VCDESIGN: FINITE-BUDGET INTERVENTION DESIGN FOR VIRTUAL CELLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtual-cell models are typically judged by how well they predict the effects of perturbations or recover the interventions associated with a cellular transition. Yet experimental planning poses a different question: given a desired cellular state, which interventions should be tested when only a limited number of experiments can be performed? Here we introduce VCDesign, which, to our knowledge, is the first framework to formulate target-conditioned cellular intervention design as finite-budget ranking over a variable candidate set. Given source and target states, a candidate library and an experimental budget, VCDesign prioritizes the interventions to test. We also introduce a decision-aligned evaluation framework that measures the quality of the selected candidates using independently measured held-out outcomes. For candidates lacking historical perturbation responses, we develop VCDesign-CED, which uses existing perturbation measurements and biological knowledge to infer candidate effects and combine them with direct candidate scoring. Across four CRISPRi Perturb-seq contexts, VCDesign-CED improves the prioritization of response-unseen candidates, increasing pooled budgeted utility at 20 selections by 0.049 and recovering 1.11 additional top-5% candidates on average. When candidate responses are already available in other contexts, direct retrieval performs better, and routing candidates according to response availability further improves mixed-library selection. Controlled analyses further show that transferable perturbation evidence and candidate knowledge contribute more to design performance than simply adding training contexts or increasing effect-model complexity, while prediction accuracy does not reliably rank nearby models by their utility for intervention design. Together, these results establish a task and evaluation framework for using virtual-cell models to make experimental decisions under finite budgets.

1 INTRODUCTION

Virtual-cell models aim to predict how cells respond to perturbations and, ultimately, to help identify interventions that can move cells towards desired states. Most work has focused on two neighbouring problems. Forward models predict the cellular consequences of a specified genetic, chemical or other perturbation (Lotfollahi et al., 2019; 2023). Inverse models start from a desired transition and seek interventions that might produce it (Wang et al., 2025). Experimental planning requires a further step, because researchers typically face a starting state, a desired target state, a set of interventions that can actually be tested and a budget that allows only a small fraction of them to be evaluated. The practical question is therefore not only what a perturbation will do, or which intervention best matches an observed transition, **but which interventions should be tested first when experimental resources are limited.** Framing cellular intervention design in these terms turns it into a decision problem in its own right.

Several lines of work already move towards this goal, but they approach it through different computational routes. Expression-profile resources such as Connectivity Map (Lamb et al., 2006) and LINCS (Subramanian et al., 2017) retrieve perturbations whose measured responses match a desired transcriptional change. Inverse models such as CellNavi predict candidate drivers directly from source and target states (Wang et al., 2025). Forward models can simulate candidate interventions and rank their predicted outcomes (Zeng et al., 2025). More recent approaches search for interventions through generative methods (Han et al., 2025) or optimization procedures (Cao et al.,

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

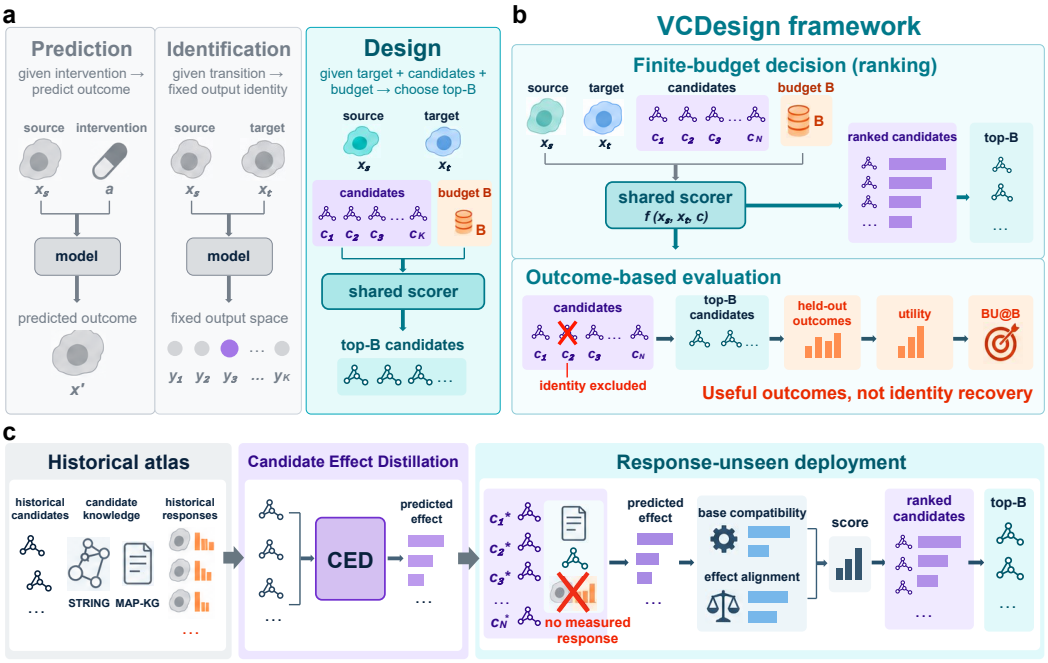

Figure 1: **VCDesign separates the intervention-design decision from its solvers and evaluates the resulting finite-budget selections. a,** Prediction maps an intervention to an outcome, identification maps a transition to a fixed output space, and design ranks variable candidates toward a target under a budget. **b,** A shared scorer $f(x_s, x_t, c)$ ranks candidates; held-out replicate outcomes evaluate the top-B prefix, excluding the query identity. **c,** CED predicts effects from historical responses and STRING/MAP-KG features, then combines base compatibility with target alignment for response-unseen candidates.

2025). Although these methods differ in how they represent interventions, what information they require and what objective they optimize, they ultimately serve the same experimental decision: which candidates in the available library should occupy the limited number of positions that can actually be tested? Existing evaluation does not always align with this decision, because prediction error measures the quality of a response estimate and identity recovery asks whether a method retrieves the intervention associated with a query transition, whereas an experiment executes only a small prefix of a ranking. A useful design system should also reward multiple distinct interventions when they produce similarly valuable outcomes, making a common task and evaluation interface necessary for comparing different approaches on the basis of the decisions they enable.

Here we introduce VCDesign (Figure 1), which formalizes this decision by asking a model to rank a feasible intervention library given source and target states and an experimental budget B, and to prioritize the B candidates to test. To our knowledge, this is **the first explicit formulation of target-conditioned cellular intervention design as finite-budget ranking over a variable candidate set.** We pair this task with a decision-aligned evaluation framework that focuses on the portion of the ranking that can actually be executed and assigns value to selected candidates using independently measured held-out outcomes, thereby allowing multiple interventions to be rewarded when they produce useful effects and placing retrieval, forward simulation, optimization and direct scoring under a common downstream objective. VCDesign also makes information availability part of the problem definition: generalizing to a new cellular context is different from ranking a candidate whose own perturbation response has never been available for effect learning. For this response-unseen setting, we develop VCDesign-CED, which uses historical perturbation measurements to learn how deployment-available biological knowledge relates to perturbational effects and then uses the inferred effects to help rank candidates for a requested cellular transition. In this way, experimental evidence from previously measured interventions can inform the prioritization of candidates that have not themselves been profiled.

We evaluate this framework across four CRISPRi Perturb-seq contexts and find that historical perturbation evidence consistently improves finite-budget selection for response-unseen candidates. Across three held-out cellular contexts, VCDesign-CED increases pooled BU@20 by 0.049 over the same base ranker and identifies, on average, 1.11 additional candidates from the top 5% of measured utility within the first 20 selections, with the same pattern also observed for held-out intervention identities in K562. The experiments further reveal a clear boundary set by information availability: when a candidate has already been measured in other cellular contexts, direct retrieval of those responses outperforms effect inference, whereas routing measured and unmeasured candidates to different solvers further improves selection in mixed libraries. Controlled analyses also show that response-unseen design depends more strongly on transferable perturbation evidence and informative candidate knowledge than on simply adding training contexts or increasing effect-model complexity; increasing the amount of measured evidence improves design utility, whereas additional contexts at matched measurement volume and the richer effect models we tested provide little or no comparable benefit. Prediction accuracy captures some large differences across data and model configurations, but it does not reliably order nearby models by their utility for intervention design. Together, these results extend virtual-cell modelling from describing cellular states towards supporting experimental decisions, and provide a common task and evaluation foundation for prioritizing high-value interventions under finite experimental budgets.

Our contributions are threefold:

- **A finite-budget formulation of cellular intervention design**, ranking a variable feasible candidate set towards a target under an experimental budget.
- **A decision-aligned evaluation framework** that scores the executable ranking prefix using independently measured held-out outcomes, enabling different computational routes to be compared through the decisions they support.
- **A response-unseen design strategy**, VCDesign-CED, that transfers historical perturbation evidence through biological knowledge and enables controlled analysis of when such transfer is useful.

2 Related Work

Forward perturbation prediction. Virtual-cell models have primarily focused on predicting molecular responses to specified interventions. scGen and CPA model perturbation-induced state changes in learned latent spaces (Lotfollahi et al., 2019; 2023). CellOT learns transport maps between cellular populations, while GEARS incorporates gene relationships to predict unseen perturbation combinations (Bunne et al., 2023; Roohani et al., 2024). More recent methods extend perturbation prediction through pretrained cellular representations, as in scGPT and CellFM (Cui et al., 2024; Zeng et al., 2025), or transfer to unseen perturbations and contexts, as in PerturbNet and State (Yu et al., 2025; Adduri et al., 2026). In their perturbation-prediction tasks, these methods share a forward formulation: given an intervention and cellular context, predict the resulting molecular state. Recent benchmarks show that common prediction metrics can be dominated by shared variation and may not isolate perturbation-specific information (Ahlmann-Eltze et al., 2025; Viñas Torné et al., 2026).

Forward models can support design by simulating candidate outcomes, but prediction accuracy and the utility of the resulting experimental choices are distinct objectives; VCDesign evaluates the latter directly.

Inverse perturbation and design. A complementary line of work starts from a desired cellular state or transition and searches for interventions that could induce it. Connectivity Map and LINCS support perturbation retrieval by matching or reversing gene-expression signatures against experimentally measured profiles (Lamb et al., 2006; Subramanian et al., 2017). CellNavi predicts genes that drive transitions between cell states, while PDGrapher predicts therapeutic perturbations toward a desired healthy state (Wang et al., 2025; Gonzalez et al., 2026). PAIRING matches perturbation effects to a desired transition in latent space (Han et al., 2025). RevPert ranks candidate drivers from a fixed genetic catalog for an observed transcriptomic contrast (Liang et al., 2026). CASCADE and U-Pert approach intervention design through optimization in learned regulatory models or perturba-

tion dynamics (Cao et al., 2025; Peng et al., 2026). These methods provide complementary routes from desired outcomes to candidate interventions.

These approaches differ in output space, access to candidate responses and evaluation objective. VCDesign instead evaluates them through a common finite-budget decision, while its response-unseen setting further distinguishes context transfer from transfer without candidate-specific response supervision.

3 THE VCDESIGN FRAMEWORK

3.1 DESIGN UNDER A FINITE BUDGET

A query $q = (x_s, x_t)$ specifies source and target states, with transition $\mathbf{d}_q = x_t - x_s$. Given a feasible candidate set $\mathcal{C}_q$ and budget B, a solver ranks candidates and selects the first B for testing (Mehrjou et al., 2022). Our reference scorer $s(q, c) = f(x_s, x_t, c)$ uses candidate knowledge $\mathbf{k}_c$ available before profiling, allowing the library to vary without a fixed output vocabulary.

3.2 EVALUATION BY MEASURED OUTCOMES

We evaluate the first B candidates that the budget allows researchers to test, rewarding useful interventions earlier in this prefix. Candidate utility is defined by independently measured held-out outcomes rather than query-identity recovery, so multiple interventions can be useful for one target. This decision-level criterion applies across solvers, while the definition of a useful outcome remains application-specific.

In our CRISPRi benchmark, useful interventions reproduce the direction of a desired transcriptomic change. Each screen is split into batch-disjoint views $v \in \{1, 2\}$; a query formed in one view is evaluated in the other. The response $\mathbf{r}_g^{(v)}$ is the log-normalized pseudobulk of knockdown g minus its view-specific control. For query identity g_q, utility is

$$U(q, c) = \cos\left(\mathbf{r}_{g_q}^{(\bar{v}_q)}, \mathbf{r}_c^{(\bar{v}_q)}\right), \tag{1}$$

where $\bar{v}_q$ is the held-out evaluation view. We exclude g_q from the pool, rewarding other interventions that produce useful outcomes; alternative utilities are compared in Section E, and other applications can use their own measured outcomes.

Budgeted utility measures the value recovered within B experiments. Let $\mathrm{RU@}B$ be their discounted mean utility, with rank weights $1/\log_2(j+1)$ (Järvelin & Kekäläinen, 2002). We normalize against random selection and the best ranking under measured utility:

$$\mathrm{BU@}B(q) = \frac{\mathrm{RU@}B(q) - \mathrm{RU}_{\mathrm{rand}}(q)}{\mathrm{RU}^{\star}\mathrm{@}B(q) - \mathrm{RU}_{\mathrm{rand}}(q)}. \tag{2}$$

Zero denotes random selection in expectation; one denotes the best ordering of that pool under U. We report BU@20 and the budget average $\overline{\mathrm{BU}} = \frac{1}{3}(\mathrm{BU@10} + \mathrm{BU@20} + \mathrm{BU@50})$. Mean selected utility $\mathrm{MU@}B$ and high-value hit counts $\mathrm{HvHit@}B$ complement this score; high-value candidates are in their pool's top 5% by utility. Queries must have reproducible responses across views (definitions and eligibility test in Sections A and A.1).

3.3 CONTEXT AND RESPONSE GENERALIZATION

A new cellular context and an unmeasured candidate pose distinct transfer problems. *Context transfer* withholds the evaluation context from training; the candidate may still have measured responses in other contexts. *Response-unseen transfer* removes an evaluated candidate's responses from the CED atlas in every training context. The base scorer remains fixed; its K562 training identities include 68% of the masked candidates, so masking tests the effect branch, not the whole system. Held-out identities, held-out contexts and response masking thus separate what is new to the scorer from what is new to the effect predictor.

4 VCDESIGN-CED

VCDesign-CED instantiates the framework where a candidate's own response is unavailable for effect learning. It combines direct candidate scoring with Candidate Effect Distillation (CED), which learns how deployment-available biological knowledge maps to perturbational effects and aligns the inferred effects with the requested transition (Figure 1b,c).

4.1 LEARNING EFFECTS FROM CANDIDATE KNOWLEDGE

Candidate features combine STRING physical-interaction information and MAP-KG features derived from protein representations (Szklarczyk et al., 2023; Feng et al., 2026; Lin et al., 2023).

CED represents training responses in a low-dimensional basis. Let the rows of $\mathbf{R}$ be training consensus responses and let $\mathbf{W}$ contain its 256 leading uncentered right singular vectors (Halko et al., 2011). Ridge regression maps standardized knowledge $\tilde{\mathbf{k}}_c$ to coefficients in this response basis:

$$(\hat{\mathbf{A}}, \hat{\mathbf{b}}) = \arg\min_{\mathbf{A},\mathbf{b}} \sum_c \|\mathbf{W}^\top \mathbf{r}_c - \mathbf{b} - \mathbf{A}\tilde{\mathbf{k}}_c\|_2^2 + \lambda \|\mathbf{A}\|_F^2, \tag{3}$$

$$\hat{\mathbf{e}}_c = \mathbf{W}(\hat{\mathbf{A}}\tilde{\mathbf{k}}_c + \hat{\mathbf{b}}), \qquad s_{\text{eff}}(q, c) = \cos(\mathbf{d}_q, \hat{\mathbf{e}}_c). \tag{4}$$

At deployment, $\hat{\mathbf{e}}_c$ is computed from knowledge alone and scored by alignment with the requested transition. This attribute-to-output mapping connects CED to zero-shot learning and learning with privileged information (Xian et al., 2019; Vapnik & Vashist, 2009; Lopez-Paz et al., 2016). Ridge regression serves as the reference effect-transfer map; richer alternatives are compared under the same evaluation setting in Section 5.4.

Held-context results use unit-direction targets $\mathbf{r}_c/\|\mathbf{r}_c\|$ with the same ridge and fusion; K562 results use raw targets (Sections B and E.1).

4.2 DIRECT CANDIDATE SCORING AND FUSION

A dual encoder maps $(x_s, \mathbf{d}_q)$ and $\mathbf{k}_c$ to unit vectors $\mathbf{u}_q$ and $\mathbf{v}_c$, respectively (Radford et al., 2021; Karpukhin et al., 2020), giving $s_{\text{base}}(q, c) = \tau\langle \mathbf{u}_q, \mathbf{v}_c \rangle$. The K562 scorer learns from cross-view utility labels using a replicate-informed pairwise loss and a top-20 set loss (Section B). We standardize each branch across the current candidate pool and add them with equal weight:

$$s(q, c) = z_q(s_{\text{base}}(q, c)) + z_q(s_{\text{eff}}(q, c)). \tag{5}$$

In a new context, CED pools responses from the other contexts on their shared $6{,}550$-gene axis; the evaluation context contributes no training response, and the base scorer stays fixed.

5 EXPERIMENTS

We use VCDesign to compare computational routes under a common finite-budget objective, test whether historical perturbation evidence can support response-unseen candidates, and identify what limits this transfer.

