WHAT DOES A DIFFERENT COUNTERFACTUAL COST? LOSS-CALIBRATED AUDITS OF LEARNED COUPLINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fitting outcome distributions does not determine individual counterfactual predictions. We audit that gap in an already trained flow. Gaussian-source rotations carried through the frozen flow change its pairings while preserving the complete conditional density path, and an exact quadratic identity prices finite alternatives under the original flow-matching objective, allowing query-specific search and independent evaluation without retraining. On a Kang stimulation task the released scCBGM-FM editor closes most of the distance between unedited and stimulated cells; an alternative gives up under one percent of that gain, costs a fraction of a percent of the fitted objective on both the global mixture and the edited condition, and still moves a majority of a 233-cell held-out panel by at least a tenth of a control standard deviation. Every cell keeps a positive predicted increment, so an aggregate reporting only the direction of effect records nothing, while the mean effect and the responder fraction both move. A costlier alternative, qualifying only at the secondary budget, reverses a responder call that four independently trained fits agree on. Learned Gaussian experiments check the accounting against a closed-form boundary price, which a strict reversal approaches without attaining, and show that the query determines which alternative is found and that optimizing an estimated price understates it. Good task performance and agreement across refits can thus coexist with sensitivity in individual predictions: audit the prediction the decision rests on, not only the aggregate reported beside it.

1 INTRODUCTION

Flow matching learns a transport by regressing the velocity of a source-to-target bridge, often an independent straight bridge (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023). Fitting one flow per intervention and pushing the same latent code through both gives a model-defined counterfactual pair

$$Y(0) = G_0(Z), \qquad Y(1) = G_1(Z), \qquad Z \sim \nu, \tag{1}$$

and a distribution of individual effects. The arm-wise fits do not determine the pairing: two models can match Q_0 and Q_1 equally well and disagree about which $Y(1)$ belongs with which $Y(0)$. Such transports are used for perturbation modelling and potential-outcome estimation (Bunne et al., 2023; Dong et al., 2023; Wu et al., 2026); conditional flow editors infer a factual latent before changing the intervention context (Andersson et al., 2026).

Prior work exposes this freedom through treatment-dependent Gaussian rotations and loss-constrained disagreement optimization (Nasr-Esfahany & Kiciman, 2023, Sections 5.3 and 6–7); bijective, cocycle and dynamic-transport restrictions identify the pairing by adding structure (Nasr-Esfahany et al., 2023; Dance & Bloem-Reddy, 2025; De Sousa Ribeiro et al., 2025). We ask about an already fitted model: *which changes to its counterfactual predictions can be exhibited without changing its density path, and what do they cost under its own fitting objective?*

Our construction changes the counterfactual map while keeping the weights frozen (Figure 1). Source symmetry preserves the complete density path, and the added velocity is linear in its coefficients even for a nonlinear flow, making the finite risk increment exactly quadratic and precomputable. Here cost means change in the original flow-matching objective, not loss of observational support or causal plausibility. Gaussian experiments show why the query and independent pricing matter (Section 4.2). In the main application the editor closes most of the distance between unedited cells and held-out

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.

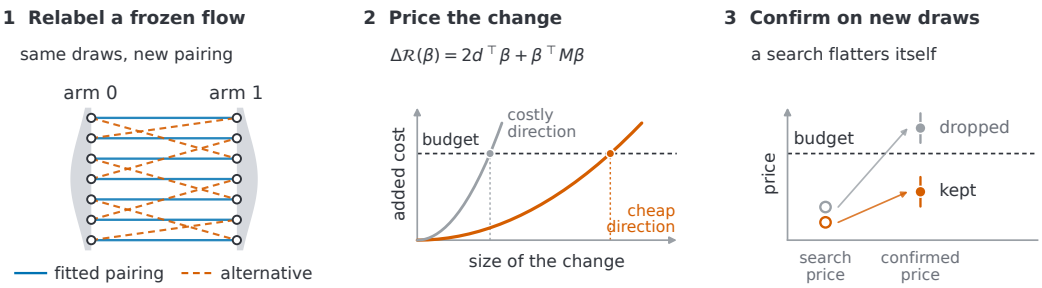

Figure 1: **Change the pairing, not the fitted density path.** Schematic. **1**: a source rotation repairs arm-0 to arm-1 draws with both arm laws left fixed. **2**: the added fitting risk is exactly quadratic, so a budget bounds how far a search travels, and cheaper directions travel further. **3**: a search flatters its own price; only the price re-confirmed on unused draws is charged against the budget.

target cells; an alternative gives up under 1% of that improvement and still moves 149 of 233 held-out cells by at least 0.1 SD (Section 5.1). Earlier applications show when sensitivity does not establish task value (Section 5.2).

2 From Marginal Fit to a Training-Rule Audit

Let $Q_a = \mathcal{L}\{Y(a)\}$ be the target law under arm $a \in \{0,1\}$ and $Z \sim \nu$. Independent conditional flow matching draws $Y_a \sim Q_a$ independently of Z, samples $t \sim \mathrm{Unif}[0,1]$, and regresses the bridge velocity

$$X_{a,t} = (1-t)Z + tY_a, \qquad U_a = Y_a - Z \tag{2}$$

under the risk

$$\mathcal{R}_a(v) = \mathbb{E}\|v(t, X_{a,t}) - U_a\|_2^2. \tag{3}$$

Integrating v_a produces maps $\Phi_{a,t}$ with endpoint $G_a = \Phi_{a,1}$. Generation requires only $G_{a\#}\nu \approx Q_a$. Sharing source labels additionally selects the transport $G_1 \circ G_0^{-1}$ and every pair-level functional

$$\Theta(G_0, G_1) = \mathbb{E}_Z\big[\ell\{G_0(Z), G_1(Z)\}\big]. \tag{4}$$

A responder fraction or dispersion of individual effects is of this kind. We report the hard query $\Pr\{s(Y(1)) - s(Y(0)) > c\}$ and use a sigmoid version only to choose a search direction.

The audit fixes the weights, source, bridge and validation law. When training standardizes each arm by $x = D_a^{-1}(y - \mu_a)$, risk is priced in x while the raw-outcome query composes G_a with $\mu_a + D_a x$. A fixed output transformation preserves distributional equality, not the loss (Appendix E.1). Changing the source would require a new population solution; Appendix A examines that dependence before Section 4.1 freezes one Gaussian model to calculate the price of reaching the decision boundary.

3 Loss-Calibrated Coupling Audit

3.1 Preserving the Model's Complete Density Path

Assume the fitted flow uses $\nu = \mathcal{N}(0, I_d)$, that $\Phi_{a,t}$ is differentiable in time and a spatial diffeomorphism, and that the risk terms below are square-integrable. Let $K_{a,1}(t), \ldots, K_{a,k}(t)$ be time-integrable skew-symmetric matrices, put $K_{a,\beta}(t) = \sum_j \beta_j K_{a,j}(t)$, and define $R_{a,\beta,t}$ by

$$\dot{R}_{a,\beta,t} = K_{a,\beta}(t) R_{a,\beta,t}, \qquad R_{a,\beta,0} = I, \tag{5}$$

which keeps $R_{a,\beta,t}$ orthogonal. Relabel the fitted trajectories as

$$\widetilde{\Phi}^{\beta}_{a,t}(z) = \Phi_{a,t}(R_{a,\beta,t} z). \tag{6}$$

Proposition 1 (A realizable density-preserving perturbation)**.** *For every a, t and β,*

$$\widetilde{\Phi}^{\beta}_{a,t\#}\mathcal{N}(0, I_d) = \Phi_{a,t\#}\mathcal{N}(0, I_d), \tag{7}$$

and the perturbed trajectories are generated by $\widetilde{v}^{\beta}_a = v_a + b_{a,\beta}$ where, for $z = \Phi^{-1}_{a,t}(x)$,

$$b_{a,\beta}(t, x) = D\Phi_{a,t}(z)K_{a,\beta}(t)z = \sum_j \beta_j b_{a,j}(t, x). \tag{8}$$

When the endpoint rotations differ across arms, Equation (1) *changes even though every arm and every time marginal is identical.*

The symmetry passes through the learned flow, so preservation holds regardless of fitting error. Computing the field requires inverse ODE solves and Jacobian–vector products; Appendix E checks their numerical implementation. For the ReLU editor the formula holds almost everywhere under the Lipschitz-flow extension in Appendix B.3.

3.2 Pricing alternatives and searching for them

Evaluate v_a and $\widetilde{v}^{\beta}_a$ on a frozen validation distribution drawn from the original bridge and write $r_a = v_a(t, X_{a,t}) - U_a$. Because Equation (8) is linear in β, the change in risk needs no expansion.

Proposition 2 (Fitted-model risk identity)**.** *For a finite perturbation basis,*

$$\Delta\mathcal{R}(\beta) = \sum_a \{\mathcal{R}_a(v_a + b_{a,\beta}) - \mathcal{R}_a(v_a)\} = 2d^\top\beta + \beta^\top M\beta, \tag{9}$$

where $d_j = \sum_a \mathbb{E}[r_a^\top b_{a,j}]$ and $M_{jk} = \sum_a \mathbb{E}[b_{a,j}^\top b_{a,k}]$, with $M \succeq 0$.

At a population regression optimum the residual is conditionally mean zero and d vanishes. A fitted network need not reach that optimum, so pricing must keep d, although using a noisy $\widehat{d}$ to steer search can exploit the sampled bridge draws.

For a conditional field $v(t, x; h)$, expectations use the original validation law P for (H, t, X_t, U), including its context weights and feature normalization. A mean-per-feature loss in dimension d_x has $d_j = \mathbb{E}_{\mathsf{P}}[r^\top b_j/d_x]$ and $M_{jk} = \mathbb{E}_{\mathsf{P}}[b_j^\top b_k/d_x]$. Encoder, sampler, targets and weights stay frozen. Unlike Proposition 3, the identity allows dependent bridge endpoints; changing the sampler would price a different objective.

