Effect-table transfer, call lists and capped-MSE calibration: an attributed account of a Virtual Cell Challenge 2026 entry

Li Meng, Team UiO

Draft of 6 November 2026, revised 23 November 2026 with the final results

Abstract

The Virtual Cell Challenge 2026 asked for zero-shot predictions of 300 CRISPRi knock-downs in three unseen human cell contexts, 400 cells each, scored by cell-eval2 on six metrics. Our entry is not a single model. It starts from AtlasShift, a public atlas-transfer baseline, and keeps its skeleton: per-target effects measured in public screens are moved onto each context's own control cells, and its count generator produces the cells. On top of that skeleton we added five parts. Noise-aware effect tables shrink each source by its own sampling noise, fuse sources by unit direction and remove the context's control axes. Residual heads on frozen embeddings from 16 foundation models only amplify changes the table already makes. A call list raises genes just past a per-gene detectability line, with its direction taken from a pooled posterior and projected on near-line axes from ARCHS4 and GPSAdb. Tables from two external predictors, UPM and PIE, are added after readings on cell lines that the predictors and our sources have not seen. A fill rule sets each context's amplitude from an estimate of the capped MSE's budget-to-claim ratio. On our local folds, after the last board reading, the steps inside our baseline gave +0.060, and the bulk mix, the call list with its near-line axes and an emission correction a further +0.036 (native score, three folds merged, eight seeds). Further steps worth about +0.02 are estimates, not measurements. Counted by what each step changed, foundation models gave about a tenth of our improvement, mostly in the first residual heads and in three external tables whose gains are estimates. Our best validation-phase upload scored 0.1882 (rank 347 of 1,635 on the frozen board). On the final leaderboard, scored on three other cell lines, the entry reached 0.2594 and ranked 41 of 807, below the roughly 0.32 that our local folds had predicted. According to the organisers' wrap-up, the three winning entries (0.4399, 0.3348 and 0.3263) share our structure: effects measured in other cell lines, adapted to the new line through its control cells. They differ from ours in the amount of reference data, in weights fitted to each new line and in response models learned across many cell contexts. We had judged public tables one at a time as extra sources of a fixed average, a test that suits a fixed table and not the data a model learns from. We separate every component by origin, list what failed, compare our entry with the winners, and release a 600-line numpy re-implementation of the parts that are ours.

1 Introduction

Perturbation-response prediction across cell types is usually framed as learning a model that maps a perturbation and a cell state to a response. In a zero-shot setting with unseen cell lines and a fixed evaluator, much of the achievable score comes from elsewhere: from which public measurements are transferred, how their noise is handled, and how the prediction is rendered against each metric. The organisers' wrap-up of the final results draws a similar conclusion from the winning entries, that reference data and the care of its transfer to a new context counted at least as much as model architecture [3]. This report describes what we did for the Virtual Cell Challenge 2026 [1]. Its first aim is attribution: our pipeline reuses other people's methods and data at almost every step, and we state for each part whose it is. Its second aim, added after the final results, is a comparison. Our entry ranked 41 of 807 on the final leaderboard with a score of 0.2594, and Section 8 sets it against the three winning entries and states what we would change.

What is ours. We claim the following and nothing else.

1. [OURS] Noise-aware construction of effect tables: per-source shrinkage of bulk-space effects with an adaptive-shrinkage prior [24] fitted per expression decile after calibrating the Poisson noise variance, unit-direction fusion across sources, per-source weights for the common response, and a completion of the genes the K562 genome-wide screen does not measure, regressed on the context's control principal components.
2. [OURS] Residual heads on frozen foundation-model features, cross-fitted, whose table enters the bulk path with its norm restored and only where it amplifies the existing change.

Table 1: Validation-phase anchors (b, a) used to normalise each metric (from the challenge's evaluator configuration). The final leaderboard normalises with anchors of the test lines, which differ from these.

	pds	mse	nmae	fid	reach	jac
b	.5079	.9878	.9967	.5128	.0783	.0306
a	.9484	.0378	.4085	.8159	.9696	.3532

3. [OURS] The call list: a per-gene detectability threshold computed from the context's own controls, a posterior direction projected on near-line axes (within-series ARCHS4 residuals [18] and GPSAdb signatures [12]), a confidence-dependent lift, and a floor that zeroes listed genes whose direction disagrees.
4. [OURS] A procedure for external predictors: direction-only use (row norms matched to ours), deflation of the predictor's own top axes, agreement weights, and a reading on clean cell lines.
5. [OURS] The fill rule: an estimator of the capped MSE's budget-to-claim ratio from the package alone and a per-context amplitude correction derived from a power law fitted to measured arms.
6. [OURS] The local evaluation design (three folds, eight-seed replacement rule, clean-line arithmetic) and the list of negative results in Section 7.
7. [OURS] The comparison of our entry with the three winning entries and the account of how we judged public data (Section 8). It was written after the final results, and none of the changes it proposes was run.

What is not ours. The pipeline's skeleton, the two-space profile construction and the count generator are AtlasShift's [17]. The evaluator and the property of its capped MSE that our fill rule uses are cell-eval2's [4]. The pooled posterior table is Sidechain's [23]. The idea of placing confident genes just past the detection threshold also appears in a public recipe (T3V200 [16]), and the idea of removing control axes was described in public board notes of several teams. All foundation models, TabPFN [13], ESM-2 [19], GenePT [7], UPM [26], PIE [5] and every dataset are other people's work. Table 2 lists them with their licences. The descriptions of the winning entries in Section 8 follow the organisers' wrap-up [3], and those methods belong to their authors.

2 Task and metric

The challenge had two phases. In the validation phase three contexts (A, B, C) were scored on a public leaderboard. That board was frozen on 22 October, when the three test contexts (D, E, F) were released. The final phase scored the teams' final entries on D, E and F, and its results were published on 19 November [3]. Each test context supplies 18,400 non-targeting control cells over 18,533 genes (10x Flex, median about 20k UMI per cell), and the validation contexts had the same layout. A submission holds 400 cells for each of 300 CRISPRi targets in each context: 360,000 cells. The method sections describe the final package for D, E and F.

Contexts were not named during the challenge. From expression alone, D resembled MOLM-13 (Pearson $r = 0.904$ with public profiles) and E resembled U-87 MG (0.836), and F had no clear match. We did not use these guesses in the build. The wrap-up names the test lines as AC16 (a cardiomyocyte line), MOLM13 and U-87MG, and the validation lines as Cal-33, HeLa and Jurkat [3]. This agrees with our readings of D and E and leaves AC16 for F. By the wrap-up, 129 of the 300 test targets were also on the validation panel and 171 were new, and all but one of the 300 had public perturbation data in some other cell line.

The score is the mean of six metrics, each normalised as $(u-b)/(a-b)$ between a baseline b and a reference a (Table 1). The wrap-up describes 0 as the cell context's mean and 1 as an independent replicate of the experiment [3]. Only mse is clipped to $[0, 1]$. Four metrics (nmae, fid, reach, jac) read the per-cell mean and the per-gene test calls, and two (pds, mse) read the pseudobulk. AtlasShift's two-space rendering follows this split: a *mean* profile and a *bulk* profile per target, both matched by its generator [17].

The capped MSE (cell-eval2's). The mse term is an unbiased squared distance between predicted and true pseudobulk means, with sampling corrections on both sides. The prediction's correction is capped per row by the reference's (cell-eval2 issue #247) and in total by a budget B, the submission's own across-perturbation centred sum of squares (issue #348), allocated to rows in proportion to what they claim. The evaluator's documentation notes the consequence that we use: injecting centred variation z into a scored gene costs $\|z\|^2$ in the plug-in distance and raises B by exactly $\|z\|^2$ [4]. Hence, while B is below the claimed correction, extra energy along the predicted direction costs nothing and only its alignment with the truth counts.

