INVERSEPERTURBBENCH: BENCHMARKING THE IDENTITY GAP IN PREDICTED PERTURBATION RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Inverse perturbation retrieval identifies the intervention behind a measured transcriptional response by ranking candidate interventions, and its success depends on separating candidates rather than on fitting responses on average. We present INVERSEPERTURBBENCH, a benchmark that measures how much intervention identity survives in a candidate library, under matched queries, matched candidate sets, and a single scoring rule. Across 41 external query actions and 1,087 candidates, measured responses recover 87.80% of interventions at rank 10, against 19.51% for an archived GEARS predictor. We call this shortfall the *identity gap*. It is not a generalization failure: on the 28 actions the predictor was fitted on, recovery is 85.7% versus 28.6%, and every predicted top-ten hit falls inside that fitted stratum. It is also not an artefact of target-gene readouts, since a target-excluded comparison retains a 16.98-point gap. Predicted first ranks concentrate on a shared translation-related response rather than on candidate-specific detail, yet removing that concentration alone does not recover accuracy. We therefore develop CARI, which rescores a frozen candidate library against the identification objective without retraining any forward model and improves genetic retrieval from 15.17% to 18.83% and drug retrieval from 24.4% to 33.3%.

1 INTRODUCTION

High-throughput perturbation screens and single-cell RNA sequencing (scRNA-seq) allow systematic measurement of cellular responses to genetic and chemical interventions (Dixit et al., 2016; Adamson et al., 2016; Datlinger et al., 2017; Srivatsan et al., 2020; Replogle et al., 2022). Dedicated predictors learn to generate unmeasured perturbation responses (Lotfollahi et al., 2019; Hetzel et al., 2022; Lotfollahi et al., 2023; Bunne et al., 2023; Roohani et al., 2024; Piran et al., 2024), while single-cell foundation models learn representations from large transcriptomic collections (Theodoris et al., 2023; Rosen et al., 2023; Cui et al., 2024; Hao et al., 2024). These advances create opportunities for computational intervention screening across a large space of genes, compounds, and biological contexts (Norman et al., 2019; McFaline-Figueroa et al., 2024; Bunne et al., 2024).

Intervention screening asks which perturbations produce a response matching a target. A direct approach uses a forward model to generate candidate responses and ranks them by their similarity to that target (Figure 1). We study the experimentally verifiable version of this, *inverse perturbation retrieval*: recovering the applied intervention from a measured response and a predefined candidate set. The *identity gap* is the difference in intervention recovery between the measured and predicted responses available for the same queries and candidates.

Average prediction error does not measure this gap. A response contains a large component shared across perturbations of the same cell type and a small component that says which perturbation was applied. A predictor can capture the shared component, drive its mean squared error toward zero, and still order the candidates wrongly, because retrieval depends on error relative to the separation between candidates: a top-1 decision survives when the candidate margin exceeds twice the worst-case prediction error (Appendix B), a comparison that average error does not make.

Existing evaluations score each model's predictions against observed responses (Wenteler et al., 2025), and PerturBench adds rank metrics that measure discrimination between interventions (Wu

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.

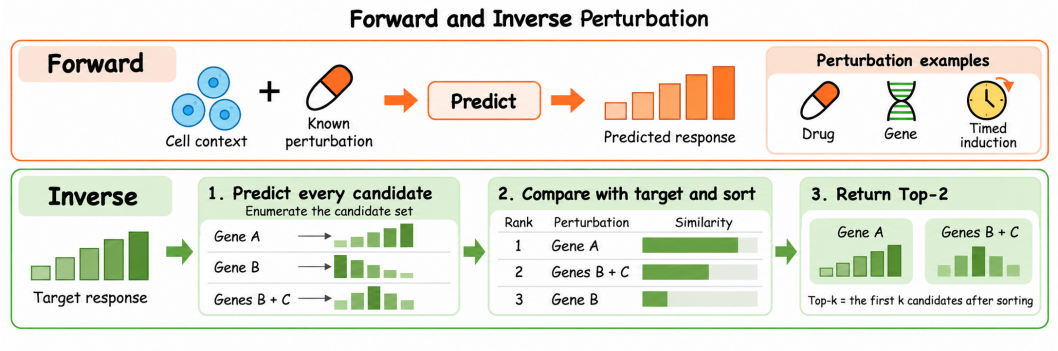


Figure 1: **Forward and inverse perturbation.** Top: predict the response to a known intervention. Bottom: rank predicted candidate responses against a target and return the top-k interventions. Profiles and rankings are schematic.

et al., 2025). We instead fix the queries and the candidate universe and ask how much identity any library supports, measured or predicted, so that the identity gap is read off directly rather than inferred across studies. Measured libraries require profiling every candidate, so predicted libraries are the practical option and the measured library is the reference. Section 4 measures the gap and tests whether it persists on actions a predictor was fitted on, Section 5 asks what rescoring recovers the lost identities, and Section 6 locates the identity signal in the readout and in the candidates. Our contributions are:

- **A benchmark for intervention identity.** INVERSEPERTURBBENCH fixes queries, candidate sets, and a single evaluator, and records candidate provenance, information use, readout, and uncertainty unit for every instance. It reports each predicted library beside a measured reference, a fitted-action stratum, and a target-excluded readout, so that a loss of identity can be attributed rather than only observed. Protocol diagnostics quantify how candidate indices and all-query scoring inflate recovery, and hit transitions evaluate model updates that raise the mean while losing established hits.
- **CARI, a method that closes part of the gap.** Rescoring a frozen candidate library against the identification objective improves genetic retrieval from 15.17% to 18.83% and drug retrieval from 24.4% to 33.3% without retraining a forward model. Both variants fit and select on training or source-cell data only.
- **An account of where the identity signal lives.** Target-panel ablations, first-rank concentration, and annotation agreement of errors identify the readouts and candidate preferences that drive retrieval. Measured errors are enriched for functional neighbours in the same complex or pathway, while predicted first ranks concentrate on a shared translation-related response.

2 RELATED WORK

Perturbation prediction and evaluation. Single-cell screens and harmonized collections such as scPerturb support the development of response predictors (Peidli et al., 2024). These include compositional autoencoders, optimal-transport models, disentangled representations, and graph-based predictors (Lotfollahi et al., 2023; Bunne et al., 2023; Piran et al., 2024; Roohani et al., 2024). Several evaluations find that simple baselines compete strongly with deep predictors (Ahlmann-Eltze et al., 2025; Csendes et al., 2025; Kernfeld et al., 2025), and assess the transferability of pretrained single-cell representations (Kedzierska et al., 2025; Wenteler et al., 2025). Systema separates perturbation-specific effects from systematic variation (Viñas Torné et al., 2026). PerturBench ranks predicted responses against each observed response and also evaluates the transposed direction (Wu et al., 2025); Cell-Eval measures perturbation discrimination with panel-wide target-gene exclusion (Arc Institute, 2026). We adopt these ranking and target-exclusion strategies and add a measured-library reference and retrieval-aware scoring under the same candidates and evaluator.

Signature retrieval and inverse design. The Connectivity Map matches query signatures to measured perturbation profiles (Lamb et al., 2006; Subramanian et al., 2017), supporting drug mode-of-action analysis (Iorio et al., 2010). Inverse-design methods extend intervention selection to desired cell states: Mogrify identifies reprogramming factors, CellOracle simulates transcription-factor perturbations, PAIRING matches latent effects to desired transitions, and PDGrapher predicts target-gene sets (Rackham et al., 2016; Kamimoto et al., 2023; Han et al., 2025; Gonzalez et al., 2026). U-Pert models transcriptomic and population outcomes (Peng et al., 2026). Our endpoint is exact recovery of a registered intervention from its measured response.

Retrieval calibration and reliable updates. Candidate popularity is related to hubness in high-dimensional retrieval (Radovanović et al., 2010); cross-lingual word retrieval addresses it by correcting for candidate neighborhoods (Conneau et al., 2018). Our drug calibration learns a popularity correction from source-cell data. The genetic retrieval method combines response similarity with a query encoder trained using a listwise objective (Cao et al., 2007). We also report hit transitions, following the concern with negative flips in model updates (Bansal et al., 2019; Yan et al., 2021).

3 INVERSEPERTURBBENCH

INVERSEPERTURBBENCH turns a perturbation screen into a retrieval task: a measured response is the query, a set of candidate actions supplies responses to compare against it, and an evaluator returns a ranking. Figure 2 summarizes the task taxonomy and instances evaluated in this paper.

3.1 QUERIES, CANDIDATES, AND EVALUATION PROTOCOLS

Each query is a measured response with its biological context, such as cell line, drug environment, dose, and replicate. An action can be a gene, a complete gene combination, a compound, or a drug-dose regimen; for McFaline, the perturbed gene is the action and the drug treatment is context. We record an evaluation instance as

$$\mathcal{C} = (\mathcal{Q}, \{\mathcal{A}_q\}_{q\in\mathcal{Q}}, \pi, \mathcal{I}, T, G, S, U), \tag{1}$$

where $\mathcal{Q}$ denotes queries; $\mathcal{A}_q$ the candidate set; π candidate provenance; $\mathcal{I}$ the information used for fitting, selection, and inference; T the expression transformation; G the gene readout; S the data separation; and U the uncertainty unit. The true action $a_q^\star$ and guide identifiers are the evaluator's join keys and never enter scoring. The experiments vary one component at a time: π in Section 4.1, the score in Section 5, and T and G in Section 6.

A measured candidate library takes responses from an external screen, another context, or separate measurements in the query context; a predicted library generates them with a forward model. Matched comparisons fix queries, candidate identities, and the evaluator, so the difference between the two libraries is the identity gap.

