GEOPERT: PERTURBATION PREDICTION AS GEOMETRY-CONSTRAINED DISTRIBUTION MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting transcriptomic responses to genetic perturbations is a fundamental objective of virtual cell modeling. However, because single-cell sequencing is inherently destructive, cellular states cannot be tracked longitudinally within the same physical cell. Instead, assays such as Perturb-seq yield snapshot measurements across cell populations, framing perturbation prediction as a conditional distribution generative modeling problem. Under this formulation, *in silico* generation follows a two-level hierarchical generative process: **1) Population-level sampling**: A condition-specific cell distribution P_c is drawn from a meta-distribution $\mathscr{P}_c$ ($P_c \sim \mathscr{P}_c$) that governs biological and technical variability across replicates, including incubation fluctuations and sequencing batch effects. **2) Cell-level sampling**: Individual cellular profiles are subsequently sampled from this realized population distribution ($x \sim P_c$). Because $\mathscr{P}_c$ defines an abstract distribution over probability measures, typically characterized via kernel mean embeddings in a latent RKHS, most existing frameworks collapse this hierarchy. By approaching in silico generation as a single-step marginal sampling task, they fail to account for replicate-level distributional shifts and are inherently incompatible with large-scale data augmentation schemes. To address these limitations, we propose GeoPert, a hierarchical virtual cell generative framework. GeoPert indirectly characterizes a cell population distribution in the latent RKHS by its energy distances, defined by the mean geodesic *Fisher-Rao* distances over a statistical manifold, against a small set of prototypical anchor populations. For a specific condition, GeoPert first generates its post-perturbation population distribution signature through conditional diffusion, supported by a foundational diffusion model via Classifier-Free Guidance (CFG). Virtual cells of a post-perturbation population are then generated through a variational inference framework conditioned on the energy distance vector signature of the population. By decoupling the virtual cell generation into a two-stage hierarchical generative process, GeoPert naturally supports large-scale data augmentation, by harnessing large-scale synthetic cell populations in the transcriptomic space with matching energy distance vectors, which improves cell generation authenticity. Benchmarks on public Perturb-seq datasets show that GeoPert achieves state-of-the-art performance in single-cell perturbation response prediction and is capable of generalizing to unseen gene perturbations.

1 INTRODUCTION

Predicting transcriptomic responses to genetic perturbations is a fundamental objective of virtual cell modeling, offering an *in silico* testbed for target discovery and cellular reprogramming. In an idealized setting, perturbation modeling aims to infer the counterfactual state of a given physical cell following a genetic intervention(Dixit et al., 2016). However, because single-cell RNA sequencing is inherently destructive, observing pre- and post-perturbation states within the same physical cell is fundamentally impossible(Bunne et al., 2023). This limitation renders cell-level pairing unreliable and precludes the direct application of standard regression frameworks at cellular level.

To bypass this barrier, recent formulations have shifted from single-cell pairing to transitions between probability measures, treating perturbation effects as functional operators mapping unperturbed to perturbed cell distributions. Since the underlying densities are unknown and accessible only through finite, discrete cell samples, these population-to-population mappings are typically learned

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

implicitly: models predict cell-level transitions while enforcing distributional alignment through statistical divergences such as Maximum Mean Discrepancy (MMD)(Gretton et al., 2012).

Nevertheless, a critical statistical reality remains overlooked: biological replicates of the same cell line exhibit non-negligible variability from stochastic incubation fluctuations and technical batch effects(Luecken & Theis, 2019). The empirical distribution under a given perturbation is thus not a fixed ground-truth measure, but a single realization from a meta-distribution over probability measures. Simulating perturbed cellular states therefore entails a two-level hierarchical process: each experimental run first draws a condition-specific population distribution $P_c \sim \mathscr{P}_c$(Fig. 1a), then samples individual cellular transcriptomes $x \sim P_c$.

Accurately capturing this hierarchical generative process is complicated by three challenges:
1, **Intractability of the Latent RKHS:** An empirical cell distribution admits a kernel mean embedding in a Reproducing Kernel Hilbert Space (RKHS)(Muandet et al., 2017), allowing perturbed states to be modeled as distributions over RKHS points. But these embeddings lack readily accessible coordinates, making parameterization of such distributions non-trivial, although pairwise divergences between observed populations remain computable.
2, **Geometric Distortion from Naïve Distance Metrics:** Distribution-level discrepancies depend on cell-to-cell distances, typically computed in size-normalized, log-transformed Euclidean space(Luecken & Theis, 2019). Yet transcriptomes lie on an intrinsic non-Euclidean manifold; imposing an ambient Euclidean metric can misrepresent geodesic distances across transcriptomic states, distorting the resulting distribution-level divergence.
3, **Data Scarcity and Lack of Principled Data Augmentation:** High-throughput perturbation data remain exceptionally scarce and costly. Generative models of raw cellular distributions risk severe overfitting and poor out-of-distribution generalization. Crucially, principled augmentation of distribution-valued targets remains lacking.

To overcome these challenges, we introduce `GeoPert`, a geometric and hierarchical generative framework for virtual cell perturbation prediction. First, treating single-cell sequencing as stochastic sampling, GeoPert interprets size-normalized gene-expression proportions as maximum-likelihood estimates of multinomial parameters, placing cells on a Riemannian statistical manifold equipped with the **Fisher-Rao information metric**. Its geodesic distance corresponds, up to scale, to spherical arc distance between square-root-transformed probability vectors (Townes et al., 2019; Amari, 2016). Second, we make implicit RKHS representations explicit through an **anchor-based multilateration scheme** (Liberti et al., 2014). Computing Fisher-Rao-based Energy Distances (Székely & Rizzo, 2013) between each target distribution and a curated dictionary of anchor populations yields a finite-dimensional distance vector, termed a *distributional multilateration fingerprint*.

Finally, GeoPert decouples cellular simulation into two stages: synthesizing a distributional multilateration fingerprint representing a condition-specific population's kernel mean embedding in the RKHS, then generating single-cell transcriptomic profiles conditioned on that fingerprint. Separating **virtual distribution generation** from **virtual cell generation** enables principled data augmentation. While experimental responses to unseen genetic perturbations remain constrained and scarce, the cell generator can be pretrained on abundant synthetic populations formed by interpolating real cells across distinct empirical populations with varying contribution weights. For distribution generation, GeoPert employs Classifier-Free Guidance (CFG) (Ho & Salimans, 2022), grounding conditional synthesis in an unconditional diffusion prior (Ho et al., 2020) pretrained on empirical and synthetic fingerprints, rather than learning conditional mappings in isolation. This prior encourages generated fingerprints to reflect the geometry of realizable population representations in the RKHS. Figure 1 illustrates the overall framework.

Our Contribution is threefold:
1, **Measuring distribution discrepancies grounded in information geometry:** We formalize cell-to-cell geodesic distances via the Fisher-Rao metric on the multinomial statistical manifold(Amari, 2016; Kass & Vos, 2011), eliminating the geometric distortion inherent to naive Euclidean distances and heuristic log-transformations. We demonstrate that this metric is computationally efficient while effectively balancing qualitative gene detection (on/off states) with quantitative changes.
2, **Distributional characterization via multilateration in abstract RKHS:** We introduce an anchor-based multilateration framework using Fisher-Rao Energy Distance, transforming abstract kernel mean embeddings in the RKHS into structured, finite-dimensional distance vectors. We show the multilateration mapping is isometric and hence injective when anchors are dense in the relevant

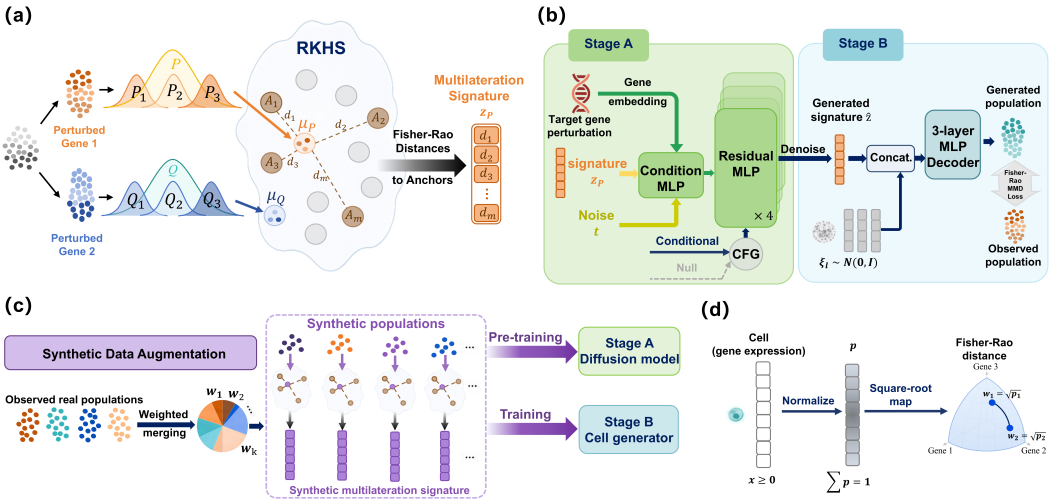


Figure 1: **GeoPert performs geometry-constrained hierarchical perturbation prediction. (a)** RKHS anchor distances define multilateration signatures. **(b)** Conditional diffusion generates population signatures, followed by cell-level decoding. **(c)** Weighted mixtures create synthetic populations and signatures to pre-train Stage A and train Stage B. **(d)** The Fisher-Rao sphere measures arc distances between normalized gene-expression distributions.

population family(detailed in Appendix B.3). For finite anchors, pairwise distances are preserved up to an error explicitly controlled by the anchor covering radius(detailed in Appendix B.4).