5.1 EXPERIMENTAL SETUP

We use CRISPRi Perturb-seq screens from K562 and RPE1 (Replogle et al., 2022) and HepG2 and Jurkat (Nadig et al., 2025). K562 provides disjoint training, development and locked evaluation identities; RPE1, HepG2 and Jurkat are held out in turn, with candidate pools of 1,273, 764 and 1,450 (Section A). Five-fold cross-fitting removes each evaluated candidate's responses from all training contexts (Section D.9); the measured setting retains them. The evaluation context never supplies training responses (Section D.9). Profile retrieval matches historical responses to the requested transition by cosine; CellNavi and GEARS adaptations are detailed in Section D.6. No optimization solver is instantiated. K562 context-transfer controls are reported in Section D.12.

The locked K562 comparison uses prespecified raw-target CED and $\overline{\text{BU}}$, fixed before scoring. For held contexts, unit-direction targets were selected on the same pools through a preregistered comparison; this choice and BU@20 were fixed before external-baseline runs (Section E.1). Levels and

Table 1: **Evaluated solvers under a common finite-budget decision objective and different information requirements.** Median BU over query-view rows: $\overline{\text{BU}}$ on the locked K562 split, whose candidates neither branch has seen, and BU@20 in held-out contexts, with each candidate's own response masked from the effect atlas or measured elsewhere. The base scorer's K562 training identities include 68% of the masked candidates; on the other 32%, VCDesign-CED gains $+0.033$ $[+0.030, +0.037]$ over it (Section D.4). Measured-response oracle: 0.814 and 0.951. –: not applicable.

		Held-out	Held-out contexts	
Solver	Interface	identities	masked	measured
Can score without a candidate-specific measured response				
Random	none	-0.005	0.002	-0.005
Goal-blind prior	effect size, ignores target	0.181	0.170	0.194
CED effect only	forward simulation + match	0.125	0.294	0.396
Text-augmented ranker	direct scoring + text	0.385		
Base scorer	direct scoring	0.355	0.440	0.440
VCDesign-CED	direct scoring + effect	$\mathbf{0.391}$	$\mathbf{0.499}$	0.508
Requires a candidate-specific measured response				
CellNavi, official*	fixed-output identification		–	-0.020
CellNavi, native*	fixed-output identification		–	0.015
Profile retrieval	retrieval		–	$\mathbf{0.675}$

*Official or candidate-library classes; one fine-tune per context (Section D.6). GEARS fits were candidate-invariant. Blank: not run.

paired gains are medians, except that hit-count gains are means; 95% cluster-bootstrap intervals keep trained models fixed (Section C).

Table 1 compares the evaluated computational routes under a common decision objective while distinguishing their candidate-response requirements. VCDesign-CED leads the evaluated response-optional solvers on held-out identities and in masked held-out contexts, whereas profile retrieval performs best when candidate responses are measured elsewhere. We next examine the source of response-unseen gains and the use of measured responses in mixed libraries.

5.2 VCDesign-CED Improves Intervention Selection

In held-out contexts, VCDesign-CED is the strongest solver we ran that can score without a measured response of the candidate (Figure 2a). With own responses masked, it improves on the identical base scorer by $+0.053$, $+0.070$ and $+0.039$ BU@20 in RPE1, HepG2 and Jurkat, and by $+0.0492$ $[+0.0456, +0.0534]$ pooled, with 78.5% of rows improving (Figure 2b). Among its first 20 selections, it includes on average 1.11 more top-5% candidates than the base scorer (median $+1$).

On the locked K562 evaluation split, whose candidates neither branch has seen, adding CED raises median $\overline{\text{BU}}$ from 0.3551 to 0.3905 (Figure 2c). The median paired gain is $+0.0319$ $[+0.0282, +0.0378]$, with 75.1% of rows improving, and $+0.0268$ $[+0.0140, +0.0346]$ over the text-augmented ranker. The high-value-hit gain appears mainly at $B = 50$ (Figure 2d): among the first 50 candidates, VCDesign-CED selects on average 1.9 $[1.6, 2.2]$ more high-value candidates than the base scorer and 2.2 $[1.8, 2.6]$ more than the text-augmented ranker; at $B \leq 20$ it exceeds the base scorer by at most 0.4 and does not separate from the text-augmented ranker.

Controls locate the source of this gain. It persists on candidates excluded from base-scorer training, depends on goal-conditioned effect alignment, and is substantially weaker when the effect branch is used alone (Sections D.4 and E.1 and Table 1). A four-quarter realized-outcome analysis is directionally consistent in mean but does not meet its preregistered median criterion, so improved directional ranking is not yet established as improved realized outcome (Section D.6).

5.3 Response Availability Defines the Solver Regime

Profile retrieval is the better solver for candidates whose responses were measured in other contexts. It reaches pooled BU@20 0.675 against 0.508 for VCDesign-CED, a paired advantage of $+0.1519$ $[+0.1424, +0.1602]$ (Figure 3a; Table 14). Measured responses should therefore be retrieved, and CED reserved for candidates without them.

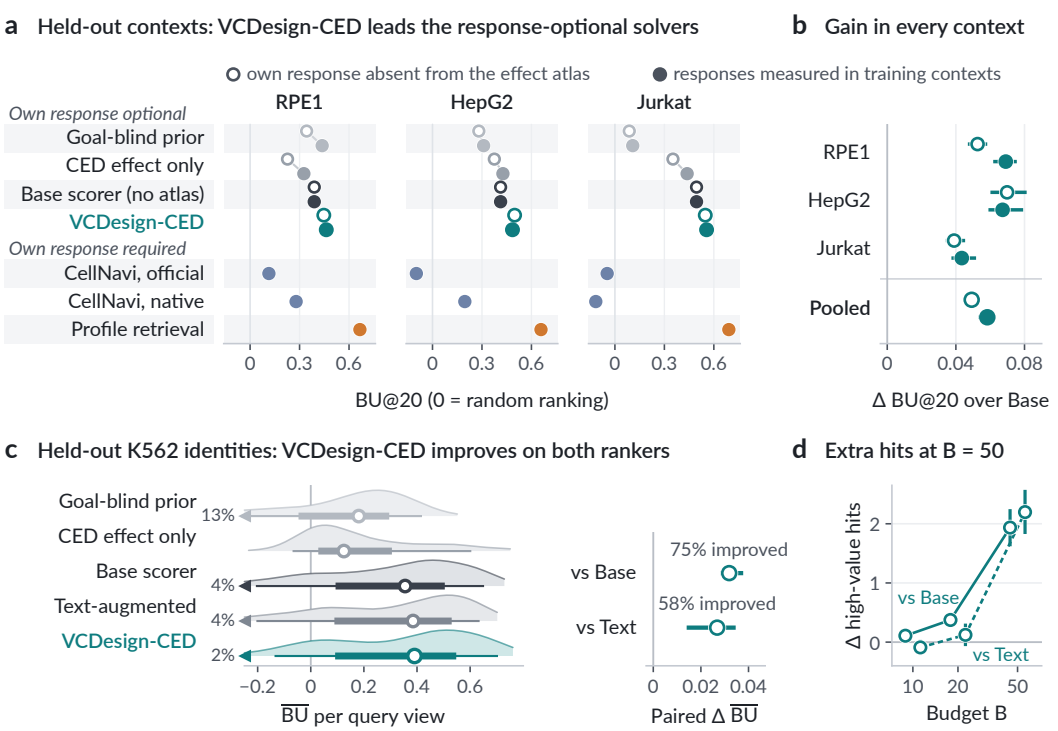

Figure 2: **VCDesign-CED improves selection in held-out contexts and on held-out identities.** **a,** Median BU@20 per held-out context (unit-direction CED); open: own response absent from the effect atlas, filled: measured in training contexts. **b,** Paired gain over the base scorer, 95% query-gene intervals. **c,** Locked K562 split (raw-target CED): row-level $\overline{\mathrm{BU}}$ (density; bars: interquartile and 5th to 95th percentiles), axis cut at -0.25 with the share of rows below; right, median paired gains. **d,** Mean paired gain in high-value (top-5%) candidates among the first B. Intervals in c, d: 95% over query identities; intervals narrower than a marker are hidden.

What does CED transfer when the candidate's own response is absent? The masking design changes one factor at a time (Figure 3b). Removing another masking fold, about 6% of atlas rows, leaves the gain unchanged: the pooled paired difference is $+0.0002$ $[-0.0001, +0.0005]$. Removing the candidates' own responses at equal atlas size lowers it by 0.0078 $[0.0061, 0.0096]$ in median (0.0154 in mean), most in RPE1 and not detectably in HepG2. Most of the gain is therefore carried by the responses of other candidates, possibly close interaction partners (Section F).

Libraries can mix both kinds of candidate. In fixed pools with 0 to 100% response-unseen candidates, scoring all of them with CED keeps a pooled gain of $+0.049$ to $+0.058$ over the base scorer (Figure 3c). A preregistered alternative routes measured candidates to profile retrieval, fused in the same way. With half the pool response-unseen it adds $+0.0125$ $[+0.0100, +0.0150]$ BU@20, shrinking as the unseen share grows ($+0.023$ at 25%, $+0.006$ at 75%). Fusion dilutes retrieval: with every candidate measured, the routed ranking scores 0.101 below retrieval alone (Section D.3).

5.4 What Limits Response-Unseen Design?

We examine how training measurements, candidate knowledge, and effect-model choices support transfer to response-unseen candidates (Figure 4a,b). Across the tested controls, transfer tracks measurement volume more consistently than context count. Atlases built from one, two and three training contexts with all their rows give pooled gains of $+0.025$, $+0.042$ and $+0.049$ over the base scorer. At a row count fixed per held context (840 on average), spreading the same rows over more contexts lowers the gain from $+0.012$ to $+0.001$ (Figure 4a; unmasked arms in Section D.11). Row matching also lowers the number of distinct identities, so this comparison separates measurement volume from context count but not from identity coverage. With identities held fixed and no mask-

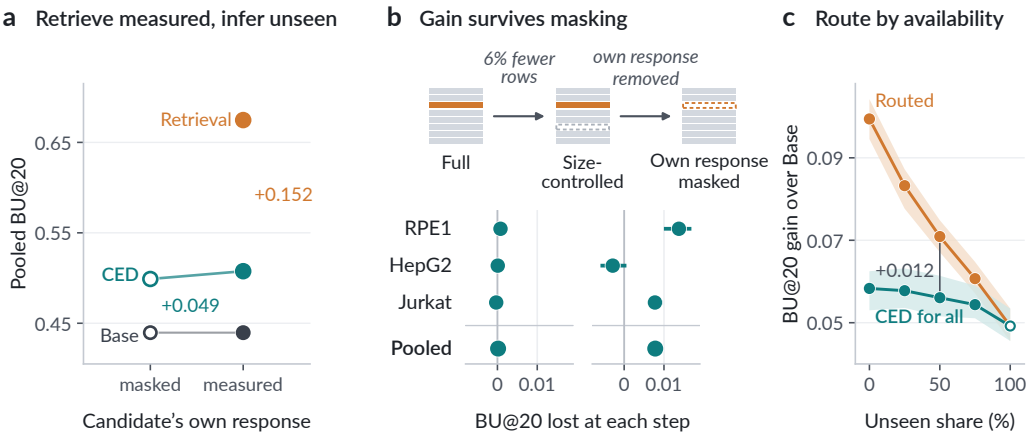

Figure 3: **Response availability sets the solver, and the response-unseen gain is transfer from other candidates' responses. a,** Pooled BU@20 with the candidate's own response masked from the effect atlas or measured elsewhere; labels are paired median gains; CED: VCDesign-CED. **b,** Paired BU@20 lost at each masking step, 95% query-gene intervals (most within the marker). **c,** Pooled BU@20 gain over the base scorer as 0 to 100% of each pool becomes response-unseen: CED scores every candidate, or measured ones are routed to retrieval (preregistered); both arms are fused with the base scorer (Section D.3). Bands: 95% query-gene intervals; label: paired median at 50%.

ing, measuring each identity in all three contexts beats one measurement per identity spread over them by $+0.028$, $+0.021$ and $+0.017$ $\overline{\text{BU}}$ (raw targets; Appendix Figure 7).

The tested effect-model extensions do not establish an advantage over the linear reference. A predictive response basis changes BU@20 by $+0.0009$ $[+0.0006, +0.0013]$, whereas the context-invariant adapter and anchored nonlinear memory show no clear gains. Unanchored memory lowers BU@20 by 0.0233 $[0.0208, 0.0257]$ (Figure 4b). Removing STRING reduces it by 0.0211 $[0.0177, 0.0235]$, compared with 0.0013 when MAP-KG is removed. These controls motivate improving the evidence available for transfer alongside the models that use it. A separate candidate- and query-representation probe is reported in Section D.5.

5.5 Better Prediction Does Not Ensure Better Design

Across masked atlas and input configurations, candidate-specific prediction accuracy and effect-only design utility are positively rank-correlated within each held context (Spearman $\rho = 0.31$, 0.69 and 0.71; stratified permutation test, $p < 10^{-4}$; Figure 4c). Within a fixed atlas, however, four of five changes to basis rank or candidate features move prediction and design in opposite directions in every held context. Thus, prediction accuracy does not reliably select the effect predictor that yields the better finite-budget ranking.

6 Discussion and Conclusion

By making experimental choice itself the common object of evaluation, VCDesign reveals a principle that is more fundamental than the performance of any single model: **the ability to design cellular interventions is not an intrinsic property of a model alone, but depends on what experimental evidence is available and how effectively that evidence can be converted into decisions under a finite budget.** Once different computational approaches are evaluated on this common decision scale, a clear information hierarchy emerges. When historical responses for a candidate are already available, those measurements provide the strongest basis for prioritization; when they are not, perturbation atlases retain value beyond the interventions they directly contain, because evidence accumulated from other perturbations can be transferred through biological knowledge to improve the ranking of response-unseen candidates across both held-out identities and cellular contexts. Mixed candidate libraries reinforce this view: intervention design is not a search for a single

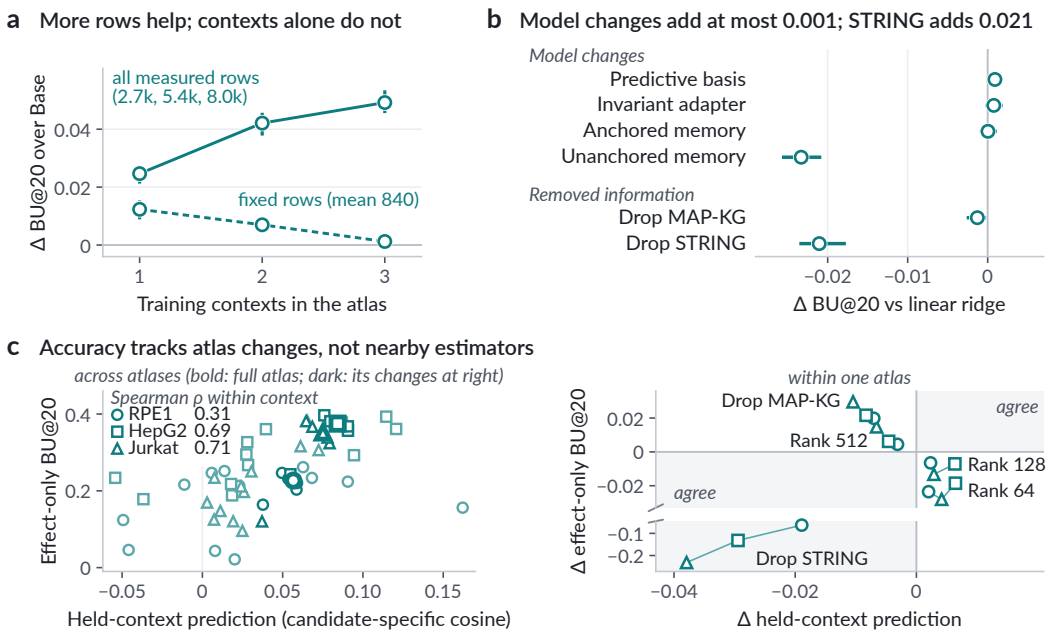

Figure 4: **Atlas composition and candidate knowledge shape response-unseen ranking; prediction accuracy does not reliably order nearby configurations by design utility.** Response-unseen candidates, unit-direction ridge, BU@20. **a,** Gain over the base scorer with all rows (solid) or a fixed row count per held context, 840 on average (dashed). **b,** Change relative to the linear ridge, fused with the base scorer. Intervals in a, b: 95% over query genes, some within the marker. **c,** Effect-only BU@20 against held-context prediction, across 18 configurations per context and within one atlas (point estimates).

model that should dominate every regime, but a problem of matching the available evidence to the appropriate computational route. Controlled analyses further suggest that the remaining bottleneck in response-unseen design lies primarily in the transferable evidence itself. Expanding perturbation measurements and providing informative candidate knowledge materially improve design quality, whereas adding training contexts alone or increasing effect-model complexity does not yield comparable gains; prediction accuracy likewise does not reliably determine which nearby models will make better experimental choices. Moving from prediction and identification to design therefore changes more than the required model output—it changes the scientific question to how biological knowledge and experimental measurements can be organized into useful evidence for deciding what to test next.

These conclusions remain bounded by the scope of the present evidence. Our experiments are restricted to single-gene CRISPRi perturbations across four cell lines and primarily establish improvements in transcriptomic directional prioritization; the realized-outcome analyses do not yet show that these ranking gains translate consistently into improved biological endpoints. Response-unseen transfer also weakens as candidate libraries approach genome scale, indicating that the value of transferable evidence does not automatically extend to arbitrarily large search spaces. At the same time, the data contain substantial and reproducible context-specific perturbation structure, yet the limited number of training contexts is insufficient to learn a context-conditioned effect model that can exploit this structure reliably. This points to a broader challenge for the field: progress may depend first on what should be measured, where it should be measured and how experimental coverage should be expanded, and only then on how increasingly powerful models organize that evidence. Addressing this challenge will require perturbation atlases with greater biological breadth and information content, better integration of candidate-specific measurements with model-based inference, and prospective experiments that test whether computational prioritization produces genuine biological benefit. **If virtual cells are to become part of experimental science, their next defining capability will be to turn the growing body of biological measurements into reliable decisions about which experiments are most worth doing next.**

AI USE STATEMENT

Generative AI tools were used to brainstorm and refine hypotheses and conceptual framing; provide feedback on methodology and experimental design; assist method implementation, data-reformatting code, and debugging; search and synthesize literature; interpret experimental results; revise scientific figures; and draft, translate, and edit manuscript text. They were not used to generate synthetic datasets, formulate or prove mathematical claims, or conduct qualitative or thematic data analysis; proof-related tasks were not applicable. The authors checked AI-assisted ideas against the literature, reviewed and tested code, verified reported values against saved artifacts, and approved every scientific decision and claim. The authors take responsibility for the final content.

