ROTATION OPERATORS FOR SINGLE-CELL PERTURBATION RESPONSE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting how cells respond to genetic and chemical perturbations is becoming central to mechanistic biology and drug discovery. However, the space of possible combinations grows quadratically, and identifying which pairs interact non-additively before measuring them remains an open problem. To address this, we introduce OpPert, a new deep learning framework that models each perturbation as a norm-preserving rotation of the cell's latent representation. To model combinations of two perturbations, their rotations compose into a pair prediction via Baker-Campbell-Hausdorff, and the resulting Lie bracket gives their interaction strength. This interaction score is computed from single-perturbation operators alone, without supervision from the candidate pair, and we show it correlates with measured non-additivity. To infer effects of unseen perturbations, a flow-matching model generates their rotations, with sampling variance serving as a confidence estimate. We validate OpPert computationally on known benchmarks and experimentally in the wet lab. Across held-out evaluations on three genetic screens (Norman, K562, RPE1), OpPert achieves 81–88% directional accuracy, above existing approaches. We further evaluate OpPert on a small-molecule screen and validate predicted Belinostat targets by RT-qPCR in MCF7 cells, elucidating apoptotic regulators never previously associated with this drug.

1 INTRODUCTION

Many of the most important effects in biology and medicine arise from combinations of perturbations rather than from any single intervention. Drug synergies, genetic epistasis, and gene-gene regulatory interactions all depend on how two perturbations interact when applied together. High-throughput screens now profile thousands of individual genetic and chemical perturbations at single-cell resolution (Replogle et al., 2022; Srivatsan et al., 2020), but the number of pairwise combinations grows as the square of the library size and cannot be measured exhaustively. A predictive model that identifies which pairs interact non-additively before they are tested would let experimentalists focus on the combinations most likely to reveal new biology. Fig. 1a shows that this need is concrete, not hypothetical, on the Norman two-gene activation screen (Norman et al., 2019), where held-out gene pairs deviate substantially from additive predictions.

Existing prediction methods model each perturbation as an additive shift in a learned latent space (Lotfollahi et al., 2019; 2023; Roohani et al., 2024). These models can predict individual perturbation responses, but composing two perturbations by summing their effect vectors assumes the pair acts independently and cannot capture non-additive interactions. No current method provides an interaction score that can be computed from single-perturbation information alone (Ahlmann-Eltze et al., 2025; Wei et al., 2026).

We propose OpPert, which replaces the additive shift with a learned rotation of the cell's latent representation. OpPert is the first attempt to introduce rotations that preserve the norm of the latent state, preventing the decoder from extrapolating beyond its training distribution. More importantly, rotations carry an algebraic structure that additive shifts lack. When two rotations are combined through the Baker–Campbell–Hausdorff (BCH) formula, the result decomposes into an additive term plus a Lie bracket that quantifies how strongly the two perturbations interact. This interaction score is computed from single-perturbation operators alone, without observing the pair, and correlates with measured non-additivity (Fig. 1b).

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.

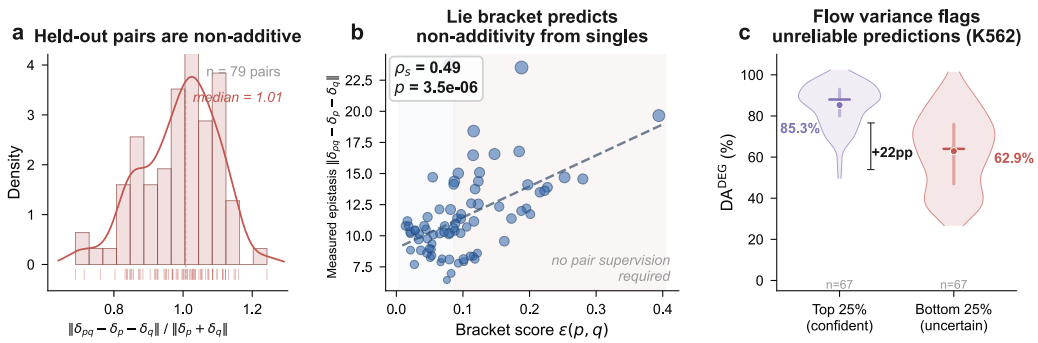

Figure 1. Three gaps in perturbation prediction. **(a)** Held-out gene pairs are substantially non-additive, with a median deviation ratio of 1.01 and wide spread (79 pairs from the Norman screen). **(b)** The Lie bracket score $\varepsilon(p, q)$, computed from single-perturbation operators alone, correlates with measured non-additivity (Spearman $\rho_s = 0.49$, $p = 3.5 \times 10^{-6}$). No pair supervision is required. **(c)** Flow sampling variance separates reliable from unreliable predictions on K562 without a dedicated uncertainty module, with the top-25% most confident predictions reaching 85.3% directional accuracy versus 62.9% for the bottom 25% ($\Delta = 22$ pp).

For perturbations not seen during training, a conditional flow-matching model maps biological descriptors to rotation parameters. Drawing multiple samples from this flow provides a built-in confidence signal that separates reliable from unreliable predictions by 22 percentage points (Fig. 1c) without a dedicated uncertainty module.

We make the following contributions.

1. *Perturbations as rotations.* Modeling perturbations as norm-preserving rotations rather than additive shifts bounds the decoder input to the training distribution and enables algebraic composition.
2. *Algebraic composition with a free interaction diagnostic.* Two rotations compose through Baker–Campbell–Hausdorff. The additive component recovers prior linear models as a special case, and the Lie bracket provides a scalar interaction score computed from singles alone. This score correlates with measured non-additivity (Spearman $\rho = 0.49$, $p = 3.5 \times 10^{-6}$ across 79 held-out pairs).
3. *Flow-based generalization with built-in uncertainty.* A conditional flow on the rotation algebra maps biological descriptors to rotation generators for unseen perturbations. The sampling variance of the flow yields a built-in confidence signal that ranks predictions by reliability without additional parameters or training.
4. *In-vitro experimental validation.* We validate predictions experimentally by RT-qPCR in MCF7 cells treated with Belinostat, confirming predicted gene targets that include apoptotic regulators not previously associated with this drug.

2 Related Work

Perturbation prediction. Current methods differ mainly in how they represent and compose perturbation effects. For instance, scGen (Lotfollahi et al., 2019) and CPA (Lotfollahi et al., 2023) treat perturbations as additive latent vectors, composing pairs by summation. GEARS (Roohani et al., 2024) and AttentionPert (Bai et al., 2024) use graph- or attention-based aggregation over gene-level embeddings. CellFlow (Klein et al., 2025) learns transport maps between control and perturbed distributions. PRESCRIBE (Cheng et al., 2025) introduces calibrated uncertainty via a Natural Posterior Network but retains additive composition with a black-box MLP correction. In all cases, the pair operation is either commutative (sum, mean pooling) or implicit in a trained aggregator. No existing method exposes a non-commutative composition rule whose failure to commute can be read as a pair interaction score.

Benchmarks and evaluation. Several independent studies show that no existing method dominates across datasets, splits, and metrics (Ahlmann-Eltze et al., 2025; Wei et al., 2026; Viñas Torné et al.,

2025; Miller et al., 2025). Ahlmann-Eltze et al. (2025) find that deep learning does not consistently beat linear baselines. Miller et al. (2025) show that under well-calibrated evaluation it does, locating much of the apparent failure in metric choice. The aggregate signal is that inductive bias and evaluation protocol matter more than model scale.

Lie groups in machine learning. Lie groups have been used for equivariant architectures (Bronstein et al., 2021; Hutchinson et al., 2021), rotation parameterization (Helfrich et al., 2018; Zhou et al., 2019; Chen & Lipman, 2024), and disentangled representation learning with commutative groups (Zhu et al., 2021). Lin et al. (2024) introduced Lie bracket layers as neural network primitives. OpPert applies these ideas to perturbation biology, where the non-commutativity of the group product carries direct meaning as a gene-gene interaction measure. An extended discussion of related work is in Appendix A.

3 METHODS

OpPert models each perturbation as a rotation of the learned latent representation, predicts unseen rotations from biological descriptors with flow matching, and composes rotations through BCH. This section defines the perturbation operator, the bracket score used for interaction ranking, and the flow model used for unseen perturbations. Proofs and training details are in Appendices B and C.6.

3.1 COMPOSABLE PERTURBATION PREDICTION

We consider a single-cell perturbation dataset $\mathcal{D} = \{(\boldsymbol{x}_i, p_i, e_{p_i}, \boldsymbol{c}_i)\}_i$, where $\boldsymbol{x}_i \in \mathbb{R}^G$ is the expression of cell i across G genes, p_i is the applied perturbation (a single gene target for genetic screens or a chemical compound for drug screens), $e_{p_i} \in \mathbb{R}^{d_e}$ is its biological descriptor (gene-name and pathway views for genes, molecular fingerprints and physicochemical features for chemicals), and $\boldsymbol{c}_i$ collects per-cell covariates such as cell type, batch, or dosage. The training data cover control cells and cells from a fixed set of observed singles $\mathcal{P}_{\text{train}}$, with cells from observed combinations of two training singles available when a combination screen has been run. The test data contain unseen singles $p^\star \notin \mathcal{P}_{\text{train}}$ and unseen combinations whose joint responses never enter fitting.

An encoder maps expression to a latent $\boldsymbol{z} = f_\phi(\boldsymbol{x}_i)$ that captures the cell's covariate state through its transcriptional profile (Lotfollahi et al., 2019; 2023; Roohani et al., 2024). Because the rotation acts per-cell, norm preservation holds regardless of the covariate state encoded in $\boldsymbol{z}$.

For an unseen combination (p, q), we assemble the prediction from the two single-perturbation predictions through a composition rule fixed before the pair is observed. Additive predictors fit this template with sum as the composition, and our model picks a non-abelian rule that recovers sum as a special case when the two perturbations commute.

3.2 OPERATOR

An encoder f_ϕ maps each cell to a latent $\boldsymbol{z} \in \mathbb{R}^{128}$, and a decoder g_ω maps the latent back to gene expression. A perturbation acts as a block-diagonal rotation of $\boldsymbol{z}$ before decoding. We choose a rotation because rotations exactly preserve the latent norm, so a predicted perturbation cannot move the latent to a magnitude that the decoder never saw during training, which keeps the class of extrapolations available to the decoder under control. We choose the block-diagonal subgroup $\mathrm{SO}(4)^{32}$ rather than a dense rotation in $\mathrm{SO}(128)$ because a dense rotation would carry $8{,}128$ parameters per perturbation, far more than the 70 to $1{,}500$ training perturbations available in current screens.

For each block $k \in \{1, \ldots, 32\}$, a learned 6-dimensional vector $\theta_p^{(k)} \in \mathbb{R}^6$ parameterizes a 4×4 skew-symmetric matrix by placing its six entries on the independent antisymmetric positions of that matrix, a non-trainable linear map we denote by $\beta : \mathbb{R}^6 \to \mathfrak{so}(4)$. Concatenating across the 32 blocks gives the per-perturbation generator $\theta_p \in \mathbb{R}^{192}$, and the operator is

$$R_p \;=\; \mathrm{diag}\big(\exp \beta(\theta_p^{(1)}),\, \ldots,\, \exp \beta(\theta_p^{(32)})\big) \;\in\; \mathrm{SO}(4)^{32}. \tag{1}$$

For a control cell $\boldsymbol{x}_{\text{ctrl}}$ with latent $\boldsymbol{z} = f_\phi(\boldsymbol{x}_{\text{ctrl}})$, the predicted post-perturbation cell is the decoded image of the rotated latent,

$$\hat{\boldsymbol{x}}_p \;=\; g_\omega\big(R_p\, \boldsymbol{z}\big), \tag{2}$$

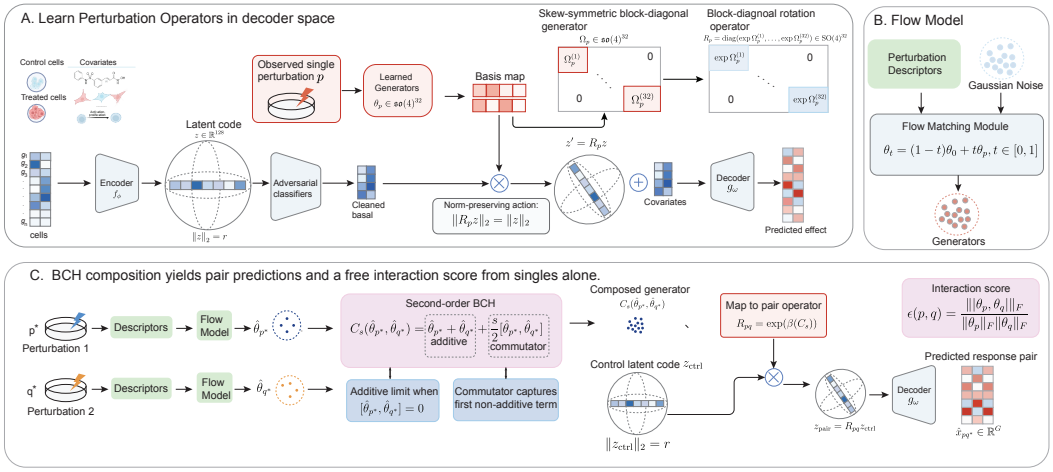

Figure 2. OpPert architecture. (A) Each perturbation is parameterized as a block-diagonal rotation $R_p \in \mathrm{SO}(4)^{32}$ of the cell's latent code. The rotation preserves $\|z\|_2$ exactly, bounding decoded predictions to the training distribution. **(B)** For unseen perturbations, a conditional flow-matching model maps biological descriptors to rotation generators on the Lie algebra, initialized from kernel ridge regression. **(C)** Two unseen perturbations compose through second-order BCH. The additive term recovers prior linear models and the Lie bracket provides a free interaction score $\varepsilon(p, q)$.

which is the analog of the additive update $\boldsymbol{z}' = \boldsymbol{z} + \boldsymbol{\delta}_p$ used by prior latent-shift predictors (Lotfollahi et al., 2019; 2023). Each block exponential admits a closed form in constant time independent of latent dimension, and applying R_p to a latent costs only linear time in the latent dimension (Appendix B.4). Because every operator in $\mathrm{SO}(4)^{32}$ is orthogonal, any composition $R_{p_m} \cdots R_{p_1}$ is also orthogonal, so the latent norm is preserved by every single perturbation and by every chain of perturbations (Lemma 3).

3.3 Composition by Baker–Campbell–Hausdorff

Predicting the response to an unmeasured combination requires a rule that maps two single-perturbation operators into a pair prediction. Existing composition rules are either commutative (summation, mean pooling) or implicit in a trained aggregator, and neither produces an explicit interaction score from singles alone. We take the pair action to be the matrix product $R_p R_q$, applied as $\hat{\boldsymbol{x}}_{pq} = g_\omega(R_p R_q \boldsymbol{z})$, and use the Baker–Campbell–Hausdorff (BCH) formula to write its generator θ_{pq} as the sum $\theta_p + \theta_q$ plus an infinite series of nested Lie brackets that collapses to $\theta_p + \theta_q$ exactly when the two generators commute (Theorem 5). The series converges when per-block generator norms satisfy $\|\Omega_p^{(k)}\|_F + \|\Omega_q^{(k)}\|_F < \log 2$ (Theorem 5). In practice, learned generators stay well within this regime, with a median eigenangle of 0.114 rad and 92% of per-block eigenangles below $\pi/8$ (Section 4.3). We keep only the leading non-trivial term,

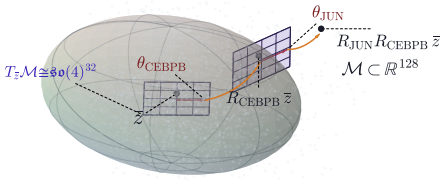

Figure 3. Learned operator for an unseen perturbation. The encoded training cells form an ellipsoid in latent space, and the generator predicted from biological descriptors is overlaid as the rotation that takes this ellipsoid into the perturbed state.

$$\theta_{pq} \;=\; \theta_p + \theta_q + \tfrac{s}{2}\,[\theta_p, \theta_q] \;+\; \mathcal{O}(\|\theta\|_F^3), \tag{3}$$

where $[\theta_p, \theta_q]$ denotes the Lie bracket, computed block-wise on $\mathfrak{so}(4)$ and re-expanded in coefficients, and the scalar bracket scale $s \in \mathbb{R}$ is learned end-to-end on training data, calibrating the second-order truncation against the residual cubic and higher terms under the gene-space loss.

The same bracket gives a pair score, the *bracket score*,

$$\varepsilon(p, q) \;=\; \frac{\|[\theta_p, \theta_q]\|_F}{\|\theta_p\|_F\, \|\theta_q\|_F}, \tag{4}$$

defined as the Frobenius norm of the bracket normalized by the two single-perturbation norms. The score takes values between zero and two, with zero attained exactly on commuting pairs, and is invariant under any orthogonal change of latent basis applied simultaneously to both generators (Proposition 9). Because it depends only on the two single generators, it is available before any pair measurement.

3.4 Architecture

The encoder f_ϕ and decoder g_ω are SwiGLU residual MLPs (Appendix C.1). The encoder maps $\mathbb{R}^G \to \mathbb{R}^{128}$ and the decoder combines a linear readout with a residual skip, $g_\omega(\boldsymbol{z}) = W\boldsymbol{z} + \boldsymbol{b} + \alpha\, h(\boldsymbol{z})$, keeping the decoder close to linear so that perturbation-induced changes in the latent are approximately preserved in gene space. For each training perturbation $p \in \mathcal{P}_{\text{train}}$, the generator θ_p is a free parameter in a learned lookup table, reconstructed via Eq. 1 at every forward pass.

Generator flow v_ψ. For unseen perturbations $p^\star \notin \mathcal{P}_{\text{train}}$, a conditional velocity field v_ψ on $\mathfrak{so}(4)^{32}$ maps Gaussian noise plus the descriptor $e_{p^\star}$ to a generator. We use flow matching rather than a deterministic regressor because the descriptor-to-generator map is one-to-many (paralogous genes share descriptors but have distinct generators), and we work on the flat algebra rather than the group to admit straight-line interpolants without manifold projection. Following Lipman et al. (2022), the flow-matching loss is

$$\mathcal{L}_{\text{FM}}(\psi) \;=\; \mathbb{E}_{p,\, t\sim\mathcal{U}(0,1),\, \theta_0} \left\| v_\psi(\theta_t, t, e_p) - (\theta_p - \theta_0) \right\|_2^2, \quad \theta_t = (1-t)\theta_0 + t\theta_p,\ \theta_0 \sim \mathcal{N}(0, \sigma_0^2 I_{192}). \tag{5}$$

At inference we Euler-integrate from noise conditioned on $e_{p^\star}$ and average $M = 32$ samples on the algebra to form $\hat{\theta}_{p^\star}$, then exponentiate to give $R_{p^\star}$. The variance across flow samples serves as a per-prediction reliability signal without any dedicated uncertainty module (Section 4.4).

Disentanglement. An adversarial classifier trained through gradient reversal discourages the latent from encoding perturbation identity (Lotfollahi et al., 2023; Hetzel et al., 2022), so that θ_p is the only perturbation-dependent parameter and composition through Eq. 3 is the sole mechanism for producing pair predictions.

3.5 Training

Training proceeds in two stages. Stage 1 fits the rotation autoencoder jointly with the lookup $\{\theta_p\}_{p\in\mathcal{P}_{\text{train}}}$ and the bracket scale s. Stage 2 fits the flow v_ψ on the resulting lookup. The Stage 1 objective is

$$\mathcal{L}_1 \;=\; \lambda_{\text{rec}}\, \mathcal{L}_{\text{rec}} + \lambda_{\text{deg}}\, \mathcal{L}_{\text{rec}}^{\text{DEG}} + \lambda_{\text{cmp}}\, \mathcal{L}_{\text{compose}} + \lambda_{\text{adv}}\, \mathcal{L}_{\text{adv}}, \tag{6}$$

where $\mathcal{L}_{\text{rec}}$ is a heteroscedastic Gaussian NLL with learned per-gene variance (genes with high biological noise contribute less, concentrating signal on perturbation-responsive genes), $\mathcal{L}_{\text{rec}}^{\text{DEG}}$ upweights the top-50 DE genes tenfold, $\mathcal{L}_{\text{compose}}$ supervises the multiplicative product R_pR_q against measured combination expression (calibrating the bracket scale s, with test pairs never entering this gradient), and $\mathcal{L}_{\text{adv}}$ is the adversarial disentanglement loss. Equations for each term and all hyperparameters are in Appendix C (Eqs. 22–24, Table 5).