To search, let Θ_s be a differentiable surrogate with $g = \nabla_\beta\Theta_s(0)$. Assume $M \succ 0$ on the retained basis; adding a ridge changes the local problem. For $g \neq 0$, the unit-quadratic-risk direction of fastest local decrease is $h = -M^{-1}g/\sqrt{g^\top M^{-1}g}$. When $d = 0$ and Θ_s is twice continuously differentiable with locally bounded Hessian, the small-budget envelope is

$$\inf_{\beta^\top M\beta \le \delta} \Theta_s(\beta) = \Theta_s(0) - \sqrt{\delta}\sqrt{g^\top M^{-1}g} + O(\delta). \tag{10}$$

If instead one minimizes the linearized query under the *full* fitted-model budget $2d^\top\beta + \beta^\top M\beta \le \delta$, completing the square gives

$$\beta^*_\delta = -M^{-1}d - \sqrt{\frac{\delta + d^\top M^{-1}d}{g^\top M^{-1}g}}\, M^{-1}g. \tag{11}$$

We use $-g$, h and Equation (11) to propose candidates, evaluating the hard query and full risk at every finite step.

To compare individual predictions, hold each factual outcome $x_i = G_0(z_i)$ fixed. Under the alternative control map its latent is $R_0^{-1}z_i$, so its counterfactual is $x^{\mathrm{cf}}_{i,\beta} = G_1(R_1R_0^{-1}z_i)$, writing $R_a = R_{a,\beta,1}$. Only the relative rotation matters. For the predicted score increment $D_\beta(x_i) = s(x^{\mathrm{cf}}_{i,\beta}) - s(x_i)$, measure disagreement by

$$H(\beta) = \frac{1}{n}\sum_{i=1}^{n} \big\{D_\beta(x_i) - D_0(x_i)\big\}^2. \tag{12}$$

Anchoring on factual units removes the spurious disagreement that indexing by the original latent z can assign to a common rotation $R_0 = R_1$.

Since H is minimized at $\beta = 0$, its gradient vanishes and we use curvature to find a direction. Write $a_i = \nabla_\beta D_\beta(x_i)|_{\beta=0}$ and $C = n^{-1}\sum_i a_i a_i^\top$. When each $\beta \mapsto D_\beta(x_i)$ is twice continuously differentiable with locally bounded second derivative, $H(\beta) = \beta^\top C\beta + O(\|\beta\|^3)$. The largest local disagreement per unit of quadratic risk is then given by the leading generalized eigenvector of

$$C\,h = \lambda\,M\,h, \tag{13}$$

whose eigenvalue gives that local ratio. This $k \times k$ problem does not require $g \neq 0$; computing the a_i is the expensive part. The expansion applies to the tanh Gaussian flows. For the ReLU editor, finite-difference C only proposes candidates, which are evaluated with the actual editor (Appendix C).

The same anchoring applies to conditional editors. Let $h = (u, c)$ be a frozen factual encoding and $h' = (u, c')$ its intervention, keeping u and every unedited concept fixed. The editor maps x to $G_{h'}G_h^{-1}(x)$, so conditional rotations induce

$$x_\beta^{\mathrm{cf}} = G_{h'}\Big(R_{h',\beta,1}R_{h,\beta,1}^{-1}G_h^{-1}(x)\Big). \tag{14}$$

The factual inverse must change with the edited generator to keep both halves of the edit in one model. A shared smooth family such as $K_{h,\beta} = c_{\mathrm{stim}}\sum_j \beta_j A_j$ with skew A_j acts on training conditions as well as intervention values, excluding gates active only at unseen contexts. Proposition 1 preserves the Gaussian-source path at fixed h, but observed cells' abducted latents need not be Gaussian under h'. Their edited population and the encoded context therefore need separate checks. Quantities depending only on the preserved output law, such as decode-only conditional expectations, cannot move.

3.3 Audit protocol and interpretation

Before search we fix the hard conclusion $\Theta \geq \tau$, surrogate, perturbation family and computational budget. Search draws determine g, d, M and candidates; frozen candidates are then evaluated independently. Gaussian-source queries share a binomial budget of 0.05 through one-sided exact bounds. Costs use fresh paired bridge draws and separate approximate simultaneous intervals (Appendix E.1). Observed-cell queries need unused factual cells or an explicit finite-population estimand: more latent draws cannot supply biological replicates. A witness qualifies against a pre-search tolerance δ only if its independent cost upper bound falls below δ.

Report a *query–cost profile*: the frozen model, query and search family; the baseline and best confirmed alternative found, with independent price uncertainty; and whether confirmation concerns generated units, a fixed panel or new subjects. If δ_* is the family's minimum conclusion-changing cost, a witness establishes $\delta_* \leq \delta(\beta)$, subject to pricing uncertainty. No flip does not establish robustness, and the audit does not choose the user's tolerance.

4 Experiments

An audit reports two numbers: how far an alternative moves the prediction, and what it costs. Both have to be trusted before either is applied to a model whose correct coupling nobody knows, so we check them first where a correct answer exists. Section 4.1 builds a case whose cheapest conclusion-changing alternative is available in closed form, giving the accounting a known value to reproduce. Section 4.2 then fits a flow, where the price must be estimated rather than derived, and separates what a search proposes from what independent evaluation returns. Section 5 audits the released scCBGM-FM editor, where neither the coupling nor its price is known in advance. The adapter also runs on TorchCFM's released sampler and MLP; Appendix C separates this software-compatibility check from application evidence. Appendix E gives configurations and confirmation protocols.

4.1 An exact minimum within the planar-rotation family

When both arms are Gaussian and the family is planar rotations, the cheapest alternative that overturns a stated conclusion can be derived rather than searched for, because the cost of an endpoint

angle is a kinetic energy with a known optimal time schedule. That derivation gives a value the executed audit has to reproduce. Continue the two-dimensional model of Appendix A under a standard Gaussian source, where the independent-CFM density path has diagonal map $L_{a,t} = \operatorname{diag}\sqrt{(1-t)^2 + t^2 \operatorname{diag}(D_a)^2}$ with $D_0 = I$ and $D_1 = \operatorname{diag}(2, 0.5)$. A pushed-forward planar rotation with angular rate $\omega_a(t)$ has energy $\int k_a(t)\omega_a(t)^2 dt$ for $k_a(t) = \operatorname{tr}(L_{a,t}^2)$, so an endpoint angle θ_a costs at least θ_a^2/I_a with $I_a = \int_0^1 k_a(t)^{-1} dt$. Splitting the relative endpoint angle $\theta = \theta_1 - \theta_0$ optimally between the arms gives

$$\Delta\mathcal{R}_{\min}(\theta) = \frac{\theta^2}{I_0 + I_1}, \qquad I_0 = \frac{\pi}{4}, \qquad I_1 = \frac{\pi}{2\sqrt{8.5}}, \tag{15}$$

against a responder probability of $\Phi\{0.7/\sqrt{3.125 - 2.5\cos\theta + 1.5\sin\theta}\}$. The responder probability reaches the 80% boundary at $\theta = 0.04299$ radians, costing $\Delta\mathcal{R} = 0.00139555$. That boundary price is attained by an executable generator under the fixed original population objective. A strict reversal needs a larger angle and therefore costs strictly more, so within this family the strict-reversal cost is an infimum rather than an attained minimum. Either value optimizes the time schedule and the split between arms, whereas the fitted-flow searches below use constant generators and so do not reach it.

4.2 Search, confirmation and query choice in a fitted flow

A fitted flow differs from an exact one in two ways the audit has to survive. Its residual need not be conditionally mean zero given $(t, X_{a,t})$, so the cross term d does not vanish and stays in every price; and both d and M are estimated from finite draws, so a search that optimizes against those estimates can fit the draws rather than the model. This experiment separates what a search proposes from what an independent sample returns, on a problem small enough to confirm on many fresh draws and large enough for a non-trivial rotation family.

We fit separate neural flows in $d = 6$, giving 15 coordinate-plane rotations per arm. In raw coordinates the control arm is $\mathcal{N}(0, I)$ and the treated arm has standard deviations $(2, .5, 1.45, .72, 1.20, .90)$ with mean shift $0.7u$ for $u = (1, 1, 0, 0, 0, 0)/\sqrt{2}$, so the arms differ in both scale and location. Each arm is standardized on its own training data and the query inverts that transformation (Appendix E.1). Selection runs on $1{,}500$ latents against a target of 77%, below the fitted model's own response fraction so that a search has something to reach for, and confirmation uses $20{,}000$ fresh ones.

The fitted model assigns a positive score effect to 82.34% of generated units, with simultaneous lower bound 81.70%, above the illustrative 80% threshold. All four frozen witnesses confirm below it (Table 3). The Euclidean $-g$ ray has the lowest estimated cost: 77.12% response, upper bound 77.81%, and fresh-bridge price 0.0221 (SE 0.0042), or 0.117% of baseline risk 18.9412. This denominator is total risk, not excess risk above the optimum (Appendix E.1). The M ray costs 0.0229 (SE 0.0043); the treated-only witness costs 0.0606. The d-aware selection price is nearly zero, but its independent price is 0.0616 (SE 0.0070), while a fresh-sample Wald test does not resolve nonzero d (30-df $p = .57$).

The scale of that fraction deserves a word, since the choice of denominator changes it by two orders of magnitude. A nine-node population reference for this model is 18.8747, so dividing the same excess by $18.9412 - 18.8747$ would report about 33% rather than 0.117%. That ratio is not an excess-risk fraction: it divides a finite-target Monte Carlo risk by a gap to a population optimum, and the two are not measured on the same sample. We therefore quote total-risk fractions throughout and give the population references in Appendix E.1; a price that is small on the total scale need not be small relative to how far the fit sits from its own optimum.