ETHICS STATEMENT

This work uses only publicly released CRISPRi Perturb-seq data from established human cell lines (K562 and RPE1 from Replogle et al. (2022); HepG2 and Jurkat from Nadig et al. (2025)) and public knowledge resources (Szklarczyk et al., 2023; Feng et al., 2026; Lin et al., 2023). No new experiments were performed, and no human-subjects or patient data were collected or used. VCDesign-CED ranks genetic interventions by how well they are predicted to realize a desired transcriptional transition in cultured cells. Its outputs are hypotheses that prioritize laboratory experiments, and they require experimental validation before any biological conclusion is drawn. The dual-use potential of such a ranking is limited, for three reasons. It ranks gene knockdowns, an intervention class already screened genome-wide in public atlases. It scores transcriptomic states of cell lines rather than organismal phenotypes. It creates no new biological agent or delivery method. Use in disease-related or therapeutic settings would call for functional validation and the biosafety and ethics review that such work normally requires.

REPRODUCIBILITY STATEMENT

Section 3 defines the design problem, the model interface and the evaluation interface, including the utility and the metrics, and Section A.1 gives their exact computation. Section 4 and Section B specify the base scorer, Candidate Effect Distillation and the fusion rule, with all hyperparameters, random seeds and compute. For Candidate Effect Distillation this covers the response basis, the features, the ridge penalty grid and how the penalty is selected. Section A describes the data sources, preprocessing, batch-disjoint replicate views, identity splits and the query-eligibility rule. Every split used in the paper (identity roles, replicate views and masking folds) is a deterministic SHA-256 hash of public identifiers, so it can be regenerated from the public data without stored index files. Section C lists the statistic, bootstrap procedure, seeds and resampling unit for every result family. Sections D and E report secondary endpoints and effect-model variants.

The anonymized supplementary code archive contains the source for preprocessing, model fitting and evaluation, a portable run configuration, protocol tests and a synthetic example.

DATA AND CODE AVAILABILITY

All experimental inputs are previously released datasets or public knowledge resources. K562 and RPE1 data come from the release associated with Replogle et al. (2022); HepG2 and Jurkat data are available from GEO under accession GSE264667 (Nadig et al., 2025). STRING v12.0, MAP-KG, and ESM-2 are available from their cited releases. An anonymized code archive accompanies this submission; it contains the model and evaluation source, a run configuration, tests, an example and a checksum manifest. Large processed inputs, checkpoints and figure source-data files are not included in that archive.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, Francis Chalissery, Rajesh Ilango, Vishvak Subramanyam, Chiara Ricci-Tam, Sanjay Nagaraj, Aidan Winters, Mingze Dong, Stefanie Fellinger, Adam Krejci, Tilmann Burckstummer, Sravya Tirukkovular,

Jeremy Sullivan, Brian S. Plosky, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 189(19):5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8): 1657–1661, 2025. doi: 10.1038/s41592-025-02772-6.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Zhi-Jie Cao, Tianzi Guo, Cheng Li, Fan Zhang, Ping Lu, Chan Gu, Fan Guo, and Ge Gao. Causality-oriented regulatory inference and rational transcriptomic reprogramming design with CASCADE. *bioRxiv*, 2025. doi: 10.1101/2025.11.20.689377.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Jinghao Feng, Ziheng Zhao, Xiaoman Zhang, Mingfei Liu, Jingyi Chen, Xingran Quan, Boyang Fu, Jian Zhang, Yanfeng Wang, Ya Zhang, and Weidi Xie. A knowledge-driven framework for predicting single-cell responses for unprofiled drugs. *Nature Machine Intelligence*, 8(9):1478–1491, 2026. doi: 10.1038/s42256-026-01286-w.

Guadalupe Gonzalez, Xiang Lin, Isuru Herath, Kirill Veselkov, Michael Bronstein, and Marinka Zitnik. Combinatorial prediction of therapeutic perturbations using causally inspired neural networks. *Nature Biomedical Engineering*, 10(5):1008–1025, 2026. doi: 10.1038/s41551-025-01481-x.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011. doi: 10.1137/090771806.

Younghyun Han, Hyunjin Kim, Chun-Kyung Lee, and Kwang-Hyun Cho. Identifying an optimal perturbation to induce a desired cell state by generative deep learning. *Cell Systems*, 16(10): 101405, 2025. doi: 10.1016/j.cels.2025.101405.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Justin Lamb, Emily D. Crawford, David Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N. Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A. Armstrong, Stephen J. Haggarty, Paul A. Clemons, Ru Wei, Steven A. Carr, Eric S. Lander, and Todd R. Golub. The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795): 1929–1935, 2006. doi: 10.1126/science.1132939.

Shiyang Liang, Cheng Yang, Jingjie Wang, and Yunqi Li. RevPert: predicting candidate drivers of transcriptomic state transitions via gallery-native reverse perturbation. *bioRxiv*, 2026. doi: 10.64898/2026.08.19.745674.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019. doi: 10.1038/s41592-019-0494-8.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6), 2023. doi: 10.15252/msb.202211517.

Arash Mehrjou, Ashkan Soleymani, Andrew Jesson, Pascal Notin, Yarin Gal, Stefan Bauer, and Patrick Schwab. GeneDisco: A benchmark for experimental design in drug discovery. In *International Conference on Learning Representations*, 2022.

Ajay Nadig, Joseph M. Replogle, Angela N. Pogson, Mukundh Murthy, Steven A. McCarroll, Jonathan S. Weissman, Elise B. Robinson, and Luke J. O'Connor. Transcriptome-wide analysis of differential expression in perturbation atlases. *Nature Genetics*, 57(5):1228–1237, 2025. doi: 10.1038/s41588-025-02169-3.

Qiangwei Peng, Yuchuan Wang, Jianzhe Li, Xinyu Wang, Yao Xiao, and Peijie Zhou. Unbalanced perturbation dynamics for cell fate design. *bioRxiv*, 2026. doi: 10.64898/2026.06.30.735555.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11671.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6.

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, David L. Lahr, Jodi E. Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C. Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M. Enache, Federica Piccioni, Sarah A. Johnson, Nicholas J. Lyons, Alice H. Berger, Alykhan F. Shamji, Angela N. Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y. Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S. Gray, Paul A. Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao,

Willis Read-Button, Xiaohua Wu, Stephen J. Haggarty, Lucienne V. Ronco, Jesse S. Boehm, Stuart L. Schreiber, John G. Doench, Joshua A. Bittker, David E. Root, Bang Wong, and Todd R. Golub. A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, 2017. doi: 10.1016/j.cell.2017.10.049.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023. doi: 10.1093/nar/gkac1000.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44(6):1050–1059, 2026. doi: 10.1038/s41587-025-02777-8.

Tianze Wang, Yan Pan, Fusong Ju, Shuxin Zheng, Chang Liu, Yaosen Min, Qun Jiang, Xinwei Liu, Huanhuan Xia, Guoqing Liu, Haiguang Liu, and Pan Deng. CellNavi predicts genes directing cellular transitions by learning a gene graph-enhanced cell state manifold. *Nature Cell Biology*, 27(10):1863–1874, 2025. doi: 10.1038/s41556-025-01755-1.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2019. doi: 10.1109/tpami.2018.2857768.

Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D Welch. PerturbNet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 21(8), 2025. doi: 10.1038/s44320-025-00131-3.

Yuansong Zeng, Jiancong Xie, Ningyuan Shangguan, Zhuoyi Wei, Wenbing Li, Yun Su, Shuangyu Yang, Chengyang Zhang, Jinbo Zhang, Nan Fang, Hongyu Zhang, Yutong Lu, Huiying Zhao, Jue Fan, Weijiang Yu, and Yuedong Yang. CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nature Communications*, 16(1):4679, 2025. doi: 10.1038/s41467-025-59926-5.

A Data and Metrics

Contexts. Table 2 summarizes the four CRISPRi Perturb-seq screens. K562 and RPE1 come from Replogle et al. (2022). HepG2 and Jurkat come from the dual-sgRNA CRISPRi screens of Nadig et al. (2025) (GEO GSE264667, day 7, 10x Genomics 3′ v3), read from the public single-cell H5AD files without reprocessing. In these two screens, identities are taken from the gene-level perturbation annotation, cells labelled non-targeting form the controls, and batches are the 10x GEM groups. All cells in the files are retained.

Two gene axes connect the contexts. K562 and RPE1 share 7,094 measured genes, 86.0% of the 8,248 K562 genes. The RPE1 zero-shot experiment places RPE1 states on the K562 gene axis at these shared genes. All four contexts share 6,550 genes. This shared 6,550-gene axis was fixed before any model was fit, and the leave-one-context-out, masking and atlas-scale experiments all use it.

HepG2 is shallower than Jurkat. Over all perturbed identities of each screen, including those below the usable depth, the median identity has 45 cells in HepG2 and 83 in Jurkat, summed over both views.

Responses and queries. Counts are normalized to a library size of 10^4 and transformed by $\log(1+x)$. For an identity in view v, the goal state is the pseudobulk of the identity's perturbed cells in the batches of v. The source state is a batch-matched control reference. It is the mean of the per-batch non-targeting control profiles, weighted by the identity's cell counts across the batches of v. The

Table 2: **Per-context data.** K562 trains the model; RPE1, HepG2 and Jurkat are held out. Sources: K562 and RPE1, Replogle et al. (2022); HepG2 and Jurkat, Nadig et al. (2025). Usable identities have at least 25 cells in each of the two batch-disjoint replicate views. The shared axis is the 6,550-gene axis common to the four contexts. Eligible queries pass the measurability rule of Section A; for K562 they are counted on the held-out evaluation split. Candidate pools are listed per experiment, and a query's own identity is always excluded from its pool. The three held-out contexts together contribute 1,333 queries (2,666 query-view rows) from 976 query genes to the pooled masking results. –: not applicable.

	K562	RPE1	HepG2	Jurkat
Role	training	held-out	held-out	held-out
Cells	1,989,578	247,914	145,473	262,956
Perturbations	9,866	2,393	2,393	2,393
Batches (views 1 / 2)	267 (131 / 136)	56 (26 / 30)	56 (26 / 30)	55 (26 / 29)
Measured genes	8,248	8,749	9,622	8,880
Usable identities	9,205	1,483	881	1,696
training and development identities on the shared axis	7,825			
Eligible queries	342 / 1,380	502	207	624
Pool, evaluation split	1,380	–	–	–
Pool, zero-shot	–	1,483	–	–
Pool, masking	1,430	1,273	764	1,450

transition is goal minus source. The consensus response $\mathbf{r}_c$, which CED uses as its target, is the mean of an identity's two view transitions and is computed on training identities only. Every gene coordinate is kept. So when the knocked-down gene is on the measured gene axis, its own expression is part of the query transition and of every view response.

Batch-disjoint replicate views. A SHA-256 hash of the batch name assigns every batch of a context to view 1 or view 2 (K562: 131 and 136 batches). A second, independent hash splits each view into two quarters (K562: 66, 65, 63 and 73 batches), so the four quarters are also batch-disjoint. The same rule is applied to each context's own batch names. K562 uses the quarters for the realized-objective analyses; RPE1, HepG2, and Jurkat additionally use them for the PHR guardrail on identities with at least 25 cells in every quarter (Section A.1).

K562 identity split. A SHA-256 hash of the identity name with a fixed salt assigns the 9,205 usable K562 identities to three splits: 6,445 training identities, a development split of 1,380 and a held-out evaluation split of 1,380 (the locked evaluation split of Section 5.1) ($6{,}445 + 1{,}380 + 1{,}380 = 9{,}205$). The splits are identity-disjoint. An evaluation-split identity is never a training target, a training query or a training candidate, and the split plays no part in checkpoint, penalty or fusion selection. Its configuration was recorded before it was scored (Section C).

Query eligibility. A transition can grade a ranking only if its two views agree, because one view forms the query and the other grades it. Consider a split of n identities with view responses $\mathbf{r}_g^{(1)}$ and $\mathbf{r}_g^{(2)}$ (Section 3). The agreement of query q is $E_q = \langle \mathbf{r}_{g_q}^{(1)}, \mathbf{r}_{g_q}^{(2)} \rangle$. Its one-sided permutation p-value against mismatched partners is

$$p_q = \frac{1 + \left| \{\, g' \neq g_q : \langle \mathbf{r}_{g_q}^{(1)}, \mathbf{r}_{g'}^{(2)} \rangle \geq E_q \,\} \right|}{n}. \tag{6}$$

A query is eligible when its Benjamini–Hochberg q-value (Benjamini & Hochberg, 1995) is at most 0.05 and $E_q > 0$. The rule and its level were fixed on the K562 training identities and are applied unchanged to every split and every context. On the evaluation split, 342 of 1,380 queries (24.78%) are eligible. In RPE1, HepG2 and Jurkat, 502 of 1,483 (33.9%), 207 of 881 (23.5%) and 624 of 1,696 (36.8%) usable identities are eligible. Eligibility filters queries only: on the evaluation split and in RPE1 zero-shot, ineligible identities remain in the candidate pool.

A.1 METRIC DEFINITIONS

Utility and ranking. The utility $U(q, c)$ is the opposite-view cosine of Equation (1). It is signed and continuous, and it is never a model input. It is computed from measurements independent of those that formed the query, because the grading view shares no batch with the query view. The pool $\mathcal{C}_q$ excludes the query's own identity, and $N_q = |\mathcal{C}_q|$. A method's ranking $\pi_q = (c_1, c_2, \dots)$ sorts $\mathcal{C}_q$ by descending score, with ties broken by ascending candidate identifier.

Budgeted utility. $\mathrm{BU}@B$ and $\overline{\mathrm{BU}}$ are defined in Equation (2) and Section 3.2, with rank weights $w_j = 1/\log_2(j+1)$ and $\mathrm{RU}^\star@B$ the rank-weighted utility of the pool sorted by U. Two details complete the definition. First, the denominator of Equation (2) is floored at 10^{-8}, so a query whose best prefix equals the random expectation never divides by zero. Second, $\mathrm{RU}_{\mathrm{rand}}(q)$, the mean utility of the pool, is the exact expectation of $\mathrm{RU}@B$ under a uniformly random ordering, because every rank then has the same expected utility. It therefore does not depend on B, and it is computed exactly, without sampled orderings.

Replicate ceiling. On the development split, ranking candidates by one replicate view of measured utility and grading in the other gives $\mathrm{BU}@10$ 0.766 and 0.797, $\mathrm{BU}@20$ 0.736 and 0.738, and $\mathrm{BU}@50$ 0.687 and 0.698 for the two grading directions. This reference uses development-split queries and candidates measured in all four quarters; it is not directly comparable with the evaluation-split measured-response oracle ($+0.8142$).

Secondary metrics. $\mathrm{MU}@B(q) = \frac{1}{B}\sum_{j=1}^{B} U(q, c_j)$ is the plain mean utility of the top B. $\mathrm{HvHit}@B(q) = \left|\{\, j \le B : U(q, c_j) \ge Q_{0.95}(q) \,\}\right|$ counts the top-B candidates at or above $Q_{0.95}(q)$, the 0.95 quantile of $\{U(q, c) : c \in \mathcal{C}_q\}$. Every metric is computed per query-view row, and tables report its median over rows.

Paired statistics and win rate. For a first and a second method evaluated on the same n rows, with per-row metric values $m_1(q)$ and $m_2(q)$, the paired difference of a row is $\Delta_q = m_1(q) - m_2(q)$. We report the median and the mean of Δ_q over rows with 95% bootstrap intervals (Section C), and the win rate $\frac{1}{n}\sum_q \mathbf{1}[\Delta_q > 0]$, the share of rows on which the first method is strictly higher.

Four-quarter realized outcomes. This secondary endpoint asks whether executing a top-ranked intervention moves the cell closer to the goal than doing nothing. The intervention may be attenuated. For an attenuation $a \in [0, 1]$,

$$J_{qc}(a) = 2a\,\alpha_{qc} - a^2 R_c, \tag{7}$$

which in expectation equals $\|\mathbf{d}_q\|^2 - \|a\,\mathbf{r}_c - \mathbf{d}_q\|^2$, the reduction in squared distance to the goal relative to no intervention. Here α_{qc} is the alignment of the candidate's response with the query's transition, and R_c is the candidate's response energy. Both are computed as products across two batch-disjoint quarters. No measurement then multiplies itself, so neither statistic carries a noise-energy bias. For a query formed in view v, the attenuation a^* that maximizes J is chosen on the two quarters of v. The value $J_{qc}(a^*)$ is then realized on the two quarters of the opposite view.

Two summaries are reported. $J\,\mathrm{RU}@B$ is $\mathrm{RU}@B$ with the realized J in place of U, averaged over the two views of a query. $J{>}0@B$ is the share of top-B picks with positive realized J, pooled over queries in the first view. For the locked secondary endpoint, both are computed on the 1,053 evaluation-split identities with at least 25 cells in every quarter, which form the candidate pool, and on the eligible queries among them; the utility comparisons of Section E apply the same four-quarter construction to development-split queries. Values are reported in Section D.

