SCAFFOLD: A SIMPLE GRAPH-BASED LINEAR MODEL ENABLES ACCURATE PREDICTION OF TRANSCRIPTOMIC RESPONSES TO UNSEEN PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting responses to untested gene perturbations could extend the reach of single-cell genetic screens. As recent work emphasizes large pretrained models, we investigate how existing biological relationships can guide prediction. Across two studies and four cell backgrounds, stronger Gene Ontology (GO) annotation overlap and STRING associations corresponded to more similar measured perturbation responses. We introduce scaffold, a graph-based linear predictor in which a GO annotation-overlap graph constrains gene representations while measured responses train a shared linear response map. In a unified comparison of 17 methods across nine scPerturBench datasets, scaffold ranked first on seven datasets and third on the remaining two, achieving the best mean dataset rank of 1.44 under the benchmark's six-metric ranking procedure. Controlled comparisons showed similar prediction quality with GO and STRING, while control-cell coexpression and much wider GO neighborhoods performed less well. These results demonstrate the predictive value of biological relationships within a compact linear model and identify relationship source and neighborhood selection as useful design choices alongside advances in model scale.

1 INTRODUCTION

Single-cell genetic perturbation screens reveal how perturbing individual genes reshapes cellular transcription (Dixit et al., 2016; Replogle et al., 2022). Although these experiments now reach genome scale, profiling every gene across diverse cell types and conditions remains experimentally demanding. Predicting responses to unmeasured gene perturbations could extend existing screens by learning from the effects of experimentally profiled perturbations within a given cellular context.

Recent work has increasingly used large pretrained models to predict perturbation responses. For example, scGPT learns gene and cell representations from large transcriptomic collections and applies them to tasks including perturbation prediction (Cui et al., 2024). Although larger models may learn richer representations, systematic evaluations have found that complex methods do not consistently outperform simple baselines (Ahlmann-Eltze et al., 2025; Wei et al., 2026). These findings motivate exploring whether biological knowledge and relationships can improve prediction of unmeasured perturbation responses, rather than relying solely on larger datasets and models.

Biological relationships provide a natural source of such prior knowledge. Genome-scale Perturb-seq analyses have shown that genes in the same protein complex tend to produce similar transcriptional phenotypes when perturbed (Replogle et al., 2022). Gene Ontology (GO) annotations describe shared molecular functions, biological processes and cellular components, whereas STRING integrates evidence for functional protein associations (Aleksander et al., 2023; Szklarczyk et al., 2023). Several perturbation-prediction methods have already incorporated such information. GEARS constructs a perturbation graph from GO annotation overlap, whereas TxPert integrates multiple knowledge graphs (Roohani et al., 2024; Wenkel et al., 2026). This growing use of biological knowledge motivates exploring how these relationships can be used more effectively to predict unmeasured perturbation responses.

We introduce scaffold, a simple and parameter-efficient graph-based linear model for perturbation-response prediction. With a 64-dimensional representation for each perturbation and a single shared

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.

linear response map, scaffold contains $64m + 8{,}192$ trainable parameters for m perturbations and requires no large-scale transcriptomic pretraining. The model learns response patterns from measured perturbations and uses a biological graph to inform predictions for related genes whose responses have not been measured.

Across two studies and four cell backgrounds, stronger GO annotation overlap and STRING associations were associated with more similar measured perturbation responses, supporting their use as predictive priors. In a unified comparison of 17 methods across nine scPerturBench datasets (Wei et al., 2026), scaffold ranked first on seven datasets and third on the remaining two, achieving the best mean dataset rank of 1.44. We further compared different sources of relationship information and graph hyperparameters, including neighborhood size, to assess how these choices affect prediction. Together, these results show that biological relationships can support strong prediction within a compact linear model and highlight the importance of how relational priors are constructed.

2 Related work

2.1 Predicting unseen perturbation responses

Single-cell perturbation-prediction methods differ in how they represent the initial cellular state, perturbation identity and response distribution. Conditional latent-variable and generative models learn transformations between control and perturbed states; graph-based models propagate information through gene networks; pretrained models use gene or cell representations learned from larger corpora; and linear or additive baselines estimate a lower-capacity response map. scGen (Lotfollahi et al., 2019), CPA (Lotfollahi et al., 2023), GEARS (Roohani et al., 2024), scGPT (Cui et al., 2024), GenePert (Chen & Zou, 2024) and TxPert (Wenkel et al., 2026) exemplify these families. scPerturBench and a separate evaluation of simple linear baselines have shown that performance varies across datasets and metrics, so architectural complexity alone is not a reliable proxy for generalization (Ahlmann-Eltze et al., 2025; Wei et al., 2026).