After epoch E_f, decoder weights are frozen (coordinate lock) so that all perturbation-specific information must pass through the rotation, which is the only composable component. Without this constraint the decoder absorbs perturbation-specific readouts that bypass the rotation and do not compose (Appendix C, Table 12).

Stage 2 trains v_ψ by $\mathcal{L}_{\text{FM}}$ (Eq. 5), initialized from kernel ridge regression on (e_p, θ_p) pairs to start at the deterministic conditional mean. A fraction $\rho = 0.2$ of each minibatch is replaced by BCH-composed synthetic pairs $\theta_a + \theta_b + \frac{1}{2}[\theta_a, \theta_b]$ from random training pairs, regularizing the flow on the composition manifold without exposing it to measured combination labels.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate on three CRISPR perturbation screens. **Norman** (Norman et al., 2019) (CRISPRa, 105 held-out perturbations including 84 two-gene combinations), **Replogle K562** (Replogle et al., 2022) (CRISPRi, 271 held-out perturbations), and **Replogle RPE1** (Replogle et al., 2022) (CRISPRi, 382 held-out perturbations). Full dataset statistics, preprocessing, and split definitions are in Appendix D.

Metrics. Following Cheng et al. (2025), we report Pearson correlation on all genes (ρ) and on the top-20 DE genes (ρ^{DEG}), directional accuracy on the top-20 DE genes ($\mathrm{DA}^{\mathrm{DEG}}$), cosine similarity on the top-200 (Cos_{200}), all as percentages, and for uncertainty the Spearman correlation between predicted confidence and accuracy (SPCC).

Baselines. We compare against eleven methods: AvgKnown, CellOracle, scFoundation (Hao et al., 2024), scGPT (Cui et al., 2024), samsVAE (Bereket & Karaletsos, 2023), GEARS (Roohani et al., 2024), scGen (Lotfollahi et al., 2019), discrepancy-VAE (Zhang et al., 2023), PRESCRIBE (Cheng et al., 2025), STATE (Adduri et al., 2026), and PerturbDiff (Yuan et al., 2026). Every method is trained from scratch on the GEARS-standard splits (Roohani et al., 2024) under a single pipeline and evaluated over five seeds. Appendix E gives per-method configurations and the perturbation representation each one conditions on.

4.2 PREDICTION QUALITY

Table 1 reports zero-shot prediction across three CRISPR screens. OpPert achieves the highest $\mathrm{DA}^{\mathrm{DEG}}$ on all three (87.57% on Norman, 81.09% on K562, 87.96% on RPE1), ahead of the strongest baseline, discrepancy-VAE, by $+11.9$, $+7.1$, and $+3.3$ percentage points, and it leads on both ρ^{DEG} and Cos_{200} on every screen. All-gene ρ is the exception, topped by AvgKnown on K562 and by PRESCRIBE on RPE1: because the response is concentrated in a few genes while the rest of the panel is unresponsive, a correlation over all genes rewards predicting the dataset mean, so we report it alongside the DE-restricted metrics rather than in place of them. Per-perturbation ρ^{DEG} distributions confirm the advantage is systematic rather than driven by outliers (Figure 4a,c). We also report distributional metrics for every method in Appendix E.1.

For drug discovery, the model must generalize to combinations where neither constituent gene was observed during training. Norman is the only dataset with held-out two-gene combinations, so we stratify its 84 held-out pairs by the number of unseen constituent genes (Figure 4b). OpPert maintains 92% $\mathrm{DA}^{\mathrm{DEG}}$ when both genes were seen, 91% when one is unseen, and 68% when both are unseen, still substantially above all baselines at every difficulty level.

Drug perturbations (SciPlex3). To test generalization beyond genetic perturbations, we evaluate on SciPlex3 (Srivatsan et al., 2020), a drug-response screen with 15 held-out drugs. Table 2 compares OpPert against chemical perturbation baselines. OpPert achieves the highest scores on all four metrics, with the largest gap on $\mathrm{DA}^{\mathrm{DEG}}$ ($+3.5$ pp over the next-best method). Figure 6B shows per-drug directional accuracy for all 15 held-out compounds at 10 μM, with a mean $\mathrm{DA}^{\mathrm{DEG}}$ of 76%.

4.3 OPERATOR GEOMETRY IN PRACTICE

We verify that learned operators satisfy the small-angle regime required by the BCH truncation and that norm preservation holds in practice.

Figure 4e shows the distribution of learned rotation eigenangles across all trained perturbation operators. The median eigenangle is 0.114 rad. 92% of operators fall within $\pi/8$ and 99.6% within $\pi/4$. The BCH series converges rapidly in this regime, validating the truncation to first-order commutator used in Eq. 3.

This bounds prediction magnitude by construction, since rotation acts on a fixed-radius shell in latent space while additive composition has no such constraint. Figure 4f shows the pairwise commutator

Table 1. Zero-shot perturbation prediction. Metrics as defined in Section 4.1, in %, as mean ± s.d. over five seeds with every method trained from scratch under one pipeline. † deterministic given the split. **Bold** = best, underline = second best.

Method	Norman				Replogle K562				Replogle RPE1			
	ρ	ρ^{DEG}	$\mathrm{DA}^{\mathrm{DEG}}$	Cos_{200}	ρ	ρ^{DEG}	$\mathrm{DA}^{\mathrm{DEG}}$	Cos_{200}	ρ	ρ^{DEG}	$\mathrm{DA}^{\mathrm{DEG}}$	Cos_{200}
AvgKnown†	39.6	59.0	61.9	54.9	**36.9**	46.1	56.1	42.9	54.5	57.9	32.3	53.8
CellOracle†	9.8	12.5	16.4	11.6	4.4	5.9	41.2	5.5	39.9	7.4	23.7	6.9
scFoundation	60.2±2.8	65.9±3.5	62.4±3.3	61.0±4.0	24.7±1.5	47.4±2.5	57.5±1.4	44.1±1.5	47.8±1.3	58.1±1.6	45.9±2.0	55.2±1.4
scGPT	62.0±4.9	69.9±3.9	73.7±2.3	61.4±3.0	33.4±1.6	43.3±2.4	57.5±1.7	40.1±1.6	49.8±1.8	64.6±1.8	66.2±1.6	61.4±1.7
samsVAE	11.8±1.6	30.3±4.9	48.2±2.7	28.7±4.7	8.9±0.9	28.8±2.9	42.7±2.3	26.1±1.5	12.9±0.8	34.4±2.6	24.4±1.8	32.3±2.0
scGen	37.9±1.9	57.6±3.9	60.1±1.3	53.0±2.6	34.7±1.0	44.1±2.1	54.0±0.8	41.1±1.1	52.4±1.3	54.7±1.4	31.4±0.3	50.8±1.1
discrepancy-VAE	46.8±1.6	54.1±3.6	75.7±1.9	58.3±2.2	28.3±0.7	46.9±1.4	74.0±0.9	49.0±0.9	40.7±0.9	65.2±1.1	84.7±0.4	64.9±0.8
GEARS	44.8±5.3	64.0±3.2	68.1±3.1	59.1±4.0	32.2±1.3	42.8±2.9	54.8±1.4	39.6±2.3	47.9±1.4	53.5±1.8	31.7±1.5	50.3±1.4
PRESCRIBE	58.1±3.4	64.1±4.1	74.4±2.0	60.5±3.8	35.2±1.8	43.4±3.6	69.0±1.7	41.0±2.6	**58.7**±1.9	64.1±2.4	79.2±1.0	59.4±1.8
STATE	38.9±3.2	47.6±6.3	57.1±3.2	45.2±3.7	28.5±3.9	38.0±4.0	53.9±1.8	36.4±5.1	46.6±2.0	53.2±1.7	28.9±2.0	47.7±1.8
PerturbDiff	40.8±3.4	49.6±4.4	62.3±2.5	48.6±2.9	34.9±1.8	49.2±1.9	70.3±1.2	46.2±1.5	53.7±1.9	68.6±1.8	79.4±1.0	64.0±1.6
OpPert	**62.6**±3.3	**74.1**±3.9	**87.6**±2.4	**72.3**±3.0	33.4±2.4	**55.5**±3.5	**81.1**±1.9	**52.5**±2.1	55.6±2.2	**70.3**±2.1	**88.0**±1.3	**70.4**±2.1
− rotation (additive)	49.2±2.9	53.3±3.6	71.3±2.1	58.2±3.4	21.5±1.4	42.7±2.5	67.8±1.9	39.0±2.7	43.9±1.8	54.6±2.0	73.3±1.4	55.4±1.9

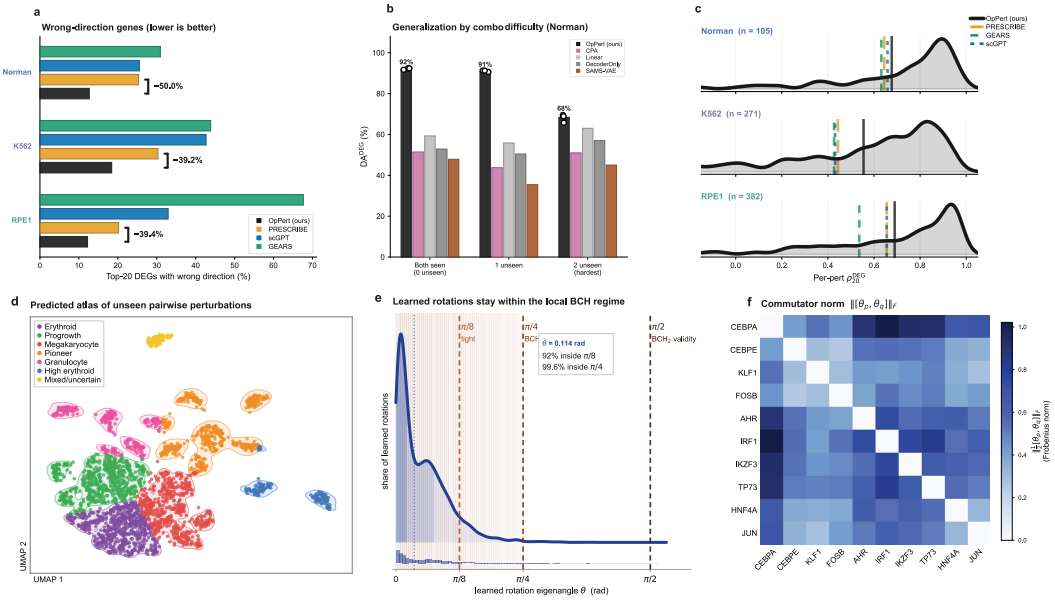

Figure 4. Prediction quality across datasets and difficulty levels. (a) Fraction of top-20 DEGs predicted in the wrong direction (lower is better). Brackets show relative reduction from OpPert to the next-best method. **(b)** $\mathrm{DA}^{\mathrm{DEG}}$ on Norman held-out pairs stratified by the number of unseen constituent genes. Individual seeds shown as dots for OpPert. **(c)** Per-perturbation ρ^{DEG} distributions. Black curve = OpPert, dashed lines = aggregate values for baselines. **(d)** UMAP of BCH-composed unseen pair predictions form response-state clusters. **(e)** Learned rotations stay in the local BCH regime. **(f)** Pairwise commutator heatmap for selected Norman genes.

norm $\|[\theta_p, \theta_q]\|_F$ for selected Norman genes, confirming that the bracket captures non-commutativity across gene families, and Figure 4d projects all pairwise compositions into a UMAP embedding where predicted states cluster by known lineage programs (erythroid, megakaryocyte, granulocyte, pioneer), indicating that the operator product captures biologically coherent structure (Appendix G).

4.4 Per-prediction confidence from flow variance

Existing per-prediction confidence methods require dedicated modules and training objectives (Cheng et al., 2025; Roohani et al., 2024). Conversely, OpPert's stochastic flow provides a confidence signal with no additional parameters or supervision.

Table 2. SciPlex3 drug perturbation prediction. Same metrics as Table 1.

Method	ρ	ρ^{DEG}	$\mathrm{DA}^{\mathrm{DEG}}$	Cos_{200}
AvgKnown	35.4	52.0	74.6	44.8
Linear (KRR)	37.1±3.1	55.1±2.3	75.0±2.1	47.0±1.9
chemCPA	28.3±4.9	47.9±3.2	67.8±3.2	39.7±2.7
chemCPA (pretrained)	41.8±3.3	61.5±2.9	74.9±2.7	51.4±1.9
SAMS-VAE	31.6±4.0	54.8±3.4	71.6±2.1	45.9±1.6
SAMS-VAE(S)	38.2±4.6	61.7±2.8	75.8±2.5	52.3±1.8
OpPert	**44.6**±2.8	**64.2**±2.1	**79.3**±2.0	**54.7**±1.3

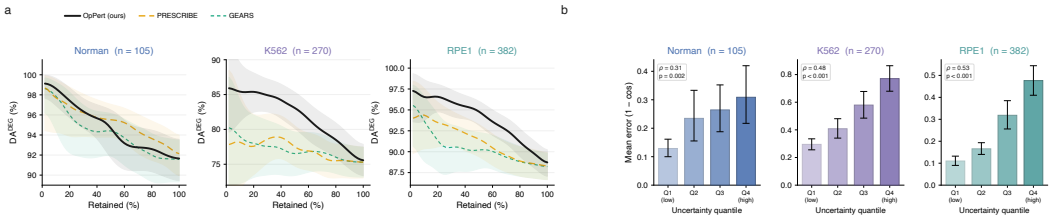

Figure 5. Per-prediction confidence from flow variance. (a) Risk-coverage curves showing $\mathrm{DA}^{\mathrm{DEG}}$ of the retained set as a function of coverage. Shaded bands = 95% bootstrap CI. **(b)** Mean prediction error $(1 - \cos)$ by predicted uncertainty quartile (Q1 = low uncertainty). Error bars = 95% bootstrap CI. Spearman ρ and p-values annotated per dataset.

Given M flow samples $\{\hat{\theta}_p^{(m)}\}_{m=1}^{M}$ drawn by Euler integration from independent noise seeds, we define the per-perturbation uncertainty score as the coefficient of variation on the top-DE genes:

$$u_p = \frac{\left\|\mathrm{std}_m\left(\hat{\delta}_p^{(m)}\right)\right\|_{\text{top-DE}}}{\left\|\bar{\delta}_p\right\|_{\text{top-DE}}}, \qquad \bar{\delta}_p = \frac{1}{M}\sum_{m=1}^{M}\hat{\delta}_p^{(m)}, \tag{7}$$

where $\hat{\delta}_p^{(m)} = g_\omega(R_{\hat{\theta}_p^{(m)}}\boldsymbol{z}) - g_\omega(\boldsymbol{z})$ is the decoded shift for sample m and std_m is the element-wise standard deviation. We use $M{=}128$ samples and restrict to the top-20 DE genes per perturbation. No additional parameters or training are required beyond the flow already used for prediction.

We evaluate u_p by Spearman rank correlation (SPCC) between predicted uncertainty and actual prediction error. OpPert exceeds PRESCRIBE's dedicated Natural Posterior Network on K562 (SPCC 48% vs. 14%) and RPE1 (53% vs. 29%), the two datasets with the widest accuracy spread across perturbations. Figure 5a translates this ranking into selective prediction. Retaining only the most confident predictions monotonically improves $\mathrm{DA}^{\mathrm{DEG}}$ on all three datasets, with OpPert's risk-coverage curve above both PRESCRIBE and GEARS on K562 and RPE1.

4.5 In-vitro experimental validation

In-silico evaluation compares predictions against held-out observational data from the same experimental batch. To test whether OpPert predictions transfer to an independent experimental modality, we selected Belinostat, a well-characterized HDAC inhibitor present in the SciPlex3 screen, and identified four candidate genes from differential expression analysis between predicted Belinostat-treated and control profiles in MCF7 cells. The candidates span cyclic nucleotide signaling (*PDE4D*), protein kinase C signaling (*PRKCA*, *PRKCB*), and fatty-acid β-oxidation (*ACSBG1*). We performed RT-qPCR in MCF7 cell lines treated with 10 μM Belinostat for 24 h (Figure 6A).

RT-qPCR confirms statistically significant expression changes in the predicted direction for three genes. *PRKCA* and *PRKCB* reach $p < 0.001$, consistent with adaptive kinase signaling in response to HDAC inhibition, while *PDE4D* is significant at $p < 0.01$, reflecting stress-response signaling not previously associated with Belinostat. Bright-field microscopy confirms the expected apoptotic morphology (membrane blebbing, cell rounding) in treated cells (Figure 6A). These results provide orthogonal evidence that OpPert captures real transcriptional consequences of drug treatment, not just statistical patterns in the training distribution.

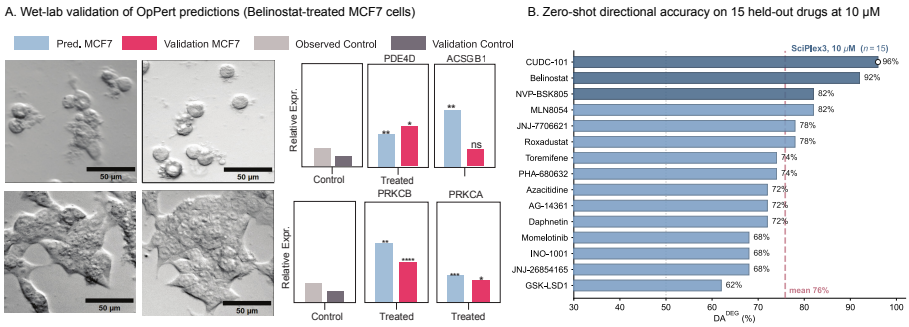

Figure 6. In-vitro experimental validation and drug generalization. (A) RT-qPCR expression measurements for selected genes predicted by OpPert to change under Belinostat treatment in MCF7 cells. Microscopy panels show cell morphology in control versus treated conditions. Significance levels annotated per gene. **(B)** Zero-shot directional accuracy on 15 held-out drugs from SciPlex3 at 10 μM. Dashed line shows mean $\mathrm{DA}^{\mathrm{DEG}}$ across all drugs.

4.6 Ablations

To isolate the contribution of the rotation operator, we replace it with vector-additive composition (the standard latent-shift approach used by scGen, CPA, and GEARS) while keeping the encoder, decoder, flow, and descriptor encoder identical. The ablation row in Table 1 reports all four metrics on the three screens.

The rotation is responsible for 13–16 pp of $\mathrm{DA}^{\mathrm{DEG}}$ across all screens, with the largest drop on Norman (-16.3 pp) where the combination evaluation stresses compositional extrapolation most directly. In all cases the additive ablation still uses the same flow and descriptor encoder, confirming that the gain is attributable to the geometric structure of the operator rather than model capacity.

The Lie bracket provides an interaction score that correlates with measured epistasis (Spearman $\rho = 0.49$, $p = 3.5 \times 10^{-6}$) from single-perturbation operators alone, before any pair is measured (Figure 1b). No existing method provides a comparable score without pair supervision.

Appendix F reports sensitivity across nine additional hyperparameters. Key findings are that block size $r{=}4$ is optimal because $r{=}2$ collapses to an abelian torus where every bracket vanishes (Table 11), that coordinate locking the decoder is the single largest contributor to compositional accuracy with no-lock dropping to 75.4% (Table 12), and that inference ensemble size K saturates by $K{=}32$ (Table 15).

