Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Exploring Representation-Aware Model Selection for Out-of-Distribution Gene Perturbation Prediction

Aman Tarikere

amantarikere1@gmail.com

University of Mysore https://orcid.org/0009-0006-9678-9449

Research Article

Keywords: Perturb-seq, single-cell RNA sequencing, gene perturbation prediction, out-of-distribution generalization, model selection, algorithm selection, GenePT, biological representation

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11768707/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Exploring Representation-Aware Model Selection for Out-of-Distribution Gene Perturbation Prediction

Aman Tarikere
Department of Studies in Computer Science, University of Mysore, Mysuru, India

Abstract

Prediction of transcriptional responses to genetic perturbations is difficult when the target perturbation is absent from model training. A further challenge is that different predictors can behave differently on different unseen perturbations. This study therefore examines out-ofdistribution (OOD) perturbation prediction as a model-selection problem rather than assuming that one predictor is uniformly optimal. Using an Adamson Perturb-seq dataset containing 65,899 cells, 5,060 genes, and 82 perturbation conditions, we analyze 21 perturbations held out from training. Four candidate predictor configurations are evaluated: GO-only MLP, GenePT-only MLP, GenePT+GO concatenation MLP, and Gated Attention Fusion. For each perturbation, the observed candidate-model performance is used to quantify the opportunity for perturbationspecific selection. The best fixed strategy has mean R^2 =0.399650, while the oracle upper bound is 0.466234, giving an available selection gap of 0.066584. Importantly, the perturbation-level retrospective adaptive values reported in the study average to R^2 =0.421228, corresponding to a 0.021578 absolute and 5.40% observed relative improvement over the fixed strategy. Embedding Coverage is also examined as a compact representation-availability descriptor. In this exploratory analysis, it shows negative Spearman associations with OOD difficulty (ρ =-0.5472, p =0.0103) and with the observed adaptive improvement (ρ = -0.4741, p = 0.0299), with both p-values treated as nominal exploratory results, whereas the paired Wilcoxon comparison between adaptive and fixed outcomes is not significant (p =0.0638). These results provide exploratory evidence that representation coverage can be useful for characterizing OOD behavior, but they do not establish universal superiority of adaptive selection. The study is deliberately positioned as a focused single-dataset analysis and identifies the validation required for broader claims.

Keywords: Perturb-seq; single-cell RNA sequencing; gene perturbation prediction; out-of-distribution generalization; model selection; algorithm selection; GenePT; biological representation

1 Introduction

Single-cell RNA sequencing (scRNA-seq) enables transcriptomic measurements at cellular resolution. Perturb-seq combines pooled genetic perturbation with single-cell RNA sequencing, allowing many perturbations to be examined simultaneously and making it possible to study transcriptional consequences of genetic intervention (Dixit et al., 2016; Adamson et al., 2016). Computational prediction of perturbation responses is attractive because experimental enumeration of all possible perturbations and perturbation combinations is infeasible.

A particularly demanding setting is perturbation-level out-of-distribution (OOD) prediction, where a target perturbation is absent from the perturbations used to train a predictor. Knowledge-informed methods such as GEARS use biological graphs to improve generalization to novel perturbations (Roohani et al., 2024). Transfer-learning and foundation-model approaches, including Geneformer, scGPT, GenePT, and related representation-learning systems, provide additional ways of incorporating biological knowledge (Theodoris et al., 2023; Cui et al., 2024; Chen and Zou, 2025). Recent benchmarking has also emphasized that perturbation-prediction performance varies across methods, datasets, and generalization scenarios; a 2026 benchmark evaluated 27 methods across 29 datasets and multiple complementary metrics (Wei et al., 2026).

These developments motivate a complementary question. Instead of asking only which architecture is best on average, can characteristics of an unseen perturbation indicate which available predictor is likely to

be more suitable? This is closely related to the classical algorithm-selection problem, where measurable characteristics of an instance are used to select an algorithm before solving that instance (Rice, 1976; Kotthoff, 2016).

In this work, each unseen genetic perturbation is treated as an instance and each candidate geneexpression predictor as an algorithm. Seven perturbation-level descriptors are examined, including Gene Ontology (GO)-derived measures, STRING-derived measures, and Embedding Coverage. The study addresses three questions: (i) whether candidate models exhibit meaningful specialization across unseen perturbations; (ii) whether perturbation-level descriptors provide useful information for retrospective adaptive selection; and (iii) whether representation coverage is associated with OOD difficulty and adaptive benefit.

The contribution is intentionally modest and testable. We do not introduce a new expressionprediction architecture. Instead, we evaluate a decision layer that analyzes when different existing predictors may be preferable. The revised analysis also explicitly separates the oracle upper bound from the deployable setting and reports the perturbation-level aggregate consistently.

2 Related Work

2.1 Perturb-seq and OOD Perturbation Prediction

Perturb-seq established a scalable experimental framework for linking genetic perturbation to transcriptional state (Dixit et al., 2016). Adamson et al. used multiplexed single-cell CRISPR screening to investigate the unfolded protein response, forming the basis of the Adamson dataset used here (Adamson et al., 2016).

GEARS introduced graph-enhanced prediction of transcriptional outcomes and demonstrated generalization to perturbations that had not been experimentally observed during training (Roohani et al., 2024). The method combines gene and perturbation relational structure with neural prediction. This motivates the present use of biological-network descriptors, but the goal here is different: rather than constructing one graph-based predictor, we ask whether biological descriptors can help decide among predictors.

2.2 Biological Representations

Geneformer demonstrated the value of transfer learning from large-scale single-cell data for biological prediction (Theodoris et al., 2023). scGPT extended foundation-model ideas to single-cell multi-omics (Cui et al., 2024), while GenePT explored gene and cell representations derived from language-model knowledge (Chen and Zou, 2025). Biolord provides a disentanglement-based approach for representing biological variation and unknown attributes (Piran et al., 2024). In the present study, representation availability is treated as a meta-level descriptor rather than directly as the prediction target.

