A Systematic Comparison of Single-Cell Perturbation Response Prediction Models

Lanxiang Li[1,2†], Yue You[1*†], Yunlin Fu[1†], Wenyu Liao[3], Xueying Fan[4,5,6,7], Shihong Lu[1], Ye Cao[1], Bo Li[1], Wenle Ren[1], Jiaming Kong[8], Shuangjia Zheng[9], Jizheng Chen[1,10], Xiaodong Liu[4,5,6,7*], Luyi Tian[1,2*]

[1*]Guangzhou National Laboratory, Guangzhou, China.
[2*]GMU-GIBH Joint School of Life Sciences, Guangzhou Medical University, Guangzhou, China.
[3]The University of Hong Kong, Hong Kong, China.
[4]School of Life Sciences, Westlake University, Hangzhou, China.
[5]Research Center for Industries of the Future, Westlake University, Hangzhou, China.
[6]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China.
[7]Westlake Institute for Advanced Study, Hangzhou, China.
[8]Renaissance Era Innovation Technology Co., Ltd, China.
[9]Global Institute of Future Technology, Shanghai Jiaotong University, Shanghai, China.
[10]State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Health, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, China.

*Corresponding author(s). E-mail(s): you_yue@gzlab.ac.cn; liuxiaodong@westlake.edu.cn; tian_luyi@gzlab.ac.cn;
†These authors contributed equally to this work.

Abstract

Predicting single-cell transcriptional responses to perturbations is central to dissecting gene regulation and accelerating therapeutic design, yet the field lacks a rigorous, task-spanning assessment of model behavior. We present a large-scale benchmark of 12 representative methods and 3 baselines across 25 datasets spanning diverse perturbation modalities and species, including two new primary immune-cell drug-response resources. We evaluated three core tasks—generalization to unseen single-gene perturbations, prediction of combinatorial interactions, and transfer across cell types—using 24 metrics covering expression-level accuracy, relative changes, differential expression recovery, and distributional similarity. Across tasks, performance depended strongly on perturbation effect size and evaluation perspective: expression-level agreement was highest for small-effect perturbations resembling controls, whereas delta- and DE-based metrics improved with larger effects, providing clearer signals. Models shared a conservative bias, with fine-tuned foundation models compressing variance and underestimating synergistic effects in combinations. PerturbNet showed superior recovery of DE signatures in Tasks 1 and 2, while no method consistently generalized across cell types in Task 3, where dataset heterogeneity dominated outcomes. This benchmark establishes current methodological limits, clarifies when metrics diverge, and provides a foundation for developing virtual-cell models that more faithfully capture heterogeneous perturbation responses.

Introduction

Single-cell RNA sequencing (scRNA-seq) offers unprecedented resolution for exploring cellular heterogeneity [1, 2]. Multiplexed perturbations together with scRNAseq, such as Perturb-seq and CROP-seq, help uncover how cells respond to diverse stimuli or interventions [3–5]. These approaches advance our understanding of cellular functions and regulatory networks, providing crucial insights for drug development and disease treatment [6]. However, current experimental systems are largely optimized for cell lines, with limited applicability to primary or in vivo systems. Furthermore, combinatorial multiplexing experiments remain prohibitively expensive and technically challenging to scale [7, 8].

In this context, in silico approaches that predict post-perturbation cell states across diverse environments and conditions present a promising, scalable alternative and provide a path to building a virtual cell [9, 10]. Early approaches focused on predicting cellular responses in cell types absent from experimental training data or responses to combinations of previously tested perturbations [11–13]. However, these methods could not be generalized to unseen perturbations. Recent advances have incorporated graph-based methods and foundational models that take advantage of prior knowledge of gene expression relationships, achieving significant improvements in generalizability [14–17].

Among them, single-cell foundation models represent a pivotal breakthrough. By leveraging vast datasets and pre-trained transformer architectures[18], these models learn robust, generalizable relationships between genes and have demonstrated potential across diverse downstream tasks, including transcriptome prediction following perturbations after models are fine-tuned on specific datasets [19]. Despite their promise, the field lacks a comprehensive framework to evaluate their capabilities systematically. Compared to tasks such as cell annotation and batch correction, perturbation prediction offers experimentally generated model input/output with well-defined objectives, making it a more robust benchmark for assessing foundation models [15, 20, 21]. Recent studies have provided initial benchmarks comparing foundation models, other deep learning architectures, and simple baselines for perturbation effect prediction, and reported that complex models do not consistently outperform the simpler alternatives [22–25]. While this work provides valuable insights, its evaluation was limited in dataset diversity and task coverage, leaving the question of whether such conclusions generalize across broader biological contexts.

At the same time, the Virtual Cell Challenge[26] has emerged as a community benchmark for perturbation prediction, attracting wide participation and fostering the development of new evaluation frameworks such as Cell-Eval[27]. Together with the release of large-scale datasets like Tahoe-100M[28], these initiatives reflect a growing momentum toward standardized, transparent, and multi-metric benchmarking in this rapidly evolving field.

Motivated by these developments, we present a comprehensive benchmark that significantly expands the scope to a more diverse set of datasets, perturbation modalities, and prediction tasks. In addition to systematically evaluating model performance, we summarize common patterns across datasets and identify key biological and technical factors influencing predictive accuracy—most notably, that the magnitude of

the perturbation effect strongly impacts both generalization and double perturbation prediction performance, and that in cell-type transfer settings, greater transcriptomic distance between source and target cell types is associated with reduced accuracy.

Results

Benchmarking models and experimental design

We comprehensively benchmarked 12 single-cell perturbation response prediction models across a broad and diverse dataset collection, together with 3 baseline models for different tasks. Our goal was to assess how well different model architectures generalize across biological contexts, perturbation types, and prediction tasks (Supplementary Fig. 1). Our dataset collection spans a broad range of biological contexts, perturbation modalities, and effect sizes. It integrates large-scale public resources, recently released high-throughput perturbation screens, and in-house generated datasets. The collection covers genetic perturbations (e.g., CRISPR knockouts and activations), small-molecule treatments (e.g., kinase inhibitors, anti-aging compounds), and immune stimuli (e.g., LPS, cytokines), measured across cancer cell lines (e.g., K562, A549), pluripotent stem cells (e.g., hESCs), and primary immune populations (e.g., T cells, NK cells, monocytes), as well as non-human immune cells (e.g., mouse, rat). Of note, we incorporated datasets that include multi-cell-type profiling from the Virtual Cell Challenge, extensive drug screens across dozens of cell lines, and primary immune cell responses to cytokines. In addition, we generated two new datasets that capture the effects of drugs on both myeloid and lymphoid subsets, enabling a more systematic evaluation of cell-type transfer (Fig. 1a).

We systematically examined perturbation modalities, numbers of differentially expressed genes, E-distances, and sample sizes both across experiments and within individual experiments. This analysis revealed broad dispersion in the number of perturbations, cell numbers, and perturbation effect sizes, with substantial heterogeneity not only between studies but also among conditions within the same study. As a modality trend, CRISPRa often induces larger transcriptome shifts, whereas CRISPRi tends to yield more moderate effects; yet well-designed, well-powered CRISPRi datasets—such as the Replogle genome-scale screen in RPE1 cells and the curated Virtual Cell Challenge—still exhibit strong signals (often > 500 DEGs for many conditions) (Fig. 1b). Beyond modality, cell-type sensitivity further modulates measurable effects, with certain immune subsets showing muted transcriptional responses to specific perturbations compared with other cell types. A full characterization of these properties and details of all datasets utilized are summarized in Table 1 and Supplementary Table 1. This heterogeneity in effect size and response patterns enables a more comprehensive and discriminative assessment of model generalization across diverse biological scenarios.

In terms of the benchmarking object, we categorized them into three tasks focusing on their ability in different scenarios, each representing a different aspect of generalization: predicting unseen perturbations, combinatorial perturbation interactions, and responses across unseen cell types. These tasks reflect the varying complexity of perturbation prediction and the diverse experimental contexts in which these models are

applied. The evaluation process is shown in Fig. 1c, where the input and output for the training and testing phases are outlined. Specifically, we focus on the following.

1. **Unseen Perturbation Prediction** (Fig. 1c, bottom left): In this task, the models were trained to predict the effects of unseen single-gene perturbations within the same cell type or cell line. Using training data that included both unperturbed and known single-gene perturbation responses, the models were tasked with inferring gene expression changes under previously unseen perturbations.
2. **Combinatorial Perturbation Prediction** (Fig. 1c, bottom middle): This task introduced additional complexity by requiring the models to predict the interaction effects of combinatorial perturbations. Based on single-gene perturbation data, models were expected to predict how new combinations of gene perturbations would influence gene expression in the same cell type or cell line.
3. **Cell state transfer** (Fig. 1c, bottom right): This task assessed the models' ability to generalize the effect of a known perturbation across different cell types or other categories. The models were trained on unperturbed and perturbed data from various cell types to learn gene expression patterns under specific perturbations. They were then tested on previously unseen cell types, predicting their responses to the same perturbations.

To evaluate the models in the three tasks, we selected 12 representative models categorized into 4 types based on their design and model structure. (1) Foundation models (scGPT, scFoundation), pre-trained on large datasets.(2) Gene regulatory network–based models (GEARS, scELMo), which use regulatory information to predict perturbation effects. (3) Variational Autoencoder(VAE)-base model (PerturbNet, scGen, trVAE, scVIDR, scPRAM). (4) Other deep learning-based models (CPA, scPreGAN, BioLord). Additionally, we established task-specific MLP-based baselines for comparison: Baseline_1 for single perturbation prediction using one-hot encoded perturbation embeddings; Baseline_2 for combinatorial perturbation prediction combining multiple one-hot encoded perturbations; and Baseline_3 for cross-cell perturbation transfer learning from other cells under the same perturbation. Detailed model descriptions are provided in Supplementary Table 2.

Next, we implemented a standardized, multi-dimensional evaluation pipeline designed to capture complementary aspects of model performance across diverse perturbation scenarios (Fig. 1c, right). Our framework integrates four categories of metrics that together assess absolute accuracy, relative effect capture, differential expression recovery, and distributional similarity. Direct gene expression–level agreement was quantified using error-based measures (MSE, RMSE, MAE, and L2 loss), goodness of fit (R2), and correlation metrics (Pearson's correlation and cosine similarity). Each of these metrics was calculated on the full gene set as well as on subsets containing the top 20, 50, or 100 genes ranked by absolute score, allowing us to assess model behavior across different levels of gene prioritization. To evaluate the ability to capture relative perturbation effects, we computed expression changes (perturbed versus control) and assessed Pearson's correlation on these delta values, together with delta direction accuracy, which measures the proportion of genes with correctly predicted up- or down-regulation. We further evaluated the recovery of differential expression by identifying

DE genes in both predicted and observed profiles, then quantifying the overlap accuracy and the Spearman correlation of effect size, defined as the number of DE genes per perturbation. Finally, to assess higher-order similarity beyond mean shifts, we calculated the Wasserstein distance and maximum mean discrepancy (MMD) between predicted and observed gene expression values. Both metrics were computed per gene by comparing the distributions of predicted and observed expression values across all cells for that gene. The per-gene distances were then aggregated to yield a single scalar that summarises overall distributional fidelity for each perturbation condition. In addition to absolute metric values, we performed a ranking-based meta-analysis in which models were ranked for each perturbation and metric, and aggregated win rates were computed to capture relative performance trends across datasets and tasks.

In summary, we systematically evaluated 15 single-cell perturbation prediction methods across three core tasks (including 3 baseline models), utilizing 25 datasets and 24 distinct evaluation metrics to provide a comprehensive comparison of model performance.

Prediction of unseen single-gene perturbations

In the first task, we assessed the ability of each model to generalize to single-gene perturbations that were not observed during training, a critical measure of whether models can extrapolate to new genetic interventions (Fig. 2a). GEARS and scELMo leverage gene regulatory priors, foundation models such as scFoundation and scGPT benefit from large-scale pretraining, and BioLord and PerturbNet integrate broad perturbation–response relationships. As a baseline, we used a simple additive MLP with one-hot perturbation encodings (see Methods).

Using direct gene expression–level metrics, we observed generally small performance differences between models, with Pearson's correlation exceeding 0.95 for most datasets (Fig. 2b, Supplementary Fig. 2-6). Datasets with smaller perturbation effects generally produced higher correlations and lower performance variance (e.g., Datlinger_2017 with many low-DE conditions), whereas large-effect datasets such as Replogle and Sunshine yielded lower scores (Fig. 2c). Within datasets, effect size quantified by TRADE (transcriptome-wide impact) showed a similar trend—smaller-effect perturbations were easier to predict (Supplementary Fig. 7).

When moving to an analysis based on expression changes relative to the control, the relationship was reversed. Delta-based Pearson correlation and delta direction accuracy were strongly positively correlated, and perturbations with larger effect sizes yielded higher scores (Fig. 2d, Supplementary Fig. 8). This likely reflects the fact that small-effect perturbations are more difficult to distinguish from technical noise, whereas large-effect perturbations provide a clearer and more reproducible signal for the models to capture.

Foundation models exhibited variance compression after fine-tuning, reducing the spread of predicted deltas toward a dataset-wide mean, as confirmed by higher MMD and Wasserstein distances relative to other model classes (Fig. 2e). This "variance compression" likely arises because the distribution preservation is not in the finetuning objectives in the foundation model.

Evaluation of DE gene overlap accuracy and effect size ranking further supported the delta-based trends: datasets with larger perturbation effects achieved higher DE overlap accuracy and stronger concordance in effect size ranking. In this setting, PerturbNet emerged as the clear top performer, achieving a mean overlap accuracy of 0.56, whereas most other methods failed to recover DEGs reliably in several datasets (Fig. 2f, Supplementary Fig. 6b).

Variance decomposition showed that dataset and perturbation identity were primary factors for direct and delta metrics, while DE accuracy was strongly driven by model choice, dominated by PerturbNet (Fig. 2g).

Perturbation interaction prediction

Task 2 evaluated each model's ability to predict gene expression responses to combinatorial perturbations. We considered three "seen levels" based on whether the individual perturbations comprising a combination had appeared in the training data: seen0 (neither perturbation seen), seen1 (one perturbation seen), and seen2 (both perturbations seen) [14] (Fig. 3a). Predicting seen0 combinations is most challenging, as it requires both generalization to previously unseen perturbations—analogous to Task 1—and inference of interaction effects, whereas seen2 primarily tests a model's ability to capture the interaction effect of combinatorial perturbation. Models included are CPA, GEARS, scELMo, scFoundation, scGPT, and the Baseline model (additive MLP with one-hot encodings for A and B (see Methods)).

Across direct gene expression–level metrics, most models showed a consistent and significant performance increase from seen0 to seen2, indicating that prior exposure to constituent perturbations in training data improves predictive accuracy. A similar trend was observed for delta Pearson correlations in some methods (Fig. 3b, Supplementary Fig. 9-13).

We further classified each combination based on its deviation from the additive expectation—comparing the observed delta for the combination to the sum of deltas for its components—into buffering (combination delta smaller than additive) and synergistic (combination delta larger than additive) categories (Fig. 3c). Across both seen0 and seen2, buffering combinations were predicted more accurately than synergistic ones, consistent with previous benchmarking studies showing that models tend to make conservative predictions for interaction effects [25]. Notably, buffering accuracy in seen0 was lower than in seen2, likely reflecting the added difficulty of generalizing to unseen perturbations while also modeling their joint effects (Fig. 3d, Supplementary Fig. 14).

At the delta and DE evaluation levels, PerturbNet again consistently outperformed all other models, replicating its dominance in Task 1, particularly in DE overlap accuracy, where it achieved substantially higher values while most other methods struggled to identify any significant DE genes (Fig. 3e).

ANOVA revealed model choice as the main determinant for MSE, with CPA deviating most from others, while both model and interaction type influenced correlation metrics (Fig. 3f).

Cell state transfer

Task 3 assessed each model's ability to generalize known perturbation responses across distinct cell types (Fig. 4a). We evaluated five methods, including scPreGAN, scGen, trVAE, scVIDR, and scPRAM, using diverse datasets in which at least two cell types shared a set of perturbations. A simple additive MLP with one-hot perturbation encodings is used as a baseline model (see Methods).

For direct gene expression–level evaluation, most models (except scPreGAN) achieved similarly high performance across datasets (Fig. 4b, Supplementary Fig. 15-18). However, notable variation was observed across datasets: in general, datasets with smaller mean E-distances between source and target cell types yielded better transfer performance (Fig. 4c,d, Supplementary Fig. 19). E-distance, measuring the separability of perturbed versus control distributions in high-dimensional space, reflects perturbation effect size and aligns with signal-to-noise ratios. Smaller E-distances generally correlated with better transfer accuracy (e.g., transfer IFN-β stimulation in human immune subsets compared with transfer various drug effect from epithelial to K562).

But E-distance is not fully predictive. For example, results from Burkhardt showed low transfer accuracy despite a small E-distance, likely due to heterogeneous drug effects and high within-group variability, while results from Hagai, with a large domain shift and strong transfer, benefited from conserved immune activation signatures. These results indicate that Task 3 performance is shaped not only by baseline domain differences between cell types, but also by the consistency of perturbation effects across cell types.

Delta-level performance mirrored direct-level trends, with trVAE leading in direction accuracy, followed by scPRAM, and scVIDR performing best in delta Pearson correlation, followed by scGen (Fig. 4e).

At the DE level, all models performed poorly, consistent with findings from previous tasks (Tasks 1 and 2), indicating a persistent challenge in accurately predicting differentially expressed genes (Fig. 4e).

Variance partitioning indicated that model, perturbation identity, and cell type each contributed significantly, highlighting that both biological and methodological factors shape transfer success (Fig. 4f).

Discussion

This study provides a comprehensive evaluation of the performance of single-cell perturbation prediction models, offering an in-depth exploration of their current capabilities and limitations. By assessing a diverse set of models using complementary criteria—including direct gene expression reconstruction, delta-based shifts, differentially expressed (DE) gene identification, and distributional similarity metrics—we uncover performance facets that single metrics alone fail to reveal. Our analysis demonstrates that model performance arises from the intricate interplay among data regime (e.g., perturbation effect size, cell-type variability), metric selection, and model architecture, each emphasizing distinct aspects of the biological signal and its capture (Fig. 5).

A key finding is that model performance is strongly modulated by perturbation effect size and the chosen evaluation metric, reflecting how these factors interact with underlying data properties. For small-effect perturbations, direct gene expression metrics (e.g., correlation or MSE) perform well, as perturbed profiles remain proximate to controls, minimizing the scope for large prediction errors. Conversely, delta-based metrics favor large-effect perturbations, where the directional signal stands out more clearly against technical noise, facilitating accurate capture of changes. However, DE gene identification imposes a more stringent requirement: models must not only predict the correct direction but also achieve sufficient effect size to attain statistical significance. This proved exceptionally challenging, and it suggests that more efforts need to be made on designing better evaluation metrics for virtual-cell perturbation tasks [29]; indeed, nearly all methods—except PerturbNet—output inflated DE-gene lists, often classifying most genes as differentially expressed across tasks.

Architecturally, many models optimize for mean reconstruction, leading to systematic variance compression that misaligns with DE requirements. Autoencoder (AE) and variational autoencoder (VAE)-based methods, including conditional variants, encourage smooth predictions with small, systematic shifts, but implicitly shrink variance to stabilize training. This results in high global correlations yet inflated DE calls, as compressed variance heightens statistical sensitivity, producing spurious significance even when biological structure is not preserved. Fine-tuned foundation models, pretrained on vast datasets, capture broad expression patterns but worsen variance compression during fine-tuning, prioritizing mean alignment. This boosts reconstruction metrics like correlation but degrades DE precision and distributional metrics like MMD, as fine-tuning collapses pretrained variability into overly smooth outputs, misrepresenting perturbation response diversity. In contrast, flow-based architectures like PerturbNet embed perturbations and cells separately, modeling full distributions via conditional normalizing flows. [30] This preserves the covariance of genes and the second moment structure, aligning the mean changes with realistic variance and enabling accurate DE recovery by resolving sharp biological signals against noise.

These architectural effects highlight the need for variance modeling to address poor DE performance, but Tasks 2 and 3 reveal further challenges. For Task 2 (combinatorial perturbations), models predict buffering combinations more accurately than synergistic ones across seen0 and seen2, consistent with prior studies showing conservative predictions for interaction effects [25]. This reflects a failure to capture synergistic covariance, suggesting exploration of advanced architectures or loss functions that prioritize non-linear interaction modeling. For Task 3 (cross-cell-type transfer), poor domain shift handling limits performance. Success requires domain adaptation, like hierarchical calibration, and diverse datasets to model cell-type-specific responses.

Despite the advances made, this study has limitations. Most datasets used are publicly available and were generated using varied technologies, introducing differences in data quality and batch effects that may influence comparisons. ANOVA analysis confirmed that performance variability arises from the dataset and perturbations, highlighting the complexity and heterogeneity of perturbation effects. Expanding the benchmark to include more perturbation datasets with standardized protocols will be critical for disentangling these influences. Additionally, the lack of genetic perturbation

datasets with diverse cell types limited our ability to assess the cell type transfer capabilities of foundation models fully.

Beyond this study, broader challenges remain in the field. A fundamental limitation is the lack of real ground truth due to the destructive nature of single-cell sequencing methods, that we cannot measure the transcriptome of the same cell before and after perturbation. Another key limitation is the lack of temporal dynamics in existing datasets, which typically include only pre- and post-perturbation time points. Furthermore, current models predominantly focus on transcriptomics, yet integrating additional modalities—such as genomic mutations, epigenomic modifications, and proteomics—is essential for a holistic understanding of perturbation responses[31]. Incorporating language representation around the central dogma (i.e., DNA, RNA, and protein) and molecular representations of perturbations at the atomic level will aid in deciphering the intricate biomolecular interactions within the cell [32], thereby offering a more generalized solution for single-cell perturbation response prediction. Spatial perturbation technology is still in its infancy, but could potentially transform the future model from single-cell to tissue level [33].

In general, our study highlights gaps and provides a roadmap for future developments of single-cell perturbation response prediction models. To facilitate practical application of our findings, we provide a decision framework to guide method selection based on experimental objectives and dataset characteristics (Supplementary Fig. 20). Optimizing training processes, improving model architectures, and incorporating multi-timepoint and multi-modal datasets will collectively drive the evolution of single-cell predictive models. Such developments will not only improve benchmarking fidelity but also accelerate the construction of 'virtual cell' [34] frameworks capable of simulating complex perturbation responses for applications in precision medicine and biotechnology.

Methods

Ethics statement

Human dermal fibroblasts (HDFs) and peripheral blood mononuclear cells (PBMCs) were collected from two male donors with informed consent. The use of both HDFs and PBMCs in this study was approved by the Westlake University Ethics Committee (approval number: 20240909LXD001).

Cell culture and drug treatment

Perturb_KHP dataset

HDFs were cultured in DMEM medium (`C11965500BT, Gibco`) supplemented with 10% fetal bovine serum (FBS; `C04002, VivaCell`), 1% non-essential amino acids (`11140-050, Gibco`), 1 mM GlutaMAX (`35050-061, Gibco`), 55 µM 2-mercaptoethanol (`21985-023, Gibco`), 1 mM sodium pyruvate (`11360-070, Gibco`), and 1% penicillin-streptomycin (`15140-122, Gibco`).

PBMCs were cultured in StemPro™-34 SFM Medium (10639-011, Gibco) supplemented with StemPro™-34 Nutrient Supplement, 1 mM GlutaMAX (35050-061, Gibco), 100 ng/mL SCF, 100 ng/mL FLT-3, 20 ng/mL IL-3, and 20 ng/mL IL-6.

K562 cells, obtained from ATCC, were cultured in RPMI 1640 medium (61870-036, Gibco) with 10% FBS and 1% penicillin-streptomycin.

HDFs from two donors were seeded into 12-well plates at a density of 1×10^5 cells per well. PBMCs from the same donors were thawed and seeded into 96-well low-attachment plates at a density of 4×10^5 cells per well. K562 cells were seeded at a density of 1×10^5 cells per well in either 96-well or 12-well plates. K562 cells were rested for 24 hours, while HDFs and PBMCs were rested for 48 hours, before drugs were added to each well to reach the specified final concentrations. Half of the medium was refreshed daily. A detailed list of drugs and concentrations used for each cell type is provided in the Supplementary Table 3. K562 cells were collected for single-cell RNA sequencing after 3 days of treatment, while HDFs and PBMCs were collected after 4 days.