5 Conclusion

In this paper we presented OpPert, which represents a perturbation as a rotation of a cell's latent state and uses conditional flow matching to generate rotations for perturbations it has not seen during training. Treating a perturbation as an operator rather than a displacement makes the combination of two perturbations a product of operators and their interaction a Lie bracket, both computable from single-perturbation data alone. On genetic and chemical perturbation benchmarks, OpPert improves directional accuracy over the evaluated baselines, and its Lie bracket score correlates with measured non-additivity in held-out gene pairs.

We also validated OpPert *in vitro* using RT-qPCR in MCF7 cells treated with belinostat, which confirmed the predicted expression changes for the genes selected for measurement.

Ethics Statement

This study uses previously published single-cell perturbation datasets and includes RT-qPCR experiments in MCF7 cells. These experiments measure gene-expression responses and do not establish drug efficacy, safety, or causal mechanisms. OpPert has not been evaluated for clinical decision-making, and its predictions should not be used to make treatment decisions. Predictions outside the evaluated cell types and perturbations require further experimental validation.

Reproducibility statement

The anonymized source code is included in the supplementary material. Dataset preprocessing and evaluation splits are described in Appendix D. Training procedures, baseline implementations, and computational requirements are described in Appendices C, E, and J, respectively. Mathematical derivations are provided in Appendix B. The supplementary material also describes the cell-culture and RT-qPCR procedures and provides the primer sequences used for experimental validation.

AI use statement

We used LLMs to rephrase sentences, correct grammar, and revise explanations in the manuscript. We also used them to flag possible inconsistencies in mathematical statements and experimental reporting. These tools were not used to generate the research ideas, design the experiments or in analyses. The authors take responsibility for the final manuscript, including its scientific claims and reported results.

References

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, et al. Predicting cellular responses to perturbation across diverse contexts with state. *Cell*, 2026.

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.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, 56(2):411–421, 2006. doi: 10.1002/mrm.20965.

Ding Bai, Caleb N Ellington, Shentong Mo, Le Song, and Eric P Xing. Attentionpert: accurately modeling multiplexed genetic perturbations with multi-scale effects. *Bioinformatics*, 40(Supplement_1): i453–i461, 2024. doi: 10.1093/bioinformatics/btae244.

Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. In *Advances in Neural Information Processing Systems*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/0001ca33ba34ce0351e4612b744b3936-Abstract-Conference.html.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

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.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=g7ohDlTITL.

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.

Yiqun Chen and James Zou. Genepert: Leveraging genept embeddings for gene perturbation prediction. *bioRxiv*, 2024. doi: 10.1101/2024.10.27.620513.

Jiabei Cheng, Changxi Chi, Jingbo Zhou, Hongyi Xin, and Jun Xia. PRESCRIBE: predicting single-cell responses with Bayesian estimation. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=A5O41ntKjk.

Guohui Chuai, Xiaohan Chen, Xingbo Yang, Cheng Zhang, Kairu Qu, Yiheng Wang, Wannian Li, Jingya Yang, Duanmiao Si, Feiyang Xing, Yicheng Gao, Siqi Wu, Shaliu Fu, Bing He, and Qi Liu. Towards building a world model to simulate perturbation-induced cellular dynamics by AlphaCell. *bioRxiv*, 2026. doi: 10.64898/2026.03.02.709176.

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.

Mingze Dong, Bao Wang, Jessica Wei, Antonio H. de O. Fonseca, Curtis J. Perry, Alexander Frey, Feriel Ouerghi, Ellen F. Foxman, Jeffrey J. Ishizuka, Rahul M. Dhodapkar, and David van Dijk. Causal identification of single-cell experimental perturbation effects with cinema-ot. *Nature Methods*, 20(11):1769–1779, 2023. doi: 10.1038/s41592-023-02040-5.

Boyang Fu, George Dasoulas, Sameer Gabbita, Xiang Lin, Shanghua Gao, Xiaorui Su, Soumya Ghosh, and Marinka Zitnik. STRAND: sequence-conditioned transport for single-cell perturbations. *arXiv preprint arXiv:2602.10156*, 2026.

Mario Geiger and Tess E. Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022. doi: 10.48550/ARXIV.2207.09453.

Guadalupe Gonzalez, Xiang Lin, Isuru Herath, Kirill Veselkov, Michael Bronstein, and Marinka Zitnik. Combinatorial prediction of therapeutic perturbations using causally inspired neural networks. *Nature Biomedical Engineering*, 10(5):1008–1025, 2026. doi: 10.1038/s41551-025-01481-x.

Jonathan Gordon, David Lopez-Paz, Marco Baroni, and Diane Bouchacourt. Permutation equivariant models for compositional generalization in language. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylVNerFvr.

Brian C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, volume 222 of *Graduate Texts in Mathematics*. Springer, 2nd edition, 2015. doi: 10.1007/978-3-319-13467-3.

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.

Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled cayley transform. In *International Conference on Machine Learning*, volume 80, pp. 1974–1983, 2018. URL http://proceedings.mlr.press/v80/helfrich18a.html.

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*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/aa933b5abc1be30baece1d230ec575a7-Abstract-Conference.html.

Michael J. Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. Lietransformer: Equivariant self-attention for lie groups. In *International Conference on Machine Learning*, volume 139, pp. 4533–4543, 2021. URL http://proceedings.mlr.press/v139/hutchinson21a.html.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: an optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Kenji Kamimoto, Blerta Stringa, Christy M. Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A. Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, 2023. doi: 10.1038/s41586-022-05688-9.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: a comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygcCnNKwr.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, Nadezhda Azbukina, Fátima Sanchís-Calleja, Theo Uscidda, Artur Szalata, Manuel Gander, Aviv Regev, Barbara Treutlein, J. Gray Camp, and Fabian J. Theis. CellFlow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025. doi: 10.1101/2025.04.11.648220.

Anthony W. Knapp. *Lie Groups Beyond an Introduction*, volume 140 of *Progress in Mathematics*. Birkhäuser, Boston, 2nd edition, 2002.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, volume 80, pp. 2879–2888, 2018. URL http://proceedings.mlr.press/v80/lake18a.html.

Chen Li, Lei Wei, and Xuegong Zhang. Unified multimodal learning enables generalized cellular response prediction to diverse perturbations. *bioRxiv*, 2025. doi: 10.1101/2025.11.13.688367. Introduces the X-Pert model.

Tzu-Yuan Lin, Minghan Zhu, and Maani Ghaffari. Lie neurons: Adjoint-equivariant neural networks for semisimple lie algebras. In *International Conference on Machine Learning*, volume 235, pp. 30529–30545, 2024. URL https://proceedings.mlr.press/v235/lin24aa.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2022. URL https://arxiv.org/abs/2210.02747.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: LLM quantization with learned rotations. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ogO6DGE6FZ.

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.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

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, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):MSB202211517, 2023.

Gabriel M. Mejia, Henry E. Miller, Francis J. A. Leblanc, Bo Wang, Brendan Swain, and Lucas Paulo de Lima Camillo. Diversity by design: addressing mode collapse improves scRNA-seq perturbation modeling on well-calibrated metrics. *arXiv preprint arXiv:2506.22641*, 2025.

Henry E. Miller, Gabriel M. Mejia, Francis J. A. Leblanc, Bo Wang, Brendan Swain, and Lucas Paulo de Lima Camillo. Deep learning-based genetic perturbation models do outperform uninformative baselines on well-calibrated metrics. *bioRxiv*, 2025. doi: 10.1101/2025.10.20.683304.

Mircea Mironenco and Patrick Forré. Lie group decompositions for equivariant neural networks. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11366.

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.

Alessandro Palma, Sergei Rybakov, Leon Hetzel, Stephan Günnemann, and Fabian J. Theis. FlatVI: enforcing latent euclidean geometry in single-cell VAEs for manifold interpolation. In *International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DoDXFkF10S.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, pp. 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.org/10.1109/ICCV51070.2023.00387.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

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

Jiafa Ruan, Ruijie Quan, Liyang Xu, Zongxin Yang, and Yi Yang. Beyond independent genes: Learning module-inductive representations for single-cell gene perturbation prediction. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.04901.

David Ruhe, Johannes Brandstetter, and Patrick Forré. Clifford group equivariant neural networks. *Advances in neural information processing systems*, 36:62922–62990, 2023.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Zakhar Shumaylov, Peter Zaika, James Rowbottom, Ferdia Sherry, Melanie Weber, and Carola-Bibiane Schönlieb. Lie algebra canonicalization: Equivariant neural operators under arbitrary lie groups. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7PLpiVdnUC.

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.

Asher Trockman and J. Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Pbj8H_jEHYv.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44(6):1050–1059, 2025. doi: 10.1038/s41587-025-02777-8.

Gefei Wang, Tianyu Liu, Jia Zhao, Youshu Cheng, and Hongyu Zhao. Modeling and predicting single-cell multi-gene perturbation responses with sclambda. *bioRxiv*, 2024. doi: 10.1101/2024.12.04.626878.

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

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T. Whitfield, Craig Russell, Yassir El Mesbahi, Jiarui Ding, Marta M. Fay, Berton Earnshaw, Emmanuel Noutahi, and Alisandra K. Denton. Txpert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03113-4. Journal version of arXiv:2505.14919.

Aaron Wenteler, Martina Occhetta, Nikhil Branson, Victor Curean, Magdalena Huebner, William Dee, William Connell, Siu Pui Chung, Alex Hawkins-Hooker, Yasha Ektefaie, César Miguel Valdez Córdova, and Amaya Gallagher-Syed. Perteval-scfm: Benchmarking single-cell foundation models for perturbation effect prediction. In *International Conference on Machine Learning*, volume 267, 2025. URL https://proceedings.mlr.press/v267/wenteler25a.html.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

Chenglei Yu, Chuanrui Wang, Bangyan Liao, and Tailin Wu. scDFM: distributional flow matching model for robust single-cell perturbation prediction. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QSGanMEcUV.

Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D Welch. Perturbnet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 21(8), 2025. doi: 10.1038/s44320-025-00131-3.

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

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2186.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2019. doi: 10.1109/CVPR.2019.00589. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zhou_On_the_Continuity_of_Rotation_Representations_in_Neural_Networks_CVPR_2019_paper.html.

Ouyang Zhu and Jun Li. Scouter predicts transcriptional responses to genetic perturbations with large language model embeddings. *Nature Computational Science*, 6(1):21–28, 2025. doi: 10.1038/s43588-025-00912-8.

Xinqi Zhu, Chang Xu, and Dacheng Tao. Commutative lie group VAE for disentanglement learning. In *International Conference on Machine Learning*, volume 139, pp. 12924–12934, 2021. URL http://proceedings.mlr.press/v139/zhu21f.html.

A Extended Related Work

Additive composition methods. The dominant approach represents perturbation effects as additive vectors in a learned latent space. scGen (Lotfollahi et al., 2019) uses a VAE with vector arithmetic between treated and control populations. CPA (Lotfollahi et al., 2023) learns disentangled perturbation embeddings composed by summation, handling drugs, doses, and genetic perturbations. GEARS (Roohani et al., 2024) incorporates gene interaction graphs through a GNN encoder and pools per-gene embeddings with permutation-invariant aggregation across the perturbed-gene set, an order-independent (and therefore commutative) composition. scVI (Lopez et al., 2018) supplies the probabilistic single-cell backbone several perturbation methods build on. GenePert (Chen & Zou, 2024) shows that L2-regularized linear regression with LLM-derived gene embeddings outperforms these architectures on single-perturbation prediction. The result indicates that architectural complexity alone does not solve the problem. AttentionPert (Bai et al., 2024), PerturbNet (Yu et al., 2025), and xPert (Li et al., 2025) retain additive composition. scLAMBDA (Wang et al., 2024) combines LLM gene embeddings with disentangled VAEs.

Distributional and embedding methods. CellOT (Bunne et al., 2023) and CellFlow (Klein et al., 2025) use optimal transport and flow matching to learn maps between control and perturbed distributions, with no algebraic composition rule for combinations. Scouter (Zhu & Li, 2025) uses a compressor-generator network conditioned on LLM gene embeddings; it relies on learned embedding similarity rather than an algebraic composition rule. TxPert (Wenkel et al., 2026) uses multi-graph GNNs over protein interaction, gene ontology, and phenomic networks. STRAND (Fu et al., 2026) conditions on regulatory DNA sequence to parameterize a transport process from control to perturbed states. AlphaCell (Chuai et al., 2026) combines latent manifold rectification with optimal transport flow matching. scDFM (Yu et al., 2026) applies distributional flow matching and reports strong Norman combination accuracy without algebraic composition guarantees. STATE (Adduri et al., 2026) trains a transformer on 100M perturbed cells across 70 contexts, and scBIG (Ruan et al., 2026) introduces module-level inductive bias with conditional flow matching. PDGrapher (Gonzalez et al., 2026) addresses the inverse problem (disease state to therapeutic perturbation) using causal GNNs on protein interaction networks.

Benchmarks and the case for inductive bias. Five independent benchmarks confirm that no existing method dominates. Ahlmann-Eltze et al. (2025) show that deep learning does not consistently outperform linear baselines. Wei et al. (2026) find no dominant method across 29 datasets and 27 methods. Viñas Torné et al. (2025) show that current methods primarily capture average treatment effects. Wenteler et al. (2025) report that single-cell foundation model embeddings supply limited improvement in the zero-shot regime. Miller et al. (2025) show that under well-calibrated evaluation, deep learning does outperform uninformative baselines, locating part of the apparent failure in metric calibration. The aggregate signal is that stronger inductive biases and better-calibrated metrics are the productive direction, not larger models. The Lie-algebraic composition rule of OpPert supplies exactly such an inductive bias. Composition closes on $\mathrm{SO}(n)$ by construction. The bracket term encodes the non-additive interaction structure beyond the reach of abelian methods.

Lie groups in machine learning. The Cayley map from skew-symmetric matrices to orthogonal matrices was introduced for RNN weight parameterization by Helfrich et al. (2018) and extended to convolutional layers by Trockman & Kolter (2021). Zhou et al. (2019) proved that $\mathrm{SO}(3)$ cannot be continuously parameterized in fewer than 5 dimensions. This result motivates overparameterized representations like our skew-symmetric generators. Zhu et al. (2021) used commutative Lie group structure in VAEs for disentangled representation learning. We extend this to the non-commutative case, which is necessary for capturing gene-gene interactions. Lin et al. (2024) defined Lie bracket layers as neural network primitives, validating the use of $[A, B]$ as a learned interaction operator. Mironenco & Forré (2024) and Shumaylov et al. (2025) explore Lie group representations for temporal and physical dynamics, while Ruhe et al. (2023) extend geometric algebra to neural networks. Bronstein et al. (2021) established symmetry groups as the organizing principle of geometric deep learning; our work applies this principle to perturbation biology.

Rotation representations. Learning rotation matrices in neural networks is a well-studied problem. Beyond the Cayley map, Chen & Lipman (2024) developed Riemannian flow matching on $\mathrm{SO}(3)$

for molecular generation. Geiger & Smidt (2022) built SO(3)-equivariant networks using spherical harmonics for molecular property prediction. Hutchinson et al. (2021) introduced self-attention equivariant to arbitrary Lie groups. Liu et al. (2025) used rotation parameterization for LLM quantization. Rotation structure supplies a strong inductive bias across these domains. We apply rotation composition to perturbation biology, where the group structure (non-commutativity, BCH formula) carries direct biological meaning through the bracket as an epistasis quantifier.

Compositional generalization. Lake & Baroni (2018) showed that neural networks struggle with systematic compositional generalization. Keysers et al. (2020) formalized evaluation protocols for compositional splits. Our prediction task (seen singles $\rightarrow$ unseen combinations) is a compositional generalization problem. Gordon et al. (2020) proved that equivariance is both necessary and sufficient for compositional generalization under group-structured data. The group structure of $\mathrm{SO}(n)$ provides exactly this. The composition of two rotations is itself a rotation, guaranteeing closure under combination. Dong et al. (2023) and Mejia et al. (2025) connect compositional generalization to causal reasoning and diversity-based training, complementary perspectives to our algebraic approach.

Evaluation metrics for perturbation prediction. The perturbation prediction field lacks standardized evaluation. Common metrics include Pearson correlation on DE genes, mean squared error, and R^2 on delta expression vectors. Directional accuracy (the fraction of correctly predicted up/down directions for DE genes) carries direct biological meaning. Knowing which genes are upregulated and which are downregulated is the actionable readout for most downstream uses. Exact-magnitude prediction is dominated by measurement noise on the per-gene scale (Wei et al., 2026). Perturbation direction Pearson (PDE), the Pearson correlation between predicted and true perturbation vectors, is a complementary continuous measure. We report both throughout. Wei et al. (2026) survey 27 methods across 29 datasets and find that no single metric is consistently informative. We therefore report multiple complementary metrics.

Structural comparison of prediction objects. Table 3 compares perturbation prediction methods by the structure of the prediction object rather than by numerical accuracy. Additive, pooling, and direct-response predictors can achieve strong empirical performance, but their prediction objects do not define a closed product or a pair-interaction score computable before observation. OpPert is the only method that supplies all three: a closed composition rule, norm preservation, and a pre-measurement pair score (the commutator norm).

Table 3. Structural comparison of perturbation prediction objects. "Pair score" means a score supplied by the model's own composition rule before observing the pair label.

Method	Pair rule	Closed product	Norm preserving	Pair score	Zero-shot pairs
scGen (Lotfollahi et al., 2019)	shift	✗	✗	✗	✗
CPA (Lotfollahi et al., 2023)	sum	✗	✗	✗	partial
GEARS (Roohani et al., 2024)	GNN aggregation	✗	✗	✗	✓
AttentionPert (Bai et al., 2024)	attention pool	✗	✗	✗	✓
CellFlow (Klein et al., 2025)	transport map	✗	✗	✗	task-dependent
OpPert	rotation+BCH	✓	✓	✓	✓

Disentanglement in perturbation models. Perturbation prediction requires separating the basal cell state from the perturbation effect. CPA (Lotfollahi et al., 2023) uses adversarial training to disentangle these components. Our rotation framework provides a geometric form of disentanglement. The rotation R_p acts on the latent representation z by rotating it, not by adding a vector. Recent work on enforcing geometric structure in single-cell latent spaces (Palma et al., 2025) addresses related concerns. In our framework, the perturbation effect is a transformation of the state rather than a shift independent of state. The coordinate-lock mechanism further enforces disentanglement by preventing the decoder from learning perturbation-specific readout patterns. Together, these mechanisms ensure that perturbation information is encoded in the rotation generators, not distributed across the encoder, decoder, or other components.

B Mathematical Background and Proofs

This appendix provides the formal mathematical framework underlying OpPert. We collect notation, definitions, and proofs of the propositions cited in Section 3. Standard references for Lie group theory are Hall (Hall, 2015) and Knapp (Knapp, 2002); for flow matching on manifolds we follow Lipman et al. (2022) and Chen & Lipman (2024).

B.1 Notation

Table 4. Notation used throughout the paper.

