PREDICTING UNSEEN SCIENTIFIC INTERACTIONS: CONDITIONAL FLOW MODELS HALLUCINATE NON-ADDITIVE EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional generative models are increasingly used for scientific prediction beyond the direct training support. Yet scientific questions often concern whether one intervention changes the effect of another. Such interactions can be of primary interest while explaining little of the total variation, allowing models to appear accurate overall without recovering the relevant scientific effect. It remains unclear how these models learn interactions and when these successfully extrapolate to unseen conditions. We therefore study interaction recovery in flow-matching models using a physiological time-series simulator, drug dose-response data, and combinatorial cell perturbations, where we separate recovery from hallucinated interaction structure. In the controlled setting, individual effects are learned much earlier than interactions and models hallucinate interactions where none exist. Across settings, we find qualitatively different generalization problems, with successful prediction depending on whether condition structure continues from observed to held-out targets. When it does, interactions are recoverable. If not, models can generate plausible responses while missing the interaction. We furthermore find that hallucinations increase as targets become less identifiable, and that successful interaction recovery in output space can fail under nonlinear readouts. Finally, we show that aggregate fidelity metrics can conceal whether models learn and extrapolate interactions.

1 INTRODUCTION

Conditional generative models such as diffusion and flow matching models have gained significant attention for generating novel biological samples and for predicting biological responses under unseen experimental conditions. Prominent use cases include single-cell transcriptomics (He et al., 2026; Yu et al., 2026), neuroscience (Gijsen et al., 2026), and protein design (Watson et al., 2023; Ingraham et al., 2023).

To be scientifically valuable, models must be able to predict how experimental conditions interact, rather than only generating broadly plausible responses. For example, administering two drugs together may be antagonistic, independent, or synergistic, and redundant pairs require predicting that a drug has a clear effect on its own but not when combined with another. Importantly, however, the bulk of overall variability across samples is often dominated by effects of individual factors. Conversely, interactions often account only for a small fraction, despite regularly being the primary object of scientific interest. For instance, in single-cell perturbation experiments, the interaction between cell-type and perturbation determines less of the variance than either factor by itself, yet it is often the central question (Schäfer et al., 2026).

As aggregate fidelity metrics (such as MSE or FID) are heavily influenced by those stronger effects of individual factors, they may tell us relatively little about how interactions are learned during training or whether an interaction learned from observed combinations can be extrapolated to unseen ones. This is critical for scientific prediction and discovery, as this implies risks where models may appear converged while still learning the interaction, or appear reasonably accurate while predicting the effect that motivated an experiment incorrectly. Furthermore, interactions are not only a property of the system but also of the data representation. A nonlinear readout of a sample can imply a different

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.

interaction, and therefore interaction recovery in output space does not transfer to a quantity which often can be more scientifically relevant.

The difficulty of this generalization problem directly depends on the available training data. A held-out target may be identifiable via smoothness of the condition space, via graph-relational knowledge, or underdetermined if only marginal support is provided. The latter problem, often called compositional generalisation (Wiedemer et al., 2023), is fundamentally different in the presence of interactions, as a training support covering each factor alone no longer determines the response in the held-out region of condition space. Simply put, observing two factors separately does not determine how they interact. In this case, the missing relationship must come from observed joint conditions, inductive biases in the model, or meta-knowledge (Shojaee et al., 2025).

Here, we study how conditional flow models learn and extrapolate non-additive interactions in three systems that differ in the manner in which the training support constrains the held-out targets. Specifically, a photoplethysmography (PPG) simulator with ground truth and tunable interaction strength, drug dose-response screens where observed doses surround the held-out conditions, and combinatorial perturbations in single cells where components of a held-out pair are optionally excluded from observed pairs.

We summarize our contributions as follows:

- **Diagnostics for interaction prediction.** We decompose predicted interactions into interaction recovery and orthogonal hallucination, with an identity to interaction error.
- **Hallucination of non-additivity.** In a simulator with a controlled interaction strength, we find flow models to generate interactions even in fully additive regimes. We additionally find stronger hallucinations for less constrained targets across settings.
- **Continuation rather than distance determines extrapolation.** We show that held-out regions of a shared condition space at equal distance from the training support can be recovered almost completely or barely at all, depending on the continuation of the condition structure off the training support.
- **Aggregate metrics conceal failures to predict interactions.** We observe distinct learning trajectories for interactions compared to main effects, that low endpoint errors can accompany failed interaction recovery, and that recovery in output space can fail after a scientific readout.

2 Related work

Predicting perturbation and combination responses. Deep models are now commonly used to predict responses to unseen chemical and genetic perturbations and their combinations (Roohani et al., 2024; Klein et al., 2025; Preuer et al., 2018). However, recent work shows that such models can fail to beat additive or mean baselines for double perturbations (Ahlmann-Eltze et al., 2025), and that common metrics are inflated by systematic variation (Viñas Torné et al., 2025). As additive baselines predict no interaction, we use these as a reference for scoring interaction errors. We complement these existing evaluations by decomposing interaction errors and relating them to the identifiability of held-out condition regions.

Compositional generalization and inductive biases of generative models. Compositional generalization has often been studied in settings where the joint response is determined by its components (Wiedemer et al., 2023; Okawa et al., 2023; Liu et al., 2022), which is an assumption that the presence of interactions violate. Prior work additionally links the order in which concepts are learned to their signal strength (Rahaman et al., 2019; Park et al., 2024), which invites the question how low-energy interactions are learned. Diffusion models have also been shown to produce unseen samples by interpolating between modes (Aithal et al., 2024), with the interaction hallucinations we study here constituting an analogous failure over condition space rather than sample space.

Interactions depend on representations. Whether factors interact depends on the data representation in which we measure, which has been long established in epidemiology, where it is described as the interaction scale (Rothman et al., 1980; VanderWeele & Knol, 2014). We extend this to the evaluation of generative models, showing that interaction recovery can differ between the output space and under a scientific readout.