3, **Decoupled hierarchical generation with geometry-guided diffusion:** We formulate a two-stage architecture decoupling virtual distribution synthesis from single-cell generation, enabling principled augmentation through synthetic populations formed by interpolating empirical cells across populations. By steering an unconditional diffusion prior via Classifier-Free Guidance (CFG), GeoPert generates multilateration fingerprints guided by the latent RKHS geometry before sampling high-fidelity single cells.

2 Preliminary

2.1 Single Cell Sequencing is a Stochastic Sampling Process

As in representation learning, a foundational challenge in scRNA-seq analysis is defining a principled cellular similarity metric, motivating re-examination of conventional preprocessing and comparison pipelines. Raw scRNA-seq measurements comprise discrete gene-wise molecular counts from stochastic capture of polyadenylated (poly-A) transcripts. Assuming largely unbiased capture across transcripts, each observed profile constitutes a sparse random sample, typically capturing only 1–5% of the cell's total mRNA repertoire(Luecken & Theis, 2019; Townes et al., 2019).

To eliminate technical disparities in sequencing depth (library size) and orient similarity measures toward relative transcriptomic composition, canonical pipelines normalize raw counts by each cell's total detected transcripts and compare profiles using Euclidean distance. Log transformation, typically $\log(1+x)$, is then applied to stabilize gene-level variance and limit dominance by abundant transcripts, recalibrating Euclidean differences to approximate fold changes in gene proportions.

However, this similarity formulation is geometrically inconsistent with the intrinsic statistical structure of count-based transcriptomic data. Given the low capture efficiency typical of scRNA-seq assays, the underlying molecular capture process is well-approximated by sampling with replacement from the cellular mRNA pool. Consequently, the observed raw count vector $\mathbf{y}_i \in \mathbb{N}^G$ of cell i is naturally governed by a multinomial likelihood(Townes et al., 2019):

$$\mathbf{y}_i \mid n_i, \mathbf{p}_i \sim \text{Multinomial}(n_i, \mathbf{p}_i), \qquad \mathbf{p}_i \in \Delta^{G-1}. \tag{1}$$

where $n_i = \sum_g^G y_{ig}$ represents total sequencing depth and Δ^{G-1} denotes the probability simplex.

Under this model, library-size normalization is precisely the maximum likelihood estimator (MLE) of relative abundances, $\hat{\mathbf{p}}_i = \mathbf{y}_i/n_i$. However, Euclidean distances between multinomial parameters, $d_E(\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_j)$, or their log-transformed counterparts, $d_E(\log \hat{\mathbf{p}}_i, \log \hat{\mathbf{p}}_j)$, do not generally reflect statistical divergence between the corresponding Multinomial distributions.

2.2 Statistical Manifold and Fisher-Rao Distance

Because normalized single-cell profiles represent maximum likelihood estimates (MLEs) of multinomial parameters, information geometry provides the natural foundation for measuring cellular dissimilarities. The Fisher Information Matrix (FIM) acts as a Riemannian metric tensor reflecting log-likelihood curvature, defining an intrinsic geodesic distance, the Fisher–Rao distance d_{FR}, over the statistical manifold(detailed in Appendix A).

For the multinomial family, the square-root map $\psi(\mathbf{p}) = (\sqrt{p_1}, \ldots, \sqrt{p_G})^\top$ isometrically embeds the probability simplex Δ^{G-1} into the positive unit hypersphere $\mathbb{S}_+^{G-1}$, reducing the Fisher-Rao distance to the spherical angular distance:

$$d_{\mathrm{FR}}(\mathbf{p}, \mathbf{q}) = \arccos\left(\sum_{g=1}^{G} \sqrt{p_g q_g}\right). \tag{2}$$

This geometric formulation offers three key advantages over canonical log-Euclidean pipelines: 1), **Intrinsic Geodesic Alignment**: It measures exact geodesic distances along the statistical manifold rather than extrinsic chords through Euclidean space. 2) **Boundary Integrity**: It natively accommodates zero expression without distorting metric geometry through arbitrary pseudo-counts ($\log(x+c)$). 3) **Equitable Dynamic Sensitivity**: It balances qualitative gene activation (binary on/off states) with quantitative changes in expression magnitude, avoiding the pathology of log transforms that over-amplify low-abundance fluctuations.

2.3 Virtual Cell Generation from a Distribution Perspective

Because destructive assays preclude observing paired pre- and post-perturbation states within the same physical cell, the supervisory paradigm necessarily shifts from instance-level cell pairing to **population-level distribution matching**. Formally, let $P_{\mathrm{ctrl}}, P_\tau \in \mathcal{P}(\mathcal{X})$ denote the control and perturbed probability distributions over cellular transcriptomic profiles $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^{|\mathcal{G}|}$, where $\mathcal{G}$ represents the set of profiled genes. Predicting cellular responses to a perturbation τ thus reduces to **learning a generative mapping that minimizes a statistical discrepancy** $\mathcal{D}$ between the predicted distribution $\hat{P}_\tau$ and the empirical ground truth P_τ:

$$Q_\theta^\tau = \mathcal{F}_\theta(P_{\mathrm{ctrl}}, e_\tau), \qquad \mathcal{L}_{\mathrm{pop}} = \frac{1}{|\mathcal{T}_{\mathrm{train}}|} \sum_{\tau \in \mathcal{T}_{\mathrm{train}}} \mathcal{D}(Q_\theta^\tau, P_\tau). \tag{3}$$

Subsequently, individual virtual cell profiles are synthesized by sampling from the predicted condition-specific distribution: $\widehat{\mathbf{x}}_i \sim Q_\theta^\tau$. In this work, each genetic perturbation τ is represented by a pre-trained gene embedding $\mathbf{e}_\tau \in \mathbb{R}^d$ encoding biological prior knowledge. Our objective is out-of-distribution (OOD) generalization: forecasting the transcriptomic responses to previously unobserved genetic perturbations, where the test perturbations are strictly disjoint from those seen during training ($\mathcal{T}_{\mathrm{train}} \cap \mathcal{T}_{\mathrm{test}} = \emptyset$).

2.4 Reproducing Kernel Hilbert Space and Maximum Mean Discrepancy

In machine learning, distribution alignment is frequently formulated via the Maximum Mean Discrepancy (MMD), a non-parametric divergence comparing probability distributions through their embeddings in a Reproducing Kernel Hilbert Space (RKHS). Given a symmetric, positive-definite kernel $k(\cdot, \cdot)$, the canonical feature map $\phi(\mathbf{x}) = k(\mathbf{x}, \cdot)$ maps samples into $\mathcal{H}_k$, embedding distributions P and Q as:

$$\mu_P = \mathbb{E}_{\mathbf{x} \sim P}[k(\mathbf{x}, \cdot)], \qquad \mu_Q = \mathbb{E}_{\mathbf{y} \sim Q}[k(\mathbf{y}, \cdot)]. \tag{4}$$

The squared MMD corresponds to the RKHS norm distance between these embeddings:

$$\mathrm{MMD}_k^2(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k}^2 = \mathbb{E}[k(\mathbf{x}, \mathbf{x}')] + \mathbb{E}[k(\mathbf{y}, \mathbf{y}')] - 2\mathbb{E}[k(\mathbf{x}, \mathbf{y})], \tag{5}$$

Provided that k is a characteristic kernel (such as the Energy Distance kernel induced by Euclidean or geodesic metrics), the embedding map is strictly injective, uniquely characterizing the distribution's characteristic function and all associated moments (mean, variance, skewness, etc.). Consequently, $\mathrm{MMD}_k(P, Q) = 0$ if and only if $P = Q$. Thus, MMD constitutes an exact surrogate for distribution equality, enabling robust population-level supervision in unpaired generative modeling.

2.5 Multilateration in High-Dimensional Space

While an RKHS lacks an explicit coordinate system, precluding direct numerical representation of the mean embedding μ_P, inter-distribution distances within $\mathcal{H}_k$ are analytically tractable. Specifically, the RKHS distance between population embeddings equates to the MMD, which is mathematically identical to the Energy Distance under negative-type metric kernels. This enables an empirical, coordinate-free parameterization of distribution geometries via multilateration.

In metric geometry, an element in a coordinate-free space is localized by its distances to designated landmarks. For a point $\mathbf{u} \in \mathbb{R}^d$, its distance signature $\Phi_M(\mathbf{u}) = (\|\mathbf{u} - \mathbf{a}_j\|_2)_{j=1}^M$ defines an injective map $\Phi_M : \mathbb{R}^d \to \mathbb{R}^M$ whenever the M anchors include $d + 1$ affinely independent points. Analogously, we represent any cell distribution P by its multilateration signature across a diverse bank of anchor distributions. Because the distance metric is Lipschitz continuous, local neighborhood structures and distributional divergences in $\mathcal{H}_k$ are faithfully preserved in the Euclidean space of these finite signatures.