The external-context PHR guardrail uses the same construction. For each query view, a^* is chosen on that view's quarter pair and $J_{qc}(a^*)$ is realized on the opposite pair. The ranking is restricted to candidates with at least 25 cells in all four quarters, and the query must meet the same requirement. For the resulting top-B prefix,

$$\mathrm{PHR}@B(q) = \frac{1}{B}\sum_{j=1}^{B} \mathbf{1}\left[J_{qc_j}(a^*_{qc_j}) > 0\right]. \tag{8}$$

Table 15 reports mean PHR@20 levels over query-view rows. Paired contrasts report both the mean and the preregistered median; pooled inference includes only contexts with at least 200 four-way candidates.

Latent endpoint value. The endpoint-distance objective compared in Section E is this same construction at full strength, $V_{qc} = J_{qc}(1) = 2\alpha_{qc} - R_c$: the estimated reduction in squared distance to the goal when the candidate's response is realized undiscounted. It is used there as the utility in place of U, so a ranking formed on the two quarters of one view is graded by V on the two quarters of the other view. Unlike Equation (1) it keeps effect magnitude, and its second term $-R_c$ is a property of the candidate alone.

B Implementation Details

This appendix lists every architectural choice and hyperparameter of VCDesign-CED, and Table 3 collects the settings in one place. Unless stated otherwise, all components operate on log-normalized pseudobulk states over the $8{,}248$ genes of the K562 screen.

Base scorer architecture. The base scorer has $30{,}111{,}233$ parameters. The source encoder maps x_s to the source embedding $\mathbf{h}_s$ through Linear(8248, 1024), LayerNorm, GELU, Dropout, Linear(1024, 512), LayerNorm and GELU. The transition encoder maps $\mathbf{d}_q = x_t - x_s$ to the transition embedding $\mathbf{h}_d$ through Linear(8248, 2048), LayerNorm, GELU, Dropout, Linear(2048, 1024), LayerNorm, GELU, Dropout, Linear(1024, 512) and LayerNorm. The FiLM layer (Perez et al., 2018) is one Linear(512, 1024) on $\mathbf{h}_s$, whose output is split into $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Its weight is initialized to zero and its bias to $(\mathbf{1}, \mathbf{0})$, so the modulation is the identity at initialization. The query embedding is Linear(512, 256) applied to $\boldsymbol{\gamma} \odot \mathbf{h}_d + \boldsymbol{\beta}$, followed by ℓ_2 normalization.

On the candidate side, each knowledge modality with input dimension d has an adapter: LayerNorm(d), Linear(d, 512), GELU, Dropout, Linear(512, 128) and LayerNorm. A learned 128-dimensional modality embedding, initialized from $\mathcal{N}(0, 0.02^2)$, is added, and each token is ℓ_2-normalized. Tokens are averaged over present modalities, projected by Linear(128, 256) and ℓ_2-normalized. A candidate with no present modality receives the zero vector; there is no learnable placeholder token. All dropout rates are 0.10. The temperature is $\tau = \min(e^{\theta}, 100)$, with θ initialized to $\log(1/0.07)$.

Candidate knowledge. Both knowledge blocks come from public static sources. The STRING block is built from the STRING v12.0 physical subnetwork (Szklarczyk et al., 2023), using the experimental and database channels with text mining set to zero. It is compressed to 512 dimensions by truncated SVD (Halko et al., 2011) and standardized on the K562 training identities where present. The MAP-KG block is the $1{,}024$-dimensional output of the released MAP-KG gene projector (Feng et al., 2026), applied to the released ESM-2 gene dictionary (Lin et al., 2023). The projector comes from the released knowledge-encoder checkpoint. That checkpoint is pretrained on 14 public static sources (gene, drug and relation databases), which contain no expression matrix or perturbation response. The separately released downstream perturbation-prediction checkpoint is not used.

On the development split, the STRING block covers 0.9232 of identities and the MAP-KG block 0.9333. Each block has a presence flag, which also serves as the pooling mask of the candidate encoder. 376 of the $6{,}445$ K562 training identities have no modality. They lie outside the knowledge universe of both sources, so they score exactly zero in both branches.

Three further modalities are excluded from the knowledge set. Gene-description text embeddings can restate published perturbation phenotypes. A sequence embedding is redundant, because its ESM-2 asset is the dictionary that the MAP-KG projector already reads. Reactome pathway membership covers only 0.5964 of identities. Finally, STRING's experimental channel and MAP-KG's pretraining graph both include BioGRID relations. These features are identical in every compared arm.

Base scorer training. The base scorer learns from pairs of candidates whose utility gap exceeds the measurement noise. Each K562 training identity serves as a query in both replicate views, and the label is the utility of Equation (1), computed among training identities. For query-view row

q, let $\mathcal{S}_q$ be the union of three candidate sets: the utility top 128, the predicted top 128 under the current weights, and 256 uniformly random candidates. The query's own identity is excluded. The measurement margin m_q is the median over candidates of $\left|\cos(\mathbf{r}_{g_q}^{(1)}, \mathbf{r}_c^{(1)}) - \cos(\mathbf{r}_{g_q}^{(2)}, \mathbf{r}_c^{(2)})\right|$. It is the typical disagreement between the utilities that the two replicate views assign to the same candidate. The pair set $\mathcal{P}_q$ holds all ordered pairs (c, c') in $\mathcal{S}_q$ with $U(q, c) - U(q, c') > m_q$, subsampled uniformly to at most 4,096 pairs. Over a minibatch $\mathcal{Q}$, the pairwise loss is

$$\mathcal{L}_{\text{pair}} = \frac{\sum_{q \in \mathcal{Q}} \sum_{(c,c') \in \mathcal{P}_q} \omega_{qcc'} \, \log\bigl(1 + e^{s_{\text{base}}(q,c') - s_{\text{base}}(q,c)}\bigr)}{\sum_{q \in \mathcal{Q}} \sum_{(c,c') \in \mathcal{P}_q} \omega_{qcc'}}, \tag{9}$$
$$\omega_{qcc'} = \text{clip}\Bigl(\frac{U(q,c) - U(q,c')}{\text{IQR}_q},\, 0.25,\, 4\Bigr)\, \rho\bigl(\text{rank}_q(c)\bigr),$$

where IQR_q is the interquartile range of $U(q, \cdot)$ over the candidate set. The rank weight ρ focuses the loss on the top of the ranking. It is 4, 2, 1 and 0.5 for a better candidate at utility rank at most 20, at most 64, at most 128, and beyond.

Training has two phases. Epochs 1 to 5 minimize $\mathcal{L}_{\text{pair}}$. Epochs 6 to 10 minimize $\mathcal{L}_{\text{pair}} + 0.5\,\mathcal{L}_{\text{set}}$, where the set loss is

$$\mathcal{L}_{\text{set}}(q) = \log \sum_{c \in \mathcal{C}_q} e^{s_{\text{base}}(q,c)} - \log \sum_{c \in \mathcal{T}_q} e^{s_{\text{base}}(q,c)}, \tag{10}$$

averaged over the minibatch. It raises the set $\mathcal{T}_q$ of the 20 highest-utility candidates above the rest of $\mathcal{C}_q$. Its size matches the intended budget of about 20 experiments. Training uses AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-4} and weight decay 10^{-4}, gradient-norm clipping at 1.0, bfloat16 autocast, and seed 20260916. Each minibatch holds 8 training identities, that is, 16 query-view rows with both views of each identity. Checkpoints from epochs 6 to 10 are compared by median $\overline{\text{BU}}$ on the development split, and epoch 8 is selected. The fused model selects the same epoch under its own score, so VCDesign-CED and the base-scorer arm share every weight.

Text-augmented ranker. The text-augmented ranker has the dual-encoder architecture of the base scorer and is trained with the same recipe: seed 20260916, AdamW, the pairwise loss in epochs 1 to 5, and the set loss added in epochs 6 to 10. It differs in two ways. Its knowledge also includes gene-description text embeddings (NCBI gene descriptions encoded with all-MiniLM-L6-v2, 384 dimensions). A candidate without knowledge receives a learnable token. The base scorer omits the text, because the text can restate published phenotypes, and gives such candidates the zero embedding.

Candidate Effect Distillation. The consensus response is $\mathbf{r}_c = \frac{1}{2}(\mathbf{r}_c^{(1)} + \mathbf{r}_c^{(2)})$, the mean of the candidate's two view transitions. The basis $\mathbf{W}$ is the rank-256 uncentered randomized SVD of the training consensus responses (seed 20260918, oversampling 32, four power iterations). The feature vector has 1,538 dimensions: the 512 STRING and 1,024 MAP-KG values, each multiplied by its presence flag, and the two flags. Features are standardized with the mean and standard deviation over training rows, and the scale is floored at 10^{-8}. The intercept is the mean target. The ridge is fit on the K562 training identities that carry at least one modality (376 of the 6,445 carry none). The penalty grid is $\lambda \in \{10^{-2}, 10^{-1}, \ldots, 10^{5}\}$ (eight values). Five identity folds (fold seed 20260918) select the penalty with the highest mean held-out gene-space cosine between $\mathbf{W}\hat{\mathbf{z}}_c$ and $\mathbf{r}_c$; this is $\lambda = 10^4$ on K562. A candidate with no modality receives $\hat{\mathbf{e}}_c = \mathbf{0}$ and an effect score of 0.

Fusion. The final fusion fits no additional weight. The per-query standardization z_q of Equation (5) uses the population standard deviation over the candidate set, floored at 10^{-12}, and the effect term enters with unit weight. Fusion form and weight were selected on the development split. In the comparison using the text-augmented ranker as the base, equal-weight fusion improved top-10 mean utility and realized-objective RU@10; an effect weight of 0.25 fitted on K562 training identities reached a higher median $\overline{\text{BU}}$ level but left those two quantities unchanged relative to the text-augmented ranker.

Table 3: Hyperparameters of VCDesign-CED. The base scorer and CED are trained on the K562 training identities. Cross-validation on these identities selects the ridge penalty, and median $\overline{\mathrm{BU}}$ on the development split selects the checkpoint epoch. Every value was fixed before the held-out evaluation split was scored.

Component	Setting	Value
Base scorer	parameters	30,111,233
	source / transition embedding width	512 / 512
	modality token width; joint embedding width	128; 256
	dropout	0.10
	temperature clamp; initial temperature	100; $1/0.07$
Training	optimizer	AdamW
	learning rate; weight decay	10^{-4}; 10^{-4}
	gradient-norm clip	1.0
	minibatch	8 identities (16 query-view rows)
	epochs (pairwise; pairwise + set)	1–5; 6–10
	pair pool: utility top; predicted top; random	128; 128; 256
	max. pairs per query	4,096
	gap-weight clip; rank weights	$[0.25, 4]$; $4/2/1/0.5$
	set-loss top size; set-loss weight	20; 0.5
	selected epoch	8
	seed	20260916
CED	basis rank r	256
	SVD seed; oversampling; power iterations	20260918; 32; 4
	feature dimension	1,538
	penalty grid; CV folds	10^{-2}–10^{5}; 5
	selected penalty (K562)	10^{4}
Fusion	effect weight; std. floor	1; 10^{-12}

Cross-context deployment. The base scorer is the frozen K562 scorer in every context. States of another context are embedded on the 8,248-gene K562 axis at the shared genes, with zeros at the remaining genes. The effect predictor depends on the experiment. In the RPE1 zero-shot experiment, it is the K562 effect predictor as fitted on K562, and the effect cosine is computed on the 7,094 genes that K562 and RPE1 share. In the leave-one-context-out, masking and atlas-scaling experiments, it is refit on the shared 6,550-gene axis of the four contexts. The refit uses one equally weighted row per (context, identity) pair, no context feature, and the rank, seed, penalty grid and cross-validation above. Per-fold refits under masking follow the protocol of Section D.9.

Compute. All experiments ran on one workstation with a single NVIDIA GeForce RTX 4090 GPU (24 GB) and an AMD Ryzen 9 7900X CPU, using Python 3.11, PyTorch 2.4.1 and NumPy 1.26. Training the base scorer takes 213 seconds for all ten epochs: about 19 seconds per pairwise epoch and 23 seconds per epoch with the set loss. CED is a closed-form ridge fit. The complete leave-one-context-out experiment, which refits CED 28 times (four held contexts, seven training atlases each) and scores and bootstraps every fold, took 4.3 minutes.

C Statistical Procedures

Every comparison in the paper is paired. All arms rank the same candidate pool for the same query-view rows under the same utility, so each row yields one paired difference per contrast. This section states the replicate structure, the statistics and the resampling unit of each result family.

Replicate structure. Each usable identity gives two query-view rows. Each row forms its query in one view and grades it with the responses of the other (Section 3). A knockdown identity is usable when it has at least 25 cells in each of two batch-disjoint replicate views. Views are assigned by a SHA-256 hash of the batch name (K562: 131 and 136 batches), and the same rule is applied to each context's own batch names. K562 identities carry about 108 cells per view. Over all perturbed identities, the median HepG2 and Jurkat identity has 45 and 83 cells summed over both views. A

second hash splits each view into two quarters (K562: 66, 65, 63 and 73 batches). These quarters support the K562 realized objective and the external-context PHR guardrail reported in Section D.

Query eligibility. A query is graded only when its two replicate views reproduce its transition. The rule is a permutation test with Benjamini–Hochberg control, defined once in Section A. Eligible transitions are far more reproducible than rejected ones. On the development split, ranking candidates by their measured-response cosine in one view and grading in the other gives $\overline{\mathrm{BU}}$ 0.7355 on eligible queries and 0.1571 on rejected ones.

Statistics. The level of an arm is its median over query-view rows. A contrast between two arms is the median of the per-row paired differences. On the held-out evaluation split, the mean paired difference is reported beside it (Table 5). The win rate is the share of rows with a strictly positive paired difference, so ties do not count as wins. Intervals are 95% percentile bootstrap intervals with 10,000 replicates and seed 20260916. In each replicate, resampling units are drawn with replacement, every row of a drawn unit enters, and the median (or mean) is recomputed. Table 4 lists the resampling unit of each result family. Intervals describe variation over queries with the trained models held fixed. The primary K562 comparison is VCDesign-CED against the base scorer on $\overline{\mathrm{BU}}$; the primary cross-context contract comparison is the same contrast on masked BU@20. The K562 configuration was fixed before the locked evaluation split was scored; the cross-context configuration, whose unit-direction targets were selected on the same held-context pools by a preregistered comparison, was fixed before any external baseline was run. Other contrasts are secondary, and their intervals are not adjusted for multiplicity.

Prediction and design. The within-context association of Figure 4c is the Spearman coefficient over the distinct configurations of each held context (18 each, after removing three fits that duplicate another because the smallest atlas is already row-matched). Its significance is a stratified permutation test: design values are permuted within each held context, the statistic is the mean of the three within-context coefficients (0.572), and 20,000 permutations with seed 20260916 give $p < 10^{-4}$. The pooled coefficient over all 57 points, 0.610, treats points as independent and is reported only as a descriptive summary (Section D.6).

Seeds and deterministic components. The base scorer is trained once, with seed 20260916. VCDesign-CED and the base-scorer arm share its epoch-8 checkpoint, so the evaluation-split comparison isolates CED. The response basis is computed by randomized SVD (Halko et al., 2011) with seed 20260918 (32 oversampling columns, 4 power iterations). The ridge fit and the fusion rule are deterministic. The random-ranking baseline is a single draw. The raw-target row-matched atlas controls average five fixed row-sampling seeds per arm. The vocabulary-size curve takes the median over three random subsamples per size, and the full class is used at its own size. Every other assignment in the paper is a deterministic SHA-256 hash of an identifier: identity roles, replicate views, context quarters, masking folds, the response-unseen share of mixed vocabularies and the row subsamples of the unit-direction row-matched atlases.

Each CellNavi level comes from one fine-tune per context and class-set configuration, so no training-seed interval is claimed. GEARS uses the one measured-setting and five masked fits prescribed by the contract; every fit is degenerate. Intervals for external comparisons resample query genes with the trained models held fixed.

Evaluation-split configuration. The checkpoints, ridge penalty, fusion rule, eligibility rule and metrics of the evaluation-split comparison were recorded in one configuration file before the held-out evaluation split was scored, and the split was scored once with that configuration. The fusion form and its unit weight were chosen on the development split before this point.

D Additional Results

This appendix reports secondary endpoints of the held-out evaluation split, results on the development split, details of the context-transfer experiments, the controlled masking and vocabulary analyses of Section 5.3, the row- and identity-matched atlas controls, the K562 fold of the held-context

Table 4: **Resampling unit of each result family.** All intervals are 95% percentile bootstrap intervals (10,000 replicates, seed 20260916) of the median paired difference, or of the mean where a mean is reported. A cluster holds every query-view row that shares its label. Pooled rows cover RPE1, HepG2 and Jurkat, each held out from training in turn. Families that pool contexts resample query genes, so the views and contexts of one gene stay together.