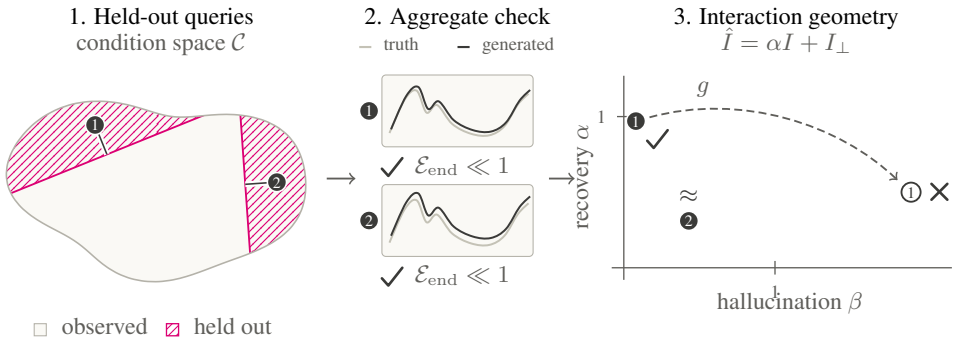


Figure 1: Aggregate generation quality does not guarantee accurate predicted interactions. Two conditions are sampled with equal distance off the training support but in different directions (1). Both samples score well on normalized endpoint error ($\mathcal{E}_{\text{end}}$), far below a trivial baseline (2). But decomposing the prediction into the component along the true interaction (α) and the orthogonal, unrelated component (β) distinguishes the samples into interaction recovery (✓) and attenuated interaction with hallucinated structure (≈) (3). Finally, applying a readout function g can considerably change the results (✓ $\xrightarrow{g}$ ×), with less recovery and yet more hallucinated structure in the readout space compared to output space.

3 Methods

3.1 Setup

We start by providing a brief general framework. Given two factors A and B taking values a and b (categorical or continuous), flow matching models predict response distributions $P_{a,b}$ over $\mathbb{R}^d$ with conditional means $\mu(a,b) = \mathbb{E}[Y \mid a, b]$. The factors are additive if $\mu(a,b) = g(a) + h(b)$ for functions g and h and the interaction here is the part of the response not captured by such decomposition. For reference levels (a_0, b_0) (Section 3.3), we define the additive interaction contrast as

$$I(a,b;\, a_0,b_0) = \mu(a,b) - \mu(a,b_0) - \mu(a_0,b) + \mu(a_0,b_0) \in \mathbb{R}^d,$$

which vanishes for all (a,b) if and only if μ is additive. In the canonical 2×2 design with a 0 baseline, individual conditions A and B, and their joint condition AB, this reduces to

$$I_{AB} = \mu_{AB} - \mu_A - \mu_B + \mu_0.$$

For continuous factors without a natural reference level, there is no scientifically meaningful choice for (a_0, b_0). We therefore use functional ANOVA (Section 3.2), which avoids choosing a reference. Namely, its interaction at each condition is simply the average of the contrast across all possible reference points in the grid. The interaction is contrasted with the total effect $\mu(a,b) - \mu(a_0,b_0)$ which is what most aggregate evaluations are sensitive to.

Interactions depend on the representation. Scientific disciplines are often concerned with representations in a semantic space (rather than the raw output space), which then form the basis for drawing conclusions. For a fixed, deterministic readout function $g : \mathbb{R}^d \to \mathbb{R}^k$, $I^g_{AB} = g(\mu_{AB}) - g(\mu_A) - g(\mu_B) + g(\mu_0)$. If g is nonlinear, then interaction recovery in the output space does not imply recovering it after applying the readout. Therefore, $\|\hat{I} - I\|$ may be considerably smaller than $\|\hat{I}^g - I^g\|$, in which case evaluating model predictions only in the output space can be incomplete or misleading.

Relating endpoint errors to interaction errors. Let $\hat{\mu}_c$ denote the predicted mean response under condition c and $e_c = \hat{\mu}_c - \mu_c$ the endpoint error. By substituting into the definition of I_{AB} we get:

$$\hat{I} - I = e_{AB} - e_A - e_B + e_0,$$

where $\hat{I}$ denotes the implied interaction based on the model predictions. These four errors have no guarantee of cancellation meaning that a small joint error does not bound the interaction error. Conversely, any errors shared across conditions may cancel out and thus small interaction errors do not indicate that individual conditional means are accurate.

Due to this, we distinguish between 'anchored' and fully generated estimates. The former relies on reference conditions while the latter relies fully on the model:

$$\hat{I}^{\text{anch}} = \hat{\mu}_{AB} - \mu_A - \mu_B + \mu_0, \hat{I}^{\text{full}} = \hat{\mu}_{AB} - \hat{\mu}_A - \hat{\mu}_B + \hat{\mu}_0.$$

The anchored error corresponds to e_{AB} and isolates model error for unseen conditions, which is analogous to experimental settings in which only single factors (but not their joint) have been measured. $\hat{I}^{\text{full}}$ instead corresponds to using the model for the whole experiment.

Distinguishing hallucination and other interaction failure modes. To expose distinct failure modes, we decompose the prediction into two components: one along the true interaction and along an orthogonal component:

$$\hat{I} = \alpha I + I_\perp, \qquad \alpha = \frac{\left\langle \hat{I}, I \right\rangle}{\|I\|^2}, \qquad I_\perp = \hat{I} - \alpha I, \qquad \beta = \frac{\|I_\perp\|}{\|I\|},$$

where $I_\perp$ is the part of the model prediction orthogonal to the true interaction. α then is the signed gain along the true interaction with β denoting the orthogonal relative energy. Because the components are orthogonal this gives the following exact identity:

$$\mathcal{E}_I^2 = \frac{\|\hat{I} - I\|^2}{\|I\|^2} = (1-\alpha)^2 + \beta^2.$$

The zero-interaction predictor is $(\alpha, \beta) = (0, 0)$ with $\mathcal{E}_I = 1$. Thus, α indicates how much of the true interaction the model recovered and β how much the model additionally hallucinated. $0 < \alpha < 1$ with a small β can be interpreted as an attenuated predicted interaction, $\alpha < 0$ a reversal, $\alpha > 1$ an exaggeration, and $\alpha \approx 0$ with a large β indicating error of which the structure is unrelated to the truth (Figure 1). We note that with $m = \|\hat{I}\|/\|I\|$ and θ denoting the angle between the predicted and true interaction, we get $\alpha = m\cos\theta$ and $\beta = m\sin\theta$. These are consequently recoverable from commonly used quantities rather than novel metrics.

