 2.48–2.85 and the direction fidelity of section 3 reaches 1.51–1.72, both far above this member's 1.17 ceiling. The fold-change error of section 6 has headroom of the same order as those two but is not an unclamped member: its floor at -6 binds at the bad end only, and leaves the good end open.

Scaled range. The scaled value is not clamped at either end. A submission below the baseline therefore scores negative and is reported so: the control-pasting submission above sits at the baseline rather than below it, since both read 0.5, and scores 0 to within a rounding error on all three contexts. The metric's own floor of 0 maps to -1.03 to -1.17, the mirror of the headroom above: with b exactly halfway between the metric's two ends, the scale is symmetric.

CHAPTER 2

Expression error

`expr_mse_unbiased_capped_norm` measures how far the predicted expression profile is from the measured one, as a fraction of how far the measured profile is from the control. A squared distance between two profiles estimated from finite samples is inflated by sampling noise on both sides, so the metric subtracts an estimate of that inflation before taking the ratio.

The sampling correction. For a group with n cells, let $S_p = \sum_g P_{p,g}$ be its total count and ℓ_i the total of cell i. Deleting one cell moves the normalizing denominator of the profile b_p and therefore every gene of it, including genes in which that cell had no reads, so the correction is a delete-one jackknife rather than a closed-form variance:

$$C_p \;=\; \frac{n-1}{n} \sum_g \sum_i \left(v_{ig} - \bar{v}_g\right)^2, \qquad v_{ig} \;=\; \log\!\left(1 + \mathrm{TS}\,\frac{P_{p,g} - y_{ig}}{S_p - \ell_i}\right),$$

with $\bar{v}_g$ the mean of v_{ig} over the group's cells, and $C_p = 0$ for a group of fewer than two cells. C_p estimates the part of a squared distance that would be present even if the two sides agreed exactly in expectation. It is computed the same way for the prediction, written $\hat{C}_p$, and for the reference control, written C_{ctrl}.

Definition. Two per-perturbation quantities are formed, both reported. The numerator is the corrected squared distance between the predicted and measured profiles,

$$N_p \;=\; \frac{1}{G_p}\left(\|\hat{b}_p - b_p\|^2 \;-\; \rho\,\min\!\left(\hat{C}_p,\; C_p\right) \;-\; C_p\right) \qquad (\texttt{expr_mse_unbiased_capped}),$$

where $\rho \le 1$ bounds the correction a submission may claim **in total** by its own across-perturbation spread,

$$\rho \;=\; \min\!\left(1,\; \frac{B}{\sum_q \frac{1}{G_q}\min(\hat{C}_q,\, C_q)}\right), \qquad B \;=\; \sum_g \sum_{p \in R_g} \frac{1}{G_p}\left(\hat{b}_{p,g} - \bar{\hat{b}}^{\,w}_{\cdot,g}\right)^2,$$

where R_g is the set of perturbations that *score* gene g — each row's own target gene is excluded from B exactly as it is from the distance — and the mean subtracted per gene is the $1/G_p$-weighted one over that same set, so both sides of the ratio are in the units the metric reports. The reason is that $\hat{C}_p$ is estimated from the *submitted cells* by a delete-one-cell jackknife, which measures how much those cells scatter within their group — equal to the error in the submitted group profile only if the cells are an i.i.d. sample. Cells whose scatter cancels in the group total have almost none of that error in $\|\hat{b}_p - b_p\|^2$, and would otherwise still be credited for it. Writing the predicted profile as truth plus independent sampling noise, B tracks the total of those per-perturbation noise terms — deliberately sitting a little below it, by $1/n_g$ of the total at equal weights (0.33 % over 300 perturbations), because a form that corrected that shortfall would pay a submitter back more than manufacturing the spread costs. So it bounds what a submission can legitimately claim, using nothing but the submission itself. A submission

whose predictions genuinely vary across perturbations — any submission with real signal — has $\rho = 1$ and is unaffected.

and the denominator is the corrected squared distance from the measured perturbation to the measured control,

$$D_p \;=\; \frac{1}{G_p}\left(\|b_p - b_{\mathrm{ctrl}}\|^2 \;-\; C_p \;-\; C_{\mathrm{ctrl}}\right) \qquad (\texttt{expr_distance_unbiased}).$$

