

The vcc2026 Metric Suite

Reference

2026/08/19

Contents

0	Scope and parameters	1
1	Perturbation discrimination	3
2	Expression error	4
3	The differential-expression table	5
4	DGE direction fidelity	6
5	DGE direction reach	7
6	DGE significant set agreement	8
7	DGE fold-change accuracy	9
8	Reference points	10

CHAPTER 0

Scope and parameters

The `vcc2026` profile scores six metrics. Parameter values are those of the packaged `configs/vcc2026.yaml`. Measured values are read from the official reference bundles built with `cell-eval2 0.15.0`, competition `rule_version 3`.

Submission format. A submission is a matrix of predicted single-cell expression.

- The gene axis and the perturbation labels match the reference exactly, including the non-targeting control label.
- The number of predicted cells per perturbation is unconstrained.
- Values are raw counts: non-negative, integral, no cell total above 10^6 .
- A matrix failing any of these is rejected rather than scored.

Evaluation data. Per context: 300 target constructs, one per gene, plus 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. Each context is scored independently against its own reference. The reference's control cells are the origin on both sides: the predicted effect is measured against them, and they are the reference group of the differential-expression test for the prediction as well as for the reference.

Gene restrictions. All six metrics exclude target genes. `pds_cosine` removes all 300 panel target genes from every distance; the other five remove the perturbation's own target gene. The four differential-expression metrics apply the further restrictions of section 3.

The six metrics.

metric	quantity	section
<code>pds_cosine</code>	separability of predicted profiles	1
<code>expr_mse_unbiased_capped_norm</code>	size of the expression error	2
<code>de_wilcoxon_direction_fidelity_yield_raw</code>	correctness of predicted directions	4
<code>de_wilcoxon_direction_reach_raw</code>	depth over which directions stay correct	5
<code>de_wilcoxon_sig_jaccard</code>	agreement of responding-gene sets	6
<code>de_wilcoxon_lfc_nmae</code>	accuracy of predicted fold changes	7

Baseline and replicate. Two reference values are measured on the same reference data and distributed with it. The baseline b is the context's mean perturbation response: an equal-weight average, over the constructs passing a cell-count and knockdown-efficiency filter, of the per-construct mean count vector, assigned identically to every perturbation and 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 are averaged. Each half uses its own control cells.

The score. For a metric with panel aggregate u ,

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

so 0 is the baseline and 1 is a repeat of the experiment. A context's score is the equal-weight mean of the six. Values above 1 and below 0 both occur and are reported rather than clipped.

Parameters.

parameter	value	applies to
bulk_target_sum TS	5×10^4	sections 1, 2
target_sum (per cell, DE)	10^6	sections 3 to 7
max_counts_per_cell	10^6	submission validation
DE test	Wilcoxon rank-sum, two-sided	sections 3 to 7
filter_gene_min_cpm_cell	5 CPM, control cells only	sections 3 to 7
p_adj_threshold α	0.05, Benjamini–Hochberg	sections 3 to 7
fdr_scope	per perturbation	sections 3 to 7
ϵ (fold change)	10^{-9}	sections 3 to 7
REACH_PURITY_FLOOR P_0	0.9	section 5
min_gate_size	10	section 7
n_splits / base_seed	5 / 0	replicate anchor
PRED_TRACE_CAP_K	1.0	section 2

Clamps. Four members are unclamped. `expr_mse_unbiased_capped_norm` is clamped to $[0, 1]$. `de_wilcoxon_lfc_nmae` is linear and floored at -6 .

CHAPTER 1

Perturbation discrimination

`pds_cosine`: whether a predicted profile is closer to its own measured profile than to any other perturbation's.

Profile. Counts are summed over a group's cells, normalized to TS, and log-transformed:

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

with $\mathcal{C}_p$ the cells labelled p and $y_{c,g}$ the raw count of gene g in cell c . Effects are taken against the reference control profile on both sides: $\delta_q = b_q - b_{\text{ctrl}}$ and $\hat{\delta}_p = \hat{b}_p - b_{\text{ctrl}}$.

Definition. For each perturbation p the predicted effect is compared with all $n = 300$ 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, so every distance is computed in the same fixed feature space, and $d_{p,q} = 1$ whenever either vector has zero norm. The score is one minus the normalized rank of the correct match:

$$\text{PDS}_p = 1 - \frac{k_p}{n-1}, \quad k_p = \#\{q : d_{p,q} < d_{p,p}\} + \frac{1}{2} (\#\{q : d_{p,q} = d_{p,p}\} - 1),$$

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

Limits. $\text{PDS}_p \in [0, 1]$, higher is better. A prediction carrying no information about which perturbation is which scores 0.5. A prediction emitting one shared profile, and a prediction whose effect is zero, each score exactly 0.5.