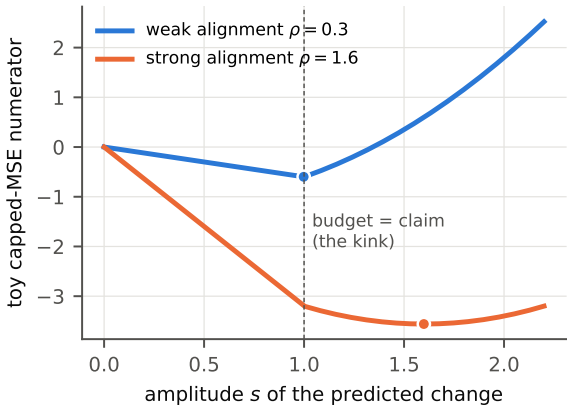


Figure 1: Toy capped-MSE numerator against amplitude (ours, with cell-eval2's definitions of budget and claim). Below the kink the numerator is linear in s, and the minimum is at the kink when alignment is weak.

A toy model (ours). Let the prediction be a fixed unit direction with amplitude s, energy s^2E, inner product $s\rho$ with the truth, and claimed correction κ. With $B = s^2E$,

$$N(s) = c_0 + s^2E - 2s\rho - \min(s^2E, \kappa) = \begin{cases} c_0 - 2s\rho, & s^2E < \kappa, \\ c_0 + s^2E - 2s\rho - \kappa, & \text{otherwise}, \end{cases} \tag{1}$$

minimised at $s^\star = \max(\sqrt{\kappa/E},\, \rho/E)$ (Figure 1). A weakly aligned prediction should be filled exactly to the kink, and past it the quadratic cost returns. This is a scalar caricature. The real term has per-row caps and weights.

3 Building blocks from others

Table 2 lists every external method, model and dataset in the final package, with what it does there. Our code release contains re-implementations only of the AtlasShift steps it needs to run (clearly marked) and none of the other items.

4 Method

Figure 2 shows the order of the steps in the final package, with each box tagged as ours or others'. All our changes act between AtlasShift's profile step and its generator.

4.1 Effect tables

Sources are summed counts per target from public screens: K562 GWPS (weight 2), HCT116 and HEK293T from X-Atlas/Orion (1 each), VCC 2025 H1 (2, three tables pooled), and KOLF2.1J (1). For target t in a source with control shares c, AtlasShift's shares are $p_t = (S_t + Pc)/\sum$ with $P = 10^5$, and its effects are taken in the two spaces above, less the common response. We changed the following.

Per-source common weight and pooled inputs. The common-response subtraction gets a weight per source, and pooled count-sum tables enter as one source.

Noise shrinkage (`-eb-shrink ash`). For each source the bulk-space effect e_{tg} has delta-method Poisson variance $v^0_{tg} = f'(p_{tg})^2\, n_t p_{tg}/\mathrm{tot}_t^2$ with $f'(p) = 5 \cdot 10^4/(1 + 5 \cdot 10^4 p)$. Genes are cut into deciles of control share. In each decile we calibrate $\phi = \mathrm{median}\big((e - \tilde{e}_g)^2/v^0\big)/\mathrm{median}(\chi^2_1)$, with $\tilde{e}_g$ the gene's median, and set $v = 0.85\,\phi\, v^0$. The factor 0.85 scales the noise variance, so the shrinkage is slightly weaker than with ϕ alone. The centred effect $d = e - \bar{e}_g$ is replaced by its posterior mean under a zero-mean normal mixture with a point mass at zero [24], fitted by EM on a grid from $0.1\times$ the median noise sd to $2 \max|d|$ (at most 60 scales). Each row is then scaled back to its norm before shrinkage, so only its direction changes. In the K562 table the calibrated ϕ rises from 1.14 in the lowest decile to 1.98 in the highest.

Table 2: Components of the final package that are other people's work. "Idea" means a concept we rewrote with no code reused.

Component	Licence	Use in our entry
AtlasShift [17]	MIT	Skeleton of the generator: source effects with prior counts $P = 10^5$, mean space $\log_2 \frac{10^6 p+1}{10^6 c+1}$ and bulk space $\log(1 + 5 \cdot 10^4 p) - \log(1 + 5 \cdot 10^4 c)$, common-response subtraction, weighted fusion, the CD4 Marson family, desired profiles $c\, 2^{\mathrm{clip}(ae)}$ and $\exp(\max(\log(1 + 5 \cdot 10^4 c) + \mathrm{clip}(ae), 0)) - 1$, promoter prior (TSS $\leq$5 kb) and the dual-moment count generator. Our final amplitudes 1.2/0.5 (defaults 0.6/0.3) and promoter fraction 0.05 (default 0.15).
cell-eval2 [4]	—	Evaluator (preset `vcc2026`). Its pseudobulk, delta and jackknife functions run inside our budget estimator.
Sidechain [23]	MIT	`pooled_delta` posterior over the board's sources: part of the bulk-mix table and the direction L of the call list.
T3V200 recipe [16]	none	Idea: shared-gene calls just past threshold with sign from a consensus direction, and the idea of a consensus table. Rewritten, no code.
Public board notes	—	Idea: deflate control axes of the target context.
K562 GWPS [22]	CC BY 4.0	Source table (weight 2).
X-Atlas/Orion [14]	CC BY-NC-SA 4.0 (data)	HCT116 and HEK293T source tables (weight 1 each). Removing both from the fusion cost -0.064 [-0.079, -0.050] on our folds (native merged). Two of the three folds are these same lines, each the other's source.
VCC 2025 H1	challenge	Source table (weight 2).
KOLF2.1J [21]	CC BY 4.0	Source table (weight 1).
CD4 T cells [27]	see source	AtlasShift's CD4 family.
DLD-1 Perturb-seq [25]	see source	Published z-score and fold-change matrices. Rows for 75 of the 300 targets, only in the consensus table of the bulk mix.
Pan H1 Perturb-seq [8]	see source	Rows for 55 of the 300 targets, only in the same consensus table. Its targets outside VCC 2025 H1 also served as a clean line (Section 4.5).
Essential-gene screens: K562, RPE1 [22], HepG2, Jurkat [20]	see sources	Training rows of the residual heads, beside H1, HCT116 and HEK293T rows.
ARCHS4 [18]	see source	Near-line within-series axes for the call list.
GPSAdb 2.0 [12]	CC BY 4.0	Near-line signature axes for the call list, and the GPSAdb-based bulk-mix table.
16 foundation models (Table 12)	per model	Frozen cell and gene embeddings of each context's controls.
ESM-2 [19], GenePT [7]	MIT, CC BY 4.0	Target-gene features (ESM-2 taken from SE-600M's protein embeddings).
TabPFN v2 [13], TabPFN-3.5	Prior Labs (v2), non-commercial (3.5)	Regressor of the 16 route heads (v2) and of the four joint heads (3.5).
adaptive shrinkage [24]	method	Prior family of our noise shrinkage (re-implemented).
UPM [26]	CC BY-NC 4.0	External table, weight 0.2 with agreement weights.
PIE [5]	CC BY-NC-SA / NC	External table $E = p_{\mathrm{de}} \times \mathrm{lfc}$ (xdataset checkpoint), weight 0.4. Earlier, the PIE line produced the call-floor mechanism.
State SE-600M [2]	see source	One of the 16 FM routes. The head table of this single route is also added as an external table, weight 0.2.