2.3 Benchmarking and Model Selection

Large-scale benchmarking has shown that performance rankings can depend strongly on the evaluation scenario. Wei et al. benchmarked 27 perturbation-response prediction methods across 29 datasets and multiple generalization settings, highlighting the importance of evaluating generalizability rather than relying on a single aggregate benchmark (Wei et al., 2026). The present study complements this direction by moving the unit of analysis from the method or dataset to the individual unseen perturbation.

Algorithm selection provides the conceptual basis for this formulation. Rice defined algorithm selection as choosing among algorithms using characteristics of the problem instance, while later work developed and surveyed automated selection strategies (Rice, 1976; Kotthoff, 2016). Here, the perturbation is the instance, the candidate neural predictor is the algorithm, and biological representation/network properties are the instance descriptors.

3 Materials and Methods

3.1 Dataset

The processed Adamson dataset contains 65,899 cells and 5,060 measured genes, including 24,263 control cells and 41,636 perturbed cells. The experiment contains 82 perturbation conditions. The data are represented as an AnnData object with expression values in the data matrix and perturbation information in cell-level metadata.

3.2 Preprocessing and Prediction Task

The underlying prediction pipeline applies quality-control processing, library-size normalization, and logarithmic transformation before model evaluation. The prediction target is the post-perturbation gene-expression profile. Evaluation is performed at the perturbation level rather than through a random cell-level split, because the objective is generalization to perturbations absent from training.

3.3 OOD Partition

The OOD analysis uses the dataset split in which the training and OOD perturbation sets have no overlap:

$$P_{train} \cap P_{OOD} = \emptyset. \quad (1)$$

The 21 OOD perturbations are ARHGAP22, BHLHE40, CAD, CARS, COPB1, CREB1, DAD1, DDIT3, DDRGK1, DHDDS, EIF2B2, FECH, GNPNAT1, HARS, MANF, SAMM50, SRP72, TARS, TMED10, TMED2, and TTI2. The underlying experiment explicitly verified zero overlap between training and OOD perturbation sets.

3.4 Candidate Predictors

Four candidate predictors are represented in the perturbation-level result matrix:

1. GO-only MLP: predictor using GO-derived biological information.
2. GenePT-only MLP: predictor using GenePT representation information.
3. GenePT+GO Concatenation MLP: predictor combining GenePT and GO representations by concatenation.
4. Gated Attention Fusion: predictor combining representation sources using an attention/gating mechanism.

The available experiment record confirms the four-model comparison and 84 perturbation-model observations (21 perturbations × 4 models). Detailed training hyperparameters for these four candidate models are not preserved in the current manuscript artifacts; therefore, no unrecorded values are introduced here.

3.5 Separate GEARS Robustness Benchmark

As a complementary benchmark, GEARS was evaluated separately on the Adamson dataset using a custom OOD split containing 17 unseen perturbation conditions: AARS+ctrl, BHLHE40+ctrl, COPZ1+ctrl, DAD1+ctrl, DHDDS+ctrl, EIF2B4+ctrl, GNPNAT1+ctrl, HSD17B12+ctrl, HSPA5+ctrl, HYOU1+ctrl, IDH3A+ctrl, PDIA6+ctrl, QARS+ctrl, SYVN1+ctrl, TARS+ctrl, TIMM44+ctrl, and TMEM167A+ctrl. The 17 test perturbations were kept disjoint from the training conditions, with six additional perturbation conditions reserved for validation. This benchmark is deliberately treated as an auxiliary robustness experiment rather than as part of the 21-perturbation model-selection analysis above.

GEARS used a hidden size of 64, one GO GNN layer, one gene GNN layer, a decoder hidden size of 16, no uncertainty head, batch size 32, 20 training epochs, learning rate 10^{-3}, and weight decay 5×10^{-4}. Three random seeds (42, 123, and 2026) were evaluated under the same 17-perturbation test set. The benchmark records absolute expression metrics and Top-K responsive-gene identification using the perturbation effect relative to control expression. Because the notebook records the proposed-model aggregate separately from the GEARS runs, the resulting cross-model comparison is reported as a benchmark summary and is not used to modify the four-model selector results.

3.6 Perturbation-Level Performance and Oracle

For perturbation p and model M_k, performance is summarized by coefficient of determination:

$$R^2_{p,k} = 1 - \frac{\sum_g (y_{p,g} - \hat{y}_{p,g,k})^2}{\sum_g (y_{p,g} - \bar{y}_p)^2}. \quad (2)$$

The perturbation-specific best model is

$$M^*(p) = \arg\max_k R^2_{p,k}. \quad (3)$$

The oracle strategy selects $M^*(p)$ independently for each perturbation using the observed perturbationlevel performance. It is an analysis upper bound and is not available before the outcome is observed.

The available experiment record confirms the four candidate models and the 21 × 4 perturbationmodel matrix, but the individual model-level mean R^2 summaries are not preserved in the current artifact set. The reported Best Fixed and Oracle values are therefore retained as directly auditable aggregate quantities, without reconstructing or inventing missing per-model means.

The best fixed strategy selects one candidate model for every OOD perturbation. Its mean R^2 is 0.399650, while the oracle mean is 0.466234. The resulting available selection gap is

$$G_{\text{oracle}} = 0.466234 - 0.399650 = 0.066584. \quad (4)$$

3.7 Biological Meta-Features

Seven perturbation-level descriptors are evaluated: Embedding Coverage, STRING Total, Total GO Weight, Neighbors with Embeddings, GO Total, GO Neighbors, and Valid Neighbors. The descriptor matrix contains these variables for the perturbation conditions. In the intended selection workflow, these descriptors are constructed from representation and biological-network information rather than from the observed post-perturbation expression outcome; outcome-dependent quantities are used only for retrospective evaluation of selection performance.

3.8 Embedding Coverage

Embedding Coverage quantifies the fraction of relevant biological entities associated with a perturbation for which an external embedding is available:

$$C_p = \frac{N_{\text{represented},p}}{N_{\text{relevant},p}}. \quad (5)$$

A value close to one indicates broad representation availability, whereas lower values indicate incomplete representation. Coverage is therefore interpreted as a representation-availability descriptor and not as a direct measure of biological correctness.

