EFFECTS OF DISTANCE METRICS AND SCALING ON THE PERTURBATION DISCRIMINATION SCORE

Qiyuan Liu
Department of Statistics
University of Chicago
Chicago, IL, USA

Qirui Zhang
Department of Biomedical Data Science
Dartmouth College
Hanover, NH, USA

Jinhong Du*
Institute of Data Science
The University of Hong Kong
Hong Kong SAR, China
Department of Statistics and Actuarial Science
The University of Hong Kong
Hong Kong SAR, China

Siming Zhao*
Department of Biomedical Data Science
Dartmouth College
Hanover, NH, USA
Dartmouth Cancer Center
Lebanon, NH, USA

Jingshu Wang*
Department of Statistics
University of Chicago
Chicago, IL, USA

ABSTRACT

Predicting gene-expression responses to genetic perturbations is a central task in functional genomics and a key application of generative models trained on single-cell CRISPR perturbation data. The Perturbation Discrimination Score (PDS) is increasingly used to evaluate whether predicted perturbation effects remain distinguishable, including in Systema and the Virtual Cell Challenge. However, its behavior in high-dimensional gene-expression settings has not been examined in detail. We show that PDS is highly sensitive to the choice of similarity or distance measure and to the scale of predicted effects. Analysis of observed perturbation responses reveals that ℓ_1 and ℓ_2-based PDS behave very differently from cosine-based measures, even after norm matching. We provide geometric insight and discuss implications for future discrimination-based evaluation metrics.

1 INTRODUCTION

Predicting transcriptional responses to genetic perturbations is a central goal of functional genomics. An important question is not only how close the predictions are on average, but also whether predicted responses for different perturbations remain distinguishable. The Perturbation Discrimination Score (PDS) formalizes this idea by evaluating whether a predicted perturbation-effect vector $\hat{\Delta}_i$ is closest to its true counterpart Δ_i for perturbation i within a collection of perturbations. Metrics of this form have been used in several recent benchmarking efforts. For example, the Systema framework (Viñas Torné et al., 2025) evaluates discriminability using centroid accuracy, which corresponds to PDS defined with the Euclidean (ℓ_2) distance. The PerturbBench framework (Wu et al., 2024) develops a class of rank-based metrics including PDS under generic distance definitions, and the Virtual Cell Challenge (VCC; Roohani et al. (2025)) further adopts this PDS metric with the Manhattan (ℓ_1) distance for large-scale benchmarking.

Although discrimination-based metrics are motivated by clear biological considerations and are increasingly used in benchmarking studies, they remain relatively new in prediction settings, and their

*Corr.: jingshuw@uchicago.edu, siming.zhao@dartmouth.edu, jinhongd@hku.hk

behavior under different distance or similarity definitions has not been systematically examined. Understanding these properties is important for interpreting benchmark results and for designing robust metrics for future perturbation-prediction studies. In our analysis, we find that the behavior of PDS depends strongly on the choice of similarity or distance measure. In particular, PDS defined using ℓ_1/ℓ_2 distance is highly sensitive to the scale of predicted effects, whereas cosine-based versions are not, leading to interactions between metric choice and scale that are not always intuitive.

In this work we examine how PDS behaves under different distance measures, with a focus on how directional agreement, magnitude, and scaling together influence discriminability. Our goal is to clarify how PDS behaves in practice and to provide insight that may guide the development of future discrimination-based evaluation metrics.

2 Variation of PDS across distance metrics

At its core, for each perturbation i, PDS ranks the true distance $d(\hat{\Delta}_i, \Delta_i)$ among the set of distances $d(\hat{\Delta}_i, \Delta_{i'})$ for all perturbation i', and linearly rescales that rank to the interval $[0, 1]$. A perfect match gives PDS $= 1$ and random guessing gives roughly 0.5, with the worst case giving 0.