Repeated audit draws isolate this optimism with one trained pair, fixed validation targets and a fixed query gradient (Figure 2). Across 20 repetitions at each of three sample sizes, the d-aware fresh-minus-selection gap falls from 0.456 to 0.090 to 0.027, with all intervals excluding zero; the M-ray intervals cover zero. Even the last gap is comparable to the 0.025 budget. The rules reach different hard queries, so this diagnoses price optimism without ranking optimizers (Appendix E.4).

The query also changes the alternative found (Table 2). The proportion-targeted M ray lowers response from 82.34% to 74.31%; the individual-disagreement direction leaves it at 80.36%, despite similar disagreement $H = 0.2183$ and 0.2036 at nearby estimated prices 0.0623 and 0.0628. These prices are not certified equal, and the latter response has a one-sided exact lower bound of 79.89%,

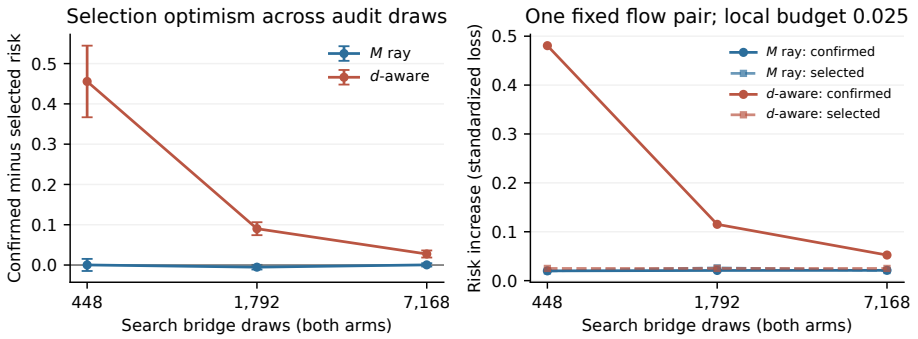

Figure 2: **Audit-sampling optimism with the network held fixed.** Twenty independent bridge repetitions at each size, all reusing one set of weights, one validation split and one query gradient. Left: the mean gap between the fresh price and the selection price, with 95% Student-t intervals across repetitions. Right: both prices at the fixed local budget of 0.025. The two rules constrain differently and reach different hard queries, so the figure diagnoses price optimism and does not rank search directions.

so it does not confirm response above 80%. The comparison shows different prediction profiles, not an optimizer ranking (Appendix E.3). Factual anchoring returns $H = 7 \times 10^{-32}$ for pure common-source relabelling. The applications below likewise audit individual predictions alongside the aggregate conclusion.

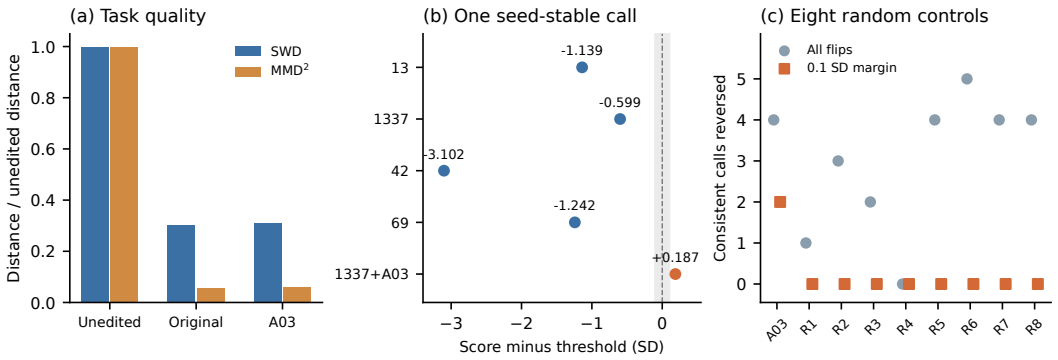

Figure 3: **Retraining agreement does not settle an individual decision.** (a) A03 retains over 99% of task improvement. (b) Four fits call the displayed cell a nonresponder; A03 crosses the threshold, qualifying at the 4% secondary budget. The cell is identified retrospectively. (c) All eight price-matched random controls pass quality and the same budget, but none reverses a consistent call with a 0.1 SD margin on both sides. Grey dots include flips without that margin.

5 Auditing a Released Single-Cell Editor

Single-cell perturbation editors are used to predict what a measured cell would have looked like under a treatment that was not applied to it. Sequencing destroys the cell, so each one is observed in exactly one condition and its counterfactual is never measured: the model supplies it. Those predictions are then read per cell — which cells respond, which to follow up — and a per-cell reading is exactly what the pairing in Equation (1) determines and the arm-wise fit does not.

The released scCBGM-FM editor (Andersson et al., 2026) infers a factual latent from a measured cell, changes the coordinate encoding the perturbation, and decodes, as in Equation (14). We fit every checkpoint locally following the released architecture and schedule, then freeze encoder and flow and edit only stimulation. The readout is a PCA-reconstructed MSigDB Hallmark program score (Liberzon et al., 2015; MSigDB Team, 2026a) in training-control SD: on Kang peripheral blood

Table 1: Six frozen candidates, independently priced on both risks. Approximate Student-t bounds cover candidates, risks and Heun-$64/128$, conditional on fixed cells. All six retain task quality. A03/A04 qualify at the 4% secondary budget only.

	Search budget	RMS change (control SD)	Cells moved ≥ 0.1 SD	Price upper bound (% of baseline) Global	Edited condition	Clears 1%
A01	0.25%	0.225	63.9%	0.282	0.607	yes
A02	0.25%	0.217	57.1%	0.297	0.634	yes
A03	1%	0.369	75.1%	0.759	1.611	no
A04	1%	0.357	73.8%	0.779	1.659	no
A05	4%	0.579	88.0%	2.618	5.533	no
A06	4%	0.565	86.7%	2.674	5.710	no

cells (Kang et al., 2018), the interferon-alpha program, whose induction is the expected response to interferon-β stimulation. Because the edit acts on observed cells rather than on generated samples, Proposition 1 does not cover the edited population, so edit quality is measured on held-out cells rather than assumed. Appendix E.5 gives preprocessing, partitions and protocols.

5.1 Task quality, retraining agreement and individual decisions

Showing that a model's individual predictions can be moved is only interesting if the model was worth trusting on the task, and only if the alternative is not simply a worse editor. Three things are therefore fixed before any search. The editor itself must work here: its edits must close at least 20% of the distance between unedited control cells and held-out stimulated cells on both target distances. An alternative then qualifies only if it gives up at most 10% of that improvement, so prediction change cannot be bought by quietly degrading the edit. Price is charged twice, on the global context mixture the model was trained under and on the edited condition alone, because the perturbation is gated by the stimulation coordinate and a global average would dilute its cost where the edit acts. Finally the budgets — 1% of each calibration risk as primary, with 0.25% and 4% alongside — and the candidate set are frozen before confirmation, so the search cannot grade its own work.

We audit Kang `CD4 T cells_1` control-to-stimulated edits on the frozen seed-1337 checkpoint, a condition whose two labels both occur in training and in every held-out partition, so the model is asked for a prediction it was trained to make. Confirmation uses 233 control cells and 229 stimulated reference cells, disjoint from search and pricing.

Two of six frozen candidates clear the primary budget on both risks and grids (Table 1). A01 retains over 99% of both task improvements while moving predictions by RMS 0.225 control SD; 149 cells (63.9%) move at least 0.1 SD. All effect signs remain positive, but the mean falls from 5.450 to 5.353 SD. Calling a cell a responder when its edited score reaches the training stimulated group's median, the responder fraction falls from 33.5% to 28.8%; just one switch has a 0.1 SD margin on both sides, so most changed calls sit close to the threshold.

A retrospective comparison asks whether retraining agreement settles the remaining calls. Four independent fits all pass task quality on this same panel and scale; 129 cellwise calls agree across every fit and both grids. A03 retains over 99% of both task improvements and reverses four of these calls, two with a two-sided 0.1 SD margin, at the 4% secondary budget but not the 1% primary budget. For the cell in Figure 3, seed 1337 is the closest baseline to the threshold, 0.599 SD below it. A03 moves it 0.786 SD to 0.187 SD above; the other three fits also call it a nonresponder. The cell was identified after inspecting the panel; Appendix E.7 gives the weaker second case and all counts.

To check whether an arbitrary rotation does the same, we freeze eight random directions before querying. All pass quality and the 4% budget; fresh paired pricing places each within a declared $\pm 10\%$ tolerance of A03 on both risks and grids. They reverse zero to five consistent calls, but none with the 0.1 SD margin, compared with two for A03. The clear reversal is therefore not reproduced by these eight price-matched controls, though this small comparison does not establish general search superiority.

None clears its original search budget: A01/A02, for example, were proposed at 0.25%. The cross-budget reporting matrix was specified after partial prices returned, with candidates and thresholds already frozen. Unlike the repeated-draw experiment in Figure 2, this does not isolate sampling optimism because samples, grids and derivatives also differ. These results concern model decisions on a fixed cohort, not validated individual biological effects. Appendix E.6 gives the full qualification and pricing protocol.

5.2 When sensitivity does not establish task value

Two earlier applications explain why sensitivity alone is not enough. On Kang `CD4 T cells_0`, all 60 candidates across four checkpoints leave every confirmation cell's predicted increment positive, an outcome of the tested candidates rather than a sign-invariance guarantee for the family. A later search finds individual changes of RMS 0.100 SD within its reference price budget, but does not pass the full numerical gate. Refitting moves predictions much more, by RMS 1.19–1.64 SD on the same cells. What distinguishes the audit is not a larger change: it keeps the learned contexts and density paths fixed, which refitting need not do. Appendix E.8 reports the failed gates, fresh pricing and refit comparison.