2.2 Biological knowledge as a perturbation prior

GO annotations, molecular associations and pretrained gene embeddings have previously been used to encode relationships between perturbations (Roohani et al., 2024; Chen & Zou, 2024; Wenkel et al., 2026). GEARS constructs a Pearson coexpression graph for gene embeddings and a separate perturbation graph from Jaccard overlap between gene-associated GO terms, followed by nearest-neighbor graph propagation (Roohani et al., 2024). TxPert learns perturbation embeddings by message passing across multiple knowledge graphs and combines them with the basal cell state through latent transfer (Wenkel et al., 2026). GO Jaccard, Top-k graph construction and knowledge-graph-based unseen prediction are therefore established components.

Existing graph-based predictors also combine interaction networks, functional annotations and transcriptional features for unseen gene perturbations (Dip & Zhang, 2026). scaffold uses these relationships through a normalized graph-Laplacian penalty on trainable perturbation representations, followed by a shared linear response map, rather than parameterized nonlinear graph message passing and a deep response decoder.

Replogle et al. used correlations between mean perturbation expression profiles to organize a genome-scale genotype–phenotype atlas and showed that genes in protein complexes or with stronger STRING associations tended to have more similar transcriptional phenotypes (Replogle et al., 2022). This work illustrates how measured response geometry can reveal biological organization alongside external molecular relationships.

3 Method

3.1 Task, notation and cell pairing

We predict the mean transcriptional response to a single-gene perturbation using experiments on other genes in the same cellular context. Let m be the number of perturbation genes, including training, validation and test genes, and let Ω_{train} denote the training subset. Each cell is represented

by an expression vector $x \in \mathbb{R}^G$, with $G = 5{,}000$ benchmark-provided highly variable genes from the `logNor` layer. A perturbation gene is a model input; the G expression genes are the outputs whose responses are predicted. We study a transductive setting: the identities and annotation-derived graph connections of held-out perturbations are available during representation learning, while their response measurements are withheld.

We fit $q = 64$ unwhitened principal components to control cells and training-perturbation cells, excluding validation and test cells. When this set exceeds 100,000 cells, we sample 100,000 without replacement using RandomState(42). Let $\mu \in \mathbb{R}^G$ be the fitted mean and $P \in \mathbb{R}^{q \times G}$ the projection matrix.

For each training perturbation p with at least ten cells, we independently sample $n_p = \min(n_0, n_p^{\text{pert}}, 300)$ control cells and perturbed cells without replacement, where n_0 and n_p^{pert} are their available counts. We pair the two sampled lists in their returned order using a single RandomState(42) generator, and reuse these pairs throughout training. Control cells can be reused for different perturbations. These are random reference pairs, not measurements of the same cell before and after perturbation. For pair index c within perturbation p, define

$$z^0_{c,p} = P(x^0_{c,p} - \mu), \qquad z^1_{c,p} = P(x^1_{c,p} - \mu), \qquad \Delta z_{c,p} = z^1_{c,p} - z^0_{c,p}.$$

scaffold learns these responses using a representation $e_p \in \mathbb{R}^q$ for each perturbation gene and a shared linear map. A GO graph couples the representations of functionally related genes, including genes without training responses (Fig. 1).

3.2 GO ANNOTATION-OVERLAP GRAPH

For perturbation genes i and j, scaffold defines the relationship weight

$$J_{ij} = \frac{|A_i^{\text{GO}} \cap A_j^{\text{GO}}|}{|A_i^{\text{GO}} \cup A_j^{\text{GO}}|},$$

where A_i^{GO} is the set of GO annotations for gene i; we set $J_{ij} = 0$ when the union is empty. We use direct gene–term membership, without ancestor propagation or term–term ontology edges. For the relationship analyses in Figures 2 and 4, we removed GO annotation records whose cited source study was the study under analysis before deduplicating the remaining gene–term pairs; independently supported annotations were retained. The benchmark predictor uses the supplied GO resource. The weighted adjacency matrix $A \in \mathbb{R}^{m \times m}$ is defined by

$$\mathcal{N}_i = \text{Top5}\{j \neq i : J_{ij} > 0\}, \qquad A_{ij} = J_{ij}\, \mathbf{1}\{j \in \mathcal{N}_i \ \vee \ i \in \mathcal{N}_j\}, \qquad A_{ii} = 1.$$

Thus each gene proposes at most five positive neighbors, and the final graph is the undirected union of those proposals. Writing $d_i = \sum_j A_{ij}$ and $D = \text{diag}(d_1, \ldots, d_m)$, we use the normalized Laplacian $L = I - D^{-1/2} A D^{-1/2}$.