Perturb_cmo_V1 dataset

Frozen human PBMCs from three healthy donors (aged 20–25; *Stem Cell Technologies*, #70025.1) were thawed and cultured in 96-well low-attachment plates at a seeding density of 4.5×10^5 cells per well. Cells from two vials originating from two donors were mixed prior to seeding. PBMCs were maintained in RPMI 1640 medium (61870-036, *Gibco*) supplemented with 10% heat-inactivated FBS (C04002, *VivaCell*) and 1% penicillin–streptomycin (15140-122, *Gibco*). Cultures were allowed to acclimate for 24 hours before drug treatment, with half of the medium replaced daily to sustain cell viability. After four days of continuous treatment, cells were harvested for downstream single-cell RNA sequencing.

A total of 111 drugs from the Drug Repurposing Compound Custom Library (L9200, *TargetMOI*) were resuspended at 10 mM in DMSO and stored at –80°C. For PBMC treatment, drugs were diluted in culture medium to a final concentration of 1 μM. Control groups received DMSO at an equivalent dilution. A detailed list of drugs used provided in the Supplementary Table 4.

Sample multiplexing for single-cell RNA sequencing

Sample multiplexing was conducted using the 3' CellPlex Kit Set A (10X Genomics, 1000261), following the manufacturer's instructions. In brief, cells from each well were collected and labeled with unique cell multiplexing oligonucleotides (CMOs). For the Perturb_KHP dataset, equal numbers of live cells from drug-treated and untreated groups were pooled into a single sample. For HDFs and PBMCs, cells from two donors treated with the same drug were mixed before incubating with the multiplexing oligos. A total of two pooled samples were prepared for K562 cells, two for HDFs, and two for PBMCs. For the Perturb_cmo_V1 dataset, cells were pooled at equal viability-adjusted numbers across eight treatment groups per sample, generating a total of 16 pooled samples. All samples underwent library preparation using the 10X Chromium Next GEM Single Cell 3' Kit v3.1 (10X Genomics, 1000123).

Datasets and preprocessing

According to the research conducted by Peidli et al. [8], we acquired the Datlinger [4, 35], Dixit [3], Frangieh [36], Papalexi [37], Replogle [38], Tian [39], and Srivatsan [40] datasets from Zenodo (access link: https://zenodo.org/records/7041849). For the Datlinger and Frangieh datasets, we curated the perturbation_2 subset, excluding perturbations related to "stimulate" and "IFNγ," as well as "Co-culture," while retaining only the CRISPR-Cas9 perturbations. The Sunshine [41] dataset was obtained from the NCBI Gene Expression Omnibus (GEO) under accession number GSE208240, where we filtered for entries classified as "exact_match," "no_match," or "scrambled_pair," and excluded entries that contained non-targeting sequences in the "guide_identity."

We further expanded our dataset collection by including the Joung and Xu datasets directly from the authors, the Virtual Cell Challenge train dataset from https://virtualcellchallenge.org/app/datasets, the Maya M. Arce dataset from GSE278572, the Tahoe dataset from https://github.com/ArcInstitute/arc-virtual-cell-atlas/tree/main, the Parse Biosciences dataset from https://parse-wget.s3.us-west-2.amazonaws.com/10m, and the Burkhardt dataset from https://www.kaggle.com/competitions/open-problems-single-cell-perturbations/data. Due to the large size of the Tahoe dataset, which consists of 14 plates, we implemented a specialized preprocessing pipeline. For each plate, we first retained combinations of cell name and drug with at least 1,000 cells, then randomly sampled up to 100 cells per group. Subsequently, we filtered conditions to retain only those with at least 40 unique cell types within each plate, while preserving control conditions. The 14 processed plates were then merged into a single integrated dataset. For the Srivatsan, Burkhardt, Tahoe, Parse Biosciences datasets, as well as our two in-house datasets, we employed TRADE(TRADEtools v 0.99.0) to compute transcriptome-wide impact scores for each perturbation condition. Based on these scores, we selected the top 5 and bottom 5 ranked conditions for retention (detailed in Supplementary Table 5). The implementation details of TRADE are provided below.

Data normalization and log transformation were performed using the scanpy(v1.10.1) toolkit. For datasets involving gene perturbations, we implemented a sample filtering process to ensure adequate coverage and minimize noise. Specifically, we applied stratified subsampling to groups with large sample sizes, limiting the maximum sample count per group to 3,000. Groups with fewer than 50 samples were excluded to reduce potential errors arising from small sample sizes. To enhance the robustness of feature selection, we calculated and retained the top 5,000 highly variable genes. Within datasets containing gene perturbations, we identified perturbed genes based on the "condition" field and extracted their names, thereby creating a subset of genes directly influenced by the experimental conditions. Ultimately, we combined the identified perturbed genes with the highly variable genes to generate the final gene set utilized for data filtering, ensuring the preservation of the most informative genes associated with perturbations. For datasets not involving gene perturbations, we retained only the top 5,000 highly variable genes for filtering. However, for the Xu

dataset, we extended the feature selection to the top 20,000 highly variable genes, as restricting to 5,000 genes resulted in insufficient perturbation conditions for analysis.

Additionally, we downloaded the Norman [42] and Adamson [43] dataset from the following links: (https://dataverse.harvard.edu/api/access/datafile/6154020, https://dataverse.harvard.edu/api/access/datafile/6154417), along with the Kang [44], Haber [45], and Hagai [46] datasets from the scGen study (https://github.com/theislab/scgen-reproducibility/blob/master/code/DataDownloader.py). The Weinreb [47] dataset was sourced from the CellOT study [48] (https://www.research-collection.ethz.ch/handle/20.500.11850/609681). Given that these datasets were appropriately preprocessed in their original studies, no further data processing steps were applied. All data analysis was performed on an in-house Linux server with Ubuntu operating system (version 22.04) and the Data Science Platform of Guangzhou National Laboratory [49].

The DE genes are calculated using the rank_genes_groups function in scanpy with method='wilcoxon'. For each perturbation condition, cells from both the perturbation and control ('ctrl') groups are extracted to form a subset using normalized expression data. After obtaining the DE results, genes are sorted by the absolute values of their scores in descending order to prioritize genes with the strongest differential expression regardless of up- or down-regulation direction. The top 20, 50, and 100 DE genes based on absolute scores are then selected and used to subset the gene expression matrix for calculating the evaluation metrics.

Models selection and setup

In this study, we selected and applied a range of advanced modeling approaches to systematically predict and compare single-cell perturbation responses. The models span diverse theoretical frameworks and algorithmic structures, ranging from baseline models to foundation models, GRN-based, VAE-based, and other deep learning-based approaches. These approaches offer strong adaptability and generalization capabilities, allowing effective application across diverse perturbation contexts through training or fine-tuning.

Baseline model

In single-cell perturbation prediction tasks, the baseline model provides a baseline performance that allows us to compare more complex models better. For task 1, we use Baseline_1, which employs one-hot encoding to represent the cell's gene expression under perturbation A as p_A. The one-hot encoding of perturbations A is embedded into z_A. The gene expression of the cell under the control (no perturbation) condition is embedded into z_{ctrl}. The gene expression after combined perturbations can be represented as $x' = \text{MLP}(z_A + z_{\text{ctrl}})$. For task 2, we use Baseline_2 and still use one-hot encoding to represent the gene expression of the cell under perturbation A as p_A and perturbation B as p_B. And the one-hot encoding of perturbations A and B is embedded into z_A and z_B, respectively. The gene expression after combined perturbations can be represented as $x' = \text{MLP}(z_A + z_B + z_{\text{ctrl}})$. For Task 3, we use Baseline_3, which

train the model for other cells under a single perturbation A, which is defined as $F_A(Z_{\text{other_ctrl}} + p_A)$, where $z_{\text{other_ctrl}}$ is the embedding of the gene expression of other cells under the control (no perturbation) condition, and p_A is the one-hot encoding of perturbation A. For the gene expression of the target cell under perturbation A, it can be represented as $F_A(z_{\text{target_ctrl}} + p_A)$, where $z_{\text{target_ctrl}}$ is the embedding of the gene expression of the target cell under the control condition.

Foundation model-based

The pre-trained models scGPT(v1.0) and scFoundation(v1.0) were employed for the perturbation prediction task, with fine-tuning performed based on the strategy used in the scGPT tutorial. In their approach, a Binary Condition Token was appended at each input gene position to denote the perturbation status of the gene. Besides, we implemented a strategy wherein a non-perturbed control cell was used as the input and a perturbed cell as the target. This was achieved by randomly matching each perturbed cell with a control cell, thereby forming input-target pairs. The models were trained using the default settings recommended in the tutorials, with scGPT undergoing 2 or 5 epochs of training for comparison and scFoundation trained for 2 epochs.

GRN-based

GEARS(v1.0) was utilized to model gene perturbations by seamlessly integrating knowledge graph information with gene embeddings. We followed the tutorial guidelines, employing the default settings as recommended.

For the scELMo(v1.0) model, we followed the tutorial's instructions and incorporated gene embeddings derived from GPT as inputs for gene embedding in GEARS.

VAE-based

The VAE-based approaches, PerturbNet(v0.0.3)[30], scGen(v2.1.0) [50] trVAE(v1.0.0) [51], scVIDR(v1.0) [52] and scPRAM(0.0.2)[53], were utilized for single-cell RNA sequencing (scRNA-seq) perturbation prediction.

PerturbNet combines a convolutional neural network (CNN) with a variational autoencoder (VAE) to capture complex patterns in single-cell gene expression data. The model learns a latent representation of gene expression and leverages this to predict perturbation effects across different conditions. We implemented PerturbNet according to the author's instructions, using the default architecture and hyperparameters.

scGen leverages a variational autoencoder (VAE) framework integrated with vector arithmetic to simulate cellular transitions between conditions, such as control and stimulation. The model encodes gene expression profiles into a latent space, reconstructing them via a decoder while learning different vectors between cellular states. This enables scGen to predict unobserved perturbation responses with high generalization. We implemented scGen following the tutorial provided by the authors, adhering to the recommended architecture and hyperparameters.

trVAE builds on the Conditional Variational Autoencoder (CVAE) framework to model perturbation effects under scenarios lacking paired control and perturbed cells. The model encodes cellular states into a shared latent space while applying condition-specific transformations via perturbation vectors. This is optimized using a masked learning objective and Maximum Mean Discrepancy (MMD) regularization to align distributions across conditions. The implementation followed the author's tutorial, with early stopping enabled and the maximum iterations set to 1000, while all other hyperparameters were kept as default.

scVIDR extends the VAE-based perturbation prediction framework by addressing cell-type-specific heterogeneity in dose-response relationships. The model employs a VAE to encode single-cell gene expression data into a latent space, then uses linear regression to predict cell-type-specific perturbation vectors ((δ_c)) based on control group latent representations. For dose-response prediction, scVIDR performs log-linear interpolation on the estimated perturbation vectors to predict intermediate doses, followed by decoding back to gene expression space. The model introduces a 'pseudo-dose' metric to characterize individual cell sensitivity to chemical perturbations by projecting cells onto the perturbation axis in latent space. We implemented scVIDR following the authors' specifications, with default hyperparameters.

scPRAM utilizes an attention mechanism-based approach for predicting single-cell perturbation responses while accounting for cellular heterogeneity. The model consists of three main components: a VAE for encoding high-dimensional gene expression data into latent space, optimal transport based on the Sinkhorn algorithm to match unpaired cells before and after perturbation, and an attention mechanism to compute perturbation vectors for test cells. The attention mechanism treats the test cell's latent representation as a query, using cosine similarity to identify the most relevant training cells, then computes weighted perturbation vectors through a softmax-normalized attention score. This approach enables scPRAM to capture individual cell-specific responses rather than relying on population averages. We implemented scPRAM with default parameters.

Other deep learning-based

The Compositional Perturbation Autoencoder (CPA,v0.8.7) [54] combines deep learning flexibility with linear model interpretability to predict cellular responses to perturbations. The model uses an encoder to map gene expression data into a disentangled latent space, removing perturbation-specific information via adversarial training. The latent embedding is then combined with perturbation and covariate embeddings, scaled by a learned dose-response function for cell-specific predictions. The decoder reconstructs gene expression profiles, enabling predictions for unseen perturbations. We followed the authors' tutorial and used the default hyperparameters for implementation.

scPreGAN(v1.0) [55] integrates autoencoders with generative adversarial networks (GANs) to predict cellular responses to perturbations. The architecture includes an encoder, two generator networks, and two discriminators. The encoder extracts features, while the generators predict perturbed data from latent vectors. Pre-training

combines reconstruction and adversarial losses, optimizing data fidelity and distribution realism. We implemented scPreGAN based on the authors' GitHub tutorial, adhering to default settings and niter set to 1000.

Biolord(v0.03) [56] utilizes a deep generative framework to disentangle single-cell data into known and unknown attributes through a decomposed latent space. The model separates each known attribute (cell type, time, perturbation) into dedicated subnetworks while capturing unknown attributes in regularized embeddings. A generator module reconstructs cellular measurements from these disentangled representations. The framework is optimized using a composite loss function that balances reconstruction accuracy with information minimality constraints, enabling counterfactual predictions by modifying specific attributes while preserving cell-specific signatures. We implemented Biolord following the official tutorial with default parameters and max_epochs set to 1000.

Evaluation metrics

In evaluating model performance, selecting the appropriate assessment metrics is essential. To provide a comprehensive evaluation, this study employs a standardized, multi-dimensional evaluation framework that integrates four complementary categories of metrics: absolute accuracy, relative effect capture, differential expression recovery, and distributional similarity. These metrics offer a holistic view of the model's predictive accuracy, biological interpretability, and distributional consistency across various perturbation scenarios.

Specifically, absolute accuracy is assessed through error-based measures (such as MSE, RMSE, MAE, and L2 loss), which quantify the deviations between predicted and actual gene expression values; goodness-of-fit metrics (such as R^2), which evaluate the model's ability to account for overall data variability; and correlation analysis (including Pearson correlation and cosine similarity), which measures the linear relationship and directional alignment between predicted and actual values. Relative effect capture is evaluated by computing Pearson correlation on delta values (expression changes between perturbed and control conditions) to assess the model's ability to predict perturbation-induced changes, along with delta direction accuracy, which quantifies the proportion of genes with correctly predicted up- or down-regulation patterns. Differential expression recovery is measured through DE overlap accuracy, which assesses the agreement between predicted and observed differentially expressed genes, as well as Spearman correlation of effect sizes, which evaluates the consistency of perturbation magnitudes across different conditions. Finally, distributional similarity metrics (such as MMD and Wasserstein distance) assess the alignment between predicted and actual gene expression distributions, capturing the nuanced distributional characteristics in high-dimensional data.

Absolute Accuracy Metrics

Mean Squared Error(MSE) is a commonly used metric for evaluating model error, quantifying the mean squared difference between predicted and actual values. MSE is

particularly suited to assessing prediction bias and overall accuracy, given its sensitivity to larger errors. In calculation, MSE squares the prediction error for each sample and then averages these values. The formula is as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of samples. **Root Mean Squared Error(RMSE)** is similar to MSE, being the square root of MSE, but its units are consistent with the original data, providing a more intuitive reflection of the magnitude of prediction errors. It is obtained by taking the square root of MSE, defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

Mean Absolute Error(MAE) measures the absolute error between predicted and actual values, is less sensitive to outliers than MSE, and is suitable for scenarios where one does not wish to penalize large errors. MAE was calculated by taking the average of the absolute difference between the predicted and actual values, defined as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

L2 Norm(L2) is used to measure the Euclidean distance between predicted and actual values, serving as an effective tool for describing the error space. It calculates the L2 norm of the difference vector between predicted and actual values, defined as follows:

$$L2 = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

R-squared(R^2) is used to measure a model's ability to explain the total variance, serving as an indicator of model fit. It represents the degree of correlation between predicted and actual values and is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

In this formula, $\bar{y}$ is the mean of the actual values. The numerator, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, represents the residual sum of squares (RSS), which is the total squared error. The denominator, $\sum_{i=1}^{n}(y_i - \bar{y})^2$, represents the total sum of squares (TSS) of the actual values.

Pearson's Correlation measures the linear relationship between predicted and actual values, making it suitable for analyzing the strength of association between

datasets. It is calculated by dividing the covariance between predicted and actual values by the product of their standard deviations, defined as:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{6}$$

Cosine similarity measures the directional similarity between two vectors, making it suitable for evaluating the similarity between model predictions and actual results. The cosine similarity between the predicted and actual value vectors is defined as:

$$\text{Cosine Similarity}(\mathbf{y_i}, \hat{\mathbf{y}}_\mathbf{i}) = \frac{y_i \cdot \hat{y}_i}{\|y_i\| \times \|\hat{y}_i\|} = \frac{\sum_{i=1}^{n} y_i \hat{y}_i}{\sqrt{\sum_{i=1}^{n} y_i^2} \times \sqrt{\sum_{i=1}^{n} \hat{y}_i^2}} \tag{7}$$

Relative Effect Capture Metrics

Pearson's Correlation Δ measures the change in the linear relationship between the predicted and actual values when adjusted for control conditions. This metric is designed to assess how the perturbation affects the correlation between predicted and actual values relative to the control condition. Specifically, Pearson's Correlation Δ is calculated by first subtracting the control values from both the predicted and actual data, and then computing the Pearson correlation coefficient on the adjusted data.

Delta Direction Accuracy evaluates a model's ability to correctly predict the directionality of gene expression changes under perturbation conditions. This metric assesses whether genes are predicted to be upregulated or downregulated in the same direction as observed experimentally. For each perturbation condition, we first compute the expression changes relative to the control by calculating delta values for both predicted and true expression profiles. The delta direction accuracy is then defined as the proportion of genes where the predicted and observed expression changes have the same directional sign, calculated as:

$$\Delta \text{ Agreement Acc.} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\text{sign}(\Delta\hat{y}_i) = \text{sign}(\Delta y_i)] \tag{8}$$

where

$$\Delta\hat{y}_i = \hat{y}_{i,\text{pert}} - \hat{y}_{i,\text{ctrl}}, \quad \Delta y_i = y_{i,\text{pert}} - y_{i,\text{ctrl}} \tag{9}$$

represent the predicted and true expression changes for gene i, respectively. Here, $\hat{y}_{i,\text{pert}}$ and $y_{i,\text{pert}}$ denote the mean predicted and true expression values under perturbation, while $y_{i,\text{ctrl}}$ represents the mean control expression. The indicator function $I[\cdot]$ equals 1 when the condition is true and 0 otherwise, and $\text{sign}(\cdot)$ denotes the sign function. This metric takes values of 0 or 1, where 1 indicates perfect directional agreement and 0 indicates directional disagreement.

Differential Expression Recovery Metrics

DE Overlap Accuracy evaluates how well perturbation prediction models recover the correct differentially expressed genes compared to ground truth experimental

data. This metric assesses the biological relevance of predictions by measuring the concordance between predicted and observed transcriptional responses.

For each perturbation condition, we first identify differentially expressed genes (DEGs) by comparing both predicted and true perturbed expression profiles against their respective control conditions using the Wilcoxon rank-sum test. DEGs are filtered based on adjusted p-values (Benjamini-Hochberg correction) with a significance threshold of $p_{adj} < 0.05$ and ranked by absolute log fold change. The DE overlap accuracy is computed at multiple granularity levels by comparing the top-ranked DEGs between predicted and true conditions. For each top-N threshold ($N \in 10, 20, 50, 100, 200, 500$), the metric is defined as:

$$\text{Accuracy}_{Top-N} = \frac{|\text{DEG}_{\text{true}}^{(N)} \cap \text{DEG}_{\text{pred}}^{(N)}|}{N} \tag{10}$$

where DEGtrue$^{(N)}$ and DEGpred$^{(N)}$ represent the sets of top-N DEGs from true and predicted data, respectively, and $|\cdot|$ denotes set cardinality. If the number of significant DEGs in either the predicted or true condition is less than N, the accuracy for that N is not calculated. A maximum of 500 cells per condition is used, with random sampling applied when necessary. The metric ranges from 0 to 1, with 1 indicating complete overlap between predicted and true DEGs.

Effect Size is defined as the number of significantly differentially expressed genes (DEGs) per condition (N_{sig}, $p_{adj} < 0.05$). The distribution of effect sizes in the true datasets is examined to characterize the transcriptional response across conditions. The relationship between predicted and true effect sizes is quantified using the Spearman rank correlation:

$$\rho_{\text{Spearman}} = \text{rank correlation}(N_{\text{sig}}^{\text{true}}, N_{\text{sig}}^{\text{pred}}) \tag{11}$$

where each vector contains paired effect sizes for all conditions in a dataset. Conditions are excluded from the calculation if there are fewer than 5 valid paired observations, or if either the predicted or true values have zero variance.

Distributional Similarity Metrics

Maximum Mean Discrepancy(MMD) is used to measure the distance between two distributions, making it suitable for evaluating the difference between predicted and actual distributions. Specifically, we computed per gene (i.e., it quantifies the distributional distance between the predicted and observed expression values across all cells for that gene). After obtaining one MMD value for every gene, we then aggregate these per-gene distances to yield a single scalar that summarises overall distributional fidelity for the perturbation condition. It calculates the mean discrepancy between the predicted and actual data using a kernel function, defined as:

$$\text{MMD}^2 = \mathbb{E}[k(y_i, y_i')] + \mathbb{E}[k(\hat{y}_i, \hat{y}_i')] - 2\mathbb{E}[k(y_i, \hat{y}_i)] \tag{12}$$

Here, y_i and $\hat{y}_i$ represent samples from two different distributions, while y_i' and $\hat{y}_i'$ are samples within the same distribution. $k(y_i, \hat{y}_i)$ is the kernel function, and we use the RBF kernel in this study.

Wasserstein Distance is used to measure the geographical distance between two probability distributions and is commonly used to assess the similarity between distributions. By solving the optimal transport problem, it calculates the cumulative discrepancy between the two distributions, defined as:

$$W(P,Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(y_i,\hat{y}_i)\sim\gamma} \left[\|y_i - \hat{y}_i\| \right] \tag{13}$$

In the formula, P and Q are two different probability distributions, and $\gamma \in \Pi(P,Q)$: represent the set of all joint distributions from P to Q.

Ranking-based meta-analysis

We conducted a comprehensive ranking-based meta-analysis to evaluate model performance across four standardized dimensions. The evaluation framework comprised four key dimensions, each weighted equally at 25%: (1) Absolute Accuracy, measuring prediction accuracy through error metrics (MSE, RMSE, L2, MAE) and correlation metrics (Pearson correlation, cosine similarity, R^2); (2) Relative Effect Capture, assessing perturbation effect detection using Pearson correlation delta and delta agreement accuracy; (3) Differential Expression Recovery, evaluating DE gene identification capability through Top50 accuracy and Spearman correlation; and (4) Distribution Similarity, measuring distributional fidelity via Maximum Mean Discrepancy (MMD) and Wasserstein distance. For the first, second, and fourth dimensions, metrics were calculated for both all genes and the top 20 differentially expressed genes, while the third dimension used all genes only.

For score calculation, models were ranked within each metric for each task separately. Error-based metrics (MSE, RMSE, L2, MAE, MMD, Wasserstein, and execution time) were reverse-ranked, where lower values received higher normalized scores. Each dimension score was computed as the average of normalized rank scores for all constituent metrics within that dimension. The Overall Score was calculated as the unweighted average of the four dimension scores, excluding time and user-friendliness metrics, which were displayed separately for reference. Missing values were imputed using the mean value of the respective metric across all models within each task to ensure complete comparisons.

The visualization employed a multi-element approach where point size represents normalized metric values within each column (larger points indicating higher values), point color indicates rank within each column using the RdGy colormap (darker colors representing better ranks), horizontal bar charts display Overall Scores with numerical labels, and star symbols represent user-friendliness levels (1 star = good, 2 stars = better). This ranking-based meta-analysis provides task-specific model comparisons while maintaining consistency across evaluation criteria and effectively handling the heterogeneity of available metrics across different models and datasets.

TRADE

To assess the transcriptome-wide impact of perturbations, we employed the TRADE (Transcriptome-wide Analysis of Differential Expression, TRADEtools v 0.99.0)[57]

method, which quantifies the total effect of a perturbation across the entire transcriptome. TRADE addresses the challenges of noise and undetected effects that are common in single-cell CRISPR screens, such as Perturb-seq, by providing a robust estimation of the true differential expression effects. The analysis was performed following the tutorial provided by the authors. Differential expression analysis was carried out using DESeq2(v 1.42.1), incorporating both condition and batch information to account for potential variations between replicates. The differential expression results were then integrated into the TRADE framework, which calculates a transcriptome-wide impact score. This score reflects the global shift in gene expression distributions caused by the perturbation relative to the control condition. TRADE's capability to summarize these effects enables the systematic ranking and comparison of perturbations across multiple datasets.