Result family	Reported in	Rows	Resampling unit
K562 locked evaluation split	Table 5, Figure 2c,d	684 (342 queries)	query identity (both views)
RPE1 zero-shot, with fusion controls	Tables 7 and 17	1,004 (502 queries)	query-view row
Pooled masking gains	Table 6	2,666 (1,333 queries; 976 genes)	query gene (all views, contexts)
Pooled estimator ablations	Table 7; pooled rows of Tables 25 and 26	2,666 (1,333 queries; 976 genes)	query gene (all views, contexts)
External-baseline headline contrasts	Table 14, Figure 3a	pooled RPE1, HepG2 and Jurkat	query gene (all views, contexts)
Held-context gains over the base scorer, masking arm differences and estimator variants	Figure 2b, Figure 3b, Figure 4a,b	per context and pooled	query gene (all views, contexts)
Mixed-pool routing and candidates the base scorer never trained on	Figure 3c, Tables 8, 10 and 11	2,666 (1,333 queries; 976 genes); untrained-query row 1,002 (361 genes)	query gene (all views, contexts)
High-value-hit differences	Figure 2d	684 (342 queries)	query identity (both views; mean)
PHR@20 contrasts	Table 15	four-way sub-pools; pooled inference uses powered contexts	query gene (all views, contexts)
Value of historical response exposure	Table 6	per held context	query identity (both views)
K562 fold at full class size (6,255 response-unseen, 1,570 measured)	Figure 5 and Tables 23 and 26	development split	query-view row
Development-split contrasts	Section E	development split	query-view row; query identity (views averaged) for utility-choice intervals on the mean

Point estimates (median paired difference, no interval): per-context raw-target masking gains; leave-one-context-out transfer; raw-target atlas-scale controls (Figures 6 and 7) (row-sampled arms averaged over five seeds); per-context mixed-pool gains (Table 9) and raw-target mixed vocabularies; vocabulary-size curve below full class size (median over three random subsamples per size); support terciles; reconstruction cosines; per-context rows of Tables 25 and 26 (raw norm alone, unit-direction targets, retention, win rate); the four-fold context-conditioned map and planted residual; CellNavi levels (one fine-tune per context and configuration).

designs, and the context-specific structure of perturbation responses. Unless stated otherwise, entries are $\overline{\mathrm{BU}}$ values or paired $\overline{\mathrm{BU}}$ differences (median over query-view rows), and intervals are 95% bootstrap intervals with the resampling unit listed in Section C.

D.1 Companion Tables for Main Results

Table 5 expands the held-out identity column of Table 1; Tables 6 and 7 are raw-target companions of the masking and ablation results. In this appendix, MASKED and SEEN denote the masked and measured settings of Table 1, BASE the base scorer, and the forward-then-match arm is the CED effect-only arm.

D.2 Response-Availability Controls

The unit-direction CED comparisons in Figure 3b separate each candidate's own historical response from the overall size of the training atlas. The full-atlas, size-controlled, and response-unseen arms

Table 5: **CED improves an identical base scorer on held-out K562 identities.** 342 queries (684 query-view rows); each row ranks 1,379 candidates. *Top:* median $\overline{\mathrm{BU}}$, MU@B and HvHit@B over rows. Bold marks a unique best value among deployable rankers; the oracle reads measured responses and is a reference. *Bottom:* paired $\overline{\mathrm{BU}}$ differences (first minus second method) with 95% paired identity-cluster bootstrap intervals (both views of a query resampled together, as the locked plan specifies) and win rates.

		MU@B		HvHit@B		
Method	$\overline{\mathrm{BU}}$	B=10	B=20	10	20	50
Random	−0.0053	+0.0072	+0.0085	0	1	2
Goal-blind prior	+0.1807	+0.0344	+0.0421	2	5	10
CED effect only	+0.1245	+0.0279	+0.0256	1	2	5
Text-augmented ranker	+0.3850	**+0.0767**	+0.0642	4	7	10
Base scorer	+0.3551	+0.0754	+0.0629	4	6	9
VCDesign-CED	**+0.3905**	+0.0745	**+0.0673**	4	7	**12**
Measured-response oracle (not deployable)	+0.8142	+0.1458	+0.1247	*9*	*16*	*30*

Paired $\overline{\mathrm{BU}}$ contrast	Median [95% CI]	Mean [95% CI]	Win rate
VCDesign-CED vs. base scorer	**+0.0319** [+0.0282, +0.0378]	+0.0342 [+0.0274, +0.0414]	0.751
vs. text-augmented ranker	+0.0268 [+0.0140, +0.0346]	+0.0254 [+0.0150, +0.0360]	0.582
vs. goal-blind prior	+0.2272 [+0.2020, +0.2521]	+0.2243 [+0.2037, +0.2450]	0.836
Base scorer vs. goal-blind prior	+0.1912 [+0.1638, +0.2175]	+0.1900 [+0.1706, +0.2096]	0.811
vs. text-augmented ranker	−0.0066 [−0.0164, +0.0031]	−0.0088 [−0.0179, +0.0004]	0.465
CED effect only vs. goal-blind prior	+0.0659 [+0.0206, +0.1114]	+0.0845 [+0.0578, +0.1115]	0.576

Table 6: **CED keeps most of its gain for response-unseen candidates.** Paired $\overline{\mathrm{BU}}$ differences of raw-target CED against the K562 base scorer, median over the query-view rows of each held context. Measured: full atlas. Response-unseen: each candidate's responses removed from the atlas in every training context. Retained: response-unseen as a share of measured. Value of response exposure: size-controlled minus response-unseen arm at equal atlas size, with 95% bootstrap intervals over query identities. Pooled: RPE1, HepG2 and Jurkat together, with 95% bootstrap intervals over query genes. A candidate's own response is worth a small fraction of the gain.

Held context (pool)	Measured	Response-unseen	Retained	Value of response exposure [95% CI]
RPE1 (1,273)	+0.0594	+0.0482	81%	+0.0063 [+0.0038, +0.0098]
HepG2 (764)	+0.0524	+0.0568	108%	−0.0025 [−0.0059, +0.0003]
Jurkat (1,450)	+0.0371	+0.0301	81%	+0.0069 [+0.0049, +0.0090]
Pooled	+0.0488 [+0.0456, +0.0526]	+0.0430 [+0.0399, +0.0470]		

share queries, candidate pools, knowledge features, the fixed base scorer, and fusion. The size-controlled and response-unseen arms have equal atlas row counts; the held context contributes no training row in any arm. Their pooled median paired BU@20 gains over the base scorer are +0.0583 [+0.0530, +0.0625], +0.0584 [+0.0541, +0.0623], and +0.0492 [+0.0456, +0.0534], respectively.

Figure 3b plots the paired differences between adjacent arms, computed row by row from the same source rows with the registered query-gene bootstrap (10,000 replicates, seed 20260916); before computing them, the analysis script reproduces every arm interval above exactly. Full atlas minus size-controlled atlas (the cost of a smaller atlas) is +0.0008 [+0.0003, +0.0013] in RPE1, +0.0001 [−0.0002, +0.0004] in HepG2, −0.0003 [−0.0008, +0.0001] in Jurkat and +0.0002 [−0.0001, +0.0005] pooled. Size-controlled minus response-unseen (the value of the candidate's own responses at equal atlas size) is +0.0138 [+0.0100, +0.0169], −0.0029 [−0.0059, +0.0006], +0.0077 [+0.0059, +0.0097] and +0.0078 [+0.0061, +0.0096]; its pooled mean is +0.0154

Table 7: **Goal-conditioned direction carries the gain.** Median paired $\overline{\mathrm{BU}}$ differences between each variant and its reference, with 95% bootstrap intervals. Upper block: terms fused with the base scorer in RPE1 zero-shot (1,004 query-view rows, resampled by query-view row). Lower block: unit-direction against raw-target CED, pooled over RPE1, HepG2 and Jurkat, each held out in turn (2,666 query-view rows, resampled by query gene). Win rate: share of rows with a strictly positive difference. Only the goal-conditioned effect score adds to the base scorer, and unit-direction targets refine it for measured and response-unseen candidates.

Variant	Reference	Setting	Median diff. [95% CI]	Win rate
VCDesign-CED	Base scorer	RPE1 zero-shot	$+0.0196$ $[+0.0146, +0.0242]$	0.629
Base + $\|\dot{\mathbf{e}}_c\|$	Base scorer	RPE1 zero-shot	-0.0016 $[-0.0104, +0.0142]$	0.498
Base + random score	Base scorer	RPE1 zero-shot	-0.0511 $[-0.0578, -0.0452]$	0.252
Unit-direction targets	CED	Pooled, measured	$+0.0048$ $[+0.0039, +0.0056]$	–
Unit-direction targets	CED	Pooled, response-unseen	$+0.0042$ $[+0.0032, +0.0051]$	–

$[+0.0127, +0.0182]$. These are unit-direction BU@20 counterparts of the raw-target value of historical response exposure in Table 6.

In the mixed-pool experiment (Figure 3c), nested identity assignments make 0, 25, 50, 75, or 100% of candidates response-unseen while preserving the query and candidate pools. When CED scores every candidate, gains remain positive at every share. The within-context range across shares is 0.0164 for RPE1, 0.0031 for HepG2, and 0.0055 for Jurkat. These unit-direction BU@20 results accompany the raw-target $\overline{\mathrm{BU}}$ vocabulary analyses in Section D.10; routing measured candidates to retrieval is analysed in Section D.3.

D.3 Routing Measured Candidates to Retrieval

This comparison was preregistered before any routed score was computed. It uses the pools, queries and nested identity hash of Figure 3c. Both arms give each response-unseen candidate its masked CED cosine. They differ only on measured candidates. *CED for all* keeps their unmasked CED cosine; *routed* replaces it by the profile-retrieval cosine. Each arm is fused with the base scorer by the frozen rule of Equation (5). Its row z-scores span every usable identity of the held context (1,483, 881 and 1,696), of which only the pool is graded. A secondary normalization z-scores the two sources separately within each row, over pool candidates only. Nothing is trained, and no weight or calibration is fitted. The primary contrast, routed minus CED for all at a 50% response-unseen share, pooled BU@20, was fixed in advance with its decision rule: routing is supported if the 95% interval lies above zero. At 100% the arms are identical, and they reproduce the masked result exactly; at 0% CED for all reproduces the measured result exactly.

Table 8: **Routing measured candidates to retrieval in mixed pools.** Pooled paired BU@20 gains over the base scorer, and routed minus CED for all, by response-unseen share. Medians with 95% query-gene intervals (976 clusters, 2,666 query-view rows). Frozen fusion unless labelled source-wise.

	Gain over the base scorer		Routed minus CED for all	
Unseen share	CED for all	Routed	Frozen fusion	Source-wise
0%	$+0.058$ $[+0.053, +0.062]$	$+0.099$ $[+0.095, +0.104]$	–	–
25%	$+0.058$ $[+0.054, +0.063]$	$+0.083$ $[+0.078, +0.087]$	$+0.0229$ $[+0.0197, +0.0261]$	$+0.0226$ $[+0.0199, +0.0257]$
50%	$+0.056$ $[+0.052, +0.061]$	$+0.071$ $[+0.067, +0.075]$	$\mathbf{+0.0125}$ $[+0.0100, +0.0150]$	$+0.0123$ $[+0.0096, +0.0150]$
75%	$+0.054$ $[+0.051, +0.059]$	$+0.061$ $[+0.057, +0.065]$	$+0.0055$ $[+0.0038, +0.0072]$	$+0.0091$ $[+0.0071, +0.0115]$
100%	$+0.049$ $[+0.046, +0.053]$	$+0.049$ $[+0.046, +0.053]$	–	–

Routing improves selection and meets the preregistered criterion (Table 8). The full series, by context and for both normalizations, is in Tables 9 and 10. At 50%, the gain is positive in every con-

Table 9: **Routing series by context and normalization.** Median paired BU@20 gain over the base scorer at each response-unseen share. Frozen: the frozen fusion (primary); source-wise: the prespecified secondary normalization. Pooled intervals of the frozen series are in Table 8.

Context	Arm	Fusion	0%	25%	50%	75%	100%
RPE1	CED for all	frozen	+0.0691	+0.0675	+0.0653	+0.0620	+0.0526
	CED for all	source-wise	+0.0771	+0.0757	+0.0727	+0.0710	+0.0601
	Routed	frozen	+0.1179	+0.0888	+0.0731	+0.0634	+0.0526
	Routed	source-wise	+0.1197	+0.1150	+0.0914	+0.0964	+0.0601
HepG2	CED for all	frozen	+0.0672	+0.0703	+0.0692	+0.0701	+0.0698
	CED for all	source-wise	+0.0810	+0.0817	+0.0789	+0.0820	+0.0770
	Routed	frozen	+0.0891	+0.0824	+0.0797	+0.0700	+0.0698
	Routed	source-wise	+0.0910	+0.0855	+0.0823	+0.0785	+0.0770
Jurkat	CED for all	frozen	+0.0434	+0.0445	+0.0434	+0.0428	+0.0390
	CED for all	source-wise	+0.0501	+0.0474	+0.0476	+0.0473	+0.0430
	Routed	frozen	+0.0892	+0.0761	+0.0672	+0.0544	+0.0390
	Routed	source-wise	+0.0898	+0.0742	+0.0652	+0.0570	+0.0430
Pooled	CED for all	frozen	+0.0583	+0.0578	+0.0561	+0.0544	+0.0492
	CED for all	source-wise	+0.0679	+0.0663	+0.0643	+0.0620	+0.0554
	Routed	frozen	+0.0994	+0.0832	+0.0709	+0.0607	+0.0492
	Routed	source-wise	+0.1004	+0.0908	+0.0781	+0.0766	+0.0554

Table 10: **Routed minus CED for all, by context and normalization.** Median paired BU@20 difference with 95% query-gene interval at each response-unseen share. At 0% the contrast is not defined by the protocol; at 100% the arms are identical.

Context	Fusion	25%	50%	75%
RPE1	frozen	+0.0222 [+0.0167, +0.0270]	+0.0070 [+0.0034, +0.0105]	+0.0030 [0.0000, +0.0055]
	source-wise	+0.0341 [+0.0280, +0.0407]	+0.0152 [+0.0090, +0.0201]	+0.0170 [+0.0133, +0.0208]
HepG2	frozen	+0.0114 [+0.0063, +0.0174]	+0.0097 [+0.0047, +0.0137]	+0.0018 [−0.0007, +0.0042]
	source-wise	+0.0093 [+0.0047, +0.0153]	+0.0052 [+0.0018, +0.0092]	+0.0023 [−0.0007, +0.0063]
Jurkat	frozen	+0.0272 [+0.0233, +0.0301]	+0.0188 [+0.0152, +0.0222]	+0.0091 [+0.0069, +0.0115]
	source-wise	+0.0205 [+0.0171, +0.0240]	+0.0133 [+0.0105, +0.0178]	+0.0056 [+0.0027, +0.0089]
Pooled	frozen	+0.0229 [+0.0197, +0.0261]	+0.0125 [+0.0100, +0.0150]	+0.0055 [+0.0038, +0.0072]
	source-wise	+0.0226 [+0.0199, +0.0257]	+0.0123 [+0.0096, +0.0150]	+0.0091 [+0.0071, +0.0115]

text: +0.0070 [+0.0034, +0.0105] in RPE1, +0.0097 [+0.0047, +0.0137] in HepG2 and +0.0188 [+0.0152, +0.0222] in Jurkat. It also holds at the other budgets: +0.0143 BU@10 and +0.0095 BU@50 in median, and +0.25 [+0.20, +0.30] high-value hits among the first 20 in mean. The benefit falls as fewer candidates are measured. At 75%, the RPE1 and HepG2 intervals include zero. At 25% and 50% the two normalizations agree pooled but not per context: source-wise scaling helps RPE1 more, and HepG2 and Jurkat less. At 75% it gives the larger pooled gain.

The fusion dilutes retrieval. With every candidate measured (0% unseen), the routed arm is the base scorer fused with retrieval. It reaches BU@20 0.567, against 0.675 for retrieval alone, a paired difference of −0.1008 [−0.1051, −0.0929]. The difference is negative in every context: RPE1 −0.135, HepG2 −0.110, Jurkat −0.074. The routed arm still exceeds VCDesign-CED on measured candidates by +0.0385 [+0.0357, +0.0421]. The frozen fusion therefore leaves part of retrieval's advantage unused. A rule that ranks measured candidates by retrieval alone would need its own calibration against the CED-scored candidates, and it was not tested. Availability here is a random identity hash; in real libraries, well-studied genes are measured more often.

D.4 CANDIDATES THE BASE SCORER NEVER TRAINED ON

The base scorer was trained on the K562 training identities. These include 68% of the masked candidates in the held contexts (64 to 70% per context), so the masked gain of Section 5.2 mixes candidates the base scorer trained on with candidates it did not. A preregistered analysis regrades the saved masked scores, unchanged, on the candidates outside the K562 training identities. Nothing is refitted or re-normalized. Neither branch trained on these candidates' responses or on the held context. Of the 1,108, 502 are K562 development identities, whose responses selected the checkpoint and the fusion form (Section B); the row "Also outside development" excludes them. The strata were built from identities alone before any score was read. A size-matched control draws three hash subsamples of the trained candidates, each the size of the untrained stratum.

Table 11: **Masked gain on candidates the base scorer never trained on.** VCDesign-CED minus the base scorer, paired $\mathrm{BU@20}$, median with 95% query-gene interval; every query in both views, except the last row, which keeps the 361 query genes outside the K562 training identities. Levels are medians. Size-matched: untrained minus the mean over three equal-size subsamples of trained candidates.

	RPE1	HepG2	Jurkat	Pooled
Untrained candidates	407 of 1,273	273 of 764	428 of 1,450	1,108 of 3,487
Base scorer level	0.257	0.383	0.387	0.349
VCDesign-CED level	0.330	0.379	0.422	0.391
Paired gain	$+0.0479$ $[+0.0422, +0.0551]$	$+0.0234$ $[+0.0099, +0.0345]$	$+0.0273$ $[+0.0224, +0.0314]$	$\mathbf{+0.0332}$ $[+0.0295, +0.0373]$
Minus size-matched	$+0.0035$ $[-0.0042, +0.0138]$	-0.0459 $[-0.0567, -0.0251]$	-0.0079 $[-0.0143, +0.0009]$	-0.0086 $[-0.0152, -0.0026]$
Also outside development	$+0.0359$ $[+0.0299, +0.0428]$	$+0.0009$ $[-0.0174, +0.0187]$	$+0.0327$ $[+0.0282, +0.0387]$	$+0.0313$ $[+0.0265, +0.0356]$
Untrained queries only	$+0.0449$ $[+0.0339, +0.0572]$	$+0.0287$ $[+0.0095, +0.0460]$	$+0.0243$ $[+0.0184, +0.0314]$	$+0.0314$ $[+0.0265, +0.0390]$

The gain persists, positive in every context (Table 11), so the masked gain does not require the base scorer to have trained on the candidate. It is smaller than on size-matched trained candidates, mostly in HepG2. Both rankers find the untrained candidates harder. On them, the base scorer reaches 0.349, against 0.450 to 0.483 on the size-matched subsamples, and VCDesign-CED reaches 0.391, against 0.512 to 0.528. These candidates may also differ in effect size or annotation depth, so the gap cannot be attributed to base-scorer training alone. HepG2 is the weak context. In HepG2, excluding the development identities as well leaves 157 candidates and an unresolved gain, $+0.0009$ $[-0.0174, +0.0187]$. On the untrained stratum at $\mathrm{BU@50}$, HepG2 gives -0.0036 $[-0.0186, +0.0145]$. Restricting the queries to untrained identities leaves the pooled gain at $+0.0314$.