3.3 GRAPH-REGULARIZED PERTURBATION REPRESENTATIONS

The trainable matrix $E \in \mathbb{R}^{m \times q}$ contains one perturbation representation per gene: E_{pk} is coordinate k of gene p's vector e_p. To initialize it, we decompose $A = U \Sigma V^\top$ and set $H_{pk} = U_{pk} \sqrt{\Sigma_{kk}}$ for $k \leq \min(m, q)$, with unused coordinates set to zero. For each latent coordinate, we calculate its standard deviation across the m genes:

$$\bar{H}_k = \frac{1}{m} \sum_{p=1}^{m} H_{pk}, \qquad s_k = \sqrt{\frac{1}{m} \sum_{p=1}^{m} (H_{pk} - \bar{H}_k)^2}, \qquad E^{(0)}_{pk} = 0.02\, \frac{H_{pk}}{s_k + 10^{-8}}.$$

Thus 0.02 sets the initialization scale of each coordinate; the benchmark initialization does not subtract its mean. The graph penalty is

$$\mathcal{R}_G(E) = \operatorname{tr}(E^\top L E) = \frac{1}{2}\sum_{i,j} A_{ij} \left\| \frac{e_i}{\sqrt{d_i}} - \frac{e_j}{\sqrt{d_j}} \right\|_2^2,$$

which encourages connected genes to have similar degree-normalized representations. It also constrains validation and test gene representations, although their response profiles do not enter the training loss. Thus, held-out perturbations are included in the graph during representation learning while their response measurements are withheld.

3.4 Linear response map and training

For a control-cell PCA state z_0 and perturbation representation e_p, define $h = [z_0^\top, e_p^\top]^\top \in \mathbb{R}^{2q}$. The shared bias-free map $W = [W_z\ W_e] \in \mathbb{R}^{q\times 2q}$, with $W_z, W_e \in \mathbb{R}^{q\times q}$, predicts

$$\widehat{\Delta z}_p(z_0) = Wh = W_z z_0 + W_e e_p, \qquad \hat{z}_1 = (I + W_z)z_0 + W_e e_p.$$

The entries of W are initialized uniformly in $[-(2q)^{-1/2}, (2q)^{-1/2}]$, using the default bias-free PyTorch linear-layer initialization. Training jointly updates the linear map and perturbation representations using the response pairs $\mathcal{P}_{\mathrm{tr}}$

$$\mathcal{L}_{\mathrm{resp}} = \frac{1}{q|\mathcal{P}_{\mathrm{tr}}|} \sum_{(c,p)\in\mathcal{P}_{\mathrm{tr}}} \left\| W_z z_{c,p}^0 + W_e e_p - \Delta z_{c,p} \right\|_2^2, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{resp}} + \lambda_G \mathcal{R}_G(E), \quad \lambda_G = 0.01.$$

At each optimization step, the response loss averages over the current minibatch and its q PCA coordinates and is combined with the full graph penalty using $\lambda_G = 0.01$. The graph penalty is not divided by the number of batches or graph nodes. We use AdamW (learning rate 10^{-3}, weight decay 10^{-5}, batch size 256, 50 epochs), cosine annealing and gradient clipping at 1.0. Weight decay applies at every step to W and all representations in E, including held-out genes. The checkpoint with the lowest validation mean-expression MSE is retained. Test responses are excluded from PCA fitting, training and checkpoint selection.

3.5 Inference

For an unseen target p and fixed control-cell set $\mathcal{C}_0$, the reported mean expression prediction is

$$\widehat{x}_p = \frac{1}{|\mathcal{C}_0|} \sum_{c\in\mathcal{C}_0} \left[P^\top \left((I + W_z) z_c^0 + W_e e_p \right) + \mu \right]_+,$$

where $[\cdot]_+$ clips expression values at zero. We sample up to 2,000 control cells without replacement using RandomState(42) and use the same set for all test targets. The evaluator uses the individual predicted cell profiles for distribution metrics and their mean for mean-expression metrics.

3.6 Benchmark and target-level evaluation

We use the three-seed scPerturBench splits (Wei et al., 2026). Seed-1 train/validation/test perturbation counts are: Replogle K562 essential (1,162/130/323), Replogle RPE1 essential (956/107/266), Tian Inhibition (129/15/36), Frangieh (109/13/31), Replogle exp7 (71/8/20), Tian Activation (64/8/19), Adamson (52/6/15), Replogle exp8 (42/5/12) and Papalexi (14/2/5). We compare scaffold and TxPert-public with the 15 released benchmark methods, including generative, graph-based, pretrained, embedding and linear predictors.