Identification and generalization regimes. Many conditional-mean functions agree on the training support while yielding distinct interactions for held-out targets. Consequently, I_{AB} is not identified by the training support alone, with inductive biases determining the prediction. Hence our experiments can be seen as evaluating this bias.

Crucially, this distinguishes an interaction setting from the additive compositional generalization problem. Namely, with interactions, the widely studied L-shaped support for grid-based contexts (Wiedemer et al., 2023) therefore requires function extrapolation that the observed marginals do not determine. We distinguish between generalization regimes based on what partially enables identification:

- **Smoothness identified:** the held-out region of condition space is surrounded by joint observations.
- **Relationally identified:** factors A and B are each observed in other combinations.
- **Underdetermined:** regime in which only marginals have been observed and thereby a training support which does not provide any constraint on interactions on the held-out region.

3.2 METRICS

We normalize errors such that a value of 1 corresponds to a trivial predictor. Specifically, we use the response scale $\text{RMS}(Y - Y_0)$ for response errors and the interaction scale $\text{RMS}(I)$ for interaction errors (Table 1). Y_0 is a domain-appropriate control or reference response.

For a 2×2 design, I_{AB} is the AB response minus the additive part $\mu_A + \mu_B - \mu_0$. Our simulator uses a large 2D grid of conditions and therefore has many condition levels per factor. We therefore obtain the same decomposition using functional ANOVA over this grid, with $Y = \tilde{Y} + I$, where $\tilde{Y}$ is the grand mean plus the two main-effect surfaces and I the residual non-additive component. We also apply this to the model's generated grid $\hat{Y}$, giving $\hat{Y} = \hat{\tilde{Y}} + \hat{I}$, where $\hat{\tilde{Y}}$ is the additive part of the model's prediction and $\hat{I}$ its interaction. All grid cells are used to compute the grand and main-effect means. $\mathcal{E}_I$, α, and β are computed afterwards on the held-out region of condition space.

Table 1: Metrics overview.

Metric	Definition	1 corresponds to
$\mathcal{E}_{\text{end}}$	$\mathrm{RMS}(\hat{Y} - Y) \,/\, \mathrm{RMS}(Y - Y_0)$	predicting the baseline
$\mathcal{E}_{\text{add}}$	$\mathrm{RMS}(\hat{\tilde{Y}} - \tilde{Y}) \,/\, \mathrm{RMS}(Y - Y_0)$	an error the size of the total response
$\mathcal{E}_I$	$\|\hat{I} - I\| \,/\, \|I\|$	predicting zero interaction
α, β	Section 3.1	-

3.3 Domain-specific interaction targets

The additive contrast is realized differently across our three studied systems (described in Section 3.5):

- **Physiological time series:** given continuous factors without a natural reference level, we use the fANOVA decomposition directly in the waveform representation, such that I_{AB} is itself a time series.
- **Physiological readout:** a nonlinear physiological readout g is applied to the generated waveforms before the fANOVA decomposition, yielding I^g_{AB}.
- **Single-cell perturbation:** we use unperturbed controls as the reference, with I_{AB} computed from condition-level mean expression vectors, thus yielding a vector over genes.
- **Drug dose–response:** the zero dose of the combined drug serves as the reference, and we define interaction relative to Bliss independence rather than additively (see below).

Additivity is not a sensible null for drug combinations as fractional effects are bounded to $[0, 1]$, meaning simple addition of two drug effects can violate these bounds. Consequently, Bliss independence is commonly used in pharmacology, which assumes that two drugs act independently (Bliss, 1939). Concretely, we define E_A and E_B as normalized (to $[0, 1]$) fractional drug effects such as inhibition, where 0 indicates no effect and 1 fully inhibiting. When drugs act independently, their combined effect is

$$E^{\text{Bliss}}_{AB} = E_A + E_B - E_A E_B.$$

Equivalently, Bliss independence means multiplication of surviving fractions, $1 - E_{AB} = (1 - E_A)(1 - E_B)$, which is additive following the readout $g(E) = \log(1 - E)$ and another example of interactions being representation-dependent (Section 3.1). Next, we can define the interaction as the deviation from this independent effect:

$$I^{\text{Bliss}}_{AB} = E_{AB} - E^{\text{Bliss}}_{AB}.$$

Positive values then indicate drug-drug synergy seeing as inhibition is larger than expected and negative values vice versa indicate antagonism.

3.4 Models

We train conditional flow matching models following Lipman et al. (2023) with minor variations across domains which we summarize in Appendix Table 4. In the standard setting, we learn a velocity field $v_\theta(x_\tau, \tau, c)$ that transports $p_0 = \mathcal{N}(0, 1)$ to the conditional data distribution $p_1(\cdot|c)$ along the probability path $x_\tau = (1 - \tau)z_0 + \tau z_1$ where $\tau \in [0, 1]$ indicates the flow matching time. Models are trained using an MSE loss and, unless stated otherwise, sampled by numerical integration of the learned flow. Table 4 also includes conditioning, optimization, model architectures, and prediction targets.

We compare three ways of obtaining a prediction for the joint condition AB:

- **Joint** model is trained and evaluated with conditioning directly on AB and explicitly represents interaction.
- **Composed** composes the learned conditional velocity fields

$$v^{\text{comp}}_{AB}(x, t) = v_A(x, t) + v_B(x, t) - v_0(x, t)$$

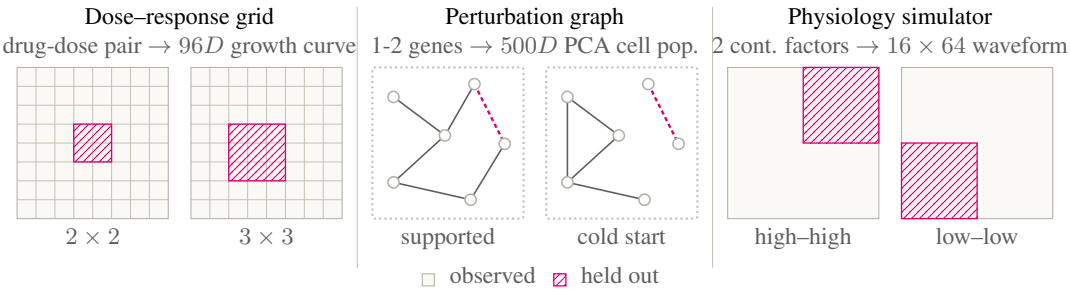


Figure 2: Held-out designs across the three systems. Please see Section 3.5 for details.

at every sampling step and is a widely-used compositional approach for unseen combinations Liu et al. (2022); Du et al. (2023). Importantly, $\hat{I}^{\text{full}}$ is not necessarily zero as the composed and single-factor fields induce different nonlinear sampling trajectories.