E-distance

E-distance is a metric used to measure the difference between two distributions in high-dimensional space, defined as the mean squared Euclidean distance between samples. In this study, we employ E-distance to quantify the high-dimensional distribution differences across different datasets, utilizing scPerturb (v0.1.0) for the calculations. For samples originating from two distributions, we first compute the Euclidean distance between each pair of samples and then calculate the average distance within each distribution as well as the average distance between the two distributions. Prior to these calculations, we standardize the single-cell data, filter out low-expression cells (min_counts=100) and low-expression genes (min_cells=50), and select the top 2000 highly variable genes as features for analysis. Detailed methodologies can be referred to in the scPerturb[8] publication.

Statistical Analysis

All statistical analyses were conducted using Python (v3.12.3) with the pandas (v1.5.3), statsmodels (v0.14.2), and NumPy (1.26.4) libraries. For each task, we performed ANOVA tests to assess the significance of performance differences across models and various experimental factors specific to each task design. Task-specific ANOVA models were implemented using ordinary least squares (OLS) regression, where all variables were treated as categorical factors using the `C()` notation in `statsmodels`.

For Task 1, we used the model:

$$\text{value} \sim C(\text{perturbation}) + C(\text{dataset}) + C(\text{model})$$

where perturbation, dataset and model served as categorical factors.

For Task 2, we used:

$$\text{value} \sim C(\text{model}) + C(\text{seen level}) + C(\text{additive type})$$

For Task 3, we used:

$$\text{value} \sim C(\text{model}) + C(\text{perturbation}) + C(\text{cell_type}) + C(\text{dataset})$$

Post-hoc analysis with Tukey's HSD test was conducted to perform pairwise comparisons between models when a significant effect ($p < 0.05$) was detected. All analyses were performed using Type II Sum of Squares for ANOVA, and results were considered statistically significant at $p < 0.05$.

Perturbation prediction on unseen genes

To assess the model's response to perturbations in single genes that it has not seen, we employ a masking strategy during model training to simulate the absence of specific genes in the training dataset. Specifically, there are multiple choices for perturbed genes in each dataset, and we select some perturbed genes as unseen genes through random sampling. During the training process, the perturbation response data for these genes are completely excluded from the training set, forcing the model to learn based solely on the perturbation features of other genes. Subsequently, the model predicts the perturbation responses for these unseen genes in the test set to evaluate its performance on unknown single-gene perturbations.

Perturbation prediction on combinatorial perturbation

For predicting combinatorial gene perturbations, we employed a random sampling strategy to select multiple gene pairs from each dataset, similar to the GEARS paper. We designed the following three training configurations to evaluate the models' generalization capabilities under varying levels of prior knowledge: Seen0: In this configuration, neither the gene pair nor their individual single-gene perturbations appear in the training data. This setup assesses the model's ability to predict responses to completely novel gene combinations. Seen1: The training set includes single-gene perturbation data for one gene of the pair, while the other gene remains unseen. This configuration tests whether the model can leverage partial prior information to predict the effects of combinatorial perturbations. Seen2: The training data contains single-gene perturbation data for both genes in the pair, but their combined perturbation is not included in the training stage. This setup evaluates the model's capacity to infer combinatorial perturbation responses based on individual gene perturbation information. By randomly selecting gene pairs, we ensured that the perturbation response of each gene combination in the test set was entirely novel to the model, thereby providing a comprehensive evaluation of its generalization performance in predicting combinatorial gene perturbations.

Perturbation prediction on unseen cell types

In the cell state transfer prediction task, we explored the models' generalization capabilities across different cell types. We trained the models using perturbation data from multiple cell types or cell lines, to predict the perturbation responses of a target cell

type that was masked during training. Specifically, the training set included perturbation response data from only a subset of cell types, while data from the target cell type were completely excluded. The models were then tasked with predicting the perturbation responses for the unseen target cell type based on the information learned from the training set. This approach allows us to evaluate the models' applicability and robustness across different biological contexts.

Data availability

The raw data of the dataset generated in this study can be accessed through the BioProject service of the China National Center for Bioinformation (https://www.cncb.ac.cn/), with the project registration number PRJCA033051. Except for the data unique to this study, all other datasets used are from public resources. For specific details on their usage, please refer to the "Methods" section of the manuscript. A comprehensive list of datasets and their download sources can be found in Table 1 and Supplementary Table 1.

Code availability

The scripts used in this study are available via GitHub at: https://github.com/TianGzlab/scPerturbBench. A website hosting the data and the visualization can be viewed at: luyitian.github.io/PerturbArena. The website is continually updated and may contains new model and show different results compared to the manusciprt.

Acknowledgements

We acknowledge the support of the Data Science Platform of Guangzhou National Laboratory and the Bio-medical Big Data Operating System (Bio-OS). We thank Lead Healthcare.AI for their help on experiments.

Author contributions

L.T. and Y.Y. designed and supervised the study. L.L.Y.Y. Y.L performed the baseline model and data analyses with help from Y.C, Y.F, S.Z, and J.K. X.F, S.L, B.L, and W.R performed experiments to generate in-house datasets with help from J.C and X.L. L.L., W.L., Y.Y., and L.T. drafted the manuscript, with contributions from all authors.

Table legends

Table 1. Overview of datasets used in this study.
Supplementary Table 1. Detailed descriptions of datasets used in this study.
Supplementary Table 2. Overview of perturbation-related models.

Supplementary Table 3. List of 22 drugs used in the Perturb_KHP dataset.

Supplementary Table 4. List of 111 drugs used in the Perturb_cmo_V1 dataset.

Supplementary Table 5. Top 5 and bottom 5 ranked drugs from the drug dataset based on TRADE calculations.

Figure legends

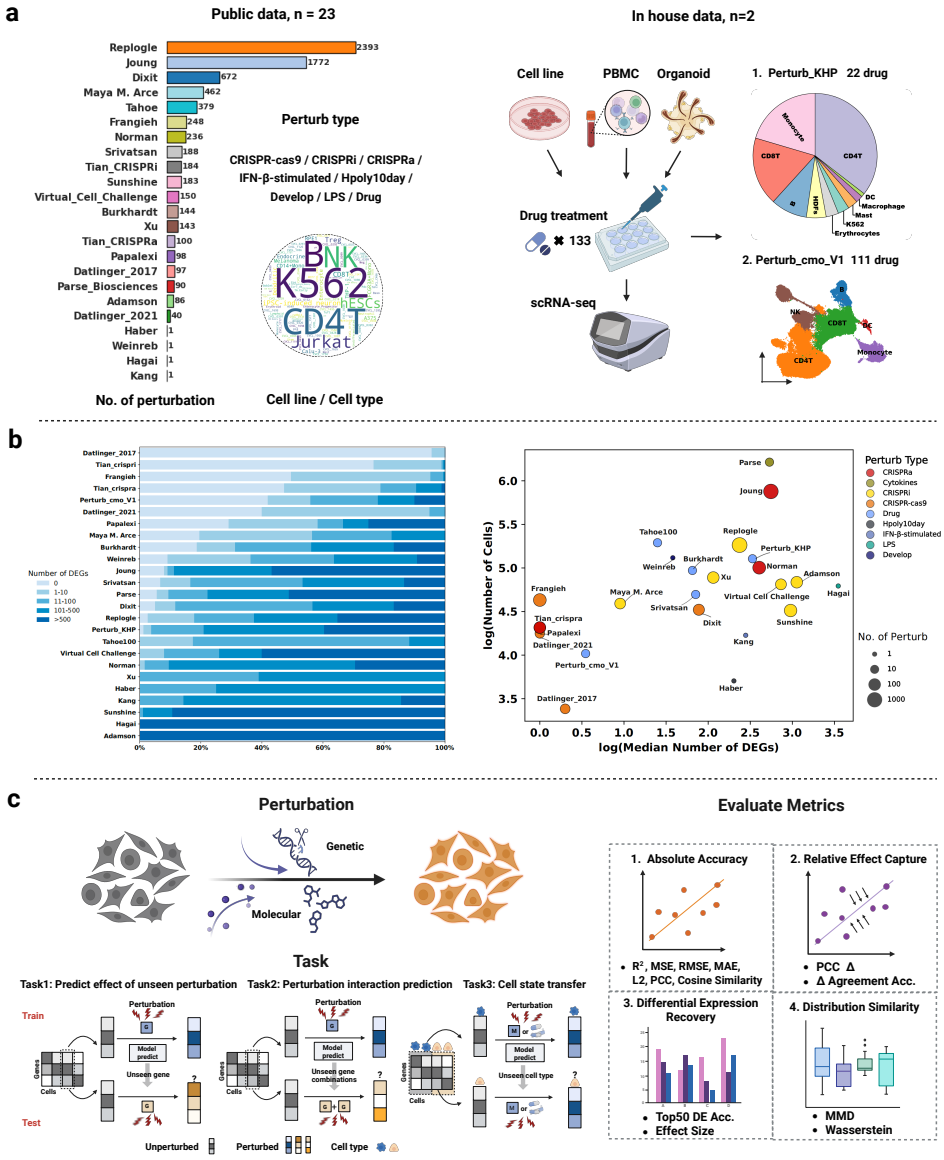


Fig. 1 Benchmarking datasets, perturbation properties, and experimental design. (a) Data overview. The benchmarking collection integrates 23 public datasets and 2 in-house datasets. Public resources include large-scale CRISPR screens, drug perturbation atlases, and immune stimulation studies. The in-house datasets profile the effects of 113 drug treatments across lymphoid immune subsets, enabling systematic evaluation of cross–cell-type generalization. Together, these datasets cover a wide range of perturbation types (CRISPR–Cas9, CRISPRi, CRISPRa, drugs, IFN-β, Hpoly, LPS) across diverse cellular systems, including cancer cell lines, pluripotent stem cells, PBMCs, organoids, and non-human immune cells. (b) Perturbation effect size and dataset properties. Left: distribution of dataset components by percentage of differentially expressed genes (DEGs) per perturbation compared to controls. Right: scatterplot showing the median number of DEGs per condition (x-axis) versus the log number of cells (y-axis). Point size indicates the number of perturbations, and color encodes perturbation type, ranging from red (strongest transcriptional effects) to blue (weakest effects). (c) Benchmarking tasks and evaluation strategy. Both genetic and molecular perturbations are used in benchmarking. Models were trained on observed perturbation–response pairs and evaluated on held-out conditions. Three tasks were benchmarked: Task 1, prediction of responses to unseen perturbations; Task 2, prediction of combinatorial perturbation interactions; and Task 3, transfer of responses across unseen cell types. Performance was assessed using multiple complementary metrics to assess absolute accuracy, relative effect capture, differential expression recovery, and distributional similarity.

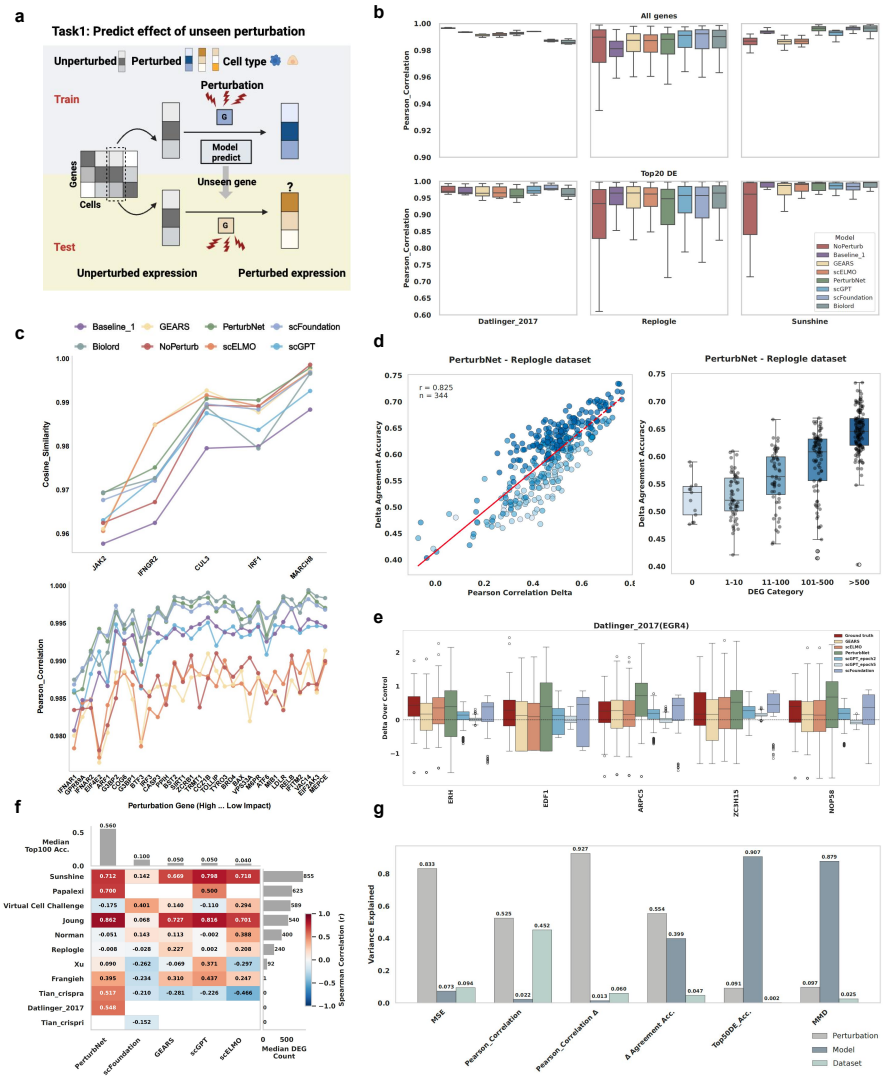


Fig. 2 Prediction of unseen single-gene perturbations across multiple models and datasets. (a) Schematic overview of the single-gene perturbation prediction task. (b) Direct gene expression-level performance metrics. Pearson's correlation coefficients on all genes and the top 20 genes across different datasets and models. (c) Cosine similarity (Papalexi dataset) and Pearson's correlation coefficients (Sunshine dataset) on selected perturbation genes across models. The perturbation genes are ranked by impact strength. (d) Delta-based analysis of perturbation effect size and prediction accuracy predicted by PerturbNet model on Repologle dataset. Pearson correlation and delta direction accuracy were positively correlated (left), with larger-effect perturbations yielding higher delta direction accuracy (right). (e) Distribution comparison of the top 5 genes between predicted and observed expression changes on EGR4 perturbation of Datilnger 2017 datasest. (f) Correlation of effect sizes between true and predicted perturbations relative to control. Median overlap accuracy of the top 100 DEGs across datasets is shown above, and median DEG count per dataset is shown on the right. Models are ranked by overlap accuracy, datasets by median effect size. (g) Variance decomposition analysis showing the relative contribution of perturbation identity, model choice, and dataset identity to prediction performance across different evaluation metrics.

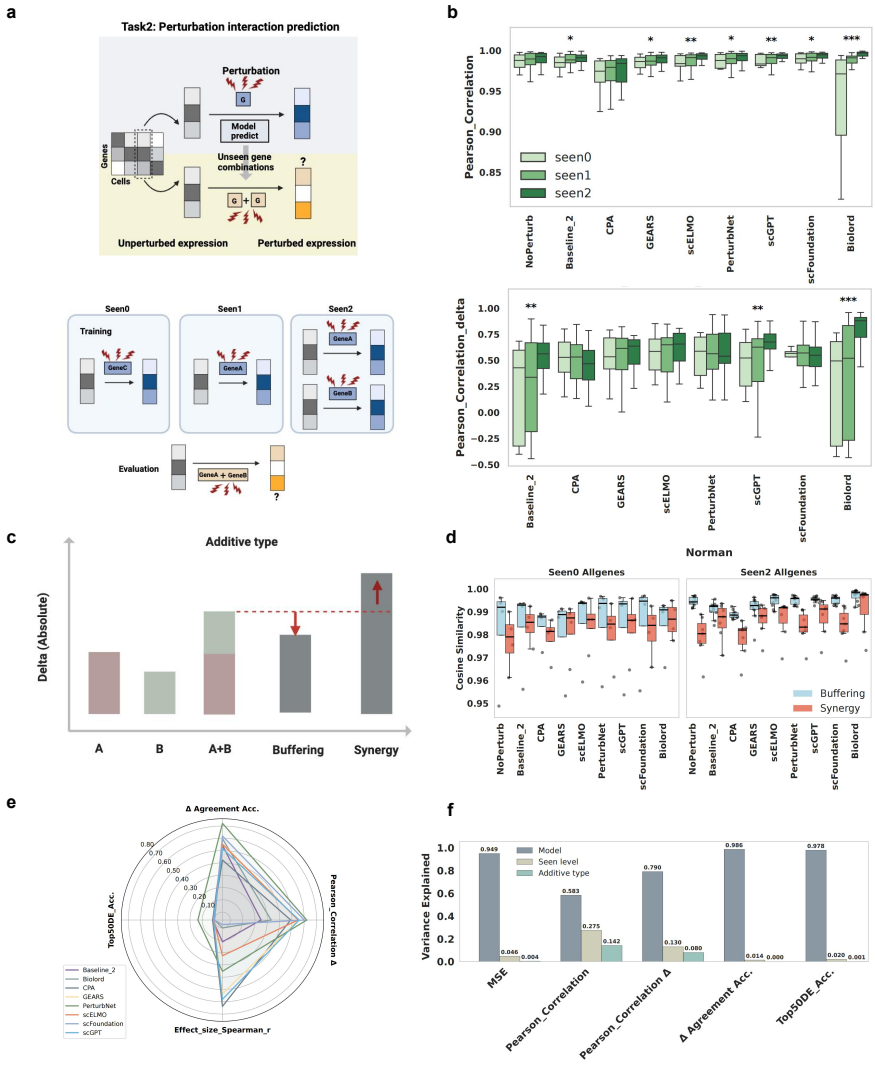


Fig. 3 Perturbation interaction prediction performance across models and interaction types (a) Schematic overview of the perturbation interaction prediction task. Bottom panel illustrates the three "seen levels" for combinatorial perturbation evaluation: seen0 (neither individual perturbation in training data), seen1 (one perturbation in training data), and seen2 (both perturbations in training data). (b) Performance metrics for Pearson's correlation coefficients on expressions and delta expressions across models comparing seen0, seen1, and seen2 conditions. Significance is shown. (c) Classification of combinatorial perturbations into buffering (combination delta smaller than additive expectation) and synergistic (combination delta larger than additive expectation) interaction types. (d) Performance comparison between buffering and synergistic combinations for seen0 and seen2 conditions across different models on all genes of Norman dataset. (e) Averaged top 50 DEG overlap accuracy, effect size correlation (Spearman r), Pearson's correlation coefficients on delta expressions, and delta direction accuracy across all tested models for perturbation interaction prediction. (f) ANOVA variance decomposition analysis results showing variance contributions from model choice and interaction type factors for different performance metrics.

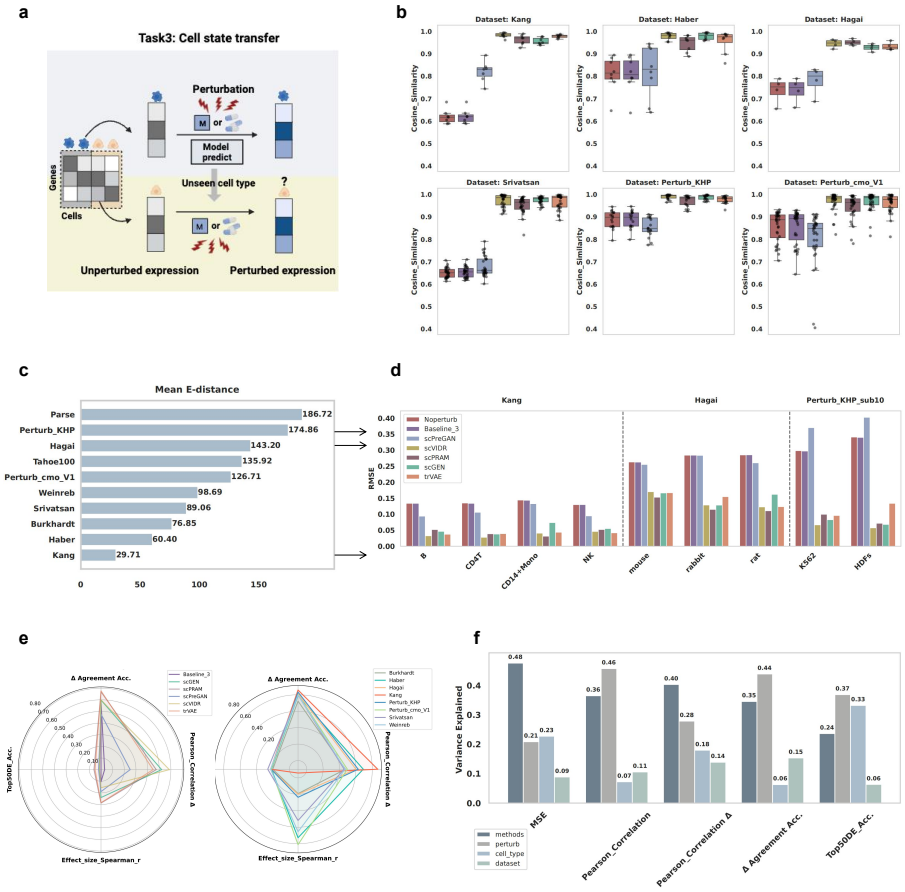


Fig. 4 Cell state transfer performance evaluation across cell types. (a) Schematic overview of the cell state transfer task framework. (b) Direct gene expression-level performance comparison across models and datasets. Cosine similarity is the evaluation metric. (c) The average E-distance among cell types for each dataset, ranked by values. (d) Model performance in predicting the responses of different cell types, evaluated using RMSE on the selected Kang, Hagai, and Perturb_KHP_sub10 datasets.. (e) Averaged top 50 DEG overlap accuracy, effect size correlation (Spearman r), Pearson's correlation coefficients on delta expressions, and delta direction accuracy across all tested models for cell state transfer prediction. (f) Variance partitioning analysis showing the relative contributions of model choice, perturbation identity, and cell type to overall transfer performance variability.

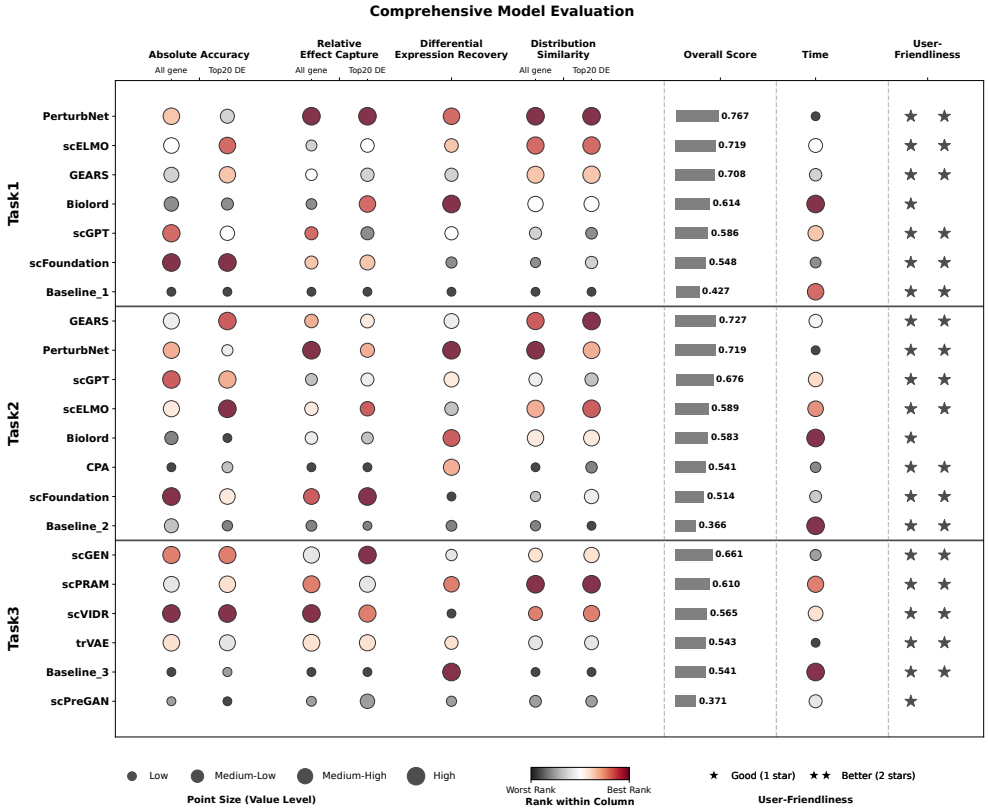


Fig. 5 Comprehensive model performance evaluation across perturbation prediction tasks. Models are evaluated across three distinct perturbation prediction tasks using four key performance dimensions: absolute accuracy, relative effect capture, differential expression recovery, and distribution similarity. Each dimension is assessed on all genes) and top20 DE, except differential expression recovery, which is evaluated only on all gene. Point size indicates the magnitude of performance values, while color intensity represents performance rank within each metric column, with larger, darker points indicating better performance. For MSE, RMSE, L2, MAE, MMD, Wasserstein, and training time, lower values indicate better performance and are reverse-ranked. Models are ranked from best (top) to worst (bottom) by overall scores calculated as the average of four dimensions (25% each), with training time and user-friendliness (1-2 star implementation ease rating) displayed but excluded from overall scoring.