Candidate and split policy. Candidate responses never include the query profile, results are reported per split, and fitting and selection use training or source data only; Appendix C shows how violating each rule inflates recovery.

3.2 EXPRESSION SPACE AND BIOLOGICAL READOUTS

For count vector x_i and normalization panel B, we use

$$T_B(x_i)_g = \log\left(1 + 10^4 \frac{x_{ig}}{\sum_{h\in B} x_{ih}}\right), \qquad \delta_{a,c} = \frac{1}{|P_{a,c}|}\sum_{i\in P_{a,c}} T_B(x_i) - \sum_{j\in N_c} w_j T_B(x_j), \tag{2}$$

where $P_{a,c}$ contains perturbed cells, N_c contains matched controls, and the nonnegative control weights sum to one. Scoring uses coordinates $g \in G$, and the normalization panel B and readout panel G can differ.

Genetic perturbations can leave a proximal signature in the targeted transcripts, so we compare all-gene, target-panel-only, and *global trans-only* readouts. The global trans-only policy removes the union of mapped candidate target genes, defined by the candidate registry, from expression inputs before fitting and score selection, so retrieval measures identity information outside the target panel.

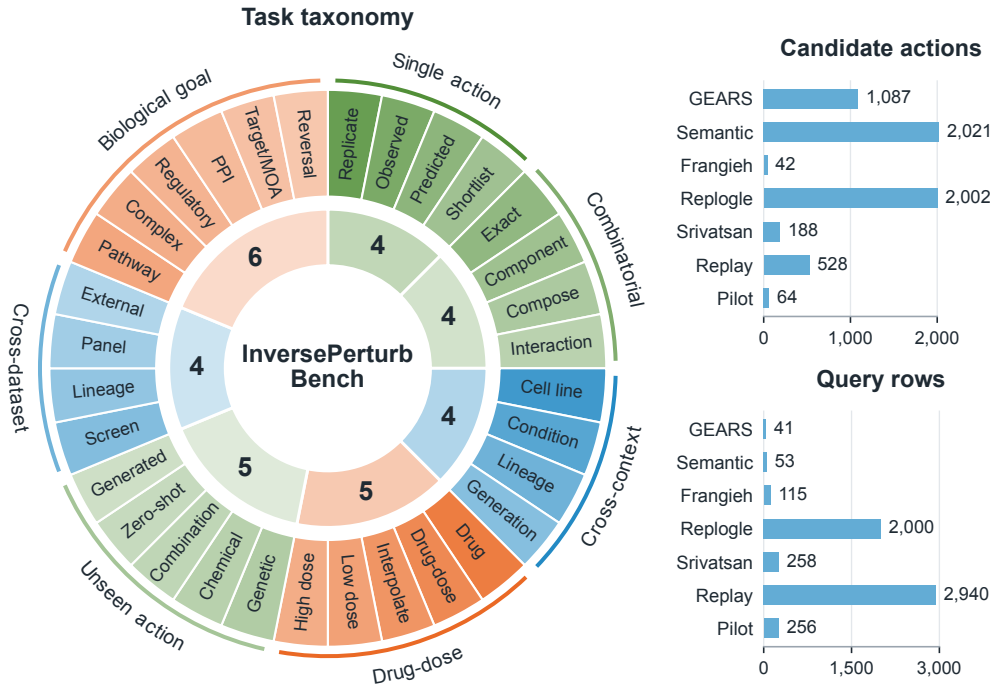


Figure 2: **Composition of the benchmark.** The inner ring groups 32 subtask definitions into seven task families; each outer sector is one subtask, with equal angular weight. Inner numerals give the number of subtasks per family. This taxonomy describes task definitions, not completed evaluations; 46 dataset-instance records instantiate 25 of the subtasks. Bars separately show candidate and query-row counts for the seven evaluated instances on linear axes. GEARS and Semantic use Adamson queries, Frangieh denotes Replogle-to-Frangieh transfer, Replogle denotes K562 retrieval, and Replay/Pilot denote the two McFaline analyses. Query rows have instance-specific units.

3.3 Ranks and uncertainty

Let $s(q, a)$ denote a candidate score and r_q the true action's rank, with $r_q = \infty$ for an absent action. We report top-k accuracy and mean reciprocal rank:

$$\text{Exact @}k = \frac{1}{|\mathcal{Q}|}\sum_q \mathbf{1}[r_q \le k], \qquad \text{MRR} = \frac{1}{|\mathcal{Q}|}\sum_q \frac{1}{r_q}. \tag{3}$$

For coverage $c_q = \mathbf{1}[a_q^\star \in \mathcal{A}_q]$ and $M_q = |\mathcal{A}_q|$, uniform random ranking gives

$$\mathbb{E}[\text{Exact @}k] = \frac{1}{|\mathcal{Q}|}\sum_q c_q \frac{\min(k, M_q)}{M_q}, \qquad \mathbb{E}[\text{MRR}] = \frac{1}{|\mathcal{Q}|}\sum_q c_q \frac{H_{M_q}}{M_q}, \tag{4}$$

where $H_M = \sum_{r=1}^{M} 1/r$. Effects are compared within a candidate universe, paired contrasts fix the query rows, candidates, and endpoint, and Table 9 gives the resampling unit of each interval. Two per-query summaries complement these means. First-rank concentration counts how often each candidate is ranked first; a few candidates absorbing many first ranks indicates candidate popularity, or hubness (Radovanović et al., 2010). Hit transitions count the top-k hits that one method retains, gains, and loses relative to another on the same rows, following the negative-flip view of model updates (Bansal et al., 2019).

4 The identity gap in predicted candidate libraries

The evaluated instances span five screens (Figure 2; Appendix Table 3). The two Adamson instances measure candidate origin, Replogle K562 and Srivatsan evaluate retrieval methods, Replogle-to-Frangieh examines readouts, and the McFaline instances test hit preservation and input coverage.

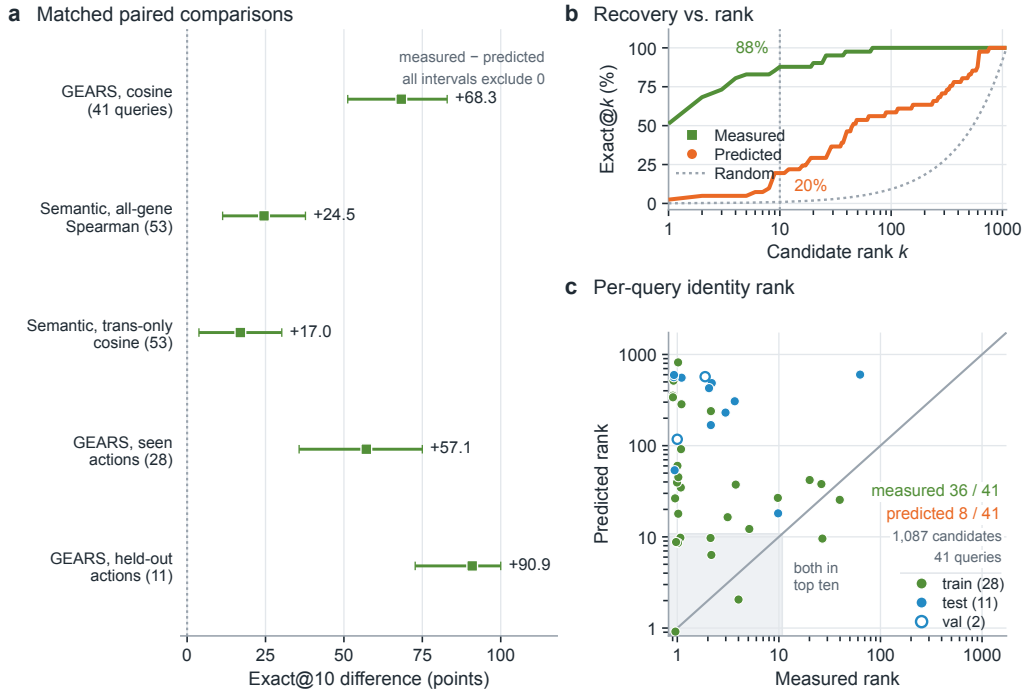


Figure 3: **The identity gap.** (a) Paired Exact@10 difference between measured and predicted candidate libraries, in percentage points, with stored 95% intervals; a positive value means measured responses recover more true interventions. (b) Recovery curves for both libraries on the 41-query GEARS panel over 1,087 candidates, against the uniform-ranking expectation. (c) True-action rank under each library for every GEARS query, coloured by the archived split, with the top-ten region shaded; counts give the queries recovered by each library.

Predicted libraries come from GEARS fitted to Replogle responses (Roohani et al., 2024), a transductive source-neighbor predictor that imputes each candidate response from the $k = 3$ most similar measured Replogle responses, and PerturBench DecoderOnly (Wu et al., 2025). Each effect compares matched rows within its protocol.

4.1 Measured references recover more identities than predicted candidates

GEARS covers 41 of the 53 Adamson query actions. For its archived h64/e20 export, the matched libraries contain 1,087 candidates on 4,573 aligned genes and share one cosine evaluator. The measured library uses the processed Replogle cells underlying model input, so the two libraries differ in whether candidate responses are measured or generated. Measured and predicted responses reach 87.80% and 19.51% Exact@10 against a random level of 0.92%, a paired gap of 68.29 points $[51.22, 82.93]$ (Figure 3a, Table 1). The gap extends across the ranking range (Figure 3b,c): the median true-action rank is 1 for measured and 46 for predicted candidates, and Exact@1 is 51.2% against 2.4%.

