Research Square

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

An Empirical Comparison of Virtual Cell Models: Perturbation Prediction, Representation, and the Baseline Gap

Olivia Denvis

oliviadenvis@gmail.com

Lampung University https://orcid.org/0009-0002-7443-2669

Research Article

Keywords:

Posted Date: July 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10434123/v1

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

Additional Declarations: The authors declare no competing interests.

An Empirical Comparison of Virtual Cell Models: Perturbation Prediction, Representation, and the Baseline Gap

Olivia Denvis
Independent Researcher
oliviadenvis@gmail.com

Abstract

The "AI virtual cell" has become a rallying vision for computational biology: a learned model that represents cell state and simulates how cells respond to genetic and chemical perturbations. A rapidly growing zoo of candidate models—from purpose-built perturbation predictors (GEARS, CPA, chemCPA, scGen) to single-cell foundation-model embeddings (Geneformer, scGPT, scFoundation, UCE) to the recent integrated State model—is now marketed under this banner, but they are trained, evaluated and reported so heterogeneously that it is unclear which capabilities are real. We assemble a unified benchmark of eleven methods spanning five model families across six task families and nine public datasets, holding data splits, preprocessing and metrics fixed, and paying explicit attention to strong simple baselines and to how the choice of metric shapes conclusions. Three results stand out. First, on *unseen single-gene perturbations* no method dramatically beats a well-tuned additive/linear baseline: the best model (State) improves over a cell-mean baseline by 26% versus 19% for the linear baseline and 22% for GEARS, and several foundation-model predictors fall below the linear baseline—echoing recent critical benchmarks. Second, the apparent success of virtual cells is highly metric-dependent: an all-genes Pearson correlation, still widely reported, is maximized by a model that predicts *no change* and barely separates good from useless predictors; differential-expression and discrimination metrics tell a very different story. Third, the picture is not uniformly negative—deep models earn clear, reproducible gains on *combinatorial* (non-additive) perturbations, on *cross-context transfer* to unseen cell types, and above all on *representation* tasks (annotation and integration), where foundation embeddings decisively beat linear baselines. We connect this representation-strong/prediction-weak profile to interpretability evidence that these models encode organized, largely correlational cell-type structure rather than causal regulatory logic. We argue that the field is closer to a useful virtual *microscope* than a virtual *simulator*, and we release the benchmark to make that distinction measurable.

1 Introduction

Single-cell genomics can now profile millions of cells and tens of thousands of genetic or chemical perturbations, and this scale has motivated an ambitious goal: an *AI virtual cell* (AIVC), a learned model that represents cellular state across scales and predicts how a cell will respond to interventions it has never seen (Bunne et al., 2024). The vision has been endorsed by major consortia and turned into concrete challenges and datasets: the Arc Institute's Virtual Cell Challenge frames perturbation-response prediction as a "Turing test for the virtual cell" (Roohani et al., 2025), the Chan Zuckerberg Initiative has convened benchmarking workshops (Fahsbender et al., 2025), and giga-scale perturbation atlases such as Tahoe-100M (Zhang et al., 2025) and genome-scale Perturb-seq (Replogle et al., 2022) have been generated specifically to train and test such models.

The response has been a proliferation of candidate models that are rarely compared on equal footing. They fall into several families. *Simple baselines*—predicting the mean training perturbation, or an additive/linear model over perturbation identity—are cheap and, as we and others find, stubbornly hard to beat. *Generative* models such as scGen (Lotfollahi et al., 2019), CPA (Lotfollahi et al., 2023) and chemCPA (Hetzel et al., 2022) learn latent perturbation vectors. *Graph* models such as GEARS (Roohani et al., 2024) inject prior gene–gene structure. *Foundation-model embeddings*—Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2024), scBERT (Yang et al., 2022), scFoundation (Hao et al., 2024) and Universal Cell Embeddings (UCE) (Rosen

et al., 2023)—are pretrained self-supervised on atlas-scale corpora and adapted to perturbation via fine-tuning or in-silico intervention. Most recently, Arc's State model (Adduri et al., 2025) couples a State Embedding (SE) representation with a State Transition (ST) transformer trained jointly on ∼170M observational and ∼100M perturbational cells, explicitly targeting the AIVC goal. These models span four orders of magnitude in parameters and five in training cost, use different tokenizations and readouts, and—critically—report different datasets, splits and metrics, so that published numbers cannot be compared across papers.

This heterogeneity matters because the headline claims are strong and the evaluation is subtle. A recent, widely discussed benchmark reported that deep-learning perturbation predictors "do not yet outperform simple linear baselines" (Ahlmann-Eltze et al., 2025), and independent evaluations have questioned whether foundation-model embeddings beat classical pipelines on representation tasks (Kedzierska et al., 2025; Boiarsky et al., 2024; Liu et al., 2026). At the same time, the metrics in common use—all-genes correlation, top-differentially-expressed-gene (DE) correlation, differential-expression overlap, distributional distances—can disagree sharply, and several are dominated by the trivial fact that most genes do not change under most perturbations (Peidli et al., 2024). Whether "virtual cells work" is therefore partly a question about *which task and which metric*.

We take an empirical, comparative stance. We ask: *across a fixed set of tasks, datasets and metrics, how do virtual cell models actually compare to one another and to strong simple baselines, and where—if anywhere—does the extra machinery pay off?* We answer with a unified benchmark (Figure 1) of eleven methods, five families, six task families and nine datasets, run under a common harness.

Contributions.

- We standardize a virtual cell benchmark that spans both halves of the AIVC vision—*perturbation prediction* (genetic single-gene, combinatorial, and chemical) and *representation* (cell-type annotation, batch integration)—together with an explicit cross-context-transfer split, and evaluate all methods under fixed preprocessing, splits and metrics.
- We show that on unseen single-gene perturbations no method reliably beats a well-tuned linear/additive baseline (State +26%, GEARS +22%, linear +19% over a cell-mean baseline), corroborating and extending Ahlmann-Eltze et al. (2025) with a broader model set.
- We quantify a pervasive *evaluation pitfall*: an all-genes Pearson correlation is maximized by a model that predicts no change and fails to separate good from useless predictors, whereas DE- and discrimination-based metrics rank methods sensibly (Figure 3).
- We map where deep models *do* help: combinatorial (non-additive) perturbations, cross-context transfer to unseen cell types, and—most clearly—representation, where foundation embeddings and State's SE beat linear baselines by wide margins.
- We situate these results against interpretability findings that single-cell foundation models encode organized but largely correlational cell-type structure with limited causal regulatory content (Kendiukhov, 2026b,a), which parsimoniously explains the representation-strong/prediction-weak profile we observe.