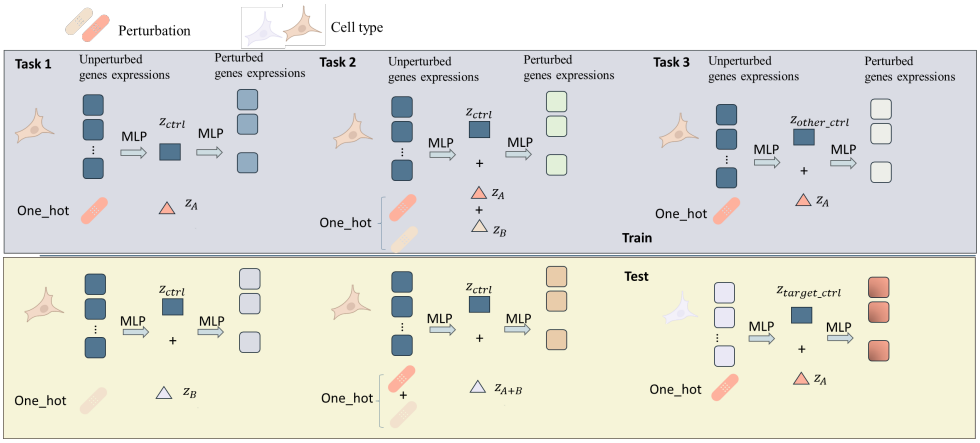


Fig. S1 Summary of baseline model used in this study

Supplementary Fig. 2

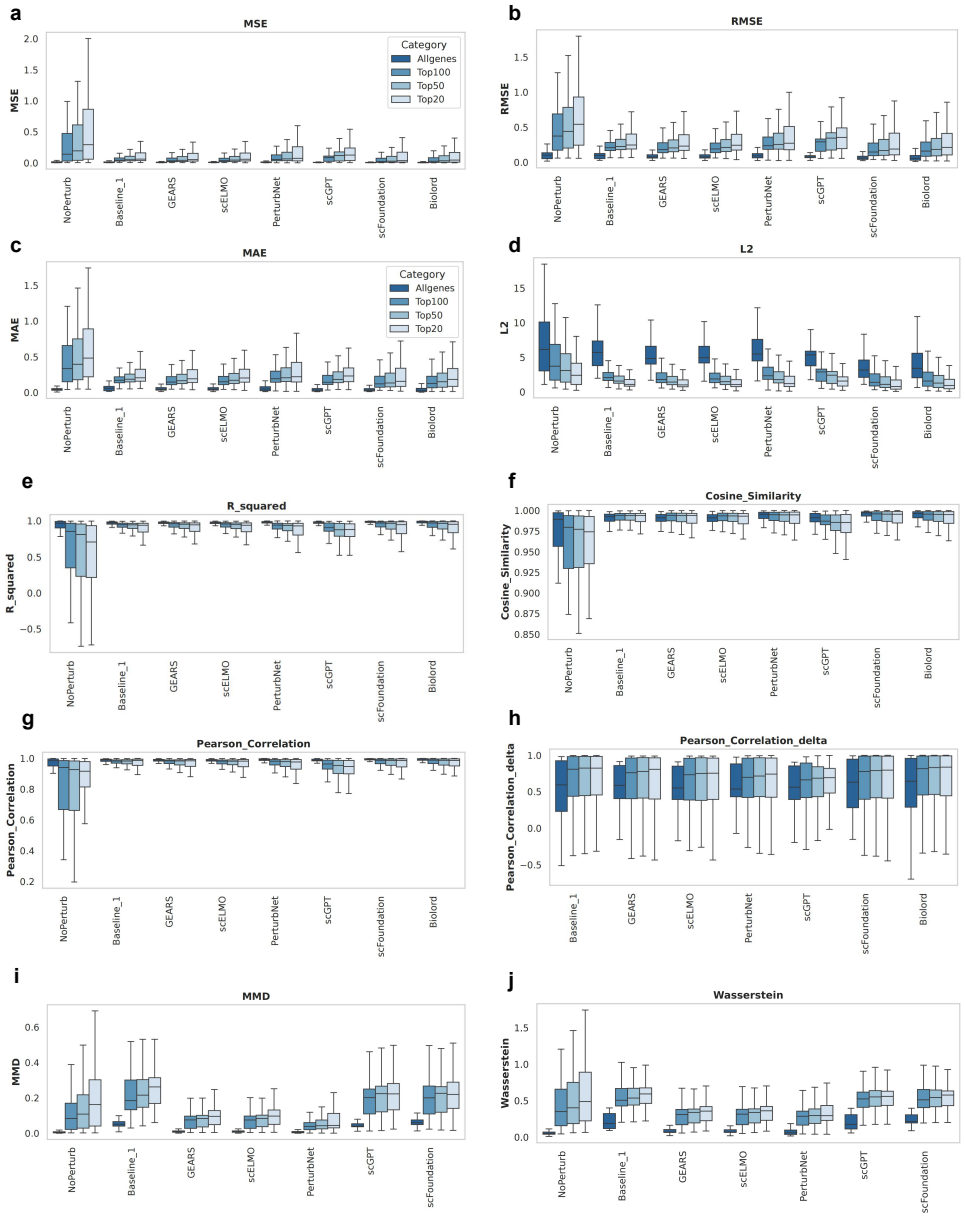


Fig. S2 Evaluation results of Task 1. Metrics aggregated on all datasets were computed across all models using four gene sets: all genes, top 100 DEGs, top 50 DEGs, and top 20 DEGs (ranked by absolute scores). Ten evaluation metrics are reported: (a) MSE, (b) RMSE, (c) MAE, (d) L2 distance, (e) R^2, (f) Cosine similarity, (g) Pearson correlation, (h) Pearson correlation delta, (i) MMD, and (j) Wasserstein distance.

Supplementary Fig. 3

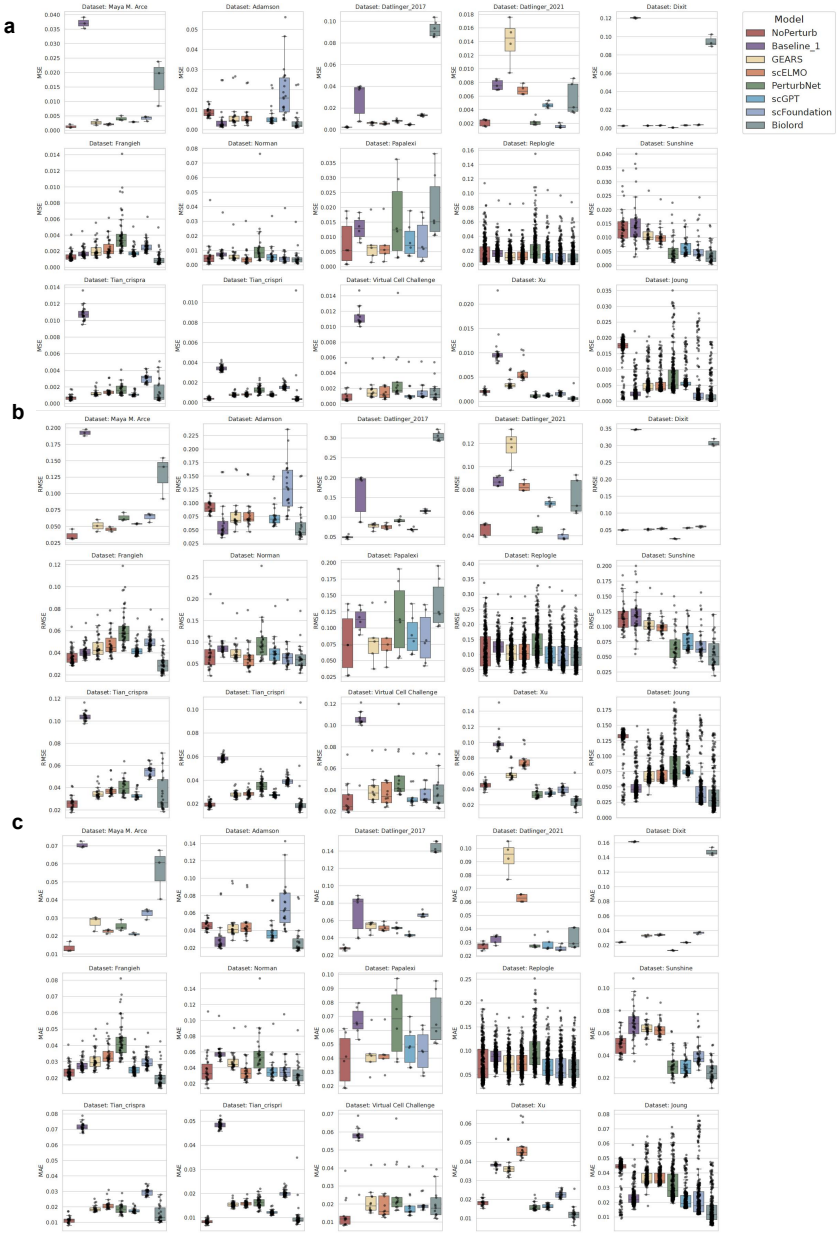


Fig. S3 Evaluation results of Task 1. Metrics were computed across all datasets and models using all genes. The evaluated datasets include Maya M. Arce, Adamson, Datlinger (2017 & 2021), Dixit, Frangieh, Norman, Papalexi, Replogle, Sunshine, Tian_crispra, Tian_crispri, Virtual Cell Challenge, Xu, and Joung. Three evaluation metrics are reported on this figure: (a) MSE, (b) RMSE, and (c) MAE.

Supplementary Fig. 4

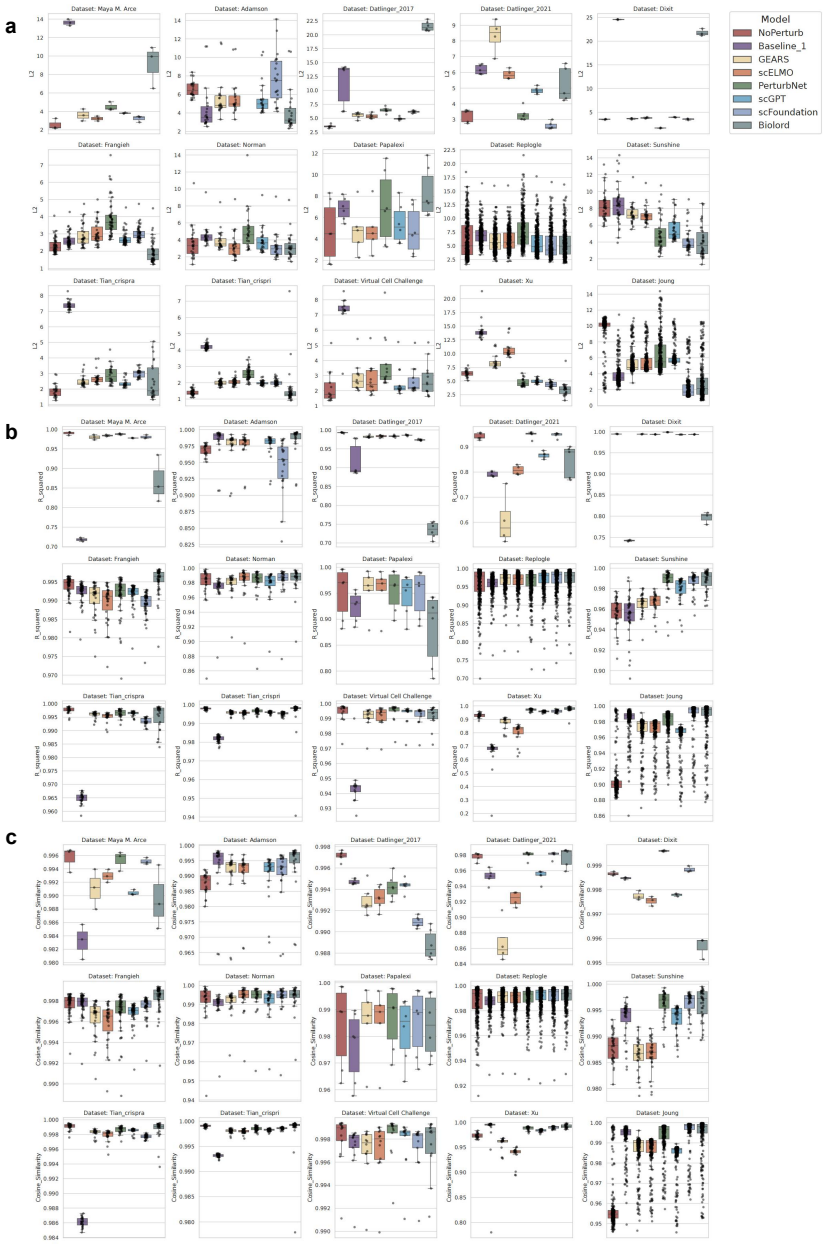


Fig. S4 Evaluation results of Task 1. Metrics were computed across all datasets and models using all genes. The evaluated datasets include Maya M. Arce, Adamson, Datlinger (2017 & 2021), Dixit, Frangieh, Norman, Papalexi, Replogle, Sunshine, Tian_crispra, Tian_crispri, Virtual Cell Challenge, Xu, and Joung. Three evaluation metrics are reported on this figure: (a) L2 distance, (b) R^2, and (c) Cosine similarity.

Supplementary Fig. 5

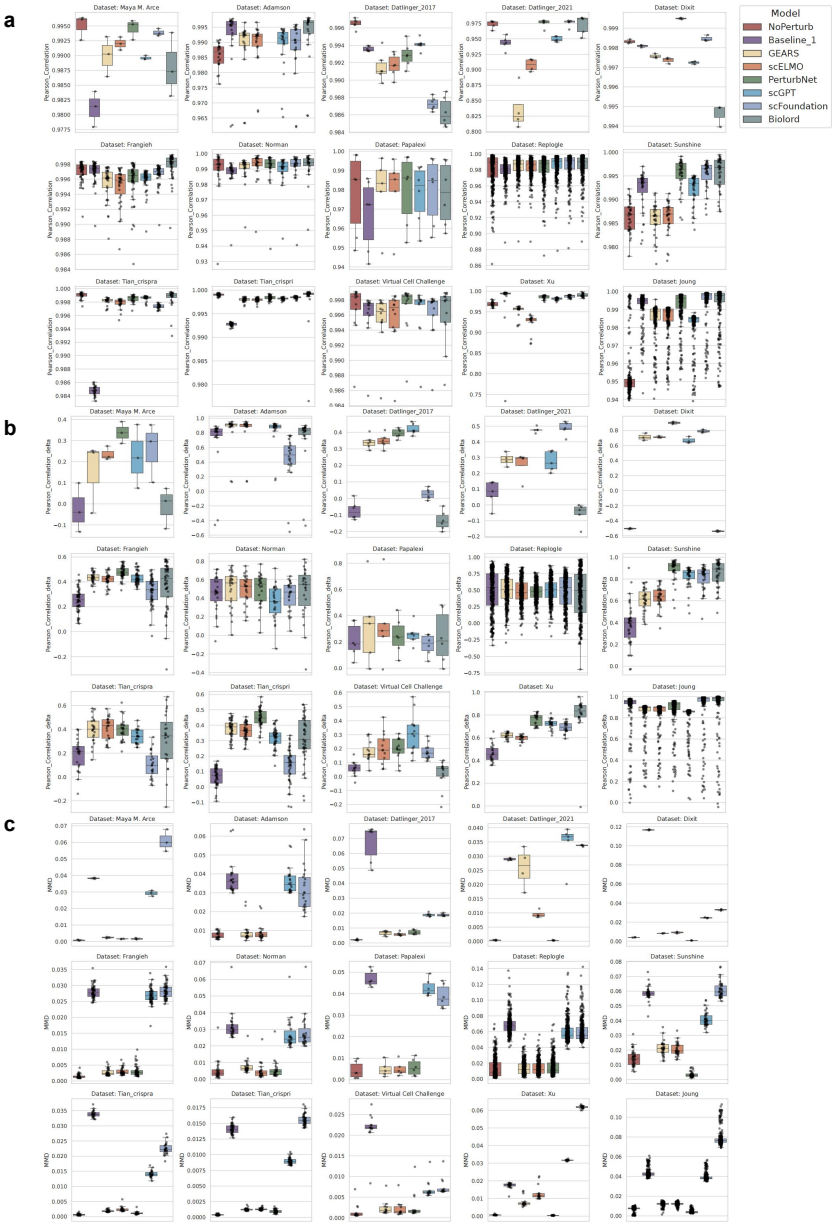


Fig. S5 Evaluation results of Task 1. Metrics were computed across all datasets and models using all genes. The evaluated datasets include Maya M. Arce, Adamson, Datlinger (2017 & 2021), Dixit, Frangieh, Norman, Papalexi, Replogle, Sunshine, Tian_crispra, Tian_crispri, Virtual Cell Challenge, Xu, and Joung. Three evaluation metrics are reported on this figure: (a) Pearson correlation, (b) Pearson correlation delta, and (c) MMD.

Supplementary Fig. 6

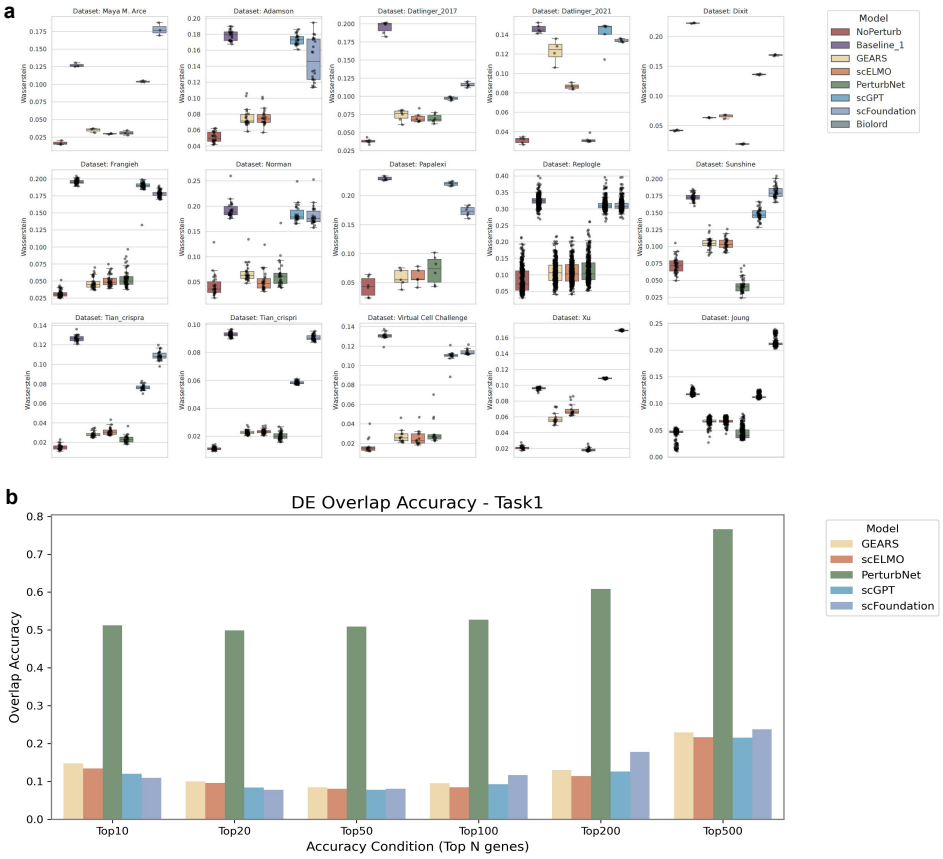


Fig. S6 Evaluation results of Task 1. (a) Wasserstein distance across all datasets and models using all genes. Datasets include Maya M. Arce, Adamson, Datlinger (2017 & 2021), Dixit, Frangieh, Norman, Papalexi, Replogle, Sunshine, Tian_crispra, Tian_crispri, Virtual Cell Challenge, Xu, and Joung. (b) Differentially expressed (DE) gene overlap accuracy between true and predicted perturbations versus control, aggregated on all datasets across all models.

Supplementary Fig.7

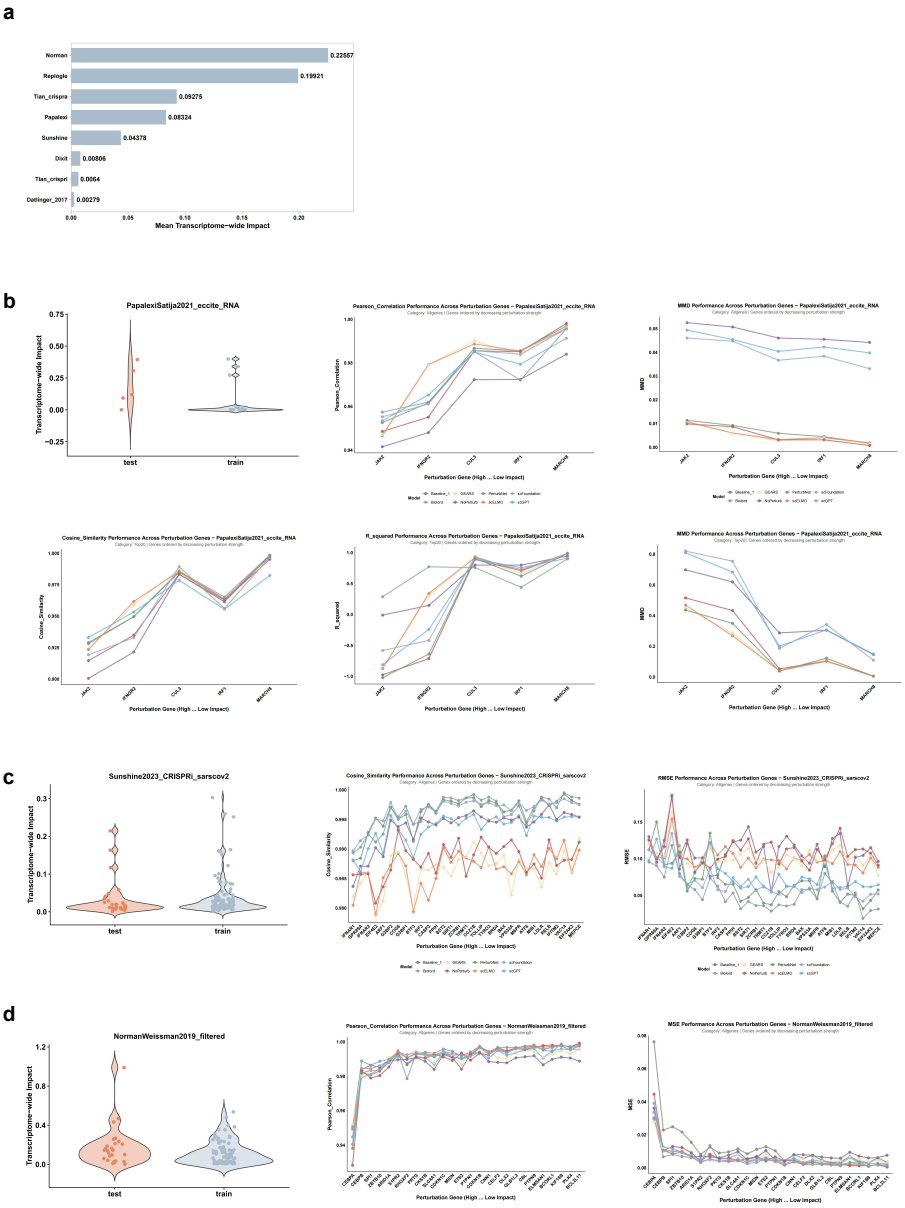


Fig. S7 Evaluation results of Task 1. (a) Mean transcriptome-wide impact across datasets. (b–d) Transcriptome-wide impact distributions on train and test data for individual datasets (left). Metrics calculated across models on gene perturbations: (b) R^2, Cosine similarity, Pearson correlation, and MMD on the Papalexi dataset; (c) Cosine similarity and RMSE on the Sunshine dataset; and (d) MSE and Pearson correlation on the Norman dataset. Perturbed genes are ordered from highest to lowest transcriptome-wide impact.