Both squared distances run over every gene except the perturbation's own target gene, and G_p is the number of genes actually summed. Both legs drop the same gene: with only the numerator excluding it, a submission that predicted the control exactly would no longer read 1. The sampling corrections are left whole, which is the one approximation here: the cached moments store C_p as a scalar, so the excluded gene's share of it is not recoverable. That share is an ordinary $1/G$ for the variance even though the gene dominates the distance, and leaving it in understates the denominator sum by 0.007 to 0.013 %, against the 6 to 11 % of the metric's range the exclusion itself removes. The $\min$ in the numerator caps the correction credited to the prediction at the correction the reference itself earns, so a submission cannot lower its error by claiming more sampling noise than the measurement has. The scored metric is the ratio of the two sums over the panel,

$$\mathrm{MSE} \;=\; \frac{\sum_p N_p}{\sum_p D_p},$$

which is why it has no per-perturbation value: a ratio of sums is not the mean of anything the per-perturbation frame could carry. D_p reads the measured data only, so it is identical for every submission scored against a given context.

Theoretical limits. Lower is better. Both N_p and D_p are signed, because each subtracts an estimate that can exceed the plug-in distance it corrects; a negative row is a statement that the correction cannot resolve a shift at that depth, not that the perturbation is null. Two reference points follow from the construction rather than from a fitted constant. A perfect prediction has expected numerator 0, so perfection is 0. A prediction that emits the control unchanged has expected numerator equal to the expected denominator, so no skill is 1. Because the denominator is itself corrected, that holds whatever the depth of the reference. There is no upper bound. Both statements are exact for the true sampling variances and approximate for the jackknife that estimates them, which is measured 0.32 % high at $\mathrm{TS} = 5 \times 10^4$; a control-emitting submission therefore reads slightly above 1 rather than exactly 1. The cap can also push a submission above 1 on its own: a prediction whose own correction would exceed the reference's forfeits the excess.

Both reference points additionally require $\rho = 1$ — perfection at 0 as well as no skill at 1: a prediction that matches the truth but emits sampled cells retains $(1-\rho)\hat{C}_p$ in expectation where the panel binds. Neither is at risk for a submission with real signal, whose predictions vary across perturbations far more than its own sampling noise (on the reference contexts, by a factor of about 3), but they are properties of the panel rather than guarantees.

The no-skill point of 1 additionally requires $\rho = 1$ for a second reason. A submission that emits an independent draw of cells per perturbation satisfies it to within a fraction of a percent. One that emits the **same** cells for every perturbation does not: its group profiles are then identical, $B = 0$, and it forfeits the whole correction, reading above 1. That is deliberate — with

identical profiles there is nothing in the submission that distinguishes real sampling noise from a systematic offset, and the correction exists only for the former. Predictions that vary across perturbations, which is every submission carrying signal, are unaffected.

Practical values. Over the three reference contexts:

quantity	range
baseline b	0.986 – 0.992
replicate r	0.028 – 0.045

The baseline sits within 0.014 of the no-skill point of 1, so on this metric a predictor handed the context's mean response is barely distinguishable from one that pastes the control. The replicate shows that half of a context's cells leave 3–5 % of the control-pasting error, so the span $r - b$ is about 0.95, the widest of the six on every context. That span is what the score divides by, which is worth separating from the replicate itself: the standard deviation of r over the five splits is 28–45 % of r, but only 0.8–2.2 % of the span. Excluding the target gene matters more here than elsewhere: an adversary predicting the control everywhere and the truth at its own target gene was taking 6–11 % of the scaled range before the exclusion was applied, because that gene is the largest-moving one in 57–66 % of perturbations and the correction consumes roughly half of the raw distance it is read against.

Scaled range. The scaled value is clamped to $[0, 1]$ at both ends, the only member clamped at either end other than the fold-change error's floor, and the only one bounded above. The upper bound follows from the metric already carrying its own no-skill point: a raw value of 1 means the prediction is no better than the control, so a scaled value past the baseline carries nothing a ranking uses. The consequence at the other end is that a submission reproducing the reference exactly scores exactly 1.000 here while the other five exceed 1.