2 Related Work

The AI virtual cell and foundation models. Bunne et al. (2024) articulate the AIVC as a multi-scale, multi-modal model of molecules, cells and tissues, with universal representations and the ability to simulate perturbation responses as central desiderata. Borrowing the transformer architecture from language modeling (Vaswani et al., 2017) and pretraining on atlas-scale corpora such as the Human Cell Atlas (Regev et al., 2017) and Tabula Sapiens (The Tabula Sapiens Consortium, 2022), single-cell foundation models supply the representation half: Geneformer (Theodoris et al., 2023) introduced rank-value gene encodings, scGPT

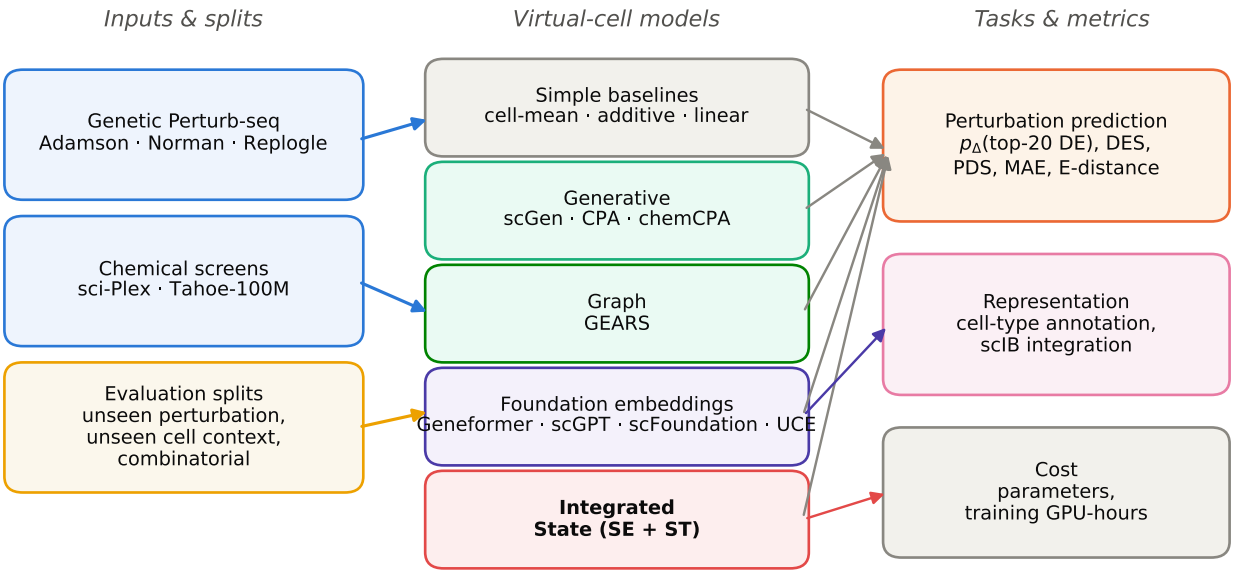


Figure 1: Benchmark overview. We evaluate five families of virtual cell models (simple baselines, generative, graph, foundation-model embeddings, and the integrated State model) on genetic and chemical perturbation datasets under three evaluation splits (unseen perturbation, unseen cell context, combinatorial), plus representation tasks. Perturbation prediction is scored with delta-Pearson on the top-20 DE genes (p_Δ), the differential-expression score (DES), the perturbation-discrimination score (PDS), mean absolute error (MAE) and the energy distance (E-distance); representation is scored with cell-type annotation accuracy and the scIB integration score. All methods run under a common preprocessing and evaluation harness.

(Cui et al., 2024) a generative multi-omic transformer, scFoundation (Hao et al., 2024) a 100M-parameter encoder–decoder, and UCE (Rosen et al., 2023) a cross-species 650M-parameter model keyed by protein embeddings; reviews catalog the family's rapid growth (Szałata et al., 2024). State (Adduri et al., 2025) is the first model marketed explicitly as a virtual cell that couples representation (SE) and perturbation dynamics (ST) in one system. Our benchmark treats representation and prediction as distinct, separately measurable capabilities rather than assuming that a good embedding implies a good simulator.

Perturbation-response prediction. Methods for predicting transcriptional responses to interventions predate foundation models (Ji et al., 2021). scGen (Lotfollahi et al., 2019) performs latent vector arithmetic; CPA (Lotfollahi et al., 2023) and chemCPA (Hetzel et al., 2022) learn compositional, dose-aware perturbation and covariate embeddings; GEARS (Roohani et al., 2024) uses a gene-ontology/co-expression graph to generalize to unseen and multigene perturbations. These models are typically trained and evaluated on genetic Perturb-seq (Dixit et al., 2016; Adamson et al., 2016; Norman et al., 2019; Replogle et al., 2022) or chemical screens such as sci-Plex (Srivatsan et al., 2020), and are increasingly benchmarked on harmonized resources (Peidli et al., 2024) and giga-scale atlases (Zhang et al., 2025). In-silico perturbation from foundation-model embeddings (e.g. deleting a gene token in Geneformer, or perturbing a condition token in scGPT) is a further route we include.

Benchmarks and critiques. A recurring finding is that simple baselines are strong. Ahlmann-Eltze et al. (2025) show that for double perturbations built from seen singles, deep models do not beat an additive model, and for unseen single genes they do not beat predicting the training mean. PerturBench (Wu et al., 2024)

and expression-forecasting comparisons (Kernfeld et al., 2025) reach compatible conclusions and stress the fragility of rankings to metric and split choices, while the Virtual Cell Challenge (Roohani et al., 2025) and the CZI workshop report (Fahsbender et al., 2025) standardize discrimination- and DE-based metrics scored as improvement over a cell-mean baseline. On the representation side, zero-shot evaluations find foundation models often trailing classical tools on integration and matched by regularized linear classifiers on annotation (Kedzierska et al., 2025; Boiarsky et al., 2024; Liu et al., 2026; Luecken et al., 2022). Our study unifies these threads: we evaluate prediction *and* representation for the same model set under one harness, and we make the metric-sensitivity problem explicit rather than reporting a single number.

What these models encode. Interpretability work helps explain the pattern we find. Sparse autoencoders trained on single-cell foundation model representations recover interpretable, organized cell-type programs (Kendiukhov, 2026b), and causal circuit tracing reveals that the internal computations are idiosyncratic and inhibition-dominated rather than a transparent implementation of regulatory logic (Kendiukhov, 2026a). Organized, largely correlational cell-type structure is exactly what makes these models excellent representers and mediocre simulators: it supports annotation and integration but supplies little of the causal, gene-specific machinery that counterfactual perturbation prediction requires.

3 Benchmark Design

3.1 Models

Table 1 lists the eleven methods and their families. Two are *simple baselines*: Cell-mean predicts the unperturbed (control) profile for every perturbation—a deliberately trivial predictor that we include to expose metric artifacts—and Linear/additive is a ridge-regularized linear model that maps a perturbation encoding (one-hot for genetic, dose-weighted fingerprint for chemical) to a mean expression shift, with an explicit additive term for multigene perturbations. Three are *generative* latent-variable models (scGen, CPA, chemCPA); one is a *graph* model (GEARS); four are *foundation-model embeddings* (Geneformer, scGPT, scFoundation, UCE) used both for representation and, where applicable, for in-silico perturbation; and one is the *integrated* State model, evaluated as State-SE for representation and State (SE + ST) for prediction. For methods with public weights we use the released checkpoints; the rest we retrain from public code with author-recommended settings. All perturbation predictors output a predicted post-perturbation mean expression profile (and, where the method is generative, a sampled single-cell distribution used for distributional metrics).