Systema used the ℓ_2 distance:

$$d_{\ell_2}(\hat{\Delta}_i, \Delta_{i'}) = \|\hat{\Delta}_i - \Delta_{i'}\|_2 = \sqrt{\sum_j (\hat{\Delta}_{ij} - \Delta_{i'j})^2},$$

and VCC used the ℓ_1 distance:

$$d_{\ell_1}(\hat{\Delta}_i, \Delta_{i'}) = \|\hat{\Delta}_i - \Delta_{i'}\|_1 = \sum_j |\hat{\Delta}_{ij} - \Delta_{i'j}|.$$

To examine how the choice of distance affects PDS in practice, we evaluate three representative prediction strategies: a mean baseline, GEARS (Roohani et al., 2024), and a cross-cell-line transfer prediction based on perturbation effects measured in the K562 cell line (Replogle et al., 2022). To make the PDS evaluation comparable across methods, we restrict the evaluation to the 115 perturbations and 7,581 genes shared between the VCC hESC dataset and the K562 screen. While the mean baseline and GEARS model are trained using the full VCC training data, PDS is computed only on this shared subset. For the mean baseline and GEARS model, we split the VCC hESC dataset into training and test sets, train the models on the training data, and evaluate PDS using predictions on the held-out test perturbations. The mean baseline predicts the average perturbation effect estimated from the training data, while GEARS is trained on the same data to generate predictions for the test perturbations. For the transfer strategy, we use the genome-wide CRISPR screen in the K562 cell line to construct predictions for the corresponding perturbations in the VCC dataset. For each perturbation i, the predicted $\hat{\Delta}_i$ is defined as the mean difference between perturbed and control cells in the K562 data after standard preprocessing, while the corresponding "true" effect Δ_i is computed analogously using the VCC hESC data under the same preprocessing pipeline.

We then compute PDS under four definitions of distance or dissimilarity:

- **(1)** ℓ_1 distance,
- **(2)** ℓ_2 distance,
- **(3)** cosine dissimilarity $1 - \cos(\hat{\Delta}_i, \Delta_{i'})$,
- **(4)** sign-based cosine dissimilarity $1 - \cos(\text{sign}(\hat{\Delta}_i), \text{sign}(\Delta_{i'}))$.

As shown in Figure 1, the resulting PDS values can vary substantially across distance metrics. For prediction strategies with limited predictive and discriminative signals, such as the mean baseline and GEARS, all similarity measures yield PDS values close to 0.5, which is consistent with random guessing. However, when using the cross-cell transfer based on the Replogle data, the choice of similarity measure leads to markedly different PDS values. In this case, cosine-based similarity produces substantially higher discrimination scores, with PDS approaching values near 0.8, while ℓ_1 and ℓ_2 distances yield values only slightly above 0.5.

This contrast raises a natural question: how can the same set of predictions appear almost random under ℓ_1/ℓ_2 distances yet become highly discriminative when evaluated with cosine-based measures?