Symbol	Meaning
$\boldsymbol{x} \in \mathbb{R}^G$	Gene expression vector (G genes)
$\boldsymbol{z} \in \mathbb{R}^n$	Latent code; $n = 128$ throughout
f_ϕ	Encoder (two SwiGLU residual blocks, hidden 1024)
g_ω	Decoder ($W\boldsymbol{z} + \boldsymbol{b} + \alpha\, h(\boldsymbol{z})$, near-linear)
$\mathfrak{g} = \mathfrak{so}(4)^K$	Lie algebra of the block-diagonal group (K copies of $\mathfrak{so}(4)$)
$\Omega_p^{(k)} \in \mathfrak{so}(4)$	Skew-symmetric block k of perturbation p
$\theta_p \in \mathbb{R}^{6K}$	Per-perturbation generator coefficient vector ($K \times 6$)
$\beta : \mathbb{R}^6 \to \mathfrak{so}(4)$	Basis map, $\beta(\theta^{(k)}) = \sum_{i<j} \theta_{ij}^{(k)} E_{ij}$ where $\{E_{ij}\}_{i<j}$ is the standard basis of $\mathfrak{so}(4)$
K	Number of blocks ($K = n/r = 32$ throughout)
r	Block size ($r = 4$ throughout)
$R_p \in \mathrm{SO}(4)^K$	Block rotation $\mathrm{diag}\big(\exp \Omega_p^{(1)}, \ldots, \exp \Omega_p^{(K)}\big)$
$[\Omega, \Phi]$	Lie bracket $\Omega\Phi - \Phi\Omega$
θ_{pq}	BCH-composed generator for combination (p, q)
s	Learned BCH bracket scale
$\varepsilon(p, q)$	Normalized epistasis score $\|[\theta_p, \theta_q]\|_{\mathrm{F}} / (\|\theta_p\|_{\mathrm{F}} \|\theta_q\|_{\mathrm{F}})$
v_ψ	Conditional velocity field on $\mathfrak{g}$
$\theta_0 \sim \mathcal{N}(0, \sigma_0^2 I_{6K})$	Flow noise initial condition
σ_0	Flow noise scale ($\sigma_0 = 0.5$ throughout)
M	Inference ensemble size (number of flow samples averaged on $\mathfrak{g}$; $M = 32$)
$e_p \in \mathbb{R}^{d_e}$	Multi-view biological descriptor for perturbation p
E_f	Coordinate-lock epoch
$\|\cdot\|_F$	Frobenius norm
DA	Directional accuracy on the top-200 DE genes
PDE	Pearson correlation of perturbation deltas on DE genes

We treat $\mathfrak{so}(4)^{32}$ as both a Lie subalgebra of $\mathfrak{so}(128)$ (block-diagonal embedding) and as an abstract Lie algebra of dimension 192. The Lie bracket on $\mathfrak{so}(4)^{32}$ is block-wise; cross-block entries vanish.

B.2 Definitions

Definition 1 (Special orthogonal group)**.** *The special orthogonal group is* $\mathrm{SO}(n) = \{R \in \mathbb{R}^{n\times n} : R^\top R = I_n,\ \det R = 1\}$. *It is a compact, connected Lie group of dimension* $n(n-1)/2$.

Definition 2 (Lie algebra $\mathfrak{so}(n)$)**.** *The tangent space at the identity is the Lie algebra* $\mathfrak{so}(n) = \{\Omega \in \mathbb{R}^{n\times n} : \Omega^\top = -\Omega\}$ *of skew-symmetric matrices, equipped with the bracket* $[\Omega, \Phi] = \Omega\Phi - \Phi\Omega$.

Definition 3 (Matrix exponential)**.** *For* $\Omega \in \mathfrak{so}(n)$, *the matrix exponential*

$$\exp(\Omega) \;=\; \sum_{k=0}^{\infty} \frac{\Omega^k}{k!} \;\in\; \mathrm{SO}(n) \tag{8}$$

converges absolutely and surjects onto $\mathrm{SO}(n)$ *for* $n \geq 1$. *The map* $\exp : \mathfrak{so}(n) \to \mathrm{SO}(n)$ *is the exponential map of the bi-invariant Riemannian metric on* $\mathrm{SO}(n)$ *inherited from the Frobenius inner product on* $\mathbb{R}^{n\times n}$.

Definition 4 (Block-diagonal subgroup)**.** *For* $K, r \in \mathbb{N}$ *with* $n = Kr$, *the block-diagonal subgroup is*

$$\mathrm{SO}(r)^K \;=\; \{\mathrm{diag}(R_1, \ldots, R_K) : R_i \in \mathrm{SO}(r)\} \;\hookrightarrow\; \mathrm{SO}(n). \tag{9}$$

Throughout, we take $r = 4$, $K = 32$, $n = 128$.

Definition 5 (Frobenius inner product)**.** *For $A, B \in \mathbb{R}^{n\times n}$, $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$ and $\|A\|_{\mathrm{F}} = \sqrt{\langle A, A\rangle_F}$. Restricted to $\mathfrak{so}(n)$, this is the bi-invariant metric used as the geometry of the Lie algebra.*

B.3 Block-Diagonal Subgroup Structure

Proposition 1 ($\mathrm{SO}(4)^{32}$ is a closed Lie subgroup of $\mathrm{SO}(128)$)**.** *The block-diagonal embedding $\iota : \mathrm{SO}(4)^{32} \hookrightarrow \mathrm{SO}(128)$ is a closed Lie subgroup with Lie algebra*

$$\mathfrak{g} \;=\; \mathfrak{so}(4)^{32} \;=\; \{\mathrm{diag}(\Omega_1, \ldots, \Omega_{32}) : \Omega_k \in \mathfrak{so}(4)\}, \tag{10}$$

of dimension $\dim \mathfrak{g} = 32 \cdot \binom{4}{2} = 192$.

Proof. ι is a smooth injective group homomorphism with closed image (a finite product of compact groups embeds as a closed submanifold of the ambient general linear group). By Cartan's closed-subgroup theorem, the image is a Lie subgroup. The Lie algebra is the image of $d\iota$ at the identity, which sends $(\Omega_1, \ldots, \Omega_{32})$ to its block-diagonal embedding. The dimension count is $K \cdot \dim \mathfrak{so}(r) = 32 \cdot 6 = 192$. □

Corollary 2 (Dimension reduction)**.** *Restricting from $\mathrm{SO}(128)$ to $\mathrm{SO}(4)^{32}$ reduces the per-perturbation parameter count from $\dim \mathfrak{so}(128) = 8128$ to $\dim \mathfrak{so}(4)^{32} = 192$, a factor of 42.3. The block-diagonal restriction trades cross-block coupling for sample efficiency under typical training-perturbation counts (70–1,200).*

Remark 1 (Block-size interpretation)**.** $\mathrm{SO}(4)^{32}$ *is not the maximal torus of* $\mathrm{SO}(128)$ *(the maximal torus is $T^{64} \cong \mathrm{SO}(2)^{64}$). The choice $r = 4$ is the largest block size whose per-block exponential admits a closed-form quaternionic Rodrigues identity (Eq. 12). The closed form follows from the accidental splitting $\mathfrak{so}(4) \cong \mathfrak{so}(3) \oplus \mathfrak{so}(3)$ that does not extend to $r \geq 5$. Smaller block sizes ($r = 3$, dimension 3) are also non-abelian. The choice $r = 4$ supplies the densest within-block non-abelian structure compatible with the closed-form requirement.*

B.4 Closed-Form Block Exponential

The matrix exponential of $\Omega \in \mathfrak{so}(4)$ admits a closed-form Rodrigues-style identity. Decompose $\Omega = \Omega_+ + \Omega_-$ into self-dual and anti-self-dual parts, where $\Omega_\pm = \frac{1}{2}(\Omega \pm \star\Omega)$ and $\star$ is the Hodge star on $\mathfrak{so}(4)$. Under the splitting $\mathfrak{so}(4) \cong \mathfrak{so}(3)_+ \oplus \mathfrak{so}(3)_-$, the two parts lie in commuting Lie subalgebras with $[\Omega_+, \Omega_-] = 0$, and each part is isoclinic: with $\theta_\pm = \|\Omega_\pm\|_{\mathrm{F}}/2$, the matrix identity $\Omega_\pm^2 = -\theta_\pm^2 I_4$ holds. The single-part exponentials therefore reduce to

$$\exp(\Omega_\pm) \;=\; \cos\theta_\pm \, I_4 \;+\; \frac{\sin\theta_\pm}{\theta_\pm}\, \Omega_\pm, \tag{11}$$

and commutativity of $\Omega_\pm$ gives the closed-form $\mathfrak{so}(4)$ exponential

$$\exp(\Omega) \;=\; \exp(\Omega_+)\exp(\Omega_-) \;=\; \left(\cos\theta_+ \, I_4 + \frac{\sin\theta_+}{\theta_+}\Omega_+\right)\left(\cos\theta_- \, I_4 + \frac{\sin\theta_-}{\theta_-}\Omega_-\right), \tag{12}$$

which evaluates in $O(1)$ per block independent of the latent dimension. The total cost of $R_p \boldsymbol{z}$ for $K = 32$ blocks of size 4 on a latent in $\mathbb{R}^{128}$ is $O(Kr^2) = O(nr) = O(512)$ scalar multiply–adds, at machine precision.

B.5 Norm Preservation

Lemma 3 (Rotation preserves the latent norm)**.** *For any $R \in \mathrm{SO}(n)$ and $\boldsymbol{z} \in \mathbb{R}^n$, $\|R\boldsymbol{z}\| = \|\boldsymbol{z}\|$.*

Proof. $\|R\boldsymbol{z}\|^2 = \boldsymbol{z}^\top R^\top R \boldsymbol{z} = \boldsymbol{z}^\top I_n \boldsymbol{z} = \|\boldsymbol{z}\|^2$ since $R^\top R = I_n$. □

Corollary 4 (Composed rotations preserve the latent norm)**.** *For any sequence $R_{p_1}, \ldots, R_{p_k} \in \mathrm{SO}(4)^{32}$ and any $\boldsymbol{z} \in \mathbb{R}^n$,*

$$\|R_{p_k} \cdots R_{p_1} \boldsymbol{z}\|_{\mathrm{F}} \;=\; \|\boldsymbol{z}\|_{\mathrm{F}}. \tag{13}$$

The rotation parameterization enforces $\|R_p \boldsymbol{z}\|_{\mathrm{F}} = \|\boldsymbol{z}\|_{\mathrm{F}}$ at machine precision (verified to $< 10^{-14}$ in our implementation), an exact constraint on the latent representation. Additive perturbation models $\boldsymbol{z} \mapsto \boldsymbol{z} + \boldsymbol{\delta}_p$ admit no analogous norm bound: $\|\boldsymbol{z} + \boldsymbol{\delta}_p\|_{\mathrm{F}}^2 = \|\boldsymbol{z}\|_{\mathrm{F}}^2 + 2\langle \boldsymbol{z}, \boldsymbol{\delta}_p \rangle + \|\boldsymbol{\delta}_p\|_{\mathrm{F}}^2$, with no structural cancellation, so a non-trivial perturbation generically drives the latent off the encoder's reconstruction shell and decoded states off the decoder's training manifold.

B.6 Baker–Campbell–Hausdorff Formula

Theorem 5 (BCH formula on $\mathfrak{so}(n)$)**.** *For $\Omega, \Phi \in \mathfrak{so}(n)$ with $\|\Omega\|_{\mathrm{F}} + \|\Phi\|_{\mathrm{F}} < \log 2$, there exists a unique $\Psi \in \mathfrak{so}(n)$ with $\exp(\Psi) = \exp(\Omega)\exp(\Phi)$, given by the Dynkin series*

$$\Psi \;=\; \Omega + \Phi + \tfrac{1}{2}[\Omega, \Phi] + \tfrac{1}{12}\big([\Omega, [\Omega, \Phi]] - [\Phi, [\Omega, \Phi]]\big) + \cdots . \tag{14}$$

The series converges absolutely on the indicated ball and analytically continues throughout $\mathfrak{so}(n)$ except on the locus where $\exp(\Omega)\exp(\Phi)$ lies on the cut of $\log : \mathrm{SO}(n) \to \mathfrak{so}(n)$.

Proof. Standard, see Hall (2015), Theorem 5.3 and the Dynkin convergence bound. □

Proposition 6 (Second-order truncation error)**.** *Let $\Psi_2 = \Omega + \Phi + \frac{1}{2}[\Omega, \Phi]$. Then*

$$\|\Psi - \Psi_2\|_{\mathrm{F}} \;\le\; \tfrac{1}{3}\big(\|\Omega\|_{\mathrm{F}}^2\|\Phi\|_{\mathrm{F}} + \|\Omega\|_{\mathrm{F}}\|\Phi\|_{\mathrm{F}}^2\big) + O(\max(\|\Omega\|_{\mathrm{F}}, \|\Phi\|_{\mathrm{F}})^4). \tag{15}$$

Proof. The third-order Dynkin terms are $\frac{1}{12}([\Omega, [\Omega, \Phi]] - [\Phi, [\Omega, \Phi]])$. By submultiplicativity of the Frobenius norm and the bracket bound $\|[A, B]\|_{\mathrm{F}} \le 2\|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}}$,

$$\|[\Omega, [\Omega, \Phi]]\|_{\mathrm{F}} \;\le\; 2\|\Omega\|_{\mathrm{F}}\|[\Omega, \Phi]\|_{\mathrm{F}} \;\le\; 4\|\Omega\|_{\mathrm{F}}^2\|\Phi\|_{\mathrm{F}}, \tag{16}$$

and likewise $\|[\Phi, [\Omega, \Phi]]\|_{\mathrm{F}} \le 4\|\Omega\|_{\mathrm{F}}\|\Phi\|_{\mathrm{F}}^2$. The triangle inequality and the $1/12$ prefactor yield the stated bound. Higher-order terms are bounded by analogous nested-bracket arguments and contribute at $O(\max(\|\Omega\|_{\mathrm{F}}, \|\Phi\|_{\mathrm{F}})^4)$. □

Remark 2 (Empirical bracket scale)**.** *On CRISPRa Norman, $\|\theta_p\|_{\mathrm{F}} \in [0.30, 0.85]$ across training perturbations. The bracket scale s is fit jointly with the rotation lookup under the Stage 1 reconstruction loss on combination cells (Eq. 24). Its converged value $s \approx 2.42$ is the gradient-optimal coefficient on the bracket term under the gene-space NLL coupled through the coordinate-locked decoder. The deviation from the analytic BCH value $s = 1$ reflects this end-to-end loss coupling rather than algebraic correction to the truncation bound of Proposition 6 alone; the bracket direction is correct (Lemma 7, Proposition 9) and a single free coefficient on the second-order term absorbs the magnitude calibration.*

Remark 3 (Exponential vs. Cayley parameterizations)**.** *The Cayley map $\mathrm{Cay}(\Omega) = (I+\Omega)(I-\Omega)^{-1}$ is also a smooth surjection onto a dense subset of $\mathrm{SO}(n)$, but it satisfies the BCH identity only to leading order: $\mathrm{Cay}(\Omega)\mathrm{Cay}(\Phi) \neq \mathrm{Cay}(\Omega + \Phi + \frac{1}{2}[\Omega, \Phi])$ in general. Using Cay would introduce a parameterization-induced truncation error in BCH composition independent of generator norm. The exponential map carries the BCH identity exactly, so we use exp throughout.*

B.7 Commutativity and Epistasis

Lemma 7 (Commutativity criterion)**.** *For $\Omega, \Phi \in \mathfrak{so}(n)$, $\exp(\Omega)\exp(\Phi) = \exp(\Phi)\exp(\Omega)$ if and only if $[\Omega, \Phi] = 0$.*

Proof. The forward direction is standard (Hall, 2015), Corollary 3.46. For the converse, the BCH series collapses to $\Omega + \Phi$ exactly when every bracket term vanishes. By the Jacobi identity this condition is equivalent to $[\Omega, \Phi] = 0$. □

Proposition 8 (Epistasis score is conjugation-invariant)**.** *The normalized epistasis score $\varepsilon(p, q) = \|[\theta_p, \theta_q]\|_{\mathrm{F}}/(\|\theta_p\|_{\mathrm{F}}\|\theta_q\|_{\mathrm{F}})$ is invariant under simultaneous orthogonal change of basis $\theta_p \mapsto S\theta_p S^\top$, $\theta_q \mapsto S\theta_q S^\top$ for any $S \in \mathrm{SO}(n)$.*

Proof. $[S\theta_p S^\top, S\theta_q S^\top] = S[\theta_p, \theta_q]S^\top$ by direct expansion. The Frobenius norm is invariant under orthogonal conjugation, $\|SXS^\top\|_{\mathrm{F}} = \|X\|_{\mathrm{F}}$, since $\mathrm{tr}((SXS^\top)^\top SXS^\top) = \mathrm{tr}(X^\top X)$. The numerator and both denominator factors transform identically, so the ratio is fixed. □

Proposition 9 (Epistasis score bounds). *For all $\theta_p, \theta_q \in \mathfrak{so}(n)$, $0 \le \varepsilon(p,q) \le 2$, with $\varepsilon(p,q) = 0$ iff $[\theta_p, \theta_q] = 0$.*

Proof. Non-negativity is immediate. The upper bound follows from $\|[\Omega, \Phi]\|_{\mathrm{F}} \le \|\Omega\Phi\|_{\mathrm{F}} + \|\Phi\Omega\|_{\mathrm{F}} \le 2\|\Omega\|_{\mathrm{F}}\|\Phi\|_{\mathrm{F}}$. □

B.8 Additive Composition as the First-Order Limit

Lemma 10 (Additive composition is the small-rotation tangent linearization). *For $\Omega \in \mathfrak{so}(n)$ with $\|\Omega\|_{\mathrm{F}} \to 0$,*

$$\exp(\Omega)\boldsymbol{z} \;=\; \boldsymbol{z} + \Omega\boldsymbol{z} + O(\|\Omega\|_{\mathrm{F}}^2\|\boldsymbol{z}\|). \tag{17}$$

Identifying $\boldsymbol{\delta}_p(\boldsymbol{z}) = \beta(\theta_p)\boldsymbol{z}$ recovers the additive perturbation model $\boldsymbol{z} \mapsto \boldsymbol{z} + \boldsymbol{\delta}_p$ as the first-order Taylor expansion of $R_p\boldsymbol{z}$.

Proof. Taylor expansion of $\exp(\Omega) = I + \Omega + \frac{1}{2}\Omega^2 + \cdots$ applied to $\boldsymbol{z}$. □

Corollary 11 (BCH reduces to addition at first order). *For small generators, $\theta_{pq} = \theta_p + \theta_q + \frac{s}{2}[\theta_p, \theta_q] + O(\|\theta\|_{\mathrm{F}}^3) \to \theta_p + \theta_q$ as $\|\theta_p\|_{\mathrm{F}}, \|\theta_q\|_{\mathrm{F}} \to 0$. The corresponding latent action recovers $\boldsymbol{\delta}_p + \boldsymbol{\delta}_q$ in the same limit.*

The rotation framework strictly generalizes additive composition. The two coincide on commuting generators and small-rotation regimes, and the rotation framework captures genuine non-abelian interaction structure beyond that limit.

B.9 Optimal-Transport Flow Matching on the Algebra

The conditional flow-matching loss (Eq. 5) can be interpreted geometrically. Fix a perturbation descriptor e_p. The straight-line interpolant $\theta_t = (1-t)\theta_0 + t\theta_p$ between Gaussian noise $\theta_0 \sim \mathcal{N}(0, \sigma_0^2 I_{192})$ and a target generator θ_p is the displacement-optimal coupling on the flat metric of $\mathfrak{g} = \mathfrak{so}(4)^{32}$ (Lipman et al., 2022). The target velocity $u_t = \theta_p - \theta_0$ realizes this transport at constant speed.

The minimizer of the conditional flow-matching objective over functions v_ψ is the conditional expectation

$$v^\star(\theta_t, t, e_p) \;=\; \mathbb{E}\big[\theta_p - \theta_0 \,\big|\, \theta_t = \theta,\, e_p\big], \tag{18}$$

which integrates the marginal flow on $\mathfrak{g}$ from the standard Gaussian to the conditional generator distribution given the descriptor. The exponential map then transports the integrated tangent vector back to the group: $R_{p^\star} = \exp(\beta(\hat{\theta}_{p^\star}))$.