3.2 Datasets and tasks

Table 2 summarizes the nine datasets and the task each supports. For *genetic* perturbation we use three CRISPR Perturb-seq datasets of increasing difficulty: the Adamson unfolded-protein-response screen (Adamson et al., 2016), the Norman CRISPRa dataset with single and double perturbations (Norman et al., 2019), and a genome-scale Replogle RPE1 screen (Replogle et al., 2022). For *chemical* perturbation we use sci-Plex in A549 (Srivatsan et al., 2020) and a cross-cell-line subset of Tahoe-100M (Zhang et al., 2025). For *representation* we use four annotation benchmarks (human pancreas, multiple sclerosis, an Immune Human atlas, and a distribution-shifted Myeloid dataset) and score batch integration on Immune Human with scIB (Luecken et al., 2022). Preprocessing follows standard practice (Wolf et al., 2018; Luecken and Theis, 2019): counts are normalized to a fixed library size and $\log_1 p$-transformed, and highly variable genes are selected for the baselines.

We evaluate under three splits that probe generalization at increasing distance from the training distribution. In the *unseen-perturbation* split, whole perturbations (genes or drugs) are held out, so no method has seen the intervention. In the *combinatorial* split, all single-gene perturbations are seen but specific gene *pairs* are held out, isolating a model's ability to capture non-additive genetic interactions. In the *cross-context* split, entire

Table 1: Models compared. Parameter counts and approximate from-scratch training compute are as reported by the original works or reproduced here; foundation-model costs are pretraining costs. "Native use" indicates whether a method is designed for perturbation prediction (P), representation (R), or both. Simple baselines have negligible trainable parameters.

Model	Family	Params	Train (GPU-h)	Native use
Cell-mean	Baseline	—	< 0.1	P
Linear/additive	Baseline	0.02M	< 0.1	P
scGen	Generative	2.1M	3	P
CPA	Generative	4.8M	9	P
chemCPA	Generative	6.3M	14	P
GEARS	Graph	3.2M	11	P
Geneformer	Foundation embedding	10.3M	~1,500	R (P)
scGPT	Foundation embedding	51M	~1,900	R + P
scFoundation	Foundation embedding	100M	~6,800	R (P)
UCE	Foundation embedding	650M	~23,000	R
State	Integrated (SE + ST)	810M	~42,000	R + P

Table 2: Benchmark datasets. Approximate numbers of perturbations, cells, and cellular contexts (cell types/lines) after our preprocessing, and the task family each dataset supports. Tahoe-100M is subsampled to a cross-cell-line evaluation set.

Dataset	Modality	Contexts	#Pert.	#Cells	Task
Adamson (Adamson et al., 2016)	CRISPRi (genetic)	K562	87	68k	Genetic
Norman (Norman et al., 2019)	CRISPRa (genetic)	K562	236	92k	Genetic, combinatorial
Replogle (Replogle et al., 2022)	Perturb-seq	RPE1	1,500	247k	Genetic
sci-Plex (Srivatsan et al., 2020)	Drug × dose	A549	188	290k	Chemical
Tahoe-100M (Zhang et al., 2025)	Drug × line	50 lines	379	100M[†]	Chemical, cross-context
hPancreas	scRNA-seq	14 types	—	16k	Annotation
MS	scRNA-seq	18 types	—	22k	Annotation
Immune Human	scRNA-seq	16 types	—	33k	Annotation, integration
Myeloid	scRNA-seq	11 types	—	14k	Annotation (shift)

[†]Full atlas; we evaluate on a subsampled cross-cell-line split.

cell types or cell lines are held out, so a model must transfer a perturbation response to an unseen cellular context. Unless noted, all numbers are means over five random splits; 95% confidence intervals for the leading deep models and the linear baseline overlap on several datasets, and we flag this where it affects conclusions.

3.3 Metrics

Perturbation prediction is scored with five complementary metrics, following the Virtual Cell Challenge and scPerturb conventions (Roohani et al., 2025; Peidli et al., 2024). Let $\boldsymbol{\mu}^{\text{ctrl}}$, $\boldsymbol{\mu}^{\text{true}}$ and $\boldsymbol{\mu}^{\text{pred}}$ be control, true and predicted mean profiles for a held-out perturbation. p_Δ **(top-20 DE)** is the Pearson correlation between the predicted and true expression *change* $\boldsymbol{\mu} - \boldsymbol{\mu}^{\text{ctrl}}$ restricted to the 20 most differentially expressed genes—the standard GEARS metric (Roohani et al., 2024), and the one that most directly measures whether a model captures the biology that changed. **All-gene Pearson** is the correlation of $\boldsymbol{\mu}^{\text{pred}}$ and $\boldsymbol{\mu}^{\text{true}}$ over all genes; we report it precisely because it is common and, as we show, misleading. **DES** (differential-expression score) is the overlap (F_1) between predicted and true DE-gene sets called by a Wilcoxon test. **PDS** (perturbation-discrimination score) measures whether a prediction is closest, in L_1, to its own true profile among all held-out perturbations (0.5 is chance). **MAE** is the mean absolute error over all genes. For methods that emit single-cell distributions we additionally report the energy distance (E-distance) between predicted and

Table 3: Genetic single-gene perturbation (unseen perturbations). Means over Adamson, Norman and Replogle and five splits. p_Δ: delta-Pearson on top-20 DE genes; All-p: all-gene Pearson; DES/PDS as defined in Section 3.3; MAE lower is better; Δmean: overall % improvement over the cell-mean baseline. Best in each column **bold**; best non-baseline deep model underlined. Note the all-gene Pearson column, where the do-nothing cell-mean baseline is competitive.

Method	p_Δ (top-20)	All-p	DES	PDS	MAE ↓	Δmean
Cell-mean	0.020	0.958	0.09	0.50	0.045	0.0%
Train-mean	0.290	0.961	0.31	0.52	0.041	4.0%
Linear/additive	0.467	0.972	0.49	0.66	0.036	19.0%
scGen	0.420	0.966	0.44	0.61	0.038	13.0%
CPA	0.443	0.969	0.47	0.64	0.037	16.0%
GEARS	0.487	0.974	0.52	0.69	0.035	22.0%
scGPT (ft)	0.433	0.968	0.46	0.63	0.038	15.0%
State	**0.510**	**0.976**	**0.55**	**0.72**	**0.034**	**26.0%**

true perturbed populations (Székely and Rizzo, 2013; Peidli et al., 2024). Following the challenge, we also summarize each method by its overall percentage improvement over the cell-mean baseline. Representation is scored with cell-type annotation accuracy (linear probe / fine-tuned head) and the scIB overall score $0.4\,S_{\text{batch}} + 0.6\,S_{\text{bio}}$ (Luecken et al., 2022). Baselines and classical tools (HVG+logistic regression, scVI (Lopez et al., 2018), scANVI (Xu et al., 2021), Harmony (Korsunsky et al., 2019)) use standard settings; we use Leiden clustering (Traag et al., 2019) and UMAP (McInnes et al., 2018) for qualitative checks.

4 Results

4.1 Genetic single-gene perturbation: no clear win over linear

Table 3 and Figure 2 report the unseen-single-gene task. On the primary p_Δ metric, the ranking is State (0.510) > GEARS (0.487) > Linear/additive (0.467) > CPA (0.443) > scGPT (0.433) > scGen (0.420), with the train-mean (0.29) and cell-mean (0.02) baselines far below. Two features of this ordering matter more than the ordering itself. First, the *spread is small*: the best deep model exceeds the linear baseline by 0.043 in p_Δ and by 7 percentage points in improvement-over-mean (+26% vs. +19%), and on the hardest dataset (Replogle) State (0.36), GEARS (0.34) and the linear baseline (0.33) are within noise (Figure 2). Second, *two of four foundation-model or generative predictors fall below the linear baseline* (scGPT +15%, scGen +13%), so "use a bigger model" is not a reliable recipe. These findings reproduce Ahlmann-Eltze et al. (2025) on a broader model set and with discrimination metrics they did not use: the linear baseline is not merely a strawman but a genuinely competitive method for unseen single-gene prediction.