On the Cui mouse cytokine data (Cui et al., 2024), three alternatives replace seven or eight cells in a twelve-cell list ranked by predicted TGF-β1 response: they retain five, five and four cells, with whole-panel rank correlations 0.22, 0.37 and 0.20. Reordering thus extends beyond the cutoff. Yet neither the baseline editor nor the displayed alternative improves the target distances over leaving cells unedited. The displayed 0.0687 SD movement is large beside the scores' 0.0236 SD dispersion and 0.0001 SD cutoff gap. Those scales help interpret turnover but do not determine it: a common shift can move every score without changing any rank. Global price estimates are unresolved from zero and dominated by a few sampled cells. This is not a task-qualified result: the ranking rule is illustrative, and shortlist turnover alone establishes neither useful editing nor practical failure. Appendix E.9 retains the price and quality diagnostics.

6 Related work

The closest predecessor, Nasr-Esfahany & Kiciman (2023), uses treatment-dependent Gaussian rotations to demonstrate non-identifiability (their Section 5.3) and optimizes a second generator for loss-constrained disagreement at the same factual evidence (their Sections 6–7). Our distinction is to construct the alternative through the existing flow, preserving its complete conditional path and pricing it by a precomputable quadratic under the original validation law (Appendix D).

This is a structured instance of predictive multiplicity (Marx et al., 2020) and underspecification (D'Amour et al., 2022): validation-equivalent models can behave differently downstream. Those literatures reach competitors by retraining or, for linear classifiers, exact combinatorial search; our alternatives keep the weights fixed.

Structural approaches instead identify counterfactuals by adding bijective, cocycle or dynamic-transport restrictions (Nasr-Esfahany et al., 2023; Dance & Bloem-Reddy, 2025; De Sousa Ribeiro et al., 2025), and causal sensitivity analysis uses interpretable curvature, rank or transport restrictions (Melnychuk et al., 2023; Masten & Poirier, 2020; De Lara et al., 2024); training-risk cost complements these by calibrating executable alternatives against the fitted objective. Within flow matching, bridge couplings serve training efficiency (Pooladian et al., 2023) and single-arm Gaussian geometry is known (Noogram, 2026), whereas our object is the cross-world coupling induced by sharing a code.

7 Limitations and conclusion

Good distribution-level performance and agreement across refits do not settle individual predictions. At the primary budget, an alternative retains over 99% of task improvement yet moves nearly two thirds of the held-out panel by at least 0.1 SD. At the 4% secondary budget, another reverses a decision shared by four fits, away from the threshold. Report the decision rule, the confirmed alternative and its independent price uncertainty; for rankings and responder calls, show cutoff gaps as well as reversals. These results cover one released method, two datasets, five checkpoints and fixed

cohorts of existing donors. They concern model predictions rather than biological effects, and the restricted rotation family cannot certify robustness.

What such an audit licenses is correspondingly narrow. A confirmed alternative shows only that the family's minimum conclusion-changing cost is at most the price paid, so the artifact worth publishing is the query–cost profile of Section 3: the frozen model and query, the family searched, the best confirmed alternative with its independent price uncertainty, and whether confirmation used generated units, a fixed panel or new subjects. Finding nothing bounds nothing, since a wider family or a better search may still succeed. What the audit supplies is a price on the table next to the prediction; which prices are acceptable remains the reader's judgement.

Reproducibility statement

The local source records public-data checksums, model provenance, frozen protocols, all candidates including failures, per-cell bridge contributions, confirmation arrays and tests. Saved arrays support replay of the reported summaries and bounds; Appendix E distinguishes population, search and confirmation samples. Compact-flow data and weights match archived SHA-256 digests, and Windows replay recovers the baseline positive-effect fraction on saved latents. This verifies artifact identity and one statistic, not regeneration of every experiment; seed-based regeneration still differs locally (Appendix E.3). No public anonymous URL is claimed in this working draft.

Superseded compact-PBMC pilot records, including the withdrawn retraining comparison, remain in the supplement with a study-status note. They document earlier TorchCFM execution, not current application performance or a search advantage; the adapter has separate executable compatibility tests.

AI use statement

Generative AI assisted literature review, mathematical development, experimental design, implementation, data preparation, interpretation, figures and revision. We have completed our review and verification of the manuscript, code and results, including the AI-assisted work, and approved this disclosure. We take responsibility for the final content, including all claims and artifacts produced with generative AI assistance.

References

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023.

Akram Aldroubi, Shiying Li, and Gustavo K. Rohde. Partitioning signal classes using transport transforms for data analysis and machine learning. *Foundations of Data Science*, 3(3):371–387, 2021. doi: 10.3934/fods.2021009.

Alma Andersson, Aya Abdelsalam Ismail, Edward De Brouwer, Doron Haviv, Tommaso Biancalani, Kyunghyun Cho, Gabriele Scalia, Aïcha BenTaieb, and Hector Corrada Bravo. scCBGM: Interpretable single-cell counterfactual editing. *arXiv preprint arXiv:2606.07760*, 2026. URL https://arxiv.org/abs/2606.07760. Accepted to ICML 2026; official implementation: github.com/almaan/scCBGM.

Franck Boyer. Ordinary differential equations and partial differential equations, 2023. URL https://www.math.univ-toulouse.fr/~fboyer/_media/enseignements/m1edoedp/poly_anglais_edo-edp_m1.pdf. Lecture notes, Paul Sabatier University, September 5; Theorem I.35.

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

Ang Cui, Teddy Huang, Shuqiang Li, Aileen Ma, Jorge L. Pérez, Chris Sander, Derin B. Keskin, Catherine J. Wu, Ernest Fraenkel, and Nir Hacohen. Dictionary of immune responses to cytokines

at single-cell resolution. *Nature*, 625:377–384, 2024. doi: 10.1038/s41586-023-06816-9. URL https://www.nature.com/articles/s41586-023-06816-9.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23:1–61, 2022. URL https://www.jmlr.org/papers/v23/20-1335.html.

Hugh Dance and Benjamin Bloem-Reddy. Counterfactual cocycles: A framework for robust and coherent counterfactual transports. *arXiv preprint arXiv:2405.13844*, 2025. URL https://arxiv.org/abs/2405.13844.

Lucas De Lara, Alberto González-Sanz, Nicholas Asher, Laurent Risser, and Jean-Michel Loubes. Transport-based counterfactual models. *Journal of Machine Learning Research*, 25(136):1–59, 2024.

Fabio De Sousa Ribeiro, Ainkaran Santhirasekaram, and Ben Glocker. Counterfactual identifiability via dynamic optimal transport. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2510.08294.

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, et al. Causal identification of single-cell experimental perturbation effects with CINEMA-OT. *Nature Methods*, 20:1769–1779, 2023. doi: 10.1038/s41592-023-02040-5.

Juha Heinonen. *Lectures on Lipschitz Analysis*, volume 100 of *Report, Department of Mathematics and Statistics*. University of Jyväskylä, 2005. URL https://www.math.jyu.fi/research/reports/rep100.pdf.

Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M. Lanata, Rachel E. Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A. Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, 2018. doi: 10.1038/nbt.4042. URL https://www.nature.com/articles/nbt.4042.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015. doi: 10.1016/j.cels.2015.12.004. URL https://pubmed.ncbi.nlm.nih.gov/26771021/.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Charles Marx, Flavio du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6765–6774, 2020. URL https://proceedings.mlr.press/v119/marx20a.html.

Matthew A. Masten and Alexandre Poirier. Inference on breakdown frontiers. *Quantitative Economics*, 11(1):41–111, 2020. doi: 10.3982/QE1288.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Partial counterfactual identification of continuous outcomes with a curvature sensitivity model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/65cbe3e21ac62553111d9ecf7d60c18e-Abstract-Conference.html.

Caroline Moosmüller and Alexander Cloninger. Linear optimal transport embedding: Provable wasserstein classification for certain rigid transformations and perturbations. *Information and Inference: A Journal of the IMA*, 12(1):363–389, 2023. doi: 10.1093/imaiai/iaac023.

MSigDB Team. HALLMARK_INTERFERON_ALPHA_RESPONSE (M5911), 2026a. URL https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARK_INTERFERON_ALPHA_RESPONSE.html. Human gene set, version 2026.1.Hs; retrieved 2026-09-05.

MSigDB Team. HALLMARK_TGF_BETA_SIGNALING (MM3865), 2026b. URL https://www.gsea-msigdb.org/gsea/msigdb/mouse/geneset/HALLMARK_TGF_BETA_SIGNALING.html. Mouse gene set, version 2026.1.Mm; retrieved 2026-09-07.

Arash Nasr-Esfahany and Emre Kiciman. Counterfactual (non-)identifiability of learned structural causal models. *arXiv preprint arXiv:2301.09031*, 2023. URL https://arxiv.org/abs/2301.09031.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25733–25754, 2023. URL https://proceedings.mlr.press/v202/nasr-esfahany23a.html.

Noogram. Flow matching on gaussians: When the gaussian flow map is the optimal-transport map, 6 2026. URL https://github.com/noogram-labs/flow-matching-gaussians. Software and technical note, version 1.0.0, released 2026-06-22.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127, 2023.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification; software: torchcfm.

Dongze Wu, David I. Inouye, and Yao Xie. Flow-based generative modeling of potential outcomes and counterfactuals. *arXiv preprint arXiv:2505.16051*, 2026. Accepted at IEEE ISIT 2026.

A Exact Gaussian calibration

The Gaussian reference has no fitting error. We first vary its source, then freeze one model for the rotation audit in Section 4.1. Let $\nu_S = \mathcal{N}(0, S)$ and $Q_a = \mathcal{N}(\mu_a, \Sigma_a)$ with $S, \Sigma_a \succ 0$; square roots are principal symmetric roots. The single-arm solution is known (Noogram, 2026).