CHAPTER 3

DGE direction fidelity

Four of the six metrics are computed from a differential-expression table rather than from the expression profiles directly. This section defines that table, then the first metric built on it: `de_wilcoxon_direction_fidelity_yield_raw`, which measures whether the genes a submission calls as responding move in the right direction, and penalizes a submission that calls far fewer genes than the measurement found.

The test. For each perturbation, each gene is tested on its own for a difference between that perturbation's cells and the control cells, by a two-sided Wilcoxon rank-sum test (equivalently the Mann–Whitney U test). The test pools the two groups' values for that gene, replaces them by their ranks, and asks whether the perturbation's cells sit systematically higher or lower in that order than the control's; the p-value comes from the normal approximation to the rank sum, which is accurate at the group sizes here. A rank test is the appropriate instrument for this data on two counts. It assumes nothing about how a gene's counts are distributed, and single-cell counts are zero-inflated and over-dispersed rather than any convenient family; and its statistic is unmoved by how large an outlying cell is, only by where that cell falls in the order, where a mean-based test would be dragged by it. The values ranked are counts normalized per cell to a total of 10^6, and that normalization matters even for a rank test: dividing each cell by its own library size reorders cells that differ in depth.

Being two-sided, the test reports only *whether* a gene moved, never in which direction or by how much. Both of those come from the $\log_2$ fold change,

$$\mathrm{lfc}_{p,g} \;=\; \log_2 \frac{m^{\mathrm{pert}}_{p,g} + \epsilon}{m^{\mathrm{ctrl}}_{g} + \epsilon}, \qquad \epsilon = 10^{-9},$$

with m the arithmetic mean of the same normalized values over the cells of the group, and ϵ keeping the ratio finite where a group's mean is zero. The same procedure with the same parameters is applied to the submission and to the reference, and on both sides the comparison group is the measured control.

Low-expression filter. A gene is tested only if its mean expression exceeds 5 counts per million in the **control** cells. Two things follow, and the second is the reason the rule is written this way.

The first is a multiple-testing tradeoff, and it is a tradeoff rather than a free choice. Each perturbation is a family of thousands of simultaneous tests, and every gene in the family makes the correction more conservative for every other gene in it. Restricting the family to genes the control expresses above the cutoff therefore improves the sensitivity available to every gene that is retained, at the price of not scoring the retained genes' excluded counterparts at all.

The second is that the surviving gene set must not depend on the answer. Because the threshold reads the control alone, that set is one property of the evaluation data: identical for every perturbation, and — since both sides are compared against the same measured control — identical for the submission and for the reference. Every reference-significant pair is therefore present in the submission's table, and which pairs exist tells a submission nothing about them.

Nothing here claims the excluded genes carry no information. A gene silent in the control and strongly induced by a perturbation is the opposite: a large, real, well-measured effect. Those effects are genuinely not scored. That is a cost of the rule, paid to keep the tested universe independent of the answer, and not a statement about the genes.

That the control decides it alone is deliberate, and the previous rule shows why. An earlier version admitted a gene whose mean exceeded the cutoff *either* in the control *or* in that perturbation's own cells. For a gene below the cutoff in the control, the pair (perturbation, gene) then existed only when the gene had risen above the cutoff in that perturbation — so its measured fold change was positive by construction. Presence in the table disclosed the sign the submission was being asked to predict: on the three reference contexts, 100.0 % of such pairs carried a positive fold change (26k–34k pairs each, around 1 % of the table, 88–113 per perturbation). A submission that raised every below-cutoff gene, identically for all perturbations and without reading which perturbation it was answering, collected a large share of the direction members through that channel alone.

Multiple testing and significance. Each perturbation is a family of thousands of simultaneous tests, one per gene that survived the filter, so a raw p-value cannot be read as evidence on its own. We use Benjamini–Hochberg correction, which controls the *false discovery rate*: the expected fraction of false positives among the genes called, rather than the probability of making any false call at all. A gene is *significant* for perturbation p when $p^{\mathrm{adj}}_{p,g} < \alpha$ with $\alpha = 0.05$, so that no more than 5 % of the genes called for that perturbation are expected to be false.

Two scoping decisions are load-bearing. The correction is computed *after* the low-expression filter, over the surviving pairs only, so discarding an uninformative gene does not merely exclude it: it lowers the number of hypotheses m for every other gene in that perturbation, and so makes the survivors easier to call. And it is computed *within* each perturbation rather than once across the panel, so m is that perturbation's own gene count and the threshold means the same thing for a strongly responding perturbation as for a weak one.