The benchmark evaluates the 100 benchmark-defined differentially expressed output genes for each target using PCC-delta, MSE, energy distance, Wasserstein distance, symmetric KL divergence and Common-DEGs (top-20 DE-gene overlap). For PCC-delta, both predicted and observed perturbation profiles are first expressed relative to the control profile and then compared by Pearson correlation;

larger PCC-delta and Common-DEGs are better, whereas smaller distances and MSE are better. For each target and seed, methods are ranked on each metric. The six ranks are averaged and these averages are ranked again to obtain a target rank. Target ranks are then averaged over all target–seed observations within each dataset, and these averages are ranked to obtain dataset ranks. The reported overall score is the mean of the nine dataset ranks; thus datasets have equal weight regardless of target count.

To compare methods across levels of functional similarity (Fig. 4), we calculate target j's mean positive GO edge weight to training perturbations:

$$R_j = \begin{cases} \dfrac{\sum_{i \in \Omega_{\text{train}}} A_{ji}\, \mathbf{1}[A_{ji} > 0]}{\sum_{i \in \Omega_{\text{train}}} \mathbf{1}[A_{ji} > 0]}, & \text{if a measured neighbor exists,} \\ 0, & \text{otherwise.} \end{cases}$$

Here A is the retained GO graph, and R_j depends only on annotations and training-set membership. We sort common seed-1 targets by R_j into five approximately equal-sized bins per dataset and use identical bins for all six displayed methods. Frangieh, Tian Inhibition, RPE1 essential and K562 essential provide 31, 36, 266 and 323 such targets, respectively, with graph coverage and predictions available for every displayed method.

Targets with $R_j = 0$ have no direct training neighbor, although some remain connected to training perturbations through multi-hop paths. Such cases were retained in the evaluation.

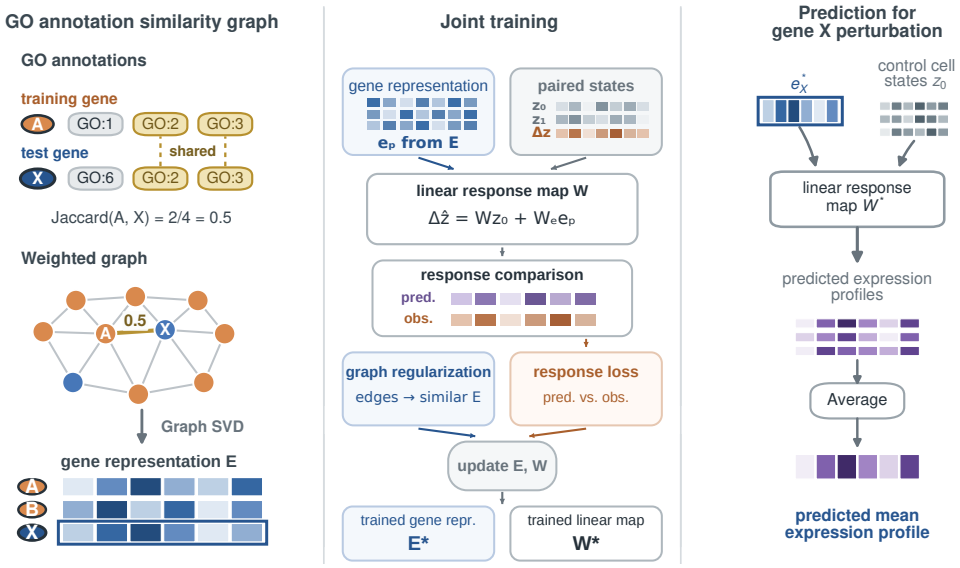

Figure 1: **scaffold uses a GO annotation overlap graph for unseen perturbation prediction.** GO annotation overlap defines the weighted graph and graph-SVD coordinates. Joint linear training combines response error on measured perturbations with a graph regularization term that also updates representations of connected held-out target genes. At inference, the held-out target representation and control-cell PCA state produce a mean expression prediction. Colours and values are schematic.

4 Experiments and Results

4.1 Gene relationships correspond to perturbation-response similarity

Previous Perturb-seq analyses showed that response-profile correlations can recover functional relationships (Replogle et al., 2022). We asked whether GO annotation overlap and STRING association scores correspond to measured response similarity. We examined all unordered measured-

gene pairs in Replogle K562 essential, Replogle RPE1 essential, Nadig HepG2 and Nadig Jurkat (GEO: GSE264667; Nadig et al., 2025) (Fig. 2).