D.5 REPRESENTATION PROBE

A factorial probe on the K562 development split asks how much of the remaining gap lies in the candidate representation and how much in the query representation. One low-rank bilinear scorer (rank 128, 256-dimensional towers, a candidate-prior branch) is trained per cell with one fixed, lighter recipe than the main scorer (60 epochs, learning rate 10^{-4}, seed 20260916) and graded on 726 development query-view rows by median $\overline{\mathrm{BU}}$. The query is either the learned query embedding of the text-augmented ranker or the raw transition under a fixed Gaussian projection. The candidate is either its knowledge features (including text) or its own measured response in the query view, projected like the raw query; the latter is a non-deployable reference. The four cells give 0.1453 and 0.1331 with knowledge features and 0.6345 and 0.6365 with measured responses (learned and raw query, respectively). Replacing knowledge with measured responses therefore adds 0.489 and 0.503, whereas changing the query representation moves $\overline{\mathrm{BU}}$ by at most 0.0121. Both query arms carry the full observed transition, so this probe shows that the learned query encoder loses little, and it bounds what better candidate information could add; it does not show that the query side carries no information. Its recipe underperforms the main scorer (deployable cells 0.133 to 0.145, against 0.344 for the main recipe), so its absolute levels are not comparable with Table 5.

D.6 External-Baseline and Realized-Outcome Companions

The external comparison follows one frozen evaluation contract across RPE1, HepG2, and Jurkat. CellNavi reproduces its released tutorial (accuracy 0.5101, weighted F1 0.4979 against 0.510 and 0.498) and is evaluated only in SEEN because its closed-set head has no class for a masked candidate (Wang et al., 2025). Each context uses one fine-tune from the released pretrained weights at the official recipe (global batch 128, 1,050 steps, step-1000 checkpoint), on cells of the identities the SEEN effect atlas may use; the held context is never used for training. The official adaptation uses about 7,000 classes; the predeclared native-scale variant uses the candidate library itself (1,273, 764, and 1,450 classes), raising per-class exposure about six-fold without changing the trainer or step budget. Both versions learn: the final cross-entropy of the native version is 6.83, 6.16 and 6.86 against chance values of 7.15, 6.64 and 7.28. These results characterize the stated fine-tuning budget, not the method's attainable performance.

GEARS can score MASKED candidates without their measured responses (Roohani et al., 2024). Its official reproduction matches the headline Top-20 differential-gene MSE (0.2456 against 0.2540), but the current package omits seven Norman test conditions. Separately, all six adaptation fits predict a bitwise-identical shift for every candidate. We therefore retain GEARS as a labelled companion and use the internal forward-then-match arm as the informative forward comparison. PDGrapher is not run: only 45% of its PPI nodes lie on the shared gene axis, and its genetic mode requires a mutation vector absent for RPE1 (Gonzalez et al., 2026).

Table 12: **External-method adaptations and information requirements.** The information regime indicates whether a method can score response-unseen MASKED candidates without changing its core model.

Method	Applicable information regime	Implementation checks	Evaluation scope
CellNavi	SEEN with interface adapter; no MASKED arm	Released tutorial matched; one fine-tune per context and class scale	SEEN companion only
GEARS	MASKED and SEEN with interface adapter	Headline MSE close; seven test conditions absent; all six adaptation fits were candidate-invariant	Labelled degenerate companion
PDGrapher	Not directly compatible without a core change	Not run: shared gene axis and mutation-vector requirements unmet	Excluded from score tables

Table 13: **Unified external-baseline evaluation.** Entries are median $\mathrm{BU@20}$ over query-view rows. MASKED is the response-unseen primary regime; SEEN allows a candidate's response in the training contexts. The oracle reads held-context responses and is not deployable. –: incompatible with the regime or companion not reported as a level.

	MASKED	SEEN			
Method	pooled	RPE1	HepG2	Jurkat	pooled
Random	$+0.002$	-0.005	-0.010	-0.003	-0.005
Goal-blind prior	0.170	0.437	0.310	0.108	0.194
GEARS (degenerate companion)	–	-0.199	-0.100	-0.037	-0.078
CellNavi, official class set	–	0.114	-0.098	-0.047	-0.020
CellNavi, native class scale	–	0.279	0.196	-0.116	0.015
CED effect only (forward-then-match)	0.294	0.326	0.427	0.438	0.396
BASE	0.440	0.389	0.413	0.495	0.440
VCDesign-CED	**0.499**	0.462	0.485	0.556	0.508
Profile retrieval	–	0.666	0.658	0.691	**0.675**
Measured-response oracle	*0.951*	*0.964*	*0.937*	*0.938*	*0.951*

Native-scale CellNavi improves over its official adaptation in RPE1 and HepG2 but not Jurkat. Its pooled $\mathrm{BU@20}$ remains below BASE. These rows are descriptive point estimates from one fine-tune

per context and configuration. The query's own class has median rank 618 of 1,273 in RPE1, 238 of 764 in HepG2, and 816 of 1,450 in Jurkat, even when other candidates achieve useful design scores. Identity recovery and outcome-based design therefore measure different orderings.

Table 14: **Paired headline contrasts under the unified contract.** Values are median per-row ΔBU@20 with 95% query-gene bootstrap intervals.

Regime	Contrast	Pooled difference [95% CI]
MASKED	VCDesign-CED − BASE	$+0.0492$ $[+0.0456, +0.0534]$
SEEN	VCDesign-CED − CellNavi, native scale	$+0.3363$ $[+0.3025, +0.3652]$
SEEN	VCDesign-CED − Profile retrieval	-0.1519 $[-0.1602, -0.1424]$

Four-quarter realized-outcome guardrail. PHR@B is the fraction of the selected prefix with independently realized positive endpoint improvement (Section A.1). It is available only when the query and candidate have at least 25 cells in each of four batch-disjoint quarters. RPE1 and Jurkat pass the preregistered minimum of 200 candidates; HepG2 is reported but underpowered.

Table 15: **PHR@20 on the four-way sub-pool.** BASE and VCDesign-CED use MASKED rankings. The oracle is not deployable. HepG2 is below the preregistered 200-candidate threshold and is excluded from the pooled powered contrast.

Context	Candidates	Queries	Powered	BASE	VCDesign-CED	Oracle
RPE1	247	86	yes	0.477	0.503	0.787
HepG2	74	16	no	0.648	0.670	0.798
Jurkat	342	135	yes	0.692	0.703	0.835

Across the powered contexts, the VCDesign-CED minus BASE contrast has mean $+0.0171$ $[+0.0083, +0.0258]$ and median 0.000. PHR@20 changes in steps of 0.05, so most rows tie and the preregistered median rule returns NULL. The oracle contrast is $+0.208$ $[+0.186, +0.232]$, showing that the endpoint itself has headroom.

Prediction and design. Within each held context, held-context prediction quality and effect-only design are rank-correlated over the 18 distinct atlas and input configurations (Spearman 0.31, 0.69 and 0.71 in RPE1, HepG2 and Jurkat; stratified permutation $p < 10^{-4}$, Section C). Pooled over all 57 fits, which include three duplicates, the coefficient is 0.610 (0.645 in SEEN), a descriptive value that treats fits as independent. Within a fixed atlas, prediction and design changes agree in sign for 3 of 15 MASKED variant-context cells (9 of 15 in SEEN): ranks 64 and 128 predict better but design worse, and rank 512 and dropping MAP-KG predict worse but design better, in every held context. Prediction quality detects large distribution shifts but does not select among nearby design estimators.

D.7 Held-Out Evaluation Split: Secondary Endpoints

Table 16 adds the four-quarter realized objective (Section A.1) to the evaluation-split comparison of Table 5. It is computed on the $1{,}053$ candidates and the eligible queries measured in all four batch-disjoint quarters, so its population is a subset of the main table's. VCDesign-CED raises the realized-objective RU@10 over the base scorer by $+0.0201$ (difference of medians) and raises the share of top-10 picks with a positive realized objective from 0.501 to 0.513; the text-augmented ranker is highest among deployable rankers on both quantities. Deeper in the budget the realized-objective difference is slightly negative (-0.0049 at $B = 20$, -0.0053 at $B = 50$), while the share of positive picks still rises at $B = 20$. Paired mean utility of the top 20 rises by a median of $+0.0043$ (mean $+0.0066$); for the top 10 the median paired change is -0.0009 and the mean $+0.0052$. The share of top-10 picks that carry no knowledge modality falls from 0.0035 to 0.0016.

Table 16: **Four-quarter realized objective on the held-out evaluation split.** Median realized-objective RU@10 and the share of top-10 picks with a positive realized objective, computed on the 1,053 candidates and the eligible queries measured in all four quarters.

Method	Realized RU@10	Share with $J > 0$ in top 10
Random	-0.03145	0.269
Goal-blind prior	$+0.01529$	0.374
CED effect only	$+0.01029$	0.383
Text-augmented ranker	$+0.08307$	0.517
Base scorer	$+0.05286$	0.501
VCDesign-CED	$+0.07296$	0.513
Measured-response oracle (not deployable)	$+0.36692$	0.732

D.8 Context Transfer Details

In the RPE1 zero-shot experiment (Table 17, first row), the mean utility of the top 10 rises from $+0.2426$ (frozen base scorer) to $+0.2682$ (VCDesign-CED) and that of the top 20 from $+0.2163$ to $+0.2437$; the measured-response oracle reaches $\overline{\mathrm{BU}}$ $+0.9606$ on the same pool. The frozen base scorer alone already exceeds random ranking by $+0.3085$ $[+0.2811, +0.3309]$ with a win rate of 0.843, so the effect branch improves a ranker that is already useful in RPE1. The split of RPE1 queries by whether their knockdown is a K562 training identity (313 and 189 queries) separates queries only: both groups rank the same 1,483 candidates. The zero-shot experiment uses the frozen K562 map on the K562 gene axis, whereas the leave-one-context-out experiments refit the map on the shared 6,550-gene axis, so the two K562-only values for RPE1 ($+0.0196$ and $+0.0201$) come from different maps on different gene axes.

Leave-one-context-out atlases. Table 17 gives every leave-one-context-out atlas and the zero-shot entry. Which contexts enter the atlas matters. A Jurkat-only atlas gives $+0.0488$ on held RPE1, more than twice the K562-only gain of $+0.0201$. A HepG2-only atlas contributes little ($+0.0057$ on RPE1, $+0.0012$ on Jurkat); HepG2 has 881 usable identities and its median perturbed identity 45 cells (both views, all identities), against 1,696 and 83 for Jurkat. The K562 column is discussed in Section D.12.

Table 17: **Context transfer.** $\overline{\mathrm{BU}}$ differences of VCDesign-CED against the K562 base scorer (median over query-view rows of the held context). Zero-shot: the K562 map on the K562 gene axis, nothing trained on RPE1, 95% bootstrap interval over query-view rows. Leave-one-context-out: the held context contributes no training row, CED is refit on the shared 6,550-gene axis, point estimates. –: the atlas contains the held context; empty: not reported. †Base scorer trained on K562; development-split queries.

Effect atlas	Held RPE1	Held HepG2	Held Jurkat	Held K562†
Zero-shot, K562 gene axis				
K562 map, unchanged	$+0.0196$ $[+0.0146, +0.0242]$			–
Leave one context out, shared 6,550-gene axis				
K562 only	$+0.0201$	$+0.0318$	$+0.0277$	–
Jurkat only	$+0.0488$		–	$+0.0071$
RPE1 only	–		$+0.0259$	-0.0021
HepG2 only	$+0.0057$	–	$+0.0012$	-0.0015
Three other contexts	$\mathbf{+0.0594}$	$\mathbf{+0.0518}$	$\mathbf{+0.0343}$	$\mathbf{+0.0076}$
Pool (usable identities)	1,483	881	1,696	7,825

Within-RPE1 effect atlases. A second RPE1 analysis asks which effect atlas serves RPE1 best when RPE1 responses may enter training. Three CED maps are fit on the 7,094 genes shared by K562 and RPE1, with no context feature: K562 only, RPE1 only, and both pooled. Predictions for candidates inside a map's own training pool are out of fold, so no candidate's own RPE1 response fits its own prediction; because other RPE1 candidates do enter training, this is not a held-context

design. Table 18 shows that pooling the two contexts more than triples the gain of the K562-only map and beats the RPE1-only map, in line with the scaling result of Section 5.4.

Table 18: **Within-RPE1 effect atlases.** Paired $\overline{\mathrm{BU}}$ differences against the frozen K562 base scorer on the RPE1 zero-shot queries and pool (Table 17), with 95% bootstrap intervals over query-view rows. The pooled atlas has a win rate of 0.880.

Effect atlas	Paired difference
K562 only	$+0.0182$ [$+0.0137$, $+0.0235$]
RPE1 only (out of fold)	$+0.0287$ [$+0.0259$, $+0.0320$]
K562 and RPE1 pooled (out of fold)	$+\mathbf{0.0657}$ [$+0.0606$, $+0.0698$]
Pooled minus K562 only	$+0.0366$ [$+0.0332$, $+0.0401$]

D.9 Controlled Masking Details

Masking protocol. A SHA-256 hash of the gene symbol assigns every candidate to one of five folds. For each fold, the response-unseen arm removes the fold's responses from the distillation atlas in every training context, recomputes the response basis $\mathbf{W}$ from the remaining rows, and refits the ridge by the recipe of Section B to predict effects for the fold. Cross-fitting over the five folds makes every candidate in the pool response-unseen. Removing a fold also shrinks the atlas, so the size-controlled arm scores each fold with the next fold's masked predictor: its atlas is reduced by the same amount, but the evaluated candidates stay measured. The measured arm uses the full atlas. All three arms share the knowledge features $\mathbf{k}_c$, the pool, the queries, the utility, the fusion rule and the metric, so the size-controlled and response-unseen arms differ only in whether a candidate's own responses supervised its predicted effect. A held context's pool holds its usable candidates that some training context measured and that carry at least one knowledge modality. All arms also share the K562 base scorer, whose training identities include 68% of the masked candidates (Section D.4), so every masking contrast measures the effect branch.

Per-fold value of historical response exposure. Table 6 cross-fits the five masking folds into one pool per context. Table 19 instead scores each fold on its own masked candidates, which gives pools of 225 to 304 candidates in RPE1 and Jurkat and 138 to 175 in HepG2. Every fold has a positive value of historical response exposure in RPE1 and Jurkat, and HepG2 folds lie between -0.0003 and $+0.0096$. In RPE1, HepG2 and Jurkat the per-fold medians exceed the cross-fitted values; in K562 the per-fold median ($+0.0081$) is below the cross-fitted value ($+0.0097$). Because $\overline{\mathrm{BU}}$ differences depend on pool size (Section D.10), Table 6 uses the cross-fitted values, which rank each context's full pool.

Table 19: **Value of historical response exposure per masking fold.** Size-controlled arm minus response-unseen arm, each fold scored on its own masked candidates. †Base scorer trained on K562; development-split queries.

Held context	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Median
RPE1	$+0.0032$	$+0.0255$	$+0.0189$	$+0.0031$	$+0.0123$	$+0.0123$
HepG2	$+0.0096$	$+0.0001$	-0.0003	$+0.0019$	$+0.0080$	$+0.0019$
Jurkat	$+0.0103$	$+0.0156$	$+0.0056$	$+0.0057$	$+0.0158$	$+0.0103$
K562†	$+0.0081$	$+0.0073$	$+0.0029$	$+0.0148$	$+0.0184$	$+0.0081$

Atlas shrinkage. Masking also removes rows from the atlas. The cost of this shrinkage alone, the measured minus the size-controlled arm, is $+0.0031$ (RPE1), -0.0003 (HepG2), -0.0000 (Jurkat) and $+0.0003$ (K562, development-split queries) on the cross-fitted pools of Table 6 and Table 24. Medians of paired differences are not additive, so measured minus response-unseen need not equal the value of historical response exposure plus shrinkage.

Comparability of the fold predictors. Cross-fitting mixes the five fold predictors' per-candidate columns into one score row, so the fold predictors should shrink alike: select the same ridge penalty,

and give the same share of each prediction to the candidate-specific part rather than the common response. In RPE1, HepG2 and Jurkat the full-atlas predictor and all five fold predictors select the same penalty, and their candidate-specific shares differ by factors of at most 1.041, 1.031 and 1.068 (1.068 also among Jurkat's five fold predictors alone); Jurkat's measured-versus-response-unseen share (81%) therefore mixes fold predictors that shrink slightly differently. In K562, the full-atlas predictor of the measured arm selects a penalty of 10^3 against 10^4 in all five fold predictors, and its candidate-specific share is 1.792 times theirs. K562's measured-versus-response-unseen contrast, including its 4% retained share, therefore compares unlike predictors. With unit-direction targets (Sections 4.1 and E.1), the shares agree within factors of 1.042, 1.046 and 1.028, and response-unseen candidates retain 0.76, 1.04 and 0.90 of the measured $\mathrm{BU@20}$ gain in RPE1, HepG2 and Jurkat (Table 25). The value of historical response exposure does not depend on this comparability: the size-controlled and response-unseen arms are built from the same five fold predictors, permuted by one fold, so any mixing distortion enters both arms alike and cancels in their difference.