Definition. Fix a perturbation p. A gene is *adjudicable* when the reference assigns it a defined, non-zero $\log_2$ fold change, so that the measurement can say which way it moved. Let n_{real} be the number of genes the reference calls significant: the budget of genes that responded. Let the submission's call set be the genes it calls significant that are also adjudicable, with n_{pred} its size and k the number of them whose predicted and reference fold changes carry the same sign. A gene the submission calls but leaves without a direction, through a null, NaN or exactly zero predicted fold change, stays in n_{pred} and scores as a miss rather than being dropped. Both counts exclude the perturbation's own target gene. The metric is

$$F_p \;=\; \frac{k}{\max\left(n_{\mathrm{pred}},\; n_{\mathrm{real}}\right)},$$

which factors exactly into a directional precision and a coverage term capped at 1:

$$F_p \;=\; \frac{k}{n_{\mathrm{pred}}} \cdot \min\!\left(1,\; \frac{n_{\mathrm{pred}}}{n_{\mathrm{real}}}\right).$$

Calling more genes than the reference found buys nothing, since the coverage term saturates; calling fewer costs proportionally. The metric's value for a context is the mean of F_p over the perturbations where it is defined.

Theoretical limits. $k \le n_{\text{pred}} \le \max(n_{\text{pred}}, n_{\text{real}})$, so $F_p \in [0, 1]$. It reaches 1 only when every call is correct *and* the submission calls at least as many genes as the reference found; either half alone is not enough. A submission assigning directions at random scores 0.5 in expectation once its coverage reaches 1, and less in proportion to its coverage below that. Three cases have explicit conventions. A submission that calls nothing has $n_{\text{pred}} = 0$ and therefore $k = 0$, so it scores 0 whenever the reference found anything: silence is not rewarded. If the reference found nothing either, the value is $0/0$ and the perturbation is dropped from the mean rather than scored, which is why the cohort is 295–300 of the 300 rather than a fixed 300. If the reference found nothing but the submission called genes anyway, the coverage term is inactive and the value is the bare fraction correct, a coin flip against noise.

Practical values. Over the three reference contexts:

quantity	range
baseline b	0.505 – 0.522
replicate r	0.795 – 0.832

The baseline sits at the chance level of 0.5: a predictor carrying no target-specific information does no better than a coin flip here. The replicate is well short of 1, so at half depth the measurement does not reproduce its own directional calls, which is the reason the scale is measured rather than assumed. The span $r - b$ is 0.28–0.33, the narrowest of the six on every context, so a given change in the raw value moves this member's contribution to the score more than it would elsewhere; the standard deviation of r over the five splits is correspondingly 1.5–4.8 % of the span, the largest of the six on two of the three contexts.

Scaled range. The scaled value is not clamped at either end. A submission below chance therefore scores negative, and the metric's own floor of 0 maps to -1.55 to -1.85: the deepest negative any of the four unclamped members can reach, which follows from this member having the narrowest span. The baseline sitting at chance rather than at zero has a consequence worth stating plainly: a genuine half of the reference scores -0.15 to -0.34 here, because a half calls fewer genes and its raw value, 0.41–0.46, falls below the mean-response arm's 0.505. On this member alone, the anchor's own halves are charged for their coverage.

CHAPTER 4

DGE direction reach

`de_wilcoxon_direction_reach_raw` measures how far down a submission's own ranking its directional calls stay reliable. Section 3 asks what fraction of the calls are correct; this asks how deep the correct ones extend before the ordering degrades.