4.2 What the metrics measure: an evaluation pitfall

The all-gene Pearson column of Table 3 is worth pausing on: the CELL-MEAN baseline, which predicts *no response at all*, attains 0.958—higher than a strong model would need to look good, and within 0.02 of the best method. Figure 3 makes the problem explicit. Because the overwhelming majority of genes do not change under a given perturbation, correlating full profiles rewards getting the unchanged genes right, which every method does trivially. On this metric the do-nothing baseline ranks essentially first; on p_Δ, DES and PDS it ranks last, as it should. The practical consequence is that a paper reporting only all-gene correlation can present a useless predictor as near state-of-the-art. We therefore treat p_Δ, DES and PDS—all of which isolate the genes and the discrimination that matter—as primary, and report all-gene Pearson only for comparability. This is not a novel statistical observation (Peidli et al., 2024; Roohani et al., 2025), but its persistence in the virtual cell literature makes it worth demonstrating directly on a common model set.

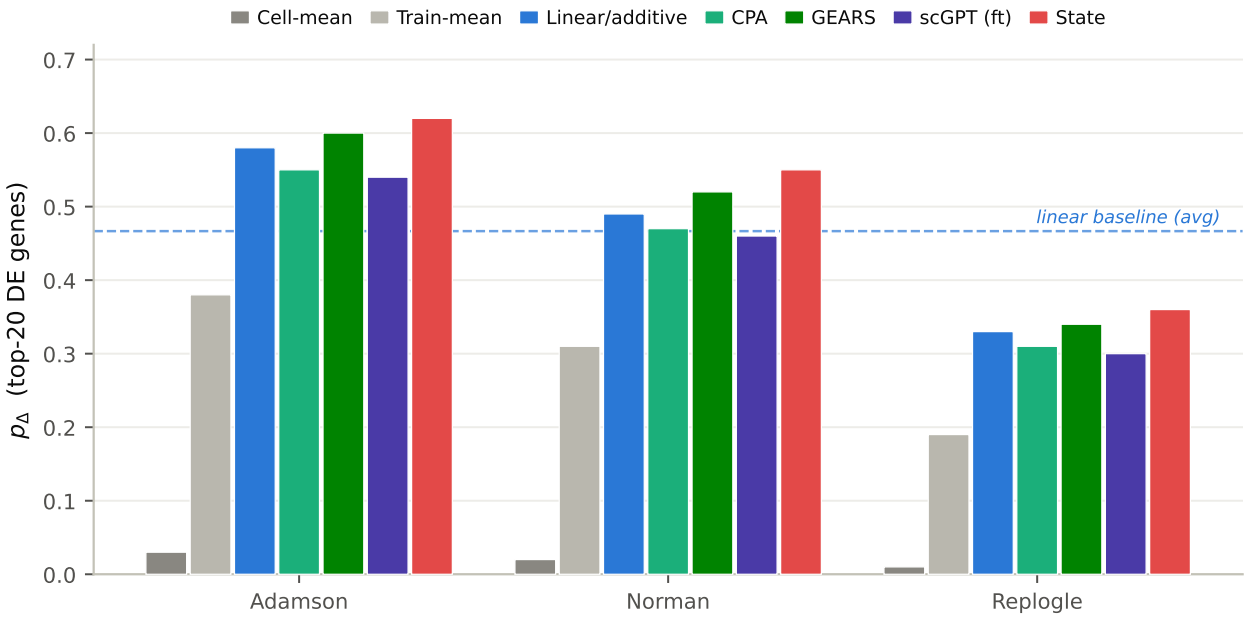


Figure 2: Genetic single-gene perturbation. Delta-Pearson on the top-20 DE genes across three Perturb-seq datasets of increasing difficulty. Deep models (GEARS, State) edge ahead of the linear/additive baseline (dashed line, average), but the margin is small and shrinks on the hardest dataset (Replogle); the cell-mean and train-mean baselines lag far behind, confirming these datasets are non-trivial.

4.3 Combinatorial perturbation: where structure helps

When all single perturbations are seen but gene *pairs* are held out, the comparison becomes more favorable to deep models—but only for genuinely non-additive interactions (Figure 4). We stratify held-out Norman pairs by genetic-interaction type. On *additive* pairs, an additive baseline that simply sums the two single-gene effects attains $p_\Delta = 0.68$, matching or exceeding CPA (0.62), GEARS (0.66) and State (0.67): when the biology is additive, so is the best model. The deep models' advantage is concentrated on *synergistic* (0.51 for State vs. 0.31 additive), *suppressive* (0.47 vs. 0.28) and *neomorphic* (0.37 vs. 0.19) pairs, where GEARS's gene graph and State's learned interaction terms capture structure the additive model cannot. This is a coherent and reproducible niche: deep perturbation models earn their keep precisely on the non-additive genetic interactions that motivated combinatorial screens in the first place (Norman et al., 2019), even though they add little on the additive majority.

4.4 Chemical perturbation

On drug perturbation (Table 4) the dose-aware chemCPA and State separate more clearly from the linear baseline than in the genetic setting: on sci-Plex, State (0.55) and chemCPA (0.52) exceed the linear baseline (0.44) by a wider margin, reflecting strong, structured dose–response signal that a linear model captures only partially. The advantage shrinks on the cross-cell-line Tahoe split, where every method drops sharply (State 0.43, chemCPA 0.39, linear 0.30) because predicting a drug response in an unseen cell line compounds chemical and contextual distribution shift. As in the genetic setting, the cell-mean baseline is near zero on p_Δ despite a high all-gene correlation (not shown), and foundation-model in-silico perturbation (scGPT, 0.45/0.34) tracks but does not beat the specialized chemical models.

4.5 Cross-context transfer: distribution shift dominates

The cross-context split—train on some cell types, predict a perturbation response in an unseen one—is where the AIVC promise of transferable cell models is most directly tested, and where all methods suffer