3.9 Retrospective Adaptive Selection Analysis

The reported retrospective adaptive outcome for perturbation p is the performance of the model selected by the perturbation-level selection analysis. Improvement relative to the fixed reference is

$$\Delta_p = R^2_{p,\hat{M}_p} - R^2_{p,M_{\text{fixed}}}. \tag{6}$$

Positive values indicate improvement and negative values indicate deterioration.

Because the current artifact set does not preserve sufficient detail to reconstruct every selectortraining step, the revised paper treats the adaptive result as a retrospective, exploratory selection analysis rather than claiming a fully reproducible deployment-time selector. The reported adaptive values are evaluated against observed perturbation-level outcomes and therefore should not be interpreted as an independently validated deployment-time selector. The perturbation-level results are the primary auditable evidence.

3.10 Feature Importance, Ablation, and Statistics

Random Forest impurity importance and permutation importance are used to assess the contribution of the seven descriptors in the recorded selection analysis. Feature ablation evaluates compact descriptor configurations. LOOCV is used for the perturbation-level classification analysis because only 21 OOD perturbations are available. Spearman correlation is used for monotonic associations, with Pearson correlation reported as a complementary linear measure. Kruskal–Wallis tests assess differences among coverage or difficulty groups, and paired Wilcoxon signed-rank tests compare perturbation-level adaptive and fixed outcomes. The nominal threshold is $\alpha = 0.05$. The reported correlation p-values are nominal exploratory values; because multiple descriptors and associations are considered, no claim of multiplicity-controlled significance is made. All inferential results are interpreted conservatively because the effective OOD sample size is only 21.

4 Results

4.1 A measurable selection opportunity

The best fixed strategy has mean $R^2 = 0.399650$, while the oracle reaches 0.466234. Thus, the candidate pool contains an observable perturbation-specific selection opportunity of 0.066584, equivalent to 16.66% relative improvement over the fixed reference. This indicates that no single candidate model is optimal for every perturbation in the evaluated 21-perturbation set.

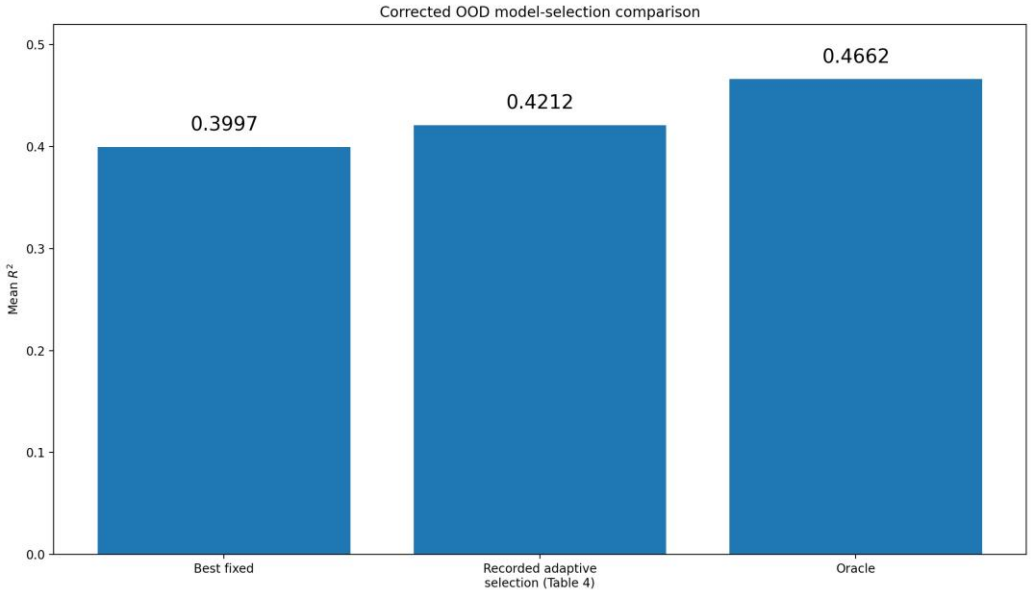


Figure 1: Corrected comparison of the fixed reference, the perturbation-level adaptive values reported in Table 4, and the oracle upper bound.

4.2 Corrected retrospective adaptive aggregate

The previous manuscript reported an adaptive mean of 0.437007. Direct recomputation from the 21 perturbation-level adaptive values shown in Table 4 gives a mean of $R^2 = 0.421228$. The revised manuscript uses this auditable value consistently. Relative to the fixed mean of 0.399650, the corrected absolute gain is 0.021578 and the relative improvement is 5.40%. Relative to the oracle opportunity of 0.066584, the perturbation-level adaptive values recover 32.41% of the available gap.

This correction is important: it prevents the aggregate claim from exceeding what is supported by the displayed perturbation-level evidence. The paired Wilcoxon test reported for the retrospective adaptive-versus-fixed perturbation outcomes gives $p = 0.063828$, so the corrected aggregate should be described as an exploratory favorable direction rather than statistically significant superiority.

Table 1: Auditable aggregate results after correcting the adaptive mean.

Strategy	Mean R^2	Gain	Relative gain	Oracle gap recovered
Best Fixed	0.399650	0.000000	0.00%	0.00%
Recorded adaptive selection	0.421228	0.021578	5.40%	32.41%
Oracle	0.466234	0.066584	16.66%	100.00%

4.3 GEARS robustness benchmark

The separate GEARS experiment provides an auxiliary check on the stability of a graph-based perturbation predictor under the custom 17-perturbation OOD split. The three directly evaluated seeds produced very similar absolute performance. The recorded R^2 values were 0.508216 (seed 42), 0.508205 (seed 123), and 0.505906 (seed 2026), while the corresponding PCC values were 0.743008, 0.743209, and 0.742751. Using the three recorded runs, the sample standard deviations are 0.0023 for MAE, 0.0006 for RMSE, 0.0013 for R^2, and 0.0002 for PCC. The resulting GEARS summary is MAE = 0.2246 ± 0.0023, RMSE = 0.4392 ± 0.0006, $R^2 = 0.5074 \pm 0.0013$, and PCC = 0.7430 ± 0.0002.