Definition. The pool is the budget of section 3: the genes the reference calls significant for perturbation p, with p's own target gene removed, of which n_{real} is the count. Those genes are ordered by the submission's own confidence, its significant calls first, then by ascending predicted adjusted p-value, then by ascending predicted p-value, then by descending $|\text{lfc}|$, with the gene name breaking what remains. Walking down that order and counting only adjudicable genes, let $P(k)$ be the fraction of the first k whose predicted sign matches the reference's. The reach depth is the deepest prefix whose purity still clears a fixed threshold,

$$k^* \;=\; \max\{\, k \;:\; P(k) \geq P_0 \}\,, \qquad P_0 \;=\; 0.9,$$

taking $k^* = 0$ when no prefix clears it, and the metric is that depth as a fraction of the budget:

$$R_p \;=\; \frac{k^*}{n_{\text{real}}}.$$

$P_0 = 0.9$ is a chosen pass mark. It reads directly: a prefix clears it while at most one call in ten is wrong, so the shallowest depth at which a single error can be tolerated is 10. A wrong call at position j leaves every prefix before it clean and therefore qualifying, so $k^* \geq j - 1$; only a wrong call at the very first position can drive k^* to zero, and only then if the ranking never recovers. "Deepest" is literal rather than "first": purity is not monotone in k, so a prefix that dips below P_0 and recovers still counts at its deeper crossing – a leading miss followed by nine matches gives $P(10) = 9/10 = 0.9$, and k^* reaches 10. The metric's value for a context is the mean of R_p over the perturbations where the budget is non-empty.

Theoretical limits. The ranking pool is the reference's own budget, so $k^* \leq n_{\text{real}}$ and $R_p \in [0, 1]$. It is 1 when the submission orders the whole budget with purity at or above 0.9 all the way down, and 0 when no prefix reaches 0.9. The metric is not a restatement of section 3: fidelity is insensitive to the order of the calls, so a submission that identifies the right genes but ranks them poorly scores well there and badly here. On the other hand, a submission that does not calibrate FDR properly, but provides correct significance ranking, will score poorly on fidelity and high on reach.

Practical values. Over the three reference contexts:

quantity	range
baseline b	0.047 – 0.097
replicate r	0.958 – 0.978

The baseline is near zero: a predictor with no target-specific information sustains the pass mark over the first gene or two and no further. The replicate is 0.96–0.98, so a half-depth repeat keeps

its ordering pure over very nearly the whole budget, and the span $r - b$ lies between 0.86 and 0.93, second only to the expression error.

Scaled range. The scaled value is not clamped at either end, but there is little room on either side of it. Below, the baseline is already near zero, so the metric's own floor of 0 maps to no worse than -0.12. Above, the raw metric is bounded by 1 and the replicate is already 0.96–0.98, so a submission reproducing the reference exactly scores between 1.02 and 1.05. It **cannot score below 1**, and that is structural rather than measured: the raw metric is bounded by 1, so $r \leq 1$, hence $(1 - b) \geq (r - b)$ and the scaled value at exact reproduction is at least 1 for any $r > b$. The same argument covers the other two members whose perfect raw value is their own upper bound, in sections 3 and 5. This is the member that sits closest to that floor: with the replicate within 0.04 of the metric's ceiling, there is almost nothing left between a half-depth repeat and a perfect one.

CHAPTER 5

DGE significant set agreement

`de_wilcoxon_sig_jaccard` measures whether a submission identifies the same genes as responding that the measurement did, charging a missed gene and an invented one alike.

Definition. Let R_p be the set of genes the reference calls significant for perturbation p and $\hat{R}_p$ the set the submission calls significant, both formed under the filter and threshold of section 3, and both with p's own target gene removed. The metric is the Jaccard index of the two,

$$J_p \;=\; \frac{\left|R_p \cap \hat{R}_p\right|}{\left|R_p \cup \hat{R}_p\right|},$$

counting each (perturbation, gene) pair once. Unlike a recall, whose denominator is $|R_p|$, and unlike a precision, whose denominator is $|\hat{R}_p|$, the Jaccard index is symmetric: a gene the submission fails to call and a gene it calls without cause cost the same. The metric's value for a context is the mean of J_p over its 300 perturbations.

Theoretical limits. $J_p \in [0, 1]$, reaching 1 when the two sets are identical and 0 when they are disjoint. An empty union, meaning neither side called anything, is defined as 1: both sides agree there was no response. That convention is why the metric returns a finite value for every perturbation, so its cohort is the full 300 where section 3's is not, and it interacts with target-gene exclusion: a perturbation whose only reference-significant gene was its own target now has an empty reference set, so a submission calling nothing for it scores 1 where it would otherwise have scored 0. There is no chance correction. Two independent sets of sizes a and $\hat{a}$ drawn from G genes overlap by about $(a\hat{a}/G)$ / $(a + \hat{a} - a\hat{a}/G)$, which at the set sizes seen here is 0.006 to 0.012, so the metric credits that much free agreement.