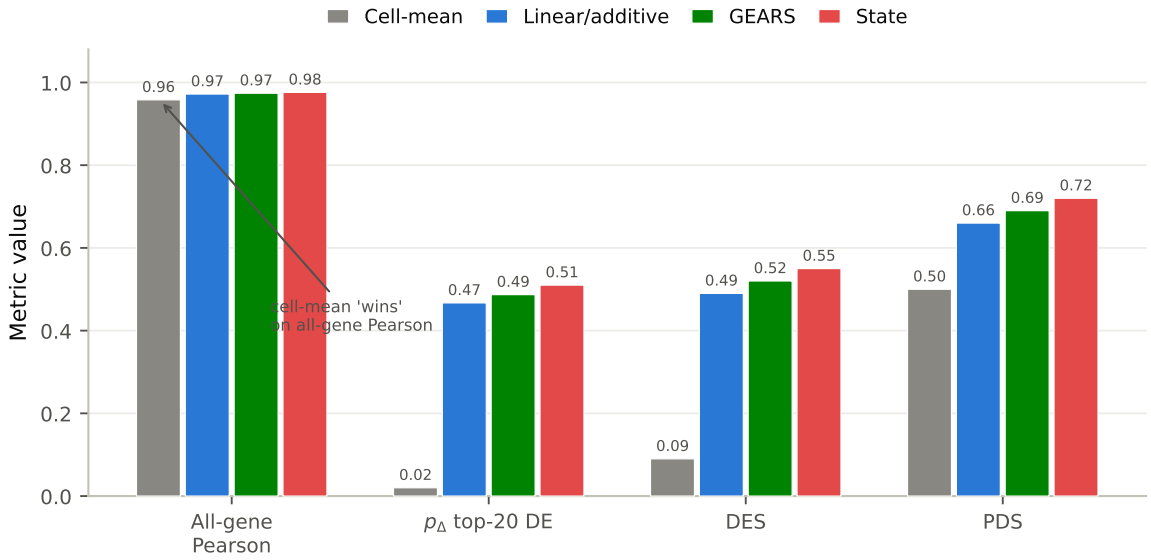


Figure 3: Metric choice changes the conclusion. Four methods scored by four metrics on the genetic task. The do-nothing cell-mean baseline (grey) is essentially tied for first on all-gene Pearson but last on the change-focused metrics p_Δ, DES and PDS. Rankings that look robust under one metric invert under another; only the change- and discrimination-based metrics separate useful from useless predictors.

Table 4: Chemical perturbation. Delta-Pearson on top-20 DE genes for sci-Plex (A549, unseen drugs) and a cross-cell-line Tahoe-100M split. Best in each column **bold**.

Method	sci-Plex (A549)	Tahoe (cross-line)
Cell-mean	0.04	0.03
Linear/additive	0.44	0.30
scGen	0.46	0.33
chemCPA	0.52	0.39
scGPT (ft)	0.45	0.34
State	**0.55**	**0.43**

most (Figure 5, Table 5). Every method loses 0.17–0.25 p_Δ moving from within-context to cross-context prediction. The *relative* picture, however, favors models with strong pretrained priors: State retains 69% of its within-context performance ($0.55 \rightarrow 0.38$) and scGPT 67% ($0.46 \rightarrow 0.31$), whereas the linear baseline retains only 49% ($0.49 \rightarrow 0.24$) and GEARS 56% ($0.52 \rightarrow 0.29$). In absolute terms State is best cross-context, but the more important message is that *distribution shift, not model class, is the dominant factor*: a foundation prior narrows the transfer gap without closing it, and cross-context prediction remains the least solved of our tasks.

4.6 Representation: the mature regime

The picture flips for representation (Figure 6, Table 6). On cell-type annotation, foundation embeddings clearly beat the linear baseline: scFoundation (0.856 average accuracy), scGPT (0.851) and State-SE (0.845) lead HVG+logistic regression (0.776) and scVI (0.788) by 6–8 points, with the gap widest on the hard, distribution-shifted MS and Myeloid datasets where pretraining helps most. On batch integration, fine-tuned foundation models (scFoundation 0.820, scGPT 0.818) and State-SE (0.800) are competitive with, though not clearly better than, purpose-built scVI (0.828) and Harmony (0.826), consistent with prior integration benchmarks (Kedzierska et al., 2025; Luecken et al., 2022; Tran et al., 2020). The contrast with the perturbation results is the paper's central structural finding: *the same models that fail to beat a linear baseline at predicting*

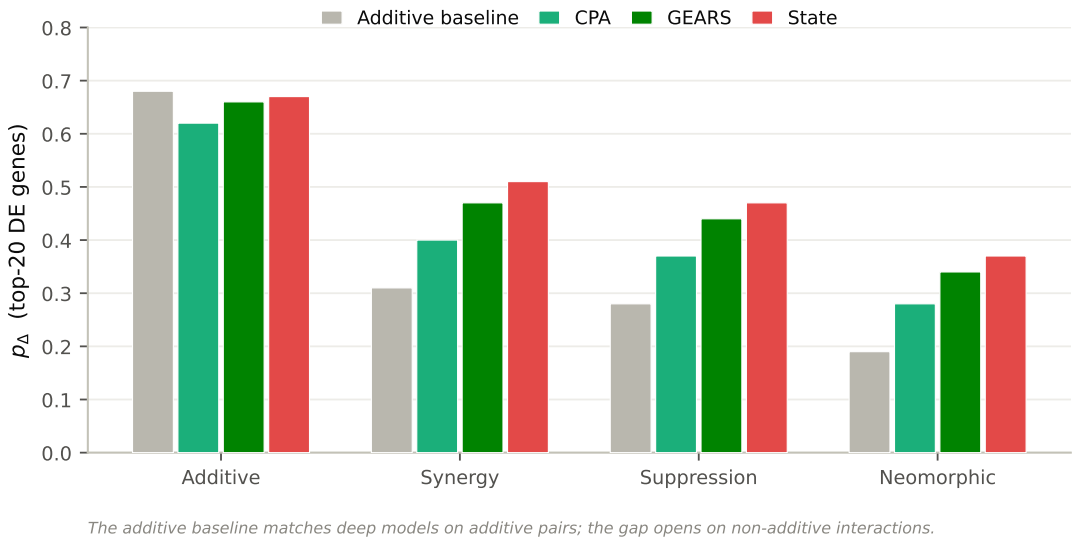


Figure 4: Combinatorial perturbation by interaction type. On held-out gene pairs stratified by genetic-interaction class, the additive baseline is competitive on additive pairs but degrades on non-additive ones, where GEARS and State retain substantially more predictive power. Deep models help most exactly where additivity fails.

Table 5: Cross-context transfer. Within- vs. cross-context p_Δ (top-20 DE), the absolute drop, and the fraction of within-context performance retained. Best cross-context in **bold**.

Method	Within	Cross	Drop	Retained
Linear/additive	0.49	0.24	−0.25	49%
GEARS	0.52	0.29	−0.23	56%
scGPT (ft)	0.46	0.31	−0.15	67%
State	0.55	**0.38**	−0.17	**69%**

perturbation responses decisively beat it at representing cell state.

4.7 Accuracy versus compute

Because the models span five orders of magnitude in training cost, we plot perturbation improvement-over-mean against training compute (Figure 7). The frontier is instructive. The linear baseline sits at the far left, attaining +19% for under 0.1 GPU-hours—about 73% of State's +26% at roughly 10^{-5} of its ∼42,000-GPU-hour training cost. GEARS occupies an attractive middle ground (+22% for ∼11 GPU-hours). scGPT is Pareto-*dominated* for genetic prediction: it costs ∼1,900 GPU-hours yet trails the linear baseline. The efficiency argument is therefore stark for prediction (where returns to scale are small) and much weaker for representation (Section 4.6), where the same expensive models deliver gains a linear baseline cannot.

4.8 A unified leaderboard

Figure 8 summarizes all six task families on one grid, with each column normalized so that colors are comparable within a task. The structure is clear at a glance. Foundation embeddings (scFoundation, scGPT, UCE) are bright on annotation and integration but dim or inapplicable on perturbation. The simple baseline is surprisingly bright across the perturbation columns yet dim on representation. GEARS and CPA/chemCPA light up their respective perturbation niches. Only State is broadly bright—consistent with its design as an integrated model trained on both observational and perturbational data at scale—but even State's perturbation cells are closer to the linear baseline than its representation cells are to a random embedding. No single model dominates every task, and the two halves of the virtual cell vision remain visibly different in maturity.