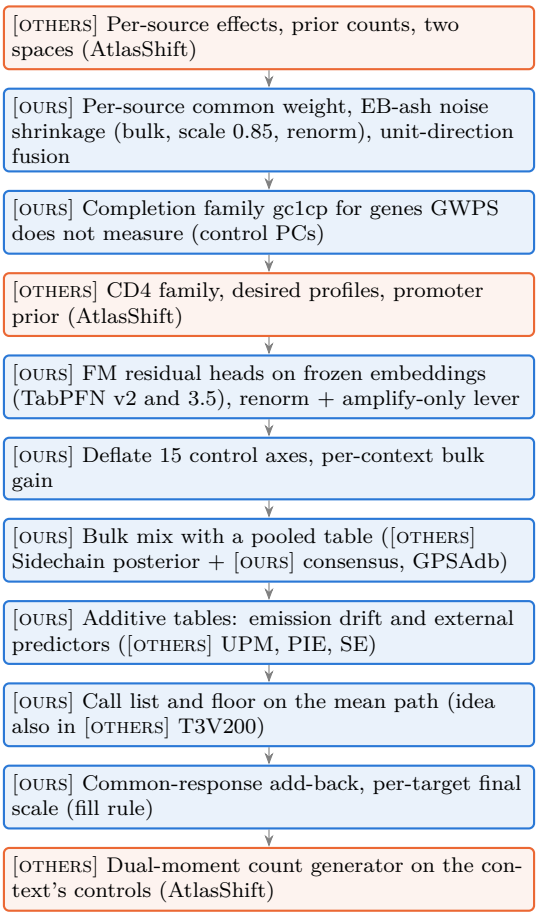


Figure 2: Order of operations in the final package. Blue boxes tagged [OURS] are our steps, and orange boxes tagged [OTHERS] are AtlasShift's. Components inside a box that belong to others are tagged in place.

Unit-direction fusion. On the bulk path each source's row enters the weighted mean divided by its norm, and the fused row is scaled to the norm of the plain weighted mean. Every source then has the same say in the direction. The median norm ratio of unit to plain mean was 0.118 on context D.

Completion family. K562 GWPS measures 7,681 genes, and in our folds 47% of the truth's energy lay on genes it does not measure. A completion of those genes, regressed on the target context's control principal components and fitted on full-panel sources, enters as an AtlasShift-style family with weight 2 and no common subtraction.

4.2 Foundation-model residual heads

Each of 16 foundation models (Table 12) embeds each context's control cells. Per target, features are the model's cell and gene representations, ESM-2 and GenePT embeddings of the target gene, a cross-fitted source-evidence table R and context axes. TabPFN v2 fits $Y - R$ on 32 label principal components (TabPFN-3.5 for the four joint heads). Training rows are targets measured in seven cell contexts: the K562 and RPE1 essential-gene screens [22], the HepG2 and Jurkat screens [20], H1, HCT116 and HEK293T (10,252 rows for the test contexts). Predictions for a target come from heads that did not see it. The final residual is

$$\mathcal{R} = 0.5\,\overline{H}_{16} + 0.4\,\overline{J}_4 - 0.9\,R_{\text{mean}}, \tag{2}$$

the mean of the 16 single-model heads, the mean of 4 heads on joint records, and the record mean. It is applied to the bulk profile as $p\,e^{\mathcal{R}}$, after which each row's change from control is scaled back to its previous norm, and the step is kept only where it has the sign of the row's existing change (the "amp" lever, which kept 32.6% of the step's energy on D). The heads are therefore a reweighting of the table's own changes, not a source of new directions. The 16 routes agree closely: each has cosine 0.9946–0.9988 with their mean. On the base of 27 October, removing the whole residual changed the native score by -0.0003 $[-0.0012, +0.0007]$ (Section 6).

4.3 Bulk path

Control-axis deflation. The top 15 principal axes of the context's control cells (centred $\log(1+\mathrm{CP50K})$) are projected out of each target's bulk change. The median removed share of energy was 2.2% on D. The idea came from public board notes, and the implementation and setting are ours.

Gain. Each context's bulk change is multiplied by one factor: D 1.458, E 1.546, F 1.527.

Pooled-table mix. With d the rows' changes and s their per-gene mean over targets (own genes excluded), the target-specific part $c = d - s$ is replaced by

$$c' = \sqrt{\|c\|^2/\|m\|^2}\, m, \qquad m = \mathrm{centre}\big(u(c) + \lambda\, u(C)\big), \quad \lambda = 2, \tag{3}$$

with $u(\cdot)$ the whole-table unit norm. C is a pooled table. In arm e1 it is the sum, each at unit norm, of the Sidechain `pooled_delta` posterior with 5 control axes removed and our consensus table with 15 removed. The consensus table is rewritten from T3V200's consensus idea. Every screen's table is shrunk entry by entry by $z^2/(z^2+25)$, with z the entry's signal-to-noise ratio, scaled to unit norm, averaged within its lab and then across labs with equal weight. The labs are those of our five sources and the CD4 family, and two screens that enter the package only here: DLD-1 [25], with rows for 75 of the 300 targets, and Pan H1 [8], with 55. In the final package (step d3) the same table is moved twice the distance toward a GPSAdb-based table on the quarter of targets where that table changes it most. On D the mixed table has cosine 0.903 with C and 0.783 with the table before.

Additive tables. A table D is added in $\log(1 + 5\cdot 10^4 p)$ space, cut at zero and renormalised. In the final package $D = 2\,\Delta_{\mathrm{emit}} + 0.2\, w \odot T_{\mathrm{UPM}} + 0.4\, T_{\mathrm{PIE}} + 0.2\, T_{\mathrm{SE}}$. Δ_{emit} is the mean difference between requested and emitted bulk profiles over repeated generator runs (yed2). Its gain on our folds was in pds, and we did not establish why. The external tables are described in Section 4.5.

4.4 Call list and floor

A gene counts toward the DE metrics only when the test calls it. For a context's controls in counts per 50k, with gene mean μ_g and sd σ_g, the change needed for a call with 400 cells is

$$\nu_g = \log_2\Big(1 + \frac{(z+1)\,\sigma_g}{0.95\sqrt{400}\,\mu_g}\Big), \qquad z = \Phi^{-1}\big(1 - \tfrac{1}{2}\cdot 0.05\cdot 500/G\big), \tag{4}$$

and $\nu_g = \infty$ for genes never seen. The direction comes from the Sidechain posterior $\log_2$ change L over the board's sources, mixed with its projections on two near-line bases and rescaled to L's whole-table norm (not row by row):

$$L' = \frac{\|L\|}{\|\tilde{L}\|}\,\tilde{L}, \qquad \tilde{L} = L + 0.5\,\Pi_A L + 0.5\,\Pi_G L. \tag{5}$$

Π_A projects on the top 1,000 axes of ARCHS4 within-series residuals of the 4,000 public samples most correlated with the context's controls (library $\geq 2\cdot 10^6$, $\log_2$ CPM, series with at least 4 samples). Π_G projects on the top 1,000 axes of 1,500 GPSAdb signatures (own-gene knock-down ≤ -0.5, same line excluded) ranked by similarity of their control means to the context's control on its 3,000 top genes. A pair (t, g) is listed when $16|L'_{tg}| \geq \nu_g$, g is not t's own gene, $L'_{tg} \neq 0$ and $1.05\,\nu_g \leq 1$. Its magnitude is $1.05\,\nu_g(1 + 0.2x)$ with $x = \mathrm{clip}(\log_2 \max(16|L'|/\nu_g, 1), 0, 4)/4$, and its sign is that of L'.

On the mean profile, for each listed pair the current change r (relative to the row's median change) is compared with the list: if their signs disagree the entry is set to no change. Otherwise, if $|r|$ is below the floor $\min(\max(0.02, \mathrm{mag}), 1)$, it is raised to the floor in its direction. On D the list held 791,964 pairs, of which 578,100 were raised and 195,584 set to no change. The sign-zero rule came from our PIE experiments (Section 7).