For each perturbation p, we first subtract the control mean from its mean expression profile, $r_p = \bar{x}_p - \bar{x}_{0,p}$. Replogle datasets use the overall control mean; Nadig datasets use control means weighted by the batch composition of each perturbation. Within each cell background, we then subtract the average response across its M measured perturbations:

$$v_p = r_p - \frac{1}{M}\sum_{u=1}^{M} r_u, \qquad \rho_{ij} = \mathrm{corr}_g(v_i, v_j).$$

The correlation is calculated across expression genes g. Subtracting the shared response prevents a transcriptional change common to many perturbations from dominating the comparison. We grouped all 7,909,817 gene pairs by GO Jaccard or STRING score, without Top-5 filtering, and estimated uncertainty using delete-one-gene jackknife intervals.

Gene pairs with GO Jaccard overlap of at least 0.6 had mean response correlations of 0.436 in K562, 0.544 in RPE1, 0.387 in HepG2 and 0.282 in Jurkat. Each value averages the correlations between measured response profiles within the high-overlap group. Pairs with weak GO overlap had response correlations near zero. Likewise, pairs with STRING scores of at least 0.9 had mean response correlations 0.199–0.266 higher than pairs with no recorded association across the four backgrounds. The substantially higher correlations among strongly connected genes show that GO and STRING relationships capture shared transcriptional consequences of gene disruption, even after removing the common response. This supports using functional relationships to connect measured perturbations to untested genes.

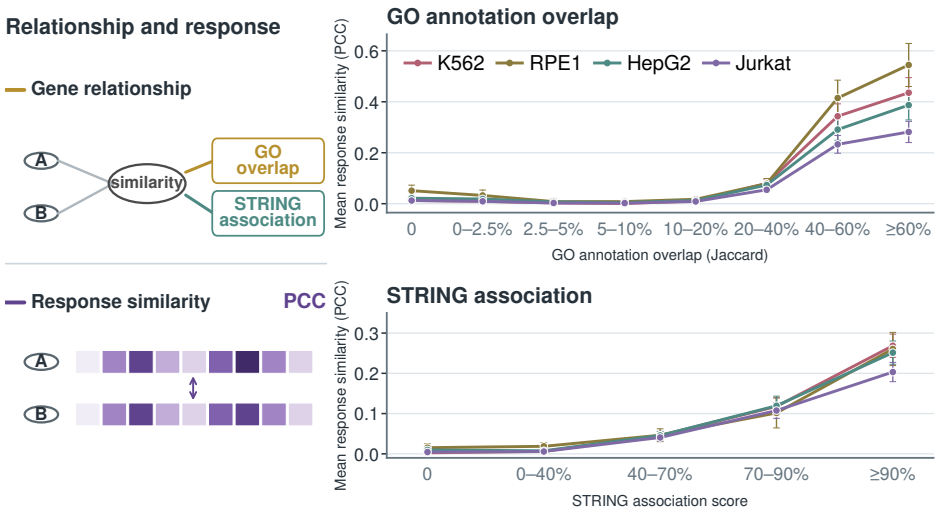

Figure 2: **GO and STRING scores track measured perturbation-response similarity.** Left, GO annotation overlap, STRING association score and measured response PCC for one gene pair. Top right, response PCC across GO Jaccard intervals; bottom right, the same analysis across STRING association-score intervals, in four cell backgrounds. Points show interval means; vertical error bars show delete-one-gene jackknife 95% intervals. All 7,909,817 measured gene pairs were included without Top-5 filtering or model predictions.

4.2 SCAFFOLD LEADS THE AGGREGATE COMPARISON ACROSS NINE DATASETS

We evaluated scaffold using the scPerturBench splits and Top-100 evaluation interface across nine datasets (Wei et al., 2026) (Fig. 3). As a check of the ranking implementation, we reproduced all 135 published dataset ranks (15 released methods $\times$ nine datasets) from the released target-level scores. We then recomputed ranks for the full 17-method pool including scaffold and TxPert-public. Figure 3 displays six representative methods, while all 17 contribute to the ranks.

scaffold ranked first in seven of the nine datasets and third in the remaining two, giving a mean dataset rank of 1.44 across the six benchmark metrics and first place in the complete 17-method comparison. GenePert and TxPert-public had mean ranks of 4.78 and 5.22, respectively. Relative to the four displayed released comparators—GenePert, scouter, scGPT and linearModel—two-sided exact Wilcoxon signed-rank tests on the nine paired dataset ranks gave Holm-adjusted $P = 0.015625$ for each comparison.