The notebook also contains a separately entered proposed-model aggregate. Because those proposed-model values are not accompanied by three corresponding prediction runs in the benchmark notebook, they are not presented here as a matched three-seed head-to-head comparison. This distinction avoids attributing unsupported replication to the proposed model. The GEARS results are therefore used as a separate auxiliary benchmark for robustness, rather than as evidence of a statistically matched superiority claim.

Table 2: Verified GEARS performance across the three recorded random seeds on the custom 17perturbation OOD setting.

Seed	MAE	RMSE	R^2	PCC
42	0.223956	0.438852	0.508216	0.743008
123	0.222683	0.438857	0.508205	0.743209
2026	0.227096	0.439882	0.505906	0.742751
Mean ± SD	0.2246 ± 0.0023	0.4392 ± 0.0006	0.5074 ± 0.0013	0.7430 ± 0.0002

Top-K responsive-gene identification was also evaluated for GEARS. Across the three seeds, Precision@20 ranged from 0.4265 to 0.4941, Precision@50 from 0.5212 to 0.5600, and Precision@100 from 0.5529 to 0.5935. These ranges indicate moderate but limited seed-to-seed variation in the ranked responsive-gene signal.

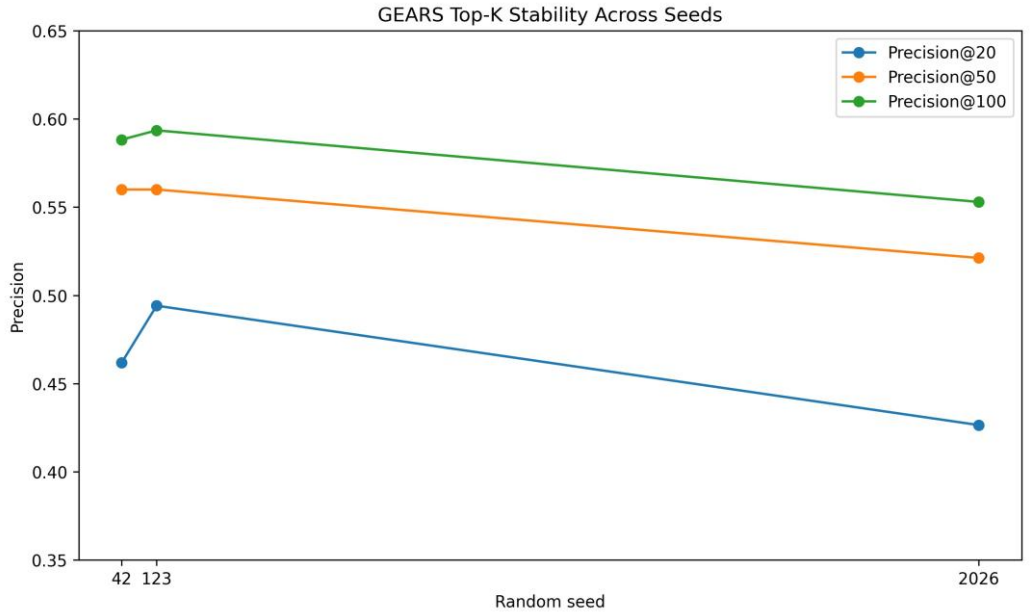


Figure 2: GEARS seed stability across the three recorded random seeds on the custom 17-perturbation OOD setting.

4.4 Feature importance

Embedding Coverage had the highest recorded Random Forest impurity-importance score (18.86%) among the seven descriptors in this exploratory analysis, followed by STRING Total (14.26%) and Total GO Weight (13.98%). In permutation analysis, Embedding Coverage has importance 0.177619, compared with 0.041905 for STRING Total and 0.006667 for Total GO Weight. The remaining recorded features have zero permutation importance in this experiment. Because the feature-importance analysis is based on only 21 OOD perturbations, these rankings should be interpreted as descriptive rather than stable estimates of population-level feature importance.

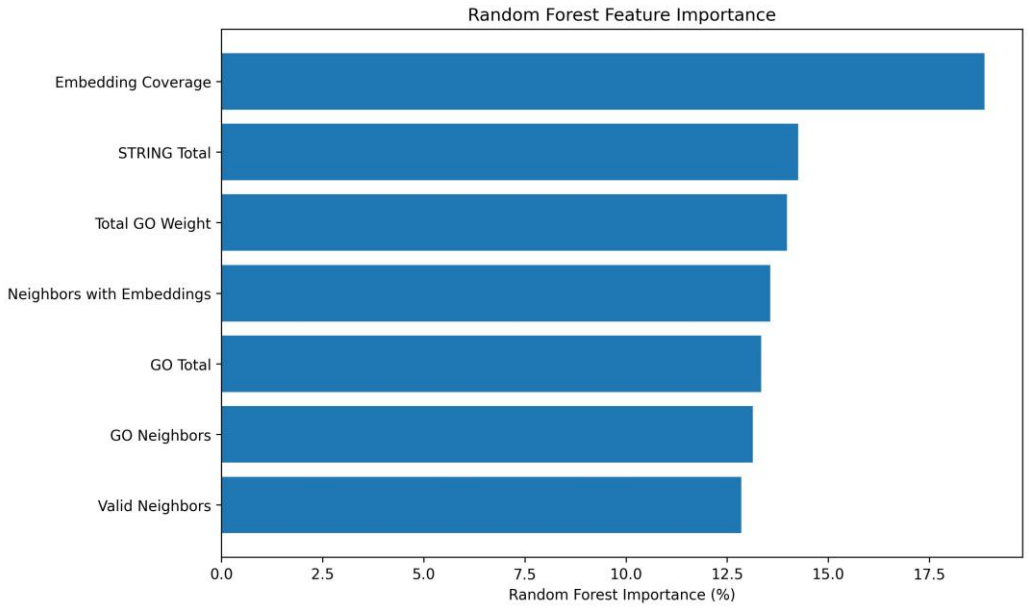


Figure 3: Recorded Random Forest impurity-based feature importance. Embedding Coverage had the highest recorded importance among the seven descriptors in this exploratory analysis.