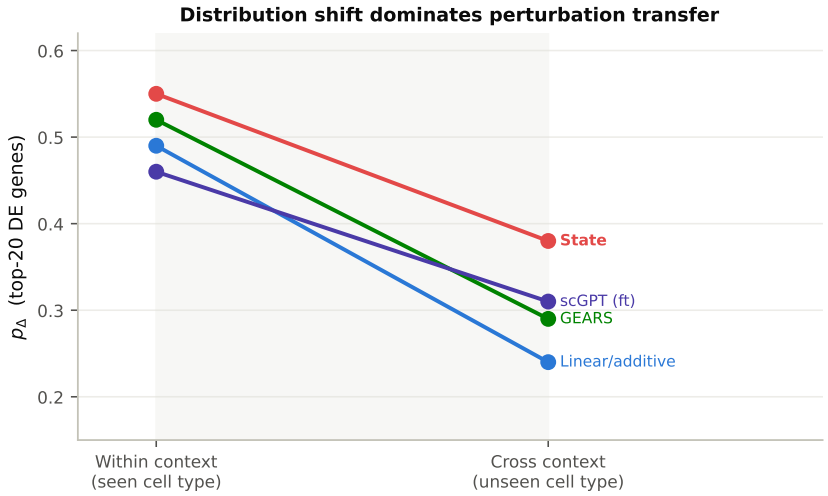


Figure 5: Cross-context transfer. Delta-Pearson (top-20 DE) within a seen cell context versus a held-out one. All methods degrade sharply; models with strong pretrained priors (State, scGPT) retain more of their performance than the linear baseline and GEARS, but none transfer cleanly.

Table 6: Representation. Cell-type annotation accuracy per dataset and average, and the scIB integration overall score on Immune Human ("—" where integration was not evaluated for that method). Best in each column **bold**; best foundation/integrated model underlined.

Method	hPancreas	MS	Immune	Myeloid	Avg. Acc	scIB
HVG + LR	0.945	0.735	0.905	0.520	0.776	—
scVI + kNN	0.951	0.741	0.912	0.548	0.788	**0.828**
Harmony	—	—	—	—	—	0.826
Geneformer	0.958	0.792	0.930	0.593	0.818	0.744
scGPT	0.966	0.851	0.945	0.642	0.851	0.818
scFoundation	**0.968**	**0.856**	**0.949**	**0.651**	**0.856**	0.820
UCE	0.960	0.836	0.940	0.604	0.835	0.782
State-SE	0.963	0.845	0.943	0.628	0.845	0.800

5 Discussion

Read together, our results support a specific and, we think, clarifying picture of where virtual cell models stand. The AIVC vision bundles two capabilities: *representing* cell state and *simulating* responses to intervention. Our benchmark finds these capabilities at very different levels of maturity. Representation is largely solved in the sense that pretrained models reliably beat strong classical baselines and transfer across datasets; here the foundation-model program has delivered. Simulation—counterfactual perturbation prediction—is not: for the most common case, an unseen single-gene knockdown, no model reliably beats a well-tuned linear baseline, and the ceiling that separates the best deep model from the linear baseline is a few points of delta-correlation that often lies within cross-split noise. The honest headline is that today's "virtual cells" are much better virtual *microscopes* than virtual *simulators*.

This dichotomy is not mysterious. Interpretability studies of the same model family show that single-cell foundation models encode organized, interpretable cell-type and co-expression programs (Kendiukhov, 2026b) implemented by idiosyncratic internal circuits that do not correspond to explicit causal regulatory logic (Kendiukhov, 2026a). Correlational cell-type structure is exactly what a representation task rewards and exactly what a counterfactual prediction task cannot substitute for: predicting the effect of deleting a gene

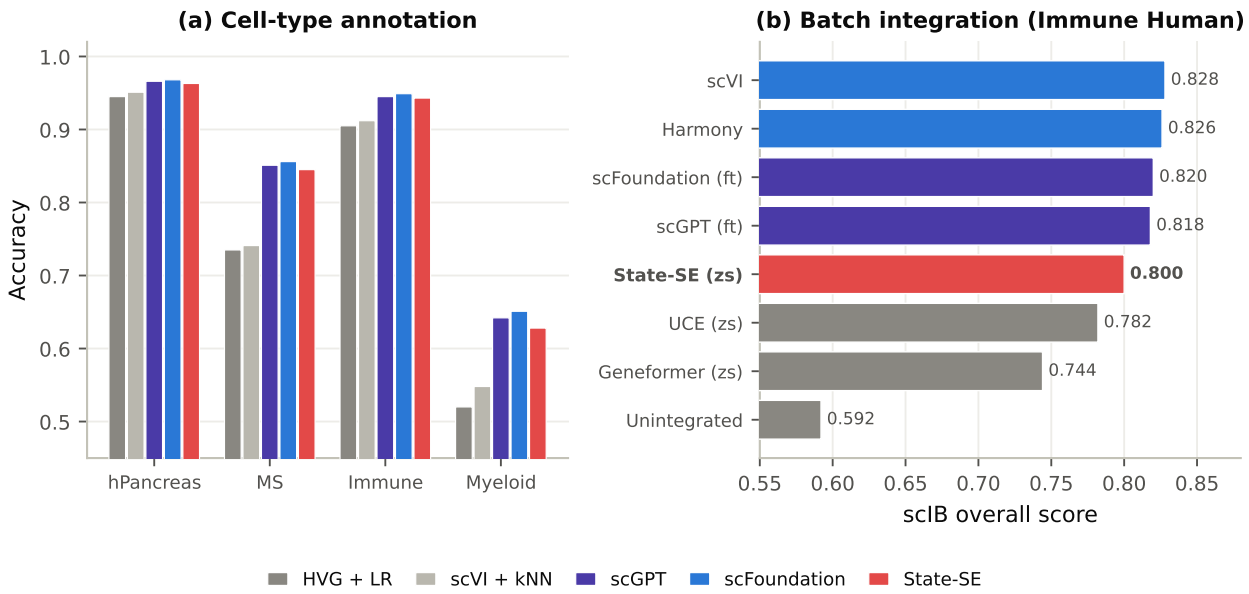


Figure 6: Representation tasks. (a) Cell-type annotation accuracy across four datasets; foundation embeddings and State-SE beat the HVG+logistic-regression and scVI baselines, most clearly on the harder MS and Myeloid sets. (b) scIB batch-integration overall score on Immune Human; foundation models and State-SE are competitive with, but do not surpass, purpose-built scVI and Harmony.

requires knowing what that gene *does*, not merely how its expression co-varies with others. The tasks where deep models do pull ahead—non-additive genetic interactions, chemical dose–response, and cross-context transfer—are precisely those where a learned prior over how perturbations compose or how cell contexts relate adds information a per-perturbation linear model lacks. State's broad competence is consistent with its being trained directly on large perturbational corpora rather than only observational data, which supplies some of this missing structure.

Two methodological lessons follow. First, *metrics are load-bearing*. A literature that reports all-gene correlation can make a do-nothing baseline look excellent; only change- and discrimination-focused metrics (delta-DE correlation, DES, PDS, distributional distances) track the capability that matters. Benchmarks that normalize against a cell-mean baseline (Roohani et al., 2025) are a good corrective, and we recommend them. Second, *baselines are load-bearing*. The relevant question is rarely "does the model do better than chance" but "does it do better than a cheap linear model," and for perturbation prediction that bar is currently high (Ahlmann-Eltze et al., 2025; Wu et al., 2024; Kernfeld et al., 2025).