Within-context reference. When the held context is also a training context, the measured arm is fit on the evaluated candidate's own response in the very context that grades it, so its value of historical response exposure is an upper bound. Fold medians of this value are then $+0.0218$ (RPE1), $+0.0228$ (HepG2), $+0.0230$ (Jurkat) and $+0.0200$ (K562, development-split queries), against the cross-context values of Table 6. Response-unseen candidates still beat the base scorer in RPE1, HepG2 and Jurkat in this design ($+0.0588$, $+0.0215$, $+0.0279$). The cross-context design avoids this self-supervision and is used in Section 5.3.

Static support. Table 20 splits response-unseen candidates into equal-count terciles of ridge leverage against the full training atlas, with identical strata in both arms. The gain over the base scorer is monotone in all three held contexts and largest where static support is weakest. We interpret leverage here as distinctiveness: a candidate near the centre of the training atlas has a predicted effect close to the common response, which the effect score can hardly separate from its neighbours, whereas a distant candidate has a distinctive predicted effect. Terciles are point estimates; the ordering is the same in all three contexts.

Table 20: **Response-unseen gain by static support.** Paired $\overline{\mathrm{BU}}$ differences against the frozen K562 base scorer (point estimates) within terciles of ridge leverage.

Held context	High support	Medium support	Low support
RPE1	$+0.0119$	$+0.0328$	$+0.0630$
HepG2	$+0.0205$	$+0.0338$	$+0.0487$
Jurkat	-0.0113	$+0.0170$	$+0.0395$

Reconstruction accuracy and design value. Table 21 compares the median cosine between predicted and measured consensus response, $\cos(\hat{\mathbf{e}}_c, \mathbf{r}_c)$, with the design gain retained under masking. Masking costs 14 to 21% of reconstruction accuracy but 0 to 19% of the design gain in RPE1, HepG2 and Jurkat; in HepG2 it costs 16% of accuracy and none of the design gain. Ranking uses the direction of a predicted effect relative to the other candidates, which is more robust to masking than its absolute accuracy.

Table 21: **Reconstruction accuracy and design value under masking.** Median $\cos(\hat{\mathbf{e}}_c, \mathbf{r}_c)$ on the evaluated candidates, its relative drop under masking, and the share of the measured design gain that response-unseen candidates retain (Table 6). [†]Base scorer trained on K562; development-split queries. [‡]Measured and response-unseen predictors differ in penalty (Section D.9). [§]Jurkat's fold predictors differ in candidate-specific share by a factor of 1.068 (Section D.9).

Held context	Measured	Response-unseen	Accuracy drop	Design gain retained
RPE1	0.2981	0.2360	21%	81%
HepG2	0.3738	0.3155	16%	108%
Jurkat	0.2158	0.1847	14%	$81\%^{\S}$
K562[†]	0.2635	0.2309	12%	$4\%^{\ddagger}$

D.10 Mixed Vocabularies and Vocabulary Size

In the mixed vocabularies of Section 5.3, a nested hash of the gene symbol makes 25, 50 or 75% of each masking pool response-unseen, with the same queries; these candidates take their predicted effect from the masking predictor that excluded them, and the rest keep the measured arm's prediction. The CED-for-all arm of Figure 3c applies this design to the unit-direction estimator and BU@20; its routed arm scores the measured candidates by profile retrieval instead (Section D.3). Table 22 is a raw-target, $\overline{\mathrm{BU}}$ companion that analyses promotions at a 50% response-unseen share. A promotion is a candidate that enters the top B of a query-view row under the fused score but not under the base scorer alone; its entrant value in that row is its utility minus the mean utility of the candidates it displaced. Each promoted candidate's entrant values are averaged over its promotions (both views, all queries), and Table 22 summarizes over promoted candidates. Promoted candidates are worth more than the candidates they displace in every context, with mean entrant values of $+0.0416$ to $+0.0629$ at $B = 10$ and $+0.0236$ to $+0.0317$ at $B = 50$. Response-unseen entrants are worth 84 to 107% of measured entrants at $B = 10$; at $B = 50$ this falls to 63% and 77% in RPE1 and Jurkat, so the cost of a response-unseen vocabulary sits in the deeper part of the budget rather than in the first picks.

Table 22: **Entrant value at a 50% response-unseen share.** Promoted: number of distinct candidates promoted at least once. Entrant value: mean over promoted candidates of each candidate's average entrant value, overall and by class. Positive: share of promoted candidates whose average entrant value is positive.

Held context	B	Promoted	Mean	Measured	Response-unseen	Positive
RPE1	10	382	$+0.0616$	$+0.0666$	$+0.0562$	0.688
RPE1	50	987	$+0.0317$	$+0.0390$	$+0.0244$	0.638
HepG2	10	251	$+0.0629$	$+0.0608$	$+0.0650$	0.661
HepG2	50	594	$+0.0316$	$+0.0306$	$+0.0327$	0.621
Jurkat	10	446	$+0.0416$	$+0.0415$	$+0.0418$	0.614
Jurkat	50	1,078	$+0.0236$	$+0.0266$	$+0.0204$	0.612

Figure 5 and Table 23 report the separate raw-target vocabulary-size stress test for the response-unseen class, the measured class and all candidates of the K562 fold, whose queries are development-split identities and whose 7,825 candidates are K562 training and development identities. The three subsamples disagree by more than the effect at 250 and 500 candidates for the response-unseen class and for all candidates, and at 250 to 1,000 for the measured class (vertical ranges in Figure 5); at 6,255 its interval is $[-0.0154, -0.0048]$, and the full measured class (1,570) gives $+0.0095$ $[+0.0066, +0.0120]$. At matched size the response-unseen class gains more than the measured class ($+0.0472$ against $+0.0023$ at 1,000 candidates, $+0.0198$ against $+0.0095$ at 1,570), so the decline with vocabulary size is not a cost of missing responses.

Table 23: **CED gain against vocabulary size (K562 fold, development-split queries).** Paired $\overline{\mathrm{BU}}$ differences against the frozen K562 base scorer, median over three random subsamples per size (the full class at its own size). –: size exceeds the class. *The three subsamples disagree by more than the effect.

Class	250	500	1,000	1,570	3,000	6,255
Response-unseen	$+0.0329^*$	$+0.0416^*$	$+0.0472$	$+0.0198$	$+0.0057$	-0.0103
Measured	$+0.0108^*$	$+0.0024^*$	$+0.0023^*$	$+0.0095$	–	–
All	$+0.0401^*$	$+0.0213^*$	$+0.0217$	$+0.0052$	-0.0023	$+0.0057$

D.11 Atlas Scale: Row- and Identity-Matched Controls

Figure 6 shows an unmasked, raw-target version of the row-matched comparison ($\overline{\mathrm{BU}}$; evaluated candidates keep their own training-context rows). At a matched number of rows, the gain over the RPE1, HepG2 and Jurkat folds, drawn as the blue line, is $+0.01679$ with one training context, $+0.02263$ with two and $+0.02190$ with three; over all four folds, including the K562 fold, it is $+0.01265$, $+0.01730$ and $+0.01675$. With all available rows the all-fold gains are $+0.01388$,

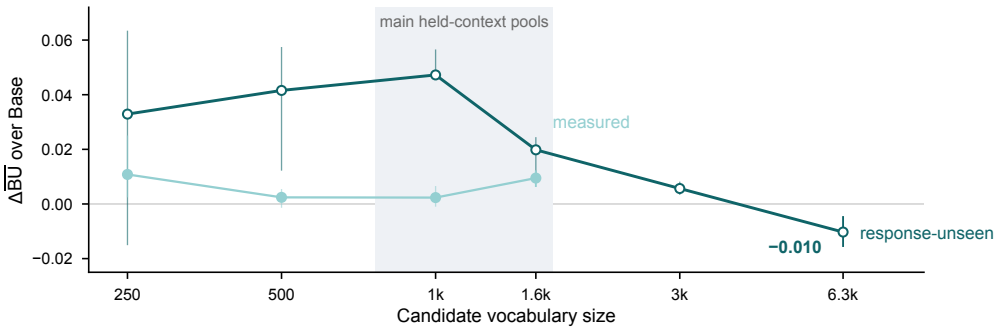

Figure 5: **Vocabulary-scale stress test on the K562 development fold.** Paired $\Delta\overline{\mathrm{BU}}$ over the frozen K562 base scorer for raw-target CED as measured or response-unseen candidate vocabularies grow. Points below the complete class size are medians over three deterministic subsamples and vertical lines span the three values; complete-class points show 95% row-bootstrap intervals. The shaded range contains the candidate-pool sizes of the main held-context experiments. The response-unseen gain is -0.0103 $[-0.0154, -0.0048]$ at 6,255 candidates, so genome-scale vocabularies remain outside the regime established by Figure 3.

$+0.01990$ and $+0.02034$. With all rows the gain rises with every added context; at a matched row count it rises from one context to two and levels off at three. With no context feature, the ridge fit of Equation (3) makes the $3M$ arm equivalent to fitting context-averaged targets at triple weight (the basis and penalty are recomputed for each arm). Against the average single context rather than the best, the spread wins in three of the four folds; the best single context of each fold (Figure 7) was identified after the fact.

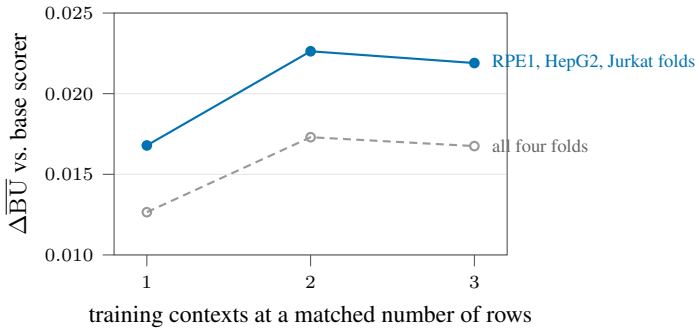

Figure 6: **Row-count matched atlases.** Every arm trains on as many rows as the smallest single-context pool in its fold, split evenly across one, two or three training contexts. Each point is the median paired $\overline{\mathrm{BU}}$ difference against the frozen K562 base scorer, over query-view rows, averaged over held-out folds, training-context combinations and five subsampling seeds; point estimates. Blue: the RPE1, HepG2 and Jurkat folds. Unmasked, raw-target counterpart of the dashed arm of Figure 4a, where the gain instead falls. Grey: all four folds, including the K562 fold, whose queries are development-split identities and whose base scorer was trained on K562. The gain rises from one context to two and levels off at three.

The natural-atlas and row-matched unit-direction experiments of Figure 4a answer the same question on response-unseen candidates. Natural atlases add perturbation rows whenever a context is added, whereas the row-matched arm keeps 840 response-unseen training rows on average. Their pooled gains are $+0.0247$, $+0.0421$, and $+0.0492$ across one, two, and three contexts for natural atlases, versus $+0.0123$, $+0.0070$, and $+0.0012$ for row-matched atlases. The same-identity control of Figure 7 fixes the identity universe instead; it uses raw targets and does not mask, so each evaluated candidate's own training-context rows are in both of its arms (one row in the spread arm, three in the $3M$ arm). Both designs agree that added measurements raise the gain. The context-count

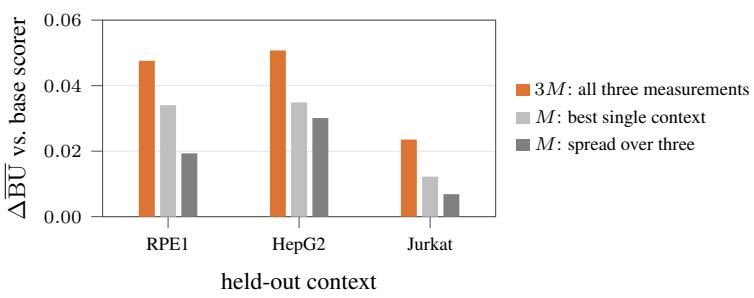

Figure 7: **Same-identity atlas-scale control (raw-target CED).** Median paired $\overline{\mathrm{BU}}$ gain of VCDesign-CED over the K562 base scorer, over query-view rows, using only identities measured in all three training contexts. $3M$: all three measurements of each. M: the same number of rows from the best single context of each fold (identified after the fact), or spread over the three contexts (mean of five row-sampling seeds; the other bars are one fit each). Tripling the measurements gains more than spreading them over contexts. No masking is applied, so an evaluated candidate's own training-context rows are in both arms; the closest response-unseen comparison, which does not fix identities, is Figure 4a. The retrospectively selected best-single arm is a descriptive reference.

comparison at a fixed number of rows depends on the design: under masking with unit-direction targets it falls with each added context, whereas in the unmasked raw-target arms above it rises from one context to two and levels off.

D.12 K562 as a Held-Out Context

K562 behaves differently because the frozen base scorer was trained on K562 responses and is far stronger there: its margin over random is $+0.6382$, against $+0.32$ to $+0.48$ in the other contexts, which leaves the effect branch little room. In every leave-one-context-out, masking, mixed-vocabulary, atlas-scale and vocabulary-size analysis, K562-fold queries are development-split identities and its candidates are K562 training and development identities; the locked evaluation split is used only in Table 1 and Section 5.2. With the three-context atlas the effect branch adds $+0.0076$ on held K562, and single-context atlases give $+0.0071$ (Jurkat), -0.0021 (RPE1) and -0.0015 (HepG2) (Table 17). Under controlled masking both arms sit near zero (Table 24): $+0.0112$ measured and $+0.0005$ response-unseen, and because these two arms use predictors with different penalties (Section D.9), K562 is compared through the value of historical response exposure, $+0.0097$ $[+0.0075, +0.0123]$, of the same order as in RPE1 and Jurkat. In mixed vocabularies the gain moves from $+0.0112$ (all measured) through $+0.0073$, $+0.0051$ and $+0.0024$ to $+0.0005$ (all response-unseen). In the same-identity atlas-scale control the $3M$ atlas gives $+0.00445$, the spread atlas -0.00006 and the best single context (Jurkat) $+0.00078$.

Table 24: **K562 held out under controlled masking (development-split queries).** Columns as in Table 6. †Base scorer trained on K562. ‡Measured and response-unseen predictors differ in penalty (Section D.9).

Held context (pool)	Measured	Response-unseen	Retained	Value of response exposure [95% CI]
K562† (1,430)	$+0.0112$	$+0.0005$	4%‡	$+0.0097$ $[+0.0075, +0.0123]$

D.13 Context-Specific Response Structure

On the 468 candidates usable in all four contexts, 0.5655 of the unit-normalized response is context specific. This residual is reproducible across batch-disjoint replicate views (median cosine $+0.2419$, against -0.0279 for a permuted control), and the median cross-context cosine of one knockdown's response ranges from $+0.1828$ (HepG2 and Jurkat) to $+0.2658$ (RPE1 and HepG2). A cross-fitted, context-specific measured-response oracle exceeds a context-blind one by $+0.0948$ (HepG2), $+0.0807$ (Jurkat), $+0.0788$ (RPE1) and $+0.0485$ (K562) on this pool. This is the open

direction named in Section F; the context-conditioned effect map cannot learn it from three training contexts (Section E).

E EFFECT-MODEL AND UTILITY VARIANTS

This appendix defines each variant behind Table 7 and Section 5.4, and it reports the variants that the main text only summarizes. Table 26 collects the numbers not shown in Table 7. Unless stated otherwise, a variant changes exactly one component of the method. The frozen K562 base scorer, candidate knowledge, pools, queries, fusion rule, utility and metric stay fixed. Effect-model variants are evaluated in the pooled held-context setting of Section 5.3, where RPE1, HepG2 and Jurkat are held out in turn. There each value is a median paired difference over 1,333 queries (2,666 query-view rows) from 976 query genes, with a 95% interval over query-gene clusters. Training-target and utility variants were compared on the development split during method design. The locked K562 result uses raw-target CED; the held-context closure and Figure 3 use the stronger unit-direction refinement.

E.1 EFFECT TARGET AND SCORE FORM

Unit-direction targets. Unit-direction targets give a small refinement that is positive in every held context. The ridge of Equation (3) is refit with each training target replaced by its direction $\mathbf{r}_c/\|\mathbf{r}_c\|$. The response basis $\mathbf{W}$ is rebuilt from the unit targets, and the penalty is re-selected by the same five-fold identity cross-validation. The effect score is the cosine of Equation (4). Fused with the base scorer, unit targets improve on raw targets by $+0.0048$ $[+0.0039, +0.0056]$ for measured and $+0.0042$ $[+0.0032, +0.0051]$ for response-unseen candidates. Per context (measured, then response-unseen), the improvement is $+0.0045$ and $+0.0022$ in RPE1, $+0.0034$ and $+0.0051$ in HepG2, and $+0.0053$ and $+0.0049$ in Jurkat, so it is positive in all six cells. Against the base scorer, unit-direction CED adds $+0.0470$ $\overline{\text{BU}}$ for pooled response-unseen candidates. The refinement holds on the masking pools of 764 to 1,450 candidates. It does not carry over to the 6,255-candidate K562 response-unseen vocabulary of Figure 5. There raw-target CED already gives -0.0103 $[-0.0154, -0.0048]$, and unit-direction CED gives -0.0250.