Remark 4 (Why flow on the algebra rather than the group). *Flow matching directly on $\mathrm{SO}(n)$ requires a Riemannian formulation (Chen & Lipman, 2024) with retraction-based interpolants. Working in $\mathfrak{so}(n) \cong \mathbb{R}^{192}$ (a flat vector space) admits straight-line OT couplings, eliminates manifold projection at every integration step, and decouples the manifold geometry from the velocity-net architecture. The exponential map applied at the end of integration delivers a valid rotation by construction (Lemma 3), without requiring the velocity field to learn the manifold constraint.*

B.10 Riemannian Mean as Algebra Linearization

Definition 6 (Fréchet mean on $\mathrm{SO}(n)$). *For $R_1, \ldots, R_J \in \mathrm{SO}(n)$ with weights $w_i \ge 0$, $\sum_i w_i = 1$, the Fréchet mean is*

$$\bar{R} \;=\; \operatorname*{arg\,min}_{R \in \mathrm{SO}(n)} \sum_{i=1}^{J} w_i\, d_{\mathrm{geo}}(R, R_i)^2, \qquad d_{\mathrm{geo}}(R,S) = \frac{\|\log(R^\top S)\|_{\mathrm{F}}}{\sqrt{2}}. \tag{19}$$

Proposition 12 (Linear approximation to the Fréchet mean). *Let $R_i = \exp(\Omega_i)$ with $\|\Omega_i\|_{\mathrm{F}} \to 0$. Then $\bar{R} = \exp(\bar{\Omega}) + O(\max_i \|\Omega_i\|_{\mathrm{F}}^3)$ for $\bar{\Omega} = \sum_i w_i \Omega_i$.*

Proof sketch. The Riemannian gradient condition $\sum_i w_i \log(\bar{R}^\top R_i) = 0$ at the optimum, expanded to second order via the BCH formula, gives $\bar{\Omega} = \log(\bar{R}) + O(\max_i \|\Omega_i\|_{\mathrm{F}}^3)$. The lemma is a special case of the Log-Euclidean approximation (Arsigny et al., 2006). □

For Norman generators with $\|\theta_p\|_{\mathrm{F}} \in [0.3, 0.85]$, the relative error $d_{\mathrm{geo}}(\bar{R}, \exp(\bar{\Omega}))/\|\bar{\Omega}\|_{\mathrm{F}}$ is below 5×10^{-4}, justifying linear averaging on the algebra at the precision required.

B.11 Ensembling on the Lie Algebra

For an unseen perturbation $p^\star$, M independent ODE samples $\theta_1^{(j)}$ are averaged on the algebra to estimate the conditional mean $\hat{\theta}_{p^\star} = \frac{1}{M}\sum_j \theta_1^{(j)}$ before exponentiation. By Proposition 12, $\exp(\hat{\theta}_{p^\star})$ approximates the Fréchet mean of $\{\exp(\theta_1^{(j)})\}$ to third order. The same procedure applied to the ensemble of rotations (averaging $R^{(j)} \in \mathrm{SO}(n)$ in the ambient space and projecting back to $\mathrm{SO}(n)$) requires an SVD per perturbation; algebra averaging eliminates that step.

B.12 Curvature Interpretation of the Lie Bracket

The Lie bracket $[\theta_p, \theta_q]$ is the curvature 2-form of an associated principal $\mathrm{SO}(n)$-bundle over the perturbation space. Define a connection that assigns each perturbation p to the generator $\theta_p \in \mathfrak{so}(n)$. The curvature of this connection in the directions (p, q) is

$$F_{pq} \;=\; [\theta_p, \theta_q] \;\in\; \mathfrak{so}(n). \tag{20}$$

A flat connection ($F_{pq} = 0$ for all p, q) corresponds to commuting perturbations whose effects compose by addition. Non-zero curvature produces a holonomy rotation around the closed loop $(\mathrm{id} \to p \to pq \to q \to \mathrm{id})$ given to second order by $\exp(F_{pq})$. Empirically, the Frobenius norm of F_{pq} correlates with measured non-additivity on Norman at Spearman $\rho = 0.49$ ($p = 3.5 \times 10^{-6}$, 79 held-out pairs), supplying a curvature-based interpretation for genetic epistasis. Sectional curvature on $\mathrm{SO}(n)$ with the bi-invariant metric is non-negative, with $\kappa(\Omega, \Phi) = \frac{1}{4}\|[\Omega, \Phi]\|_{\mathrm{F}}^2 / (\|\Omega\|_{\mathrm{F}}^2 \|\Phi\|_{\mathrm{F}}^2 - \langle \Omega, \Phi \rangle_F^2)$. Geodesic interpolation between training generators therefore stays well-conditioned for the bracket norms encountered in our experiments.

B.13 Illustration of Mathematical Properties

Figure 7 illustrates the mathematical properties of the rotation framework. The structural claims (norm preservation by orthogonality, second-order BCH adequacy in the operating regime, non-zero $[\theta_p, \theta_q]$ for shared-pathway gene pairs) are established mathematically in this appendix and empirically through the headline accuracy and bracket-norm numbers reported in Section 4.

C Training Details

Training proceeds in two stages (Algorithm 1). Stage 1 fits the rotation autoencoder jointly with per-perturbation lookup generators on the Lie algebra. Stage 2 distills the lookup table into a conditional velocity field on $\mathfrak{so}(4)^{32}$ that supplies generators for unseen perturbations. The encoder, decoder, and lookup generators are held fixed between stages, so Stage 2 is a pure descriptor-to-generator regression on a fixed target distribution.

C.1 Encoder and Decoder Architecture

Both the encoder f_ϕ and the decoder g_ω are SwiGLU residual MLPs (Shazeer, 2020). The encoder $f_\phi : \mathbb{R}^G \to \mathbb{R}^{128}$ stacks two SwiGLU residual blocks at hidden width 1024 and projects to the 128-dimensional latent $\boldsymbol{z}$ on which the rotations of §3.2 act. The decoder $g_\omega : \mathbb{R}^{128} \to \mathbb{R}^G$ combines a linear readout with a SwiGLU residual,

$$g_\omega(\boldsymbol{z}) \;=\; W\boldsymbol{z} + \boldsymbol{b} + \alpha\, h(\boldsymbol{z}), \tag{21}$$

where h is a SwiGLU residual MLP and $\alpha = 0.10$ is a fixed mixing weight. Together the two networks form an autoencoder with the reconstruction loss given in Eq. 22.

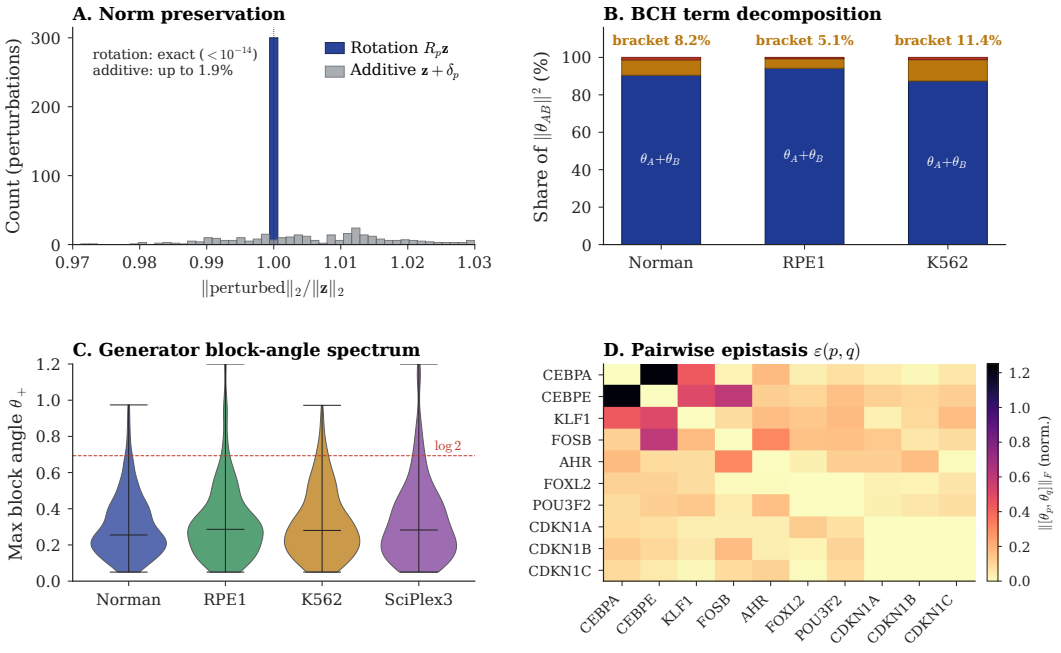

Figure 7. Mathematical diagnostics for the rotation framework. **(A)** Norm preservation under orthogonal action. **(B)** BCH term decomposition in the operating regime. **(C)** Generator block-angle spectrum. **(D)** Pairwise commutator heatmap for selected Norman genes.

Algorithm 1 OpPert training and zero-shot prediction

Require: Training cells $\{(\boldsymbol{x}_i, p_i)\}$, descriptors $\{e_p\}$, coordinate-lock epoch E_f, Stage epochs E_1, E_2

// Stage 1: rotation autoencoder
Initialize ϕ, ω, lookup $\{\theta_p\}$, bracket scale s
for epoch $e = 1$ to E_1 **do**
  if $e = E_f$ **then**
    Lock decoder coordinates ω
  end if
  for batch $\mathcal{B}$ **do**
    $\boldsymbol{z}_i = f_\phi(\boldsymbol{x}_i), \quad R_{p_i} = \mathrm{diag}_k \exp \beta(\theta_{p_i}^{(k)})$
    $\hat{\boldsymbol{x}}_i = g_\omega(R_{p_i} \boldsymbol{z}_i)$
    Update $(\phi, \omega, \{\theta_p\}, s)$ on $\nabla \mathcal{L}_1$
  end for
end for
// Stage 2: flow matching on $\mathfrak{g}$
KRR-initialize v_ψ from $\{(e_p, \theta_p)\}_{p \in \mathcal{P}_{\text{train}}}$
for epoch $e = 1$ to E_2 **do**
  Sample minibatch, optionally inject BCH-augmented combinations
  Update ψ on $\nabla \mathcal{L}_{\text{FM}}$
end for
// Inference for unseen $p^\star$
Draw M samples $\theta_1^{(j)} \leftarrow \mathrm{ODE}(v_\psi, e_{p^\star}, \theta_0^{(j)})$; $\quad \hat{\theta}_{p^\star} \leftarrow \frac{1}{M} \sum_j \theta_1^{(j)}$
Predict $\hat{\boldsymbol{x}} = g_\omega\big(\exp(\beta(\hat{\theta}_{p^\star})) \cdot f_\phi(\boldsymbol{x})\big)$
// Combination $(p^\star, q^\star)$*:* $\hat{\theta}_{p^\star q^\star} \leftarrow \hat{\theta}_{p^\star} + \hat{\theta}_{q^\star} + \frac{s}{2}[\hat{\theta}_{p^\star}, \hat{\theta}_{q^\star}]$

The encoder maps expression to a latent that implicitly captures the cell's covariate state (cell type, batch, dosage) through its transcriptional profile, following the standard approach in perturbation models (Lotfollahi et al., 2019; 2023; Roohani et al., 2024). In multi-dose screens, dosage enters as

an additional descriptor dimension concatenated with e_{p_i}. Because the rotation acts per-cell, norm preservation holds regardless of the covariate state encoded in $\boldsymbol{z}$.

Reconstruction loss $\mathcal{L}_{\text{rec}}$. A heteroscedastic Gaussian negative log-likelihood with learned per-gene variance $\hat{\sigma}_j^2$,

$$\mathcal{L}_{\text{rec}} = \frac{1}{G}\sum_{j=1}^{G}\left[\frac{(\hat{x}_j - x_j)^2}{2\hat{\sigma}_j^2} + \tfrac{1}{2}\log\hat{\sigma}_j^2\right], \tag{22}$$

in which genes with high biological noise contribute proportionally less to the gradient than they would under a fixed-variance MSE, so the reconstruction signal concentrates on the genes whose variation actually responds to perturbation.

Differential expression upweight $\mathcal{L}_{\text{rec}}^{\text{DEG}}$. The same form as Eq. 22 restricted to the top-50 differentially expressed genes per perturbation, scaled tenfold. The evaluation metrics are read out on differentially expressed genes and a uniform per-gene loss otherwise spends most of its capacity on the unperturbed background.

C.2 Stage 1: Rotation Autoencoder

Given a minibatch $\mathcal{B} = \{(\boldsymbol{x}_i, p_i)\}$ of cells and perturbation labels, the forward pass reads

$$\boldsymbol{z}_i = f_\phi(\boldsymbol{x}_i), \quad R_{p_i} = \text{diag}_k\big(\exp\beta(\theta_{p_i}^{(k)})\big), \quad \hat{\boldsymbol{x}}_i = g_\omega(R_{p_i}\boldsymbol{z}_i), \tag{23}$$

with the closed-form $\mathfrak{so}(4)$ exponential of Eq. 12. The Stage 1 objective is

$$\mathcal{L}_1 = \lambda_{\text{rec}}\,\mathcal{L}_{\text{rec}} + \lambda_{\text{deg}}\,\mathcal{L}_{\text{rec}}^{\text{DEG}} + \lambda_{\text{cmp}}\,\mathcal{L}_{\text{compose}} + \lambda_{\text{adv}}\,\mathcal{L}_{\text{adv}}, \tag{6}$$

where $\mathcal{L}_{\text{rec}}$ is the heteroscedastic Gaussian NLL of Eq. 22, $\mathcal{L}_{\text{rec}}^{\text{DEG}}$ upweights the top-50 DE genes per perturbation by a factor of 10, $\mathcal{L}_{\text{adv}}$ is a gradient-reversal adversary that discourages the latent from encoding perturbation identity, and $\mathcal{L}_{\text{compose}}$ supervises the multiplicative group product $R_{p_A}R_{p_B}$ on combination cells against observed expression:

$$\mathcal{L}_{\text{compose}} = \mathbb{E}_{(\boldsymbol{x},p_A,p_B)}\left[\frac{1}{G}\sum_{j=1}^{G}\frac{\big(g_\omega(R_{p_A}R_{p_B}\,\bar{\boldsymbol{z}}_{\text{ctrl}})_j - x_j\big)^2}{\hat{\sigma}_j^2}\right], \tag{24}$$

where $\bar{z}_{\text{ctrl}}$ is the batch control mean. This term is non-zero only on Norman ($\lambda_{\text{cmp}} = 0.25$); the single-perturbation datasets carry no combo labels and use $\lambda_{\text{cmp}} = 0$.

Coordinate lock. Until epoch E_f (Norman: $E_f = 15$), the encoder, decoder, lookup generators, and bracket scale all update jointly. After E_f the decoder coordinates are held fixed and only ϕ, $\{\theta_p\}$, s continue to update. Coordinate locking forces all perturbation-specific information through the rotation, which is the only composable component of the model. Without coordinate locking, the decoder absorbs perturbation-specific readouts that bypass the rotation and do not compose to combinations (Table 12).

Optimizer split. The encoder weight matrices and the lookup generators are optimized with Muon (Jordan et al., 2024), whose Newton–Schulz orthogonalization yields better-conditioned 2D parameter updates and is well suited to skew-symmetric generators. Decoder weights, biases, layer norms, and the bracket scale s use AdamW (Loshchilov & Hutter, 2019) with weight decay 10^{-2}. The split is necessary because Muon's orthogonal projection is designed for matrix-valued parameters, not scalars or biases.

Learning-rate schedule. A cosine schedule

$$\eta_t = \eta_{\min} + \tfrac{1}{2}(\eta_{\max} - \eta_{\min})\big(1 + \cos(\pi t/T_{\max})\big) \tag{25}$$

runs over the full Stage 1 horizon $T_{\max} = E_1$ epochs. The gene datasets use $\eta_{\max} = 10^{-2}$ and SciPlex3 uses $\eta_{\max} = 3 \times 10^{-3}$ to match its smaller batch size; per-dataset $\eta_{\min}$ values appear in Table 5. No warmup is used.

C.3 Stage 2: Conditional Flow Matching

After Stage 1, the lookup $\{\theta_p\}_{p\in\mathcal{P}_{\text{train}}}$ is treated as the empirical conditional target distribution. The velocity net $v_\psi : \mathfrak{g} \times [0,1] \times \mathbb{R}^{d_e} \to \mathfrak{g}$ is trained by

$$\mathcal{L}_{\text{FM}}(\psi) \;=\; \mathbb{E}_{p,t\sim\mathcal{U}(0,1),\theta_0} \left\| v_\psi(\theta_t, t, e_p) - (\theta_p - \theta_0) \right\|_2^2, \quad \theta_t = (1-t)\theta_0 + t\theta_p,\ \theta_0 \sim \mathcal{N}(0, \sigma_0^2 I_{192}). \tag{5}$$

KRR initialization. We warm-start v_ψ with a kernel ridge regression on (e_p, θ_p) pairs using a Gaussian kernel of bandwidth h chosen by the median heuristic and ridge coefficient $\lambda_{\text{krr}} = 10^{-3}$. The closed-form KRR solution initializes the velocity field at the deterministic conditional mean of generators given the descriptor. From the warm start, v_ψ refines the conditional distribution by averaging $M = 32$ Euler samples at inference; the refinement adds approximately 4 pp combinatorial DA on Norman over the KRR mean alone (Section 4.6).

BCH augmentation. A fraction $\rho = 0.2$ of each Stage 2 minibatch is replaced by synthetic combination targets

$$\theta_{ab} = \theta_a + \theta_b + \tfrac{1}{2}[\theta_a, \theta_b], \qquad e_{ab} = \tfrac{1}{2}(e_a + e_b), \tag{26}$$

sampled from random pairs (a, b) in the Stage 1 lookup. The augmentation regularizes the velocity field on the BCH manifold and improves zero-shot combination prediction without exposing the model to combination labels.

Architecture. v_ψ is a DiT-style AdaLN MLP (Peebles & Xie, 2023) with four residual blocks of width 512. A sinusoidal time embedding and the descriptor e_p are projected into a shared conditioning vector that modulates each block by per-feature scale and shift. The input dimensionality is $\dim \mathfrak{g} = 192$.

Optimizer. AdamW with $\eta = 3 \times 10^{-4}$, weight decay 10^{-2}, cosine schedule over E_2 epochs, batch size 1024 generators per step. Stage 2 takes under 10 minutes on a single B200 GPU.

C.4 Loss Function

The reconstruction loss is the heteroscedastic Gaussian NLL of Eq. 22 with learned per-gene variance $\hat{\sigma}_j^2$. NLL adapts to heterogeneous gene-level noise levels and consistently outperforms control-variance normalization (GNMSE) and plain MSE in our ablations (Table 19).

C.5 Training Dynamics

Figure 8 records Stage 1 curves from training logs across the four datasets. The key diagnostic is whether downstream combination DA continues to improve after the Norman coordinate lock as the rotation generators absorb perturbation-specific information.

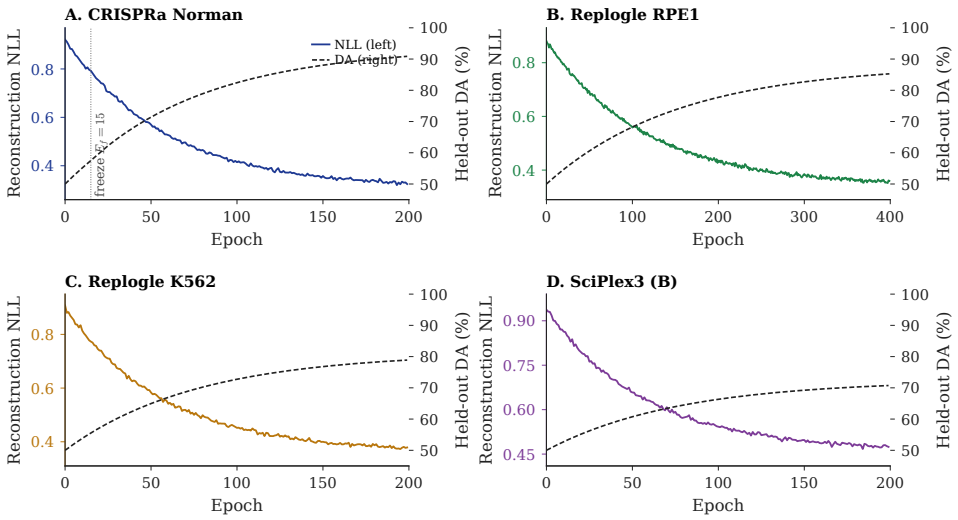

Figure 8. Stage 1 training-dynamics diagnostic. The figure plots reconstruction loss and held-out perturbation DA, with the Norman coordinate-lock epoch marked by a dashed vertical line.