Practical values. Over the three reference contexts:

quantity	range
baseline b	0.021 – 0.037
replicate r	0.375 – 0.423

The baseline is close to the chance level above: the mean-response predictor's significant set overlaps the reference's by only a few per cent. The replicate is 0.38–0.42, far short of 1, so at half depth the measurement's two halves agree on well under half of the genes in their union. Most of the set disagreement a submission is charged with is therefore the reference's own instability rather than the submission's error, which is exactly what the scale removes. The span $r - b$ is 0.34–0.39 and the standard deviation of r over the five splits is 0.8–2.2 % of it.

Scaled range. The scaled value is not clamped at either end. Below, the baseline is near zero, so the metric's own floor of 0 maps to only -0.05 to -0.11. Above, because the replicate is low, the headroom is the largest of the six: a submission reproducing the reference exactly scores 2.48–2.85 here, against 1.00 for the expression error and 1.03–1.17 for perturbation discrimination. A submission whose set agreement exceeds a half-depth replicate's therefore gains more on the average than an equally large excess would gain elsewhere.

CHAPTER 6

DGE fold-change accuracy

`de_wilcoxon_lfc_nmae` measures whether a submission gets the size of the response right, where sections 3 to 5 read only its direction and its membership.

Definition. The gate is the reference's own significant set for perturbation p: genes the reference calls significant that also carry a finite reference $\log_2$ fold change, with p's target gene removed. Write S_p for that set. The metric is the mean absolute error of the predicted fold changes over the gate, normalized by the mean absolute reference fold change over the same gate,

$$\mathrm{NMAE}_p \;=\; \frac{\sum_{g \in S_p} \left| \widehat{\mathrm{lfc}}_{p,g} - \mathrm{lfc}_{p,g} \right|}{\sum_{g \in S_p} |\mathrm{lfc}_{p,g}|}.$$

A gene in the gate that the submission's table does not carry, or carries as null or non-finite, is read as a predicted fold change of zero, that is, as a prediction of no change. The metric's value for a context is the mean of NMAE_p over the perturbations it returns.

Gate size. A perturbation is scored only if its gate holds at least 10 genes; a ratio over a handful of just-over-threshold genes is noise rather than measurement. A perturbation whose gate is empty, or whose denominator is zero, is omitted for the same reason. All three conditions read the reference alone, so the surviving set is identical for every submission: it ranges 209–261 of the 300 on the three reference contexts. That property is load-bearing rather than incidental, and it is why a non-finite prediction is filled with zero rather than masked. Masking would let a submission emit non-finite values until a perturbation fell below the gate size and disappeared from its own aggregate.

Theoretical limits. Lower is better and there is no upper bound. Perfection is 0. The no-skill point is 1 and follows from the definition rather than from the data: a submission predicting zero fold change everywhere makes the numerator equal the denominator term by term. That identity is exact in fold-change space. Because the competition computes the submission's fold changes against the measured control rather than against the submission's own control cells, a submission broadcasting the true control is not compared against itself and lands near 1 rather than at it: measured 1.008, 1.008 and 1.010 on the three reference contexts.

Practical values. Over the three reference contexts:

quantity	range
baseline b	1.0009 – 1.0017
replicate r	0.369 – 0.431

The baseline sits at 1.000 to within 0.002, so the mean-response predictor's fold changes are, on average over the gate, no closer to the truth than predicting no change at all would be. The replicate is 0.37–0.43: a half-depth repeat reproduces its own fold changes with between a third and a half of the error a null prediction makes. The span $r - b$ is 0.57–0.63, and the standard deviation of r over the five splits is 0.7–2.7 % of it, despite this member running the smallest and most variable cohort of the six.

The replicate estimator. This is the one member whose replicate is not the ordinary split-half comparison. A half of the data calls far fewer genes significant, so a split-half gate would be 21–35 % smaller than the gate the metric itself uses, and the two numbers would be means over different cohorts. Its anchor instead takes the gate and the denominator from the full reference table and only the two fold-change vectors from the halves, so its cohort matches the metric's exactly. The reference bundle records which estimator produced each anchor, and a mismatched pairing is refused rather than scored.