3 Related Works

Perturbation response prediction.

Ahlmann-Eltze et al. (2025) establish linear aggregate-expression predictors as strong baselines. GEARS (Roohani et al., 2024) and scGPT (Cui et al., 2024) use biological graphs and pretrained gene-expression representations, respectively, but their commonly used pipelines randomly pair control and perturbed cells, favoring average-effect prediction. CPA separates basal states and perturbation effects in latent space through an adversarial compositional autoencoder (Lotfollahi et al., 2023). State uses kernel-based set-to-set prediction without cell pairing (Adduri et al., 2026). However, population matching alone does not explicitly model variability across population-level responses.

Population-level generative modeling. PerturbDiff models response variability through a perturbation-conditioned distribution over cell-population embeddings $\mu_{P_\tau} \in \mathcal{H}_k$ (Yuan et al., 2026). It trains a denoising model to generate **cell collections** under distributional supervision. While pioneering the concept, it approximates the sampling in RKHS by a first-order feature-map expansion under suitable differentiability assumptions, thus only providing an interpretation of training stochasticity arisen from data sparsity and fails to establish an equivalence to genuine random sampling in RKHS.

Diffusion models in function spaces. Function-space diffusion reverses Gaussian corruption of functions (Kerrigan et al., 2023; Pidstrigach et al., 2024; Lim et al., 2025). Applied to implicit kernel mean embeddings, it requires discretization through function evaluations or truncated bases and a covariance ensuring RKHS-valued noise. Standard isotropic Gaussian noise is not RKHS-valued in infinite dimensions, making noise construction a mathematical requirement rather than merely an implementation choice.

4 Methods

Motivation. Destructive sequencing precludes before-after cell pairing, motivating distribution-level prediction. Many methods match a **single pooled target** $\overline{P}_\tau$ per perturbation τ, represented by one RKHS point. Yet batch effects and cell-cycle composition can produce different cell-group distributions, motivating a distribution over embeddings rather than one pooled point(Fig. 1 a). PerturbDiff targets this variability but approximates RKHS noising through a first-order expansion of cell-space operations, without guaranteeing equivalence along the full diffusion trajectory (Yuan et al., 2026).

We therefore model **distributions over cell-group distributions**, decoupling population prediction from cell generation. Let $P_{\tau,r}$ denote the empirical distribution of group r under τ, with anchor-distance signature $z_{\tau,r} = T_M(P_{\tau,r})$. Our objectives are

$$\begin{aligned} q_\theta(\cdot \mid e_\tau) &\approx P_z(\cdot \mid \tau), \quad \text{population-level modeling,} \\ p_\phi(\cdot \mid z_{\tau,r}) &\approx P_{\tau,r}, \qquad \text{cell-expression generation,} \end{aligned} \tag{6}$$

$P_z(\cdot \mid \tau)$ is the distribution of group signatures and e_τ is an externally pretrained gene embedding.

4.1 Representing Population Distributions as Multilateration Vectors

We represent each population by its distances to fixed anchor populations $\mathcal{A}_M = \{A_1, \ldots, A_M\}$. Let $d(P, Q) := \mathrm{MMD}_{k_{\mathrm{FR}}}(P, Q)$. The kernel k_{FR} is defined in Section 4.2. Then

$$T_M(P) = \big(d(P, A_j)\big)_{j=1}^{M} = \big(\|\mu_P - \mu_{A_j}\|_{\mathcal{H}_{k_{\mathrm{FR}}}}\big)_{j=1}^{M} \in \mathbb{R}^M_{\geq 0}. \tag{7}$$

With distortion controlled by anchor coverage:

Proposition 1 (Coverage-dependent distortion)**.** *Let $(\mathcal{S}, d)$ be compact and $\mathcal{A}_M \subseteq \mathcal{S}$. Define the covering radius $\varepsilon_M = \sup_{P \in \mathcal{S}} \min_{1 \leq j \leq M} d(P, A_j)$. Then, for all $P, Q \in \mathcal{S}$,*

$$0 \leq d(P, Q) - \|T_M(P) - T_M(Q)\|_\infty \leq 2\varepsilon_M. \tag{8}$$

Proof in Appendix B.4.

Farthest-first selection. Proposition 1 motivates minimizing the empirical covering radius over training populations $\mathcal{C}$:

$$\mathcal{A}_M^\star \in \operatorname*{arg\,min}_{\substack{\mathcal{A} \subseteq \mathcal{C} \\ |\mathcal{A}| = M}} \max_{P \in \mathcal{C}} \min_{A \in \mathcal{A}} d(P, A). \tag{9}$$

We approximate this objective using medoid-initialized farthest-first traversal (Gonzalez, 1985):

$$A_{s+1} \in \operatorname*{arg\,max}_{P \in \mathcal{C} \setminus \mathcal{A}_s} \min_{A \in \mathcal{A}_s} d(P, A), \qquad \mathcal{A}_{s+1} = \mathcal{A}_s \cup \{A_{s+1}\}. \tag{10}$$

Each step targets the least-covered population, seeking to tighten the representation-error bound. On K562, Euclidean distances between 100-anchor vectors strongly correlate with population MMD distances(Pearson $r = 0.88$). Correlation remains stable across 20–140 anchors; we therefore treat anchor count as a hyperparameter, with a default value set to 100 (Fig. 5).

4.2 Measuring MMD on a Statistical Manifold with Fisher-Rao Kernel

The Fisher-Rao angular distance d_{FR} is conditionally negative definite on the probability simplex. For any fixed reference $p_0 \in \Delta^{|\mathcal{G}|-1}$, it induces the positive-semidefinite kernel

$$k_{\mathrm{FR}}(x, y) = \frac{1}{2}\left[d_{\mathrm{FR}}(x, p_0) + d_{\mathrm{FR}}(y, p_0) - d_{\mathrm{FR}}(x, y)\right]. \tag{11}$$

Reference terms cancel, yielding the MMD between empirical measures $\widehat{P} = n^{-1}\sum_i \delta_{x_i}$ and $\widehat{Q} = m^{-1}\sum_j \delta_{y_j}$:

$$d^2(\widehat{P}, \widehat{Q}) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} d_{\mathrm{FR}}(x_i, y_j) - \frac{1}{2n^2}\sum_{i,i'=1}^{n} d_{\mathrm{FR}}(x_i, x_{i'}) - \frac{1}{2m^2}\sum_{j,j'=1}^{m} d_{\mathrm{FR}}(y_j, y_{j'}). \tag{12}$$

This equals half the empirical Fisher-Rao energy distance, requiring only pairwise cell distances. The kernel is characteristic on the simplex, so $d(P, Q) = 0$ only when $P = Q$ (Appendix B.2).

4.3 Generating Population Distributions with Diffusion

Conditional diffusion. We learn $P_z(\cdot \mid \tau)$ through diffusion in anchor-distance space. Each training group provides a clean target $z_0 = z_{\tau,r}$, corrupted as

$$z_t = \alpha_t z_0 + \sigma_t \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_M), \qquad \alpha_t^2 + \sigma_t^2 = 1, \tag{13}$$

with $\alpha_T \approx 0$ and $\sigma_T \approx 1$. The denoiser predicts velocity conditioned on $h = e_\tau$:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0,h,t,\epsilon}\left[\|v_\theta(z_t,t,h) - (\alpha_t\epsilon - \sigma_t z_0)\|_2^2\right]. \tag{14}$$

Augmentation-based pretraining. To supplement limited population-level supervision, we first learn an unconditional prior from real and synthetic signatures. We select two training anchors A_i, A_j, independently sample $x_\ell^{(i)} \sim A_i$ and $x_\ell^{(j)} \sim A_j$, and construct

$$\widetilde{x}_\ell = w x_\ell^{(i)} + (1-w)x_\ell^{(j)}, \qquad w \in [0,1], \tag{15}$$

with one weight shared across the synthetic population. Its empirical distribution $\widetilde{P}$ yields $\widetilde{z} = T_M(\widetilde{P})$. Real and synthetic signatures pretrain the denoiser through equation 14 with $h = \emptyset$, initializing subsequent conditional training(details in Appendix D.4).