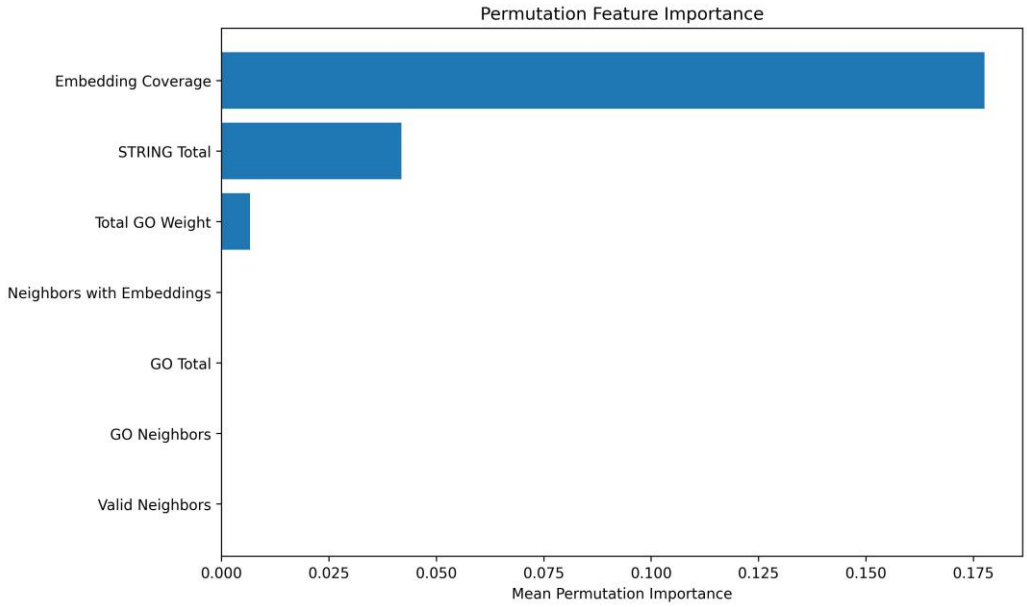


Figure 4: Recorded permutation importance of the seven perturbation-level descriptors.

4.5 Feature ablation

The original ablation table mixed aggregate adaptive scores with the corrected perturbation-level result set. To avoid reproducing that inconsistency, the revised main-text table reports the auditable LOOCV classification accuracy only. Embedding Coverage alone and the configuration without STRING Total both reach 0.380952, while the all-feature configuration reaches 0.333333. These results are exploratory because the classification sample contains only 21 perturbations.

Table 3: Feature-ablation results using the recorded LOOCV selection-classification accuracy.

Configuration	Number of features	LOOCV accuracy
Embedding Coverage Only	1	0.380952
Embedding + STRING	2	0.238095
Without Embedding Coverage	6	0.333333
STRING Total Only	1	0.333333
Without STRING Total	6	0.380952
All Features	7	0.333333

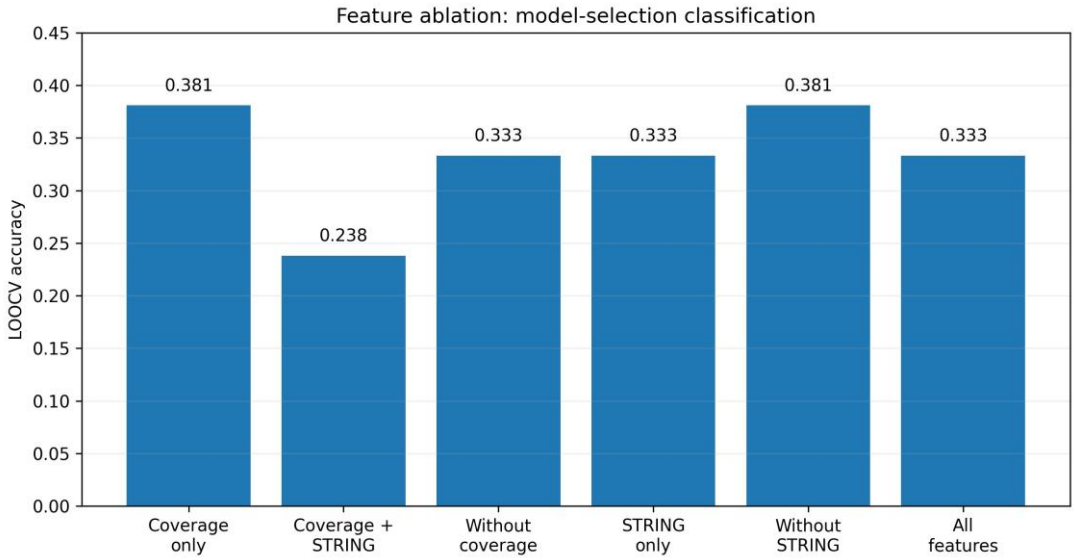


Figure 5: LOOCV classification accuracy across the recorded feature-ablation configurations.

4.6 Embedding Coverage and OOD difficulty

The 21 OOD perturbations are categorized in the recorded analysis as 10 Easy, 7 Moderate, and 4 Hard cases. Mean Embedding Coverage is 0.9973 for Easy cases and 0.9892 for Hard cases. The Moderate group is heterogeneous because it contains partially represented perturbations.

Spearman correlation between Embedding Coverage and difficulty is $\rho = -0.547171$ with nominal $p = 0.010254$. This indicates a monotonic association in the present dataset, but the result is not presented as multiplicity-controlled significance. A LOOCV classifier using only Embedding Coverage reaches 52.38% accuracy. It identifies Easy cases relatively well (recall 0.90) but does not reliably identify the Moderate class. A learned split near 0.99 coverage is therefore an empirical threshold for this dataset only, not a universal biological cutoff.