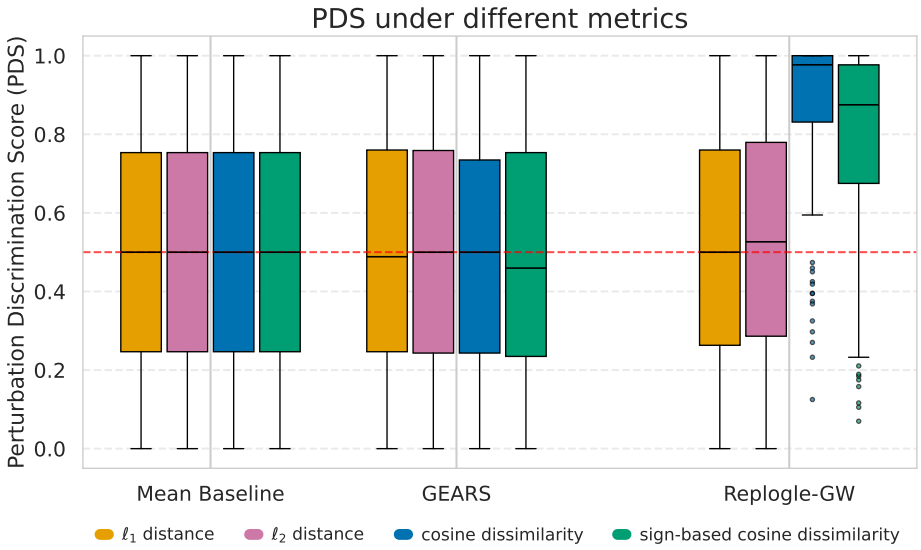

Figure 1: **Effect of distance metrics on PDS across prediction methods.** Boxplots of PDS computed using ℓ_1 distance, ℓ_2 distance, cosine dissimilarity, and sign-based cosine dissimilarity for three prediction strategies: training mean baseline, GEARS, and K562 cross-cell-line transfer. Each point corresponds to one perturbation, with the target gene excluded from the perturbation-effect vectors.

3 Norm matching does not rescue ℓ_1/ℓ_2-based PDS

One possible explanation for the discrepancy observed for the cross-cell-line transfer predictions in Figure 1 is that the predicted perturbation effects may have a different global scale from the observed effects when evaluated under ℓ_1/ℓ_2 distance. To test this possibility, we considered rescaled predictions

$$\tilde{\Delta}_i = c_i \hat{\Delta}_i,$$

where c_i is chosen such that $\|\tilde{\Delta}_i\|_1 = \|\Delta_i\|_1$ (for ℓ_1) or $\|\tilde{\Delta}_i\|_2 = \|\Delta_i\|_2$ (for ℓ_2). This enforces exact norm matching between each prediction and its true perturbation effect.

However, as shown in Figure 2, even after this normalization, the resulting mean PDS scores under ℓ_1 and ℓ_2 remain close to 0.5 across the prediction strategies considered. Thus, simply correcting the global scale does not substantially improve discriminability when PDS is defined using magnitude-sensitive metrics.

The underlying reason is geometric. Consider the ℓ_2 case. Even when the rescaled prediction has the correct length and is more directionally aligned with the true effect Δ_i than with any other perturbation, the ℓ_2 distance to another perturbation $\Delta_{i'}$ may still be smaller if $\Delta_{i'}$ has a sufficiently short norm. Figure 3 illustrates this in two dimensions: although $\Delta_{i'}$ is orthogonal to $\tilde{\Delta}_i$, its shorter length creates a "red" segment that lies entirely within the circle of radius $d_{\ell_2}(\tilde{\Delta}_i, \Delta_i)$. Points along this segment remain closer to the prediction in Euclidean distance despite having a larger angular deviation from it.

In this 2D setting, even if $\tilde{\Delta}_i$ is orthogonal to $\Delta_{i'}$, we can guarantee that

$$d_{\ell_2}(\tilde{\Delta}_i, \Delta_i) < d_{\ell_2}(\tilde{\Delta}_i, \Delta_{i'})$$

for every $\Delta_{i'}$ (i.e., the circle around $\tilde{\Delta}_i$ does not intersect the ray in the direction of $\Delta_{i'}$) only when