Proposition 3 (Population independent-CFM map)**.** *For the bridge in Equation* (2)*, the population square-loss minimizer has endpoint* $G^{\mathrm{CFM}}_{a,S}(z) = \mu_a + F_{a,S}(1)z$*, where*

$$F_{a,S}(1) = S^{1/2}(S^{-1/2}\Sigma_a S^{-1/2})^{1/2}S^{-1/2}.$$

It transports ν_S *to* Q_a *exactly; Appendix B.1 gives the velocity and full time map.*

Sharing codes selects the two-arm map $T^{\mathrm{CFM}}_S = F_{1,S}(1)F_{0,S}(1)^{-1}$. For quadratic-cost optimal transport, write $T^{\mathrm{OT}}_S = B_1(S)B_0(S)^{-1}$, where $B_a(S) = S^{-1/2}(S^{1/2}\Sigma_a S^{1/2})^{1/2}S^{-1/2}$.

Corollary 1 (Reciprocal-prior identity)**.** $F_{a,S}(1) = B_a(S^{-1})S^{-1}$*, and therefore*

$$T^{\mathrm{CFM}}_S = T^{\mathrm{OT}}_{S^{-1}}. \tag{16}$$

Either transport is invariant to every *Gaussian source* $S \succ 0$ *exactly when* $\Sigma_1 = c^2\Sigma_0$ *for some* $c > 0$*, in which case it equals* cI*.*

Inverting the source covariance exchanges the rules' counterfactual transports while preserving exact marginal generation. Source invariance requires proportional arm covariances, specializing a known compatibility condition in linear optimal transport (Aldroubi et al., 2021; Moosmüller & Cloninger, 2023).

Along the two-dimensional reciprocal source path in Appendix B.1, increasing ϵ from -0.35 to $+0.35$ raises $p_B = \Pr\{u^\top(Y_1 - Y_0) > 0\}$ from 76.12% to 87.27% under optimal transport and lowers it under independent CFM. Both give 81.20% at $S_0 = I$ (Figure 4(a)). Every point is a population optimum with identical target marginals and average score effect 0.7; the largest covariance error is 1.3×10^{-13}. The eleven-point responder swing therefore comes from the training rule and source alone.

B Proofs

B.1 Proof of Proposition 3

Joint Gaussian regression gives $\mathbb{E}[U_a \mid X_{a,t} = x]$ directly. Its mean is μ_a; the bridge mean is $t\mu_a$; the bridge covariance is $C_t = (1-t)^2 S + t^2\Sigma_a$; and $\mathrm{Cov}(U_a, X_{a,t}) = t\Sigma_a - (1-t)S$. The population square-loss minimizer is therefore

$$v^*_a(t,x) = \mu_a + \{t\Sigma_a - (1-t)S\}\{(1-t)^2 S + t^2\Sigma_a\}^{-1}(x - t\mu_a). \tag{17}$$

For the flow, set $W_a = S^{-1/2}\Sigma_a S^{-1/2}$ and work in source-whitened coordinates. The centered covariance path is $H_t = (1-t)^2 I + t^2 W_a$. Both H_t and $\dot{H}_t$ are functions of W_a and commute, so

$$\frac{d}{dt}H_t^{1/2} = \frac{1}{2}H_t^{-1/2}\dot{H}_t = \{tW_a - (1-t)I\}H_t^{-1/2}.$$

Thus $q_t = H_t^{1/2}q_0$ solves the whitened version of Equation (17). Transforming back gives the centered time map

$$F_{a,S}(t) = S^{1/2}\left[(1-t)^2 I + t^2 S^{-1/2}\Sigma_a S^{-1/2}\right]^{1/2}S^{-1/2}. \tag{18}$$

At $t = 1$, $F_{a,S}(1)SF_{a,S}(1)^\top = \Sigma_a$, proving the endpoint claim.

For Figure 4(a), we use $\Sigma_0 = I$, $\Sigma_1 = \mathrm{diag}(4, 0.25)$, $u = (1,1)^\top/\sqrt{2}$, $\mu_1 - \mu_0 = 0.7u$, and $S_\epsilon = \exp(2\epsilon K)$ with $K_{11} = K_{22} = 0$ and $K_{12} = K_{21} = 1$. Since $S_{-\epsilon} = S_\epsilon^{-1}$, Equation (16) makes the responder-probability curves reflections of one another.

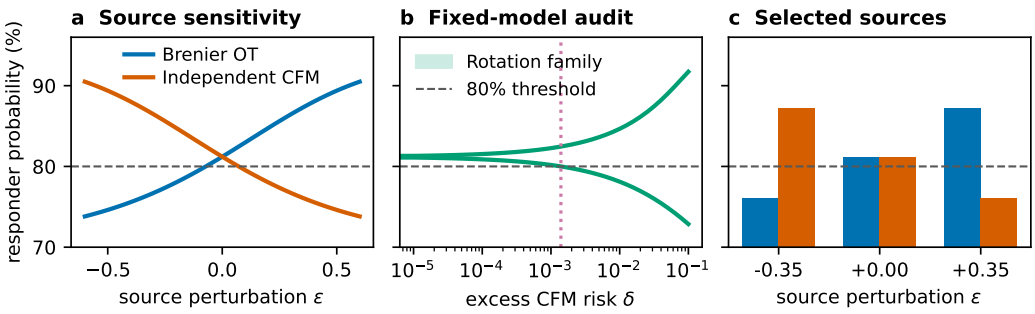

Figure 4: **Exact calibration.** (a) OT and independent CFM have reciprocal source sensitivity (Equation (16)). (b) Density-path-preserving rotations at fixed source and objective; the vertical line marks the first 80% flip. (c) Selected sources. Panels share the probability scale, and target marginals and the average score effect are identical throughout.

B.2 Proof of Corollary 1

Substituting S^{-1} into the Brenier form gives

$$B_a(S^{-1}) = S^{1/2}(S^{-1/2}\Sigma_a S^{-1/2})^{1/2}S^{1/2} = F_{a,S}(1)S.$$

Hence $F_{a,S}(1) = B_a(S^{-1})S^{-1}$; the common right factor cancels when the two arms are composed.

If $\Sigma_1 = c^2\Sigma_0$, homogeneity of the square root yields $B_1(S) = cB_0(S)$ for every S, proving sufficiency. Conversely, suppose $B_1(S)B_0(S)^{-1} = A$ for every $S \succ 0$. Take $S = \Sigma_0$ to see that A is symmetric positive definite. For any symmetric positive definite H, choose $S = H^{-1}\Sigma_0 H^{-1}$; uniqueness of the Brenier map gives $B_0(S) = H$. Invariance forces $B_1(S) = AH$. Since every Brenier map is symmetric, $AH = HA$ for every $H \succ 0$. Therefore $A = cI$, and pushing S forward gives $\Sigma_1 = c^2\Sigma_0$. The CFM statement follows from the reciprocal identity.

B.3 Proof of Proposition 1

Skew symmetry implies $\frac{d}{dt}(R_{a,\beta,t}^\top R_{a,\beta,t}) = 0$, hence $R_{a,\beta,t}$ is orthogonal and preserves $\mathcal{N}(0, I_d)$. Pushforward composition immediately yields Equation (7). Along a perturbed trajectory $x_t = \Phi_{a,t}(R_{a,\beta,t}z_0)$, write $z_t = R_{a,\beta,t}z_0 = \Phi_{a,t}^{-1}(x_t)$. The chain rule gives

$$\dot{x}_t = v_a(t, x_t) + D\Phi_{a,t}(z_t)K_{a,\beta}(t)z_t,$$

which is Equation (8). Linearity in β follows from the definition of $K_{a,\beta}$.

For the scCBGM editor we need a Lipschitz extension. At fixed context its evaluation-mode decoder is continuous in time and globally Lipschitz in state: affine, ReLU and residual layers preserve Lipschitz continuity, and LayerNorm has bounded derivative because its normalization epsilon is positive. Its exact ODE defines spatial bi-Lipschitz homeomorphisms, jointly locally Lipschitz in (t, z), though not necessarily C^1 (Boyer, 2023, Theorem I.35). Spatial derivatives are interpreted almost everywhere.

For the continuously differentiable rotations used here, $\Psi_t(z) = \Phi_t(R_t z)$ has absolutely continuous trajectories. Rademacher's theorem (Heinonen, 2005, Theorem 3.1), the chain rule and Fubini give $\partial_t\Psi_t(z) = v_t(\Psi_t(z)) + D\Phi_t(R_t z)\dot{R}_t z$ for almost every (t, z); $(t, z) \mapsto (t, R_t z)$ preserves Lebesgue null sets. Orthogonal invariance preserves the Gaussian path at every time. Conditional on each frozen target/context, the original Gaussian bridge has a Lebesgue density for $t < 1$, so uniform-time pricing ignores derivative exceptional sets and retains the square-integrable risk identity. The composed generator is defined everywhere, but this argument gives neither uniqueness for arbitrary solutions of its nonsmooth Eulerian field, numerical error bounds, nor a classical Jacobian at every factual cell.

B.4 Proof of Proposition 2 and Equation (10)

Expanding the square under the fixed original validation law gives

$$\|r_a + \sum_j \beta_j b_{a,j}\|^2 - \|r_a\|^2 = 2\sum_j \beta_j r_a^\top b_{a,j} + \sum_{j,k} \beta_j \beta_k b_{a,j}^\top b_{a,k}.$$

Taking expectations and summing arms proves Equation (9); M is a Gram matrix. At the population optimum $v_a^*(t, X) = \mathbb{E}[U_a \mid t, X]$, so the conditional expectation of r_a given (t, X) is zero and $d = 0$.

For the local envelope, suppose $M \succ 0$, $d = 0$, and Θ_s is twice continuously differentiable on a neighborhood of zero with Hessian operator norm bounded by C. Taylor's theorem gives $\Theta_s(\beta) = \Theta_s(0) + g^\top \beta + r(\beta)$ with $|r(\beta)| \le C\|\beta\|^2/2$. On $\beta^\top M \beta \le \delta$, $\|\beta\|^2 \le \delta/\lambda_{\min}(M)$, so the remainder is uniformly $O(\delta)$. Cauchy–Schwarz in the M inner product gives the leading term in Equation (10).