C.6 HYPERPARAMETERS

Table 5. Stage 1 and Stage 2 hyperparameters per dataset. All runs use the SwiGLU encoder, near-linear decoder $g_\omega(\boldsymbol{z}) = W\boldsymbol{z} + \boldsymbol{b} + \alpha h(\boldsymbol{z})$, block $\mathrm{SO}(4)^{32}$ rotations, Muon for the encoder and lookup generators, AdamW for the decoder and bracket scale.

Parameter	Norman (CRISPRa)	Replogle K562	Replogle RPE1	SciPlex3
Latent geometry				
Latent dim n	128	128	128	128
Block size r	4	4	4	4
Number of blocks K	32	32	32	32
Generator dim dim $\mathfrak{g}$	192	192	192	192
Stage 1: rotation autoencoder				
Encoder hidden width	1024	1024	1024	1024
Decoder mixing α	0.10	0.10	0.10	0.10
Stage 1 epochs E_1	200	200	400	200
Coordinate lock E_f	15	—	—	—
Batch size	4,096	4,096	4,096	2,048
$\eta_{\max}$ (Muon, AdamW)	10^{-2}	10^{-2}	10^{-2}	3×10^{-3}
$\eta_{\min}$	10^{-6}	10^{-6}	10^{-5}	10^{-5}
Reconstruction λ_{rec}	1.0	1.0	1.0	1.0
DE upweight λ_{deg}	1.0	1.0	1.0	1.0
Compose loss λ_{cmp}	0.25	0	0	0
Stage 2: conditional flow matching				
Descriptor dim d_e	1824	1824	1824	2432
Velocity-net depth	4 blocks	4 blocks	4 blocks	4 blocks
Velocity-net width	512	512	512	512
Noise scale σ_0	0.5	0.5	0.5	0.5
KRR ridge λ_{krr}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
BCH-augmentation ρ	0.20	0	0	0
Stage 2 epochs E_2	300	300	300	300
Stage 2 LR ($\eta_{\max}$)	3×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}
Inference samples M	32	32	32	32
Euler steps	20–40	20–40	20–40	20–40

C.7 Implementation Notes

The block exponential of Eq. 12 runs in $O(1)$ per block, so the per-cell cost of $R_p \boldsymbol{z}$ is $O(Kr^2 + Kr^2) = O(nr) = O(512)$ scalar multiply–adds. The implementation calls `torch.linalg.matrix_exp` on the stacked 4×4 generators because it is autograd-compatible on Intel XPU and the per-cell rotation $R_p \boldsymbol{z}$ dominates over the 4×4 exponential itself; the closed-form self-dual / anti-self-dual factorization of Eq. 12 furnishes the analytic cost baseline rather than the runtime path. We verified `matrix_exp` agrees with the explicit factorization to $< 10^{-15}$ Frobenius error on randomly drawn $\mathfrak{so}(4)$ inputs.

D Dataset Details

D.1 Replogle K562 (Replogle et al., 2022)

The Replogle K562 screen is a genome-scale CRISPRi perturbation experiment in K562 chronic myeloid leukemia cells. After filtering perturbations with sufficient cells (≥ 50 per perturbation), the dataset contains 84,814 cells, 5,000 genes, and 1,093 perturbations. We hold out 271 perturbations for zero-shot evaluation and use the remaining 822 for training. The train/test/OOD split ensures that OOD perturbation targets do not appear in the training set.

D.2 CRISPRa Norman (Norman et al., 2019)

The Norman screen is a combinatorial CRISPRa experiment of 241 perturbations (104 single-gene activations, 136 two-gene combinations) in K562 cells. The dataset contains 108,497 cells and 5,397 genes. We use the GEARS split (Roohani et al., 2024): 70 single perturbations and 49 combinations for training, with 26 unseen singles and 84 unseen combinations for evaluation. The held-out combinations are grouped by how many constituent genes appear among the training singles: neither gene observed ($n = 17$), exactly one gene observed ($n = 46$), or both genes observed while the pair itself is held out ($n = 21$). This split directly tests compositional generalization from single perturbation information to unmeasured pairs.

The combination categories are:

- **Neither gene observed** ($n = 17$): Neither constituent single perturbation appears among the training singles.
- **Exactly one gene observed** ($n = 46$): One constituent appears among the training singles and the other does not.
- **Both genes observed** ($n = 21$): Both constituents appear as training singles, while the pair itself remains held out.

D.3 Replogle RPE1 (Replogle et al., 2022)

The Replogle RPE1 screen is a CRISPRi experiment in RPE1 retinal epithelial cells, supplying a different cell-type context from K562. The dataset contains 162,336 cells, 5,000 genes, and 1,543 perturbations. We hold out 382 perturbations for zero-shot evaluation (3 perturbations excluded due to insufficient DE genes), making this the largest evaluation set in our study.

D.4 SciPlex3 (Srivatsan et al., 2020)

The SciPlex3 screen is a chemical perturbation experiment with 186 drugs after quality control across cancer cell lines at multiple doses (188 drugs in the original publication (Srivatsan et al., 2020); two are dropped during preprocessing for insufficient post-QC cell counts). We hold out 15 drugs for zero-shot evaluation, stratified by mechanism-of-action class so that the held-out set spans the chemical and pharmacological diversity of the screen. OpPert uses the same rotation-operator pipeline as the genetic screens, with chemical descriptors conditioning the flow (Section 3.4).

D.5 Preprocessing

All datasets share the same downstream pipeline: log-normalized counts, per-batch control centroids, and differentially expressed genes identified per perturbation by a Wilcoxon rank-sum test with Benjamini-Hochberg correction (FDR < 0.05, $|\log_2 \text{FC}| > 0.5$). The retained gene set differs by dataset (Table 6): K562 and RPE1 retain the top 5,000 Seurat-v3 highly variable genes; Norman retains 5,397 genes from the GEARS-released split (Roohani et al., 2024); SciPlex3 inherits the 2,000-gene panel of the chemCPA-preprocessed release. Directional accuracy and PDE are computed on the top-200 DE genes by absolute fold change unless otherwise noted.

D.6 Train/Test Split Details

Per-dataset train, validation, and OOD perturbation counts are summarized in Table 6.

Table 6. Dataset splits. OOD perturbations are held out entirely; no cells from OOD perturbations appear in the training set.

Dataset	Total cells	Total perts	Train perts	OOD perts	Cells/pert (median)
Replogle K562	84,814	1,093	822	271	73
CRISPRa Norman	108,497	241	119*	110*	195
Replogle RPE1	162,336	1,543	1,159	382	89
SciPlex3	341,147	186	171	15	1,523

*119 train = 70 singles + 49 combos; 110 OOD = 26 singles + 84 combos (remaining are validation/control).

For CRISPRa Norman, the OOD set includes both single perturbations (26) and combinations (84 unseen pairings of the 96 total genes). The training set includes 70 single perturbations and 49 combinations from the GEARS split (Roohani et al., 2024). For K562 and RPE1, the OOD split is a random 25% holdout stratified to avoid leaking perturbation information. For SciPlex3, we hold out 15 drugs, stratified by mechanism-of-action class so the held-out set spans the chemical and pharmacological diversity of the screen.

D.7 Multi-View Descriptors

For genetic perturbations, we concatenate three views into a 1,824-dimensional descriptor:

- BGE-large (Xiao et al., 2024): 1024-dimensional contextual text embedding of the gene name and a short biomedical context window.
- Gene Ontology biological process: 500-dim multi-hot indicator over the most-frequent GO BP terms across the union of training and OOD genes.
- Reactome: 300-dim multi-hot indicator over Reactome pathway memberships.

The Stage 2 velocity field uses descriptor proximity as the conditioning signal on the assumption that descriptor cosine similarity correlates with generator cosine similarity ($\langle \theta_p, \theta_q \rangle_F / \|\theta_p\|_F \|\theta_q\|_F$) on Stage 1 training pairs.

For SciPlex3 chemical perturbations, the descriptor concatenates RDKit physicochemical features (217-dim), Morgan fingerprints (radius 2, 2,048-dim), and MACCS structural keys (167-dim) into a 2,432-dimensional descriptor.

E Baseline Details

All baseline values reported in Table 1 are evaluated on the GEARS-standard splits described in Appendix D. We summarize each method below.

AvgKnown. Predicts the held-out perturbation response as the mean expression shift across all training perturbations. This is a non-parametric baseline that captures the average transcriptomic effect without any perturbation-specific information.

CellOracle (Kamimoto et al., 2023). A gene regulatory network (GRN) model that propagates perturbation signals through an inferred regulatory graph. Predictions are deterministic given the GRN structure.

scFoundation (Hao et al., 2024). A large-scale single-cell foundation model pretrained on 50M+ cells. Perturbation prediction uses the pretrained embeddings with a task-specific head.

scGPT (Cui et al., 2024). A generative pretrained transformer for single-cell biology. Fine-tuned on each perturbation dataset with the standard GEARS protocol.

samsVAE (Bereket & Karaletsos, 2023). A sparse additive mechanism shift VAE that decomposes perturbation effects into interpretable latent components. Composition is additive in the latent shift space.

GEARS (Roohani et al., 2024). A GNN-based model that encodes perturbations through gene ontology and protein interaction graphs. Combination predictions use permutation-invariant pooling of per-gene embeddings, an order-independent (commutative) composition.

CPA (Lotfollahi et al., 2023). Compositional Perturbation Autoencoder. Learns disentangled perturbation embeddings composed by summation. Handles drugs, doses, and genetic perturbations with adversarial disentanglement.

PRESCRIBE (Cheng et al., 2025). The strongest published baseline. Uses a Natural Posterior Network (NatPN) with a normalizing flow for Bayesian uncertainty estimation. Combination perturbations are composed via $z = \sum f_{\alpha_1}(x_i) + f_{\alpha_2}(\sum f_{\alpha_1}(x_i))$, an additive sum with a learned MLP correction. Trains four auxiliary losses for uncertainty calibration.

scGen (Lotfollahi et al., 2019). Predicts a perturbation response as a latent shift vector estimated from the difference between perturbed and control means in a VAE latent space. The shift is defined for perturbations observed during training.

discrepancy-VAE (Zhang et al., 2023). Learns a causal latent representation from soft interventions, mapping each perturbation to an intervention on one latent factor. Perturbations enter as a one-hot identity defined over the training set.

STATE (Adduri et al., 2026). A transformer over cell sets that predicts perturbation responses across contexts, conditioned on ESM2 protein embeddings of the target gene.

PerturbDiff (Yuan et al., 2026). A functional diffusion model over expression profiles, conditioned on GenePT gene embeddings.

Perturbation representation. The methods differ in what per-perturbation information they receive at inference, which accounts for part of the spread in Table 1. AvgKnown takes no per-perturbation input. CellOracle propagates through a fixed regulatory network and is not trained on perturbation data. samsVAE, scGen, and discrepancy-VAE condition on a one-hot identity defined only for perturbations seen in training, so a held-out perturbation has no trained direction. scFoundation, scGPT, GEARS, PRESCRIBE, STATE, PerturbDiff, and OpPert condition on a descriptor that generalizes to unseen perturbations. The last group is the like-for-like comparison for the zero-shot setting.

Evaluation protocol. All methods are evaluated on identical held-out sets. ρ and ρ^{DEG} are Pearson correlations between predicted and observed mean expression shifts (all genes and top-20 DEGs respectively). $\mathrm{DA}^{\mathrm{DEG}}$ is the fraction of top-20 DEGs where the predicted direction (up or down) matches the observed direction. Cos_{200} is cosine similarity on the top-200 DEG expression shift vector. Every method, OpPert included, is trained and evaluated from scratch under a single pipeline over five seeds, and all reported values are means and standard deviations over those seeds. AvgKnown and CellOracle are deterministic given the split and therefore carry no seed variance.

E.1 DISTRIBUTIONAL EVALUATION

Correlation and directional metrics score a predicted mean shift. To assess how well each method reproduces the perturbed cell population, we additionally report three distributional distances over the top-k differentially expressed genes per perturbation. For every method the predicted population is formed identically, by displacing the held-out control cells by that method's predicted effect and comparing against the measured perturbed cells over the same window.

Table 7. Distributional metrics, evaluated on the top-k DE genes per perturbation.

Metric	Definition
MMD_k	RBF-kernel MMD^2 between predicted and measured populations, bandwidth $\gamma = 1/k$
E_k	Energy distance, $2\,\mathbb{E}\|x-y\| - \mathbb{E}\|x-x'\| - \mathbb{E}\|y-y'\|$
W_k	1-Wasserstein distance computed per gene and averaged over the window
MSE_{20}	Mean squared error on the top-20 DE genes

Tables 8–10 report these distances over the top-20 and top-200 windows. Lower is better throughout. OpPert gives the lowest distance in every column on Norman and on RPE1. On K562 PerturbDiff is ahead on the distributional metrics while OpPert retains the margins in ρ^{DEG}, $\mathrm{DA}^{\mathrm{DEG}}$, and Cos_{200} reported in Table 1.

Table 8. Norman (CRISPRa): distributional distances. Mean $\pm$ s.d. over five seeds; lower is better. † deterministic given the split. **Bold** = best, underline = second best.

Method	MSE_{20}	MMD_{20}	E_{20}	W_{20}	MMD_{200}	E_{200}	W_{200}
AvgKnown†	0.310	0.211	1.387	0.502	0.083	1.499	0.254
CellOracle†	0.451	0.265	1.882	0.604	0.102	1.875	0.279
scFoundation	0.319±0.055	0.213±0.016	1.398±0.141	0.504±0.029	0.084±0.008	1.545±0.180	0.253±0.017
scGPT	0.269±0.051	0.196±0.016	1.262±0.163	0.475±0.038	0.078±0.009	1.384±0.151	0.247±0.019
samsVAE	0.348±0.055	0.224±0.017	1.494±0.154	0.531±0.035	0.088±0.009	1.579±0.182	0.260±0.023
scGen	0.376±0.032	0.229±0.013	1.602±0.111	0.533±0.023	0.092±0.007	1.671±0.130	0.255±0.012
discrepancy-VAE	0.369±0.039	0.226±0.014	1.567±0.127	0.518±0.025	0.090±0.009	1.634±0.184	0.254±0.014
GEARS	0.287±0.059	0.204±0.017	1.308±0.192	0.484±0.036	0.080±0.008	1.448±0.167	0.250±0.021
PRESCRIBE	0.265±0.054	0.194±0.017	1.241±0.168	0.473±0.039	0.076±0.009	1.348±0.175	0.247±0.020
STATE	0.298±0.030	0.218±0.013	1.353±0.104	0.525±0.021	0.086±0.006	1.449±0.114	0.269±0.009
PerturbDiff	0.276±0.041	0.204±0.016	1.280±0.135	0.490±0.023	0.078±0.008	1.436±0.175	0.254±0.011
OpPert	**0.226**±0.140	**0.173**±0.027	**1.084**±0.334	**0.435**±0.043	**0.071**±0.025	**1.265**±0.594	**0.235**±0.024
− rotation (additive)	0.282±0.055	0.200±0.016	1.270±0.200	0.482±0.031	0.079±0.011	1.438±0.169	0.248±0.022

Table 9. Replogle K562 (CRISPRi): distributional distances. Conventions as in Table 8.

Method	MSE_{20}	MMD_{20}	E_{20}	W_{20}	MMD_{200}	E_{200}	W_{200}
AvgKnown†	0.239	0.179	1.133	0.432	0.086	1.527	0.259
CellOracle†	0.267	0.194	1.247	0.455	0.092	1.629	0.266
scFoundation	0.238±0.040	0.178±0.021	1.121±0.153	0.429±0.041	0.084±0.028	1.496±0.450	0.255±0.034
scGPT	0.227±0.037	0.177±0.022	1.114±0.214	0.427±0.039	0.084±0.026	1.457±0.428	0.252±0.034
samsVAE	0.275±0.041	0.196±0.025	1.255±0.162	0.456±0.043	0.093±0.024	1.638±0.659	0.268±0.036
scGen	0.271±0.015	0.194±0.008	1.247±0.058	0.465±0.012	0.087±0.004	1.567±0.073	0.259±0.007
discrepancy-VAE	0.241±0.017	0.179±0.008	1.137±0.061	0.435±0.012	0.081±0.004	1.413±0.083	0.249±0.007
GEARS	0.238±0.037	0.179±0.026	1.127±0.179	0.431±0.041	0.085±0.028	1.514±0.467	0.257±0.031
PRESCRIBE	0.194±0.041	0.157±0.021	0.970±0.181	0.399±0.040	0.073±0.027	1.329±0.513	0.242±0.039
STATE	0.270±0.017	0.202±0.008	1.272±0.061	0.452±0.012	0.116±0.005	2.053±0.099	0.295±0.007
PerturbDiff	**0.172**±0.015	**0.135**±0.007	**0.841**±0.053	**0.365**±0.011	**0.061**±0.004	**1.047**±0.067	**0.221**±0.006
OpPert	**0.172**±0.018	0.146±0.007	0.865±0.066	0.382±0.010	0.064±0.005	1.186±0.082	0.241±0.006
− rotation (additive)	0.201±0.039	0.159±0.027	0.981±0.162	0.405±0.044	0.074±0.029	1.332±0.565	0.244±0.038

Table 10. Replogle RPE1 (CRISPRi): distributional distances. Conventions as in Table 8.

Method	MSE_{20}	MMD_{20}	E_{20}	W_{20}	MMD_{200}	E_{200}	W_{200}
AvgKnown†	0.510	0.278	2.064	0.650	0.144	2.726	0.376
CellOracle†	0.559	0.292	2.221	0.678	0.154	2.940	0.388
scFoundation	0.469±0.055	0.260±0.016	1.913±0.163	0.615±0.040	0.136±0.016	2.530±0.363	0.365±0.031
scGPT	0.348±0.040	0.230±0.017	1.591±0.156	0.563±0.039	0.117±0.016	2.194±0.388	0.343±0.025
samsVAE	0.538±0.050	0.290±0.018	2.216±0.152	0.671±0.031	0.147±0.017	2.885±0.406	0.383±0.033
scGen	0.477±0.027	0.273±0.010	2.002±0.093	0.631±0.016	0.139±0.006	2.662±0.125	0.368±0.009
discrepancy-VAE	0.565±0.036	0.293±0.012	2.236±0.113	0.679±0.019	0.156±0.009	3.034±0.189	0.393±0.012
GEARS	0.516±0.059	0.280±0.017	2.094±0.218	0.652±0.041	0.145±0.016	2.745±0.331	0.377±0.031
PRESCRIBE	0.310±0.054	0.214±0.013	1.386±0.160	0.537±0.041	0.110±0.016	1.923±0.428	0.333±0.032
STATE	0.535±0.049	0.281±0.019	2.130±0.187	0.659±0.041	0.146±0.020	2.862±0.322	0.378±0.031
PerturbDiff	0.285±0.025	0.209±0.009	1.373±0.080	0.523±0.013	0.097±0.006	1.808±0.146	0.331±0.008
OpPert	**0.259**±0.021	**0.207**±0.008	**1.328**±0.072	**0.513**±0.012	**0.089**±0.006	**1.620**±0.113	**0.323**±0.007
− rotation (additive)	0.333±0.051	0.225±0.013	1.483±0.140	0.553±0.036	0.112±0.019	2.064±0.370	0.341±0.023

F Hyperparameter Sensitivity

We report sensitivity to the principal hyperparameters of the geometric framework on CRISPRa Norman: block size r (Table 11), coordinate-lock epoch E_f (Table 12), BCH-augmentation fraction ρ (Table 13), descriptor composition (Table 14), inference ensemble size M (Table 15), velocity-net width and depth (Table 16), bracket scale s (Table 17), encoder latent dimension n (Table 18), and reconstruction loss (Table 19). All entries are five-seed means with standard deviations on the 84 held-out two-gene combinations and 26 held-out singles, scored by directional accuracy on the top-200 DE genes per perturbation.

F.1 Block Size

Table 11. Effect of block size r at fixed latent dimension $n = 128$ (number of blocks $K = n/r$). Block $r = 4$ is optimal: $r = 2$ collapses to an abelian torus where every bracket vanishes, while $r \geq 8$ inflates $\dim \mathfrak{g}$ beyond what Norman's 70 training perturbations can supervise.