$$\cos\big(\tilde{\Delta}_i, \Delta_i\big) > \cos(60^\circ) = 0.5.$$

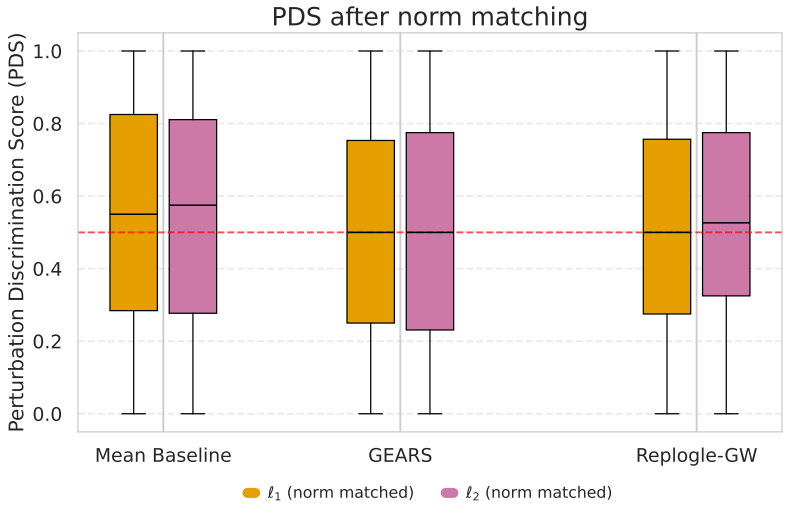

Figure 2: **Norm-matched ℓ_1/ℓ_2 PDS.** Boxplots of PDS computed using ℓ_1 and ℓ_2 distances after rescaling each predicted perturbation-effect vector $\hat{\Delta}_i$ to to match the ℓ_1 or ℓ_2 norm of the corresponding true effect vector Δ_i. Results are shown for the same three prediction strategies as in Figure 1. The target genes are excluded from the perturbation-effect vectors.

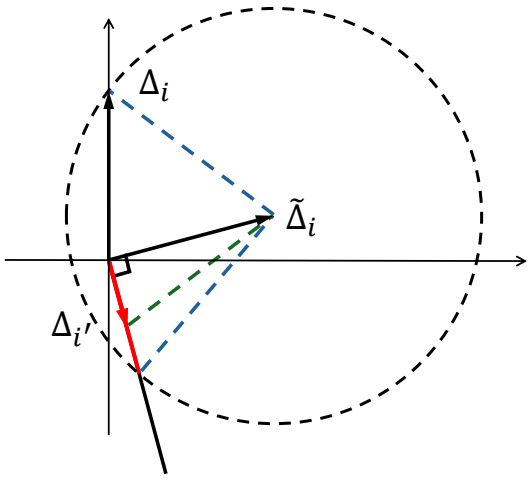

Figure 3: **Geometric illustration of the sensitivity of ℓ_2-based PDS in two dimensions.** Even though the rescaled predicted effect $\tilde{\Delta}_i$ is orthogonal to $\Delta_{i'}$, the shorter magnitude of $\Delta_{i'}$ (red segment) places it closer in Euclidean distance to $\tilde{\Delta}_i$ than Δ_i. This demonstrates how ℓ_2-based rankings can favor vectors with smaller norms despite poorer directional alignment.

In higher dimensions, this effect becomes even more pronounced, because vectors with smaller norms occupy a larger region in which they remain closer in ℓ_1/ℓ_2 distance. As a result,magnitude-sensitive versions of PDS may remain relatively insensitive to directional accuracy unless the cosine similarity between predicted and true perturbation effects is very high.

4 Why ℓ_1/ℓ_2-based PDS is intrinsically scale-sensitive

Another striking feature of ℓ_1/ℓ_2-based PDS is its sensitivity to the overall scale of the predicted vectors. In practice, multiplying all predicted effects $\hat{\Delta}_i$ by a constant c can substantially change the resulting PDS values, even though scaling leaves all directional information unchanged. This behavior is illustrated in Figure 4. For the cross-cell-line transfer predictions based on the Replogle

dataset, ℓ_1/ℓ_2-based PDS increases sharply as the scaling factor c grows and then gradually stabilizes. In contrast, when predictions lack discriminative power across perturbations, such as the mean baseline that produces identical predictions for all perturbations, PDS remains close to random guessing regardless of scaling. Thus, scaling does not create discriminative power by itself, but it can strongly amplify the contribution of directional alignment when such signal is present. This behavior can be explained through simple asymptotic calculations.