4.4 Virtual Cell Generation Guided by Population Distribution Signatures

The decoder G_ϕ realizes $p_\phi(\cdot \mid z)$ as a population-conditioned pushforward of a Gaussian prior. For each signature, it generates an empirical distribution of m cells:

$$\widehat{P}_{\phi,\tau,r} = \frac{1}{m}\sum_{i=1}^{m}\delta_{G_\phi(z_{\tau,r},\xi_{\tau,r,i})}, \qquad \xi_{\tau,r,i} \overset{\text{iid}}{\sim} \mathcal{N}(0,I), \tag{16}$$

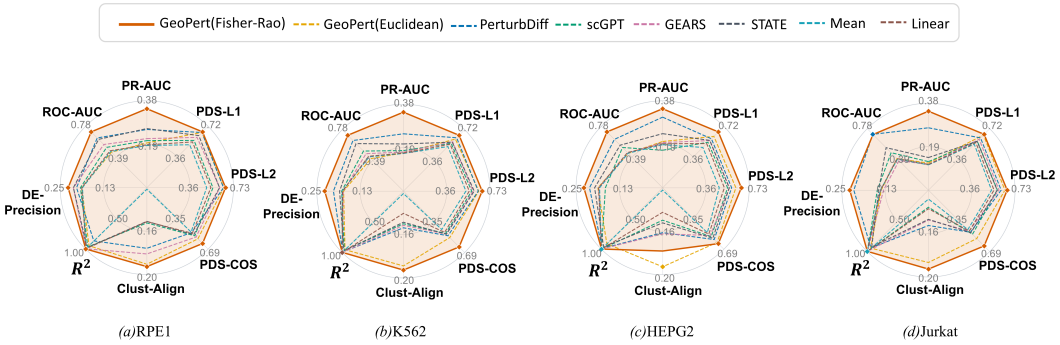
δ_x denotes a point mass. For cell-group indices $\mathcal{I}$, we minimize the Fisher-Rao MMD loss

$$\mathcal{L}_{\text{dec}} = \frac{1}{|\mathcal{I}|}\sum_{(\tau,r)\in\mathcal{I}} d^2\left(\widehat{P}_{\phi,\tau,r}, P_{\tau,r}\right), \tag{17}$$

Decoder augmentation. Unlike perturbation labels, population signatures are computable for synthetic data. Hierarchical decoupling therefore enables label-free decoder augmentation: pairs $(\widetilde{z}, \widetilde{P})$ constructed by equation 15 supplement real pairs under the same objective in equation 17.

5 Results

5.1 Predicting Single-Gene Perturbation Transcriptional Responses



*(a)*RPE1 *(b)*K562 *(c)*HEPG2 *(d)*Jurkat

Figure 2: **Prediction of unseen single-gene perturbations across methods, metrics, and datasets.** Each axis represents a higher-is-better metric. **GeoPert achieves the best overall performance across datasets and metrics.**

GeoPert generalizes across cellular contexts. We evaluated unseen gene perturbation prediction in RPE1, K562 (Replogle et al., 2022), HepG2, and Jurkat (Nadig et al., 2025) against six baselines with eight metrics covering average expression, population structure, and differential-expression recovery (Fig. 2). GeoPert ranked first or among the leading methods across datasets and metrics.

Table 1: Norman double-perturbation benchmark.

	0/2 seen		1/2 seen		2/2 seen	
Method	**PDS-L1 ↑**	**ROC-AUC ↑**	**PDS-L1 ↑**	**ROC-AUC ↑**	**PDS-L1 ↑**	**ROC-AUC ↑**
PerturbDiff	0.7083	0.7813	0.7583	0.7821	0.8978	0.7956
GEARS	0.7292	0.4572	0.7774	0.4559	0.9067	0.4884
STATE	0.6667	0.7531	0.7536	0.7716	0.9422	0.7949
CPA	0.5417	0.3293	0.5418	0.3814	0.7022	0.4272
Train Mean	0.5417	0.4051	0.5114	0.4199	0.5333	0.4307
Linear	0.5347	0.5748	0.5744	0.6285	0.7467	0.6089
GeoPert	**0.7708**	**0.8299**	**0.8166**	**0.8440**	**0.9689**	**0.8717**
GeoPert(Euclidean)	0.5412	0.7650	0.6078	0.7701	0.6570	0.7867

GeoPert recovers gene responses while preserving population structure. GeoPert excelled in differential-expression recovery, frequently leading PR-AUC, ROC-AUC, and DE-Precision, while strong PDS and Clust-Align scores indicated preservation of perturbation identity and population structure. PerturbDiff remained competitive on selected gene-level metrics but was less consistent at the population level. Replacing the final Fisher-Rao MMD loss with Euclidean distance-induced MMD weakened multiple differential-expression metrics despite competitive population-level scores, proving that Fisher-Rao supervision for recovering perturbation-responsive genes.

5.2 Predicting combined dual-gene perturbation outcomes

We evaluated 2-gene generalization in the Norman dataset (Norman et al., 2019), with 131 2-gene perturbations. Three settings distinguish whether neither (**0/2 seen**), one (**1/2 seen**), or both (**2/2 seen**) constituent genes were observed as single-gene perturbations during training.

GeoPert achieved the highest PDS-L1 and ROC-AUC across all settings (Table 1), improving perturbation discrimination and, particularly, differential-expression recovery. Even in the hardest 0/2-seen setting, its ROC-AUC reached 0.8299 versus 0.7813 for the strongest baseline, indicating better identification of responsive genes. Replacing the final Fisher-Rao MMD loss with Euclidean distance-induced MMD markedly reduced both metrics across all settings, supporting information-geometric distribution matching for combinatorial response prediction.

5.3 Unconditional pretraining improves population-level modeling

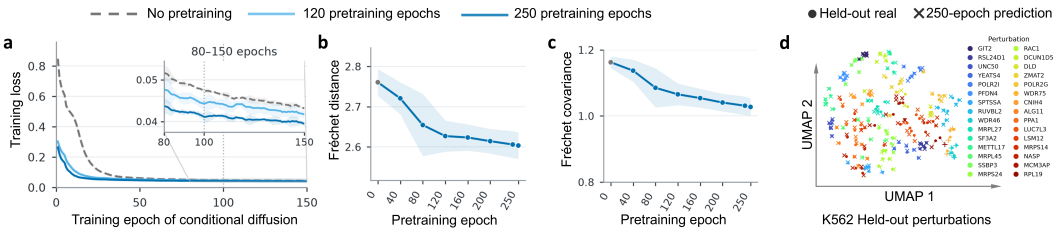


Figure 3: **Effect of population pretraining on conditional diffusion and held-out prediction. (a)** Conditional diffusion loss with 0, 120, or 250 pretraining epochs. **(b,c)** Fréchet distance and its covariance component between predicted and observed signature distributions versus pretraining epochs. **(d)** UMAP of real populations and predictions, colored by perturbation. **More pretraining improves optimization and held-out distributional matching (a–c), while the 250-epoch predictions faithfully recover the perturbation-specific population structure (d).**

Unconditional pretraining captures shared structure in anchor-distance space, providing a transferable prior for population-level modeling. Across datasets, we varied the pretraining duration while keeping the subsequent conditional-training budget fixed. Pretrained models converged faster and achieved lower conditional diffusion losses, with longer pretraining providing further improvements even during late-stage training (Fig. 3a).

On held-out perturbations, the Fréchet distance between predicted and observed population-signature distributions progressively decreased with additional pretraining (Fig. 3b). Its covariance component measures discrepancies in distributional spread and cross-coordinate covariation, separately from differences in the means. This component also decreased as pretraining increased (Fig. 3c), indicating better recovery of the magnitude and structure of variation across cell-group signatures, rather than merely their average location. Joint UMAP visualization of held-out K562 perturbations showed that predicted anchor-distance signatures remained distinguishable across perturbations while aligning closely with observed signatures for the same perturbation. This agreement supports GeoPert's generalization to unseen perturbations while preserving perturbation-specific population responses.(Fig. 3d). These results support unconditional pretraining as a useful prior for accelerating conditional optimization and improving the fidelity of population-level distributional predictions.

5.4 Balanced Population Augmentation Improves Cell-level Generation

We varied the fraction of virtual populations in decoder training while holding other settings fixed (Fig. 4). Moderate augmentation improved DE-Precision, PR-AUC, and ROC-AUC, with the strongest overall performance at intermediate proportions, including the default $\rho_{\text{virtual}} = 0.5$. Further increasing the virtual fraction reduced performance; at the highest proportions, DE-Precision and PR-AUC fell below the no-augmentation baseline. This pattern suggests virtual populations enrich distributional supervision, whereas excessive reliance on synthetic data may bias learning away from real expression distributions, supporting a balanced mixture of real and virtual populations.

We also compared other six gene embeddings across four cell lines (Cui et al., 2024; Rosen et al., 2026; Du et al., 2019; Chen & Zou, 2023; Szklarczyk et al., 2025)(Appendix Table 2). Although rankings varied across individual metrics, the overall performance profiles were broadly similar.