4.5 External predictors

A table u from an external predictor and our row a (both centred over targets, own genes excluded) give the increment $\lambda\, w_t\, T_t$ with

$$T = \mathrm{defl}_{3,0.25}\Big(u\,\frac{\|a_t\|}{\|u_t\|}\Big), \qquad w_t = 1 + \kappa\,\big(\mathrm{rank}_{[0,1]}\cos(a_t, u_t) - \tfrac{1}{2}\big)\,2, \tag{6}$$

Table 3: Local folds. N_{conf}: number of genes called per target in the truth.

Fold	Line	Targets	Cells per target (median)	N_{conf} mean (median)
f1	HCT116 (X-Atlas/Orion)	299	151	13.7 (2)
f2	HEK293T (X-Atlas/Orion)	300	200	26.1 (1)
f3t	H1, thinned to ∼20k UMI	150	∼370	1040.9 (204.5)

where $\text{defl}_{3,0.25}$ keeps a quarter of the table's own top three across-target axes. The predictor contributes direction only. Deflation was needed because on one fold the external table put back the shared axes our deflation had removed, and these were the truth's hub genes there (top-5 slot share 0.074 to 0.242, pds -0.065). The settings are $\lambda = 0.2$ and $\kappa = 0.5$ for UPM, $\lambda = 0.4$ with flat weights for PIE, and $\lambda = 0.2$ with flat weights for the SE table, the head table of the single SE-600M route.

Clean lines. Both predictors were trained on lines in our folds, so fold reads cannot separate leakage from gain. The SE table exists only where the 16 routes were run (our folds and D, E, F). On the clean lines we therefore read a non-FM proxy in its place, the mean of six source tables, which has cosine 0.96 with the processed SE table, so the SE model's own part is unmeasured. We read them on five lines outside their training data and outside our source pool: H1 targets outside VCC H1 [8], two A549 screens [10, 11], TeloHAEC [15] and the strong subset of a macrophage screen [9]. The rows read on these lines were built without the bulk mix and without the call floor, so the Pan H1 screen, which enters the package only through the bulk mix, was not in the rows read against it. KOLF2.1J looked clean but is in our source pool, which made its reading void. On a clean line we convert a cosine ratio and a pds change into a planning number

$$v = 2.83 \cdot 0.15\,(c/c_0 - 1)/0.95/6 + \Delta\text{pds}/0.4405/6, \tag{7}$$

where 2.83 is our linearisation of the normalised capped term in the cosine below the kink and 0.15 is an assumed, unmeasured board cosine of the base package. v is not a score.

4.6 Fill rule

For a package we estimate the budget-to-claim ratio

$$\widehat{r} = B \Big/ \textstyle\sum_p w_p \min\left(C_p^{\text{pred}},\ \widehat{C}\right), \tag{8}$$

using cell-eval2's own budget B, row weights and jackknife corrections. The truth's unknown noise term $\widehat{C}$ is replaced by the mean jackknife of five random 400-cell subsets of the context's controls. Over measured arms B scaled as k^p with the amplitude factor k, $p = 1.76, 1.60, 1.40$ for D, E, F. The fill multiplies each target's final bulk scale by $(1.04/\widehat{r})^{1/p}$. The target 1.04 is where the best measured arms sat (1.03–1.05). Our earlier package a sat at 0.730, 0.709 and 0.717, well below the kink. The ratio is re-read after every added table, and the final package reads 1.040, 1.024 and 1.027. On one rebuild the predicted ratios were 1.038/1.07/1.023 and the measured 1.037/1.064/1.012.

5 Evaluation design

Folds. We built three local folds that imitate the board (Table 3). The primary reading is the native score merged over the three folds, and the validation-anchor scale is reported as description only. A candidate is screened on two seeds and replaces the incumbent only after eight seeds. f1 and f2 share 299 targets and serve as each other's sources, so gene-level statistics fitted there are circular. One gene weight that gained +0.011 turned out to be that leak.

Known biases. f1 and f2 are easier than D, E, F (same lab, each other's source), and a learned mapping lost 60–75% of its gain when sibling labels were removed. Seed-to-seed spread of one arm can exceed the difference between candidates (f1 reach from +0.0106 to −0.0141). No public line matches the test contexts in both depth and target count. Under our reading of the lines, only 26 of the 900 target–line pairs of the test contexts had a same-line public measurement, 21 of them in a shallow U-87MG kinome screen that the wrap-up also names [3].

Table 4: Validation-phase board scores of our uploads, in order.

Upload	Content	Score
kd-h1avg	first upload	−0.2332
kd		−0.3019
gwps-tgt		0.0656
gwps-tgt-l1p5		0.0635
dr-es	FM residual heads	0.0778
as-es-amp1	[OTHERS] AtlasShift generator + our heads	0.1820
as-s3a-pac		0.1874
s8c9cfw		0.1744
s8n7pk1g	+ budget-based amplitude	0.1882

Table 5: Final leaderboard, scored on the test contexts D, E and F with the test lines' anchors [6]. Team names follow the organisers' wrap-up [3]. The last row is the median over all 807 entries, taken column by column.

Rank	Entry	Score	pds	mse	nmae	fid	reach	jac
1	BioInsight AI Lab (Illumina)	.4399	.8313	.6842	.4399	.1908	.4288	.0642
2	FlowState	.3348	.8363	.3632	.3286	.1695	.3128	−.0018
3	X-team (Xaira)	.3263	.7346	.4681	.3038	.0659	.3186	.0667
10		.3002	.7609	.3537	.2905	.1218	.2587	.0154
41	Spirited (ours)	.2594	.7627	.2805	.2099	.0975	.2095	−.0035
100		.2209	.7502	.1508	.2002	.0722	.1776	−.0256
	median of 807	.1018	.4620	.0000	.0988	.0160	.0981	−.0120

6 Results

Validation board. Table 4 lists our validation-phase uploads, scored on contexts A, B and C. The largest jump (0.078 to 0.182) came when we adopted AtlasShift's generator, which is not our contribution. dr-es differs from gwps-tgt in its base as well as in its heads, so their gap of 0.0122 is not the heads' value. On the frozen board, which listed 1,635 entries when we read it, our best, 0.1882, ranked 347. The wrap-up counts 1,469 teams on that board [3], and we have not resolved the difference. For reference, rank 1 was 0.472, rank 10 0.3193, rank 100 0.2711. AtlasShift's README reports 0.1546. None of these uploads holds the steps measured after 22 October.

Final leaderboard. The final package (dgxzpsg) was submitted on 5 November and scored on D, E and F. It reached 0.2594 and ranked 41 of the 807 entries on the final leaderboard (Table 5). The wrap-up counts 805 teams with a final entry [3]. Our rank is one of a band. Thirteen other entries lie within 0.004 of ours (ranks 33 to 46), and the wrap-up reports that redrawing an entry's cells under the same pseudobulk moved its score by up to 0.004.

Our distance to the first place is 0.1804 in the mean of the six metrics. The mse term carries 37% of it (a gap of 0.404), nmae 21% (0.230), reach 20% (0.219), fid 9% (0.093), and pds and jac 6% each (0.069 and 0.068). Against the second place we trail on every metric, most on nmae (0.119) and reach (0.103). Against the third we lead on pds (by 0.028) and fid (0.032) and trail on mse (0.188), reach (0.109), nmae (0.094) and jac (0.070). By metric our rank is 40 on pds, 44 on mse, 58 on nmae, 38 on fid, 51 on reach and 304 on jac. Our jac is slightly below the level of the cell-context mean. The wrap-up reports that entries in general called far too many DE genes, and our call list holds about 2,640 genes per target on D.