Against TxPert-public, we compared matched targets using the symmetric relative MSE difference $2(\mathrm{MSE}_{\mathrm{TxPert}} - \mathrm{MSE}_{\mathrm{scaffold}})/(\mathrm{MSE}_{\mathrm{TxPert}} + \mathrm{MSE}_{\mathrm{scaffold}})$ and the PCC-delta difference (scaffold minus TxPert-public). Averaging within each dataset and then equally across datasets gave an MSE effect of 0.067 (95% bootstrap interval, 0.030–0.106) and a PCC-delta difference of 0.009 (-0.008–0.027). Thus the advantage over TxPert-public was clearer in expression error than in correlation. Together with the aggregate ranks, these results show that the shared linear map yields strong predictions across datasets, while two predictors using biological relationships recover similar response patterns.

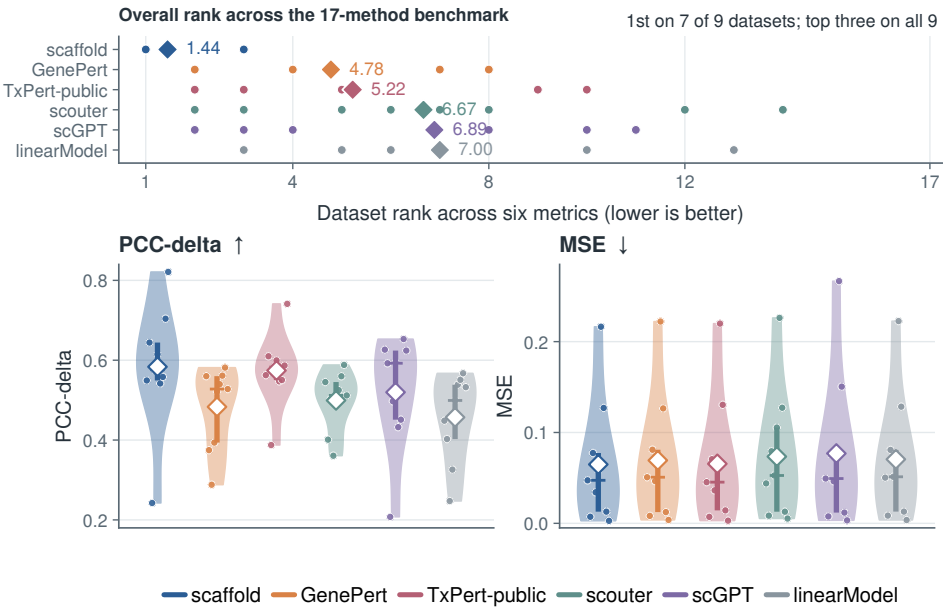

Figure 3: **Consistent predictive performance across nine perturbation-prediction datasets.** Top, dataset ranks computed with the scPerturBench rule in the complete 17-method pool; six methods are displayed for legibility. Each point is one dataset rank, open diamonds mark the nine-dataset mean rank, and adjacent values report that mean. Bottom, PCC-delta and MSE across the nine datasets; violin shapes show the empirical distribution across datasets, filled points show individual datasets, horizontal marks show medians, vertical bars show interquartile ranges, and open diamonds show means. scaffold ranked first on seven datasets and third on two. The displayed summaries use the same nine-dataset evaluation interface throughout.

A separate mean-only comparator that averages GO training-neighbor responses showed a trade-off: scaffold had higher mean PCC-delta, but neighbor averaging had lower mean MSE (Supplementary Table S7). This comparator is not part of the 17-method six-metric ranking.

4.3 GO SIMILARITY AND ABSOLUTE PREDICTION QUALITY

We compared scaffold and TxPert-public with linearModel, GEARS, GenePert and scGPT in the five GO-similarity bins defined above (Fig. 4). In K562, scaffold and TxPert-public both improved from PCC-delta values near 0.41 in the lowest-similarity bin to 0.783 and 0.796 in the highest bin. The other four methods reached 0.382–0.575 in that highest bin. RPE1 showed a similar separation:

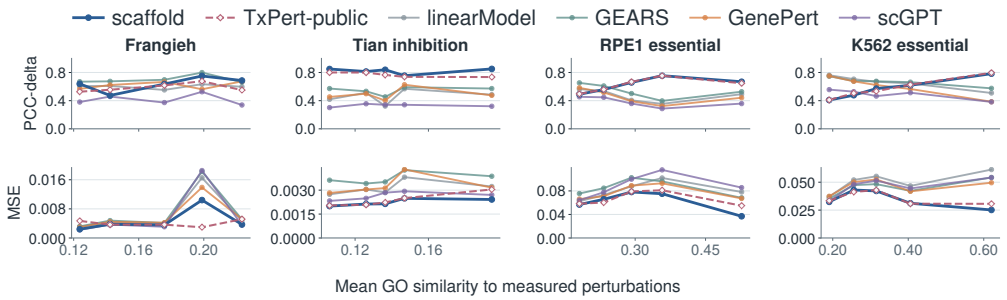

Figure 4: **Prediction quality across levels of GO similarity to measured perturbations.** Frangieh, Tian Inhibition, RPE1 essential and K562 essential contain 31, 36, 266 and 323 common seed-1 test targets, respectively. Targets are grouped into five approximately equal-sized bins by mean retained GO similarity to training perturbations. Lines show mean target-level PCC-delta (top) and MSE (bottom) for six methods on identical targets; the horizontal coordinate is the median GO similarity in each bin. Grouping uses annotations and training-set membership only.

scaffold and TxPert-public reached 0.669 and 0.647 in the highest bin, compared with 0.359–0.527 for the other methods. Both relationship-based predictors therefore showed an advantage for targets with strong functional connections to training perturbations, despite their different model architectures.

The MSE results provide a complementary comparison of the relationship-based predictors. In the highest-similarity bins, scaffold achieved MSEs of 0.025 in K562 and 0.037 in RPE1, compared with 0.031 and 0.055 for TxPert-public; both methods also had lower MSE than the other four methods in these bins. On Tian Inhibition, scaffold had the highest PCC-delta in all five bins, ranging from 0.755 to 0.851, followed by TxPert-public, which ranged from 0.735 to 0.801. Frangieh showed a different ordering: GEARS had the highest PCC-delta in the first four bins, whereas scaffold ranked first in the last bin. Thus, scaffold and TxPert-public showed their clearest shared advantage in K562, RPE1 and Tian Inhibition, but the curves do not establish a universally monotonic relationship between GO similarity and prediction quality.

4.4 Ablation of prior source and GO neighborhood size

We conducted an ablation study to evaluate how the prior supplied to scaffold affects prediction (Fig. 5). We compared GO, STRING, control-cell coexpression and scGPT gene-vector cosine graphs, each retaining Top-5 neighbors, plus a variant using frozen scGPT vectors directly without a graph penalty. The latter is an input variant of scaffold, not the full scGPT predictor. All variants use 64-dimensional perturbation inputs, the same response map, splits, three seeds and evaluation procedure. For this matched ablation, graph-SVD coordinates are centered before scaling, $E_{pk}^{(0)} = 0.02(H_{pk} - \bar{H}_k)/(s_k + 10^{-8})$, and trainable perturbation representations have zero weight decay; the map retains weight decay 10^{-5}. These settings are shared across the graph variants.

GO and STRING gave nearly identical mean PCC-delta across the nine datasets (0.586 and 0.586). Coexpression was lower at 0.569, a difference of -0.0168 from GO (paired target-bootstrap $P = 0.0056$). The scGPT graph and direct-vector variants averaged 0.576 and 0.569, respectively, without a clear statistical separation from GO. These comparisons indicate that functional annotations and protein associations provide useful priors for perturbation effects; correlations in unperturbed cells gave weaker predictions under the same model.

To test neighborhood size, we retained the GO source and changed Top-5 to Top-50 or Top-500. Mean PCC-delta changed little at Top-50 (0.580) but fell to 0.569 at Top-500 (difference -0.0169, $P = 0.0002$). Coexpression and Top-500 remained lower after Holm correction across the six comparisons with GO Top-5 (adjusted $P = 0.0280$ and 0.0012). Under this matched training protocol, the compact GO Top-5 construction therefore outperformed the broader alternatives. Because changing k also changes the normalized Laplacian and graph-SVD initialization, this comparison evaluates the resulting graph construction as a whole rather than isolating edge content at a fixed regularization operator.