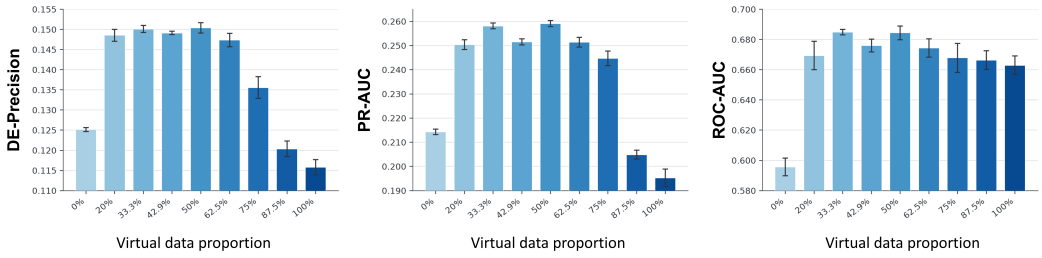


Figure 4: **Effect of population augmentation on differential-expression recovery.** DE-Precision, PR-AUC, and ROC-AUC are reported across different virtual data proportions, defined as $\rho_{\text{virtual}} = N_{\text{virtual}}/(N_{\text{real}} + N_{\text{virtual}})$. **Moderate augmentation yields the best overall performance, while excessive virtual data degrades performance.**

6 Discussion and Conclusion

Single-cell perturbation screens yield unpaired populations across batches, demanding a distribution-matching rather than cell-pairing paradigm (Yuan et al., 2026). Yet, measure-level modeling is bottlenecked by the coordinate-free nature of the RKHS. To resolve this, here we propose GeoPert. GeoPert parameterizes population distributions via information-geometric multilateration. Hierarchically decoupling population synthesis from cell realization unlocks metric-space data augmentation, facilitating robust zero-shot generalization to unseen perturbations.

Our work is not without limitations. First, our framework assumes that perturbation effects are sufficiently captured by off-the-shelf gene embeddings prior to generative modeling. Systematically learning task-specific, context-aware perturbation representations, or optimizing them end-to-end alongside the generative backbone, remains an important open challenge. Second, to maintain computational efficiency during cell-level realization (Stage 2), we implemented a variational decoder. While effective, our hierarchical formulation is fundamentally architecture-agnostic and natively compatible with foundation models, including Transformer-based decoders and Diffusion Transformers (DiTs). Comprehensively benchmarking these alternative architectures within our framework is a key priority for future work. Finally, the performance of multilateration signatures on atlas-scale cell collections warrants deeper investigation. We plan to further investigate it in our future work.

AI USE STATEMENT

We used generative AI tools GPT-5.6 sol for three tasks with recommended disclosure: (i) drafting portions of the algorithmic code, subsequently reviewed, tested, and in part rewritten by the authors to ensure correctness and conformity with our design; (ii) checking the manuscript for formatting and typographical issues, such as LaTeX layout, cross-references. (iii) validating if the mathematical proofs made by the authors are correct. We did not use generative AI tools to generate synthetic datasets or for any other mandatory-disclosure task beyond those described above; mandatory-disclosure tasks not relevant to this work were not applicable.

All AI-assisted output was reviewed by the authors, who take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made our code publicly available at `https://anonymous.4open.science/r/GeoPert-A36F`. All datasets used in this study are open source, and their associated publications are cited in the manuscript.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, Francis Chalissery, Rajesh Ilango, Vishvak Subramanyam, Chiara Ricci-Tam, Sanjay Nagaraj, Aidan Winters, Mingze Dong, Stefanie Fellinger, Adam Krejci, Tilmann Burckstummer, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, 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. *Cell*, 189(19):5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052.

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.

Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer Japan, Tokyo, 2016. doi: 10.1007/978-4-431-55978-8.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Yiqun Chen and James Zou. GenePT: A simple but effective foundation model for genes and cells built from ChatGPT. *bioRxiv*, 2023. doi: 10.1101/2023.10.16.562533. URL `https://www.biorxiv.org/content/10.1101/2023.10.16.562533`.

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.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038.

Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, 20(Suppl 1):82, 2019. doi: 10.1186/s12864-018-5370-x.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. doi: 10.1016/0304-3975(85)90224-5.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Robert E Kass and Paul W Vos. *Geometrical foundations of asymptotic inference*. John Wiley & Sons, 2011.

Gavin Kerrigan, Justin Ley, and Padhraic Smyth. Diffusion generative models in infinite dimensions. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9538–9563, 2023. URL https://proceedings.mlr.press/v206/kerrigan23a.html.

Leo Liberti, Carlile Lavor, Nelson Maculan, and Antonio Mucherino. Euclidean distance geometry and applications. *SIAM Review*, 56(1):3–69, 2014. doi: 10.1137/120875909.

Jae Hyun Lim, Nikola B. Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, Christopher Pal, Arash Vahdat, and Anima Anandkumar. Score-based diffusion models in function space. *Journal of Machine Learning Research*, 26(158):1–62, 2025. URL https://jmlr.org/papers/v26/23-1472.html.

Nian Liu, Xiaoxin He, Thomas Laurent, Francesco Di Giovanni, Michael M Bronstein, and Xavier Bresson. A general graph spectral wavelet convolution via chebyshev order decomposition. *arXiv preprint arXiv:2405.13806*, 2024.

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.

Russell Lyons. Strong negative type in spheres. *Pacific Journal of Mathematics*, 307(2):383–390, 2020.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

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

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.

Jakiw Pidstrigach, Youssef Marzouk, Sebastian Reich, and Sven Wang. Infinite-dimensional diffusion models. *Journal of Machine Learning Research*, 25(414):1–52, 2024. URL https://jmlr.org/papers/v25/23-1271.html.

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.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Stephen R. Quake, and Jure Leskovec. Universal cell embedding provides a foundation model for cell biology. *Nature*, 656 (8126):183–191, 2026. doi: 10.1038/s41586-026-10689-z.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013. doi: 10.1214/13-AOS1140.

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.

Damian Szklarczyk, Katerina Nastou, Mikaela Koutrouli, Rebecca Kirsch, Farrokh Mehryary, Radja Hachilif, Dewei Hu, Matteo E. Peluso, Qingyao Huang, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Research*, 53(D1):D730–D737, 2025. doi: 10.1093/nar/gkae1113.

F. William Townes, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry. Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model. *Genome Biology*, 20(1):295, 2019. doi: 10.1186/s13059-019-1861-6.

Xinyu Yuan, Xixian Liu, Ya Shi Zhang, Zuobai Zhang, Hongyu Guo, and Jian Tang. PerturbDiff: Functional diffusion for single-cell perturbation modeling. *arXiv preprint arXiv:2602.19685*, 2026. URL https://arxiv.org/abs/2602.19685.

A Information-Geometric Foundations

A.1 Fisher Information Metric for the Multinomial Distribution

Let θ denote the n-dimensional Cartesian coordinates of points on a simplex, where $|\theta|_1 = 1$. Let ω denote the square-root transformed Cartesian coordinates, in which the simplex is transformed into a sphere with $|\omega|_2 = 1$. For a point on the sphere, let the bases $\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_n}$ denote the orthogonal bases of the local tangent space. A vector $\mathbf{z}$ in the local tangent space, represented as $\mathbf{z} = \sum_{i=1}^{n} \beta_i \mathbf{e_i}$, can be converted to a representation in the ω space as $\mathbf{z} = \sum_{i=1}^{n} \alpha_i \omega_\mathbf{i}$ with chain rule, where $\alpha_i = \sum_{i=1}^{n} \beta_i \frac{\partial \omega_\mathbf{i}}{\partial \mathbf{e_i}}$. The inner product of a vector $\mathbf{z}$ in the ω coordinate system can thus be translated to the local tangent space as following:

$$|\mathbf{z}|_2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \beta_i \beta_j \frac{\partial \omega_\mathbf{k}}{\partial \mathbf{e_i}} \frac{\partial \omega_\mathbf{k}}{\partial \mathbf{e_j}} \tag{18}$$

The above equation can be perceived as a translation of L-2 norm between the local tangent space and the ω coordinate system, as $|\mathbf{z}|_2 = \mathbf{z}\top\mathbf{G}\mathbf{z}$, where $\mathbf{G}$ is the metric tensor defined as $G_{i,j} = \sum_{k=1}^{n} \frac{\partial \omega_\mathbf{k}}{\partial \mathbf{e_i}} \frac{\partial \omega_\mathbf{k}}{\partial \mathbf{e_j}}$. By substituting in $\omega_i = \sqrt{\theta_i}$ and applying chain rule in reverse ($\theta_i = \omega_i^2$), we have:

$$\mathbf{G}_{i,j} = \sum_{k=1}^{n} \frac{\partial \omega_\mathbf{k}}{\partial \mathbf{e_i}} \frac{\partial \omega_\mathbf{k}}{\partial \mathbf{e_j}} = \sum_{k=1}^{n} (\frac{1}{2\omega_\mathbf{k}} \frac{\partial \theta_\mathbf{k}}{\partial \omega_\mathbf{k}} \frac{\partial \omega_\mathbf{k}}{\partial \mathbf{e_i}})(\frac{1}{2\omega_\mathbf{k}} \frac{\partial \theta_\mathbf{k}}{\partial \omega_\mathbf{k}} \frac{\partial \omega_\mathbf{k}}{\partial \mathbf{e_j}}) = \frac{1}{4\theta_k} \sum_{k=1}^{n} \frac{\partial \theta_\mathbf{k}}{\partial \mathbf{e_i}} \frac{\partial \theta_\mathbf{k}}{\partial \mathbf{e_j}} \tag{19}$$

given that $\frac{\partial \theta_{\mathbf{k}}}{\partial \mathbf{e_i}} = \frac{\partial \theta_{\mathbf{k}}}{\partial \omega_{\mathbf{k}}} \frac{\partial \omega_{\mathbf{k}}}{\partial \mathbf{e_i}} = 2\omega_{\mathbf{k}} \frac{\partial \omega_{\mathbf{k}}}{\partial \mathbf{e_i}}$. Equation 19 establishes the metric tensor between a simplex and its square-root-transformed unit hypersphere, where the geodesic distance on the equivalent hypersphere manifold between two points θ_1 and θ_2 on the hyper-simplex can be computed as:

$$D_{\mathrm{FR}}(\theta_1, \theta_2) = arccos(\omega_1 \cdot \omega_2) \tag{20}$$

with $\theta_1 = (\omega_1)^2$ and $\theta_2 = (\omega_2)^2$.