Forecast and outcome. Our forecast was too high. Before the deadline we had one pair of scores for calibration. Our best validation upload read 0.2622 on the folds (validation-anchor scale) and 0.1882 on the validation board, an offset of −0.074. Applied to the final package, whose fold value of about 0.40 includes the estimated steps, the offset predicted about 0.32. The package scored 0.2594. Three things separate forecast and outcome, and we cannot tell them apart. The offset came from a single pair, made on the validation lines. The folds had been reused for hundreds of comparisons, so the fold value is optimistic by an unknown amount. The test lines and their anchors differ from the validation ones. Set beside the fold means of arm a in Table 8, which are on the validation anchors and belong to an earlier package, the final values are close on pds, fid and jac and lower on mse (0.28 against 0.38), nmae (0.21 against 0.42) and reach (0.21 against 0.52). The shortfall is thus in mse, nmae and reach and not in pds.

Table 6: Steps inside B10, native merged, each against the state when it was measured (not additive).

Step	Gain	Step	Gain
EB-ash shrinkage	+0.0151	bulk gain 1.5	+0.0076
noise scale 0.85	+0.0033	common subtract 1.0	+0.0041
completion family gc1cp	+0.0077	control-axis deflation	+0.0021
unit-direction fusion	≈ +0.0017	KOLF2.1J source	+0.0015
common add-back (pool)	+0.0032	add-back detrend	+0.0010

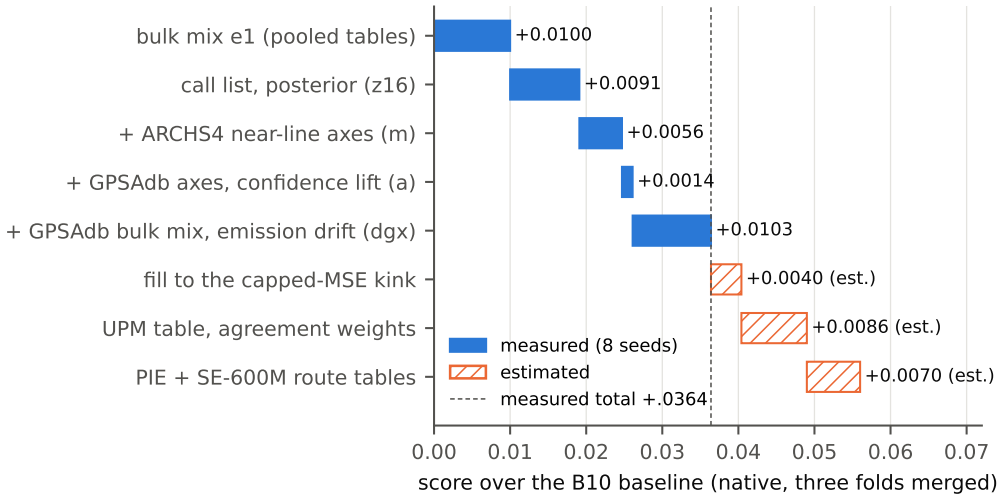


Figure 3: From B10 to the final package. Solid bars were measured on the folds with eight seeds. The e1 bar is read against B10, the a and dgx bars are paired against the previous row, and the call-list and ARCHS4 bars are differences of two readings against B10. Hatched bars are clean-line estimates and the fill's expected gain, not measurements.

Table 7: Measured chain over B10, native merged over three folds with 90% intervals over seeds, and the validation-anchor scale (description only).

Arm	Adds	Native vs B10	Val scale vs B10
e1	pooled-table bulk mix	+0.0100	+0.0236
z16	+ call list from the raw posterior	+0.0191 [+0.0044, +0.0316]	+0.0403
m	+ ARCHS4 near-line axes	+0.0247 [+0.0097, +0.0374]	+0.0468
a	+ GPSAdb axes, confidence lift	+0.0261 [+0.0111, +0.0391]	≈ +0.0491
dgx	+ GPSAdb bulk mix (d3), emission drift	+0.0364 (+0.0103 vs a, se 0.0005)	≈ +0.069

Local folds. After the last board reading, the steps inside our baseline B10 (all effect-table changes, no bulk mix, no call list) gave +0.0601 [+0.0501, +0.0717] over the package of 22 October. Table 6 lists the measured parts of that gain. Each was measured once against the state at the time, so they do not add up. The same gain also holds changes of the head weights, bundled with other steps. On top of B10, Figure 3 and Table 7 give the chain to the final package.

The steps from m to a are small (+0.0014, se 0.0006, positive on 7 of 8 seeds) and were negative on f2 (−0.0045). They passed our replacement rule but may not survive on the board. The call-list step was not measured as a paired arm. The z16 and e1 readings against B10 differ by +0.0091 (−0.0095 on f1, +0.0032 on f2, +0.0337 on f3t), so the list's gain comes from f3t, where the truth calls many genes, and it costs on f1. Table 8 shows that the gains sit in mse, nmae, fid and reach, and pds barely moves.

Estimated steps. Three further steps were packaged on estimates. The fill was expected to give about +0.004 (about +0.0033 per 0.1 of amplitude below the kink). UPM with agreement weights read $v = +0.0086$ on the five clean lines (flat weights +0.0068, per line H1 +0.0069, A549 biohub +0.0192, A549 myllia +0.0048, TeloHAEC +0.0163, macrophage −0.0041). PIE at 0.4 read +0.0058 (permuted control −0.0039, and +0.0017 without the macrophage line). The SE table at 0.2 read +0.0026, but that is a reading of its non-FM proxy (Section 4.5). Together about +0.02, so roughly +0.056 over B10 in all, and the estimated part rests on an assumed board cosine.

Table 8: Per-metric means on the validation-anchor scale, three folds and eight seeds.

Arm	Score	pds	mse	nmae	fid	reach	jac
B10	.3203	.7964	.2534	.3389	.0197	.5066	.0068
z16	.3606	.8029	.3838	.3984	.0671	.5149	−.0035
m	.3671	.8029	.3842	.4224	.0790	.5173	−.0033
a	.3693	.8028	.3844	.4234	.0860	.5215	−.0019

Table 9: Truth-substitution ceilings for the call list (native merged, against list nk1).

Factor taken from the truth	Gain	Best without the truth
direction	+0.0727	no predictor above our direction
magnitude	+0.0630	confidence lift, about +0.0007
list length	+0.0266	no length predictor above nk1
ordering	+0.0114	sign-agreement ordering, +0.0001
N_{conf} budget	−0.0032	—

Table 10: Gains split by whether a step added or changed a foundation-model (FM) part. The scales differ between stages, so the last row adds them only roughly.

Stage	Scale	Gain	FM	Non-FM
gwps-tgt to s8n7pk1g	board	+0.123	≈ +0.010 (8%)	≈ +0.113
22 October package to dgx	native merged	+0.097	≈ +0.002 (2%)	≈ +0.094
dgx to final package	clean-line v, fill estimate	≈ +0.021	≈ +0.017 (81%)	≈ +0.004
All three	mixed	≈ +0.240	≈ +0.029 (12%)	≈ +0.211

Where the room is. Replacing one factor of the call list with the truth gives the ceilings in Table 9. The room is in direction and magnitude, not in list length: among called, truly significant genes our sign was right for 0.885, 0.792 and 0.767 of them on f1, f2 and f3t, and no predictor we tried raised these rates.