- **Shared-model AB** relies on same network as used for the Composed evaluation, but instead directly queries AB. It thereby functions as a control which allows us to separate the effect of the shared training from the effect of the compositional step.

3.5 DATASETS

We study three systems, which cover different aforementioned identification regimes while having well defined interaction targets (Figure 2). The physiological simulator of photoplethysmogram (PPG; widely used in wearables) maps melanin and blood-volume fractions to 16-channel PPG waveforms with 64 timesteps and provides exact simulator ground truths and a tunable interaction strength. Specifically, we use fANOVA to decompose the simulated grid as $Y = \tilde{Y} + I$ (Section 3.2) and define $Y_\gamma = \tilde{Y} + \gamma I$. Consequently, γ modulates the interaction while leaving both main effects and the grand mean unchanged. Then, $\gamma = 1$ refers to the original grid, with $\gamma = 0.5$ denoting a halved interaction strength and $\gamma = 0$ a purely additive setting. We construct a simplified version of PPGen (Behrmann et al., 2026), where tissue composition determines optical properties in a multilayer skin model. For each channel (ch), we apply a standard normalized pulse-amplitude (AC/DC) readout (Allen, 2007; Coutrot et al., 2021):

$$\left(\frac{\text{AC}}{\text{DC}}\right)_{ch}(Y) = \frac{\max_t Y_{ch,t} - \min_t Y_{ch,t}}{\frac{1}{T}\sum_{t=1}^{T} Y_{ch,t}}.$$

Briefly, AC is the heartbeat-synchronous change, mainly attributed to pulsatile blood volume. Meanwhile, DC is the baseline optical signal from skin, melanin, and other slowly varying effects. The readout therefore asks the physical question of how large the pulsatile optical change is relative to the baseline reflectance, and how this jointly depends on melanin and dermal blood volume.

We include two further datasets enabling investigations in measured data under smoothness-identified and relationally-identified holdouts. Specifically, the dose-response dataset consists of longitudinal yeast growth responses (96-dimensional) to pairs of drugs administered on 8×8 dose grids (Cokol et al., 2011). We hold-out and predict 2×2 and 3×3 blocks for seen drug pairs. The Norman CRISPRa Perturb-seq dataset contains single-cell responses to single- and paired gene perturbations (Norman et al., 2019). For the held-out double perturbations, other combinations of the targets are either included in the training set ('graph-supported' regime) or excluded ('cold-start' regime). We predict in a 500-dimensional PCA space. We provide more details of these datasets in Appendix B.

For dose-response and perturbation data, we report the anchored estimate to approximate more realistic scientific workflows. For PPG, only generated estimates are applicable.

Interaction magnitudes vary across these three systems, with interaction-to-total norm ratios of 5.6% (PPG), 23.2% (dose-response; median assay), and 46.1% (perturbation; median pair). This corresponds to 0.3%, 5%, and 21% of response energy. The aggregate error is therefore most severely dominated by main effects for the PPG data. For Y_0, we use the lowest melanin/blood volume value, untreated responses, and unperturbed conditions.

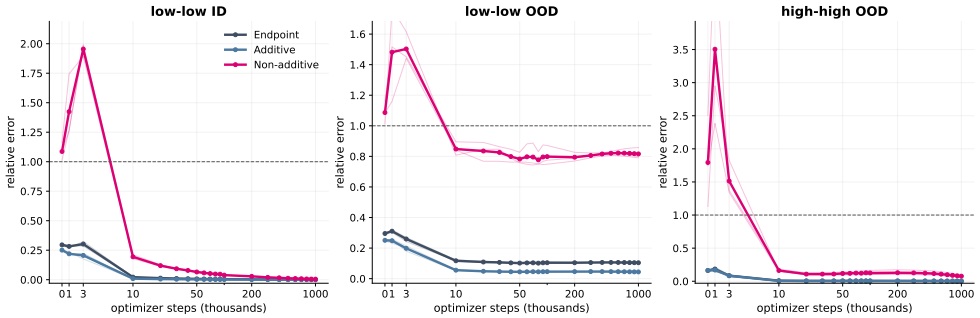


Figure 3: Learning dynamics for PPG targets in the training set (ID) and unseen targets (OOD) for separate regions of condition space. We show mean dynamics of the joint model across three random seeds, with faint lines indicating single-seed runs.

4 RESULTS

4.1 AGGREGATE ENDPOINT ERROR CONCEALS INTERACTION LEARNING DYNAMICS

For the synthetic PPG dataset, we hold-out two separate corners of the condition space (*low-low* (LL) and *high-high* (HH)), referring to the melanin and blood volume values) and investigate errors across training (Figure 3). Whereas additive and endpoint errors appear to have largely converged by 10K optimizer steps, the interaction error has not. Specifically, for seen targets, the model needs at least an order of magnitude longer training. For unseen targets, we observe two distinct dynamics: when LL is held-out, interaction error stays high and plateau's by 50K steps. In contrast, for unseen HH targets, interaction error is approximately 7 times smaller by 20K steps, with minor continued improvements at 1M steps.

We also find such decoupling between trajectories of endpoint and interaction errors in measured data, albeit in the opposite direction. For held-out Norman pairs, $\mathcal{E}_I$ begins to increase after $\sim$30K steps while $\mathcal{E}_{\text{end}}$ continues to decrease until $\sim$50K steps (Appendix Figure 7). Consequently, checkpoint selection using endpoint error selects models whose interaction predictions are already degrading. In the well-constrained setting of dose-response prediction, no clear decoupling is observed.