CHAPTER 2

Expression error

`expr_mse_unbiased_capped_norm`: the squared distance from the predicted to the measured profile, as a fraction of the measured profile's distance from the control, with sampling noise subtracted from both.

Sampling correction. For a group of n cells with total count S_p and per-cell totals ℓ_i , a delete-one jackknife:

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

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. It is computed the same way for the prediction, $\hat{C}_p$, and for the reference control, C_{ctrl} .

Numerator.

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

where the prediction's credited correction is bounded both by the reference's own correction and by the submission's across-perturbation spread:

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

with R_g the perturbations retaining gene g and $\bar{b}_{\cdot,g}^w$ their weighted mean.

Denominator.

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

Both squared distances run over every gene except the perturbation's own target gene, and G_p is the number of genes summed. The sampling corrections are left whole. D_p reads the measured data only, so it is identical for every submission scored against a given context.

Definition. The scored metric is the ratio of the two sums over the panel:

$$\text{MSE} = \frac{\sum_p N_p}{\sum_p D_p},$$

so it has no per-perturbation value.

Limits. Lower is better. Both N_p and D_p are signed. Perfection is 0. A prediction emitting the control unchanged has expected numerator equal to expected denominator, so no skill is 1. There is no upper bound. The jackknife is measured 0.32 % high at $\text{TS} = 5 \times 10^4$, so a control-emitting submission reads slightly above 1.

CHAPTER 3

The differential-expression table

Sections 4 to 7 read a table computed as follows, identically for the submission and for the reference, with the measured control as the comparison group on both sides.

Test. Each gene is tested on its own by a two-sided Wilcoxon rank-sum test (Mann–Whitney U), p-value from the normal approximation to the rank sum. The values ranked are counts normalized per cell to a total of 10^6 .

Fold change. Direction and magnitude come from the $\log_2$ fold change:

$$\text{lfc}_{p,g} = \log_2 \frac{m_{p,g}^{\text{pert}} + \epsilon}{m_g^{\text{ctrl}} + \epsilon}, \quad \epsilon = 10^{-9},$$

with m the arithmetic mean of the same normalized values over the group's cells.

Low-expression filter. A gene is tested only if its mean expression exceeds 5 counts per million in the control cells. The gate reads the reference's control group alone, so it admits the same gene set for every submission.

Significance. Benjamini–Hochberg correction is applied after the filter, over the surviving genes only, and within each perturbation rather than across the panel. A gene is significant for perturbation p when $p_{p,g}^{\text{adj}} < \alpha$ with $\alpha = 0.05$.

Adjudicability. A gene is adjudicable when the reference assigns it a defined, non-zero $\log_2$ fold change: non-null, non-NaN and non-zero. An infinite fold change carries a direction.

CHAPTER 4

DGE direction fidelity

`de_wilcoxon_direction_fidelity_yield_raw`: whether the genes a submission calls as responding move in the right direction, penalizing a submission calling far fewer genes than the reference found.

Definition. Fix a perturbation p . Let n_{real} be the number of genes the reference calls significant. 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 whose predicted and reference fold changes carry the same sign. Both counts exclude the perturbation's own target gene.

$$F_p = \frac{k}{\max(n_{\text{pred}}, n_{\text{real}})},$$

equivalently a directional precision times a coverage term capped at 1:

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

The metric's value for a context is the mean of F_p over the perturbations where it is defined.

Limits. $F_p \in [0, 1]$, higher is better. It reaches 1 only when every call is correct and the submission calls at least as many genes as the reference found. Random directions score 0.5 in expectation once coverage reaches 1.

Conventions. A gene the submission calls but leaves without a direction stays in n_{pred} and counts as a miss. A submission calling nothing scores 0 whenever the reference found anything. If the reference found nothing either, the value is 0/0 and the perturbation is dropped from the mean. If the reference found nothing and the submission called genes, the value is the bare fraction correct.

CHAPTER 5

DGE direction reach

`de_wilcoxon_direction_reach_raw`: how far down a submission's own ranking its directional calls stay reliable.