Supplementary Fig. 8

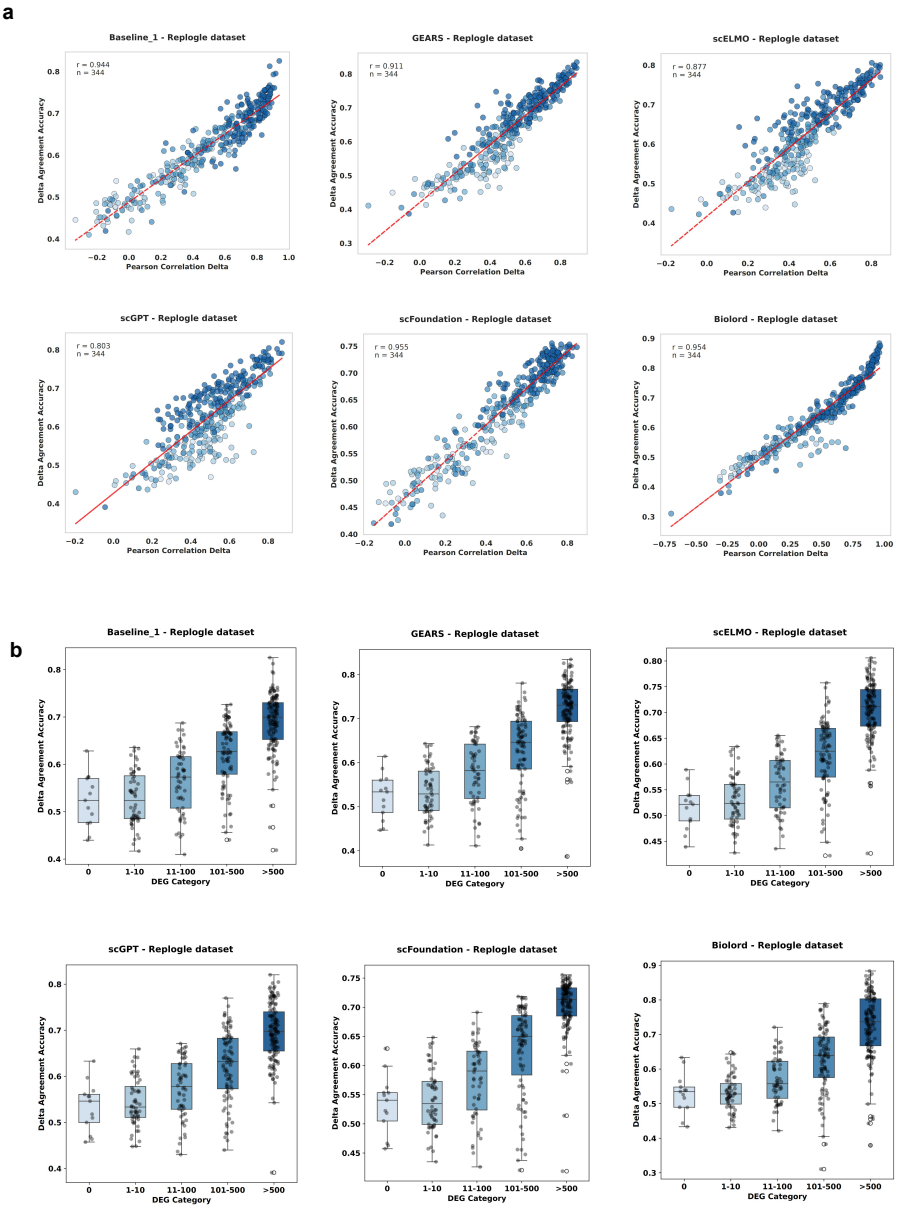


Fig. S8 Evaluation results of Task 1. (a) Delta direction agreement and prediction accuracy for the Baseline_1, GEARS, scELMO, scGPT, scFoundation, and Biolord models on the Repologle dataset. (b) Delta direction agreement across different effect size groups for the Baseline, GEARS, scELMO, scGPT, scFoundation, and biolord models on the Repologle dataset.

Supplementary Fig. 9

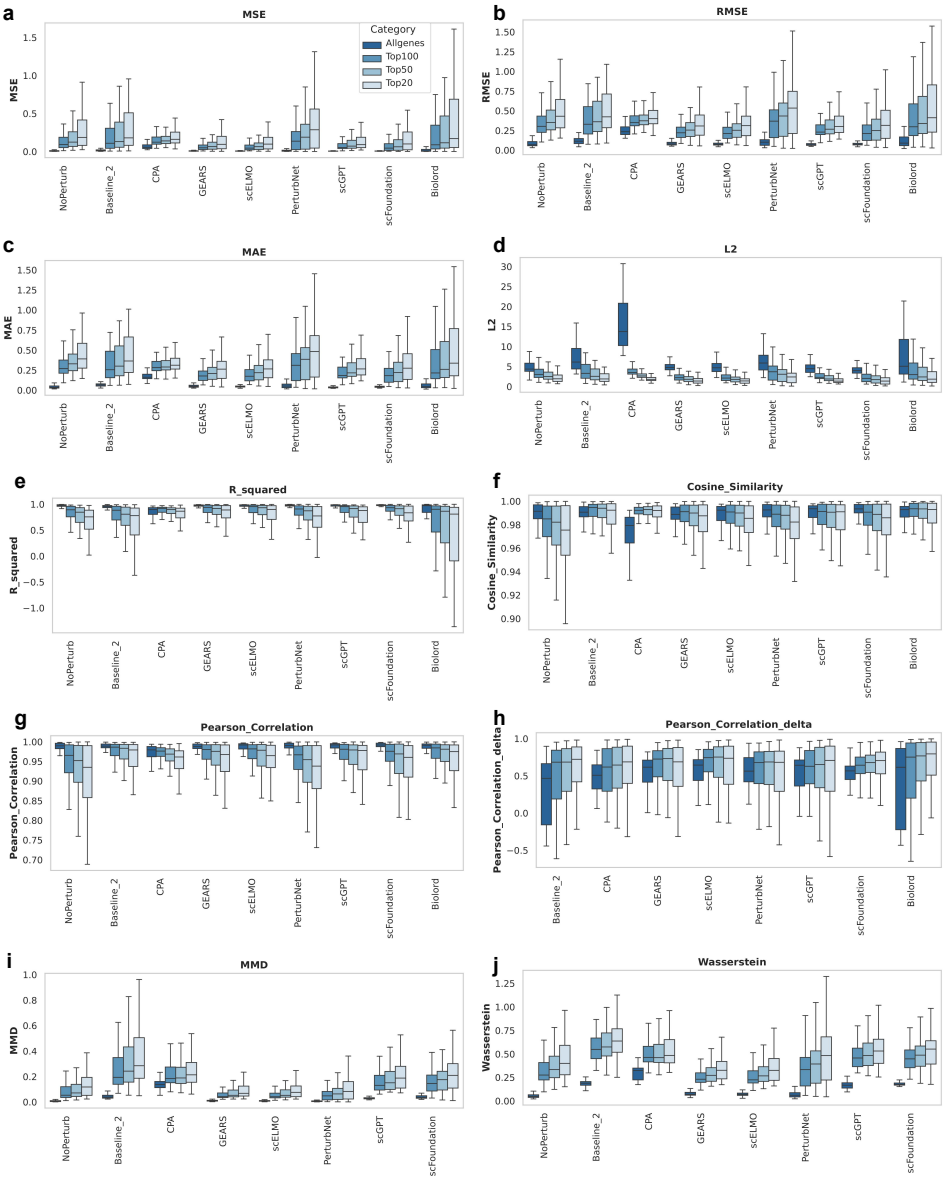


Fig. S9 Evaluation results of Task 2. Metrics aggregated on all datasets were computed across all models using four gene sets: all genes, top 100 DEGs, top 50 DEGs, and top 20 DEGs (ranked by absolute scores). Ten evaluation metrics are reported: (a) MSE, (b) RMSE, (c) MAE, (d) L2 distance, (e) R^2, (f) Cosine similarity, (g) Pearson correlation, (h) Pearson correlation delta, (i) MMD, and (j) Wasserstein distance.

Supplementary Fig. 10

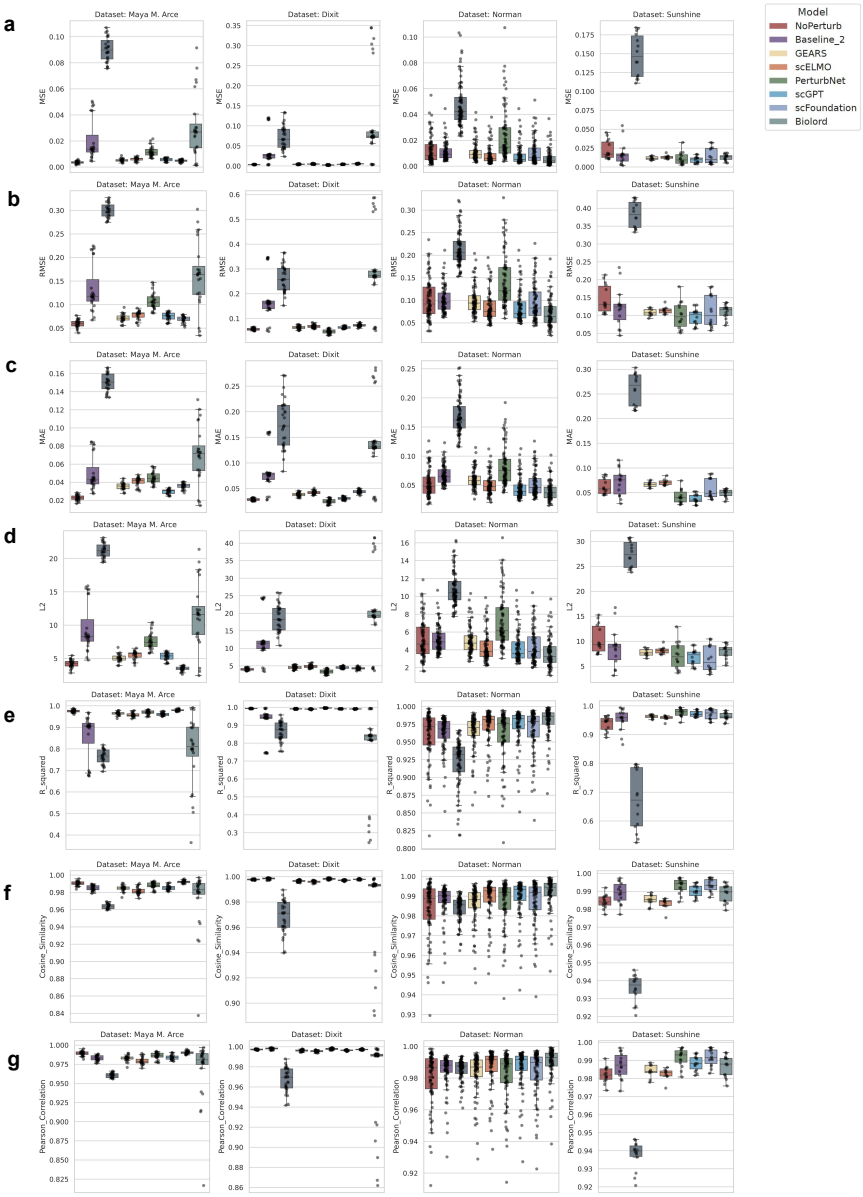


Fig. S10 Evaluation results of Task 2. Metrics were computed across all datasets and models using all genes. The evaluated datasets include Maya M. Arce, Dixit, Norman, and Sunshine. Seven evaluation metrics are reported on this figure: (a) MSE, (b) RMSE, (c) MAE, (d) L2 distance, (e) R^2, (f) Cosine similarity, and (g) Pearson correlation.

Supplementary Fig. 11

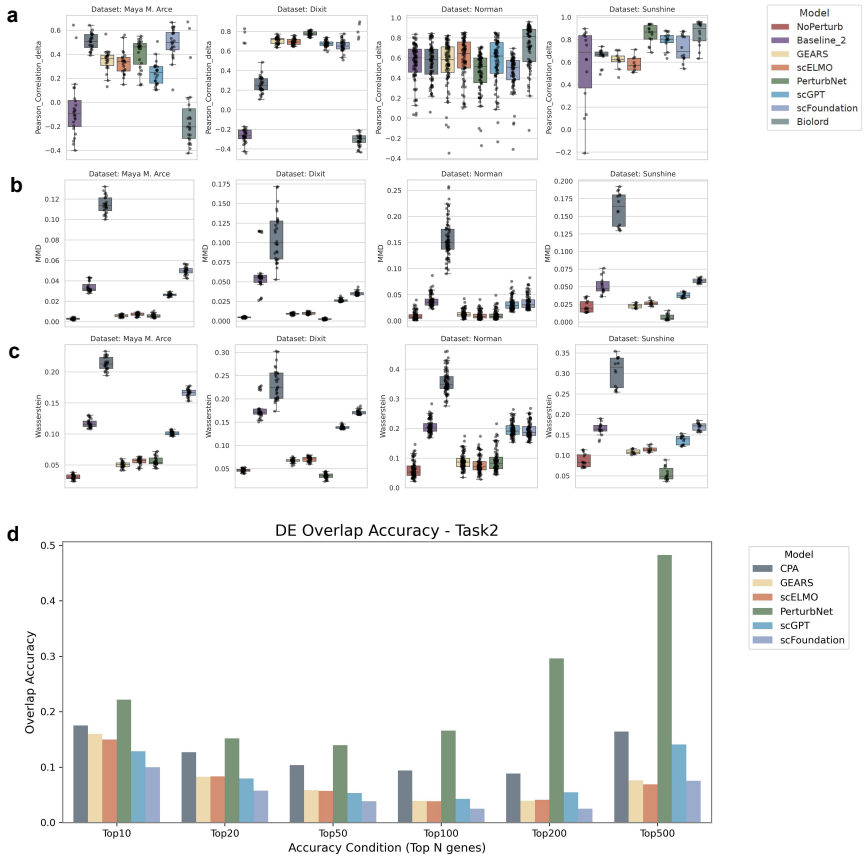


Fig. S11 Evaluation results of Task 2. (a) Pearson correlation delta, (b) MMD, and (c) Wasserstein distance across all datasets and models using all genes. Datasets include Maya M. Arce, Dixit, Norman, and Sunshine. (d) Differentially expressed (DE) gene overlap accuracy between true and predicted perturbations versus control, aggregated on all datasets across all models.

Supplementary Fig. 12

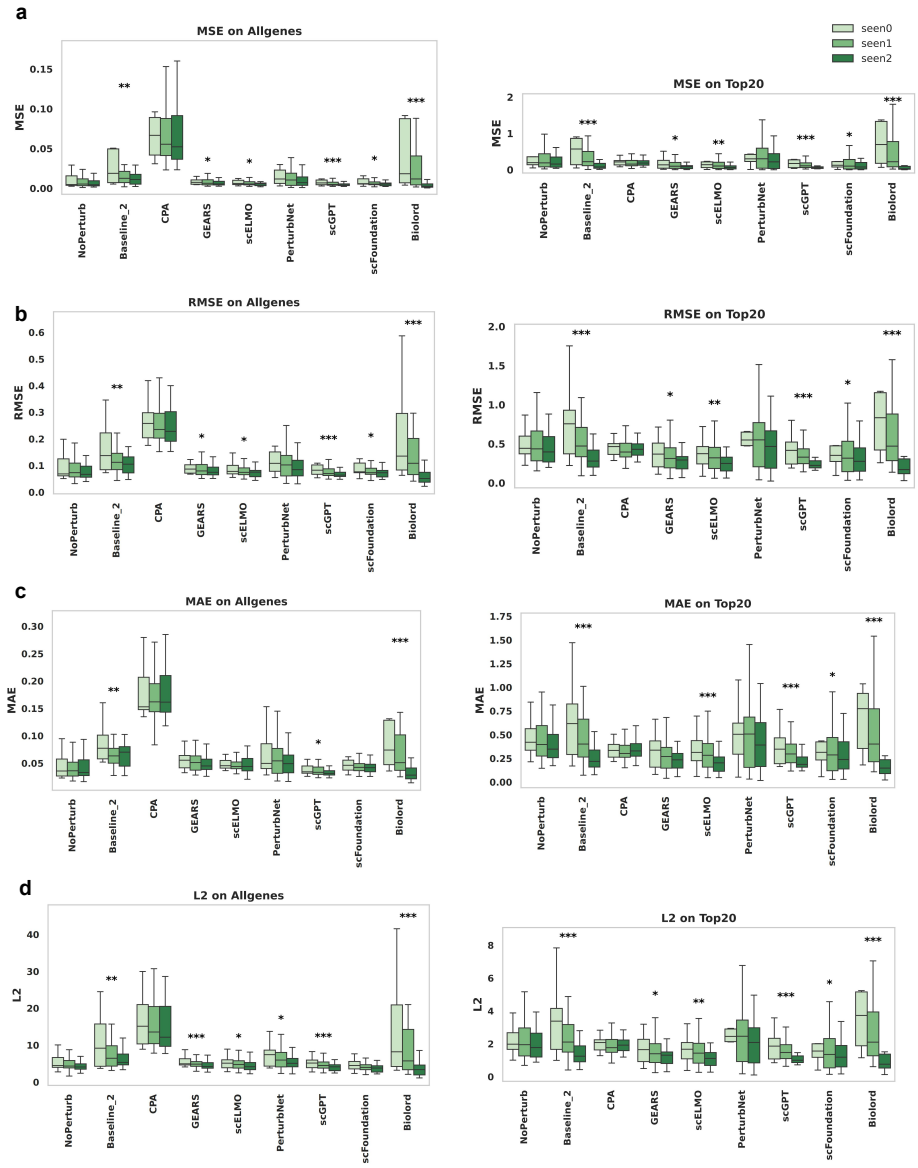


Fig. S12 Evaluation results of Task 2. Metrics aggregated across all datasets were computed for each model and seen level using all genes and the top 20 DEGs. Each model is plotted with seen0, seen1, and seen2. Four evaluation metrics are reported in this figure: (a) MSE, (b) RMSE, (c) MAE, and (d) L2 distance.

Supplementary Fig. 13

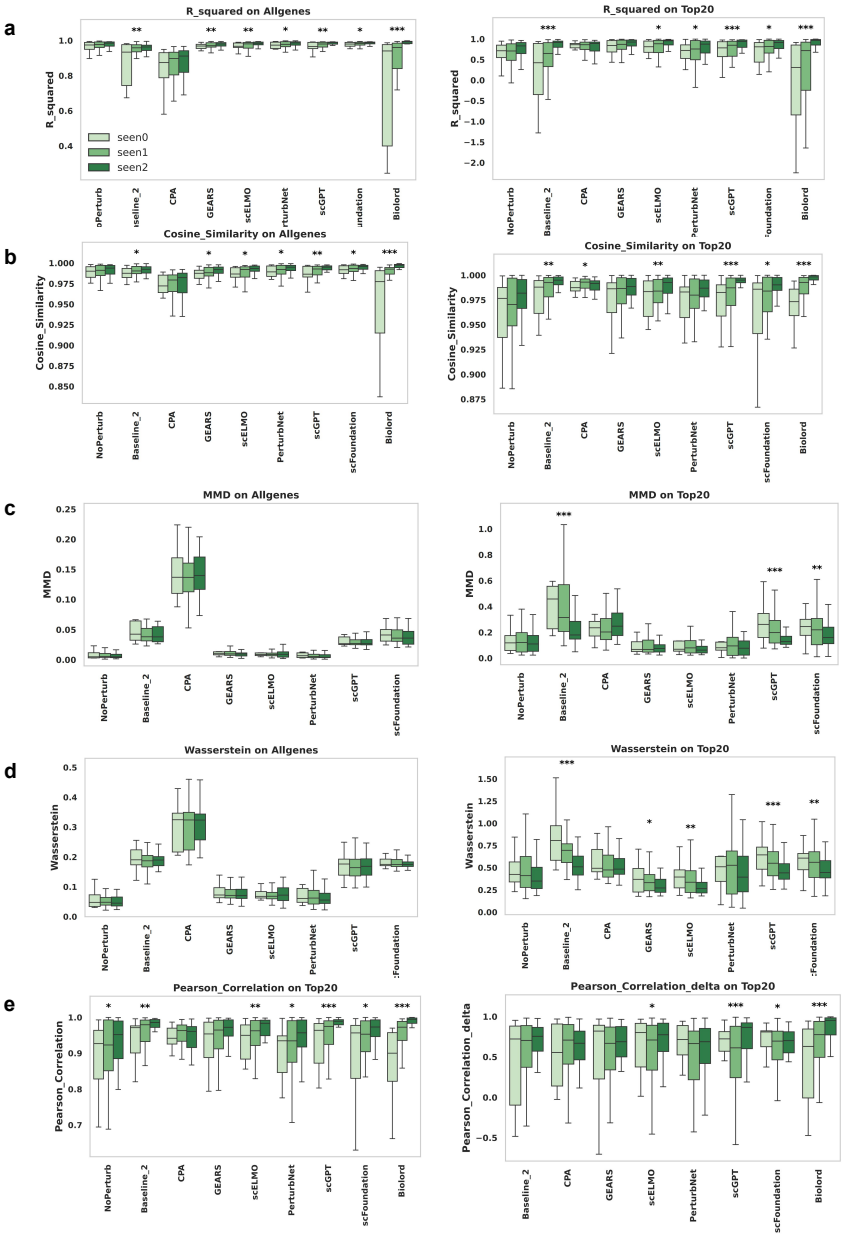


Fig. S13 Evaluation results of Task 2. Metrics aggregated across all datasets were computed for each model and seen level using all genes and the top 20 DEGs. Each model is plotted with seen0, seen1, and seen2. Four evaluation metrics are reported in this figure: (a) R^2, (b) Cosine similarity, (c) MMD, and (d) Wasserstein distance. (e) Pearson correlation and Pearson correlation delta calculated using the top 20 DEGs only.

Supplementary Fig. 14

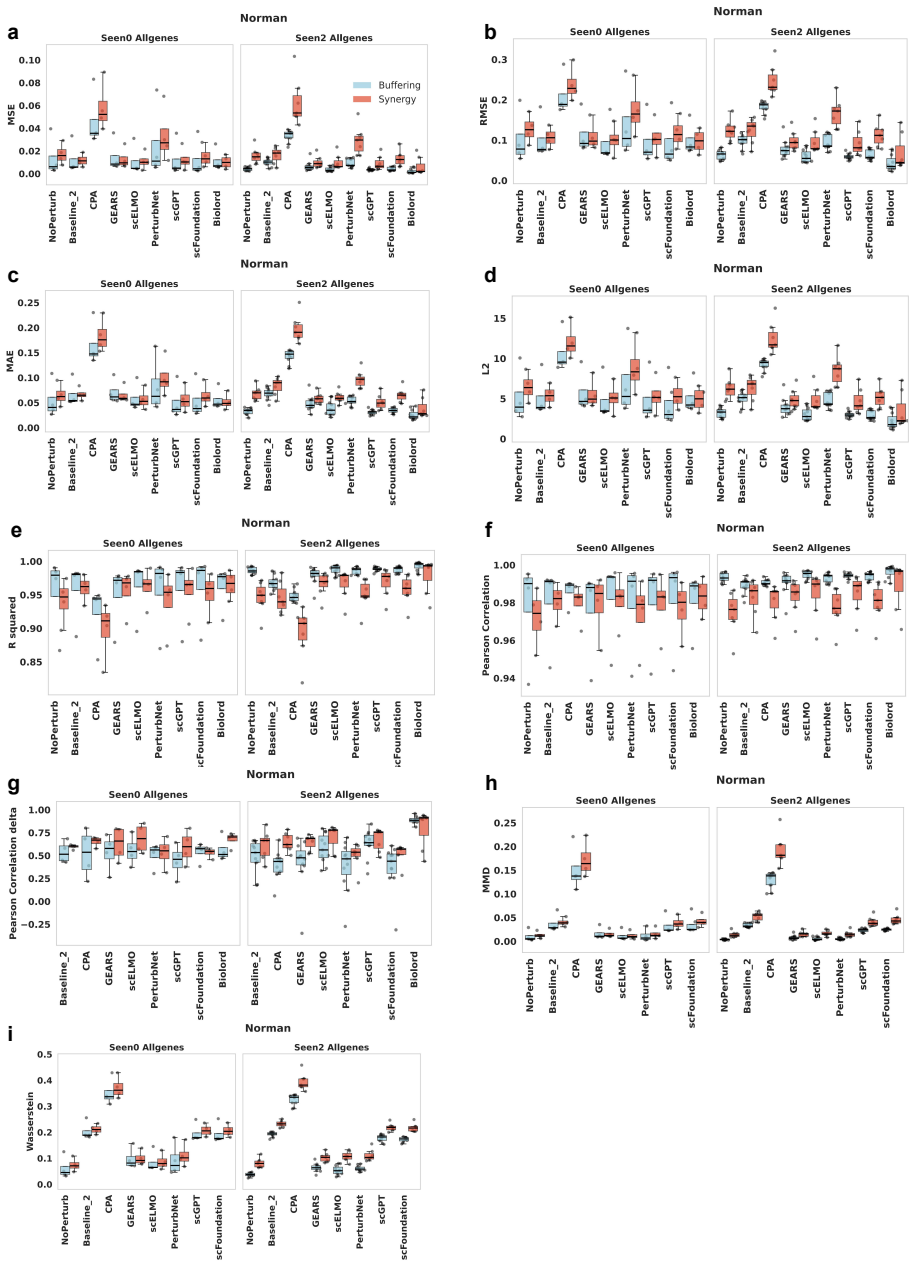


Fig. S14 Evaluation results of Task 2. Performance comparison between buffering and synergistic combinations for seen0 and seen2 conditions across different models on all genes of the Norman dataset. Metrics compared are (a) MSE, (b) RMSE, (c) MAE, (d) L2 distance, (e) R^2, (f) Pearson correlation, (g) Pearson correlation delta, (h) MMD, and (i) Wasserstein distance.