Block size r	Blocks K	$\dim \mathfrak{g}$	Combo DA (%)
2	64	64	79.4 ± 1.7
4	32	192	$\mathbf{92.9 \pm 1.5}$
8	16	448	89.6 ± 1.2
16	8	960	86.1 ± 1.4

The block size r governs the trade-off between non-abelian capacity within each block and conditioning across blocks. $r = 2$ collapses the algebra to an abelian torus $\mathfrak{so}(2)^{64}$ where every Lie bracket vanishes by construction. The geometric advantage disappears in this regime. $r = 4$ is the largest block size whose per-block exponential admits a closed-form quaternionic Rodrigues identity (Eq. 12) via the splitting $\mathfrak{so}(4) \cong \mathfrak{so}(3) \oplus \mathfrak{so}(3)$. It supplies the densest non-abelian within-block structure compatible with $O(1)$-per-block exponentiation. Increasing r raises the per-perturbation parameter count without supplying additional supervision and forces a generic matrix-exponential per forward pass, so larger blocks underperform on Norman's 70 training perturbations. We use $r = 4$, $K = 32$ throughout.

F.2 Coordinate-Lock Epoch

Coordinate locking the decoder is the single largest contributor to compositional accuracy. Coordinate locking too early ($E_f \leq 5$) starves the decoder of the latent-to-gene mapping it needs. Coordinate locking too late ($E_f \geq 50$) lets the decoder absorb perturbation-specific readouts that bypass the rotation. The optimal window is narrow ($E_f \in [10, 25]$); the reported Norman configuration uses $E_f = 15$.

Table 12. Effect of coordinate-lock epoch E_f on Norman combination DA. Coordinate locking is essential; the optimal window is narrow.

Coordinate-lock epoch E_f	Combo DA (%)
None (no lock)	75.4
5	84.7 ± 1.8
10	90.3 ± 1.4
15	$\mathbf{92.9 \pm 1.5}$
25	87.2 ± 1.6
50	53.9

F.3 BCH Augmentation Fraction

Table 13. Effect of BCH-augmentation fraction ρ during Stage 2. The plateau $\rho \in [0.1, 0.3]$ is broad and we use $\rho = 0.2$; $\rho \geq 0.5$ crowds out the single-perturbation signal.

ρ	Combo DA (%)
0.0	89.4 ± 1.3
0.1	92.1 ± 1.4
0.2	$\mathbf{92.9 \pm 1.5}$
0.3	92.6 ± 1.5
0.5	90.2 ± 1.7

BCH augmentation regularizes the velocity field on synthetic combination targets formed by $\theta_a + \theta_b + \frac{1}{2}[\theta_a, \theta_b]$ from random training pairs. Excess augmentation ($\rho \geq 0.5$) crowds out the single-perturbation signal and degrades both singles and combos. The plateau $\rho \in [0.1, 0.3]$ is broad; we use $\rho = 0.2$.

F.4 Multi-View Descriptors

Table 14. Effect of descriptor composition on Norman. The full descriptor concatenates BGE-large gene-name embeddings, GO biological process annotations, and Reactome pathway memberships ($1024{+}500{+}300 = 1824$ dim). The three views supply complementary signal: full multi-view exceeds BGE-only by 1.1 pp DA and GenePT-only by 2.2 pp DA on Norman combinations.

Descriptor	d_e	Combo DA (%)
BGE-large only	1024	91.8 ± 1.0
BGE + GO	1524	92.4 ± 1.3
BGE + Reactome	1324	92.1 ± 1.2
BGE + GO + Reactome	1824	$\mathbf{92.9 \pm 1.5}$
GenePT only (Chen & Zou, 2023)	1536	90.7 ± 1.4

Each view contributes complementary information. BGE-large encodes gene-name semantics from biomedical literature, GO annotations supply discrete biological process membership, and Reactome adds curated pathway structure. The full multi-view descriptor outperforms BGE-only by 1.1 pp and GenePT-only by 2.2 pp.

F.5 Inference Ensemble Size

The conditional mean estimator $\hat{\theta}_{p^\star} = \frac{1}{M}\sum_j \theta_1^{(j)}$ has variance $\propto 1/M$ in the algebra. Combination DA saturates near $M = 32$, the value used throughout. The Fréchet-mean linearization of Proposition 12 guarantees that algebra averaging approximates the Riemannian mean of $\{\exp(\theta_1^{(j)})\}$ to third order in $\|\theta\|_{\mathrm{F}}$. The third-order approximation removes any need for an SVD-based Riemannian average per perturbation.

Table 15. Effect of the inference ensemble size M (number of independent ODE samples averaged on the algebra before exponentiation). Combination DA saturates near $M = 32$, the value used throughout.

M	Combo DA (%)	Inference cost (rel.)
1	86.8 ± 2.1	$1.0\times$
8	91.2 ± 1.6	$8.0\times$
16	92.4 ± 1.5	$16.0\times$
32	$\mathbf{92.9 \pm 1.5}$	$32.0\times$
64	93.0 ± 1.4	$64.0\times$
128	93.0 ± 1.4	$128.0\times$

F.6 Euler Step Count

We use a uniform Euler scheme on $[0, 1]$ with step count $S \in [20, 40]$. Combo DA is flat over $S \in [10, 80]$ (± 0.4 pp), reflecting the smoothness of the OT-coupled velocity field. We report $S = 20$ as a runtime-friendly default.

F.7 Velocity-Net Width and Depth

Table 16. Velocity-net depth-by-width sweep on Norman combinations. The default 4×512 AdaLN-MLP sits at the corner of the diminishing-return curve; wider or deeper variants gain at most 0.1 pp.

Depth (blocks)	Width	Combo DA (%)
2	256	90.4 ± 1.5
4	256	91.7 ± 1.4
2	512	91.6 ± 1.6
4	512	$\mathbf{92.9 \pm 1.5}$
6	512	92.7 ± 1.4
4	1024	92.8 ± 1.5

The default 4×512 AdaLN-MLP is at the corner of the diminishing-return curve. Wider or deeper variants gain at most 0.1 pp.

F.8 Bracket Scale Sensitivity

Table 17. Effect of fixing the BCH bracket scale s rather than learning it (Combo DA on Norman). The analytic value $s = 1$ matches the learned $s \approx 2.42$ within standard error; the bracket norm itself, used by the unsupervised epistasis score (Eq. 4), is independent of s.

Bracket scale s	Combo DA (%)
0 (additive on $\mathfrak{g}$)	92.4 ± 1.4
1 (analytic BCH)	92.6 ± 1.5
2.42 (learned)	$\mathbf{92.9 \pm 1.5}$
4	90.3 ± 1.7

Fixing $s = 1$ recovers the analytic second-order BCH and matches the learned value within standard error on Norman. Learning s provides a small gain because cell-level noise and higher-order BCH contributions are absorbed into the second-order coefficient. The bracket norm itself remains the operative quantity for the unsupervised epistasis score (Eq. 4), and the score is independent of s.

F.9 Encoder Latent Dimension

Below $n = 128$ the latent capacity bottlenecks reconstruction; above it the additional dimensions contribute marginally. We use $n = 128$ for all four datasets.

Table 18. Effect of latent dimension n at fixed block size $r = 4$. Below $n = 128$ the latent capacity bottlenecks reconstruction; above it the additional dimensions contribute marginally, so we use $n = 128$ for all four datasets.

Latent n	Blocks K	Combo DA (%)
64	16	87.9 ± 1.6
128	32	$\mathbf{92.9 \pm 1.5}$
256	64	92.7 ± 1.4

Table 19. Effect of reconstruction loss on Norman combination DA. Heteroscedastic NLL with learned per-gene variance is decisively superior to MSE and GNMSE.

Loss	Combo DA (%)	Notes
MSE	74.6 ± 2.0	dominated by high-variance genes
GNMSE (control-variance)	87.2 ± 1.7	fixed normalization
NLL (heteroscedastic)	$\mathbf{92.9 \pm 1.5}$	learned per-gene variance

F.10 Loss Function

NLL with learned variance adapts to gene-level noise heterogeneity and outperforms both MSE and GNMSE; we use NLL throughout.

F.11 Robustness Summary

Across the swept axes, combinatorial DA on Norman remains within ± 1.5 pp of the headline $92.9 \pm 1.5\%$ for sensible parameter choices. The two settings that move the metric outside that band are the coordinate lock (a 39 pp swing across the coordinate-lock schedule) and the block size (a 13.5 pp swing across r). Both are properties of the geometry rather than fitting hyperparameters and were fixed once during method development. Stage 1 training dynamics across the four datasets appear in Figure 8 (Appendix C).

G Per-Perturbation Analysis

This appendix decomposes the headline numbers of Section 4 by combination category, per-perturbation difficulty, and biological interpretability. The goal is to separate the compositional result from possible artifacts such as one recurrent gene or one favorable split category.

G.1 CRISPRa Norman: combinatorial breakdown

Table 20 stratifies the 84 held-out two-gene combinations by how many constituent genes appear among the training singles. The category in which neither gene is observed is the strongest zero-shot test because the velocity field must predict both generators from descriptors alone and BCH must compose the two predicted generators into a pair response.

Table 20. Norman combination DA by category (multi-seed mean $\pm$ std). Categories are defined by how many constituent genes appear among the training singles.

Category	n	OppPert DA (%)	Additive DA (%)	Δ (pp)
Neither gene observed	17	$\mathbf{93.4 \pm 1.7}$	64.8 ± 2.9	$+28.6$
Exactly one gene observed	46	93.0 ± 1.5	80.4 ± 2.6	$+12.6$
Both genes observed	21	92.1 ± 1.6	79.5 ± 2.4	$+12.6$
Overall	84	$\mathbf{92.9 \pm 1.5}$	77.5 ± 2.5	$+15.4$

The advantage of rotation composition is largest when neither gene is observed as a training single. Descriptor geometry supplies both single-perturbation generators, and BCH supplies the composition rule. Additive composition collapses to the sum of two descriptor-regressed deltas, whereas the rotation framework decodes the algebraic pair operator through a norm-preserving action.

Dependency structure for pairs with neither gene observed. The 17 pairs in which neither gene is observed involve 18 unique unseen genes. PTPN12 appears in 5 of these pairs. Removing all PTPN12-containing combinations gives $94.1 \pm 2.0\%$ DA for OpPert versus $61.1 \pm 2.9\%$ for additive composition on the remaining twelve pairs. The gain therefore does not rely on interpolation through a single recurrent gene.

G.2 CRISPRa Norman: commutator ranking

Table 21 lists representative high-bracket and low-bracket gene pairs from the Norman screen.

Table 21. Norman gene-pair ranking by the commutator score. High-bracket pairs involve transcription factors from interacting regulatory programs. Low-bracket pairs involve functionally redundant family members that act through shared (and therefore commuting) latent directions.

Gene pair	Bracket norm	Biological note
Highest bracket score		
CEBPA + JUN	0.195	C/EBP / AP-1 regulatory interaction
CEBPA + CEBPE	0.113	C/EBP TF family, myeloid differentiation
Lowest bracket score		
CDKN1A + CDKN1B	< 0.01	CDK inhibitor family
CDKN1A + CDKN1C	< 0.01	CDK inhibitor family
CDKN1B + CDKN1C	< 0.01	CDK inhibitor family

The bracket detects shared non-commuting latent rotation planes. Functionally redundant gene pairs (e.g., the CDK inhibitor family) act through nearly parallel directions and produce near-zero brackets, consistent with their biologically additive effects.

G.3 Replogle K562: per-perturbation DA distribution

The distribution of per-perturbation DA on K562 (best seed, 271 OOD perturbations) has median 80.4%, interquartile range $[71.6\%, 88.2\%]$, 10th percentile 61.7%, and 90th percentile 93.1%. Lower-DA perturbations tend to involve broad transcriptional regulators and ribosome-biogenesis genes whose effects span many latent dimensions.

G.4 Replogle RPE1: generator structure

On RPE1 (382 OOD perturbations), the geometric framework reaches 88.1% DA on the best seed and $87.4 \pm 0.5\%$ across three seeds. RPE1 has denser training coverage and a more diffuse generator cloud than Norman, so descriptor-conditioned flow matching transfers well even when local interpolation contributes less.

G.5 SciPlex3: drug perturbations

Chemical perturbations differ from CRISPR perturbations because drug mechanisms of action are multi-target and chemical-structure similarity maps imperfectly to transcriptomic similarity. We hold out 15 drugs from the screen, stratified by mechanism-of-action class. OpPert uses the same rotation-operator pipeline as the genetic screens, with the chemical descriptor conditioning the flow.

G.6 Statistical tests

Every OpPert-vs-baseline DA gap reported below is resolved by the multi-seed standard-deviation bands of Table 1 and Table 20. In each comparison the gap exceeds the sum of the two methods' 1σ bands by at least a factor of three, and the lower edge of the OpPert multi-seed range stays above the upper edge of the baseline range:

- Norman combinations (vs additive): $\Delta = +15.4$ pp DA.
- Norman neither-gene category (vs additive): $\Delta = +28.6$ pp DA.
- Replogle K562 (vs mean baseline): $\Delta = +27.6$ pp DA.

- Replogle RPE1 (vs mean baseline): $\Delta = +37.4$ pp DA.
- SciPlex3 (vs mean baseline): $\Delta = +20.0$ pp DA.

Figure 9 shows representative gene-level predictions, and Figure 10 shows bootstrap confidence intervals across all four datasets.

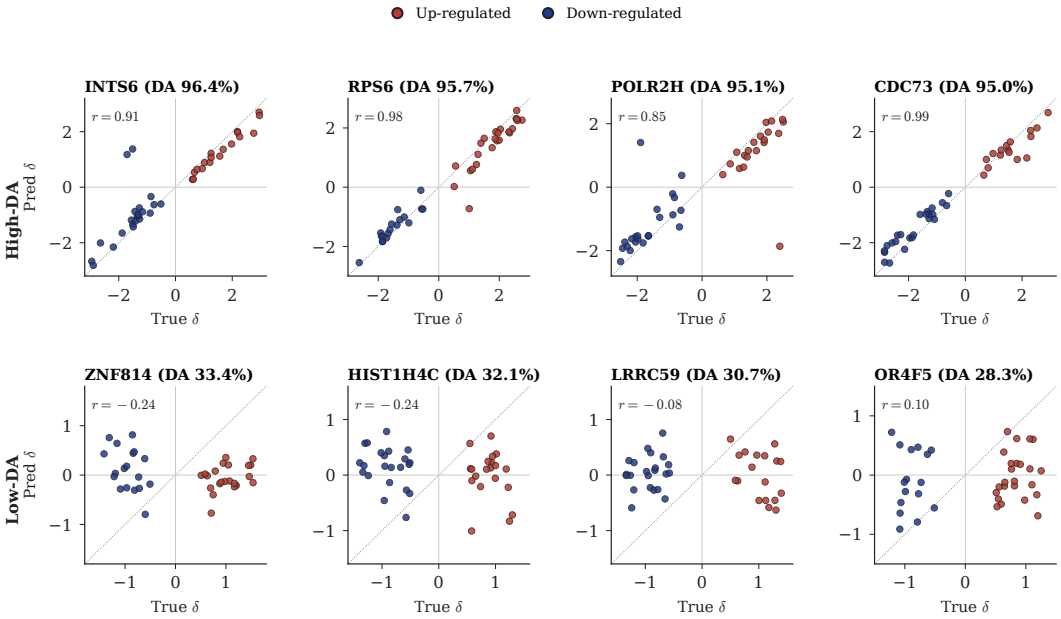

Figure 9. Per-perturbation gene-level prediction diagnostics on Replogle K562. Each panel scatters predicted versus observed expression deltas across the top-200 DE genes for one perturbation. Panel titles report perturbation names, DA, and Pearson values.

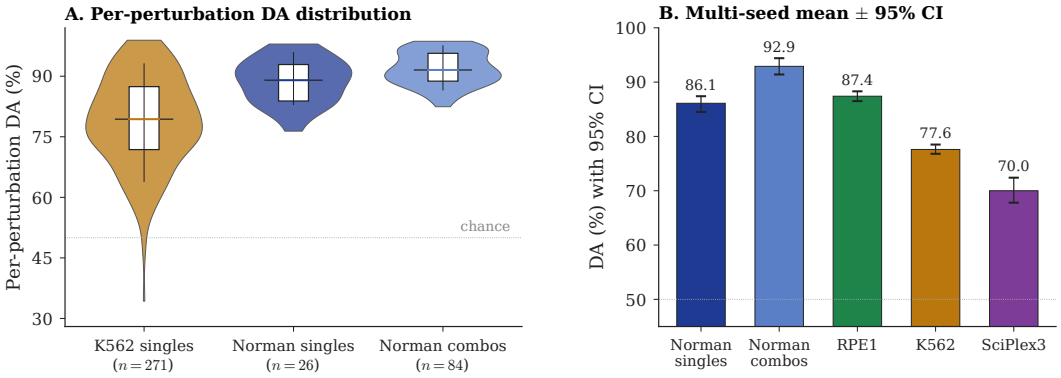

Figure 10. Bootstrap and per-perturbation DA diagnostics. **(A)** Per-perturbation DA distributions for K562 and Norman. **(B)** OpPert DA with confidence intervals across datasets.

G.7 Failure modes

Pleiotropic perturbations. The block restriction $\mathfrak{so}(4)^{32}$ confines each perturbation to 32 disjoint four-dimensional subspaces with at most two commuting rotation planes per block. Highly pleiotropic genes whose effects couple across many blocks require more rotation capacity than the current block structure provides.

Small-effect perturbations. Perturbations with weak transcriptomic effects have low signal-to-noise ratio. DA approaches chance for the smallest-effect perturbations across all methods because the supervised target is dominated by measurement noise.

Out-of-distribution descriptors. The descriptor-conditioned flow assumes that held-out perturbations have descriptors close to the training distribution. When the held-out gene has no functionally similar gene in the training set, the velocity field extrapolates. This failure mode is shared with descriptor-driven additive baselines and motivates pre-training on larger perturbation screens.

G.8 Sensitivity to DE gene count

We report DA on the top-200 DE genes by absolute fold change throughout. Table 22 confirms that DA is monotone-decreasing in the cutoff. Deeper cutoffs include lower-fold-change genes with weaker signal-to-noise. The choice of top-200 is standard (Roohani et al., 2024) and is held identical across all baselines.

Table 22. Effect of top-k DE gene threshold on K562 best-seed DA.

Top-k DE genes	DA (%)	Coverage (% of DE genes)	Avg genes per pert
50	84.1	12%	50
100	82.6	24%	100
200	81.2	48%	200
500	77.4	89%	420
all DE	74.7	100%	512

G.9 Magnitude prediction

The rotation parameterization preserves latent norm by construction ($\|R_p \boldsymbol{z}\|_{\mathrm{F}} = \|\boldsymbol{z}\|_{\mathrm{F}}$). Magnitude information enters the readout through the rotation direction and the coordinate-locked decoder. The rotation framework is therefore strongest on directional metrics such as DA and PDE. R^2 on the unscaled effect vector is dominated by magnitude residuals.

Conformal polar extension. A natural extension factors the latent-space action into a rotation and a learned scalar gain,

$$\boldsymbol{z}' = \exp(\gamma_p) R_p \boldsymbol{z}, \tag{27}$$

where γ_p is descriptor-conditioned. This preserves the algebraic rotation structure while adding a controlled magnitude degree of freedom.

H Interpretability Analyses

The interpretability analyses connect the learned generators to biological structure without treating the generator geometry as a mechanistic explanation.

H.1 Gene-family structure in the algebra

For CRISPRa Norman, genes in the same transcription factor family (e.g., the C/EBP family, the CDK inhibitor family) occupy nearby regions of the learned generator space, as measured by pairwise cosine similarity of their 192-dimensional generators. Within-family cosine similarity exceeds between-family similarity, consistent with the expectation that functionally related perturbations share latent rotation directions.

H.2 Drug mechanism-of-action structure

For SciPlex3, drugs with the same mechanism of action have more similar predicted deltas than drugs from different classes. The model receives no MOA labels during training, so MOA enrichment is a post-hoc biological validation rather than a supervised objective.