The semantic comparison holds 53 Adamson queries and 2,021 Replogle actions fixed; its predictor is fitted without Adamson outcomes, and static action features are available for all candidates. Measured and source-only semantic-predicted libraries reach 39.62% and 15.09% Exact@10 under all-gene Spearman scoring, a gap of 24.53 points $[11.32, 37.74]$, with MRR 0.2540 against 0.0811. Under global target-transcript exclusion and cosine scoring the gap is 16.98 points $[3.77, 30.19]$, so it extends beyond the target transcripts.

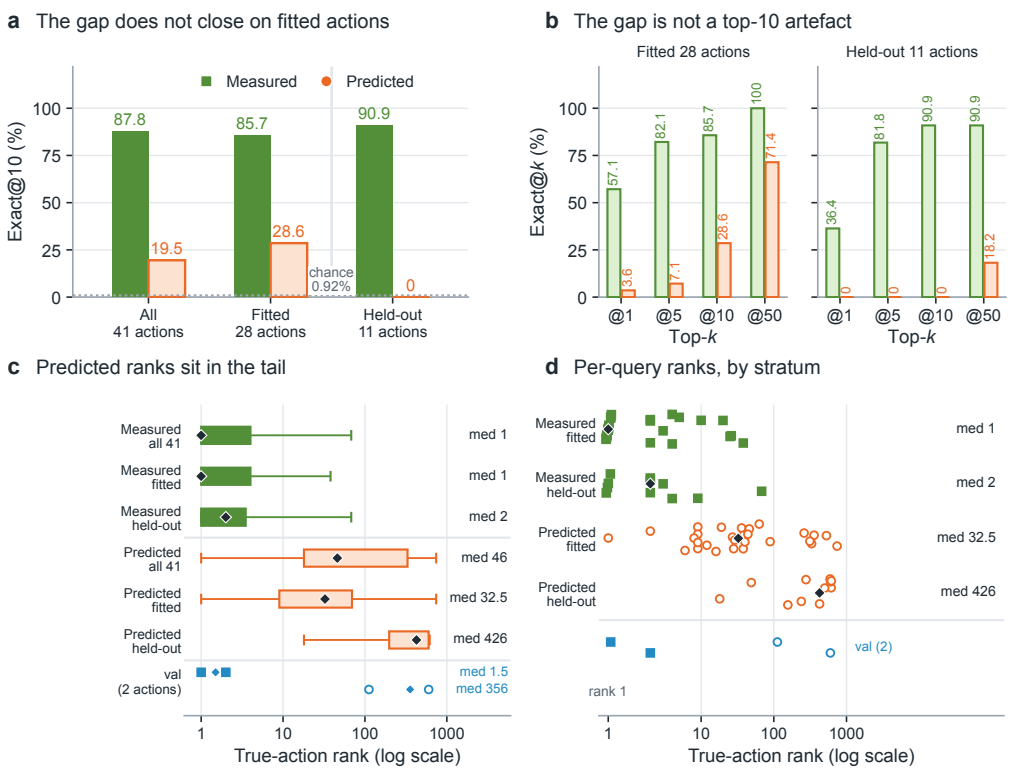


Figure 4: **The gap is not a generalization failure.** (a) Exact@10 for measured (green) and predicted (orange) candidates on all 41 GEARS queries, the 28 fitted actions, and the 11 held-out actions; the dotted line is the uniform-ranking expectation. (b) Accuracy at $k = 1, 5, 10, 50$ by stratum. (c) Per-stratum true-action rank distributions with medians; the two validation actions are shown separately. (d) Per-query true-action ranks by stratum on a log axis, with medians.

4.2 The gap survives on actions the predictor was fitted on

If predicted candidates failed only on unseen interventions, the gap would measure generalization and would close with more training data. Splitting the GEARS panel by its archived training labels tests this directly. On the 11 held-out actions, measured and predicted Exact@10 are 90.91% and 0%; on the 28 fitted actions they are 85.7% and 28.6% (Figure 4). The predictor recovers fewer than a third of the interventions it was trained on, and all eight of its top-ten hits fall inside that fitted stratum. The loss is a loss of candidate-specific detail in the predicted responses themselves, not a generalization deficit that more data would remove.

5 CARI: Candidate-Aware Rescoring for Identification

CARI targets the ordering of candidates rather than the fit of their responses. It leaves the candidate library frozen and learns only how to turn available similarities into an identification score, so no forward model is retrained. Its two forms, for genetic and drug perturbations, share one interface: a function from a query and a candidate to a score, fitted on training or source-cell data and applied to a fixed library.

5.1 CARI-fusion: genetic perturbations

On Replogle K562, each of five fixed splits holds out 20% of actions, and a fixed neighbor predictor represents each held-out candidate by averaging the responses of its three nearest training actions in feature space, which blurs action-specific changes. A query encoder trained with a listwise objective

Table 1: **Paired Exact@10 differences.** Each entry is the change in Exact@10, in percentage points of the metric, with a 95% interval computed by resampling biological units: held-out actions for Replogle and Adamson, and drugs for Srivatsan. Repeated seeds and contexts are averaged within a unit before resampling.

Contrast	Difference	95% interval
Measured − GEARS, cosine	+68.29	[51.22, 82.93]
Measured − semantic, all-gene Spearman	+24.53	[11.32, 37.74]
Measured − semantic, trans-only cosine	+16.98	[3.77, 30.19]
Frangieh: matched log − matched raw	+23.48	[10.83, 35.30]
Replogle: CARI-fusion − fixed, global control	+3.66	[1.98, 5.36]
Replogle: CARI-fusion − fixed, batch-linear	+4.37	[2.76, 6.00]
Srivatsan: CARI-cal (fixed) − raw	+10.85	[1.94, 19.77]
Srivatsan: CARI-cal (source-selected) − raw	+8.91	[1.16, 16.28]

instead maps each query to semantic action features and optimizes identification directly. We combine their scores as

$$s_\lambda(q, a) = (1 - \lambda)\mathcal{N}(s_{\text{fixed}}(q, a)) + \lambda\mathcal{N}(s_{\text{list}}(q, a)), \tag{5}$$

where $\mathcal{N}$ normalizes each score row. An inner training split selects λ and the normalization rule; the selected setting is refitted on outer training actions and evaluated on held-out actions. Expression inputs use the global trans-only readout throughout, so gains come from responses outside the target panel.

Score fusion raises Exact@10 from 15.17% to 18.83% under global controls and from 15.33% to 19.70% under batch-linear controls (Figure 5a,b), paired gains of 3.66 and 4.37 points (Table 1). Both policies evaluate the same 2,002 candidates and 2,000 query-by-split rows, which cover 1,358 distinct held-out actions; levels and paired changes are action-equal means. Global-control MRR rises from 0.0652 to 0.0753. The curves cross: fusion is slightly below the fixed predictor at $k = 1$ and above it from $k = 10$ onward, so the encoder mainly rescues correct candidates that the fixed predictor leaves just outside the top of the list.

5.2 CARI-CAL: DRUG PERTURBATIONS

On Srivatsan (A549, K562, and MCF7; three model seeds), DecoderOnly predicts each drug's response in a held-out cell type, and raw cosine scores place 92.2% of first ranks on training drugs. Drug results are means over the 86 distinct drugs. Calibration with features selected by source-cell validation MRR, the main result, raises Exact@10 from 24.42% to 33.33%, a paired gain of 8.91 points (Table 1); MRR rises from 0.098 to 0.130, and the training-drug share falls to 79.8% (Figure 5c,d). The fixed score–popularity–norm feature set, a secondary variant with no selection step, reaches 35.27% Exact@10 and 0.162 MRR, a paired gain of 10.85 points, and lowers the training-drug share to 58.5%. Appendix E resolves the calibration by feature set.

6 ANALYSIS

6.1 EXPRESSION SPACE AND THE TARGET PANEL

With the same 42 candidates and 115 queries, matching Replogle and Frangieh in per-cell $\log(1 + \text{CPM}_{10k})$ raises Exact@10 from 34.8% in raw-count space to 58.3% (Table 1). The 41 target transcripts, just 1.3% of 3,065 genes, alone reach 75.7%, consistent with proximal repression in both screens (Gilbert et al., 2014; Replogle et al., 2022). Removing them lowers recovery to 40.9%, versus 59.1% for a matched control panel; the remaining genes still exceed random ranking (23.8%) and a label-permutation null ($p = 1/5001$; Appendix F). We therefore use the global trans-only readout in the Replogle and semantic model comparisons.