For Multinomial distributions, the Fisher information matrix takes the same form as the metric tensor for the simplex-to-sphere pullback mapping. To see this, we use an alternative form of the Fisher information matrix

$$[I(\theta)]_{i,j} = -\mathbb{E}_{x \sim p(x;\theta)} \left[\left(\frac{\partial}{\partial \theta_i \partial \theta_j} \log p(x;\theta) \right) \middle| \theta \right] \tag{21}$$

We have:

$$\begin{aligned} \frac{\partial}{\partial p_i \partial p_j} \log p(x;\mathbf{p}) = \sum_{k=1}^{G} \frac{\partial}{\partial p_i} \left(\frac{\partial p_k}{\partial p_j} \cdot \left(\frac{\partial \log p(x;\mathbf{p})}{\partial p_k} \right) \right) = \sum_{k=1}^{G} \frac{\partial}{\partial p_i} \left(\frac{\partial p_k}{\partial p_j} \cdot \left(\frac{x_k}{p_k} \right) \right) \\ = \sum_{k=1}^{G} \frac{\partial}{\partial p_i \partial p_j} \left(p_k \frac{x_k}{p_k} \right) - \sum_{k=1}^{G} \frac{\partial p_k}{\partial p_i} \cdot \frac{\partial p_k}{\partial p_j} \cdot \left(\frac{x_k}{p_k^2} \right) \end{aligned} \tag{22}$$

Take 22 into 21, we have:

$$[I(\theta)]_{i,j} = -\sum_{k=1}^{G} \frac{\partial \mathbb{E}[x_k]}{\partial p_i \partial p_j} + \sum_{k=1}^{G} \frac{\partial p_k}{\partial p_i} \cdot \frac{\partial p_k}{\partial p_j} \cdot \left(\frac{\mathbb{E}[x_k]}{p_k^2} \right). \tag{23}$$

Notice that $\mathbb{E}[x_k] = N \cdot p_k$ with N being the total RNA molecule count of a cell. Thus Equation 23 becomes

$$[I(\theta)]_{i,j} = \frac{\partial}{\partial p_i \partial p_j} \sum_{k=1}^{G} (x_k) + \sum_{k=1}^{G} \frac{\partial p_k}{\partial p_i} \cdot \frac{\partial p_k}{\partial p_j} \cdot \left(\frac{N}{p_k} \right) = \frac{N}{p_k} \sum_{k=1}^{G} \frac{\partial p_k}{\partial p_i} \cdot \frac{\partial p_k}{\partial p_j}. \tag{24}$$

Notice that Equation 24 has the same form as Equation 19 but amplified by a factor of $4N$. This establishes the geometric equivalence between information distance and arc distance: The Fisher-Rao distance between two Multinomial distributions is a constant multiple of the arc distance between ω_i and ω_j on the hypersphere, after taking square-root transformation: $\omega_i^2 = \mathbf{p}_i$ and $\omega_j^2 = \mathbf{p}_j$.

A.2 Chentsov's Theorem and Information-Preserving Transformations

Statistical manifolds. Let $\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$ be a parametric family of probability distributions, where Θ is a smooth parameter space. The family $\mathcal{P}$ forms a statistical manifold, on which one may define Riemannian metrics to measure infinitesimal distances between distributions.

Chentsov's Theorem. *Up to a positive constant scaling, the Fisher information metric is the unique Riemannian metric on a statistical manifold that is invariant under sufficient-statistics transformations (equivalently, under probabilistically meaningful mappings such as aggregation, coarse-graining, or reparameterization of outcomes).*

B Theoretical Guarantees and Proofs

Throughout, $G = |\mathcal{G}| \geq 2$, $\mathcal{X} = \Delta^{G-1}$, and P, Q are Borel probability measures on $\mathcal{X}$. We use the unit-sphere convention $d_{\mathrm{FR}}(x,y) = \arccos\langle \sqrt{x}, \sqrt{y} \rangle$. Thus $0 \leq d_{\mathrm{FR}} \leq \pi/2$; a different positive constant scaling leaves the conclusions unchanged.

B.1 Fisher-Rao Kernel Validity and Reference Invariance

Claim. For any $p_0 \in \mathcal{X}$, the kernel

$$k_{p_0}(x,y) = \frac{1}{2} \left[d_{\mathrm{FR}}(x,p_0) + d_{\mathrm{FR}}(y,p_0) - d_{\mathrm{FR}}(x,y) \right] \tag{25}$$

is positive semidefinite, and its squared MMD is independent of p_0. This is a distance-induced kernel, not an exponential kernel (Sejdinovic et al., 2013).

Proof. Let σ be uniform probability measure on S^{G-1} and set $b_\omega(x) = \mathbf{1}\{\langle \omega, \sqrt{x}\rangle > 0\}$. A uniformly oriented hyperplane separates two unit vectors with probability equal to their angle divided by π. Hence

$$d_{\mathrm{FR}}(x, y) = \pi \int \big(b_\omega(x) - b_\omega(y)\big)^2 \, d\sigma(\omega). \tag{26}$$

For coefficients satisfying $\sum_i c_i = 0$, expansion gives

$$\sum_{i,j} c_i c_j d_{\mathrm{FR}}(x_i, x_j) = -2\pi \int \left(\sum_i c_i b_\omega(x_i)\right)^2 d\sigma(\omega) \le 0, \tag{27}$$

establishing conditional negative definiteness. Substitution into equation 25 yields

$$k_{p_0}(x, y) = \pi \int [b_\omega(x) - b_\omega(p_0)][b_\omega(y) - b_\omega(p_0)] \, d\sigma(\omega), \tag{28}$$

which is an inner-product kernel.

For independent $X, X' \sim P$ and $Y, Y' \sim Q$, expanding MMD cancels every reference-dependent term:

$$\begin{aligned} \mathrm{MMD}^2_{k_{p_0}}(P, Q) &= \mathbb{E} d_{\mathrm{FR}}(X, Y) - \frac{1}{2}\mathbb{E} d_{\mathrm{FR}}(X, X') - \frac{1}{2}\mathbb{E} d_{\mathrm{FR}}(Y, Y') \\ &= \frac{1}{2}\mathcal{E}_{\mathrm{FR}}(P, Q). \end{aligned} \tag{29}$$

All expectations exist because the distance is bounded. Thus the kernel reference is immaterial to MMD and is distinct from the anchor populations defining T_M.

B.2 Characteristicness on the Probability Simplex

Claim. The kernel in equation 25 is characteristic on $\mathcal{X}$: its MMD vanishes if and only if $P = Q$.

Proof. The square-root image $B = \sqrt{\mathcal{X}} \subset S^{G-1}$ lies in an open hemisphere. Indeed, for $a = G^{-1/2}(1, \dots, 1)$ and $u \in B$, $\langle a, u\rangle \ge G^{-1/2} > 0$. By the strong-negative-type theorem for open hemispheres (Lyons, 2020, Theorem 1.2), for probabilities R, S supported on B,

$$\iint \arccos\langle u, v\rangle \, d(R - S)(u) \, d(R - S)(v) = 0 \quad \Longleftrightarrow \quad R = S. \tag{30}$$

Apply this to $R = (\sqrt{\cdot})_{\#}P$ and $S = (\sqrt{\cdot})_{\#}Q$. By equation 29, zero MMD implies $R = S$. The square-root map is invertible on B, with inverse $u \mapsto u \odot u$, so $P = Q$. The converse is immediate.

Fixing p_0, identify $k_{\mathrm{FR}} = k_{p_0}$, write $\mathcal{H} = \mathcal{H}_{k_{\mathrm{FR}}}$, and define $d(P, Q) = \mathrm{MMD}_{k_{p_0}}(P, Q)$ as a metric on population distributions. Negative type alone establishes kernel validity; strong negative type establishes characteristicness (Sejdinovic et al., 2013). The restriction to the simplex is important: the entire sphere does not have strong negative type.

B.3 Dense-Anchor Isometry and Injectivity

Let $(\mathcal{S}, d)$ be a compact family of population distributions, and let $\{A_j\}_{j \ge 1}$ be dense in $\mathcal{S}$. Define

$$T_\infty(P) = \big(d(P, A_j)\big)_{j \ge 1} \in \ell^\infty. \tag{31}$$

Then

$$\|T_\infty(P) - T_\infty(Q)\|_\infty = d(P, Q), \qquad P, Q \in \mathcal{S}. \tag{32}$$

Proof. The reverse triangle inequality gives $|d(P, A_j) - d(Q, A_j)| \le d(P, Q)$ for every j. By density, choose $A_{j_k} \to P$. Then $d(P, A_{j_k}) \to 0$ and $d(Q, A_{j_k}) \to d(Q, P)$. The supremum of coordinate differences therefore equals $d(P, Q)$. Thus T_∞ is isometric and injective, giving a bijection between $\mathcal{S}$ and its image, not all of ℓ^∞.