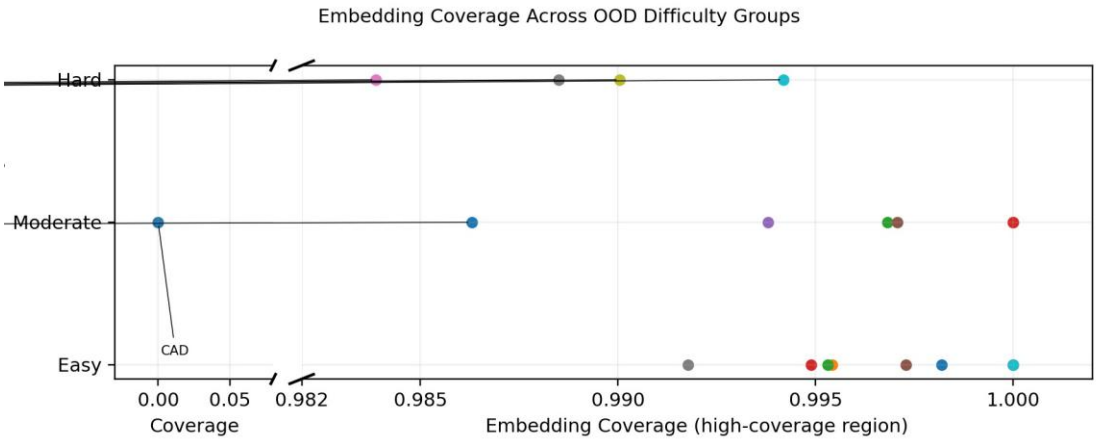


Figure 6: Embedding Coverage and the recorded OOD difficulty categories, with the high-coverage region expanded to make small differences visible.

4.7 Embedding Coverage and adaptive improvement

Using the corrected perturbation-level adaptive values, Embedding Coverage shows a negative Spearman association with adaptive improvement ($\rho = -0.474077$, nominal $p = 0.029919$). Pearson correlation is not

significant ($r = -0.085892$, $p = 0.711241$), indicating that the observed relationship is better represented as a rank-based or potentially nonlinear pattern than as a simple linear trend.

The seven low-coverage perturbations have mean improvement 0.094911, compared with -0.023554 for the medium-coverage group and -0.006623 for the high-coverage group. The Kruskal–Wallis comparison is not significant ($H = 3.488351$, $p = 0.174789$). Thus, the group-level evidence does not establish a universal coverage threshold for adaptive benefit.

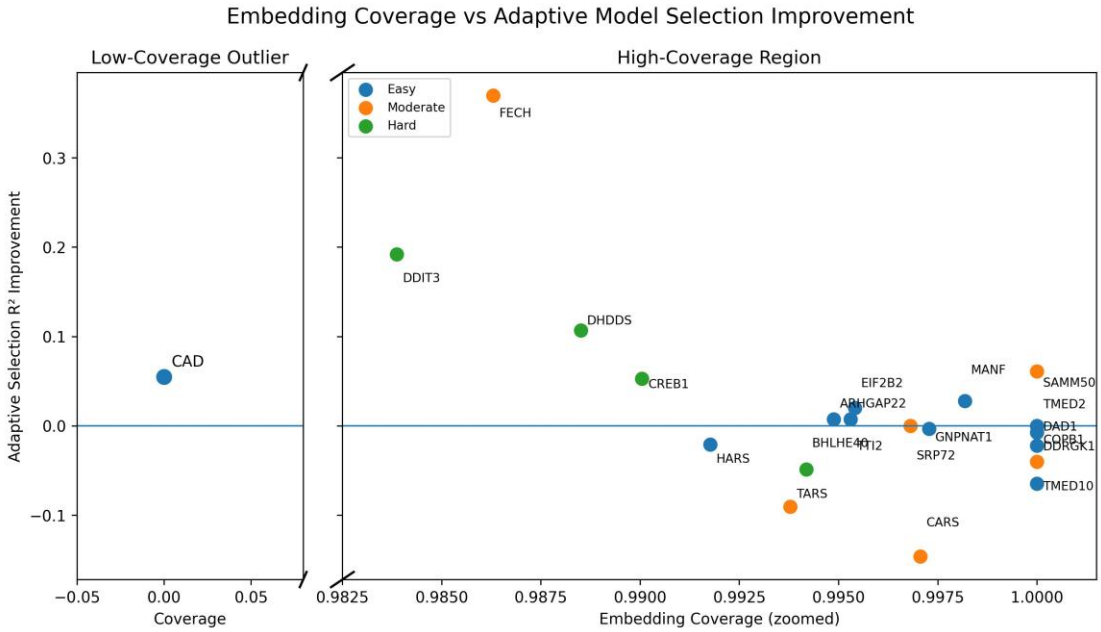


Figure 7: Embedding Coverage versus corrected perturbation-level adaptive improvement. Lowercoverage perturbations contain the largest positive gains, but variability is substantial.

4.8 Difficulty alone does not explain adaptive gain

Mean adaptive improvements are -0.005718 for Easy perturbations, 0.029686 for Moderate perturbations, and 0.075630 for Hard perturbations. The association between difficulty and adaptive improvement is not significant ($\rho = 0.265967$, $p = 0.243889$), and the Kruskal–Wallis test is also not significant ($H = 1.891302$, $p = 0.388427$). Difficulty category alone therefore does not adequately explain adaptive benefit.