Supplementary Fig. 15

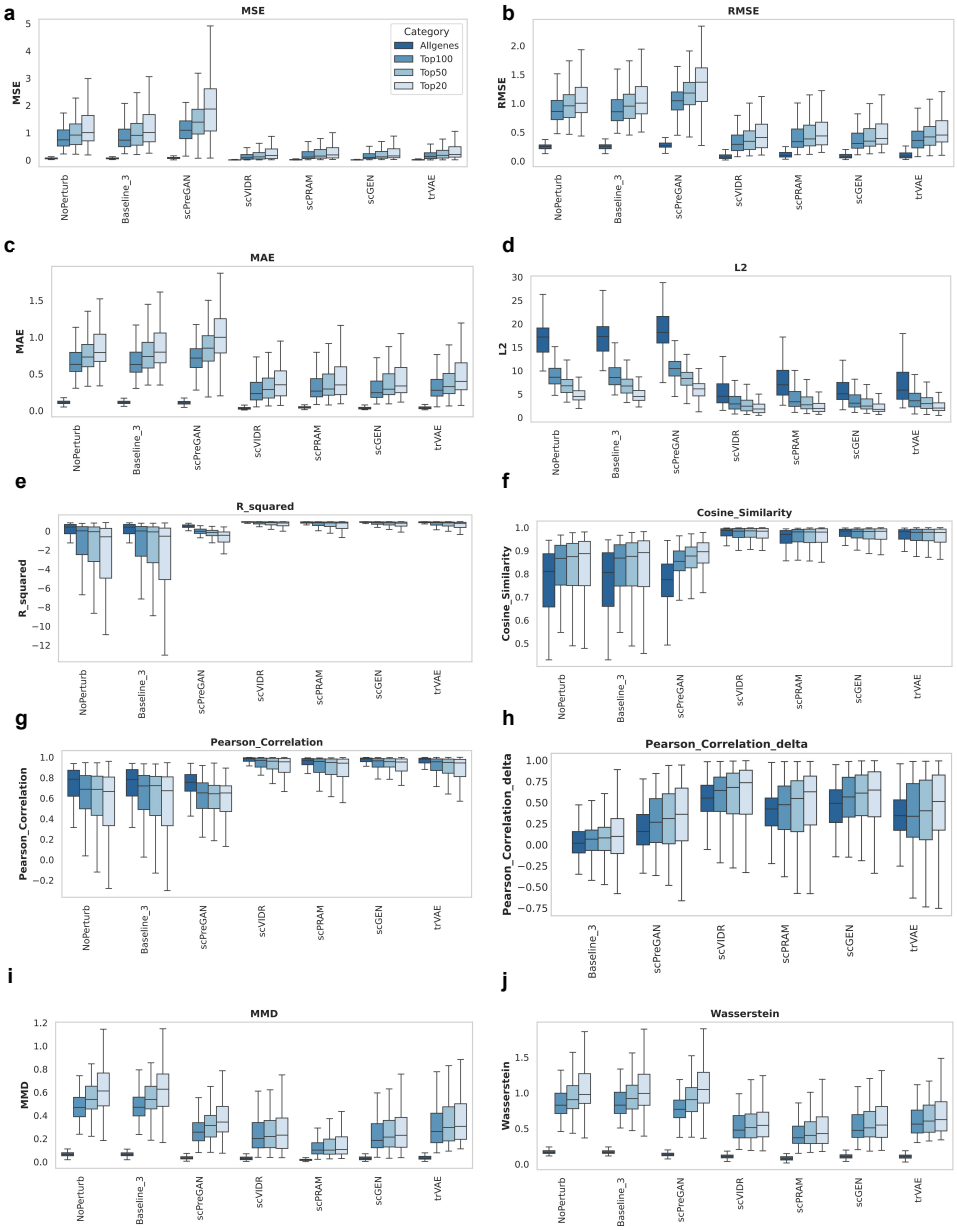


Fig. S15 Evaluation results of Task 3. Metrics aggregated on all datasets were computed across all models using four gene sets: all genes, top 100 DEGs, top 50 DEGs, and top 20 DEGs (ranked by absolute scores). Ten evaluation metrics are reported: (a) MSE, (b) RMSE, (c) MAE, (d) L2 distance, (e) R^2, (f) Cosine similarity, (g) Pearson correlation, (h) Pearson correlation delta, (i) MMD, and (j) Wasserstein distance.

Supplementary Fig. 16

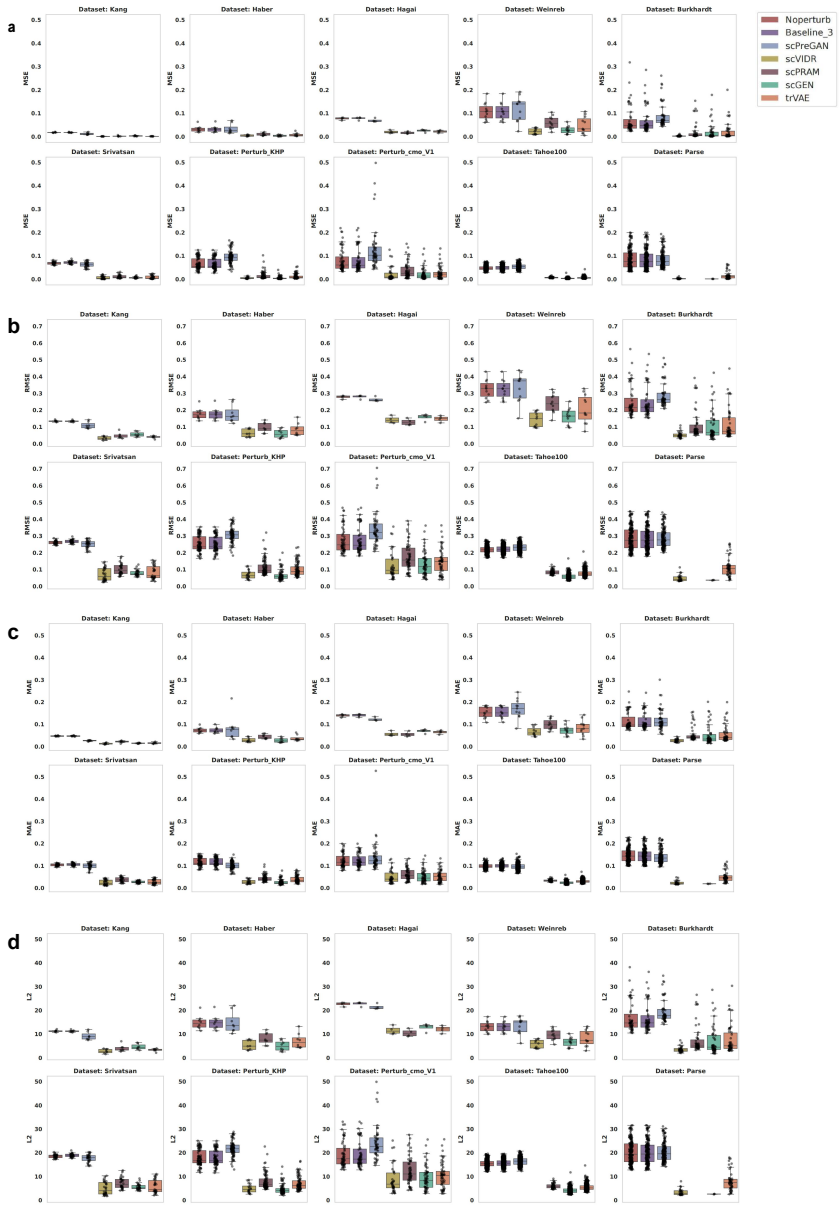


Fig. S16 Evaluation results of Task 3. Metrics were computed across all datasets and models using all genes. The evaluated datasets include Kang, Haber, Hagai, Weinreb, Burkhardt, Srivatsan, Perturb_KHP, Perturb_cmo_V1, Tahoe100, and Parse. Four evaluation metrics are reported on this figure: (a) MSE, (b) RMSE, (c) MAE, and (d) L2 distance.

Supplementary Fig. 17

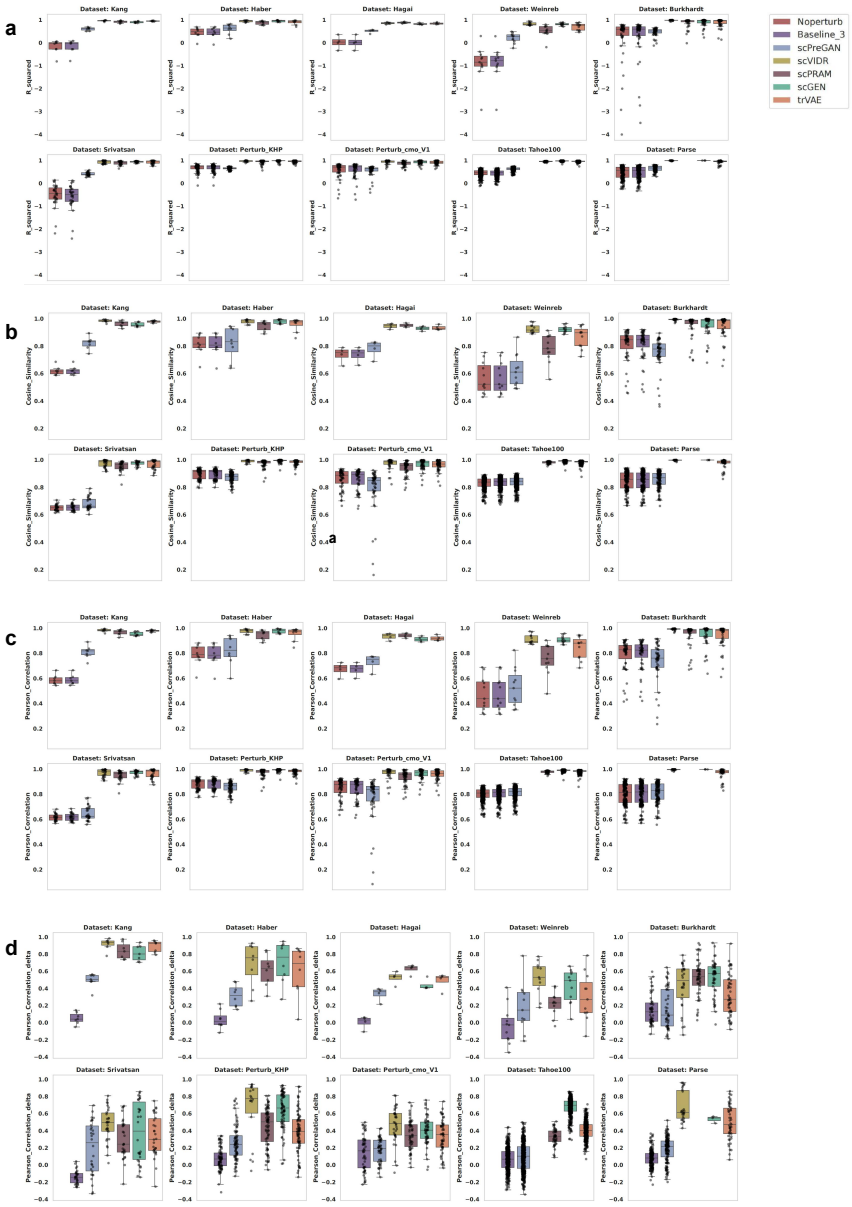


Fig. S17 Evaluation results of Task 3. Metrics were computed across all datasets and models using all genes. The evaluated datasets include Kang, Haber, Hagai, Weinreb, Burkhardt, Srivatsan, Perturb_KHP, Perturb_cmo_V1, Tahoe100, and Parse. Four evaluation metrics are reported on this figure: (a) R^2, (b) Cosine similarity, (c) Pearson correlation, and (d) Pearson correlation delta.

Supplementary Fig. 18

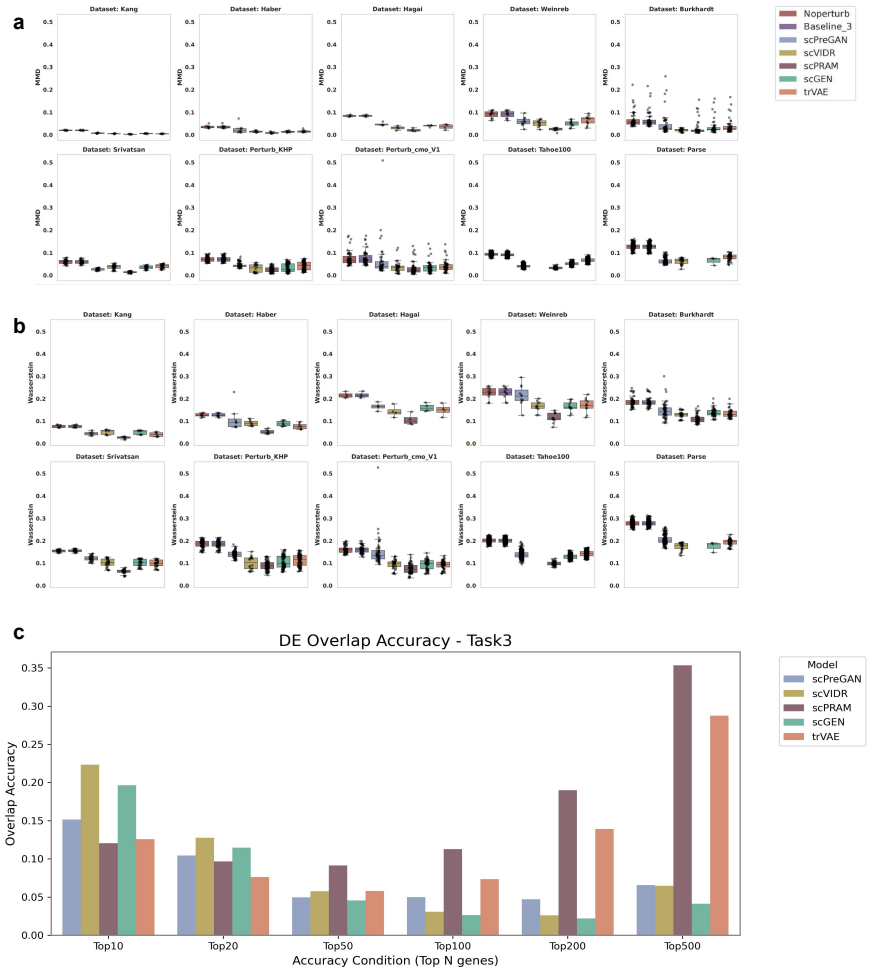


Fig. S18 Evaluation results of Task 3. (a) MMD and (b) Wasserstein distance across all datasets and models using all genes. Datasets include Kang, Haber, Hagai, Weinreb, Burkhardt, Srivatsan, Perturb_KHP, Perturb_cmo_V1, Tahoe100, and Parse. (c) Differentially expressed (DE) gene overlap accuracy between true and predicted perturbations versus control, aggregated on all datasets across all models.

Supplementary Fig. 19

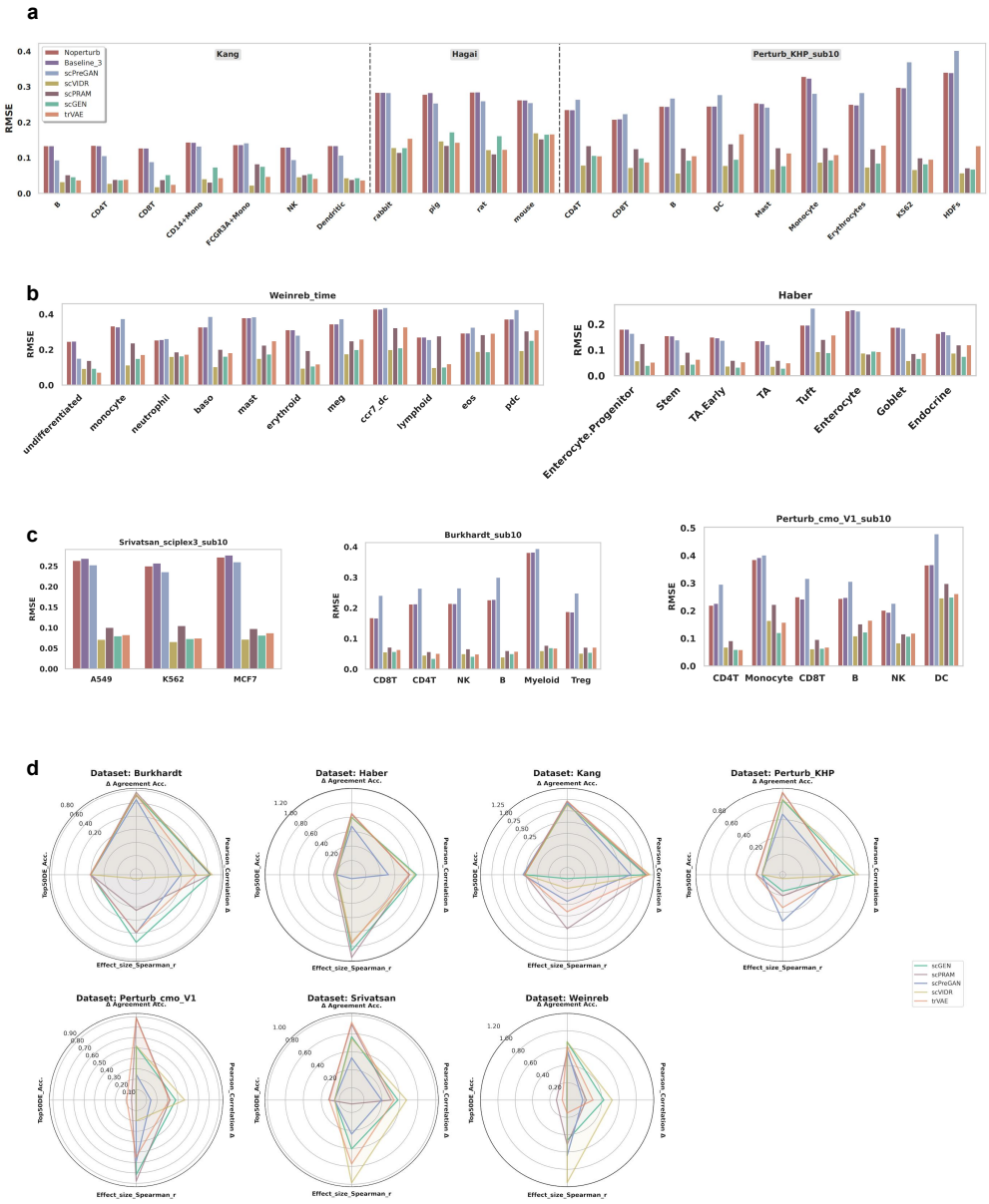


Fig. S19 Evaluation results of Task 3. (a-c) RMSE evaluation across all models and cell types of the eight datasets. (d) Top 50 DEG overlap accuracy, effect size correlation (Spearman r), Pearson's correlation coefficients on delta expressions, and delta direction accuracy across all tested models and datasets for cell state transfer prediction.

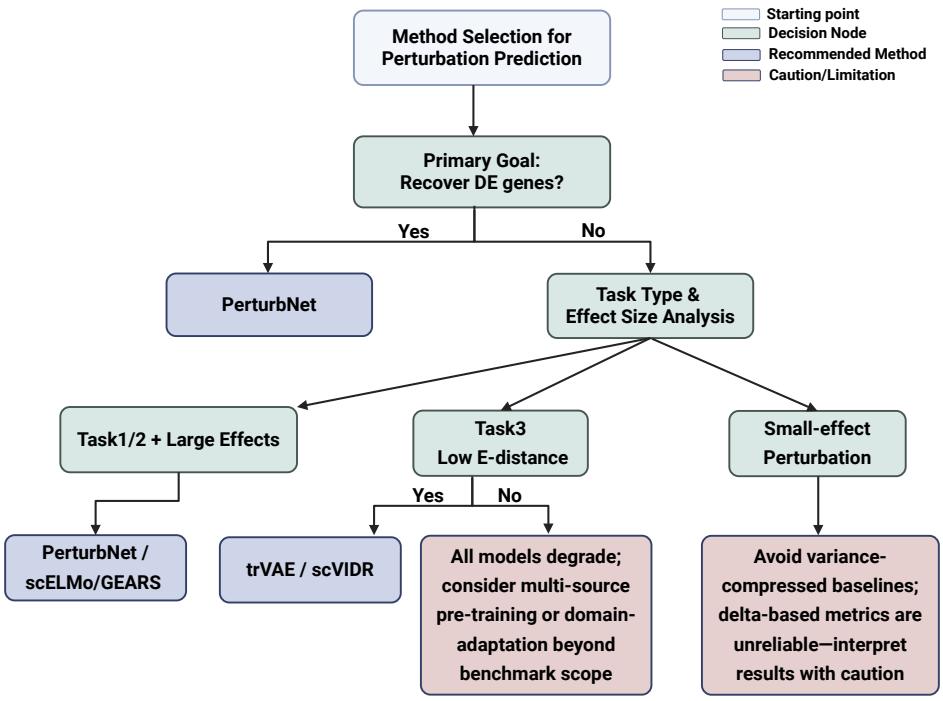


Fig. S20 Decision framework for perturbation prediction method selection.

Competing interests

Although not directly related to this paper, X.L. is a cofounder of iCamuno Biotherapeutics. The remaining authors declare no competing interests.