6.2 CANDIDATE PREFERENCES AND THE STRUCTURE OF ERRORS

GEARS predictions rank DARS, EIF2B2, or EIF2S2 first for 22 of 41 queries, concentrating on translation-related actions (Adamson et al., 2016; Pakos-Zebrucka et al., 2016); measured first

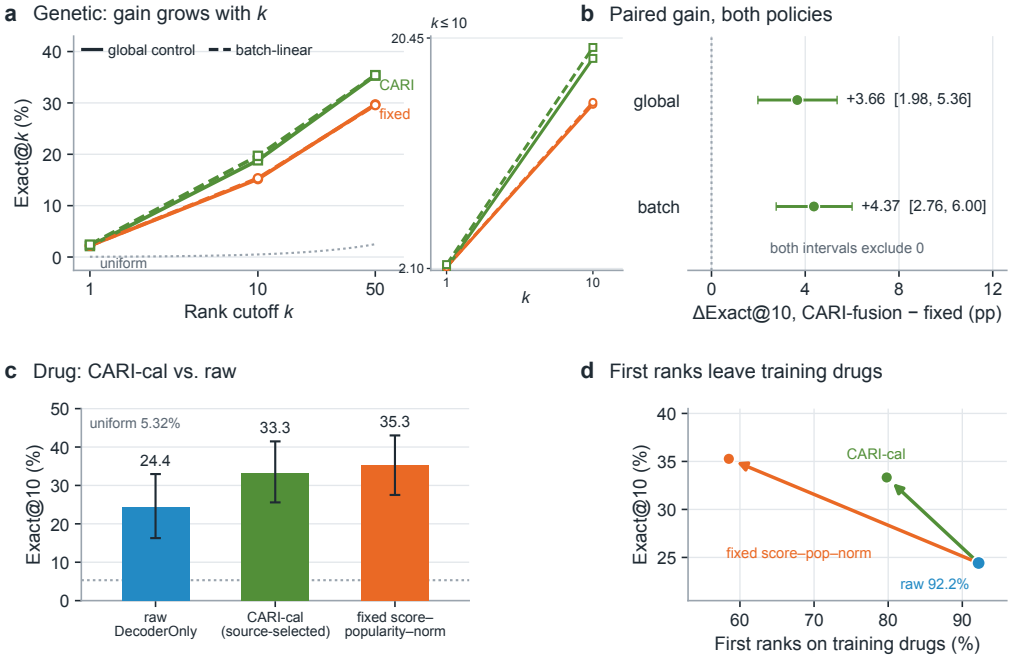


Figure 5: **CARI improves candidate ranking.** (a) Replogle, 2,002 candidates, 1,821 mapped target transcripts excluded: Exact@k for the fixed neighbor predictor (orange) and CARI-fusion (green) under global-control (solid) and batch-linear (dashed) signature policies. (b) Paired Exact@10 gain of CARI-fusion over the fixed predictor, by policy. (c) Srivatsan, 188 drug candidates, 258 query-by-seed rows: Exact@10 for raw DecoderOnly scores, CARI-cal selected on source-cell validation, and the fixed score–popularity–norm feature set. (d) Exact@10 against the share of first ranks placed on training drugs, with an arrow from the raw score to each calibration. Error bars are stored 95% intervals; dotted lines mark random ranking.

ranks span 29 actions versus 21 for predictions. Measured errors include functional neighbors in eIF2B (EIF2B2/EIF2B4 → EIF2B5), splicing (CHERP → U2SURP) (De Maio et al., 2018), and N-glycosylation (DDOST → DHDDS) (Kelleher & Gilmore, 2006; Park et al., 2014). Across all measured-library errors, 20% share a CORUM complex and 30% a Reactome pathway with the true action, against degree-matched nulls of 0.52% and 3.15% (Holm $p < 0.001$). Predicted errors show no comparable enrichment (Appendix Table 6).

Hit preservation in the McFaline replay. On McFaline's held-out action–context combinations (McFaline-Figueroa et al., 2024), a development-selected fallback between a hybrid predictor and Geneformer V2 raises replay Exact@10 from 3.57% to 4.01% over 2,940 queries, yet gains 48 hits and loses 35, exceeding the recorded loss limit of 29 (Appendix G.1). Mean accuracy improves, but hit preservation fails. Expanding the input from 15,009 to 58,347 columns yields no resolved MRR gain: all paired intervals include zero (Appendix H).

6.3 Why the identity gap persists

Figure 6 shows that neither small error nor lower candidate concentration guarantees identity recovery. In the analytic swap construction (bottom left), MSE $= \eta^2$ approaches zero, but Exact@1 stays zero for every $\eta > 0$: error must be small relative to candidate separation (Appendix B). GEARS centering (bottom center) lowers the largest first-rank share from 26.83% to 14.63% over 41 queries, while Exact@10 falls from 19.51% to 17.07% and the 11 test actions remain at 0%. By contrast, Srivatsan calibration (bottom right) raises Exact@10 from 24.42% to 33.33% with source-selected features or 35.27% with fixed features, while lowering the training-drug share. Dispersing first ranks is therefore not sufficient for better identification.

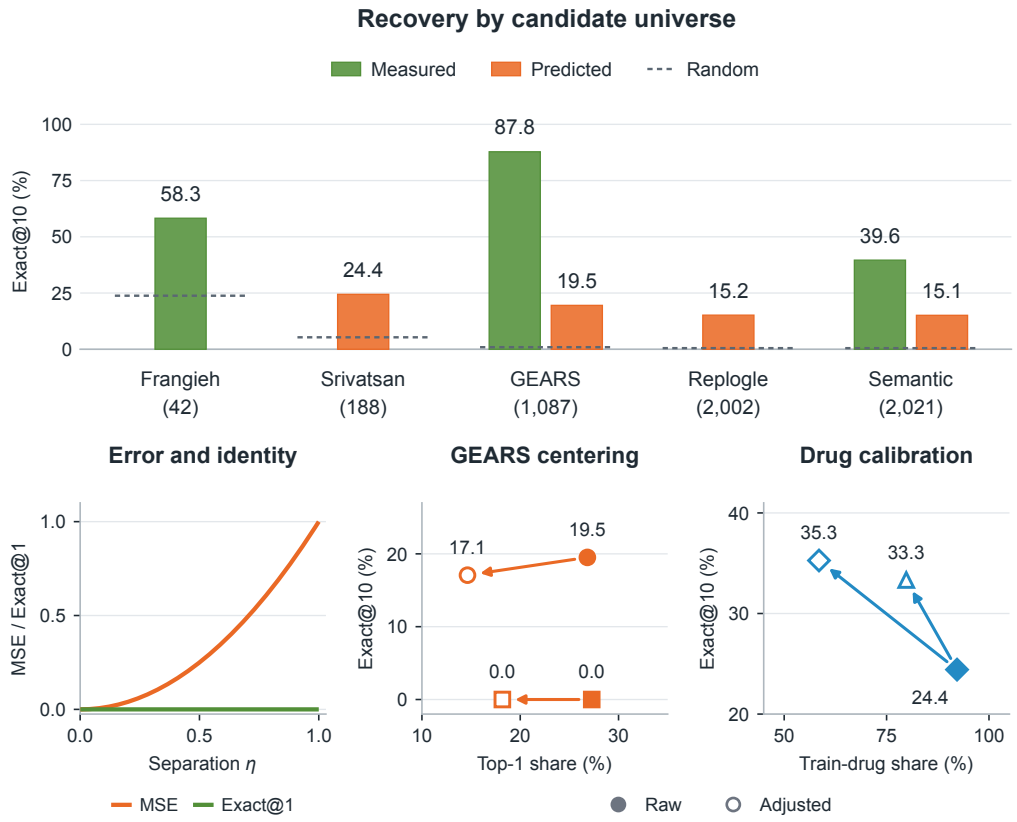


Figure 6: **Why the gap persists.** Top: recovery across candidate universes (counts in parentheses); dashed marks show uniform-random performance. Only paired bars are matched origin contrasts. Bottom left: the analytic swap example, with MSE $= \eta^2$ and Exact@1 $= 0$ for $\eta > 0$ (Appendix B). Bottom center: GEARS centering on all 41 queries (circles) and 11 test actions (squares); additional strata are in Appendix D.3. Bottom right: Srivatsan calibration with fixed (diamond) or source-selected (triangle) features. Arrows run from raw (filled) to adjusted (open) scores. Concentration denotes the largest single-candidate first-rank share for GEARS but the training-drug share for Srivatsan; these axes are not directly comparable.

The top row spans 42–2,021 candidates, with random Exact@10 ranging from 23.8% to 0.49%. This is not a controlled candidate-size ablation: datasets and query protocols differ. The two paired origin contrasts hold queries and candidate universes fixed, yet measured libraries outperform predicted ones in both. Library size cannot account for these matched gaps.

7 Conclusion

InversePerturbBench measures how much of an intervention's identity survives in the responses available for ranking it. Measured references recover far more identities than the predicted libraries we evaluate, and the identity gap persists on the actions a predictor was fitted on, so it is a property of the predicted responses rather than of generalization. CARI, which rescores a frozen library, improves genetic and drug retrieval on held-out actions, and the experiments show that suppressing candidate popularity alone does not. Reporting each predicted library beside a measured reference, a fitted-action stratum, a target-excluded readout, and hit transitions gives a concrete basis for perturbation models that support intervention selection.

AI USE STATEMENT

Generative AI assisted literature search, interpretation of archived records, drafting and revision, figure code, and LaTeX assembly. It also summarized supplied metadata and per-query ranks and simulated reviewer feedback. Performance measurements come from archived experiment records; the authors are responsible for all content.

REPRODUCIBILITY STATEMENT

Section 3 defines evaluation objects, and Appendix Table 3 lists candidates and query units. Appendices D–I describe implementation and uncertainty. The source contains numerical summaries, per-query GEARS ranks, figure code, and a McFaline metadata inventory. Reconstructing original model runs additionally requires the cell-level data, weights, configurations, and prediction assets described in Appendix J. Appendix C reports the protocol and leakage diagnostics.

REFERENCES

Britt Adamson, Thomas M. Norman, Marco Jost, Min Y. Cho, James K. Nuñez, Yuwen Chen, et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, 167:1867–1882.e21, 2016. doi: 10.1016/j.cell.2016.11.048.

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: 1657–1661, 2025. doi: 10.1038/s41592-025-02772-6.

Arc Institute. cell-eval2: Metrics reference. Software documentation, 2026. URL https://github.com/ArcInstitute/cell-eval2/blob/main/docs/metrics.md. Main branch, accessed September 22, 2026.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:2429–2437, 2019. doi: 10.1609/aaai.v33i01.33012429.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, et al. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20:1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187: 7045–7063, 2024. doi: 10.1016/j.cell.2024.11.015.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 129–136, 2007. doi: 10.1145/1273496.1273513.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1710.04087.