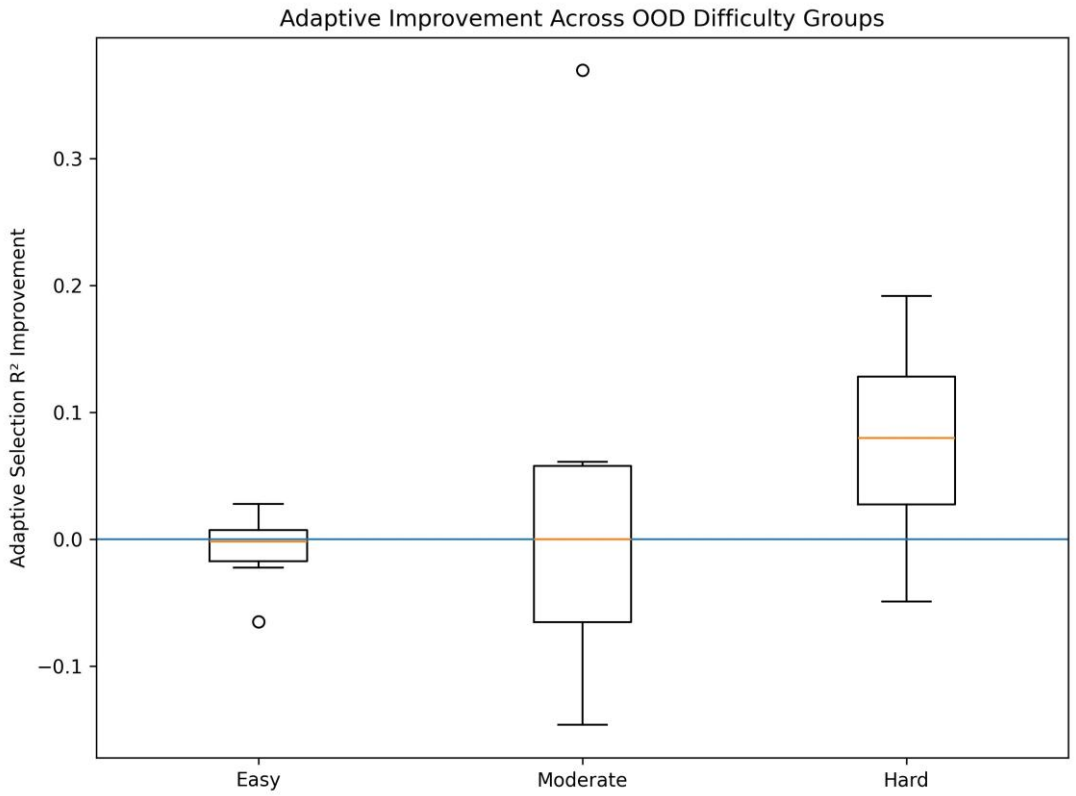


Figure 8: Corrected perturbation-level adaptive improvement summarized by OOD difficulty category.

4.9 Perturbation-level analysis

Table 4 is the primary audit table for the adaptive analysis. The largest positive gain occurs for FECH (+0.36974), followed by DDIT3 (+0.19187) and DHDDS (+0.10689). The largest negative gains occur for CARS (-0.14649), TARS (-0.09062), and TMED10 (-0.06500). This heterogeneity supports the premise that aggregate performance alone can conceal substantial perturbation-specific variation.

Table 4: Perturbation-level fixed and adaptive performance used in the corrected aggregate.

Perturbation	Coverage	Fixed R^2	Adaptive R^2	Difficulty	Δ
MANF	0.998182	0.84701	0.87482	Easy	0.02781
EIF2B2	0.995413	0.93558	0.95531	Easy	0.01973
TTI2	0.995305	0.82877	0.83602	Easy	0.00725
ARHGAP22	0.994881	0.81295	0.82016	Easy	0.00721
DDRGK1	1.000000	0.89892	0.89892	Easy	0.00000
GNPNAT1	0.997283	0.81779	0.8142	Easy	5 -

			0.00354	
TMED2	1.000000	0.85910	0.85162 - 0.00748	Easy
HARS	0.991770	0.80688	0.78602 - 0.02086	Easy
DAD1	1.000000	0.82320	0.80090 - 0.02230	Easy
TMED10	1.000000	0.87704	0.81204 - 0.06500	Easy
FECH	0.986301	0.37097	0.74071	0.36974 Moderate
SAMM50	1.000000	0.69778	0.75868	0.06090 Moderate
CAD	0.000000	0.42313	0.47769	0.05456 Moderate

Perturbation Coverage Fixed R^2			Adaptive R^2	Difficulty
SRP72	0.996815	0.75032	0.75032 - 0.00000	Moderate
COPB1	1.000000	0.71744	0.67715 - 0.04029	Moderate
TARS	0.993789	0.72047	0.62985 - 0.09062	Moderate
CARS	0.997059	0.51190	0.36541 - 0.14649	Moderate
DDIT3	0.983871	-1.89404	-1.70217 - 0.19187	Hard
DHDDS	0.988506	0.22923	0.33612 - 0.10689	Hard
CREB1	0.990050	-1.50777	-1.45497 - 0.05280	Hard

BHLHE40	0.994186	-1.13402	-1.18306 -0.04904	Hard

5 Discussion

The corrected analysis supports three cautious observations. First, candidate predictors show heterogeneous behavior across unseen perturbations. The difference between the best fixed strategy and the oracle demonstrates that model specialization creates a measurable, although bounded, selection opportunity. Second, the directly auditable adaptive values show a positive aggregate direction relative to the fixed strategy, but the paired Wilcoxon test does not reach conventional significance. Third, Embedding Coverage is repeatedly informative: it has the highest recorded individual feature-importance score and shows nominal Spearman associations with both OOD difficulty and adaptive improvement.

The coverage result should not be interpreted as evidence that lower representation coverage necessarily causes better retrospective adaptive selection. Coverage is a descriptor of representation availability, and it may correlate with other biological or computational properties of perturbations. In addition, the Pearson analysis is not significant and the coverage-group Kruskal–Wallis test is not significant. These results favor interpreting coverage as a diagnostic signal rather than as a deterministic decision rule.

The perturbation-level table illustrates why this distinction matters. FECH, DDIT3, and DHDDS obtain substantial positive gains, while CARS, TARS, and TMED10 deteriorate under the recorded adaptive choice. A practical future selector should therefore be able to abstain or revert to a fixed model when its confidence is low.

The study also complements recent broad benchmarking. Large benchmarks evaluate many methods across many datasets and scenarios (Wei et al., 2026); our narrower analysis instead asks whether information about the perturbation itself can guide selection among a small candidate pool. This is a different axis of generalization and should be evaluated alongside, rather than substituted for, large-scale benchmarking.