For Equation (11), complete the square:

$$2d^\top \beta + \beta^\top M \beta = (\beta + M^{-1}d)^\top M (\beta + M^{-1}d) - d^\top M^{-1} d.$$

The feasible set is therefore an M-ellipsoid centered at $-M^{-1}d$. Minimizing the linear form $g^\top \beta$ over its boundary by the same Cauchy–Schwarz argument gives Equation (11). This is the exact solution of the *linearized-query* problem, not of the finite hard-query problem.

B.5 Derivation of the analytic rotation price

For $x = L_{a,t} z$, a planar latent rotation of instantaneous rate $\omega_a(t)$ pushes to the field $b_a = \omega_a L_{a,t} J L_{a,t}^{-1} x$. Under the path law $z \sim \mathcal{N}(0, I)$,

$$\mathbb{E}\|b_a\|^2 = \omega_a(t)^2 \operatorname{tr}(L_{a,t}^2) = k_a(t)\omega_a(t)^2.$$

Subject to $\int_0^1 \omega_a(t)dt = \theta_a$, weighted Cauchy–Schwarz gives minimum θ_a^2/I_a, attained at $\omega_a(t) = \theta_a/[I_a k_a(t)]$. Minimizing $\theta_0^2/I_0 + \theta_1^2/I_1$ subject to $\theta_1 - \theta_0 = \theta$ yields Equation (15). Direct Gaussian variance calculation gives the responder formula in the main text.

C Algorithm and implementation details

The curvature expansion for H requires the twice continuous differentiability stated in Section 3. The Gaussian flows use tanh activations, but scCBGM's almost-everywhere ReLU derivatives do not imply the cubic remainder. Its central-difference C therefore proposes candidates for evaluation with the actual composed editor. Gaussian queries use separate latent draws; scCBGM follow-ups reuse a fixed factual panel with a new price sample. At the smallest Gaussian amplitude the quadratic approximation is within 1.4% on the search panel. For scCBGM, the Jacobian's column sums match the autograd derivative of the summed effect within 8×10^{-4} relative. This aggregate check does not certify every per-cell derivative or the Gram matrix.

For each validation bridge point (t_i, x_i, u_i), invert $z_i = \Phi_t^{-1}(x_i)$ and estimate

$$b_k(t_i, x_i) = D\Phi_t(z_i) K_k z_i \approx \frac{\Phi_t(z_i + \eta K_k z_i) - \Phi_t(z_i - \eta K_k z_i)}{2\eta}.$$

The compact implementation batches all basis directions into two ODE solves, accumulates empirical d, M and the covariance of $\widehat{d}$, and symmetrizes M. The ridge is recorded below. Query gradients use centered differences with common search latents. The six-dimensional experiments use separate 15-coefficient arm blocks and constant generators with endpoint $G_a(\exp K_{a,\beta} Z)$; the conditional editor uses Equation (14) with a shared context gate.

The compact result JSON records coefficients, queries, simultaneous exact-binomial bounds, pointwise Wilson diagnostics, selection and fresh prices, Monte Carlo SEs, the $\widehat{d}$ Wald diagnostic, runtime and marginal checks. A combined-direction finite-difference price uses two perturbed solves without reconstructing M. Conditional applications retain independent per-draw d_i, M_i arrays and donor-aware intervals for every frozen candidate.

An adapter must expose the original loss coordinates and normalization, bridge law, velocity, conditioning representation, forward/inverse transport and symmetric source, and distinguish conditional generation from factual editing. A Gaussian source supports the rotations; the fixed validation law supports Proposition 2, including for coupled or noisy bridges. ODE, Jacobian and factual-inverse consistency require implementation-specific checks.

The adapter accepts TorchCFM 1.0.7's `ConditionalFlowMatcher` and `MLP` (Tong et al., 2024) with an isotropic Gaussian source and independent straight bridge ($\sigma = 0$). Tests check forward–inverse consistency and quadratic versus direct finite-rotation pricing on shared draws. With a local training loop and locally fitted weights, this establishes library compatibility, not validation of an author-released checkpoint.

D Relation to loss-constrained disagreement

	Nasr-Esfahany & Kiciman (2023), Sections 6–7	This audit
Variable optimized	Parameters of a second conditional generator/inference model	Source-symmetry coefficients of one fixed fitted flow
Fit control	Equality ideal; scalarized or threshold penalty in implementation	Complete model density path exact by construction; original CFM risk priced exactly in β
Counterfactual target	User-chosen discrepancy between predictions at the same factual evidence	A hard conclusion or factual-anchored individual disagreement
Statistical report	Training trade-off curve	Query-specific confirmation; separate risk uncertainty and fixed-panel summaries
Scope	General learned SCM mechanism class	Restricted executable alternatives around a flow with a symmetric source

The predecessor's Section 5.3 provides the endpoint-rotation argument, and Section 6.1.2 permits chosen discrepancies at the same factual evidence. Our distinction is complete-path preservation and original-FM pricing around frozen weights. The table claims neither lower prices nor faster audits than retraining. Seed refits in Appendix E.8 measure prediction sensitivity without benchmarking loss-constrained disagreement optimization.

E Experimental details and additional checks

E.1 Loss units, quadrature, and uncertainty

The compact flows set `standardize_outcomes=True`, storing each arm's training mean and coordinatewise standard deviation (`ddof=0`) for validation. Reported squared residuals are summed over coordinates and arms; queries invert the affine transformation. This is $2d$ times the sum of scikit-learn's per-arm half mean-squared data losses, excluding the parameter penalty (`alpha=1e-5`); `loss_` is not the denominator. scCBGM-FM retains its native mean-per-feature loss and context-mixture weighting, so raw prices are not compared across implementations.

For isotropic source and diagonal target covariance with standard deviations σ_{aj}, Gaussian conditional variance gives

$$\mathcal{R}_{\mathrm{Bayes}} = \sum_{a,j} \int_0^1 \frac{\sigma_{aj}^2}{(1-t)^2 + t^2\sigma_{aj}^2}\, dt = \frac{\pi}{2} \sum_{a,j} \sigma_{aj}.$$

Target means do not enter. Raw-coordinate risk is $\pi(6 + 6.77)/2 = 20.0591$, ideal population-standardized risk is $6\pi = 18.8496$, and conditioning on actual training scales gives 18.8500. Seven- and nine-node midpoint approximations give 18.8908 and 18.8747. The fixed-witness baseline is 18.9412, hence the total-risk fraction $0.0221438/18.9412 = 0.116908\%$. Population values do not bound risk under finite held-out targets. Subtracting the nine-node reference would give $0.0221438/(18.9412 - 18.8747) \approx 33\%$, but mixes a population optimum with finite-target Monte Carlo risk and does not estimate the true excess-risk fraction. A small total-risk fraction need not be small on that scale.

Training samples $t \sim \mathrm{Unif}[0,1]$. Pricing uses $t_j = (j+1/2)/T$, $j = 0, \dots, T-1$, with weight $1/T$, independent source draws, and uniform resampling of frozen validation targets at each node. Gaussian selection uses $(T, n) = (7, 192)$ per arm; the diagnostic $\widehat{d}$ uses $(9, 256)$ and final fixed-witness pricing $(9, 1024)$. For a pointwise loss difference with within-node sample variance s_{aj}^2, the variance of the two-arm estimate is $\sum_{a,j} s_{aj}^2/(T^2 n)$. The covariance of $\widehat{d}$ is accumulated with the same within-node centering. These SEs condition on networks and held-out targets, and do not include quadrature, numerical integration, or training error.

Within each study, baseline lower and candidate upper bounds divide a binomial error budget 0.05 equally across one-sided exact tails. Bonferroni coverage allows shared confirmation latents. The JSON also records approximate simultaneous Monte Carlo intervals $\widehat{\Delta\mathcal{R}} \pm z_{1-0.05/(2K)}\widehat{\mathrm{SE}}$ for K frozen costs. Query and cost bands have separate 95% targets, not a joint guarantee. All candidates are reported; the lowest estimated price does not establish population superiority.

E.2 Exact Gaussian experiments

All matrix powers use eigendecompositions of symmetrized positive-definite matrices. The source path $S_\epsilon = \exp(2\epsilon K)$ is evaluated on 49 equally spaced points from -0.6 to 0.6. Responder probabilities are analytic normal probabilities. Numerical quadrature for I_a uses absolute and relative tolerance 10^{-13}, and the flip angle is root-solved to tolerance 10^{-14}. The risk envelope uses 41 logarithmically spaced budgets from 10^{-5} to 10^{-1} plus zero. The largest covariance reconstruction error across the whole sweep is 1.3×10^{-13}.

E.3 Learned Gaussian experiment

Each arm uses 8,000 training observations. The MLP has two 80-unit tanh layers and input features $(x, t, \sin \pi t, \cos \pi t)$; it is trained for 180 epochs with 10 minibatches of size 384, Adam step size 2×10^{-3}, and 36-step Heun integration. Selection uses 4,000 validation targets, seven time points, 192 bridge draws per time, 1,500 query latents, 30 arm-specific coefficients, gradient step 0.012, and sigmoid temperature 0.12. The pre-specified search target is 77%. Confirmation uses 20,000 new latents and 18,432 fresh bridge contributions per candidate.

The compact tables report supplied experiment records. The artifact contains four training/validation blocks, three query-latent arrays and both serialized flows, with per-file SHA-256 hashes and the producing environment. Data and weight digests recomputed in the drivers' order match the archive. Windows replay with NumPy 2.5 and scikit-learn 1.9 exactly recovers the baseline fraction 0.82335 on all $20{,}000$ saved confirmation latents, without refitting or fresh draws. This verifies object identity and one statistic, not all searches, prices or sampling experiments. Local seed-based regeneration still produces different data and weight digests; explicit diagonal factors remove an SVD basis choice but not all cross-stack variation. The verifier records both routes and environments without overwriting archived expectations.