The ℓ_2 case. Consider the squared ℓ_2 distance between a scaled prediction $c\hat{\Delta}_i$ and another perturbation effect $\Delta_{i'}$:

$$d_{\ell_2}\big(c\hat{\Delta}_i, \Delta_{i'}\big)^2 = \|c\hat{\Delta}_i - \Delta_{i'}\|_2^2 = c^2\|\hat{\Delta}_i\|_2^2 + \|\Delta_{i'}\|_2^2 - 2c\hat{\Delta}_i^\top \Delta_{i'}.$$

The difference in squared distances between the true perturbation i and another perturbation i' is therefore

$$d_{\ell_2}\big(c\hat{\Delta}_i, \Delta_i\big)^2 - d_{\ell_2}\big(c\hat{\Delta}_i, \Delta_{i'}\big)^2 = \big(\|\Delta_i\|_2^2 - \|\Delta_{i'}\|_2^2\big) - 2c\big(\hat{\Delta}_i^\top \Delta_i - \hat{\Delta}_i^\top \Delta_{i'}\big).$$

Dividing by c and taking $c \to \infty$ yields

$$\lim_{c\to\infty} \frac{d_{\ell_2}\big(c\hat{\Delta}_i, \Delta_i\big)^2 - d_{\ell_2}\big(c\hat{\Delta}_i, \Delta_{i'}\big)^2}{c} = -2\big(\hat{\Delta}_i^\top \Delta_i - \hat{\Delta}_i^\top \Delta_{i'}\big).$$

Using the identity

$$\hat{\Delta}_i^\top \Delta_{i'} = \|\hat{\Delta}_i\|_2 \, \|\Delta_{i'}\|_2 \, \cos\big(\hat{\Delta}_i, \Delta_{i'}\big),$$

the condition for the true perturbation i to be ranked closer than i' in the limit $c \to \infty$ becomes

$$\cos\big(\hat{\Delta}_i, \Delta_i\big) > \frac{\|\Delta_i'\|_2}{\|\Delta_i\|_2} \cos\big(\hat{\Delta}_i, \Delta_{i'}\big).$$

This inequality shows that the comparison becomes increasingly governed by directional alignment (cosine similarity) as scaling increases. In particular, if $\cos(\hat{\Delta}_i, \Delta_{i'}) = 0$, meaning the prediction is orthogonal to the wrong perturbation, then the condition reduces to $\cos(\hat{\Delta}_i, \Delta_i) > 0$, regardless of the relative magnitudes $\|\Delta_i\|_2$ and $\|\Delta_i'\|_2$. More generally, when $\cos(\hat{\Delta}_i, \Delta_{i'})$ is near 0, which is common in high-dimensional settings, only modest positive cosine similarity with the true perturbation is needed to outrank many alternatives. Thus, scaling amplifies the directional term relative to magnitude differences, causing the ℓ_2-based ranking to behave increasingly like cosine similarity.

The ℓ_1 case. A similar limit arises for the ℓ_1 norm. For each coordinate j,

$$\big|c\hat{\Delta}_{ij} - \Delta_{i'j}\big| = c\big|\hat{\Delta}_{ij}\big| - \operatorname{sign}(\hat{\Delta}_{ij}) \operatorname{sign}(\Delta_{i'j})\big|\Delta_{i'j}\big| \quad (c \to \infty),$$

so that

$$\lim_{c\to\infty} \Big\{ d_{\ell_1}\big(c\hat{\Delta}_i, \Delta_i\big) - d_{\ell_1}\big(c\hat{\Delta}_i, \Delta_{i'}\big) \Big\} = -\Big[\sum_{j=1}^{p} \operatorname{sign}(\hat{\Delta}_{ij}) \operatorname{sign}(\Delta_{ij})\big|\Delta_{ij}\big| - \sum_{j=1}^{p} \operatorname{sign}(\hat{\Delta}_{ij}) \operatorname{sign}(\Delta_{i'j})\big|\Delta_{i'j}\big| \Big].$$

Thus, in this limit, ℓ_1-based PDS becomes driven by a weighted sign cosine similarity. This links ℓ_1-based PDS to a form of sign-based similarity in which agreement in sign on large-magnitude coordinates contributes most strongly to the ranking.