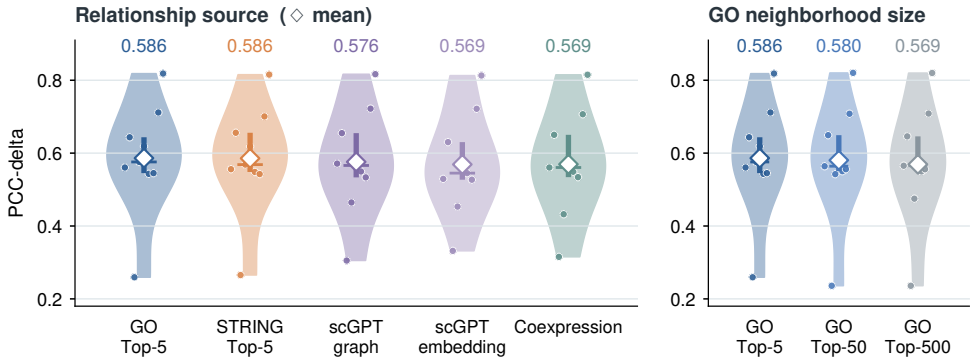

Figure 5: **Ablation of prior source and neighborhood size in scaffold.** Each point represents one of nine datasets, averaging Top-100 PCC-delta over targets within seed and then over three seeds; violins show the distribution of dataset means. Internal horizontal marks and thick vertical bars indicate medians and interquartile ranges. Left, five prior variants: GO Top-5, STRING Top-5, scGPT-cosine Top-5, frozen scGPT vectors without a graph, and control-cell coexpression Top-5. Right, GO Top-5, Top-50 and Top-500. Open diamonds and labels give nine-dataset means. Comparisons use a paired target bootstrap within datasets, retaining seed observations together, with GO Top-5 as reference.

Matched component controls, training-connectivity coverage and graph-label randomization are reported in Supplementary Tables S8–S10. Removing the graph penalty reduced PCC-delta and increased MSE in the three datasets tested, whereas fixed SVD coordinates retained substantial predictive information across nine datasets.

5 Discussion

scaffold shows that functional relationships can support accurate prediction of untested gene perturbations through a shared linear response map. Strong GO and STRING relationships correspond to similar measured responses. In the benchmark comparison, the GO-based scaffold model achieved the best aggregate rank across nine datasets; in the matched prior-source ablation, GO and STRING produced similar PCC-delta. The advantage shared by scaffold and TxPert-public in several target-level comparisons further supports biological relationships as a useful basis for generalization. The prior-source ablation shows that which relationships are retained matters: control-cell coexpression and much broader GO neighborhoods were less effective.

Our evaluation focuses on single-gene perturbations within the measured cellular contexts. The results therefore support compact biological priors for this setting; extending the framework to new cellular contexts or higher-order perturbations would require context-specific relationships or interaction terms. The model is most effective for perturbation-group mean expression, whereas its current additive formulation captures cell-to-cell variation less strongly (Supplementary Table S1). Within the evaluated task, these results support improving the use of existing functional knowledge alongside continued development of larger pretrained models.

AI Assistance

During the preparation of this manuscript, the authors used large language models to assist with language polishing, literature retrieval, research ideation and execution, and drafting of selected sections. The authors reviewed and verified all generated content and remain responsible for the manuscript, analyses and conclusions.

REFERENCES

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

Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The Gene Ontology knowledgebase in 2023. *GENETICS*, 224(1), 2023. doi: 10.1093/genetics/iyad031. URL https://doi.org/10.1093/genetics/iyad031.

Yiqun Chen and James Zou. GenePert: Leveraging GenePT Embeddings for Gene Perturbation Prediction. *bioRxiv*, 2024. doi: 10.1101/2024.10.27.620513. URL https://doi.org/10.1101/2024.10.27.620513.

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

Sajib Acharjee Dip and Liqing Zhang. Predicting Unseen Gene Perturbation Response Using Graph Neural Networks with Biological Priors. *bioRxiv*, 2026. doi: 10.64898/2026.03.23.713780. URL https://doi.org/10.64898/2026.03.23.713780.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038. URL https://doi.org/10.1016/j.cell.2016.11.038.

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

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517. URL https://doi.org/10.15252/msb.202211517.

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

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

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

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

Zhiting Wei, Yiheng Wang, Yicheng Gao, Shuguang Wang, Ping Li, Duanmiao Si, Yuli Gao, Siqi Wu, Danlu Li, Kejing Dong, Xingbo Yang, Chen Tang, Shaliu Fu, Xiaohan Chen, Wannian Li, Yuzhou You, Chen Zhang, Aibin Liang, Guohui Chuai, and Qi Liu. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 23(2):451–464, 2026. doi: 10.1038/s41592-025-02980-0. URL https://doi.org/10.1038/s41592-025-02980-0.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T. Whitfield, Craig Russell, Yassir El Mesbahi, Jiarui Ding, Marta M. Fay, Berton Earnshaw, Emmanuel Noutahi, and Alisandra K. Denton. TxPert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03113-4. URL https://doi.org/10.1038/s41587-026-03113-4.