On the recorded CPU rerun, risk-matrix construction takes 27 seconds, the query gradient 65 seconds, ray searches 45–51 seconds, and the d-aware path 5 seconds after precomputation. Concurrent workstation load prevents controlled speed comparisons. For the Euclidean witness, 24, 36 and 48 Heun steps give fixed-draw prices 0.0240649–0.0240664; JVP steps 10^{-4}–5×10^{-4} agree to displayed precision. Five to nine time points give 0.024–0.032, within Monte Carlo SEs 0.012–0.016; the 5,000-latent query stays 76.88%. Integrating $v + b$ reduces endpoint RMSE relative to $G_a(\exp K_a Z)$ from about 10^{-3} at 8 outer steps to 6.0×10^{-5} (control) and 5.0×10^{-5} (treatment) at 24.

Seven amplitudes per ray also give descriptive interpolated curves, with H evaluated on $20{,}000$ separate latents. Price SEs of 0.0016–0.0090 are not propagated through interpolation. Individual-direction H is 2.07–2.60 times random and 0.85–0.93 times the proportion ray; these are not executable equal-cost comparisons. The smallest-amplitude quadratic approximation errs by 1.4% on the search panel, with regularized exchange rate $\lambda = 3.28$.

For executed comparisons, search uses the individual-curvature direction, proportion-targeted M ray and a random direction chosen from 24 proposals by search-panel H at budget 0.025, then fixed across seven amplitudes. Against random, the selected pair costs 0.0100 and 0.0096 (paired

Table 2: **The audit objective changes the prediction profile.** Two executed Gaussian candidates with nearby price estimates; H is factual-anchored disagreement and parenthesized costs are one standard error. These are descriptive profiles, not a cost-equivalence test or an optimizer ranking.

Search target	Response (%)	H	Price (1 SE)
Individual (H)	80.36	0.2036	0.0628 (0.0066)
Response fraction	74.31	0.2183	0.0623 (0.0068)

Table 3: **Independent confirmation on the learned Gaussian model.** U_{sim} is the exact-binomial Bonferroni upper bound and the baseline lower bound is 81.70%. Costs use fresh bridge draws with separate approximate simultaneous bands. These are generated-unit queries, not factual-cell edits, with losses summed over standardized coordinates and arms (Appendix E.1).

Alternative	$\widehat{p}$ (%)	U_{sim} (%)	$\widehat{\Delta\mathcal{R}}$ (SE)	Sim. cost interval
$-g$	77.120	77.808	0.0221 (0.0042)	[0.0116, 0.0327]
M ray	77.055	77.744	0.0229 (0.0043)	[0.0122, 0.0336]
Treated only	76.650	77.343	0.0606 (0.0064)	[0.0445, 0.0766]
d-aware	76.245	76.942	0.0616 (0.0070)	[0.0442, 0.0791]

difference $+0.0004 \pm 0.0037$) and reaches $H = 0.0354$ and 0.0167 (difference $+0.0188 \pm 0.0004$). Against the M ray, costs are 0.0628 and 0.0623 ($+0.0005 \pm 0.0105$), with $H = 0.2036$ and 0.2183 (deficit 0.0147 ± 0.0030); $\pm$ denotes one SE. At matched selection budgets, the individual/proportion H ratio is 1.28–2.84. Its contrast with the nearby-estimated-price comparison shows dependence on the pricing convention without isolating sampling, selection and time-grid effects.

Each reported pair minimizes the absolute estimated-price gap over 49 executed combinations. Within arm/time strata, paired risk variance uses per-draw differences; H variance uses per-latent differences of squared effect changes. Ordinary normal 95% intervals for individual minus random are $[-0.00683, 0.00760]$ for price and $[0.01806, 0.01945]$ for H; individual minus proportion gives $[-0.02008, 0.02098]$ and $[-0.02044, -0.00887]$. Price intervals are unadjusted for nearest-price selection, and no equivalence margin was specified. Independent latents resolve these selected-pair H differences conditional on the fitted models and bridge-based selection, but support neither simultaneous claims across pairs nor a ranking at equal true cost. Interpretation records preserve the source JSON hashes and numerical estimates.

E.4 Repeated audit sampling on fixed networks

All 60 audits reuse one pair trained with the Gaussian configuration above, its training/validation sets, and the gradient from 1,500 search latents. A weight-and-preprocessing SHA-256 identifies this fit without assuming identity with another training run. We use 20 independent repetitions at each of $n = 32, 128, 512$ draws per arm and midpoint node. Seven nodes give 448, 1,792 or 7,168 total search draws; samples are neither pooled nor nested across sizes or repetitions.

Both methods minimize the fixed linear surrogate $g^\top \beta$. The M-ray constraint is $\beta^\top(\widehat{M} + \lambda I)\beta \leq 0.025$; the d-aware constraint adds $2\widehat{d}^\top\beta$. Here $\lambda = \max\{10^{-6}, 10^{-5}\,\mathrm{tr}(\widehat{M})/30\}$. We always report the full un-ridged risk, including $\widehat{d}$, for both methods. The budget is fixed before this run, not publicly preregistered. These are local-surrogate candidates, not repetitions of the nonlinear threshold-search algorithm.

After freezing both candidates, each repetition uses 7,168 new bridge contributions for paired pricing and 4,000 new query latents. Methods share confirmation draws within repetitions; repetitions and sizes have distinct seeds. Confirmation never filters or updates candidates. Analytic MLP derivatives propagate tangents through the discrete Heun map; eight central-difference checks give maximum relative JVP discrepancy 8.9×10^{-10}, alongside endpoint and risk checks. The recorded CPU study takes about 24 minutes under concurrent workstation load.

For d-aware search, mean confirmation-minus-selection prices are 0.4556 (95% mean interval $[0.3668, 0.5445]$), 0.0902 ($[0.0741, 0.1063]$), and 0.0274 ($[0.0186, 0.0361]$). The corresponding M-

ray means are 0.0002, -0.0053, and 0.0004, with all mean intervals containing zero. Intervals use a Student-t approximation over independent repetitions and remain conditional on this one fitted pair and fixed target data. Per-candidate query bounds are pointwise diagnostics, not a simultaneous claim across all 120 evaluations.

E.5 Shared scCBGM preparation and the earlier Kang protocol

The preprocessing and model configuration below are shared across the Kang studies. The search family, score and confirmation protocol here describe the earlier `CD4 T cells_0` audit. Appendix E.6 gives the current `CD4 T cells_1` task's rank-12 family, quality checks and pricing.

The official implementation is pinned to commit `5436a2bc3f0e341e755cfa6de2c9636095f2854c`. The OSF real-data archive has SHA-256 `bcabad08320756878a9b59c1b4cf58c7416f67a882b467100addd178361659d2`; its Kang file contains $24{,}054$ cells, $3{,}000$ genes and eight donors (`obs.ind`). We normalize its integer counts to library size 1214 and apply `log1p` once; the released expression matrix is already log-normalized. Retaining the released genes and subtype annotations retains their transductive full-data selection. A new 128-dimensional PCA uses training cells only, randomized solver and seed 20260905; this solver and the audit split differ from the upstream loader.

Before encoder or PCA fitting, we split eligible cells by donor and `L2_stim` stratum into train/calibration/search/confirmation at $60/15/12.5/12.5\%$. Unlike the upstream training split, this excludes all $2{,}743$ stimulated `CD4 T cells_0` cells for target evaluation. That subtype's factual control counts are $1{,}646/411/344/342$, with all eight donors in each partition. The two-stage concept-bottleneck and conditional FM models use a 128-dimensional representation, four 1024-unit hidden layers, 200 epochs per stage, batch size 128, Adam learning rate 3×10^{-4} and decay 0.997. Predeclared seeds are 13, 42, 1337 and 69. The adapter retains released model definitions and training loops through a new array-based launcher, rather than the Hydra benchmark. Editing is deterministic from time zero, without added noise or classifier-free guidance. Pricing uses the frozen evaluation-mode network's native mean-per-feature error on the original independent Gaussian bridge and context mixture, excluding dropout-averaged training-state risk.

A random orthonormal $P \in \mathbb{R}^{128\times 6}$ (seed 90605) embeds 15 coordinate-plane skew generators as $A_j = PJ_jP^\top$. The shared gate uses the actual floating stimulation coordinate: $R_{h,\beta,t} = \exp(tc_{\text{stim}}B_\beta)$, $B_\beta = \sum_j \beta_j A_j$, with coefficient norm capped at four. This restricted family has 15 coefficients, compared with 8,128 for an unrestricted skew matrix in dimension 128; all 128 unknown concepts and unedited known concepts stay frozen. Search uses 128 factual controls, sigmoid temperature one training-control score SD and 256 original bridge draws. Euclidean, M-ray and the best of six frozen random directions at five budgets give 15 frozen candidates and 40 retained search points. Budget roots and final prices keep the full linear term; confirmation never changes candidates.

The frozen MSigDB set has 97 genes, 27 in the released features, above the predeclared minimum of 20. The equal-weight score uses the same frozen PCA reconstruction for factual and edited outcomes, scaled by `CD4 T cells_0` training-control SD (0.07211 log-expression units). Its factual-confirmation projection RMSE is 0.557 SD, so it measures the representation rather than unprojected gene expression. Genes and zero-effect/majority thresholds are fixed independently of audit results. Summaries condition on the cohort and model, with donor-aware, finite-cell and numerical uncertainty distinguished.

The one-SE budget uses 3,202 calibration cells and 32 bridge draws each. All checkpoints query the same 342 new factual controls from the eight training donors; neither checkpoints nor repeated latent or bootstrap draws increase the biological sample size. Fresh pricing uniformly samples 128 of the 2,638 original-mixture confirmation cells, with four independent bridge/time draws each. The 512 paired differences are averaged within cell before computing cell-weighted CR1 donor-cluster SEs. Query and price one-sided bounds each allocate 0.025 over respectively 128 and 120 endpoints, covering four checkpoints, two resolutions, baseline lower/query upper and price upper bounds. Student-t critical values use seven degrees of freedom, giving approximate conditional summaries. All-zero/all-one responses or degenerate donor variation yield unavailable bounds. Protocols are locally frozen before their searches and confirmations, not publicly preregistered.