References

[1] Rood, J.E., Wynne, S., Robson, L., Hupalowska, A., Randell, J., Teichmann, S.A., Regev, A.: The Human Cell Atlas from a cell census to a unified foundation model. Nature (2024) https://doi.org/10.1038/s41586-024-08338-4

[2] Xu, X., Hua, X., Mo, H., Hu, S., Song, J.: Single-cell RNA sequencing to identify cellular heterogeneity and targets in cardiovascular diseases: from bench to bedside. Basic Research in Cardiology **118**(1), 7 (2023) https://doi.org/10.1007/s00395-022-00972-1

[3] Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., *et al.*: Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. cell **167**(7), 1853–1866 (2016)

[4] Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., Bock, C.: Pooled crispr screening with single-cell transcriptome readout. Nature methods **14**(3), 297–301 (2017)

[5] Duan, B., Zhou, C., Zhu, C., Yu, Y., Li, G., Zhang, S., Zhang, C., Ye, X., Ma, H., Qu, S., *et al.*: Model-based understanding of single-cell crispr screening. Nature communications **10**(1), 2233 (2019)

[6] Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P.C., Li, M.J., Wang, J., *et al.*: The support of human genetic evidence for approved drug indications. Nature genetics **47**(8), 856–860 (2015)

[7] Fang, X., Lloyd, C.J., Palsson, B.O.: Reconstructing organisms in silico: genome-scale models and their emerging applications. Nature Reviews Microbiology **18**(12), 731–743 (2020)

[8] Peidli, S., Green, T.D., Shen, C., Gross, T., Min, J., Garda, S., Yuan, B., Schumacher, L.J., Taylor-King, J.P., Marks, D.S., *et al.*: scperturb: harmonized single-cell perturbation data. Nature Methods **21**(3), 531–540 (2024)

[9] Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., AlQuraishi, M., Brennan, P., Burkhardt, D.B., Califano, A., Cool, J., Dernburg, A.F., Ewing, K., Fox, E.B., Haury, M., Herr, A.E., Horvitz, E., Hsu, P.D., Jain, V., Johnson, G.R., Kalil, T., Kelley, D.R., Kelley, S.O., Kreshuk, A., Mitchison, T., Otte, S., Shendure, J., Sofroniew, N.J., Theis, F., Theodoris, C.V.,

Upadhyayula, S., Valer, M., Wang, B., Xing, E., Yeung-Levy, S., Zitnik, M., Karaletsos, T., Regev, A., Lundberg, E., Leskovec, J., Quake, S.R.: How to Build the Virtual Cell with Artificial Intelligence: Priorities and Opportunities (2024). https://arxiv.org/abs/2409.11654

[10] Quake, S.R.: The cellular dogma. Cell **187**(23), 6421–6423 (2024) https://doi.org/10.1016/j.cell.2024.10.029 . Publisher: Elsevier. Accessed 2024-12-02

[11] Fiers, M.W., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z., Aerts, S.: Mapping gene regulatory networks from single-cell omics data. Briefings in functional genomics **17**(4), 246–254 (2018)

[12] Lopez, R., Gayoso, A., Yosef, N.: Enhancing scientific discoveries in molecular biology with deep generative models. Molecular systems biology **16**(9), 9198 (2020)

[13] Ji, Y., Lotfollahi, M., Wolf, F.A., Theis, F.J.: Machine learning for perturbational single-cell omics. Cell Systems **12**(6), 522–537 (2021)

[14] Roohani, Y., Huang, K., Leskovec, J.: Predicting transcriptional outcomes of novel multigene perturbations with gears. Nature Biotechnology **42**(6), 927–935 (2024)

[15] Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., Wang, B.: scgpt: toward building a foundation model for single-cell multi-omics using generative ai. Nature Methods, 1–11 (2024)

[16] Rood, J.E., Hupalowska, A., Regev, A.: Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. Cell **187**(17), 4520–4545 (2024)

[17] Gavriilidis, G.I., Vasileiou, V., Orfanou, A., Ishaque, N., Psomopoulos, F.: A mini-review on perturbation modelling across single-cell omic modalities. Computational and Structural Biotechnology Journal (2024)

[18] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019). https://arxiv.org/abs/1810.04805

[19] 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., *et al.*: Transfer learning enables predictions in network biology. Nature **618**(7965), 616–624 (2023)

[20] Shao, X., Yang, H., Zhuang, X., Liao, J., Yang, P., Cheng, J., Lu, X., Chen, H., Fan, X.: scdeepsort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. Nucleic acids research **49**(21), 122–122 (2021)

[21] Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., Song, L.: Large-scale foundation model on single-cell transcriptomics. Nature Methods, 1–11 (2024)

[22] Wenteler, A., Occhetta, M., Branson, N., Huebner, M., Curean, V., Dee, W.T., Connell, W.T., Hawkins-Hooker, A., Chung, S.P., Ektefaie, Y., Gallagher-Syed, A., Córdova, C.M.V.: Perteval-scfm: Benchmarking single-cell foundation models for perturbation effect prediction. bioRxiv (2024)

[23] Wu, Y., Wershof, E., Schmon, S.M., Nassar, M., Osi'nski, B.z., Eksi, R., Zhang, K., Graepel, T.: Perturbench: Benchmarking machine learning models for cellular perturbation analysis. ArXiv **abs/2408.10609** (2024)

[24] Kedzierska, K.Z., Crawford, L., Amini, A.P., Lu, A.X.: Zero-shot evaluation reveals limitations of single-cell foundation models. Genome Biology **26**(1), 101 (2025)

[25] Ahlmann-Eltze, C., Huber, W., Anders, S.: Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Nature Methods **22**(8), 1657–1661 (2025) https://doi.org/10.1038/s41592-025-02772-6

[26] Roohani, Y.H., Hua, T.J., Tung, P.-Y., Bounds, L.R., Yu, F.B., Dobin, A., Teyssier, N., Adduri, A., Woodrow, A., Plosky, B.S., *et al.*: Virtual cell challenge: Toward a turing test for the virtual cell. Cell **188**(13), 3370–3374 (2025)

[27] Adduri, A.K., Gautam, D., Bevilacqua, B., Imran, A., Shah, R., Naghipourfar, M., Teyssier, N., Ilango, R., Nagaraj, S., Dong, M., et al.: Predicting cellular responses to perturbation across diverse contexts with state. bioRxiv, 2025–06 (2025)

[28] Zhang, J., Ubas, A.A., Borja, R., Svensson, V., Thomas, N., Thakar, N., Lai, I., Winters, A., Khan, U., Jones, M.G., et al.: Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. BioRxiv, 2025–02 (2025)

[29] Mejia, G.M., Miller, H.E., Leblanc, F.J., Wang, B., Swain, B., Camillo, L.P.d.L.: Diversity by design: Addressing mode collapse improves scrna-seq perturbation modeling on well-calibrated metrics. arXiv preprint arXiv:2506.22641 (2025)

[30] Yu, H., Qian, W., Song, Y., Welch, J.D.: Perturbnet predicts single-cell responses to unseen chemical and genetic perturbations. Molecular Systems Biology, 1–23 (2025)

[31] Moon, H.C., Herschl, M.H., Pawluk, A., Konermann, S., Hsu, P.D.: A combinatorial domain screening platform reveals epigenetic effector interactions for transcriptional perturbation. bioRxiv (2024) https://doi.org/10.1101/2024.10.28.620683

https://www.biorxiv.org/content/early/2024/11/05/2024.10.28.620683.full.pdf

[32] Zheng, S., Rao, J., Zhang, J., Zhou, L., Xie, J., Cohen, E., Lu, W., Li, C., Yang, Y.: Cross-modal graph contrastive learning with cellular images. Advanced Science **11**(32), 2404845 (2024)

[33] Saunders, R.A., Allen, W.E., Pan, X., Sandhu, J., Lu, J., Lau, T.K., Smolyar, K., Sullivan, Z.A., Dulac, C., Weissman, J.S., Zhuang, X.: A platform for multimodal in vivo pooled genetic screens reveals regulators of liver function (2024) https://doi.org/10.1101/2024.11.18.624217 . Publisher: Cold Spring Harbor Laboratory Section: New Results. Accessed 2024-12-03

[34] Metzner, E., Southard, K.M., Norman, T.M.: Multiome Perturb-seq unlocks scalable discovery of integrated perturbation effects on the transcriptome and epigenome (2024) https://doi.org/10.1101/2024.07.26.605307 . Publisher: Cold Spring Harbor Laboratory Section: New Results. Accessed 2024-08-29

[35] Datlinger, P., Rendeiro, A.F., Boenke, T., Senekowitsch, M., Krausgruber, T., Barreca, D., Bock, C.: Ultra-high-throughput single-cell rna sequencing and perturbation screening with combinatorial fluidic indexing. Nature methods **18**(6), 635–642 (2021)

[36] Frangieh, C.J., Melms, J.C., Thakore, P.I., Geiger-Schuller, K.R., Ho, P., Luoma, A.M., Cleary, B., Jerby-Arnon, L., Malu, S., Cuoco, M.S., *et al.*: Multimodal pooled perturb-cite-seq screens in patient models define mechanisms of cancer immune evasion. Nature genetics **53**(3), 332–341 (2021)

[37] Papalexi, E., Mimitou, E.P., Butler, A.W., Foster, S., Bracken, B., Mauck III, W.M., Wessels, H.-H., Hao, Y., Yeung, B.Z., Smibert, P., *et al.*: Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. Nature genetics **53**(3), 322–331 (2021)

[38] Replogle, J.M., Saunders, R.A., Pogson, A.N., Hussmann, J.A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E.J., Adelman, K., Lithwick-Yanai, G., *et al.*: Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. Cell **185**(14), 2559–2575 (2022)

[39] Tian, R., Abarientos, A., Hong, J., Hashemi, S.H., Yan, R., Dräger, N., Leng, K., Nalls, M.A., Singleton, A.B., Xu, K., *et al.*: Genome-wide crispri/a screens in human neurons link lysosomal failure to ferroptosis. Nature neuroscience **24**(7), 1020–1034 (2021)

[40] Srivatsan, S.R., McFaline-Figueroa, J.L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H.A., Jackson, D.L., Daza, R.M., Christiansen, L., *et al.*: Massively multiplex chemical transcriptomics at single-cell resolution. Science **367**(6473), 45–51 (2020)

[41] Sunshine, S., Puschnik, A.S., Replogle, J.M., Laurie, M.T., Liu, J., Zha, B.S., Nuñez, J.K., Byrum, J.R., McMorrow, A.H., Frieman, M.B., *et al.*: Systematic functional interrogation of sars-cov-2 host factors using perturb-seq. Nature Communications **14**(1), 6245 (2023)

[42] Norman, T.M., Horlbeck, M.A., Replogle, J.M., Ge, A.Y., Xu, A., Jost, M., Gilbert, L.A., Weissman, J.S.: Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science **365**(6455), 786–793 (2019)

[43] 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., *et al.*: A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. Cell **167**(7), 1867–1882 (2016)

[44] Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., *et al.*: Multiplexed droplet single-cell rna-sequencing using natural genetic variation. Nature biotechnology **36**(1), 89–94 (2018)

[45] Haber, A.L., Biton, M., Rogel, N., Herbst, R.H., Shekhar, K., Smillie, C., Burgin, G., Delorey, T.M., Howitt, M.R., Katz, Y., *et al.*: A single-cell survey of the small intestinal epithelium. Nature **551**(7680), 333–339 (2017)

[46] Hagai, T., Chen, X., Miragaia, R.J., Rostom, R., Gomes, T., Kunowska, N., Henriksson, J., Park, J.-E., Proserpio, V., Donati, G., *et al.*: Gene expression variability across cells and species shapes innate immunity. Nature **563**(7730), 197–202 (2018)

[47] Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., Klein, A.M.: Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science **367**(6479), 3381 (2020)

[48] Bunne, C., Stark, S.G., Gut, G., Castillo, J.S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A., Rätsch, G.: Learning single-cell perturbation responses using neural optimal transport. Nature Methods, 1–10 (2023) https://doi.org/10.1038/s41592-023-01969-x . Publisher: Nature Publishing Group. Accessed 2023-11-05

[49] Liu, J.-L., Yang, L.-G., Xiao, Q.-Y., Li, Z.-Q., Cheng, J.-Y., Chen, Z.-R., Zhou, J.-W., Wan, X., Chang, T.-H., Zhang, X., Li, Y.-X.: Bio-medical Big Data Operating System (Bio-OS): An Integrated Data Mining Environment for Data Intensive Scientific Research (2024) https://doi.org/10.1101/2024.10.17.612997 . Publisher: Cold Spring Harbor Laboratory Section: New Results. Accessed 2024-12-03

[50] scgen predicts single-cell perturbation responses. Nature methods **16**(8), 715–721 (2019)

[51] Lotfollahi, M., Naghipourfar, M., Theis, F.J., Wolf, F.A.: Conditional out-of-distribution generation for unpaired data using transfer vae. Bioinformatics **36**(Supplement_2), 610–617 (2020)

[52] Kana, O., Nault, R., Filipovic, D., Marri, D., Zacharewski, T., Bhattacharya, S.: Generative modeling of single-cell gene expression for dose-dependent chemical perturbations. Patterns **4**(8) (2023)

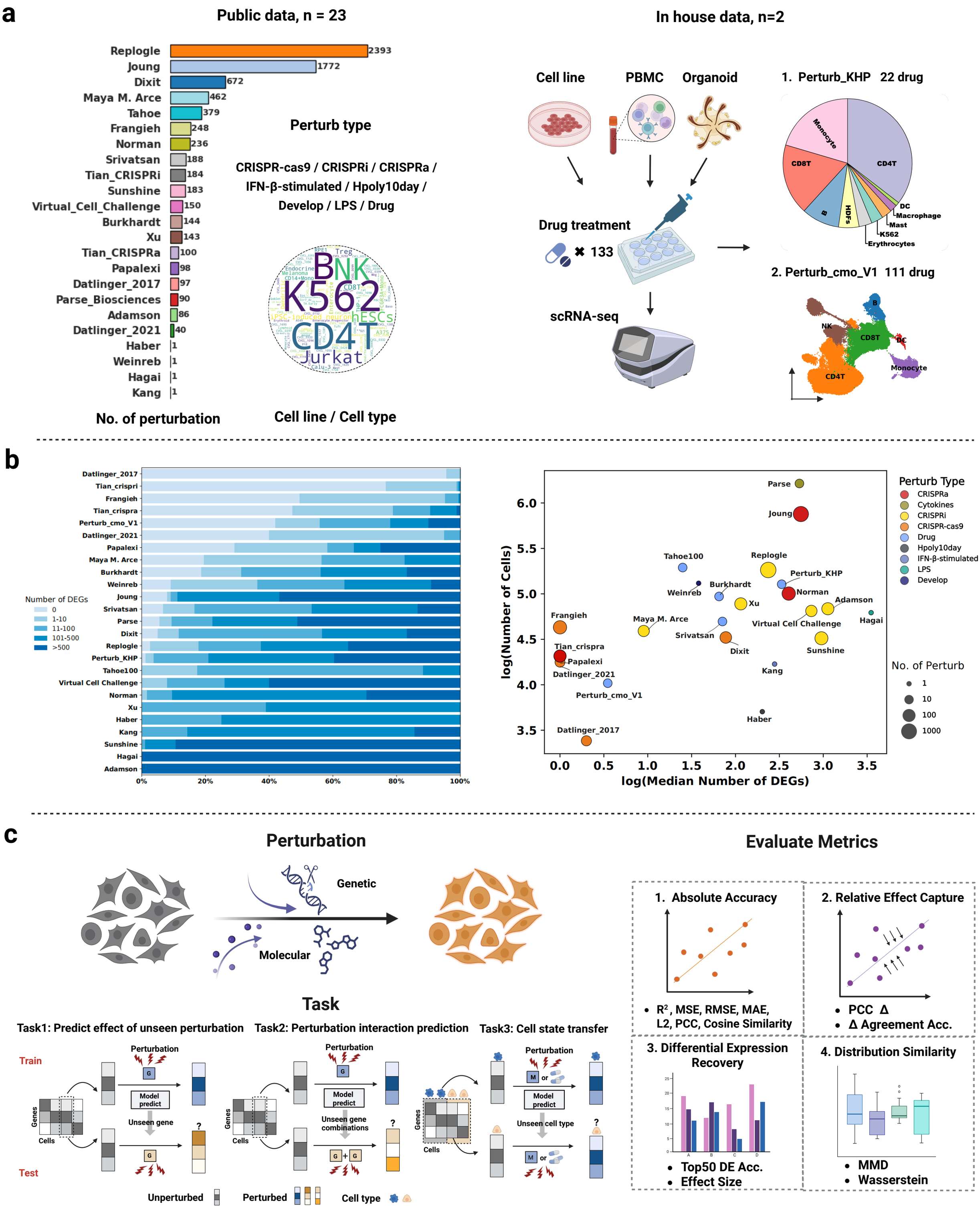
[53] Jiang, Q., Chen, S., Chen, X., Jiang, R.: scpram accurately predicts single-cell gene expression perturbation response based on attention mechanism. Bioinformatics **40**(5), 265 (2024)

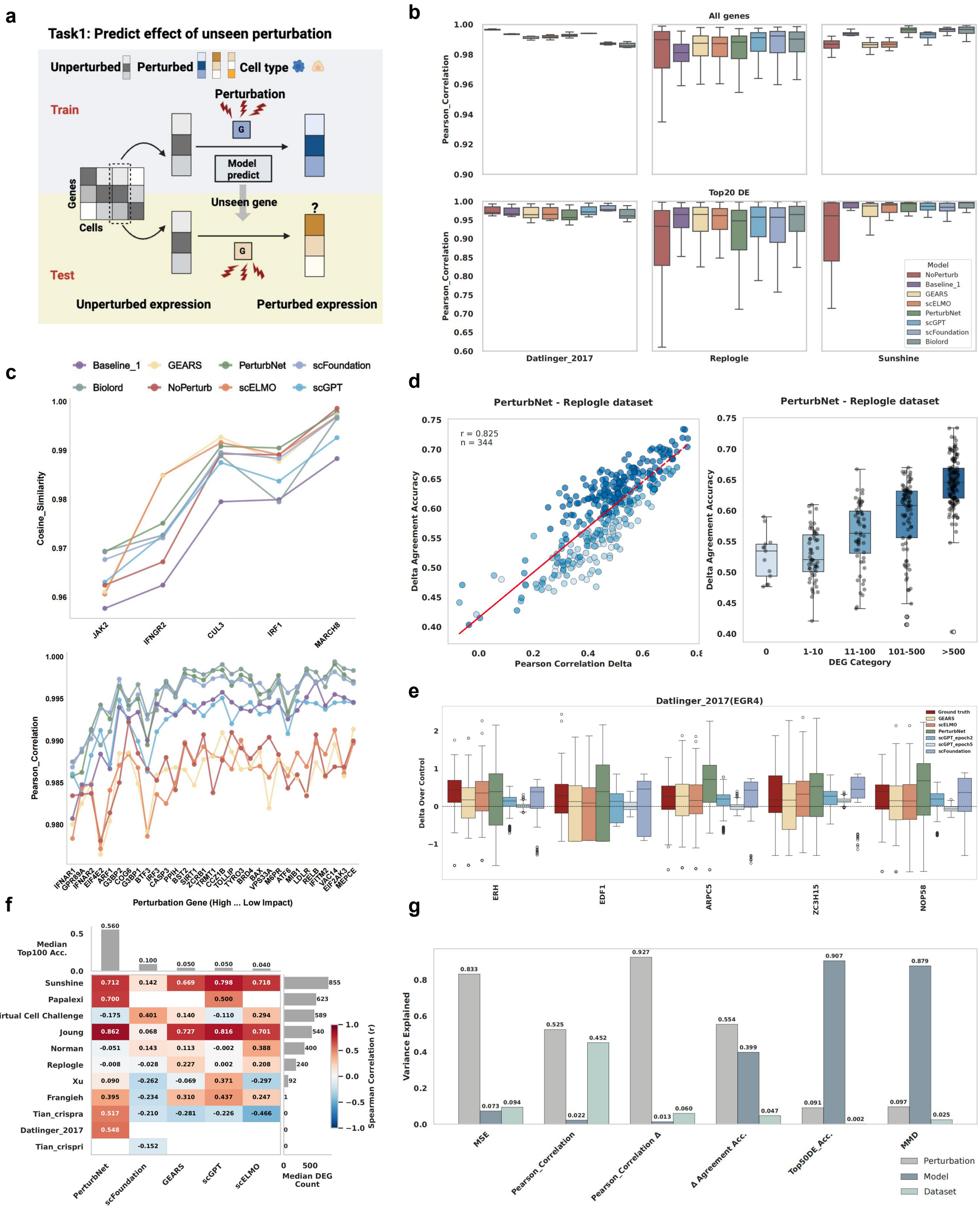
[54] Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., Hetzel, L., Ji, Y., Ibarra, I.L., Srivatsan, S.R., Naghipourfar, M., Daza, R.M., Martin, B., *et al.*: Predicting cellular responses to complex perturbations in high-throughput screens. Molecular systems biology **19**(6), 11517 (2023)

[55] Wei, X., Dong, J., Wang, F.: scpregan, a deep generative model for predicting the response of single-cell expression to perturbation. Bioinformatics **38**(13), 3377–3384 (2022)

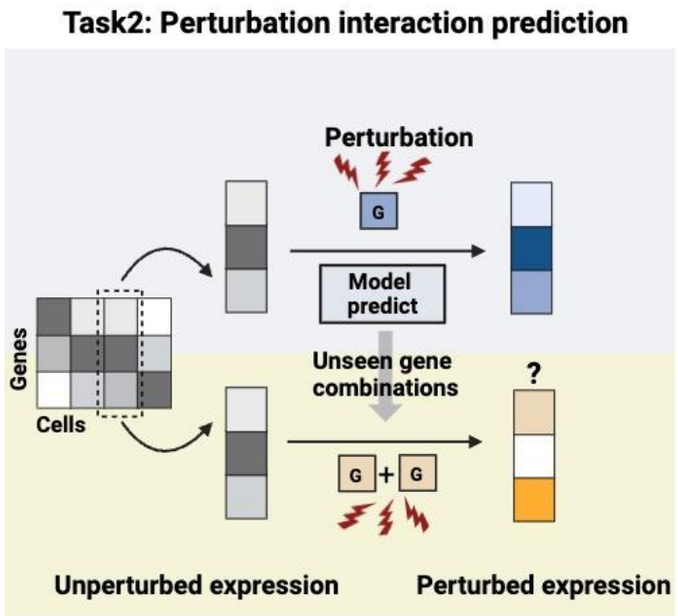
[56] Piran, Z., Cohen, N., Hoshen, Y., Nitzan, M.: Disentanglement of single-cell data with biolord. Nature Biotechnology **42**(11), 1678–1683 (2024)

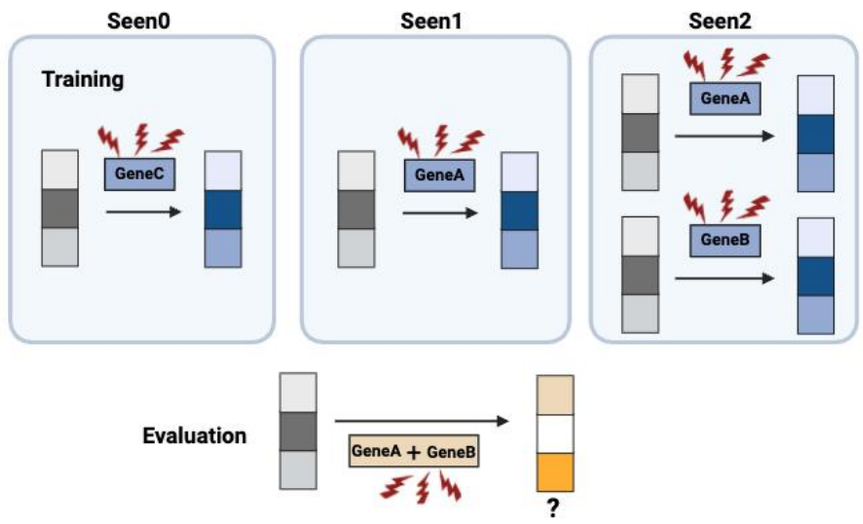
[57] Nadig, A., Replogle, J.M., Pogson, A.N., Murthy, M., McCarroll, S.A., Weissman, J.S., Robinson, E.B., O'Connor, L.J.: Transcriptome-wide analysis of differential expression in perturbation atlases. Nature Genetics, 1–10 (2025)


a
Public data, n = 23
Replogle 2393
Joung 1772
Dixit 672
Maya M. Arce 462
Tahoe 379
Frangieh 248
Norman 236
Srivatsan 188
Tian_CRISPRi 184
Sunshine 183
Virtual_Cell_Challenge 150
Burkhardt 144
Xu 143
Tian_CRISPRa 100
Papalexi 98
Datlinger_2017 97
Parse_Biosciences 90
Adamson 86
Datlinger_2021 40
Haber 1
Weinreb 1
Hagai 1
Kang 1
No. of perturbation
Perturb type
CRISPR-cas9 / CRISPRi / CRISPRa /
IFN-β-stimulated / Hpoly10day /
Develop / LPS / Drug
BNK
K562
CD4T
Jurkat
hESCs
Cell line / Cell type
In house data, n=2
Cell line
PBMC
Organoid
Drug treatment
✖ 133
scRNA-seq
1. Perturb_KHP 22 drug
CD4T
CD8T
Monocyte
B
HDFs
DC
Macrophage
Mast
K562
Erythrocytes
2. Perturb_cmo_V1 111 drug
B
NK
CD8T
DC
Monocyte
CD4T
b
Number of DEGs
0
1-10
11-100
101-500
>500
Datlinger_2017
Tian_crispri
Frangieh
Tian_crispra
Perturb_cmo_V1
Datlinger_2021
Papalexi
Maya M. Arce
Burkhardt
Weinreb
Joung
Srivatsan
Parse
Dixit
Replogle
Perturb_KHP
Tahoe100
Virtual Cell Challenge
Norman
Xu
Haber
Kang
Sunshine
Hagai
Adamson
0%
20%
40%
60%
80%
100%
log(Number of Cells)
log(Median Number of DEGs)
Perturb Type
CRISPRa
Cytokines
CRISPRi
CRISPR-cas9
Drug
Hpoly10day
IFN-β-stimulated
LPS
Develop
No. of Perturb
1
10
100
1000
c
Perturbation
Genetic
Molecular
Task
Task1: Predict effect of unseen perturbation
Task2: Perturbation interaction prediction
Task3: Cell state transfer
Train
Test
Perturbation
Model predict
Unseen gene
Unseen gene combinations
Unseen cell type
Genes
Cells
Unperturbed
Perturbed
Cell type
Evaluate Metrics
1. Absolute Accuracy
• R², MSE, RMSE, MAE, L2, PCC, Cosine Similarity
2. Relative Effect Capture
• PCC Δ
• Δ Agreement Acc.
3. Differential Expression Recovery
• Top50 DE Acc.
• Effect Size
4. Distribution Similarity
• MMD
• Wasserstein