Gerold Csendes, Gema Sanz, Kristóf Z. Szalay, and Bence Szalai. Benchmarking foundation cell models for post-perturbation RNA-seq prediction. *BMC Genomics*, 26:393, 2025. doi: 10.1186/s12864-025-11600-2.

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

Paul Datlinger, André F Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14:297–301, 2017. doi: 10.1038/nmeth.4177.

Antonia De Maio, Hari Krishna Yalamanchili, Carolyn J. Adamski, Vincenzo A. Gennarino, Zhandong Liu, Jun Qin, et al. RBM17 Interacts with U2SURP and CHERP to Regulate Expression and Splicing of RNA-Processing Proteins. *Cell Reports*, 25:726–736.e7, 2018. doi: 10.1016/j.celrep.2018.09.041.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167:1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038.

Luke A. Gilbert, Max A. Horlbeck, Britt Adamson, Jacqueline E. Villalta, Yuwen Chen, Evan H. Whitehead, et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*, 159:647–661, 2014. doi: 10.1016/j.cell.2014.09.029.

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*, 10:1008–1025, 2026. doi: 10.1038/s41551-025-01481-x.

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*, 16:101405, 2025. doi: 10.1016/j.cels.2025.101405.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, et al. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21:1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26711–26722, 2022. doi: 10.52202/068431-1937.

Francesco Iorio, Roberta Bosotti, Emanuela Scacheri, Vincenzo Belcastro, Pratibha Mithbaokar, Rosa Ferriero, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, 107:14621–14626, 2010. doi: 10.1073/pnas.1000138107.

Kenji Kamimoto, Blerta Stringa, Christy M. Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A. Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614:742–751, 2023. doi: 10.1038/s41586-022-05688-9.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26:101, 2025. doi: 10.1186/s13059-025-03574-x.

Daniel J. Kelleher and Reid Gilmore. An evolving view of the eukaryotic oligosaccharyltransferase. *Glycobiology*, 16:47R–62R, 2006. doi: 10.1093/glycob/cwj066.

Eric Kernfeld, Yunxiao Yang, Joshua S. Weinstock, Alexis Battle, and Patrick Cahan. A comparison of computational methods for expression forecasting. *Genome Biology*, 26:388, 2025. doi: 10.1186/s13059-025-03840-y.

Justin Lamb, Emily D. Crawford, David Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*, 313:1929–1935, 2006. doi: 10.1126/science.1132939.

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

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19:e11517, 2023. doi: 10.15252/msb.202211517.

José L. McFaline-Figueroa, Sanjay Srivatsan, Andrew J. Hill, Molly Gasperini, Dana L. Jackson, Lauren Saunders, et al. Multiplex single-cell chemical genomics reveals the kinase dependence of the response to targeted therapy. *Cell Genomics*, 4:100487, 2024. doi: 10.1016/j.xgen.2023.100487.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365:786–793, 2019. doi: 10.1126/science.aax4438.

Karolina Pakos-Zebrucka, Izabela Koryga, Katarzyna Mnich, Mila Ljujic, Afshin Samali, and Adrienne M Gorman. The integrated stress response. *EMBO Reports*, 17:1374–1395, 2016. doi: 10.15252/embr.201642195.

Eon Joo Park, Kariona A. Grabińska, Ziqiang Guan, Viktor Stránecký, Hana Hartmannová, Kateřina Hodaňová, et al. Mutation of Nogo-B Receptor, a Subunit of cis-Prenyltransferase, Causes a Congenital Disorder of Glycosylation. *Cell Metabolism*, 20:448–457, 2014. doi: 10.1016/j.cmet.2014.06.016.

Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, et al. scPerturb: harmonized single-cell perturbation data. *Nature Methods*, 21:531–540, 2024. doi: 10.1038/s41592-023-02144-y.

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. Preprint.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42:1678–1683, 2024. doi: 10.1038/s41587-023-02079-x.

Owen J L Rackham, Jaber Firas, Hai Fang, Matt E Oates, Melissa L Holmes, Anja S Knaupp, et al. A predictive computational framework for direct reprogramming between human cell types. *Nature Genetics*, 48:331–335, 2016. doi: 10.1038/ng.3487.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010. URL https://www.jmlr.org/papers/v11/radovanovic10a.html.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185:2559–2575.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*, 42:927–935, 2024. doi: 10.1038/s41587-023-01905-6.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorcan, Tabula Sapiens Consortium, Stephen R. Quake, et al. Universal Cell Embeddings: A Foundation Model for Cell Biology. *bioRxiv*, 2023. doi: 10.1101/2023.11.28.568918. Preprint.

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367:45–51, 2020. doi: 10.1126/science.aax6234.

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, 171:1437–1452.e17, 2017. doi: 10.1016/j.cell.2017.10.049.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, et al. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, 2023. doi: 10.1038/s41586-023-06139-9.

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

Aaron Wenteler, Martina Occhetta, Nikhil Branson, Victor Curean, Magdalena Huebner, William Dee, et al. PertEval-scFM: Benchmarking single-cell foundation models for perturbation effect prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66633–66677, 2025. URL https://proceedings.mlr.press/v267/wenteler25a.html.

Yan Wu, Esther Wershof, Sebastian M. Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. PerturBench: Benchmarking machine learning models for cellular perturbation analysis. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3225. Datasets and Benchmarks Track.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, et al. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14294–14303, 2021. doi: 10.1109/CVPR46437.2021.01407.

A TASK TAXONOMY AND EVIDENCE SCOPE

INVERSEPERTURBBENCH separates task definitions from instantiated datasets and from completed model evaluations. The seven task families are listed in Table 2. The archived registry has 32 subtask definitions and 46 dataset-instance records covering 25 unique subtasks. These counts describe coverage of definitions and data. A method can be eligible but unevaluated, incompatible with an action ontology, or available only as a diagnostic, so scores are reported per instance.

Table 2: Task families in the registry; the main paper reports selected instances.

ID	Family	Decision question
T1	Single-action identification	Which legal intervention matches a target response?
T2	Combinatorial attribution	Which complete combination or constituent actions explain a response?
T3	Cross-context transfer	Which action can be identified when source and target contexts differ?
T4	Drug-dose / regimen resolution	Which compound and dose or regimen matches the response?
T5	Unseen-action discovery	Can a held-out action be ranked from allowed features or generated responses?
T6	Cross-dataset transfer	Does an inverse rule transfer across independently produced screens?
T7	Biological-goal consistency	Do selected actions agree with an independently specified biological goal?

In T5, "discovery" refers to an action held out from permitted training responses. T7 requires an endpoint independent of the ranker's own errors. The main-text experiments instantiate exact-identity tasks.

Table 3: **Evaluated instances.** M is the number of candidates. Query rows can share actions, contexts, or model seeds. The candidate library of each instance is measured or predicted, and the last column gives the interval resampling unit. The task registry is listed in Table 2, and Section 4 reports the results for each instance.

Instance	M	Evaluation unit	Comparison
Adamson → GEARS library	1,087	41 query actions	Measured / predicted responses
Adamson → Replogle	2,021	53 query actions	Measured / semantic predictions
Replogle → Frangieh	42	115 action-context queries	Expression space and gene panel
Replogle K562	2,002	2,000 query-by-split rows	Fixed / nested score fusion
Srivatsan highest dose	188	258 query-by-seed rows	Raw / calibrated ranking
McFaline replay	528	2,940 guide-context profiles	Mean accuracy and hit transitions
McFaline fixed-cell pilot	64	256 profiles	Old / full expression axis

B AN ELEMENTARY MARGIN ARGUMENT

This metric-space argument relates prediction error, candidate margin, and action recovery.

Proposition 1 (Sufficient condition for preserving a top-1 decision)**.** *Let q be a query in a metric space $(\mathcal{Z}, d)$, and suppose $a^\star$ is the unique nearest measured candidate in a finite library $\{z_a : a \in \mathcal{A}\}$. Define*

$$\gamma(q) = \min_{a \neq a^\star} d(q, z_a) - d(q, z_{a^\star}).$$

If $d(z_a, \hat{z}_a) \leq \epsilon$ for every candidate and $\gamma(q) > 2\epsilon$, then $a^\star$ is the unique nearest predicted candidate.

Proof. For every alternative a, the reverse triangle inequality gives $d(q, \hat{z}_a) \geq d(q, z_a) - \epsilon$, while $d(q, \hat{z}_{a^\star}) \leq d(q, z_{a^\star}) + \epsilon$. Subtracting yields $d(q, \hat{z}_a) - d(q, \hat{z}_{a^\star}) \geq \gamma(q) - 2\epsilon > 0$. □

The condition bounds the worst-case error by the margin, which average error leaves free. Let two one-dimensional observed responses be $z_1 = 0$ and $z_2 = \eta$, with queries $q_1 = z_1$ and $q_2 = z_2$. Swap the predicted responses: $\hat{z}_1 = \eta$ and $\hat{z}_2 = 0$. The mean squared prediction error is η^2, which tends to zero as $\eta \to 0$, but nearest-candidate Exact@1 is zero for both queries for every $\eta > 0$. The observed margin shrinks together with the prediction error, so correct identity requires error small relative to the margin. For cosine scoring, the proposition applies to the chord metric $\sqrt{2 - 2\cos}$ on unit-normalized responses.

C PROTOCOL AND LEAKAGE DIAGNOSTICS

Table 4 reports the diagnostics. These diagnostics come from the archived inverse-perturbation protocol runs. They test whether a reported score can be recovered from the candidate index or the split rather than from a response prediction, and each is scoped to the interface named beside it.