4.2 DISTINGUISHING INTERACTION-PREDICTION FAILURES

Distinct recoverability of a shared condition space. We further investigate these bifurcated results by inspecting model interaction recovery and hallucination (Figure 4). We observe stark differences, with joint models largely recovering the interaction for HH but not LL targets ($\alpha \approx 1.00$ and $\alpha \approx 0.31$ at $\gamma = 1$, respectively). Hallucinated structure is also higher for LL ($\beta \approx 0.08$ vs 0.43). For the LL corner, we provide mapsof interaction produced by models over condition space, which show that models manage to capture only broad structure and that the composed and shared-model AB baselines significantly underestimate interaction amplitude (Figure 5).

These results showcase that the extend to which interactions can be extrapolated based on a joint training support can strongly depend on the modeled domain. The two corners are geometrically matched but not physically equivalent. We find that a simple log-additive continuation rule selected based on the training support transfers to the held out HH corner. LL, in contrast, approaches a physical boundary where it is not (Appendix C). This maps well to the aforementioned identifiability problem.

Modulating interaction strengths. The PPG simulator allowed us to do a controlled investigation of different interaction strengths (γ). However, at $\gamma = 0$, we have $I = 0$ and thus $\mathcal{E}_I$ and α are not applicable. As such, to quantify what interaction a model predicts in the absence of a true interaction to recover, we compare against the full-strength $\gamma = 1$ truth $I^{(1)}$, with

$$\alpha^{(1)} = \frac{\langle \hat{I}, I^{(1)} \rangle}{\|I^{(1)}\|^2}, \qquad \beta^{(1)} = \frac{\|\hat{I} - \alpha^{(1)} I^{(1)}\|}{\|I^{(1)}\|}.$$

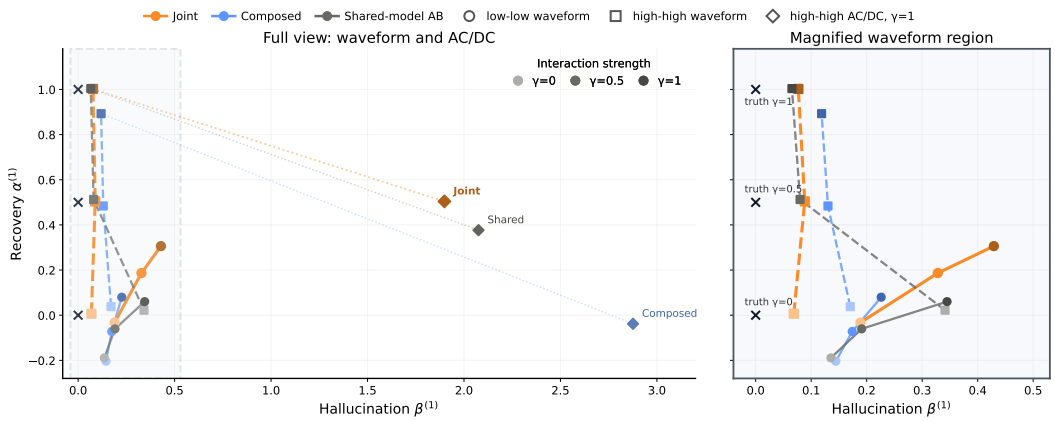


Figure 4: Interaction errors for unseen PPG targets. The left panel shows errors in the output (waveform) representation and readout (AC/DC) representation, while the right panel shows a magnified view of the (same) waveform results. The interaction strength γ was modulated $1 \to 0$ and is indicated by a dark $\to$ bright marker color gradient. Mean over training and inference seeds.

Ideally, a model shows $\alpha^{(1)} \approx \gamma$ and $\beta^{(1)} \approx 0$.

First, we observe that reducing interaction strength leads to monotonic decreases in $\alpha^{(1)}$, indicating models do somewhat calibrate their magnitudes based on the training support. For LL, hallucinated structure similarly decreases. Importantly, at $\gamma = 0$, $\beta^{(1)} > 0$ for all models. Thus, when the true interaction has been fully removed, models still generate structure, with joint $\beta^{(1)}_{LL} \approx 0.19$ and $\beta^{(1)}_{HH} \approx 0.08$. This suggests that models' inductive bias produces some baseline level of non-additivity.

Strong predictions in output space can mask failed recovery in readout space. While the joint model predicts interaction for unseen HH targets well in waveform space, results can differ significantly after a physiological readout, with $\alpha^{wf} \approx 0.97 \to \alpha^{AC/DC} \approx 0.50$ and $\beta^{wf} \approx 0.08 \to \beta^{AC/DC} \approx 1.90$.

4.3 Interaction errors across regimes

Finally, we study interaction recovery and hallucination across generalization regimes and varying training supports. For PPG, whereas previously we held out 25% of the condition grid, we now extend this to 50% and 75% and train only for 100K steps due to computational costs. For dose-response data, we hold out a 2×2 or 3×3 dose block. For single-cell perturbation, we predict responses to unseen gene-perturbations pairs using either single factors or additional combinations.

As identification gets harder, we observe reduced interaction recovery (α), with distinct delta magnitudes across systems. For PPG ($\gamma = 1$) LL this nearly entirely removes weak recovery ($0.30 \to 0.05$), indicating the importance of locality for extrapolation. Meanwhile, HH is recoverable even far from the training support (1.03). For dose-response, we observe only a very weak drop ($0.98 \to 0.97$), with sufficient smoothness constraint persisting to the larger held-out block. We confirm this by fitting a standard (deterministic) regression baseline and find nearly identical recovery ($0.98 \to 0.97$). For perturbations, recovery is again considerably reduced ($0.42 \to 0.17$), highlighting the importance of other observed combinations.

Interestingly, we observe increased hallucination across systems: PPG HH β $0.12 \to 0.25$, dose-response $0.18 \to 0.24$, and perturbations $0.80 \to 0.86$. This suggests hallucinated structure not to be a fixed property of the model and that they do not merely add a baseline level of hallucinated structure, but that this scales as targets get less identifiable.