Definition. The pool is 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 confidence: its significant calls first, then ascending predicted adjusted p-value, then ascending predicted p-value, then descending $|\text{log fold change}|$, then gene name. 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.

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

taking $k^* = 0$ when no prefix clears P_0 , and

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

“Deepest” is literal rather than “first”: purity is not monotone in k , so a prefix that dips below P_0 and recovers counts at its deeper crossing. The metric's value for a context is the mean of R_p over the perturbations with a non-empty pool.

Limits. $R_p \in [0, 1]$, higher is better. At $P_0 = 0.9$ the shallowest depth at which one wrong call is tolerated is 10: no prefix of nine or fewer genes containing a wrong call qualifies, but a deeper one can, because a leading miss followed by nine matches gives $P(10) = 9/10$ and k^* reaches 10. A wrong call at position j leaves every shorter prefix clean, 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.

CHAPTER 6

DGE significant set agreement

`de_wilcoxon_sig_jaccard`: whether a submission identifies the same responding genes as the reference, charging a missed gene and an invented one alike.

Definition. With R_p the reference's significant set and $\hat{R}_p$ the submission's, both under the filter and threshold of section 3 and both with p 's own target gene removed:

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

counting each (perturbation, gene) pair once. The metric's value for a context is the mean of J_p over its 300 perturbations.

Limits. $J_p \in [0, 1]$, higher is better, 1 when the sets are identical and 0 when disjoint. An empty union is defined as 1, so the metric returns a value for every perturbation. There is no chance correction.

CHAPTER 7

DGE fold-change accuracy

`de_wilcoxon_lfc_nmae`: whether a submission gets the size of the response right.

Definition. The gate S_p is the reference's own significant set for perturbation p , restricted to genes carrying a finite reference $\log_2$ fold change, with p 's target gene removed.

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

A gene in the gate that the submission does not carry, or carries as null or non-finite, is read as a predicted fold change of zero. 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 perturbation whose gate is empty, or whose denominator is zero or non-finite, is omitted. All conditions read the reference alone, so the surviving set is identical for every submission.

Limits. Lower is better, no upper bound. Perfection is 0. A submission predicting zero fold change everywhere makes numerator equal denominator term by term, so no skill is 1.

Replicate estimator. This member's anchor 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. The reference bundle records which estimator produced each anchor, and a mismatched pairing is refused rather than scored.

CHAPTER 8

Reference points

Measured on the three official reference bundles at `cell-eval2 0.15.0`, `rule_version 3`. Ranges are over contexts A, B and C.

member	b	r	span $ r - b $	cohort
pds_cosine	0.500	0.927 – 0.984	0.43 – 0.48	300
expr_mse_unbiased_capped_norm	0.986 – 0.992	0.028 – 0.045	0.95 – 0.96	panel ratio
de_wilcoxon_direction_fidelity_yield_raw	0.505 – 0.522	0.795 – 0.832	0.28 – 0.33	295 – 300
de_wilcoxon_direction_reach_raw	0.047 – 0.097	0.958 – 0.978	0.86 – 0.93	290 – 300
de_wilcoxon_sig_jaccard	0.021 – 0.037	0.375 – 0.423	0.34 – 0.39	300
de_wilcoxon_lfc_nmae	1.0009 – 1.0017	0.369 – 0.431	0.57 – 0.63	209 – 261

pds_cosine’s baseline is exactly 0.500000 on all three contexts: with all 300 panel target genes removed, the mean-response submission carries no discriminating information and every row ties.

Scaled value of an exact reproduction of the reference.

member	scaled score	clamp
pds_cosine	1.03 – 1.17	none
expr_mse_unbiased_capped_norm	1.000	[0, 1]
de_wilcoxon_direction_fidelity_yield_raw	1.51 – 1.72	none
de_wilcoxon_direction_reach_raw	1.02 – 1.05	none
de_wilcoxon_sig_jaccard	2.48 – 2.85	none
de_wilcoxon_lfc_nmae	1.58 – 1.75	floor –6

Seed noise. The standard deviation of r over the five splits, as a fraction of the span the score divides by: 0.5 – 1.3 % (pds_cosine), 0.8 – 2.2 % (expr_mse_unbiased_capped_norm), 1.5 – 4.8 % (fidelity), 0.2 – 0.9 % (reach), 0.8 – 2.2 % (sig_jaccard), 0.7 – 2.7 % (lfc_nmae). The splits are seeded and reproducible, so this is the scale on which a difference is too small for the metric to resolve, not a run-to-run error bar.