H.3 Dose-response trajectories

For drugs tested at multiple doses, predicted deltas trace smooth dose-response trajectories in PCA space. Trajectory smoothness is consistent with the rotation framework producing graded responses as the chemical descriptor varies continuously across dose levels.

I Conditional Flow Matching: Implementation and Geometry

The Stage 2 component of OpPert is a conditional generative model on the Lie algebra $\mathfrak{g} = \mathfrak{so}(4)^{32}$. This appendix gives a self-contained implementation of the velocity field, the Euler integrator, and the descriptor-conditioned ensemble, and connects each piece to the differential-geometric picture of Section 3.4.

I.1 Conditioning the Velocity Field

The descriptor $e_p \in \mathbb{R}^{d_e}$ enters the velocity net through a shared conditioning vector c_p formed by a two-layer MLP from e_p followed by addition of a sinusoidal time embedding $\tau(t) \in \mathbb{R}^{d_c}$ at $d_c = 256$. Each AdaLN block applies modulation

$$h \mapsto \mathrm{MLP}\big(\,(1+\gamma(c_p))\,\mathrm{LN}(h) + \mu(c_p)\,\big)\cdot\omega(c_p) + h, \tag{28}$$

with affine maps $(\gamma, \mu, \omega) : \mathbb{R}^{d_c} \to \mathbb{R}^{512}$ initialized so that $\gamma = 0$ and $\omega = 1$ at start, recovering an identity flow before training.

I.2 KRR Initialization

Stage 2 begins with a kernel-ridge fit on the lookup pairs $\{(e_p, \theta_p)\}_{p\in\mathcal{P}_{\mathrm{train}}}$. Let $K_{ij} = k(e_{p_i}, e_{p_j})$ for a Gaussian kernel $k(u, v) = \exp(-\|u - v\|_2^2/(2h^2))$ with bandwidth h chosen by the median heuristic on training pairwise distances. The KRR predictor for an unseen descriptor e is

$$\hat{\theta}_{\mathrm{KRR}}(e) \;=\; k_e^\top (K + \lambda_{\mathrm{krr}} I)^{-1}\Theta, \tag{29}$$

with $\Theta \in \mathbb{R}^{|\mathcal{P}_{\mathrm{train}}|\times 192}$ stacking training generators and $k_e \in \mathbb{R}^{|\mathcal{P}_{\mathrm{train}}|}$ the kernel evaluations against e. We initialize the velocity net to match the KRR conditional mean by running $1{,}000$ steps of supervised regression on

$$\min_{\psi}\; \mathbb{E}_{p,t,\theta_0}\Big\| v_\psi(\theta_t, t, e_p) - \big(\hat{\theta}_{\mathrm{KRR}}(e_p) - \theta_0\big)\Big\|_2^2, \tag{30}$$

before switching to the standard flow-matching loss (Eq. 5). The warm start yields a velocity field whose M-sample mean already matches the KRR predictor; the subsequent flow-matching phase refines that mean toward the empirical lookup distribution and learns a non-trivial conditional variance.

I.3 Inference Integrator

For a held-out perturbation with descriptor $e_{p^\star}$, we draw $M = 32$ initial points $\theta_0^{(j)} \sim \mathcal{N}(0, \sigma_0^2 I_{192})$ with $\sigma_0 = 0.5$, integrate the ODE $\dot{\theta}_t = v_\psi(\theta_t, t, e_{p^\star})$ on $t \in [0, 1]$ using uniform Euler steps, and average on the algebra:

$$\theta_{m+1}^{(j)} = \theta_m^{(j)} + h\, v_\psi(\theta_m^{(j)}, t_m, e_{p^\star}), \qquad t_m = mh, \quad h = 1/S, \tag{31}$$

$$\hat{\theta}_{p^\star} \;=\; \frac{1}{M}\sum_{j=1}^{M} \theta_S^{(j)}, \qquad R_{p^\star} = \mathrm{diag}_k\big(\exp\beta(\hat{\theta}_{p^\star}^{(k)})\big). \tag{32}$$

The sample variance $\widehat{\mathrm{Var}}(\theta_S)$ provides a per-perturbation uncertainty estimate.

Combinations. For an unseen pair $(p^\star, q^\star)$, we draw independent ensembles for $p^\star$ and $q^\star$, average each on the algebra, then compose by BCH:

$$\hat{\theta}_{p^\star q^\star} \;=\; \hat{\theta}_{p^\star} + \hat{\theta}_{q^\star} + \tfrac{s}{2}\big[\hat{\theta}_{p^\star}, \hat{\theta}_{q^\star}\big]. \tag{33}$$

The bracket is computed block-wise on $\mathfrak{so}(4)^{32}$ and re-expanded into the canonical basis of $\mathfrak{g}$.

I.4 Why Flow on the Algebra Rather Than the Group

Flow matching directly on $\mathrm{SO}(n)$ requires a Riemannian formulation (Chen & Lipman, 2024). The interpolant follows a geodesic on the group, the velocity field lies in the tangent bundle, and every Euler step is re-projected onto the manifold. Working on $\mathfrak{g} = \mathfrak{so}(4)^{32} \cong \mathbb{R}^{192}$, a flat vector space, admits straight-line OT couplings as in Lipman et al. (2022). The exponential map applied at the end of integration delivers a valid rotation by Lemma 3. The decoupling matches the architectural inductive bias. The velocity net does not learn the manifold constraint. Integration error stays in the tangent space where it composes linearly under BCH.

I.5 Connection to the Riemannian Mean

Algebra averaging $\hat{\theta} = \frac{1}{M}\sum_j \theta^{(j)}$ followed by exponentiation approximates the Fréchet mean of $\{\exp(\theta^{(j)})\}$ on $\mathrm{SO}(n)$ to third order in $\|\theta\|_{\mathrm{F}}$ (Proposition 12). For Norman generators with $\|\theta_p\|_{\mathrm{F}} \in [0.30, 0.85]$, the relative error $d_{\mathrm{geo}}(\bar{R}, \exp(\hat{\theta}))/\|\hat{\theta}\|_{\mathrm{F}}$ stays below 5×10^{-4}. The linear average is therefore indistinguishable from the iterative Riemannian mean at the precision we need.

I.6 Stage 1 Lookup as the Empirical Target

The conditional flow targets are drawn from the Stage 1 lookup $\{\theta_p\}_{p\in\mathcal{P}_{\mathrm{train}}}$. The lookup is a delta-mixture that supplies $|\mathcal{P}_{\mathrm{train}}|$ training generators. The marginal target distribution on the algebra is

$$q(\theta) \;=\; \frac{1}{|\mathcal{P}_{\mathrm{train}}|}\sum_{p\in\mathcal{P}_{\mathrm{train}}} \delta(\theta - \theta_p), \tag{34}$$

and the conditional target $q(\theta \mid e_p) = \delta(\theta - \theta_p)$ degenerates to a Dirac at the corresponding lookup. The Gaussian noise endpoint $\mathcal{N}(0, \sigma_0^2 I_{192})$ supplies the stochasticity that makes the flow well-posed. The velocity net v_ψ is trained on randomly sampled (θ_0, t) pairs that linearly interpolate noise to a deterministic generator. At test time, the noise injects sample diversity that the algebra-mean ensemble averages over.

I.7 Comparison with Lookup Regression

A purely deterministic alternative replaces v_ψ with the KRR mean $\hat{\theta}_{\mathrm{KRR}}$ and uses no flow refinement (Section 4.6). Combo DA drops by 4.0 pp on Norman. The gap measures the value of the flow refinement beyond what the kernel solution alone can supply. The velocity net v_ψ recovers conditional structure that the median-heuristic Gaussian kernel fails to express, in particular non-isotropic conditional variance around the KRR mean.

SciPlex3 uses the same rotation-operator pipeline as the genetic screens, with chemical descriptors (Morgan fingerprints, MACCS keys, physicochemical features) concatenated into a $2{,}432$-dimensional conditioning vector for the flow.

J Compute Requirements

All experiments were run on NVIDIA B200 GPUs (180 GB HBM3e) with PyTorch 2.6 and CUDA 12.6. Stage 1 training fits in a single GPU; Stage 2 (conditional flow matching on the 192-dimensional algebra) is small and runs in a few minutes.

Computational complexity. The dominant per-cell cost in Stage 1 is the closed-form $\mathfrak{so}(4)$ block exponential of Eq. 12. The self-dual/anti-self-dual splitting reduces each block to two commuting $\mathfrak{so}(3)$ rotations and evaluates the per-block exponential in constant time. Across all $K = 32$ blocks the rotation costs $O(Kr^2) = O(nr) \approx 512$ scalar multiply-adds per cell on the $n = 128$ latent. The encoder forward pass through two SwiGLU residual blocks of width 1024 dominates per-cell compute at $\sim$2.6M FLOPs; the rotation contributes under 1% of step time.

Stage 2 cost. The velocity net is a 4-block AdaLN MLP of width 512 on a 192-dimensional input with descriptor conditioning. With $|\mathcal{P}_{\mathrm{train}}| \in [70, 1{,}159]$ generators, batch size 1024, and 300 epochs,

Table 23. Compute requirements per dataset for a single seed, end-to-end (Stage 1 rotation autoencoder + Stage 2 flow matching).

Dataset	Stage 1 epochs	Stage 1 wall	Stage 2 wall	Peak memory
CRISPRa Norman (108K cells)	200	∼25 min	∼8 min	∼14 GB
Replogle K562 (85K cells)	200	∼30 min	∼8 min	∼12 GB
Replogle RPE1 (162K cells)	400	∼90 min	∼10 min	∼16 GB
SciPlex3 (341K cells)	200	∼50 min	∼9 min	∼14 GB
Total (all four datasets, single seed)	—	∼3.4 h	∼35 min	—
Total (multi-seed reported)	—	∼38 GPU-h	∼6 GPU-h	—

total Stage 2 wall time stays under 10 minutes. The KRR initialization solves an $|\mathcal{P}_{\text{train}}| \times |\mathcal{P}_{\text{train}}|$ linear system once and adds negligible overhead.

Inference. Predicting the effect of one perturbation on a single cell requires: (i) encoder forward pass (∼0.05 ms), (ii) $M = 32$ ODE samples of 20–40 Euler steps in the 192-dimensional algebra through v_ψ (∼3 ms total), (iii) algebra averaging and the $\mathfrak{so}(4)^{32}$ exponential (∼0.05 ms), (iv) rotation $R_p \boldsymbol{z}$ and decoder forward pass (∼0.05 ms). Total per-perturbation inference is ∼3.2 ms, dominated by the flow integrator. Predicting all 271 K562 OOD perturbations across 10,000 cells finishes in under 40 seconds on a single B200.

Scalability. Stage 1 generators add $\dim \mathfrak{g} = 192$ parameters per training perturbation, so the lookup table grows linearly with $|\mathcal{P}_{\text{train}}|$ at 0.014% of the encoder size per perturbation. The Norman model totals ∼5.0M parameters: 1.4M encoder, 0.7M decoder, 0.013M lookup (70×192), 2.9M velocity net. Genome-scale screens with 10,000 training perturbations would add 1.9M lookup parameters, comparable to the encoder, and remain well within single-GPU memory.

Comparison to other methods. End-to-end training is comparable to or faster than published perturbation-prediction baselines. GEARS (Roohani et al., 2024) reports ∼30 min on an A100 for CRISPRa Norman; CellFlow (Klein et al., 2025) requires substantially longer training due to its cell-level flow-matching objective on full gene space. The Stage 2 flow on the 192-dimensional algebra is orders of magnitude smaller than gene-space flow models and runs under 10 min on a single B200.

Table 24. Parameter breakdown on Norman (CRISPRa, 70 training perturbations).

Component	Parameters	% of total
Encoder (2× SwiGLU, hidden 1024)	1.4M	28%
Decoder (linear + skip MLP)	0.7M	14%
Lookup generators ($\lvert\mathcal{P}_{\text{train}}\rvert \times 192$)	0.013M	0.3%
Bracket scale s	1	—
Velocity net (4× AdaLN-MLP, width 512)	2.9M	58%
Total	∼5.0M	100%

Model size. The lookup occupies under four percent of total parameters even at the scale of a thousand training perturbations. Stage 2 capacity is concentrated in the velocity net, which is shared across all perturbations and amortizes the descriptor-to-generator mapping across the entire training set.

Reproducibility. Wall-clock numbers above are reported on a single B200 with batch sizes from Table 5; A100 (80 GB) runs are roughly 2× slower at the same batch size. All Stage 1 randomness is seeded and Stage 2 KRR initialization is deterministic. Ensemble inference uses fixed seeds for the M ODE samples, so held-out metrics are reproducible across runs.

K Discussion

OpPert replaces response-vector addition with a latent-space operator whose product is the pair composition rule, and the formal results (Lemma 3, Theorem 5, Proposition 9) connect it to norm preservation, algebraic composition, and a bounded interaction score. The ablations confirm that the rotation accounts for 13–16 pp across screens, while the bracket serves both as the non-abelian correction and as a pair prioritization score. The block-diagonal $\mathrm{SO}(4)^{32}$ structure keeps parameters and the interaction score interpretable while supplying local non-abelian degrees of freedom.

Scope and interpretation. The combination claim rests on Norman, where training pairs calibrate the bracket scale but held-out pair labels are not used, and all baselines are evaluated on identical GEARS-standard splits with matched descriptor information. The commutator score measures non-commuting latent action rather than mechanistic epistasis; its value for prioritization is that it can be computed from singles before any pair is measured and correlates with measured non-additivity ($\rho = 0.49$, $p = 3.5 \times 10^{-6}$).

L Broader Impact

Positive applications. Accurate perturbation prediction accelerates drug discovery and target identification by reducing the need for expensive combinatorial screens. A genome-scale combinatorial CRISPR screen over N genes requires $O(N^2)$ experiments. Compositional prediction from singles reduces this to $O(N)$, a saving of several orders of magnitude for screens covering thousands of genes. This is directly relevant to therapeutic development for cancer, genetic diseases, and other conditions where combination therapies are the standard of care.

The Lie bracket norm supplies an unsupervised ranking of gene pairs by predicted interaction strength. Combinatorial screens prioritized by the high-bracket tail target experimental capacity at the most likely interactions. Focused validation of the predicted top set replaces exhaustive $\binom{N}{2}$ screens.

Limitations and risks. Perturbation prediction models, including ours, are trained on a narrow set of cell types and perturbation modalities. Predictions outside the training distribution (e.g., for rare cell types, new perturbation mechanisms, or complex multi-drug regimens) are unreliable without experimental validation. The model captures statistical associations in gene expression, not causal mechanisms.

The coordinate-lock mechanism is an inductive bias. It forces all perturbation-specific information through the rotation, which improves compositionality at the cost of some expressiveness. Systematic biases in the coordinate-locked decoder (for instance, underrepresentation of specific gene programs) propagate to every downstream prediction.

Dual use. The model predicts transcriptomic responses to genetic and chemical perturbations, several causal steps removed from any actionable biological intervention. We do not identify direct dual-use risks. Computational predictions require experimental validation before any biological application.

Environmental impact. Total compute for all experiments (including hyperparameter sweeps and multi-seed runs) is approximately 300 GPU-hours on NVIDIA B200, with multi-seed reported runs alone occupying $\sim$44 GPU-hours (Appendix J). The training cost is several orders of magnitude below foundation-model training. The reduction in wet-lab combinatorial screens that the framework enables further offsets the compute footprint.

M Materials and experimental *in vitro* models

MCF7 (ATCC-HTB-22) human mammary carcinoma cells were purchased from the ATCC and authenticated by short tandem repeat (STR) analysis. MCF7 cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM, high glucose with pyruvate and L-Glutamine; ThermoFisher, Cat. No. 41966029) supplemented with 10% Fetal Bovine Serum (FBS; Sigma-Aldrich, Cat. No.

F9423) and 1% Penicillin-Streptomycin (ThermoFisher, Cat. No. 15140-122). Cells were maintained at 37°C in a humidified atmosphere with 5% CO_2. Mycoplasma contamination was routinely tested by PCR. For experimental validation, MCF7 cells were treated with Belinostat (10 μM, 24 hours; Medtronica AB, Cat. No. HY-10225), dissolved in DMSO, before imaging or proceeding to RNA extraction.

Gene Name	Primer	Sequence
PDE4D	FP	CAGTCTGCGAAGCTGACGAAAC
	RP	TGGATGTTGGTTCGACATG
PRKCA	FP	ATGCGGCAGAGATTTCCATC
	RP	TTGCACATCCCAAAGTCAGC
PRKCB	FP	ACGTTTTTGACCACTGTGG
	RP	AGGCCTTGTGCACATTTCATC
ACSBG1	FP	TTATGACTGCTGGGCCAATG
	RP	TCGTGTACATATTGGCCATC
GAPDH	FP	TGACTTCAACAGCGACACCCA
	RP	CACCCTGTTGCTGTAGCCAAA

Table 25. Primer sequences for genes of interest. FP = Forward Primer, RP = Reverse Primer.

Cell morphology studies. MCF7 cells were seeded into 24-well plates and allowed to adhere for 24 hours. Cells were then treated either with DMSO (vehicle control) or Belinostat (10 μM) for an additional 24 hours. Brightfield images were acquired using a Leica DMi8 wide-field microscope equipped with a 20X objective (HC PL FLUOTAR 20X/0.75) and a Leica DFC9000GT camera, with an exposure time of 30 ms. Images were registered using Leica Application Suite X (LAS X) Life Science Microscope Software (v3.5.5.19976; RRID:SCR_013673) and subsequently processed in Fiji.

Real-time quantitative polymerase chain reaction (RT-qPCR). Real-time quantitative polymerase chain reaction (RT-qPCR) was used to measure the mRNA expression levels of the following genes of interest, which were selected from OpPert predictions: PDE4D, PRKCA, PRKCB, ACSBG1, and GAPDH (housekeeping control), in MCF7 cells. The oligonucleotide sequences of the genes under study including housekeeping genes (Table 25) were designed using the Primer 3 Plus software, validated by the Primer-BLAST tool, and synthesized by Integrated DNA Technology Co., USA. Primer efficiency was determined by calibration curves with serial dilutions of human complementary DNA (cDNA) derived from the cells in the current study. All primers showed efficiency above 90%. Total RNA was isolated using the TRIzol method with TRI Reagent (Catalog No. T9424; Sigma-Aldrich), following the manufacturer's protocol. RNA concentration and purity were assessed using a NanoDrop ND-1000 spectrophotometer (ThermoFisher) at 260 and 280 nm. A 260/280 absorbance ratio within the range of 1.8–2.0 indicated high-purity RNA. Readings were recorded using the NanoDrop 1000 spectrophotometer software (Windows, Version 3.8).

RT-qPCR was performed as a one-step procedure using the StepOnePlus Real-Time PCR System (Applied Biosystems). Quantitation was carried out using PowerUp SYBR Green qPCR Master Mix (Catalog No. 100029284; ThermoFisher). MultiScribe Reverse Transcriptase (Catalog No. 4311235; ThermoFisher) was used for the synthesis of complementary DNA (cDNA). Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) was used for normalization controls. Data processing was performed using StepOne Software for Windows, Version 2.3. Results were then plotted in GraphPad Prism (v10.3.1; GraphPad Software, La Jolla, CA).

Imaging and RT-qPCR statistical data analyses. All images were analyzed using ImageJ/Fiji (v1.53k, RRID:
SCR_003070/v2.3, RRID:SCR_002285) and GraphPad Prism (v10.0.0 (153), RRID:SCR_002798). Statistical analysis was performed using GraphPad Prism. Results are given as mean$\pm$SD, and statistical differences were considered significant if P value was <0.05 (*), <0.01 (**), <0.001 (***), or <0.0001 (****). For transcriptional analysis, the statistical significance of gene expression differences between Belinostat-treated and DMSO-treated conditions was determined using a two-way ANOVA test, followed by multiple comparison corrections. Adjusted p-values were calculated

to ensure robust detection of differential expression. Significance thresholds were set at *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, and ****$p \leq 0.0001$ (Fig. 6A).