Comparing endpoint and interaction errors instead, reveals an important further insight (Table 2). Namely, for PPG LL, enlarging the held-out region of condition space maintains non-trivial endpoint

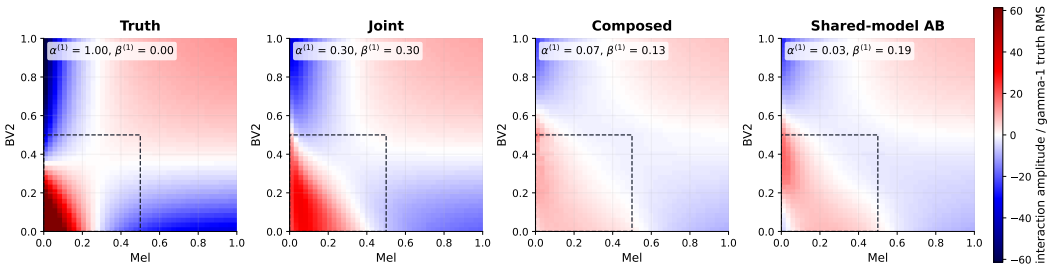


Figure 5: Maps showing the non-additive interaction projected onto the primary truth interaction mode. Colors indicate the signed interaction amplitude, normalized by the $\gamma = 1$ low-low truth RMS. The dotted square indicates the held-out region of condition space (corresponding to PPG LL 25%). To aid visualisation, we clip the colormap to [0.5%, 99.5%] of the maximum amplitude.

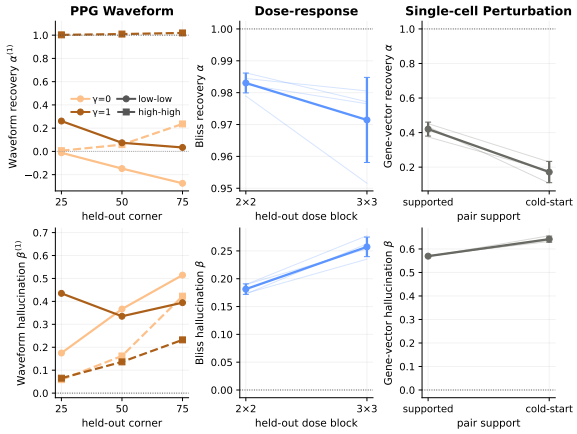


Figure 6: Interaction recovery and hallucination for unseen targets across the PPG simulator, drug dose-response, and single-cell perturbation datasets for joint models.

System / holdout	$\mathcal{E}_{\text{end}}$	$\mathcal{E}_I$
PPG LL, 25%	0.104 ± 0.006	0.798 ± 0.065
PPG LL, 50%	0.144 ± 0.0004	0.960 ± 0.009
PPG LL, 75%	0.165 ± 0.007	0.992 ± 0.031
PPG HH, 25%	0.006 ± 0.001	0.125 ± 0.035
PPG HH, 50%	0.015 ± 0.001	0.205 ± 0.021
PPG HH, 75%	0.027 ± 0.002	0.251 ± 0.026
Dose-response 2×2	0.045 ± 0.002	0.216 ± 0.016
Dose-response 3×3	0.057 ± 0.003	0.300 ± 0.007
Perturb. supported	0.582 ± 0.027	0.919 ± 0.018
Perturb. cold-start	0.561 ± 0.021	1.052 ± 0.044

Table 2: Normalized endpoint and relative interaction error of joint models, with 1 indicating prediction of a control response or zero interaction respectively. Mean and SDs across training seeds. For PPG, we always use 100K step checkpoints for consistency.

prediction ($\mathcal{E}_{\text{end}} \ll 1$), while making the interaction prediction no better than predicting no interaction at all ($\mathcal{E}_I \approx 1$).

5 Discussion

We have investigated how flow models learn interactions and generate them for unseen targets using error decompositions. We have found multiple interesting dynamics: aggregate endpoint errors can conceal interaction errors, scientifically-relevant readouts can materially change interaction prediction scores, and found that both recovery and hallucination depend on identifiability regimes.

Our findings suggest that scientific evaluation should test interaction recovery directly in the representation space which is used to draw conclusions. With respect to data, individual factors appear insufficient; either collecting observations or appropriate structural assumptions are necessary to constrain unseen interactions. Computing interaction recovery and hallucination in addition to aggregate errors can help distinguish succesful extrapolation from plausible predictions which miss the scientific effects of interest.

Limitations of the presented work include our restriction to two-factor interactions, a reliance on a mean-based (rather than distributional) contrast, and the usage of a single flow-matching recipe.

AI USE STATEMENT

In this work, we used generative AI tools for evaluating our proposed evaluation framework, provide feedback on research methodology and experiments, implement methods, and support data analysis. Additionally, we used generative AI tools for modifying scientific figures, edit software code, editing for improving readability, analyse existing literature, and brainstorming. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work proposes an evaluation framework rather than developing novel model capabilities or applications. Furthermore, it analyzes synthetic as well as existing data without collecting new biological data.

REPRODUCIBILITY STATEMENT

To enable reproducibility, we will publicly release the code used for model training and evaluation upon acceptance. In addition, this work references the used datasets and includes descriptions of analysed data (Appendix sections B.1, B.2, B.3), model architecture (Appendix D), model optimization (Appendix D), and evaluation (Section 3.1).

REFERENCES

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.

Sumukh K Aithal, Pratyush Maini, Zachary C Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. *Advances in neural information processing systems*, 37:134614–134644, 2024.

John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28(3):R1–R39, 2007.

Jens Behrmann, Maria R Cervera, Antoine Wehenkel, Andrew C Miller, Albert Cerussi, Pranay Jain, Vivek Venugopal, Shijie Yan, Guillermo Sapiro, Luca Pegolotti, et al. Inferring optical tissue properties from photoplethysmography using hybrid amortized inference. *Machine Learning: Health*, 2(1):015012, 2026.

Chester I. Bliss. The toxicity of poisons applied jointly. *Annals of Applied Biology*, 26(3):585–615, 1939. doi: 10.1111/j.1744-7348.1939.tb06990.x.

Murat Cokol, Hon Nian Chua, Murat Tasan, Beste Mutlu, Zohar B. Weinstein, Yo Suzuki, Mehmet E. Nergiz, Michael Costanzo, Anastasia Baryshnikova, Guri Giaever, Corey Nislow, Chad L. Myers, Brenda J. Andrews, Charles Boone, and Frederick P. Roth. Systematic exploration of synergistic drug pairs. *Molecular Systems Biology*, 7:544, 2011. doi: 10.1038/msb.2011.71.