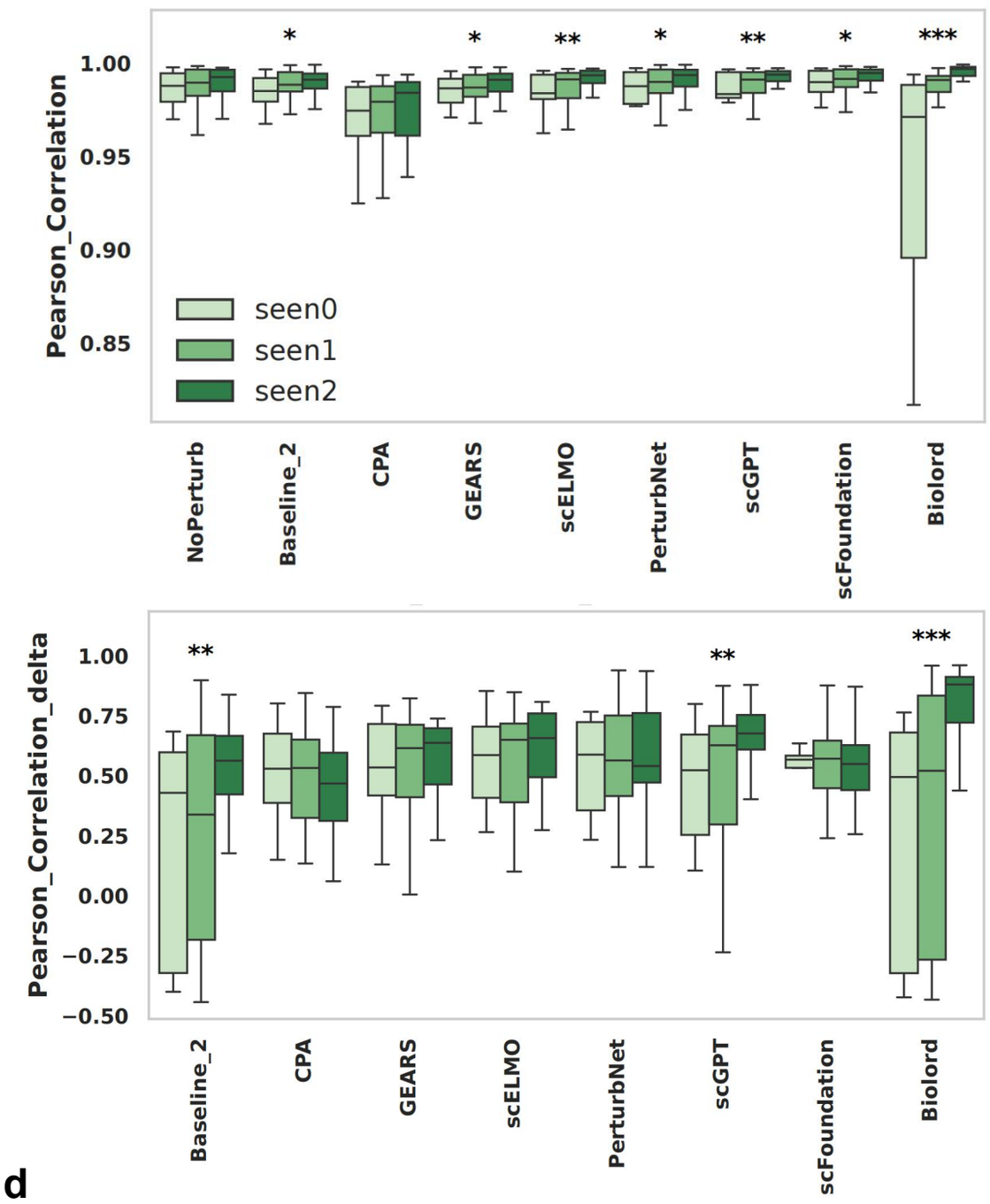

a
Task1: Predict effect of unseen perturbation
Unperturbed
Perturbed
Cell type
Train
Perturbation
G
Model predict
Genes
Cells
Unseen gene
?
Test
Unperturbed expression
Perturbed expression
b
All genes
Top20 DE
Pearson_Correlation
Datlinger_2017
Replogle
Sunshine
Model
NoPerturb
Baseline_1
GEARS
scELMO
PerturbNet
scGPT
scFoundation
Biolord
c
Baseline_1
GEARS
PerturbNet
scFoundation
Biolord
NoPerturb
scELMO
scGPT
Cosine_Similarity
JAK2
IFNGR2
CUL3
IRF1
MARCH8
Pearson_Correlation
Perturbation Gene (High ... Low Impact)
d
PerturbNet - Replogle dataset
r = 0.825
n = 344
Delta Agreement Accuracy
Pearson Correlation Delta
DEG Category
0
1-10
11-100
101-500
>500
e
Datlinger_2017(EGR4)
Ground truth
GEARS
scELMO
PerturbNet
scGPT_epoch2
scGPT_epoch5
scFoundation
Delta Over Control
ERH
EDF1
ARPC5
ZC3H15
NOP58
f
Median Top100 Acc.
Sunshine
Papalexi
Virtual Cell Challenge
Joung
Norman
Replogle
Xu
Frangieh
Tian_crispra
Datlinger_2017
Tian_crispri
PerturbNet
scFoundation
GEARS
scGPT
scELMO
Median DEG Count
Spearman Correlation (r)
g
Variance Explained
MSE
Pearson_Correlation
Pearson_Correlation Δ
Δ Agreement Acc.
Top50DE_Acc.
MMD
Perturbation
Model
Dataset

a

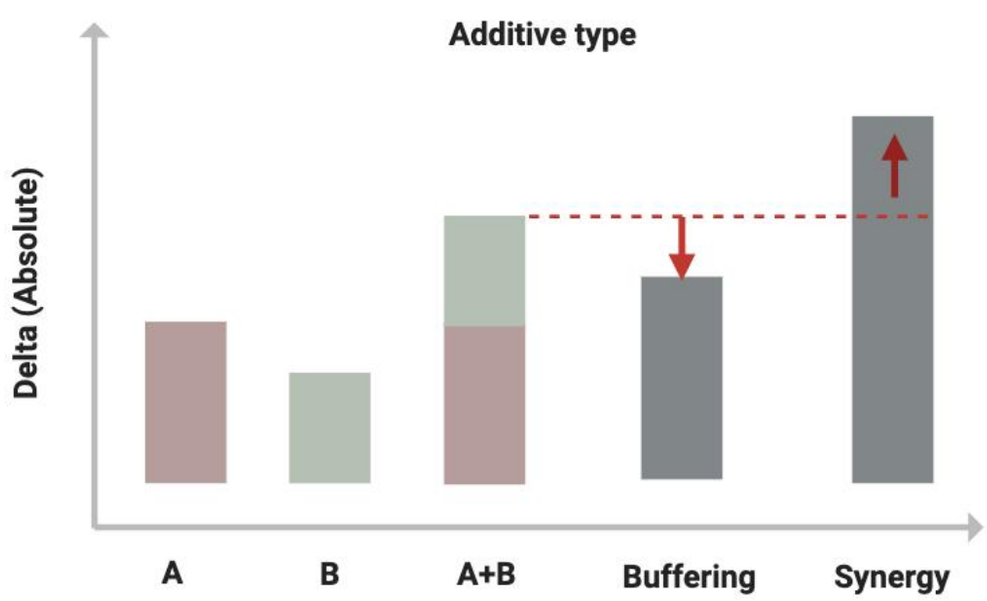

Task2: Perturbation interaction prediction
Perturbation
G
Model predict
Genes
Cells
Unseen gene combinations
?
G + G
Unperturbed expression
Perturbed expression


Seen0
Seen1
Seen2
Training
GeneC
GeneA
GeneA
GeneB
Evaluation
GeneA + GeneB
?

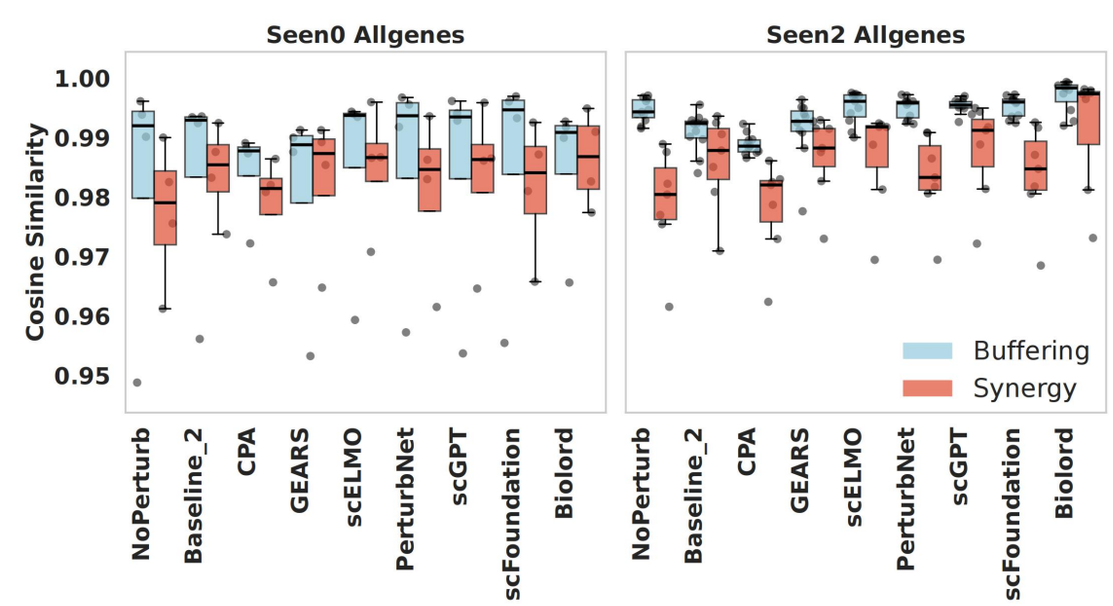
b


Pearson_Correlation
1.00
0.95
0.90
0.85
seen0
seen1
seen2
NoPerturb
Baseline_2
CPA
GEARS
scELMO
PerturbNet
scGPT
scFoundation
Biolord
Pearson_Correlation_delta
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
Norman

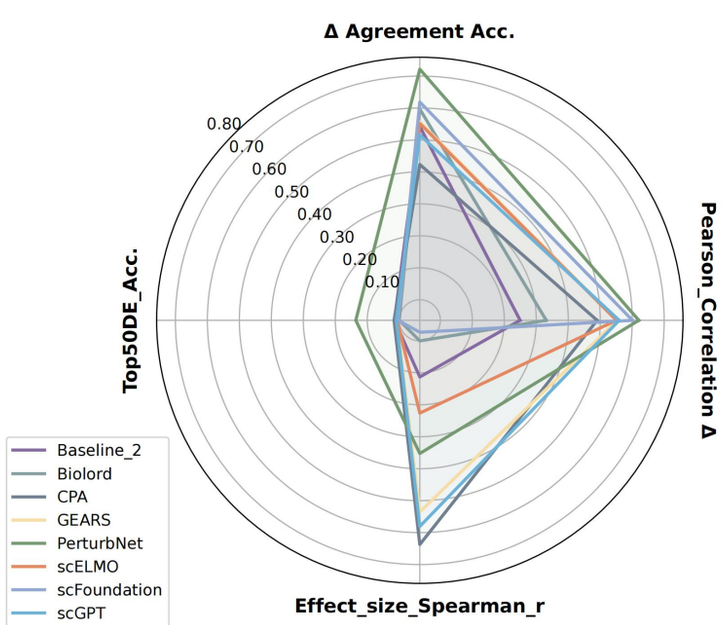
c


Additive type
Delta (Absolute)
A
B
A+B
Buffering
Synergy

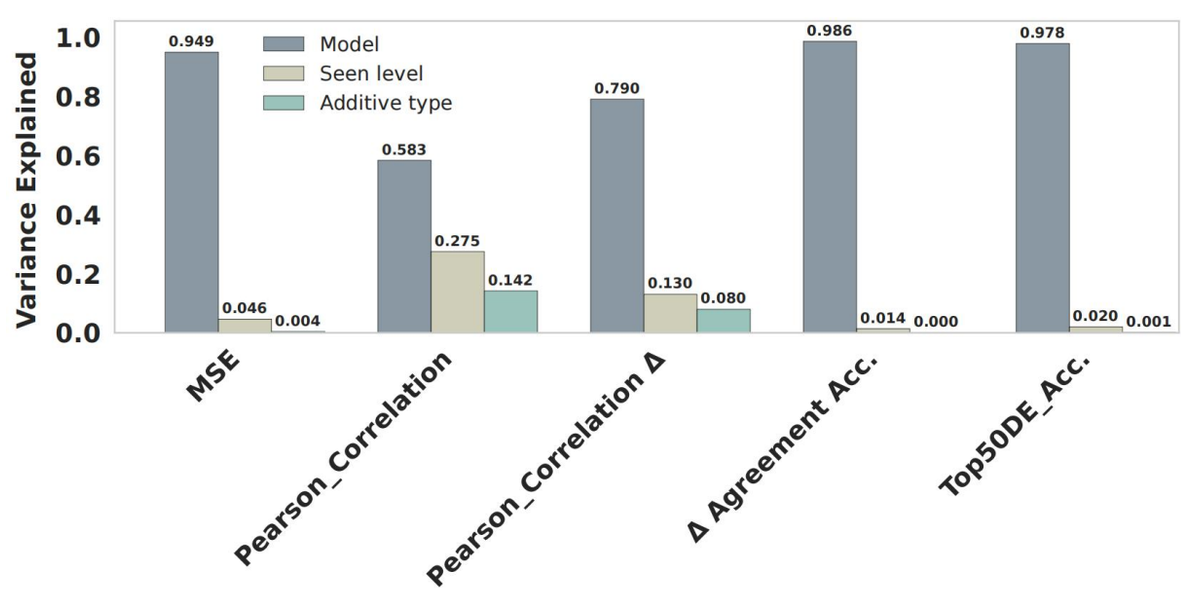
d

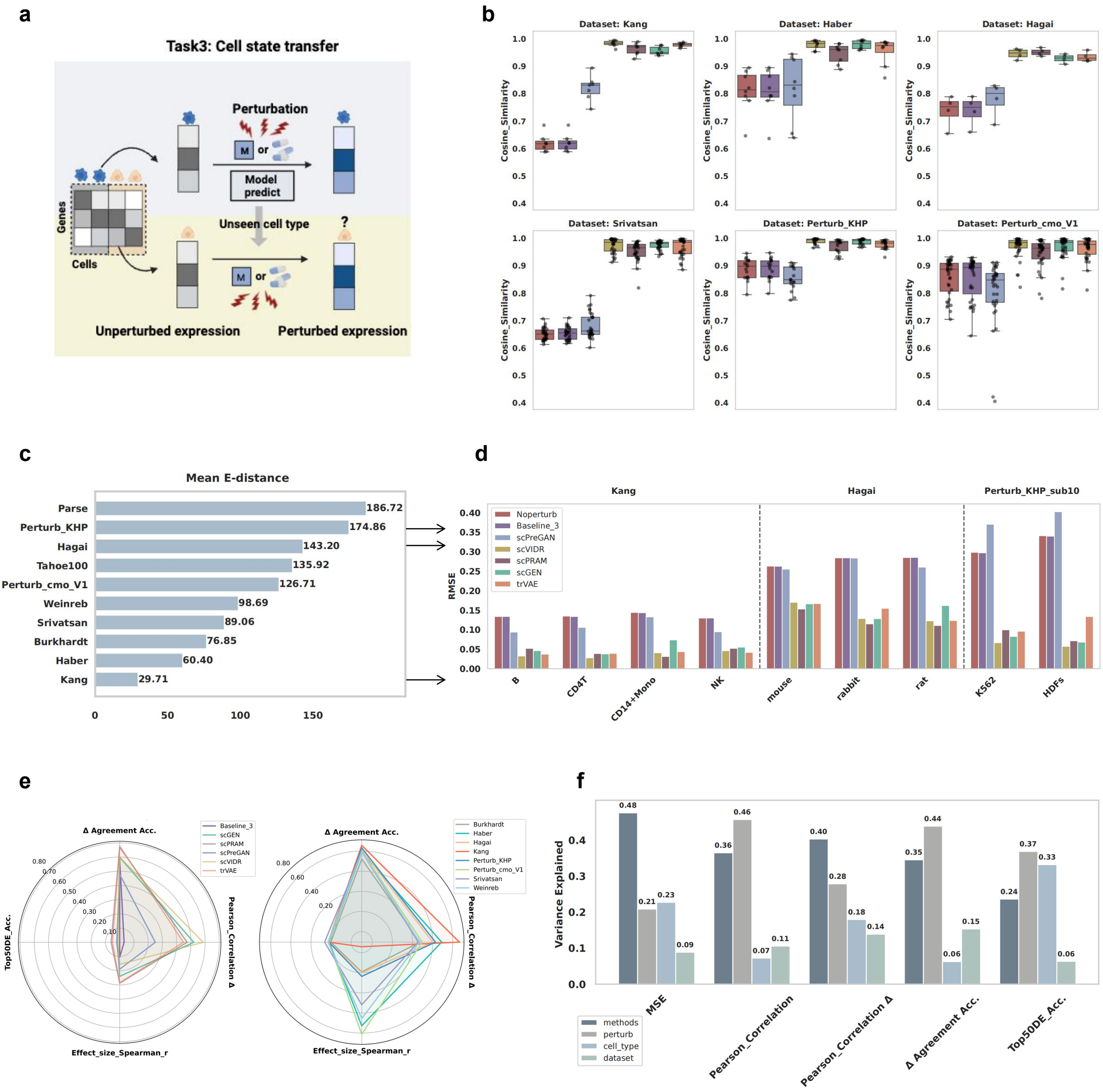

Seen0 Allgenes
Seen2 Allgenes
Cosine Similarity
1.00
0.99
0.98
0.97
0.96
0.95
Buffering
Synergy
NoPerturb
Baseline_2
CPA
GEARS
scELMO
PerturbNet
scGPT
scFoundation
Biolord

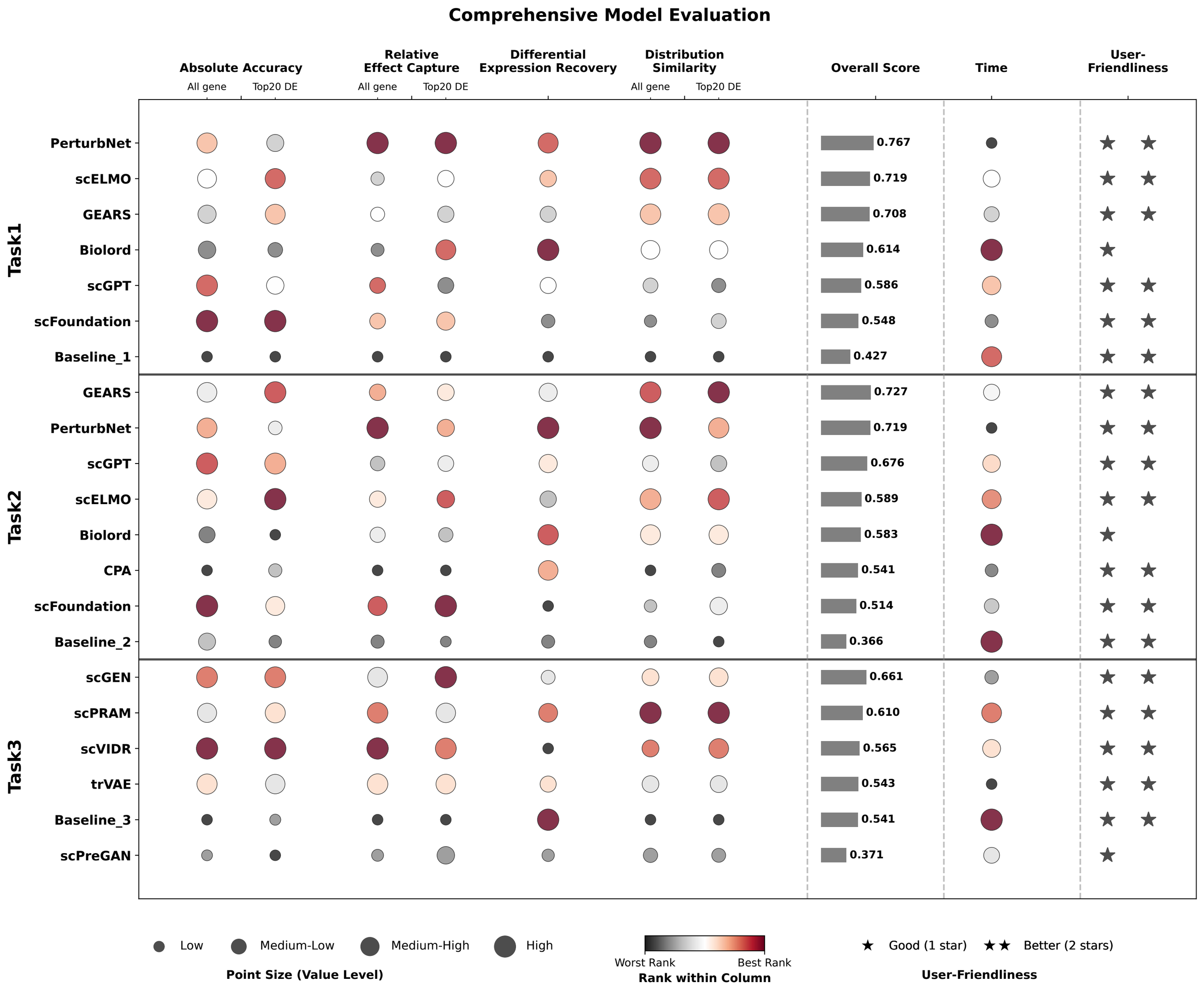
e


Δ Agreement Acc.
Pearson_Correlation Δ
Effect_size_Spearman_r
Top50DE_Acc.
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
Baseline_2
Biolord
CPA
GEARS
PerturbNet
scELMO
scFoundation
scGPT

f


Variance Explained
1.0
0.8
0.6
0.4
0.2
0.0
Model
Seen level
Additive type
0.949
0.046
0.004
0.583
0.275
0.142
0.790
0.130
0.080
0.986
0.014
0.000
0.978
0.020
0.001
MSE
Pearson_Correlation
Pearson_Correlation Δ
Δ Agreement Acc.
Top50DE_Acc.


a
Task3: Cell state transfer
Perturbation
M or
Model predict
Genes
Cells
Unseen cell type
?
M or
Unperturbed expression
Perturbed expression
b
Dataset: Kang
Dataset: Haber
Dataset: Hagai
Dataset: Srivatsan
Dataset: Perturb_KHP
Dataset: Perturb_cmo_V1
Cosine_Similarity
c
Mean E-distance
Parse 186.72
Perturb_KHP 174.86
Hagai 143.20
Tahoe100 135.92
Perturb_cmo_V1 126.71
Weinreb 98.69
Srivatsan 89.06
Burkhardt 76.85
Haber 60.40
Kang 29.71
d
Kang
Hagai
Perturb_KHP_sub10
RMSE
Noperturb
Baseline_3
scPreGAN
scVIDR
scPRAM
scGEN
trVAE
B
CD4T
CD14+Mono
NK
mouse
rabbit
rat
K562
HDFs
e
Δ Agreement Acc.
Baseline_3
scGEN
scPRAM
scPreGAN
scVIDR
trVAE
Burkhardt
Haber
Hagai
Kang
Perturb_KHP
Perturb_cmo_V1
Srivatsan
Weinreb
Top50DE_Acc.
Pearson_Correlation Δ
Effect_size_Spearman_r
f
Variance Explained
methods
perturb
cell_type
dataset
MSE
Pearson_Correlation
Pearson_Correlation Δ
Δ Agreement Acc.
Top50DE_Acc.


Comprehensive Model Evaluation
Absolute Accuracy
All gene
Top20 DE
Relative Effect Capture
All gene
Top20 DE
Differential Expression Recovery
Distribution Similarity
All gene
Top20 DE
Overall Score
Time
User-Friendliness
Task1
PerturbNet 0.767
scELMO 0.719
GEARS 0.708
Biolord 0.614
scGPT 0.586
scFoundation 0.548
Baseline_1 0.427
Task2
GEARS 0.727
PerturbNet 0.719
scGPT 0.676
scELMO 0.589
Biolord 0.583
CPA 0.541
scFoundation 0.514
Baseline_2 0.366
Task3
scGEN 0.661
scPRAM 0.610
scVIDR 0.565
trVAE 0.543
Baseline_3 0.541
scPreGAN 0.371
Low
Medium-Low
Medium-High
High
Point Size (Value Level)
Worst Rank
Best Rank
Rank within Column
Good (1 star)
Better (2 stars)
User-Friendliness