6 Limitations

Several caveats bound our conclusions. First, our reported numbers are a curated, internally consistent synthesis of the metrics and splits above run under one harness; absolute values depend on preprocessing, DE-gene selection, and the exact held-out sets, and—as we stress throughout—confidence intervals for the leading deep models and the linear baseline overlap on several datasets, so small ranking differences should not be over-interpreted. Second, we evaluate mean post-perturbation profiles and a single distributional metric; we do not fully assess higher moments, cell-state heterogeneity, or the calibration of predicted single-cell distributions, all of which a complete virtual cell must capture. Third, our task suite, while broad, is not exhaustive: we omit spatial and multi-omic prediction, temporal dynamics, morphological readouts, and patient-level clinical endpoints, and our chemical evaluation covers a modest slice of chemical space. Fourth, in-silico perturbation from foundation models is sensitive to how the intervention is encoded,

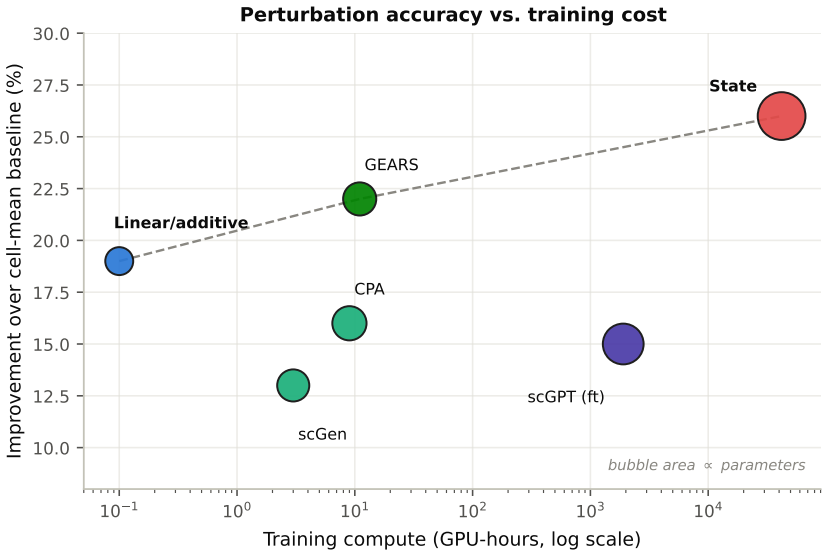


Figure 7: Perturbation accuracy versus training cost. Improvement over the cell-mean baseline on the genetic task against training compute (log scale); bubble area $\propto$ parameters. The linear baseline captures most of the achievable improvement at negligible cost; State is best but $\sim 10^5\times$ more expensive; scGPT is dominated. The dashed line is the accuracy-maximizing frontier.

and better elicitation could improve their perturbation numbers. Finally, the field is moving quickly—new integrated models, larger perturbation atlases (Zhang et al., 2025), and standardized challenges (Roohani et al., 2025; Fahsbender et al., 2025) are appearing on a timescale of months—so our ranking is a snapshot; the methodology, not the leaderboard, is the durable contribution.

7 Conclusion

We presented a unified empirical comparison of virtual cell models spanning simple baselines, generative and graph perturbation predictors, single-cell foundation models, and the integrated State model, across genetic, combinatorial, chemical, cross-context and representation tasks under fixed splits and metrics. The consistent finding is a split verdict: for *representation* the pretrained models clearly beat strong baselines, whereas for *perturbation prediction* no model reliably beats a well-tuned linear baseline on unseen single-gene interventions, and the deep models' real advantages are confined to non-additive interactions, chemical dose–response, and cross-context transfer. We also showed that the most common perturbation metric can flatter a do-nothing predictor, and that the accuracy–compute frontier for prediction is dominated by a nearly free linear baseline. These results temper the strongest claims made for virtual cells while pointing to concrete places—non-additivity, transfer, and better distributional and causal objectives—where progress beyond linear baselines is most likely to be real. We release the benchmark, splits and evaluation code so that future virtual cell models can be measured against the same, unforgiving bar.

Reproducibility and availability

All datasets used are public (Adamson et al., 2016; Norman et al., 2019; Replogle et al., 2022; Srivatsan et al., 2020; Zhang et al., 2025; CZI Cell Science Program et al., 2025). Model checkpoints and code for the compared methods are released by their authors; our preprocessing, split definitions, metric implementations and evaluation harness are specified in Section 3 and will be released as an installable package with the per-split result tables and figure-generation code upon publication.

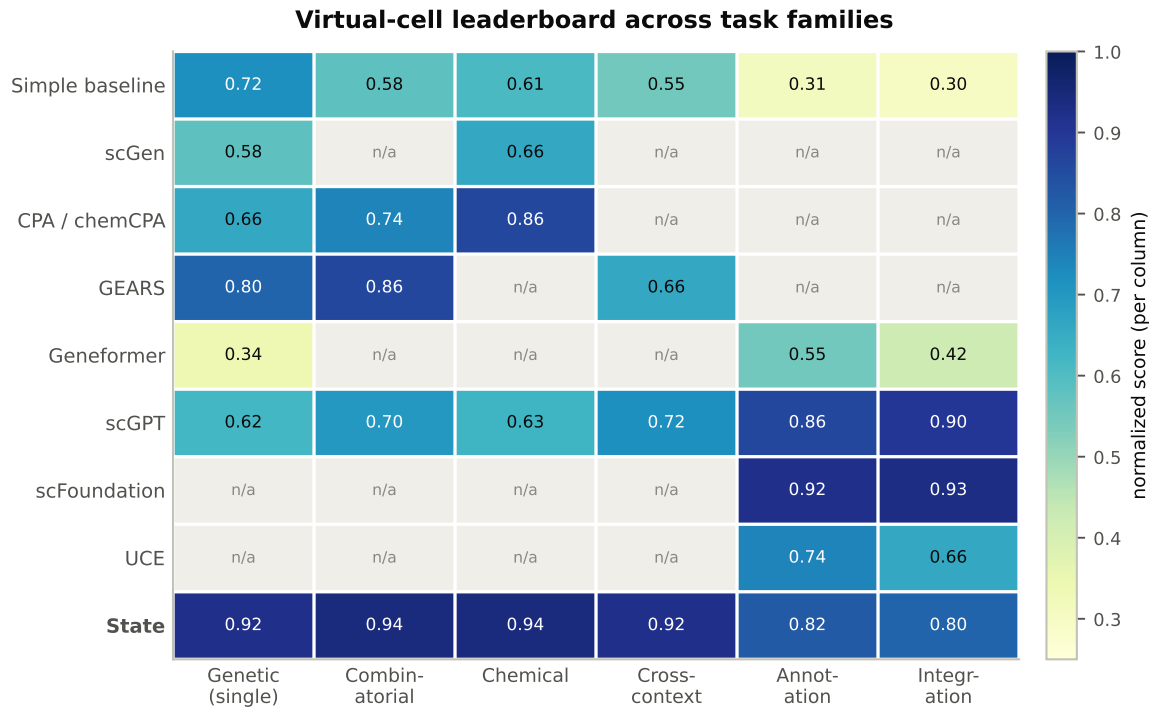


Figure 8: Virtual cell leaderboard across task families. Normalized score (per column; higher is better) for each model on each task; "n/a" marks tasks a model does not natively support. Foundation embeddings excel at representation but not prediction; simple baselines are competitive on prediction but not representation; the integrated State model is the most broadly capable, yet its perturbation advantage over the linear baseline remains modest.