Scaled range. The scaled value is linear, as for every other member, and floored at -6. This is the one floor in the suite that is not simply a metric's own range: the fold-change error is unbounded above, so without it a single badly scaled submission could move the average without limit. Linear means a submission further from the truth than the baseline is charged in proportion: a raw value of 2.0 scores -1.58 to -1.75 on the reference contexts, and the floor binds only at a raw value of 4.4 to 4.8. There is no explicit high clamp, but the member is bounded above all the same: a normalized absolute error cannot fall below zero, so no submission can score past the value earned at a raw error of 0 – 1.58–1.75 on the reference contexts, which is what a submission reproducing the reference exactly earns.

CHAPTER 7

Submission file requirements

Two contracts, at two stages — and they deliberately differ. Section 0's *Submissions* paragraph is `cell_eval2`'s SCORING contract: what the scorer must be handed. This chapter is the Arc `vcc` platform's UPLOAD contract: what a participant uploads, as `vcc prep` enforces it. They are not the same file, and two of the differences look like contradictions until the middle step is named:

	upload (`vcc prep`)	scoring (`cell_eval2`, section 0)
non-targeting rows	forbidden	required, label must match the reference
cells per perturbation	exactly 400	unconstrained
perturbation column	`target_gene`	`pert_col`, whatever the run declares

The platform bridges them. It reads the held-out real control cells out of the reference and concatenates them onto the prediction, and it stamps the scorer's perturbation column from `target_gene`. So the control label is absent on upload and present at scoring, and that is also why a participant may omit controls at all: the control side of every metric is then provably held-out reference data rather than anything the submission supplied (`control_source: real`).

Not all of these protect a metric's value. The counts rules do, and so does the perturbation set. The column name, the file packaging and the size cap are format and resource constraints of the upload — `cell_eval2` treats the perturbation column name as non-semantic (the packaged preset says `pert_col: target`, the production bundles were built with `perturbation`), so getting it wrong fails the upload rather than mis-scoring. Every one of them is a rejection rather than a repair.

One file, all three contexts. A single file covers every cell context, and every cell carries a `context` label in `.obs`, reused exactly as it appears in the control files supplied for the phase: `A/B/C` during the validation phase, `D/E/F` in the final phase. The contexts are scored separately and averaged, so a mislabelled cell is scored against the wrong reference.

The perturbation column is `target_gene`, holding **gene symbols** — `ADNP` — not construct identifiers — `ADNP-1`. The panel is gene-level, and downstream the label is what resolves each row's own target gene for the exclusions sections 1 to 6 rely on: a label that does not resolve keeps the perturbation's own transcript in the scored vector, a coordinate predictable from the label alone.

Exactly the listed perturbations, and exactly the listed shape.

An extra perturbation is rejected rather than ignored, and the cell count is the same for every perturbation, so a different one is rejected too. Rows labelled `non-targeting` are rejected: the control cells supplied with the phase are model **inputs** only and must not be copied into the prediction.

Raw counts in `.X`. The same rule section 0 states, and the one requirement here that is purely about scoring: values must be non-negative, **integral**, finite, and with **no cell total above** 10^6 (`max_counts_per_cell`). Counts are required because the expression metrics compare

requirement	value
perturbations	exactly those in `pert_counts.csv` — no extras, none missing
non-targeting rows	none
cells per perturbation	exactly 400, in every context
genes in `.var`	exactly 18,533, in the order given by `gene_names.csv`
stored entries in `.X`	at most 4,750,000,000

group-summed profiles, which cannot be recovered from log-normalized data — so do not apply `normalize_total` or `log1p`. A matrix failing any of these is rejected rather than scored in the wrong space.

The size cap is on density, not on file size. At the standard shape — 300 perturbations $\times$ 400 cells $\times$ 3 contexts, so 360,000 cells — 4,750,000,000 stored entries is about 13,200 per cell. Predictions compress extremely well, so a small file can still be far over the limit. The count is of entries STORED, which for a dense array is every slot including the zeros: a dense matrix carries all 18,533 genes for every cell, so 6,671,880,000 entries, or $1.40\times$ the cap, and is over it on its own with no adversarial content at all. Store the prediction sparse, without explicit zeros.