These asymptotic results clarify why ℓ_1/ℓ_2-based PDS is sensitive to the global scale of predicted effects. As the scaling factor c increases, the contribution of magnitude differences becomes dominated by inner-product terms that reflect directional alignment between vectors. Consequently, the rankings induced by ℓ_2-based PDS approach those determined by cosine similarity, while ℓ_1-based PDS approaches a weighted sign-based similarity. This limiting behavior explains the empirical pattern in Figure 4: scaling increasingly emphasizes directional alignment between predicted and true perturbation effects.

Importantly, this phenomenon is not an artifact of manipulating the evaluation metric. Instead, it is an inherent property of magnitude-sensitive norms in multi-dimensional spaces, where distances combine both scale and direction. Because scaling modifies the magnitude component while leaving direction unchanged, ℓ_1/ℓ_2-based PDS naturally interpolates toward cosine- or sign-based similarity measures as predictions are globally rescaled.

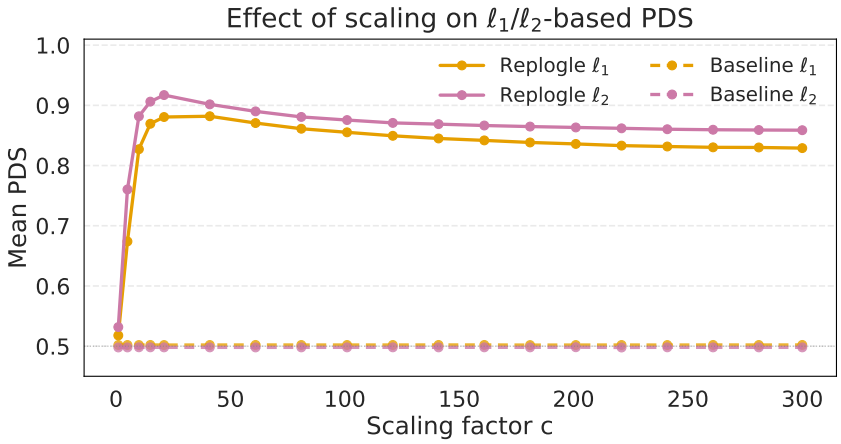

Figure 4: **ℓ_1/ℓ_2-based PDS metrics are sensitive to the magnitude of predicted effects.** Mean PDS of scaled predictions $c\hat{\Delta}_i$ is shown as a function of the scaling factor c. Solid curves correspond to the cross-cell-line transfer predictions based on the Replogle dataset, while dashed curves correspond to the mean baseline.

5 Implications for PDS-based evalution

The analyses above suggest two potential directions for refining PDS in future perturbation-effect benchmarks. A first option is to define PDS using cosine-based similarity measures. Because cosine similarity depends only on direction, these metrics are invariant to global rescaling and robust to differences in normalization or preprocessing. Cosine-based PDS therefore evaluates whether predicted perturbation effects capture the correct pattern of up- and down-regulated genes, without being influenced by inconsistencies in total effect magnitude across studies or platforms.

A potential concern is that cosine-based PDS may be easier to score well on: achieving high discrimination does not necessarily require high correlation with the true effects, only that predictions outperform random guessing. A stricter alternative is therefore to retain ℓ_1/ℓ_2-based PDS, but to fix the norm of the predicted vectors, for example by enforcing $\|\hat{\Delta}_i\|_1 = \|\Delta_i\|_1$ or $\|\hat{\Delta}_i\|_2 = \|\Delta_i\|_2$. This "norm-matched" PDS removes the possibility of improving scores through arbitrary rescaling while still focusing on directional discriminability. Under this formulation, achieving high PDS requires genuinely close directional alignment with the true perturbation effects, yielding a more stringent evaluation criterion.