Foundation-model share. Counted by what each step changed, foundation models gave about a tenth of our improvement (Table 10). Most of that tenth sits at the two ends of the project, in the first residual heads and in the three external tables of the final package, and the latter are estimates. We count as foundation-model parts the 16 embedding models and the TabPFN heads fitted on them, the TabPFN-3.5 joint heads, the ESM-2 and GenePT features, State, and the UPM, PIE and SE tables. A step's gain goes to the foundation models when the step added or changed one of these parts, and to the rest otherwise. The stages are on different scales, so shares are computed within each stage and the total is rough.

In the first stage the board cannot split a jump, so the FM part comes from local readings of the same changes: the 16-model ensemble head against the record mean (+0.0167, +0.0038 and +0.0035 on three single folds, mean +0.0080) and the two head settings added in as-s3a-pac (+0.0007 and +0.0011 on f3t). The rest of that stage is AtlasShift's generator (+0.1043 on the board) and non-FM settings. In the second stage the FM part is the sum of the adopted head-tuning steps. One more head change, dropping sibling-line targets from the heads' training rows (+0.0021 [+0.0006, +0.0037] on the folds), went only into the heads of D, E and F, and counting it raises the overall share to 13%. In the third stage the FM part is the UPM, PIE and SE tables (Section 4.5).

The share depends on the accounting. Without the estimated third stage it is about 5%. Counting the call-list step (+0.0091) as foundation-model work, because its floor and sign-zero rule came from our PIE experiments although the list holds no PIE data, gives 16%. Starting from our knock-down-only upload (−0.3019) instead of gwps-tgt adds +0.37 of non-FM baseline building and gives 5%.

Removal gives a smaller number than addition. On the base of 27 October, removing the whole residual changed the native score by −0.0003 [−0.0012, +0.0007], while removing single non-FM parts from the same base cost −0.0055 to −0.0280. This suggests that later non-FM steps took over most of the early head gain, which we did not test. The head weights changed after that reading, and removing all foundation-model parts from the final package was not measured. Within the heads, the single-cell embeddings were not separable from the rest: on AtlasShift a head of the same form without them, still with a TabPFN or TabICL regressor and ESM-2 features, added as much (single folds, differences from −0.0057 to +0.0034).

7 Negative results

These were measured and dropped unless stated. Each figure is a change against the arm current at the time, on the native merged scale unless the item names another.

- State's perturbation model zero-shot: -0.39 to -0.40 against the record mean on the first fold. Two residual State variants, -0.2123 and -0.2239 (first fold, one training run each). LoRA fine-tuning, -0.0179 against the unadapted model on held-out H1. State as a further source (ST-11), $v = +0.0002$ on the H1 clean line. Stack scored -0.012 zero-shot and -0.049 fine-tuned on the first fold (absolute scores, against 0.286 for copying the HEK293T effects).
- Open-X-Cell could not be told apart from its permuted control. The FM heads were kept, but their tuning was small: the adopted steps summed to about $+0.002$, none was separable from zero on its own, and this is the value of tuning, not of the heads' presence (Section 6). A learned entry MLP gave $+0.0021$, and $+0.0007$ without sibling-line labels.
- PIE's table mixed into our effects: -0.0131, -0.0262 and -0.0466 on three lines PIE had not seen. The same line of work produced the call floor, which on the mean path alone gained $+0.0132$ [$+0.0087$, $+0.0178$] on those lines merged, so our first conclusion that the PIE route was closed was too broad. It had tested one use. With both arms' directions made as inaccurate as PIE's were on board targets, the floor's gain was $+0.0000$ [-0.0031, $+0.0031$].
- Public recipes run as published: T3V200 -0.0939, Sidechain's code -0.0493 against our arm. Several board recipes as list variants did not exceed our list.
- Additional knock-down resources as list axes (KnockTF, CREEDS, PerturbAtlas, RenormCell, disease-matched tables, a MOLM-13 screen): screening estimates from $+0.0036$ to $+0.0114$ against z16, none above the list's own $+0.0117$, and within 0.0015 of their permuted controls where we ran one. CRISPRa screens are not mirrors of knock-down.
- Gene-partner columns (STRING, complexes, DepMap): -0.0012 to -0.0018. Amplitude predictors: Spearman with true strength at most 0.18.

8 Comparison with the top three entries

The three winning entries have our structure and differ from our entry in how much reference data they use and in how much of the transfer is learned. Our only source for their methods is the organisers' wrap-up [3], and the descriptions below are our summary of it. Everything this section says we would change is hindsight. None of it was run.

Shared structure. By the wrap-up, each winning entry is a modular pipeline with three steps. It estimates a knock-down's effect from measured responses to the same knock-down in other cell types, adapts the estimate to the new line using only that line's control cells, and generates 400 cells per target. Our entry has the same three steps (Figure 2). The design was open to everyone because, by the wrap-up's count, all but one of the 300 test targets had been knocked down before in some other line. All three winners also predict, for each target, which genes the test will call. In our entry the call list does that job with a fixed rule, a detectability threshold from the controls and a direction from a pooled posterior, and no part of it is fitted on held-out contexts. Table 11 sets the three entries beside ours.

The organisers' reading. The wrap-up concludes that reference data and the care of its transfer counted at least as much as architecture, that carrying measured responses over without any learning already recovers a large part of the signal, and that a podium place did not require private data. It also notes that the only foundation model on the podium, X-Cell, selected and scaled measured effects [3]. Our accounting points the same way. Foundation models gave about a tenth of our improvement, as additions to measured effects, and State's perturbation model on its own scored far below a copy of the measurements (Sections 6 and 7).

Overlap of data. Part of the winners' data was also in our package. The wrap-up lists three published screens for the first-place entry. Two of them match, by cell type and first author, the KOLF2.1J and CD4 T-cell screens among our sources. The third, in hPSC lines, we tested as an extra table and did not keep. The second-place entry used 222 datasets, and the wrap-up names 13 of its sources. Five are in our final package in some role. These are the Replogle screens (the K562 genome-wide screen as a source and the essential-gene screens as training rows of the heads), the Nadig screens (training rows), X-Atlas/Orion, the 2025 challenge data and ARCHS4. Five more we tested as extra tables or list axes and did not keep (HipSci, VIPerturb-seq, LINCS, PerturbAtlas and KnockTF). One screen from IGVF served as a clean line. Adamson

Table 11: The three winning entries as described in the organisers' wrap-up [3], and how each differs from ours. Scores are those of the final leaderboard.

Entry	Reference data	Transfer to the new line	Difference from ours
1. BioInsight AI Lab (Illumina), CellularAI, 0.4399	Illumina's Billion Cell Atlas and three published genome-scale CRISPRi Perturb-seq datasets (iPSC, hPSC lines, primary CD4 T cells).	One covariate-adjusted effect per screen. A random-effects meta-analysis with weights per gene and perturbation, from control-cell similarity to the new line, noise and knock-down efficiency. An SVD and ridge model of how a dataset's control expression shifts its effects, applied to the new line's controls. An estimate of which genes respond detectably. The final model was selected on the challenge's validation data.	An atlas as added reference. Weights fitted per gene, perturbation and new line, where ours are fixed per screen. A learned shift from control expression.
2. FlowState, 0.3348	222 perturbation-response datasets from 128 studies in 86 cellular contexts. About half are single-cell and the rest bulk, across CRISPRi, CRISPR knockout, siRNA and shRNA.	No published or pretrained model. The inputs are the perturbed gene and the new line's control cells. Three heads, each trained with whole cell contexts held out, predict whether a gene responds, in which direction and by how much. Datasets are weighted by reliability and noisy estimates shrunk.	The response is learned across many contexts. Our heads saw seven and only amplify the table's own changes.
3. X-team (Xaira), X-Cell Transfer, 0.3263	X-Cell, a masked diffusion model over sets of cells, trained on more than 37 million Perturb-seq cells from in-house and external screens, and reference CRISPRi screens.	The model's raw output was not submitted. Measured effects are pooled with weights for cell count and control similarity and a cap per study. Model confidence and directional agreement decide, gene by gene, how much of each measured effect carries over. A gradient-boosted classifier predicts which genes will be significantly DE and guides effect sizes against the noise of a 400-cell pseudobulk.	A perturbation model trained at scale gates and scales measured effects. Our external predictors add direction with fixed weights.

and ENCODE we never tested. The entry's full list is not public, so we cannot say which of the tables we dropped it used.