Maxime Coutrot, Emmanuel Dudoignon, Jona Joachim, Etienne Gayat, Fabrice Vallee, and François Depret. Perfusion index: physical principles, physiological meanings and clinical implications in anaesthesia and critical care. *Anaesthesia Critical Care & Pain Medicine*, 40(6):100964, 2021.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8489–8510. PMLR, 2023.

Sam Gijsen, Michał Łukomski, Marc-André Schulz, and Kerstin Ritter. Flow matching with in-context priors for out-of-distribution brain dynamics, 2026.

Siyu He, Yuefei Zhu, Daniel Naveed Tavakol, Haotian Ye, Yeh-Hsing Lao, Zixian Zhu, Cong Xu, Shradha Chauhan, Guy Garty, Raju Tomer, et al. Squidiff: predicting cellular development and responses to perturbations using a diffusion model. *Nature methods*, 23(1):65–77, 2026.

John B. Ingraham, Max Baranov, Zak Costello, Karl W. Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M. Lord, Christopher Ng-Thow-Hing, Erik R. Van Vlack, Shan Tie, Vincent Xue, Sarah C. Cowles, Alan Leung, João V. Rodrigues, Claudio L. Morales-Perez, Alex M. Ayoub, Robin Green, Katherine Puentes, Frank Oplinger, Nishant V. Panwar, Fritz Obermeyer, Adam R. Root, Andrew L. Beam, Frank J. Poelwijk, and Gevorg Grigoryan. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023. doi: 10.1038/s41586-023-06728-8.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, et al. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, pp. 2025–04, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision – ECCV 2022*, volume 13677 of *Lecture Notes in Computer Science*, pp. 423–439. Springer, 2022. doi: 10.1007/978-3-031-19790-1_26.

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

Maya Okawa, Ekdeep S. Lubana, Robert P. Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50173–50195, 2023. doi: 10.52202/075280-2182.

Core F Park, Maya Okawa, Andrew Lee, Hidenori Tanaka, and Ekdeep S Lubana. Emergence of hidden capabilities: Exploring learning dynamics in concept space. *Advances in Neural Information Processing Systems*, 37:84698–84729, 2024.

Kristina Preuer, Richard PI Lewis, Sepp Hochreiter, Andreas Bender, Krishna C Bulusu, and Günter Klambauer. Deepsynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, 34(9):1538–1546, 2018.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

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

Kenneth J. Rothman, Sander Greenland, and Alexander M. Walker. Concepts of interaction. *American Journal of Epidemiology*, 112(4):467–470, 1980. doi: 10.1093/oxfordjournals.aje.a113015.

Philipp S. L. Schäfer, Kendall A. Reid, Zach Boldyga, Ekin D. Aksu, Hugo Hakem, and Julio Saez-Rodriguez. Towards principled evaluation of single-cell perturbation prediction models. *bioRxiv*, 2026. doi: 10.64898/2026.07.23.740433.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan Reddy. Llm-sr: Scientific equation discovery via programming with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 16054–16085, 2025.

Tyler J VanderWeele and Mirjam J Knol. A tutorial on interaction. *Epidemiologic methods*, 3(1):33–72, 2014.

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

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. In *Advances in Neural Information Processing Systems*, volume 36, pp. 6941–6960, 2023. doi: 10.52202/075280-0304.

Chenglei Yu, Chuanrui Wang, Bangyan Liao, and Tailin Wu. scDFM: Distributional flow matching model for robust single-cell perturbation prediction, 2026. ICLR 2026 poster.

A Additional Learning Dynamics

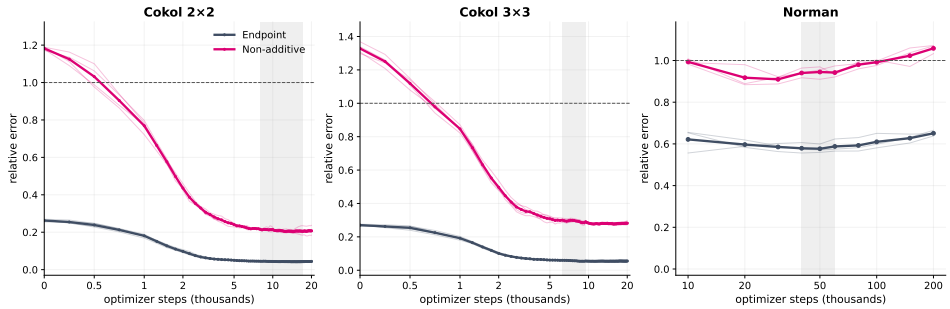


Figure 7: Held-out evaluation across the training trajectory for the Cokol et al. (2011) dose-response setting and for the Norman et al. (2019) perturbation data. While the former shows synchronous dynamics for aggregate and non-additive metrics due to a smoothness-constrained held-out region of condition space, the latter shows decoupling. Specifically, whereas $\mathcal{E}_I$ begins to increase after 30K steps, $\mathcal{E}_{end}$ continues to decrease up to 50K steps. The shading indicates checkpoint selection based on validation endpoint error, highlighting that using endpoint error for model selection can be suboptimal for interaction prediction.

B Dataset details

B.1 PPG simulator.

The PPG data are generated using deterministic replay of a fixed Monte Carlo photon-path bank using a three-layer skin model (simplified from Behrmann et al. (2026)). A path bank contains 10^8 photon histories for 35 scattering settings with linear interpolation between adjacent settings, and Beer-Lambert reweighting applying absorption. This makes the simulator deterministic conditional on the path bank. We generate 64 time points for four source-detector distances (3-6 mm) across four wavelengths (525, 660, 850, 940 nm).

The condition space is made up of epidermal melanin (0.25–14%) and dermal blood volume (BV2; 0.1–3.9%). Meanwhile, arterial oxygen saturation is fixed at 80% and subcutaneous blood volume at 4.05%. The temporal blood-volume modulation follows a normalized arterial luminal-area waveform (from the PWDB database; age-45). Dermal and subcutaneous blood volumes are multiplied by $1 + 0.02w(t)$ with $w(t) \in [0, 1]$.