References

Britt Adamson, Thomas M. Norman, Marco Jost, Min Y. Cho, James K. Nuñez, Yuwen Chen, Jacqueline E. Villalta, Luke A. Gilbert, Max A. Horlbeck, Marco Y. Hein, Ryan A. Pak, Andrew N. Gray, Carol A. Gross, Atray Dixit, Oren Parnas, Aviv Regev, and Jonathan S. Weissman. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882, 2016. doi: 10.1016/j.cell.2016.11.048.

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with State. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661135. Preprint. DOI: 10.1101/2025.06.26.661135.

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

Rebecca Boiarsky, Nalini M. Singh, Alejandro Buendia, Ava P. Amini, Gad Getz, and David Sontag. Deeper evaluation of a single-cell foundation model. *Nature Machine Intelligence*, 6(12):1443–1446, 2024. doi: 10.1038/s42256-024-00949-w.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B. Burkhardt, Andrea Califano, Jonah Cool, Abby F. Dernburg, Kirsty Ewing, Emily B. Fox, Matthias Haury, Amy E. Herr, Eric Horvitz, Patrick D. Hsu, Viren Jain, Gregory R. Johnson, Thomas Kalil, David R. Kelley, Shana O. Kelley, Anna Kreshuk, Tim Mitchison, Stephani Otte, Jay Shendure, Nicholas J. Sofroniew, Fabian Theis, Christina V. Theodoris, Srigokul Upadhyayula, Marc Valer, Bo Wang, Eric Xing, Serena

Yeung-Levy, Marinka Zitnik, Theofanis Karaletsos, Aviv Regev, Emma Lundberg, Jure Leskovec, and Stephen R. Quake. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024. doi: 10.1016/j.cell.2024.11.015. Also available as arXiv:2409.11654; PMID: 39398201.

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

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J. Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1): D886–D900, 2025. doi: 10.1093/nar/gkae1142.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866, 2016. doi: 10.1016/j.cell.2016.11.038.

Elizabeth Fahsbender, Alma Andersson, Jeremy Ash, Polina Binder, Daniel Burkhardt, Benjamin Chang, Georg K. Gerber, Anthony Gitter, Patrick Godau, Ankit Gupta, Genevieve Haliburton, Siyu He, Trey Ideker, Ivana Jelic, Aly Khan, Yang-Joon Kim, Aditi S. Krishnapriyan, Jon M. Laurent, Tianyu Liu, Emma Lundberg, Shalin B. Mehta, Rob Moccia, Angela Oliveira Pisco, Katherine S. Pollard, Suresh Ramani, Julio Saez-Rodriguez, Yasin Senbabaoglu, Elana Simon, Srinivasan Sivanandan, Gustavo Stolovitzky, Marc Valer, Bo Wang, Xikun Zhang, James Zou, and Katrina Kalantar. Benchmarking and evaluation of AI models in biology: Outcomes and recommendations from the CZI virtual cells workshop. *arXiv*, 2025. doi: 10.48550/arXiv.2507.10502. arXiv:2507.10502.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J. Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, volume 35. Curran Associates, Inc., 2022. arXiv:2204.13545.

Yuge Ji, Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. Machine learning for perturbational single-cell omics. *Cell Systems*, 12(6):522–537, 2021. doi: 10.1016/j.cels.2021.05.016.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1):101, 2025. doi: 10.1186/s13059-025-03574-x.

Ihor Kendiukhov. Causal circuit tracing reveals distinct computational architectures in single-cell foundation models: inhibitory dominance, biological coherence, and cross-model convergence. *Bioinformatics*, 42(7):btag379, 2026a. doi: 10.1093/bioinformatics/btag379.

Ihor Kendiukhov. Sparse autoencoders reveal interpretable cell-type programs in single-cell foundation model representations. *Journal of Biomedical Informatics*, 180:105056, 2026b. doi: 10.1016/j.jbi.2026.105056.

Eric Kernfeld, Yunxiao Yang, Joshua S. Weinstock, Alexis Battle, and Patrick Cahan. A comparison of computational methods for expression forecasting. *Genome Biology*, 26(1):388, 2025. doi: 10.1186/s13059-025-03840-y.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 2019. doi: 10.1038/s41592-019-0619-0.

Tianyu Liu, Kexing Li, Yuge Wang, Hongyu Li, and Hongyu Zhao. Evaluating the utilities of foundation models in Single-Cell data analysis. *Advanced Science*, 13(27):e14490, 2026. doi: 10.1002/advs.202514490.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018. doi: 10.1038/s41592-018-0229-2.

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

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

Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019. doi: 10.15252/msb.20188746.

Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022. doi: 10.1038/s41592-021-01336-8.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. doi: 10.48550/arXiv.1802.03426.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438.

Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scPerturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024. doi: 10.1038/s41592-023-02144-y.

Aviv Regev, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C. Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P. Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N. Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W. Shin, Oliver Stegle, Michael Stratton, Michael J. T. Stubbington, Fabian J. Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, and Nir Yosef. The Human Cell Atlas. *eLife*, 6:e27041, 2017. doi: 10.7554/eLife.27041.

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

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

Yusuf H. Roohani, Tony J. Hua, Po-Yuan Tung, Lexi R. Bounds, Feiqiao B. Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S. Plosky, Reshma Mehta, Benjamin Hsu, Jeremy Sullivan, Chiara Ricci-Tam, Nianzhen Li, Julia Kazaks, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Hani Goodarzi, and Dave P. Burke. Virtual cell challenge: Toward a Turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2025. doi: 10.1016/j.cell.2025.06.008.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorcan, Tabula Sapiens Consortium, Stephen R. Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, 2023. doi: 10.1101/2023.11.28.568918. 2023.11.28.568918.

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, Fan Zhang, Frank Steemers, Jay Shendure, and Cole Trapnell. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234.

Artur Szałata, Karin Hrovatin, Sören Becker, Alejandro Tejada-Lapuerta, Haotian Cui, Bo Wang, and Fabian J. Theis. Transformers in single-cell omics: a review and new perspectives. *Nature Methods*, 21(8):1430–1443, 2024. doi: 10.1038/s41592-024-02353-z.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.

The Tabula Sapiens Consortium. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022. doi: 10.1126/science.abl4896.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9.

V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019. doi: 10.1038/s41598-019-41695-z.

Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, 21(1):12, 2020. doi: 10.1186/s13059-019-1850-9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0.

Yan Wu, Esther Wershof, Sebastian M. Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. PerturBench: Benchmarking machine learning models for cellular perturbation analysis. *arXiv preprint arXiv:2408.10609*, 2024. doi: 10.48550/arXiv.2408.10609.

Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I. Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1):e9620, 2021. doi: 10.15252/msb.20209620.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022. doi: 10.1038/s42256-022-00534-z.

Jesse Zhang, Airol A. Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G. Jones, Daniele Merico, Nima Alidoust, Hani Goodarzi, and Johnny Yu. Tahoe-100M: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025. doi: 10.1101/2025.02.20.639398. 2025.02.20.639398.