B.4 Proof of the Finite-Coverage Bound

For $M \geq 1$ and $\mathcal{A}_M = \{A_1, \ldots, A_M\} \subseteq \mathcal{S}$, define

$$T_M(P) = \big(d(P, A_j)\big)_{j=1}^{M}, \qquad \varepsilon_M = \sup_{P \in \mathcal{S}} \min_{1 \leq j \leq M} d(P, A_j). \tag{33}$$

The bound stated in the main text is

$$0 \leq d(P, Q) - \|T_M(P) - T_M(Q)\|_\infty \leq 2\varepsilon_M. \tag{34}$$

Proof. The reverse triangle inequality proves the left inequality. Choose $A_{j^\star}$ nearest to P. Then

$$\begin{aligned} d(P, Q) &\leq d(P, A_{j^\star}) + d(Q, A_{j^\star}) \\ &\leq 2d(P, A_{j^\star}) + |d(Q, A_{j^\star}) - d(P, A_{j^\star})| \\ &\leq 2\varepsilon_M + \|T_M(P) - T_M(Q)\|_\infty. \end{aligned} \tag{35}$$

The argument only needs a finite covering radius; compactness additionally ensures finite covers at arbitrary resolutions. For nested anchors with dense union, compactness gives $\varepsilon_M \to 0$, and equation 34 yields uniform convergence of represented pairwise distances.

C Evaluation Metrics

C.1 Average-Expression Accuracy

For each perturbation, the coefficient of determination across genes is

$$R^2_\tau = 1 - \frac{\sum_{g \in \mathcal{G}_{\text{eval}}} (\widehat{a}_{\tau,g} - a_{\tau,g})^2}{\sum_{g \in \mathcal{G}_{\text{eval}}} (a_{\tau,g} - \overline{a}_\tau)^2}, \qquad \overline{a}_\tau = \frac{1}{|\mathcal{G}_{\text{eval}}|} \sum_g a_{\tau,g}. \tag{36}$$

Higher values indicate better mean-expression prediction, with 1 denoting exact agreement.

C.2 Perturbation Identity and Population Structure

Perturbation discrimination score. For prediction τ, rank all observed perturbation effects by $\ell(\widehat{\Delta}_\tau, \Delta_\nu)$, $\nu \in \mathcal{E}$. Let $r_\tau \in \{0, \ldots, K-1\}$ be the zero-based rank of the correct target under the evaluator's tie ordering. The score is

$$\text{PDS}_\ell = \frac{1}{K} \sum_{\tau \in \mathcal{E}} \left(1 - \frac{r_\tau}{K}\right). \tag{37}$$

We distinguish ℓ_1, ℓ_2, and cosine-distance variants. A higher score indicates that predicted mean effects better identify their perturbations.

Clustering alignment. Let c^{obs} be Leiden labels of observed centroids at resolution γ_0, and c^{gen}_γ labels of predicted centroids at resolution γ.

$$\text{Clust-Align} = \max\left\{0, \max_{\gamma \in \Gamma} \text{AMI}(c^{\text{obs}}, c^{\text{gen}}_\gamma)\right\}. \tag{38}$$

C.3 Differential-Expression Recovery

ROC-AUC. Let $\mathcal{D}_\tau = \{g : y_{\tau,g} = 1\}$ and $\mathcal{N}_\tau = \mathcal{G}_{\text{eval}} \setminus \mathcal{D}_\tau$. Then

$$\text{ROC-AUC}_\tau = \frac{\sum_{g \in \mathcal{D}_\tau} \sum_{h \in \mathcal{N}_\tau} \left[\mathbf{1}\{s_{\tau,g} > s_{\tau,h}\} + \frac{1}{2}\mathbf{1}\{s_{\tau,g} = s_{\tau,h}\}\right]}{|\mathcal{D}_\tau|\,|\mathcal{N}_\tau|}. \tag{39}$$

It measures how often a true responsive gene outranks a nonresponsive one.

PR-AUC. Let Prec_j and Rec_j denote precision and recall, with $\text{Rec}_0 = 0$. Then

$$\text{PR-AUC}_\tau = \sum_j (\text{Rec}_j - \text{Rec}_{j-1})\,\text{Prec}_j\,. \tag{40}$$

DE-Precision. Let $\widehat{\mathcal{D}}_\tau$ be the predicted significant set and $k_\tau = |\widehat{\mathcal{D}}_\tau|$. If $\mathrm{Top}_k(S)$ retains the first k ranked genes of S (or all available genes when fewer than k), then

$$\text{DE-Precision}_\tau = \frac{\left|\mathrm{Top}_{k_\tau}(\mathcal{D}_\tau) \cap \mathrm{Top}_{k_\tau}(\widehat{\mathcal{D}}_\tau)\right|}{k_\tau}, \qquad k_\tau > 0. \tag{41}$$

C.4 Population-Signature Fidelity

For each held-out perturbation, let $\{z^{\mathrm{obs}}_{\tau,r}\}_{r=1}^{R}$ and $\{z^{\mathrm{gen}}_{\tau,s}\}_{s=1}^{S}$ be observed and predicted population signatures evaluated using the same anchors and scaling. Let $b^{\mathrm{obs}}_\tau, b^{\mathrm{gen}}_\tau$ and $C^{\mathrm{obs}}_\tau, C^{\mathrm{gen}}_\tau$ be their sample means and covariance matrices (each using the $R-1$ or $S-1$ denominator). The Gaussian Fréchet discrepancy is

$$\begin{aligned} \mathrm{FD}_\tau &= \|b^{\mathrm{obs}}_\tau - b^{\mathrm{gen}}_\tau\|_2^2 + \mathrm{FD}_{\mathrm{cov},\tau}, \\ \mathrm{FD}_{\mathrm{cov},\tau} &= \mathrm{tr}\left[C^{\mathrm{obs}}_\tau + C^{\mathrm{gen}}_\tau - 2\left((C^{\mathrm{obs}}_\tau)^{1/2} C^{\mathrm{gen}}_\tau (C^{\mathrm{obs}}_\tau)^{1/2}\right)^{1/2}\right]. \end{aligned} \tag{42}$$

C.5 Anchor-Representation Fidelity

On a fixed collection of populations excluded from anchor selection, compare original distances $D_{ij} = d(P_i, P_j)$ with $\Delta^{(M)}_{ij} = \|T_M(P_i) - T_M(P_j)\|_\infty$. Distance agreement and normalized contraction are

$$\begin{aligned} \rho_M &= \mathrm{Corr}_{i<j}(D_{ij}, \Delta^{(M)}_{ij}), \\ E_M &= \frac{\sum_{i<j}(D_{ij} - \Delta^{(M)}_{ij})}{\sum_{i<j} D_{ij}}, \end{aligned} \tag{43}$$

D Implementation and Training Details

D.1 Cell-Group Construction and Sampling

By default, we group cells within each perturbation τ using dataset-provided experimental batch labels. If these labels are unavailable or all batch-specific groups contain fewer than $n_{\min}$ cells, we instead randomly permute the N_τ cells at each epoch and partition them into $K \le N_\tau$ disjoint groups of sizes $\lfloor N_\tau/K \rfloor$ or $\lceil N_\tau/K \rceil$. All cells in each retained population are used; populations with fewer than two cells are excluded.

Let R_τ denote the number of annotated batch groups, or K under random partitioning. Each group defines the empirical distribution

$$P^{(e)}_{\tau,r} = \frac{1}{|\mathcal{B}^{(e)}_{\tau,r}|} \sum_{x \in \mathcal{B}^{(e)}_{\tau,r}} \delta_x, \qquad r = 1, \ldots, R_\tau. \tag{44}$$

Batch-defined memberships remain fixed, whereas random partitions are regenerated each epoch.

D.2 Model Architectures and Gene Embeddings

GeoPert comprises two generative stages: a conditional diffusion model predicts population signatures, and a stochastic decoder maps each signature to a cell-expression distribution.

Population-level diffusion. The denoiser is a residual MLP with four blocks and hidden width 256. It receives a noisy anchor-distance vector $z_t \in \mathbb{R}^M$, a timestep embedding, and a projected perturbation representation h_τ. These inputs are mapped into a shared hidden space and combined to predict the velocity $v_\theta(z_t, t, h_\tau) \in \mathbb{R}^M$. Reverse diffusion transforms Gaussian noise into a population signature z conditioned on the perturbation.

Cell-level generation. The decoder is an MLP with two hidden layers of width 256. It concatenates z with an independent latent vector $\xi_i \sim \mathcal{N}(0, I_{16})$ and generates $\widehat{x}_i = G_\phi(z, \xi_i) \in \Delta^{G-1}$, a

nonnegative, unit-sum expression profile over G genes. Cells within a generated group share z but receive independent latent samples. The decoder therefore models within-group variation without directly receiving the perturbation embedding and is trained using Fisher-Rao MMD.