Table 4: Protocol diagnostics. Scores are quoted on their own interface and are not comparable across rows.

Diagnostic	Interface	Result
Candidate index	Norman (Norman et al., 2019), observed signatures as candidates	An index that includes the query key returns Exact@1 of 100%. Under a strict train-candidate policy with any-gene recovery, linear decomposition reaches 43.3% $[34.0, 52.6]$ and a GEARS forward-to-inverse bridge 23.7% $[15.5, 32.0]$. Exact combination recovery is zero throughout.
All-query against split-aware retrieval	PerturBench (Wu et al., 2025) LinearAdditive, 2,160-cell 8-drug Srivatsan subset	All-query Exact@1 is 66.7%, and all 16 successes are training query groups. Split-aware retrieval over the same predictions gives Exact@1 of 0/8 on held-out groups with mean exact rank 7/8; every held-out top-1 candidate is drawn from the training split, against a held-out MRR of 14.4% for the model and 34.0% for random ranking.
Paper-scale split boundary	PerturBench LinearAdditive e1, 188 Srivatsan drugs	All-query Exact@10 is 34.04% against 5.32% random, falling to 30.34% after adjustment. Held-out Exact@10 is about 9.9% while training groups reach 44.5%, and about 85% of held-out top-1 candidates are training candidates. A train-mean baseline reaches 32.62% all-query.

The three diagnostics share one failure mode: a candidate index or an all-query evaluation lets a ranker recover an intervention from group membership rather than from the measured response. The first row shows the extreme case, where including the query signature makes the answer a lookup. The second and third show the same effect at smoke and paper scale in an official forward-model bridge, where predicted candidate signatures concentrate on training prototypes even though the true held-out candidate is present for every query. Split-aware retrieval and a strict candidate policy are therefore part of the task definition rather than presentation choices, and the all-query figures in the archived bridge runs are reported here as diagnostics.

On Norman, canonicalizing combination labels gives 108 training, 26 validation, and 97 test conditions with empty pairwise intersections, so no condition crosses the split. Test-label selection is bounded on the 258-query Srivatsan protocol. The deployable configuration selects features on cross-cell validation without target test labels; an oracle that selects on target test labels, used only as a forbidden reference, gains 1.9 points of Exact@10 $[0.0, 4.3]$ (38.8% to 40.7%) and 0.8 points of MRR $[0.2, 1.7]$, so test-label selection adds at most about two points on this protocol. The comparison uses the target-cell-trained configuration; the source-selected result in Section 5 selects on source-cell validation.

The audit covers candidate policy, split routing, and model selection; pretraining exposure of foundation-model checkpoints is a separate question.

D IMPLEMENTATION AND INFORMATION USE

D.1 REPLOGLE NESTED FUSION

The strict refit removes 1,821 mapped candidate target transcripts from an 8,563-gene readout panel, leaving 6,742 expression features. Candidate identities remain the same 2,002 actions. The source action features concatenate the archived GO biological-process, GenePT model-3, and Reactome-2024 representations; their availability for all candidates is part of the information contract. They come from external annotations and gene-description embeddings, independent of query responses.

The fixed predictor averages the responses of three training-action neighbors using positive cosine weights in the action-feature space. The archive calls this baseline `fixed_fusion`; in the score-fusion implementation it corresponds to $\lambda = 0$. The listwise route maps a query signature into the action-feature space, with training-set standardization for both spaces. It optimizes a temperature-scaled classification objective over training actions. The registered encoder has hidden dimension zero, making it a linear query map.

The recorded refit uses five outer split seeds, 101–105, each holding out 20% of actions, and a 20% inner validation fraction within the remaining actions. The model seed is 1, training lasts 150 epochs, batch size is 256, temperature is 0.05, learning rate is 10^{-3}, and weight decay is 10^{-2}. The configuration records dropout 0.1. The inner grid contains row z-score or rank-percentile normalization and $\lambda \in \{0, 0.1, \ldots, 1\}$. Selection uses Exact@10 with recorded secondary tie-breaks. The selected setting is refitted on the outer training actions before evaluating the held-out actions against the full candidate library. These details come from the archived refit contract.

The two signature policies require at least 20 cells and use either global controls or a batch-adjusted linear effect estimate. Both undergo the same global trans-only refit, which closes the direct target-transcript route in the expression inputs; static candidate features keep their permitted action identity.

D.2 SRIVATSAN SOURCE-CELL CALIBRATION

For candidate response z_a and query q, the compact fixed feature set is

$$\phi(q, a) = \big(s(q, a),\, h(a),\, s(q, a)h(a),\, \|z_a\|_2\big), \qquad s(q, a) = \cos(q, z_a).$$

Here $h(a)$ is the frequency with which the candidate occurs among the top five alternatives for permitted training queries in the corresponding context, normalized by its opportunities to appear. Exact self-matches are excluded when computing this frequency. It measures candidate popularity, the hubness quantity the head corrects.

The logistic head standardizes features using its fitting rows and gives the positive and negative classes equal total weight. The source implementation uses L2 regularization, an unpenalized intercept, and gradient updates; its recorded defaults are L2 coefficient 1, learning rate 0.1, and 1,000 iterations. Source cell types contribute validation-labeled candidate/query pairs to fit the head, and the held-out cell type contributes evaluation queries. Permitted target-cell training profiles enter the popularity statistic.

The source-selected variant compares the recorded feature sets using leave-one-source-cell validation, then refits on all source cells. The sets comprise raw score only; score and rank features; score, popularity, and their interaction; the latter plus candidate norm; and the full combination. The source-MRR rule selects the full feature set in six of nine seed-by-target-cell contexts, score–popularity in two, and score–popularity–norm in one. We report both the fixed-feature and source-selected results; source-cell validation controls fitting and feature selection, while target-context training profiles contribute to popularity.

D.3 MEASURED-SOURCE AND GEARS COMPARISONS

The semantic source comparison uses a 2,021-action library and 53 Adamson query actions. It permits source-screen responses and semantic action features, and fitting uses no Adamson outcomes. The measured and predicted rows share the candidate registry. The source and target screens differ in biological context, so the origin gap is a transferred-response gap.

The GEARS comparison uses the archived h64/e20 export, 1,087 supported candidate actions, and 4,573 shared exact-symbol genes. The 41 covered query actions comprise 28 training, two validation, and 11 test actions under the GEARS split. The run label records architecture width (h64) and nominal training budget (e20), and results describe this archived export. Measured comparison candidates use the exact processed source cells, isolating candidate origin within this interface. The two validation actions are reported with the covered set. Per-query ranks give median true-action ranks of 1 for measured and 46 for predicted candidates; on the 11 test actions the medians are 2 and 426. Exact@1 is 51.2% (21 of 41) for measured and 2.4% (1 of 41) for predicted candidates.

The same 1,087×4,573 predicted matrix yields cross-screen response agreement alongside those ranks. Mean per-action Pearson correlation between predicted and observed response differences is 0.375 over all 41 queries and 0.293 on the 11 test actions. Response-difference MSE is 0.00559 over all 41 queries against 0.00542 for a zero-response baseline, and 0.00575 against 0.00571 on the test actions, so the predicted responses do not beat the no-change baseline on this error measure. Both sides of this comparison are external cross-screen quantities: the predictor was fitted to Replogle responses and is scored against Adamson responses, not against a held-out portion of its own training distribution. Correlation with the observed response direction and error magnitude therefore dissociate, which is the behavior Proposition 1 describes, and neither quantity alone determines the ranking result.

A candidate-popularity correction tests how much of the predicted-library gap is positional. Subtracting, with coefficient one, each candidate's mean score over training queries only leaves the 11 held-out test actions at 0% Exact@10 with MRR falling from 0.919% to 0.749%, and leaves the 13 validation and test actions at 0% with MRR falling from 0.859% to 0.713%. Over all 41 queries the correction lowers Exact@10 from 19.51% to 17.07% and the largest first-rank share from 26.83% to 14.63%, sweeping DARS from 11 first ranks to 3 and EIF2S2 from five to none while leaving EIF2B2 at six. The correction therefore changes which candidates lead without recovering accuracy, and it fits on target-side training queries, so it is a target-adapted diagnostic rather than a target-free method.

E SCORE-ABLATION DETAIL

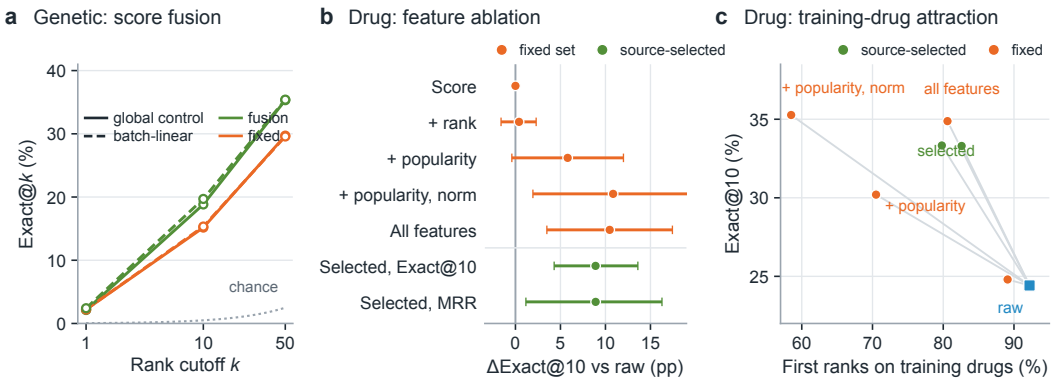


Figure 7: **Score fusion and drug calibration in detail.** (a) Replogle, 2,002 candidates: Exact@k at $k = 1, 10, 50$ averaged over five fixed splits, under global-control (solid) and batch-linear (dashed) signature policies, for the fixed neighbor predictor and nested score fusion; the dotted line is the uniform expectation. (b) Srivatsan, 188 drug candidates: paired change in Exact@10 against the raw score for each recorded feature set over 258 query-by-seed rows, with stored 95% intervals; filled markers are fixed feature sets and open markers the source-selected variants. (c) Exact@10 against the share of first ranks placed on training drugs; the square is the raw score. Dotted lines mark random ranking.