The auxiliary GEARS experiment provides an additional reproducibility check for the broader OOD prediction setting. Across three seeds on a separate 17-perturbation split, GEARS produced very similar absolute R^2 and PCC values and retained moderate Top-K responsive-gene precision. This supports treating GEARS as a stable separate auxiliary benchmark in this study, while the different split and evaluation purpose mean that these numbers should not be merged with the 21-perturbation selector analysis.

6 Practical Implications

A potential future workflow is: (i) identify an unseen perturbation; (ii) calculate representation and network descriptors without accessing its outcome; (iii) estimate candidate-model suitability; (iv) select a predictor when confidence is sufficient; and (v) otherwise use a fixed reference or abstain.
Such a system would treat model choice as an explicit decision problem.

At present, however, the study does not demonstrate a deployment-ready selector. The 21 OOD perturbations are also the principal limitation for perturbation-level model selection, and the exact selector-training configuration is not fully preserved in the current artifact set. These constraints are explicitly retained rather than hidden.

7 Limitations

The principal limitation is the small number of independent OOD perturbations ($n = 21$). This limits statistical power and makes feature-importance and classifier estimates sensitive to individual cases. The Hard category contains only four perturbations, so within-category analyses are especially uncertain.

Second, the study uses a single Adamson dataset and one OOD partition. Generalization to other cell types, perturbation screens, and embedding resources remains untested. Recent benchmarking demonstrates the value of multi-dataset evaluation (Wei et al., 2026); repeated OOD partitions and independent datasets are therefore required before making broad claims.

The auxiliary GEARS benchmark uses a separate 17-perturbation custom split, so it adds seed-level robustness evidence but does not constitute independent-dataset validation or a matched head-to-head evaluation of the selector.

Third, the oracle uses observed test performance and is an upper bound rather than a deployable method. Fourth, the candidate-model training configurations are not fully preserved in the manuscript artifacts. The revised manuscript therefore avoids inventing hyperparameters and recommends releasing the exact training configurations with any final submission.

Finally, the adaptive aggregate had to be corrected because the previously reported mean did not match the perturbation-level values. The revised manuscript resolves this by using the directly recomputed value $R^2 = 0.421228$ throughout the main adaptive analysis.

8 Conclusion

This study examines OOD gene perturbation prediction from a model-selection perspective. Across 21 unseen Adamson perturbations, the best fixed strategy achieves mean $R^2 = 0.399650$, while an oracle over the candidate pool reaches 0.466234, demonstrating a measurable selection opportunity. The auditable perturbation-level retrospective adaptive values average to $R^2 = 0.421228$, giving a 5.40% observed relative improvement over the fixed reference, although the paired comparison is not statistically significant at $\alpha = 0.05$.

Embedding Coverage showed the highest recorded individual feature-importance score among the evaluated descriptors and showed nominal Spearman associations with both OOD difficulty and adaptive improvement in this exploratory analysis. Nevertheless, group-level tests and the paired adaptive comparison do not establish universal superiority, and the difficulty classifier itself is only moderately accurate. The appropriate conclusion is therefore that representation coverage is a promising diagnostic for perturbation-level OOD analysis, not a validated universal model-selection rule.

The separate three-seed GEARS benchmark further shows that a graph-based baseline can be evaluated reproducibly on a custom OOD split, but it does not remove the need for repeated partitions, independent datasets, larger candidate pools, and strictly nested selector training. Such experiments would determine whether representation-aware model selection can become a general component of robust single-cell perturbation prediction.

Data and Code Availability

The study uses the Adamson Perturb-seq dataset. For reproducibility, a final public release should contain the preprocessing configuration, exact OOD perturbation list, all four candidate-model configurations, selector code, feature-construction code, and the perturbation-level result table. Dataset use should follow the original data-release terms and attribution requirements.

Declarations

Funding: No external funding was reported.

Competing interests: The author declares no competing interests.

Author contributions: The author was responsible for conceptualization, methodology, implementation, analysis, interpretation, and manuscript preparation.

References

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T.M., Lander, E.S., Weissman, J.S., Friedman, N., and Regev, A. (2016). Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7), 1853–1866.e17. doi:10.1016/j.cell.2016.11.038.

Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., Pak, R.A., Gray, A.N., Gross, C.A., Dixit, A., Parnas, O., Regev, A., and Weissman, J.S. (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7), 1867–1882.e21. doi:10.1016/j.cell.2016.11.048.

Roohani, Y., Huang, K., and Leskovec, J. (2024). Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42, 927–935. doi:10.1038/s41587-02301905-6.

Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., Al Sayed, Z.R., Hill, M.C., Mantineo, H., Brydon, E.M., Zeng, Z., Liu, X.S., and Ellinor, P.T. (2023). Transfer learning enables predictions in network biology. *Nature*, 618, 616–624. doi:10.1038/s41586-023-06139-9.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. (2024). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21, 1470–1480. doi:10.1038/s41592-024-02201-0.

Piran, Z., Cohen, N., Hoshen, Y., and Nitzan, M. (2024). Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42, 1678–1683. doi:10.1038/s41587-023-02079-x.

Chen, Y. and Zou, J. (2025). Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9, 483–493. doi:10.1038/s41551-024-01384-3.

Rice, J.R. (1976). The algorithm selection problem. *Advances in Computers*, 15, 65–118. doi:10.1016/S0065-2458(08)60520-3.

Kotthoff, L. (2016). Algorithm selection for combinatorial search problems: a survey. In *Data Mining and Constraint Programming*, 149–190. Springer. doi:10.1007/978-3-319-50137-6_7.

Wei, Z., Wang, Y., Gao, Y., Wang, S., Li, P., Si, D., Gao, Y., Wu, S., Li, D., Dong, K., Yang, X., Tang, C., Fu, S., Chen, X., Li, W., You, Y., Zhang, C., Liang, A., Chuai, G., and Liu, Q. (2026). Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 23, 451–464. doi:10.1038/s41592-025-02980-0.