Notably, neither of these refinements requires PDS to evaluate the total magnitude accuracy of predicted perturbation effects, and in many settings this is disirable. Total effect magnitudes are intrinsically difficult to estimate because they depend on factors such as guide RNA efficiency, experimental design, measurement noise in single-cell RNA-seq, and preprocessing choices.

In practice, the total ℓ_1 and ℓ_2 norms of perturbation-effect vectors can vary dramatically under different normalization pipelines. For example, Figure 5 compares mean perturbation effects derived from the same raw counts of the VCC data under two common preprocessing choices: (i) median library-size normalization and (ii) per-10k scaling followed by log1p transformation. Although these transformations yield perturbation effects with high cosine similarity, their total ℓ_1 and ℓ_2 norms differ substantially.

These observations suggest that direction-based metrics may provide more stable and meaningful comparisons across contexts, particularly when integrating data across experiments or platforms. In contrast, ℓ_1/ℓ_2-based PDS is sensitive to global scaling in a way that does not consistently reflect biological magnitude accuracy: increasing the scale of the predicted effects can artificially inflate the score, while magnitude mismatch may be penalized or rewarded depending on their interaction with vector norms and angular relationships. Thus, even if one wished to evaluate total magnitude accuracy, ℓ_1/ℓ_2-based PDS would not be the appropriate tool. Focusing PDS on directional infor-

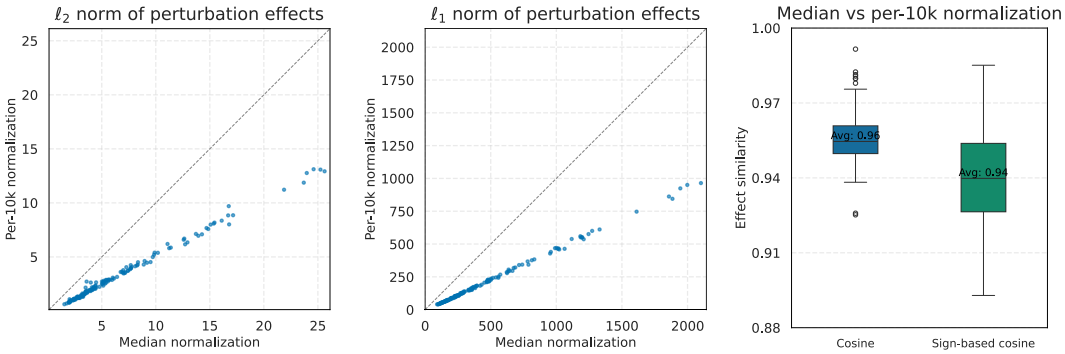

Figure 5: **Perturbation-effect magnitudes depend strongly on preprocessing.** The first two panels show the ℓ_2 and ℓ_1 norms of mean perturbation effects obtained using per-10k versus median normalization; each point corresponds to one perturbation. The third panel shows boxplots of the cosine and sign-based cosine similarities between the effects produced by the two preprocessing pipelines.

mation avoids these instabilities and more directly evaluates the aspects of perturbation effects that are most reproducible across contexts.

6 Conclusion

Our analysis clarifies how the Perturbation Discrimination Score behaves across different similarity measures and under global prediction scaling. In particular, ℓ_1/ℓ_2-based PDS can be strongly influenced by the magnitude of predicted vectors, while cosine-based versions are more robust to these effects. These findings do not critique the use of PDS in benchmarking the prediction performance; rather, they illuminate the metric's geometry and offer guidance for designing future discriminability-based evaluations. As perturbation-prediction benchmarks continue to grow in scale and complexity, a deeper understanding of PDS-style metrics will help ensure that evaluation criteria reflect the biological and statistical goals of each task.

References

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, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024.

Yusuf H Roohani, Tony J Hua, Po-Yuan Tung, Lexi R Bounds, Feiqiao B Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S Plosky, et al. Virtual cell challenge: Toward a turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2025.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, pp. 1–10, 2025.

Yan Wu, Esther Wershof, Sebastian M Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Kun Zhang, and Thore Graepel. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*, 2024.