An archived all-readout run resolves the fusion into its components on a related protocol, where fixed and listwise scores were both retained. Exact@10 is 15.70%, 18.15%, and 19.85% under global controls and 16.50%, 18.15%, and 20.50% under batch-linear controls for the fixed, listwise, and fused scores; the fused score exceeds both components, and its advantage over the listwise score alone

is positive on action clusters for Exact@10 and MRR. The gain over the fixed score is positive for Exact@10 and crosses zero for MRR, so the fusion result is an Exact@10 result. This run evaluates the historical all-readout protocol rather than the global trans-only protocol of Section 5, and the learned listwise branch of that protocol was not retained, so its components cannot be recomputed from stored scores. A sweep of the listwise weight over this protocol, at 11 points from 0 to 1 and on a 23-point grid for each of three score normalizations, is released with the evidence package; it is a sensitivity analysis, and no weight is selected on test performance.

Figure 7 gives the detail. Panel (a) shows the fusion gain growing with k, and panel (b) shows that score, popularity, and norm carry the calibration: the score-only and score–rank sets leave Exact@10 at or near the raw level, while adding popularity and then candidate norm raises it. Panel (c) relates the same configurations to first ranks placed on training drugs; source-selected rows move the raw share from 92.2% down to 79.8%, and the fixed score–popularity–norm set to 58.5%.

F Additional transfer and model results

F.1 Expression value space and target transcripts

Table 5 and Figure 8 report the 42-action Replogle-to-Frangieh comparison. There are 115 action-context queries and 3,065 aligned genes. Removing all mapped candidate target transcripts removes 41 genes; the candidate universe remains 42 actions. The action-only and action-excluded rows split the ranking signal between the target panel and the remaining genes.

Table 5: Replogle-to-Frangieh transfer. Exact metrics and their intervals are percentages; MRR is on $[0, 1]$. Exact@10 intervals cluster 115 queries by 42 actions.

Value space / readout	Exact@1	Exact@10	MRR	Exact@10 interval
Raw / raw; all	6.1	34.8	0.1608	[24.0, 45.5]
Log / log; all	29.6	58.3	0.4145	[46.1, 69.9]
Log / log; action genes excluded	7.8	40.9	0.2105	[29.9, 51.9]
Log / log; action genes only	47.0	75.7	0.5644	[63.2, 86.5]
Uniform random	2.4	23.8	0.1030	—

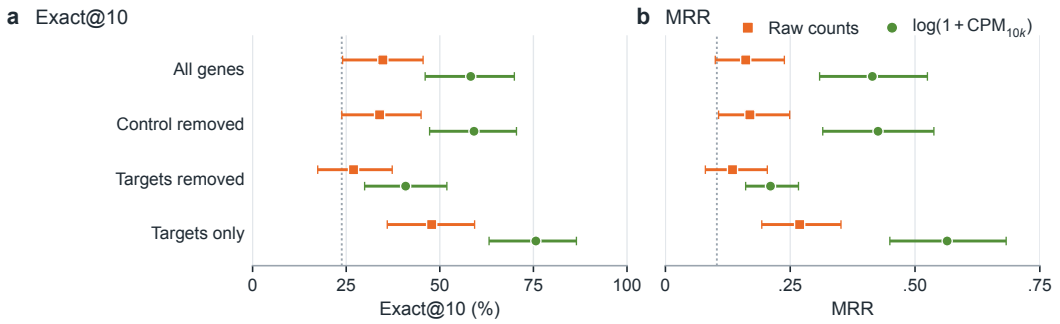


Figure 8: Value-space and readout contrasts on a fixed measured-source transfer task, with the same 42 candidates and 115 queries in every row. Panels show Exact@10 and MRR for raw counts (squares) and $\log(1 + \mathrm{CPM}_{10k})$ (circles); lines are action-cluster 95% intervals and dotted lines mark random ranking.

The archived log-space action-label permutations give $p = 1/5001$ for the all-readout Exact@1, Exact@5, Exact@10, and MRR endpoints. After action-transcript exclusion, Exact@10 and MRR also have $p = 1/5001$. A single matched-control panel gives 59.1% Exact@10. Additional source-only matched-control deletion panels provide a sensitivity analysis: across three local-neighborhood settings, their mean Exact@10 is approximately 59.7–59.8%, compared with 40.9% after deleting the action transcripts. The drop is therefore specific to the action genes relative to those matched panels. Permutation results are reported per dataset.

G PATHWAY AND COMPLEX CONSISTENCY OF RANKING ERRORS

The examples of functional near-misses in Section 6 come from annotated cases. Table 6 reports the complete error set, which supports the same reading under a rule frozen before analysis: CORUM 5.3 human complexes and the Reactome human pathway snapshot, exact HGNC symbol matching, and a 5–200 gene size rule for primary pathways. For every query whose first-ranked candidate is wrong, we ask whether that candidate and the true action share at least one annotation. Unannotated errors stay in the denominator and are also reported separately.

Table 6: **Annotation agreement of first-rank errors.** Proportion of erroneous first ranks whose candidate shares a complex or pathway with the true action. The matched null resamples wrong candidates within annotation-degree strata; Holm correction covers the four primary tests.

Library	Annotation	Errors	Agreement	Null	p_{Holm}
GEARS	CORUM	40	0%	0.72%	1.000
GEARS	Reactome	40	5%	3.51%	0.844
Measured	CORUM	20	20%	0.52%	< 0.001
Measured	Reactome	20	30%	3.15%	< 0.001

Measured-library errors carry functional structure: 20% of the 20 erroneous first ranks share a CORUM complex with the true action against a matched-null mean of 0.52%, and 30% share a Reactome pathway against 3.15%. Predicted-library errors do not show the same structure: 0% and 5% for CORUM and Reactome, neither distinguishable from the null after Holm correction. The contrast is between two error sets that differ in size and composition, so it describes the character of each library's errors rather than a paired method advantage. The reading is also sensitive to annotation breadth: permitting Reactome pathways of any size raises GEARS agreement to 42.5

G.1 MCFALINE DEVELOPMENT HISTORY

The matrices contain 35,745/2,910/2,941 train/development/original-test profiles on 15,009 genes, with median support 13/18/17 cells. Training covers 528 actions, 2,841 guides, and 15 contexts. The partitions' 6,631/542/543 action–context combinations are disjoint, while all development/test actions, guides, and contexts occur in training. Aggregating 8,523 candidate guide–context rows gives 528 actions across three target contexts. In Table 7, "Frozen" denotes a fixed comparator interface, "Selected" development selection, and "Action OOF" action-group out-of-fold predictions. Guide support, selection procedures, and seeds differ across rows, with most OOF runs using one seed, so the table records development history.

Table 7: McFaline development history, 528 candidates. Exact metrics are percentages.

Method	Protocol	Exact@1	Exact@10	MRR
Centroid	Frozen	0.447	2.715	0.01840
Ridge	Frozen	0.275	2.577	0.01673
scGPT frozen cosine	Frozen	0.172	2.921	0.01652
Geneformer V1 frozen cosine	Frozen	0.103	3.024	0.01709
Geneformer V2 frozen cosine	Frozen	0.344	3.127	0.01892
Protected neural ranker	Selected	0.550	2.887	0.01962
Candidate residual ranker	Action OOF	0.241	3.162	0.01731
Guide-invariant shared encoder	Action OOF	0.103	3.162	0.01749
Geneformer V2 partial fine-tuning	Action OOF	0.309	3.058	0.01859
Raw-expression shared encoder	Action OOF	0.309	1.649	0.01276
V2 / action-semantic dual encoder	Action OOF	0.309	2.371	0.01642
Fold-local mean shift	Action OOF	0.584	3.299	0.02018
Measurement-aware hybrid	Action OOF	0.619	3.780	0.02361
Selective fallback	Action OOF	0.412	3.746	0.02123

Geneformer V2 uses the frozen 104M cancer-oriented checkpoint, official gene dictionary, a 4,096-token cap, and final-layer mean pooling over non-padding tokens; the full-axis pilot includes special tokens. In the archived centroid representation interface, each action prototype averages its training

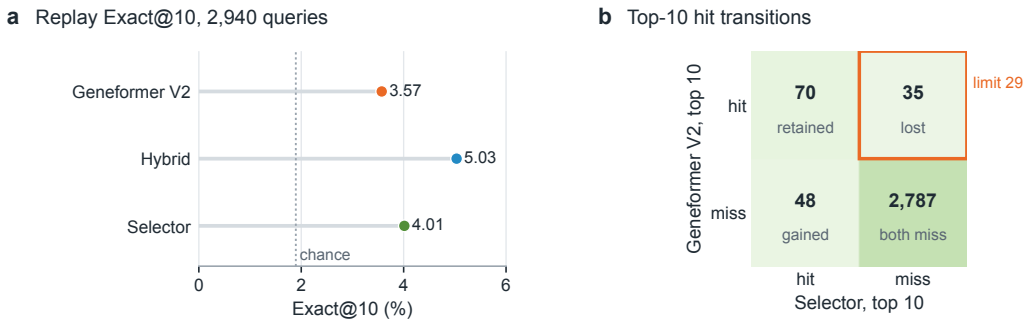


Figure 9: **Accuracy and hit preservation on the McFaline replay.** (a) Exact@10 on 2,940 queries and 528 candidates; the dotted line is random ranking. (b) Top-10 transitions between Geneformer V2 and the selector. The outlined cell holds the 35 hits lost against the recorded limit of 29.

profiles outside the query context before cosine scoring. Thus guide/context rows supply 528 action prototypes. The supplied foundation inputs contain 38,655 train/development raw-count sums and 151 encoding batches, with no test count-sum profiles. The transfer inputs instead contain signed deltas for all partitions. Their 422,104 directed training pairs preserve guide identity while changing context, so supervision is within-guide. Per-cell transformations require the original cells, and the two input interfaces are distinct.