First, we simulate a uniform 101×101 Mel-BV2 grid. We apply balanced functional ANOVA to the clean $\gamma = 1$ reference grid, writing $Y_1 = \tilde{Y} + I$, where $\tilde{Y}$ contains the grand mean and the Mel and BV2 main-effect functions, and I is the residual interaction. The controlled interaction-strength grids are then defined as $Y_\gamma = \tilde{Y} + \gamma I$. As such, $\gamma = 1$ is the original grid, with $\gamma < 1$ reducing its interaction while preserving the mean and main effects. 10^6 Mel-BV2 pairs are drawn independently and uniformly across the condition space. We obtain the PPG waveform for every pair by bilinearly interpolating each detector-wavelength-time coordinate of the reference grid.

Table 3: PPG simulator configuration.

Component	Configuration
Tissue geometry	Epidermis/dermis/subcutaneous thicknesses: $0.20/1.50/18.30$ mm; 0.05-mm isotropic voxels; $18 \times 18 \times 20$-mm domain.
Source and detectors	Normally incident uniform disk source, radius 1.6 mm; four spherical detectors of radius 0.25 mm at 3, 4, 5, and 6 mm.
Optical constants	Refractive index $n = 1.4$; anisotropy $g = 0.9$; shared reduced-scattering law $\mu_s'(\lambda) = 6.25(\lambda/1000\,\mathrm{nm})^{-1.4}$ cm^{-1}.
Absorbers	Eumelanin in epidermis; water and lipid in all layers; collagen in dermis; oxygenated/deoxygenated hemoglobin in dermis and subcutaneous tissue.
Photon-path bank	35 scattering settings; 10^8 photons per setting; 5-ns MCX time window; fixed paths reused for absorption replay.

Splitting. We use nested low-low or high-high rectangles to create increasingly larger hold-out regions of condition space: $25\% \to 50\% \to 75\%$. These contain 735k/15k, 490k/10k, and 245k/5k training/validation samples, respectively. For final evaluation, we use a consistent held-out corner of condition space such that the test region does not change as the training support shrinks.

B.2 Drug-response data

The Cokol et al. (2011) yeast screenings consist of 200 drug-pair assays across 33 drugs, with 175 different-drug assays and 25 same-drug controls. Each assay is an 8×8 dose grid, with every cell consisting of a 96D growth curve. This results from 15 minute measurements across 24 hours, yielding 12800 total trajectories.

Splitting. For each assay we hold out a fixed set of blocks (rather than cross-validation). The 2×2 hold-out yields [11577, 759, 464] train/validation/test trajectories, with 3×3 sample sizes being [11077, 769, 954]. The same 106 assays are used to compare across the two block sizes. Untreated, single-drug, and same-drug conditions are included in both training sets.

B.3 Single-cell perturbation data

The Norman et al. (2019) K562 CRISPRa Perturb-seq dataset consists of 91205 cells and 5045 genes, including controls, single perturbations, and 131 double perturbations which involves 73 genes. The double-perturbation conditions contain 49-1030 cells (median=257). Models use a 500D PCA representation fitted to 20000 control and single-perturbation cells. For evaluation, we decode predictions to the 5045-gene log-expression space.

Splitting. We use cross-validation with four outer folds holding out each double perturbation once, while all controls and single perturbations are always included in the training set. Train/validation/test sample sizes range between [78539-80391, 2628-3770, 7348-9992] cells. Of the 131 test pairs, 113 are 'graph-supported' meaning that both genes are included in other training pairs. For the remaining 18 'cold-start' pairs, at least one of the two genes has no observed training pairs. The four folds are repeated across three training seeds.

B.4 Evaluation

Prior to computing endpoint and interaction metrics, we average across sampling seeds within each trained model. For PPG, we average across four sampling seeds, for drug-response data we average

eight predictions, and for single-cell perturbation data we average eight batches of 24 generated cells. We report variation across training seeds, with PPG using three seeds, drug-response data uses four seeds, and single-cell perturbation data using three.

C WHY DO THE PPG HOLD-OUT CORNERS DIFFER?

To understand why flow models trained on PPG data extrapolated well into the held-out high-high (HH) corner, but not into the geometrically matched low-low (LL) corner, we performed a small check on spatial continuation. Specifically, we use the training support to select from a small set of simple continuation rules. This procedure selected a raw-log-additive rule which constructed HH highly accurately (RMSE$= 0.007$, interaction cosine=1.00). While the same rule was selected when LL was held out, it failed to extrapolate (RMSE=0.771). This suggests that the response dynamics in the held-out HH corner continue smoothly from observed support, while the LL region has interaction structure which cannot be accurately inferred from the L-shaped training support.

D IMPLEMENTATION DETAILS

Table 4: Implementation details for the three experimental settings.

Detail	**PPG**	**Dose-response**	**Perturbation**
Output	16×64 waveform	96-point trajectory	PCA-500 expression
Architecture	Conv-patch DiT	Symmetric MLP	Residual MLP
Width / depth	128 / 4	256 / 3	512 / 4
Attention heads	4	N/A	N/A
Parameters	0.862M (both)	0.509M each	3.031M
Conditioning	2 factors; + presence bits (shared)	Drug-token set	Gene-token set
Condition set	AB (Joint); all four (shared)	0–2 drugs	0–2 genes
Flow source	Gaussian noise	Gaussian noise	Measured control (OT-paired)
Prediction target	Endpoint x_1	Endpoint x_1	Velocity $x_1 - x_0$
Training loss	MSE; 4-mask MSE (shared)	Endpoint MSE	Velocity MSE
Optimizer	AdamW	AdamW	AdamW
Learning rate	2×10^{-4}	2×10^{-4}	2×10^{-4}
Weight decay	10^{-8}	10^{-4}	10^{-4}
LR schedule	5k warm-up + cosine	Constant	Constant
Gradient clipping	None	5	1
EMA decay	0.999	0.999	0.999
Batch size	512	512	512
Training budget	1M steps	20k steps	200k steps / fold
Checkpoint	Final EMA	Best covered-validation EMA	Best inner-validation EMA
Primary inference	Euler	$t = 0$ endpoint	Heun
Inference steps	128	1	20
Joint prediction	Direct AB query	Direct AB query	Direct AB query
Composed prediction	$\hat{x}_A + \hat{x}_B - \hat{x}_\emptyset$	N/A	$v_A + v_B - v_\emptyset$