What we would change. Four changes follow from the comparison. None was run, and we do not know their size.

1. *More practice lines, from more than one lab.* Two of our three folds come from one lab and share 299 of 300 targets, and the third is the stem-cell line of the 2025 challenge. A fourth fold on a line of the Replogle screens was planned and never built. Every learned step was therefore selected on what amounts to two cell contexts.
2. *A response model trained across many public screens, with whole contexts held out.* This is the second-place design, with 86 contexts. Our learned steps saw far fewer. The residual heads were fitted on rows from seven contexts and only amplify the table's own changes. A gradient-boosted classifier that picked which genes to lift past the floor did worse on the stem-cell fold than a random pick from the same candidates (-0.0049 [-0.0070, -0.0029]). A gene-level boosted model, trained on two folds and tested on the third, lost to the recipe's effects on the stem-cell fold (-0.061 [-0.076, -0.044] in pooled cosine after centring per target). On the two sibling folds it read higher ($+0.136$ and $+0.079$), but not on the entries the truth calls significant, where its correlation with the true log fold change was 0.36 and 0.31 against the recipe's 0.49 and 0.39. Cross-validated within one fold it was positive on all three ($+0.145$, $+0.095$, $+0.020$). We read this as learnable within a context and not across the contexts we had, and stopped. Neither model had other public screens as training contexts, so the reading does not bear on a model trained across dozens of them.
3. *Source weights from control similarity to the new line.* The first- and third-place entries weight each reference by the similarity of its control cells to the new line's. We tested the idea on 20 October, with four sources. Tilting their weights by foundation-model or expression similarity read $+0.0020$ to $+0.0053$ merged over the folds in four arms. The gain sat on the two sibling folds, and the stem-cell fold lost in all four arms (-0.0038 to -0.0114). We kept fixed weights. What such weights are worth over dozens of sources was not measured.
4. *More public data, used as training contexts.* The final package draws effects from seven single-cell screens and two bulk resources. We judged about 30 further tables, most of them one at a time, and kept none. ENCODE's knock-down experiments were counted and never tested.

How we judged data. The four gaps have one root, the way we tested data. After AtlasShift's generator had lifted the board score, each new public table was tested as one more source of a fixed weighted average, or as one more axis of the call list, against the current package on the three folds. One table moves such an average little. Its reading was a few thousandths, often inside the spread of its permuted control, and the table was dropped. This test measures what a table adds to a fixed average. It does not measure what a model trained across all the tables could learn, where every table is one more training context and their number is what counts. We also picked tables by how many test targets they covered. Four essential-gene screens were not made sources because they cover none of the test targets, although their rows trained the heads. The U-87MG kinome screen, with 21 test targets at a median of 83 cells each, was sized by that coverage at +0.001 at most and not built into context E. A training context does not need to cover the test targets, because what a model learns from it is how an effect moves from one line to another. The question why the remaining public data were not used together was raised twice inside the team, on 3 and 4 November. Our answer put 22 knock-down tables into one pooled posterior and fitted a low-rank map per table, both still inside the fixed pipeline and screened by a proxy on the folds, and no arm was run. A model with those tables as held-out training contexts was never built, and we do not claim to have tried the second-place design.

Where we would look first. The mse term carries 37% of our distance to the first place. It reads the size and the direction of every predicted pseudobulk change. Our sizes came from rules, one gain per context and a fill computed from the metric's budget, and within a target the relative sizes of genes were those of the fused table after shrinkage. The first-place entry estimates effect sizes with weights per gene and perturbation and corrects them from the new line's controls, and the third-place entry calibrates them against the noise of a 400-cell pseudobulk. A size learned per gene and target is where we would look first. This is a guess from the comparison, not a measurement.

Lessons. We take three lessons. First, data should be tested the way it will be used. Adding one table at a time tests a fixed table. Data meant for a model go in as a batch with whole contexts held out, so that the reading is of the model. Second, a model earns credit only against a control of the same form. That comparison is how the single-cell embeddings in our heads turned out not to be separable from a head without them (Section 6). Third, asking why a model predicts what it does can yield a method that outlives the model. PIE's detection probabilities led to the call floor, which stayed in the package after PIE's direction had been replaced.

9 Limitations

- Much of the prediction is transferred measurement. By our accounting foundation models gave about a tenth of the improvement (Table 10), and their largest part in the final package is estimated, not measured.
- The comparison with the winning entries rests on the organisers' summary of their methods, and the changes it suggests were not run.
- Two of our three folds come from one lab and share 299 of 300 targets, so learned steps and data tables were judged on what amounts to two cell contexts.
- The two-path rendering and the call floor are tuned to the metrics and have no biological meaning. The rules allow the organisers to change the scoring if they see manipulation.
- The final candidate was chosen after hundreds of comparisons on reused folds. Intervals are not corrected for selection, and the true gain is probably below the measured one.
- The estimated steps depend on an assumed board cosine and on clean lines that differ from D, E and F, and one of five lines was negative for UPM.
- Our forecast of the final score was too high (about 0.32 against 0.2594). It rested on one pair of local and board scores, made on the validation lines, and on folds reused for selection. We cannot separate these two causes from each other or from the change of cell lines, so the fold gains reported here are not estimates of gains on the final leaderboard.
- The largest remaining error is in mse: the predicted common part of all targets scores worse than predicting none, and the truth's common part is nearly uncorrelated across cell lines.

10 Licences and use

The final package contains numbers generated by UPM (CC BY-NC 4.0), PIE (code CC BY-NC-SA, weights non-commercial), TabPFN-3.5 (the four joint heads, weights under a non-commercial licence) and State SE-

600M (non-commercial licence), and is therefore for non-commercial use. Open-X-Cell is not in it. T3V200 has no licence, and we used ideas only and reused no code. Our code release is MIT, and its AtlasShift-derived file carries AtlasShift's MIT notice.

Code and data availability

The supplementary package `vccmini` re-implements our steps in about 600 lines of numpy with a synthetic end-to-end demo and tests (`code/`). It does not include challenge data, external tables, model weights or our full generator. The AtlasShift steps it needs are in `vccmini/atlasshift.py`, marked as not ours. Appendix B maps each paper step to its function. The code is available at https://github.com/mengli11235/vcc2026.

Acknowledgements

We thank the authors of AtlasShift, Sidechain, cell-eval2, the foundation models, UPM, PIE, and the public datasets listed in Table 2, and the teams who shared notes on the board. Much of the code, the experiment logs and this draft were produced with Claude (Anthropic), an AI assistant, working under the direction of the authors, who are responsible for the content.

References

[1] Abhinav Adduri, Dave Burke, Joseph Caputo, et al. Virtual Cell Challenge 2026: Benchmarking zero-shot generalization across cellular contexts. *Cell*, 189(19):5827–5830, 2026. doi: 10.1016/j.cell.2026.08.004.

[2] Arc Institute. State embedding model SE-600M. https://huggingface.co/arcinstitute/SE-600M, 2025.

[3] Arc Institute. Virtual Cell Challenge 2026: Winners and wrap-up. https://arcinstitute.org/news/virtual-cell-challenge-2026-wrap-up, 2026. Published 19 November 2026, read 23 November 2026.