The selector retains 70 baseline hits, recovers 48 misses, loses 35 hits, and misses jointly on 2,787 queries. V2 has 105 hits and the selector 118, but 35 losses exceed the recorded limit of 29. The 2,940-row reporting subset omits one original query, and the results are a retrospective replay.

Figure 9 shows the accuracy and the transition counts. Representation diagnostics show mean sampled Geneformer cosine near 0.922, approximately 92.18% of uncentered squared singular-value energy in the leading direction, and 768 numerically nonzero singular values. Raw and centered same-action cross-guide AUC values are approximately 0.5014 and 0.5066. These observations indicate strong anisotropy and weak action separability under the evaluated interface.

H FIXED-CELL FULL-AXIS DIAGNOSTIC

The pilot uses 64 actions, two guides, and two contexts: A172 without drug and U87MG with lapatinib. At 16 perturbed cells per profile, $64 \times 2 \times 2 = 256$ profiles contain 4,096 perturbed cells. The 1,530 true non-targeting controls are matched by partition, environment, dose, and replicate using recorded profile weights; guide folds have disjoint control selections. Candidates always use the other guide. “Same” preserves context and “cross” changes it. This pilot therefore separates guide identity from action identity, which the broader development panel shares across partitions.

The full count axis has 58,347 columns versus 15,009 in the old axis. The archived input comparison reports identical counts at every old-axis intersection entry for the fixed cells. We compare expression differences using old-axis normalization, full-axis normalization and scoring, and full-axis normalization restricted back to the old intersection. The last variant separates normalization-denominator effects from the contribution of extra features. The comparison uses fixed feature sets. Table 8 gives the endpoint values and Figure 10 the paired contrasts.

For same-context retrieval, full-axis expression MRR is 0.07677 and full-axis Geneformer MRR is 0.06935. Their one-sided label-permutation p-values are approximately 0.3443 and 0.6892. The corresponding random MRR is 0.07412. Same-minus-cross-context MRR intervals include zero for both representations. Cell line and drug change together between the two contexts.

The full Geneformer axis remains limited to 4,096 tokens after mapping and rank-based tokenization. At most 4,096 of the 58,347 columns enter the transformer for each profile. Increasing support to 32 cells per guide while retaining both common guides leaves only four eligible actions in the recorded selection analysis. That would define a different four-action task. Knockdown efficiency is measured by independent guide-quality evidence, beyond target-gene zero counts.

Table 8: Fixed-cell McFaline pilot. Exact@1 and Exact@5 are percentages; MRR and intervals are on $[0, 1]$. "Intersection, full norm." scores only the old columns after normalization on all columns.

Representation / axis	Context	Exact@1	Exact@5	MRR	95% interval
Expression / old	same	0.781	8.203	0.06710	[0.05183, 0.08483]
Expression / old	cross	0.391	5.469	0.05578	[0.04084, 0.07268]
Expression / full	same	1.953	7.422	0.07677	[0.05686, 0.10021]
Expression / full	cross	1.172	7.422	0.06214	[0.04489, 0.08089]
Expression / intersection, full norm.	same	1.562	7.422	0.07286	[0.05508, 0.09334]
Expression / intersection, full norm.	cross	0.391	5.469	0.05248	[0.03920, 0.06768]
Geneformer / old	same	2.344	7.812	0.07982	[0.05891, 0.10255]
Geneformer / old	cross	1.172	5.859	0.07058	[0.05545, 0.08805]
Geneformer / full	same	1.172	6.641	0.06935	[0.05322, 0.08870]
Geneformer / full	cross	2.344	7.422	0.08036	[0.06214, 0.10128]
Uniform random	either	1.5625	7.8125	0.07412	—

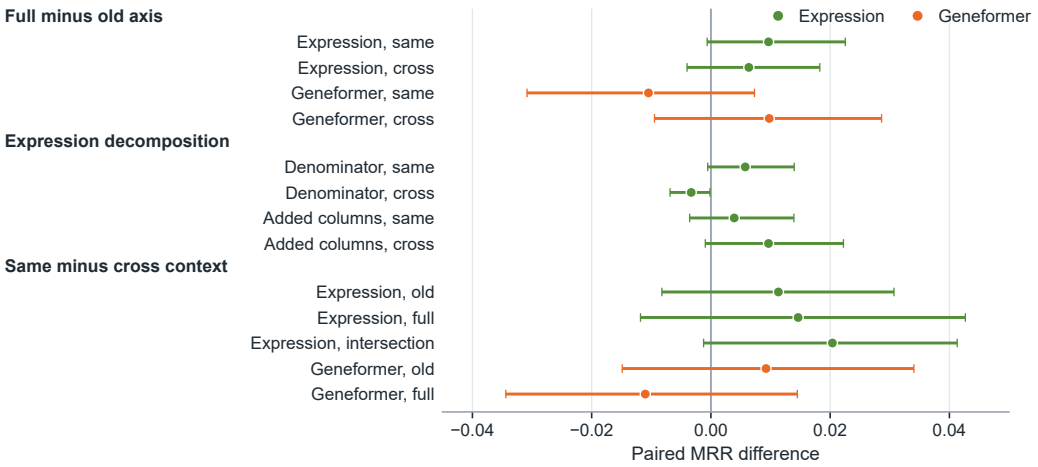


Figure 10: **Paired MRR differences on fixed cells.** The 256 profiles cover 64 actions and candidates always use the other guide; same/cross denotes the context relation. Rows are grouped into full-minus-old-axis contrasts, the decomposition of that difference into normalization denominator and added columns, and same-minus-cross-context contrasts. Lines show stored 95% action-cluster intervals.

I Metrics and Statistical Interpretation

I.1 Tie Handling and Missing Candidates

Stable sorting in the registered candidate order is retained for the archived Replogle fusion implementation. The full-axis pilot instead averages uniformly over positions occupied by tied candidates. If b candidates have scores strictly above the true action and e candidates, including the truth, are tied with it, then

$$\mathbb{E}[\mathbf{1}(r \le k) \mid \text{tie}] = \frac{\max\{0, \min(e, k-b)\}}{e}, \qquad \mathbb{E}[1/r \mid \text{tie}] = \frac{H_{b+e} - H_b}{e}.$$

The second quantity differs from the reciprocal of the average tied rank. An absent true action receives zero hit probability and reciprocal rank. If absent actions are removed before evaluation, the resulting covered-action subset and its coverage must be reported explicitly, as for the 41-action GEARS comparison. Each archived evaluator keeps its recorded tie rule.

I.2 Uncertainty Units and Conditionality

Table 9 lists the resampling unit and scope of each stored interval.

Table 9: Resampling units of the reported uncertainty. Counts and procedures describe archived analyses. Intervals in Table 1 come from a separate 20,000-draw recomputation that averages repeated seeds or contexts within each biological unit and resamples units.

Analysis	Recorded procedure	Scope of interval
Adamson semantic	5,000 paired query resamples	53 query actions; fixed libraries and scorer
Adamson GEARS	5,000 action-cluster resamples	41 covered actions; one archived export
Frangieh transfer	5,000 action-cluster resamples	115 queries grouped into 42 actions; fixed source/target processing
Replogle fusion	10,000 crossed query-by-split resamples; 20,000 action-cluster resamples	Exploratory sensitivity over five fixed splits; 2,000 rows group into 1,358 held-out actions
Srivatsan calibration	10,000 evaluation-row resamples; 20,000 drug-cluster resamples	258 query-by-seed rows group into 86 drugs, each repeated three times
McFaline full axis	2,000 action-cluster resamples	64 actions; conditional on fixed cells and controls; seed 47
McFaline final replay	Descriptive scores and hit transitions	Retrospective replay

Permutation values apply to their specified diagnostic families without global multiplicity adjustment. Effects are reported per protocol.

J EVIDENCE PACKAGE AND EXTENSIONS

`data/evidence.json` and `data/source_map.json` store performance summaries and their source locations, `data/protocol_diagnostics/` holds the tables behind Appendix C, `data/supplement_20260926/` holds the biological-unit intervals, calibration, forward-metric, fusion, and annotation tables behind Table 1, Appendix G, and Section D, and the included scripts regenerate figures and tables. `scripts/describe_mcfaline_bundle.py` produces `data/mcfaline_bundle_inventory.json` from metadata, array dimensions, and pair indices. The bundle's `full` directory names a split with 15,009 columns; the 58,347-column axis belongs to the pilot.

The supplied materials include selected code, a full-axis pilot, and historical records. The new matrices omit individual-cell counts, batch membership, weights, and prediction caches. Target contexts are A172/nintedanib, T98G/lapatinib, and U87MG/untreated; each has other training actions. Query IDs and metadata bearing labels are evaluator join keys, excluded from scoring (Section 3). The manuscript archive contains derived summaries and authoring code; input matrices remain separate.