Gene embeddings. Unless otherwise stated, we use pretrained gene embeddings from *KG-WaveGC* Liu et al. (2024) as perturbation descriptors. The embedding $e_\tau \in \mathbb{R}^{d_e}$ remains frozen, while a trainable linear projection adapts it to the diffusion condition:

$$h_\tau = f_\psi(e_\tau) = W_\psi e_\tau + b_\psi. \tag{45}$$

The projection is optimized with the denoiser, separating task-specific conditioning from external gene-representation learning.

D.3 Training Schedule

The two-stage generative architecture is fitted through four optimization phases: unconditional pretraining, conditional adaptation, decoder training, and joint fine-tuning.

Component training. Unconditional pretraining uses batches of at most 256 real or virtual group signatures and a 100-step cosine diffusion schedule. Conditional adaptation initializes the backbone from the selected pretraining checkpoint and optimizes velocity prediction on real groups. Timesteps are stratified across the noise schedule. Perturbation conditioning is dropped with probability 0.5 using one Bernoulli decision for the entire update batch. An exponential moving average (EMA) with decay 0.99 provides conditional checkpoints. The decoder is trained separately on real and virtual signature-population pairs. Its loss is the Fisher-Rao energy statistic MMD.

Joint fine-tuning. The independently selected components are jointly optimized using real training perturbations only, with at most 16 perturbations per update. The active objective is

$$\mathcal{L}_{\text{joint}} = 0.1\mathcal{L}_{\text{diff}} + 2\mathcal{E}_{\text{FR}}\big(\widehat{P}_{\text{gen,pool}}, \widehat{P}_{\text{real,pool}}\big), \tag{46}$$

averaged over perturbations.

D.4 Hyperparameter Settings

We use 100 anchors and 2,000 expression genes per dataset. Unconditional pretraining uses mini-batches of 256 signatures, while conditional training processes up to 128 perturbations per updates. Decoder updates contain up to eight source populations, with 32 observed cells sampled with replacement and 32 cells generated per group. Joint fine-tuning uses 16 perturbations per update, generating four groups of 16 cells per perturbation.

Using AdamW, unconditional pretraining runs for 200 epochs at a learning rate of 10^{-3}, conditional training for 400 epochs at 2×10^{-4}, and decoder training for 200 epochs with an initial learning rate of 10^{-3}. Joint fine-tuning runs for 120 epochs, with learning rates of 5×10^{-5} for diffusion and 10^{-4} for the decoder. Checkpoints are selected using validation data.

Synthetic-population weights. For each randomly selected anchor pair, we uniformly choose one anchor as dominant and sample $u \sim \text{Beta}(1, \alpha_{\text{mix}})$, setting $w = 0.5 + 0.5u$ with default $\alpha_{\text{mix}} = 0.1$. Draws with $w > 0.99$ are rejected and resampled. The dominant and remaining anchors receive weights w and $1 - w$, respectively, shared across all cells in the synthetic population.

Virtual-data budgets. We generate 3,000 synthetic populations of 64 cells each for unconditional pretraining. Decoder training generate data with a default virtual fraction $\rho_{\text{virtual}} = 0.5$, yielding equal numbers of real and virtual source-population visits per epoch.

Batch-size threshold. We set $n_{\min} = 30$ cells per perturbation-batch group. If batch labels are unavailable or all groups within a perturbation fall below this threshold, we use epoch-wise random partitioning.

E Additional Results

E.1 Anchor Coverage and Group Size

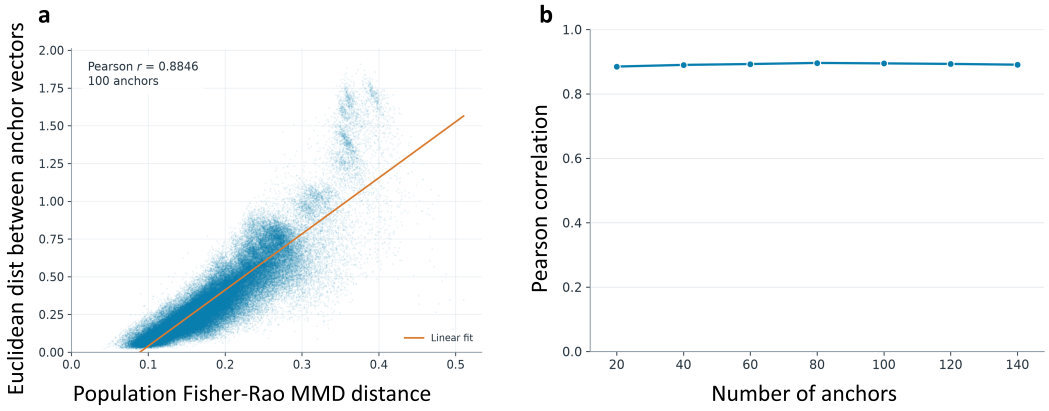


Figure 5: **Preservation of population geometry by anchor distance representations.**(a) Population Fisher–Rao MMD versus Euclidean distance between 100-anchor representations for K562 test perturbations. Each point represents a population pair; the orange line indicates a linear fit (Pearson r = 0.8846). (b) Corresponding Pearson correlations using 20–140 anchors. MMD denotes the square root of the Fisher–Rao energy statistic.

E.2 GENE EMBEDDING COMPARISON

Table 2: Gene embedding benchmark across four datasets. Higher is better for all metrics.

RPE1							
Gene embedding	**PR-AUC ↑**	**PDS-L1 ↑**	**PDS-L2 ↑**	**PDS-Cos ↑**	R^2 ↑	**DE-Prec. ↑**	**ROC-AUC ↑**
scGPT(Cui et al., 2024)	0.3368	0.5679	0.5730	0.5385	0.9971	0.2286	0.6971
UCE(Rosen et al., 2026)	0.3210	0.5918	0.5989	0.5892	0.9976	0.2195	0.6778
Gene2Vec(Du et al., 2019)	0.3505	0.5614	0.5669	0.5362	0.9973	**0.2320**	0.7039
GenePT(Chen & Zou, 2023)	0.3474	0.6039	0.6110	**0.6004**	**0.9981**	0.2296	0.7054
STRING-Slim(Szklarczyk et al., 2025)	**0.3560**	0.6368	0.6458	0.5876	0.9960	0.2310	0.7145
STRING-512(Szklarczyk et al., 2025)	0.3514	**0.6419**	**0.6509**	0.5976	0.9950	0.2313	**0.7152**
K562							
Gene embedding	**PR-AUC ↑**	**PDS-L1 ↑**	**PDS-L2 ↑**	**PDS-Cos ↑**	R^2 ↑	**DE-Prec. ↑**	**ROC-AUC ↑**
scGPT	0.2109	0.5640	0.5675	0.6183	**0.9983**	0.1279	0.6327
UCE	0.2289	0.5655	0.5689	0.6747	0.9976	0.1267	0.6398
Gene2Vec	0.2097	0.5404	0.5438	0.6085	0.9981	0.1157	0.6161
GenePT	0.2504	0.5901	0.5951	0.6817	0.9980	0.1439	0.6587
STRING-Slim	**0.2651**	0.6230	0.6297	0.7203	0.9981	**0.1485**	0.6780
STRING-512	0.2599	**0.6411**	**0.6489**	**0.7313**	**0.9983**	0.1458	**0.6820**
HEPG2							
Gene embedding	**PR-AUC ↑**	**PDS-L1 ↑**	**PDS-L2 ↑**	**PDS-Cos ↑**	R^2 ↑	**DE-Prec. ↑**	**ROC-AUC ↑**
scGPT	0.2795	0.5811	0.5817	0.5634	0.9961	0.1816	0.6556
UCE	0.2760	0.5989	0.5999	0.6142	0.9961	0.1814	0.6414
Gene2Vec	0.2712	0.5718	0.5723	0.5645	0.9968	0.1733	0.6416
GenePT	0.2879	0.5999	0.6043	0.5937	0.9972	0.1851	0.6640
STRING-Slim	**0.2992**	**0.6820**	**0.6839**	0.6350	**0.9979**	**0.1889**	**0.6843**
STRING-512	0.2963	0.6648	0.6688	**0.6486**	0.9978	0.1886	0.6779
JURKAT							
Gene embedding	**PR-AUC ↑**	**PDS-L1 ↑**	**PDS-L2 ↑**	**PDS-Cos ↑**	R^2 ↑	**DE-Prec. ↑**	**ROC-AUC ↑**
scGPT	0.1844	0.5424	0.5439	0.5726	0.9979	0.1166	0.6612
UCE	0.2038	0.5553	0.5610	0.6132	0.9979	0.1285	0.6603
Gene2Vec	0.2006	0.5567	0.5590	0.5683	**0.9983**	0.1274	0.6660
GenePT	0.2238	0.5865	0.5943	0.6427	0.9980	0.1342	0.6874
STRING-Slim	**0.2468**	**0.6329**	**0.6438**	0.6647	**0.9983**	0.1406	0.7047
STRING-512	0.2440	0.6155	0.6248	**0.6784**	0.9981	**0.1409**	**0.7097**