Keep predicted norm. The predicted norm acts as a context-dependent candidate prior, so VCDesign-CED uses directional cosine for its effect branch. With unit-direction targets, the score $\|\hat{\mathbf{e}}_c\| \cos(\mathbf{d}_q, \hat{\mathbf{e}}_c)$ keeps the predicted norm as a per-candidate factor and changes nothing else. It helps measured candidates ($+0.0142$ $[+0.0088, +0.0181]$) but not response-unseen ones (-0.0004 $[-0.0043, +0.0044]$). The norm by itself ignores the query. Fused with the base scorer, the unit-target norm alone changes $\overline{\text{BU}}$ by -0.0116 $[-0.0180, -0.0062]$ for response-unseen candidates, whereas the goal-conditioned unit-direction score adds $+0.0470$. Two scores that carry the norm do gain on pooled rows against raw-target CED. Keeping the norm with unit targets gives $+0.0215$ $[+0.0159, +0.0261]$ for measured and $+0.0069$ $[+0.0023, +0.0117]$ for response-unseen candidates. Scoring by the raw-target dot product $\langle \mathbf{d}_q, \hat{\mathbf{e}}_c \rangle$ gives $+0.0393$ $[+0.0322, +0.0451]$ and $+0.0295$ $[+0.0240, +0.0353]$. A post hoc per-context control shows where these gains come from. Fused with the base scorer, the goal-blind raw predicted norm alone gives $+0.0434$, $+0.0679$ and -0.0790 for response-unseen candidates in RPE1, HepG2 and Jurkat. CED gives $+0.0482$, $+0.0568$ and $+0.0301$ in the same cells (Table 6). The norm is therefore strong in RPE1 and HepG2 and harmful in Jurkat, whereas the goal-conditioned cosine gains in every held context.

Supervision restricted to measurable identities. Breadth of supervision matters more than target cleanliness. This variant restricts the ridge's training targets to K562 training identities whose two replicate views agree, using the eligibility rule of Section A. The targets are then cleaner but come from fewer identities. On the development split, the effect score alone reaches $\overline{\text{BU}}$ $+0.0691$, against $+0.1377$ when all training identities supervise the ridge. The typical K562 target is noisy (median cross-view cosine 0.0494 over training identities). Even so, a ridge fit on all of them transfers better than one fit on the reliable subset.

E.2 EFFECT MAP

The linear ridge matches or beats every richer effect map we tested. The first three variants below are compared with the unit-direction ridge, the stronger linear reference, on pooled response-unseen candidates. The context-conditioned map is compared with the raw-target context-blind map in the four leave-one-context-out folds.

Reliability-weighted ridge. Reweighting the ridge by axis reliability has no effect, because reliability is nearly uniform across axes. The variant reweights the ridge penalty per axis of the response basis by the cross-view reliability of that axis, estimated on the training contexts, so that reproducible axes are shrunk less. The change is -0.0001 $[-0.0002, +0.0000]$ $\overline{\mathrm{BU}}$. Inside the rank-256 basis, the median cross-view correlation of an axis is 0.49 to 0.50 for every held context. The top 10% of axes hold 12% of the total reliability, close to the 10% that uniform reliability would give. With reliability this flat, isotropic ridge shrinkage is already close to the reliability-aware solution.

Nonlinear effect memory. A nonlinear effect memory scores below the ridge and matches it only when anchored to it. The model replaces the linear map by a nonlinear predictor. It routes each candidate's knowledge features to a sparse combination of learned response programs. It also adds responses transported from similar training candidates, which it retrieves from a memory. A candidate never retrieves itself, and response-unseen candidates are removed from the memory as well as from the training rows. Trained end to end, the model changes $\overline{\mathrm{BU}}$ by -0.0253 $[-0.0282, -0.0228]$ (-0.0233 $[-0.0257, -0.0208]$ BU@20). The anchored version predicts a sparse correction that is added to the unit-direction ridge, so it starts exactly at the ridge. It matches the ridge, with $+0.0002$ $[-0.0007, +0.0012]$ $\overline{\mathrm{BU}}$ ($+0.0000$ $[-0.0010, +0.0011]$ BU@20). Anchored versions that keep only the parametric program branch or only the memory branch give -0.0016 and -0.0053 BU@20.

Context-invariant adapter. A context-invariant adapter leaves design unchanged, because little of the ridge's residual is shared across contexts. The adapter corrects the unit-direction ridge's predicted direction within the tangent space of the unit sphere. Its correction is scaled by a learned scalar initialized at zero, so before training its predictions equal those of the ridge exactly. It is trained either by pooled empirical risk over the training contexts or with an added invariance penalty. The penalty acts on the worst-context excess risk and on the variance of excess risk across contexts. The epoch is selected by leaving one training context out. The adapter gives $+0.0008$ $[-0.0003, +0.0019]$ BU@20 with pooled risk and -0.0012 $[-0.0019, -0.0003]$ with the invariance penalty. Only 0.315, 0.322 and 0.287 of the ridge's reproducible residual is shared across the training contexts of RPE1, HepG2 and Jurkat. A correction that must hold across contexts therefore has little to learn.

Context-conditioned effect map. Three training contexts do not identify how responses depend on context. The context-specific structure itself is large and reproducible (Section D.13). This variant adds a low-rank modulation of the effect map by a context representation built from control cells. It is trained on three contexts and evaluated on the fourth. In all four leave-one-context-out folds, the cross-validated penalty shrinks the modulation away, and the difference from the context-blind map is zero to four decimals. A positive control explains this null. We planted a context residual of the model's own form, at a share of 0.25, into synthetic responses, with the real Jurkat control cells as the held context. The model recovers it only partly ($+0.3885$), and the best design value anywhere on the penalty grid is $+0.00001$.

E.3 RESPONSE BASIS, MODALITY, RANK AND ATLAS SIZE

The ridge depends mainly on STRING and on atlas size, and little on the response basis or the rank. Each variant in Table 25 changes one input of the unit-direction ridge and is scored on pooled response-unseen candidates, fused with the base scorer. No alternative response basis improves on the randomized-SVD basis of Section 4.1 by a margin that matters. The reason is that, out of fold, only 0.012 to 0.018 of the centred candidate-specific response signal is predictable from the knowledge features in any basis. STRING carries most of the transferable signal. Dropping MAP-KG changes BU@20 by -0.0013, and dropping STRING changes it by -0.0211. The rank matters little: ranks 64 to 512 move BU@20 by at most 0.0023 from rank 256. Atlas size, rather than context

count, tracks the gain. Natural atlases gain more with each added training context. At a matched row count and with these targets, however, added contexts lower the gain. Transfer is positive in 20 of the 21 combinations of held context and source atlas (unit-direction targets, BU@20). The source atlases are every non-empty subset of the held context's three training contexts, and the one null is HepG2 alone predicting Jurkat. Response-unseen candidates retain 0.76, 1.04 and 0.90 of the measured-candidate gain in RPE1, HepG2 and Jurkat.

Table 25: **Sensitivity of the unit-direction ridge.** Median paired BU@20 differences on pooled response-unseen candidates (RPE1, HepG2 and Jurkat held out in turn), fused with the base scorer, with 95% intervals over query-gene clusters where recorded. Atlas arms average per query over the atlases in each class; row-matched arms subsample every atlas to the smallest single-context atlas. The transfer row counts, per held context, the source atlases formed by every non-empty subset of its three training contexts (21 cells, each against the base scorer). Retention is the response-unseen over the measured BU@20 gain against the base scorer. STRING and atlas size move the gain; the response basis and the rank barely do.

Variant	Reference	Result
Shared-covariance basis	Unit-direction ridge	-0.0050 $[-0.0072, -0.0034]$
Predictive basis	Unit-direction ridge	$+0.0009$ $[+0.0006, +0.0013]$
Shared and predictive basis combined	Unit-direction ridge	$+0.0013$ $[-0.0003, +0.0030]$
Drop MAP-KG (STRING only)	Unit-direction ridge	-0.0013 $[-0.0026, -0.0002]$
Drop STRING (MAP-KG only)	Unit-direction ridge	-0.0211 $[-0.0235, -0.0177]$
Rank 64 / 128 / 512	Rank 256	-0.0023 / -0.0006 / $+0.0008$
Natural atlas, 1 training context	Base scorer	$+0.0247$ $[+0.0212, +0.0273]$
Natural atlas, 2 training contexts	Base scorer	$+0.0421$ $[+0.0378, +0.0457]$
Natural atlas, 3 training contexts	Base scorer	$+0.0492$ $[+0.0456, +0.0534]$
Row-matched, 1 training context	Base scorer	$+0.0123$ $[+0.0089, +0.0154]$
Row-matched, 2 training contexts	Base scorer	$+0.0070$ $[+0.0046, +0.0095]$
Row-matched, 3 training contexts	Base scorer	$+0.0012$ $[-0.0011, +0.0036]$
Transfer, held context $\times$ source subset	Base scorer	positive in 20 of 21
Retention, RPE1 / HepG2 / Jurkat	Measured candidates	0.76 / 1.04 / 0.90
Win rate, RPE1 / HepG2 / Jurkat	Base scorer	0.805 / 0.877 / 0.740

Table 26: **Effect-model and utility variants not shown in Table 7.** Each row gives the variant and, in parentheses, its reference. Held-context rows are median paired differences with 95% intervals over query-gene clusters, pooled over 1,333 queries (2,666 query-view rows) unless a context is named; rows without an interval are point estimates. Development-split rows are median paired differences, except rows written "a vs b", which give the levels of the variant and its reference. Pairs a / b give two regimes or two grading directions as labelled in the text. Measured resp.: rankings built from measured responses, nothing trained. Real. obj.: graded by the four-quarter realized objective. Richer effect maps and alternative utilities offer no consistent design advantage.

Variant (reference)	Setting	Metric	Result
Score composition and effect target			
Effect score alone (base scorer)	Pooled, measured	$\overline{\mathrm{BU}}$	-0.0272 $[-0.0439, -0.0116]$
Effect score alone (base scorer)	Pooled, unseen	$\overline{\mathrm{BU}}$	-0.1003 $[-0.1137, -0.0831]$
Goal-blind norm, unit targets (base)	Pooled, unseen	$\overline{\mathrm{BU}}$	-0.0116 $[-0.0180, -0.0062]$
Unit-direction CED (base scorer)	Pooled, unseen	$\overline{\mathrm{BU}}$	$+0.0470$
Unit targets, effect only (raw targets)	Pooled, measured	$\overline{\mathrm{BU}}$	$+0.0396$ $[+0.0362, +0.0433]$
Unit targets, effect only (raw targets)	Pooled, unseen	$\overline{\mathrm{BU}}$	$+0.0262$ $[+0.0236, +0.0305]$
Keep predicted norm (unit targets)	Pooled, measured	$\overline{\mathrm{BU}}$	$+0.0142$ $[+0.0088, +0.0181]$
Keep predicted norm (unit targets)	Pooled, unseen	$\overline{\mathrm{BU}}$	-0.0004 $[-0.0043, +0.0044]$
Keep norm, unit targets (raw targets)	Pooled, measured	$\overline{\mathrm{BU}}$	$+0.0215$ $[+0.0159, +0.0261]$
Keep norm, unit targets (raw targets)	Pooled, unseen	$\overline{\mathrm{BU}}$	$+0.0069$ $[+0.0023, +0.0117]$
Raw-target dot product (raw targets)	Pooled, measured	$\overline{\mathrm{BU}}$	$+0.0393$ $[+0.0322, +0.0451]$
Raw-target dot product (raw targets)	Pooled, unseen	$\overline{\mathrm{BU}}$	$+0.0295$ $[+0.0240, +0.0353]$
Raw norm alone, post hoc (base)	RPE1, unseen	$\overline{\mathrm{BU}}$	$+0.0434$
Raw norm alone, post hoc (base)	HepG2, unseen	$\overline{\mathrm{BU}}$	$+0.0679$
Raw norm alone, post hoc (base)	Jurkat, unseen	$\overline{\mathrm{BU}}$	-0.0790
Unit-direction targets (raw targets)	RPE1	$\overline{\mathrm{BU}}$	$+0.0045$ / $+0.0022$
Unit-direction targets (raw targets)	HepG2	$\overline{\mathrm{BU}}$	$+0.0034$ / $+0.0051$
Unit-direction targets (raw targets)	Jurkat	$\overline{\mathrm{BU}}$	$+0.0053$ / $+0.0049$
Unit-direction CED (base scorer)	K562, 6,255 unseen	$\overline{\mathrm{BU}}$	-0.0250
Raw-target CED (base scorer)	K562, 6,255 unseen	$\overline{\mathrm{BU}}$	-0.0103 $[-0.0154, -0.0048]$
Effect map (unit-direction ridge; pooled, response-unseen)			
Reliability-weighted ridge	Pooled	$\overline{\mathrm{BU}}$	-0.0001 $[-0.0002, +0.0000]$
Nonlinear effect memory	Pooled	$\overline{\mathrm{BU}}$	-0.0253 $[-0.0282, -0.0228]$
Nonlinear effect memory	Pooled	BU@20	-0.0233 $[-0.0257, -0.0208]$
Nonlinear memory, anchored to ridge	Pooled	$\overline{\mathrm{BU}}$	$+0.0002$ $[-0.0007, +0.0012]$
Nonlinear memory, anchored to ridge	Pooled	BU@20	$+0.0000$ $[-0.0010, +0.0011]$
Memory, parametric branch, anchored	Pooled	BU@20	-0.0016
Memory, memory branch, anchored	Pooled	BU@20	-0.0053
Invariant adapter, pooled risk	Pooled	BU@20	$+0.0008$ $[-0.0003, +0.0019]$
Invariant adapter, invariance penalty	Pooled	BU@20	-0.0012 $[-0.0019, -0.0003]$
Context-conditioned map (raw-target context-blind map; four leave-one-context-out folds)			
Context-conditioned map (context-blind)	Four held contexts	$\overline{\mathrm{BU}}$	0.0000 in every fold
Planted residual, share 0.25	Jurkat held	recovery	$+0.3885$
Planted residual: best design value	Jurkat held	$\overline{\mathrm{BU}}$	$+0.00001$
Development split			
Measurable-identity targets (all)	Effect score alone	$\overline{\mathrm{BU}}$	$+0.0691$ vs $+0.1377$
Latent endpoint utility (cosine)	Measured resp.	$\overline{\mathrm{BU}}$	-0.1222
Attenuation-aware ranking (direction)	Measured resp.	real. obj. $\overline{\mathrm{BU}}$	-0.1778 / -0.2662
Kernel distributional transition	Measured resp.	real. obj. $\overline{\mathrm{BU}}$	-0.0202 / -0.0133
Distributional tie-breaking (mean only)	Measured resp.	real. obj. $\overline{\mathrm{BU}}$	$+0.000018$

E.4 Alternative Utilities and Response Representations

Utility and score form. The directional utility of Section 3.2 was compared with four alternatives to the utility or response representation. All four rank candidates from measured responses on development-split queries restricted to identities measured in all four batch-disjoint quarters; no model is trained and no model score enters any value. The latent endpoint value replaces the direction cosine as the utility, so a ranking formed on one quarter pair is graded by that quantity on the other pair (Section A.1). The attenuation-aware ranking scores a candidate by realized $J_{qc}(a^*)$ rather than by direction and, like the two distributional variants, is graded by the realized objective rather than U. The kernel (RBF) distributional transition replaces each pseudobulk transition by a kernel mean embedding of the identity's single-cell population in a frozen low-dimensional projection of the gene axis. The tie-breaking variant keeps the pseudobulk ordering and reorders only within a pre-declared score tolerance, using the directly measured distributional score to bound what tie-breaking could contribute. Intervals for these contrasts resample query identities with both grading directions averaged and are reported on the mean (Section C).

The alternatives lower or change little under the available measurements, so we use the signed direction cosine of pseudobulk transitions as the benchmark utility and effect score. A latent endpoint utility changes $\overline{\mathrm{BU}}$ by a median of -0.1222 against the cosine, and 80.8% of its variance does not depend on the query. An attenuation-aware objective that also selects per-candidate knockdown strength changes realized-objective $\overline{\mathrm{BU}}$ by medians of -0.1778 and -0.2662 in the two grading directions. Kernel distributional transitions change it by -0.0202 and -0.0133; using distributional scores only to break mean-ordering ties changes it by $+0.000018$ (mean -0.000118) in the first direction. These comparisons diagnose measurement resolution; they do not establish directional cosine as a universal biological objective.

F Scope and Limitations

Our evidence concerns single-gene CRISPRi in four cell lines with transcriptomic readouts. Evaluation is restricted to queries whose transitions reproduce across replicate views, representing 23.5 to 36.8% of usable identities. This restriction applies to queries: on the evaluation split and in RPE1 zero-shot, ineligible identities remain in the candidate pool (Section A). Directional $\mathrm{BU}@B$ is the primary endpoint because the measurements support reproducible directional evaluation. Realized-outcome analyses remain mixed: PHR@20 improves in mean but not median, and the K562 realized-objective gain is $+0.020$ at $B = 10$ but slightly negative at larger budgets (Sections D.6 and D.7). Directional ranking gains are therefore not yet established as realized endpoint improvements.

The informative response-unseen comparisons rely on internal baselines; the published-method runs did not provide an informative masked-candidate comparator under the implemented protocol (Section D.6). Random gene folds leave most complex partners in the atlas, and STRING accounts for most of the observed knowledge-feature contribution, so transfer may partly exploit close biological neighbors. Module-level masking would test this possibility. In the K562 stress test, whose context trained the base scorer, the candidate library contains $6{,}255$ identities absent from the other screens. The gain declines from $+0.047$ at $1{,}000$ candidates to $+0.006$ at $3{,}000$ and becomes negative at full size: -0.0103 $[-0.0154, -0.0048]$ $\overline{\mathrm{BU}}$ with raw targets and -0.0250 with unit targets (Figure 5). Preserving retrieval's advantage when combining it with CED, extending reliable ranking to genome-scale libraries, and learning context-aware effect maps remain directions for future work (Sections D.3 and D.13).