[4] Arc Institute. cell-eval2, version 0.18.0. Python package, the challenge's evaluator with preset `vcc2026`, 2026. Capped expression MSE: `metrics/delta.py`, issues #247 and #348.

[5] Arc Institute. PIE. https://github.com/ArcInstitute/pie; weights https://huggingface.co/arcinstitute/PIE_replogle_xdataset, 2026. Code CC BY-NC-SA; weights non-commercial.

[6] Arc Institute. Virtual Cell Challenge 2026: final leaderboard. https://virtualcellchallenge.org/, 2026. 807 entries, read 23 November 2026.

[7] Yiqun Chen and James Zou. GenePT: a simple but effective foundation model for genes and cells built from ChatGPT. Embeddings v2, Zenodo 10833191 (CC BY 4.0), 2024.

[8] GSE295214. H1 hESC CRISPRi Perturb-seq, GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE295214, 2025.

[9] GSE301119. Primary human macrophage CRISPRa/i screen, GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE301119, 2025.

[10] GSE337804. A549 CRISPR Perturb-seq, GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE337804, 2026.

[11] GSE345058. A549 gene knock-out Perturb-seq, GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE345058, 2026.

[12] Shipeng Guo, Yanshuang Jiang, Jieya Zou, et al. GPSAdb 2.0: an expanded atlas of gene-perturbation transcriptomes with enhanced tools for regulatory gene discovery. *Nucleic Acids Research*, 54(D1): D1082–D1086, 2026. doi: 10.1093/nar/gkaf1077. Data: Zenodo 23585243, CC BY 4.0.

[13] Noah Hollmann, Samuel Müller, Lennart Purucker, et al. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025.

[14] Ann C. Huang, Tsung-Han S. Hsieh, Jiang Zhu, et al. X-Atlas/Orion: Genome-wide Perturb-seq datasets via a scalable fix-cryopreserve platform for training dose-dependent biological foundation models. bioRxiv, doi:10.1101/2025.06.11.659105, 2025. Data: figshare+, https://doi.org/10.25452/figshare.plus.29190726, CC BY-NC-SA 4.0.

[15] IGVFDS7943IIWZ. TeloHAEC endothelial CRISPRi Perturb-seq, IGVF analysis set (Schnitzler et al., Nature 2024). https://data.igvf.org/analysis-sets/IGVFDS7943IIWZ, 2024.

[16] Ji-Yuanlan. T3V200_final. https://github.com/Ji-Yuanlan/T3V200_final, 2026. No licence file; read for ideas only, no code reused.

[17] kaipengm2. AtlasShift: A simple atlas transfer method. https://github.com/kaipengm2/Virtual-Cell-Challenge-2026, 2026. Commit d24ce4fdae0cd7cba3ba29546cd8737094eae98a, MIT licence.

[18] Alexander Lachmann, Denis Torre, Alexandra B. Keenan, et al. Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications*, 9:1366, 2018.

[19] Zeming Lin, Halil Akin, Roshan Rao, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023.

[20] Ajay Nadig, Joseph M. Replogle, Angela N. Pogson, et al. Transcriptome-wide analysis of differential expression in perturbation atlases. *Nature Genetics*, 57(5):1228–1237, 2025.

[21] Sami Nourreddine, Yesh Doctor, Amir Dailamy, et al. A genome-scale CRISPRi perturbation atlas of human induced pluripotent stem cells. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03199-w. Preprint doi:10.1101/2024.11.03.621734. Data: figshare+, https://doi.org/10.25452/figshare.plus.27261219.v1, CC BY 4.0.

[22] Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575, 2022.

[23] saberhq. Sidechain. https://github.com/saberhq/sidechain, 2026. MIT licence; `pooled_delta`.

[24] Matthew Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

[25] Jake Yeung, Jenille Tan, Liang Wang, et al. Joint analysis of multiply perturbed cells improves statistical power and cost efficiency in Perturb-seq screens. GEO accession GSE337988, CRISPRi Perturb-seq in DLD-1 cells, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE337988, 2026.

[26] youlab2025. UPM. https://github.com/youlab2025/UPM; weights https://huggingface.co/youlab2025/UPM, 2026. Preprint doi:10.64898/2026.10.25.761625; CC BY-NC 4.0.

[27] Ronghui Zhu, Emma Dann, Jun Yan, et al. Genome-scale perturb-seq in primary human CD4+ T cells maps context-specific regulators of T cell programs and human immune traits. bioRxiv, doi:10.64898/2025.12.23.696273, 2025. Differential expression statistics, `GWCD4i.DE_stats.h5ad`.

A Foundation models

Table 12: The 16 foundation-model routes (frozen embeddings of each context's controls).

Route	Model	Source
scf	scFoundation	https://huggingface.co/genbio-ai/scFoundation
gbc	GB.Cell-100M	https://huggingface.co/genbio-ai/GB.Cell-100M
tx	Tahoe-x1 (70m)	https://huggingface.co/tahoebio/Tahoe-x1
sc	scGPT whole-human	official checkpoint (Google Drive, via README)
uce	UCE 4-layer	figshare 24320806
se	State SE-600M	https://huggingface.co/arcinstitute/SE-600M
sp	scPRINT-2	https://huggingface.co/jkobject/scPRINT
cf	CellFM 800M	https://huggingface.co/ShangguanNingyuan/CellFM
gf	Geneformer V2-316M	Hugging Face (ctheodoris/Geneformer)
st	Stack-Large-Aligned	https://huggingface.co/arcinstitute/Stack-Large-Aligned
gj	GeneJEPA	https://huggingface.co/elonlit/GeneJEPA
pg	PerturbGen	https://huggingface.co/lotfollahi-lab/PerturbGen
gc	GeneCompass	https://github.com/xCompass-AI/GeneCompass
c2s	C2S-Scale Gemma-2 2B	https://huggingface.co/vandijklab/C2S-Scale-Gemma-2-2B
nf	Nicheformer	Hugging Face release
tf	TranscriptFormer	CZI release (transcriptformer 0.6.1)

B Final settings and code map

Table 13: Final setting of each step, the generator option that sets it, and the `vccmini` function that re-implements it.

Step (final setting)	Generator option	`vccmini` function
Source effect, prior 10^5 [OTHERS]	—	`atlasshift.source_effect`
Common subtraction 1.0, per source	`-common-subtract`, source spec	`atlasshift.subtract_common`
EB-ash, scale 0.85, renorm	`-eb-shrink ash`	`bulk.eb_shrink`
Unit fusion (bulk)	`-fuse-unit`	`bulk.fuse_unit`
Desired profiles, amp 1.2/0.5 [OTHERS]	`-amp-mean`, `-amp-bulk`	`atlasshift.desired_profiles`
FM heads .5/.4/−.9, renorm, amp	`-add-log-ratio`	`fmhead.add_log_ratio_bulk`
Deflate 15 axes	`-deflate 15`	`bulk.control_axes`, `bulk.deflate`
Bulk gain 1.458/1.546/1.527	`-bulk-gain`	`bulk.bulk_gain`
Bulk mix $\lambda = 2$	`-bulk-mix`	`bulk.bulk_mix`
Additive tables, $k = 1$	`-bulk-add`	`bulk.bulk_add`, `external.*`
Call list nk1	(list builder)	`calllist.need_of`, `call_list`
Call floor, sign zero, mag 1, cap 1, $m = 0.02$	`-call-floor*`	`calllist.call_floor_mean`
Fill, target ratio 1.04	`-bulk-final-scale`	`bulk.fill_correction`, `bulk.bulk_final_scale`
Budget-to-claim estimate	(separate script)	`evaluate.ratio_est`
Count generator, fit retry 1000 [OTHERS]	`-fit-retry`	not included