Queries use Euler-1000/2000; fresh prices use autograd JVPs through Heun-64/128. Numerical gates require forward/inverse relative RMSE at most 10^{-3}, maximum effect refinement difference 0.01 control SD, sign disagreement 1%, and mean absolute per-cell price refinement $0.1\delta_{\mathrm{SE}}$. Direct integration of the priced field on the first two confirmation cells requires endpoint RMSE at most 10^{-3} for the search-selected candidate. These sensitivity checks do not certify ODE error. CUDA graphs replay unchanged float32 Euler operations with TF32 disabled, checked against ordinary Python execution on fixed rows at both intervals and grids. The aborted first seed-13 attempt keeps its frozen search archive; replay must reproduce sample IDs, calibration and every coefficient before confirmation.

Matched-size SWD and released rMMD compare edits with the same 342 target cells per checkpoint. rMMD uses float64 kernel accumulation and the minimum eligible training `L2_stim` class discrepancy, retaining the original bandwidth and floor. Original-condition price strata and abducted-latent/gate summaries are descriptive. No non-inferiority margin was fixed for these distances, so they cannot establish biological or edited-population equivalence. Numerical failure, unavailable population inference and no observed flip remain separate outcomes.

E.6 Supported-context application protocol

The seed-1337 checkpoint, encoder, PCA, 27-gene interferon-alpha program and training-control SD stay frozen. Both `CD4 T cells_1` task labels occur in training and every held-out partition. Search uses control rows for queries and stimulated rows for development quality; confirmation uses 233 control and 229 stimulated cells. Pricing uses the disjoint calibration partition with fresh source and time draws. Rotations act in a fixed rank-12 subspace (seed 9101337), with 66 skew-generator coefficients and norm cap four.

On 235 control search cells, central differences with step 0.01 estimate the score Jacobian, C and the gradient g of a sigmoid response surrogate with temperature 0.25 SD. The global risk quadratic uses 512 resampled search cells; the target quadratic uses all 230 stimulated search cells, each with one fresh source/time pair. At each budget, SLSQP minimizes $g^\top\beta$ and maximizes $\beta^\top C\beta$ separately, subject to both full risks $2\widehat{d}_j^\top\beta + \beta^\top\widehat{M}_j\beta \le \delta_j$ and $\|\beta\|_2 \le 4$. Each objective gets up to four starts (gradient, averaged-risk, generalized-eigenvector and seeded random) and 150 iterations. A 10^{-6} ridge affects starting directions only. Proposals are checked against both quadratics and, if necessary, scaled inward along their rays; no optimality is certified.

Finite edits and development quality select up to two proposals per budget: greatest H and smallest hard response fraction, with fixed-order tie breaking and duplicate removal. The resulting six vectors are frozen before confirmation. Thus, unlike Equation (13), this search uses d in both constraints. Search takes 65.9 seconds on the local GPU; this excludes confirmation and is not a retraining comparison.

Before search, preflight selects the first Heun pair among $128/256$, $256/512$ and $512/1024$ with baseline inverse–forward relative RMSE at most 10^{-4} and score-refinement RMS/maximum at most $0.005/0.02$ SD. Heun-$128/256$ qualifies on search cells: round-trip errors are $8.7\times10^{-7}/1.8\times10^{-7}$ and refinement RMS/maximum $9.1\times10^{-5}/2.9\times10^{-4}$ SD. Confirmation refinement is $9.5\times10^{-5}/3.4\times10^{-4}$ SD. Search uses Heun-32 finite differences; pricing uses autograd Jacobian–vector products at Heun-$64/128$.

The global and edited-condition calibration risks are 0.4452 and 0.5063, estimated from 32 bridge draws per cell on all $3{,}202$ cells and the 273 target cells, respectively. The floating stimulation gate has global median 3.3×10^{-4} but target median above 0.999, with 98.9% of target values above one half. It therefore attenuates many global contributions without switching them off. A01's target price is about 2.3 times its global price in native units. The target risk still averages over observed encodings rather than pricing each edited individual's exact context.

Sliced Wasserstein distance uses 128 projections (seed 910); MMD^2 uses an unbiased RBF estimator with bandwidth $1/128$ and a nonnegative clamp for gating. The 20% baseline-improvement and 10% alternative-loss gates are evaluated again on confirmation. These are fixed-panel operating tolerances, not equivalence tests or biological validation. All six candidates pass; A01 retains 99.48% and 99.74% of the two improvements.

Table 4: One clear reversal and a weaker case. Entries are scores minus the responder threshold in control SD at Heun-256. Both grids give the same decisions with two-sided 0.1 SD margins. The cells were identified retrospectively, not independently selected for validation.

Cell	Seed 13	Seed 1337	Seed 42	Seed 69	1337 + A03
GAGCAGGAAACCTG-1	-1.139	-0.599	-3.102	-1.242	$+0.187$
AGGAAATGAGACTC-1	-1.254	-0.124	-1.381	-0.897	$+0.108$

Prices average 16 full-cohort bridge replicates; SEs use independent replicate means, not cells or donors. One-sided Student-t bounds use 15 degrees of freedom and Bonferroni correction over six candidates, two risks and two grids (critical value 3.375). A01's native global/target estimates are 0.001183 ± 0.000021 and 0.002679 ± 0.000117 (mean $\pm$ SE), with upper bounds $0.282\%/0.607\%$ of calibration risk. The largest single-cell shares of bridge variance are $0.5\%/2.0\%$ and of quadratic price $0.2\%/0.9\%$. These approximate Monte Carlo bounds condition on the fixed eight-donor cohort; they do not provide exact finite-sample coverage or population inference.

Qualification requires both quality gates, both risk budgets on both grids, and the numerical checks. Material sensitivity additionally requires RMS at least 0.1 SD and at least 10% of cells moving 0.1 SD on both query grids. A01/A02 qualify at 1% and are material; A03/A04 qualify only at 4%; A05/A06 qualify at none of the three budgets. None passes 0.25% or its own search budget. Cross-budget reporting followed partial pricing, with candidates, budgets, queries, cohorts and thresholds already frozen. The same simultaneous bounds apply to these nested thresholds. Saved-array replay verifies summaries and bounds, not new data or coverage.

The baseline closes 69.66% of the SWD and 94.18% of the MMD^2 gap. A01's maximum change is 0.825 SD and its mean change -0.098 SD. Removing the mean shift leaves RMS 0.202 SD against baseline cross-cell SD 1.488: the movement is not just a common shift. Donor RMS ranges from 0.165 to 0.282 SD.

E.7 Four-seed agreement and random-direction controls

The four fits (seeds $13, 1337, 42, 69$) share prepared data, PCA, program weights, training-control SD, threshold and 233 CD4_1 cell IDs. Three baselines are newly evaluated; seed 1337 and A01–A06 use archived outputs. All baselines pass the original quality and Heun-$128/256$ numerical gates. A responder's edited score is at least the training stimulated median, 9.6815624. Responder fractions vary across seeds (29.61%, 33.48%, 19.31%, 24.46%), so the analysis concerns cellwise agreement: 16 consistent responders and 113 nonresponders on both grids, with 104 other calls disagreeing.

A03 reverses four consistent calls, two with 0.1 SD margins for every baseline and both grids. The main cell moves 0.786 SD from seed 1337, eighth largest among 233 movements; the weaker case starts 0.124 SD below threshold and moves 0.232 SD. Seed 1337 is empirically the closest baseline at every observed flip; perturbing that seed does not require this. Other fits establish agreement, not extra distance to cross. All/margin reversal counts for A01–A06 are $2/0$, $2/0$, $4/2$, $3/0$, $4/2$, $3/1$. A01/A02 reverse the same two calls at 1%, with margins 0.093 and 0.035 SD. A03 qualifies only at 4%, retaining $99.04\%/99.57\%$ of task improvements with archived global/target upper bounds $0.759\%/1.611\%$. Top-twelve lists and top-one/top-two donor priorities are not seed-stable.

For the random controls, eight standard-normal directions in the same 66-coefficient subspace (seed 911080) retain their sampled signs. Each ray is scaled to A03's target price using its full signed quadratic on 32 fresh target-cohort bridge replicates at Heun-64, subject to norm cap four. Coefficients are frozen before queries and confirmation; none is selected by decision changes. This follow-up was fixed after inspecting the original result.

Independent Heun-64/128 pricing shares fresh sources and times across the controls and unchanged A03. Global pricing uses 16 full $3{,}202$-cell replicates; target pricing adds 48 replicates of the 273 target cells, for 64 total. Approximate Student-t upper tails use $0.05/36$ (nine candidates, two risks, two grids). Matching uses paired replicate differences with two-sided tails $0.05/64$ (eight comparisons, two risks, two grids). Every interval must lie within $\pm 10\%$ of A03's archived estimate (0.00323392 global, 0.00734805 target); merely containing zero is insufficient. Price and matching

Table 5: All eight random controls and unchanged A03, independently repriced. Means and SEs use Heun-128; upper bounds take the maximum across grids. All nine pass quality and numerical checks. Matching requires the paired 10% rule for both risks and grids; margin flips require 0.1 SD on both sides for every seed and grid among the same 129 consistent calls.

	Price estimate (SE), 10^{-3}		Price UB (%)		Matched	Flips	
	Global	Target	Global	Target	both risks	All	Margin
A03	3.215 (0.033)	7.372 (0.101)	0.749	1.518	anchor	4	2
R01	3.202 (0.044)	7.303 (0.089)	0.755	1.497	yes	1	0
R02	3.227 (0.042)	7.590 (0.085)	0.759	1.552	yes	3	0
R03	3.136 (0.028)	7.257 (0.086)	0.727	1.486	yes	2	0
R04	3.243 (0.040)	7.450 (0.095)	0.760	1.530	yes	0	0
R05	3.104 (0.028)	7.175 (0.093)	0.720	1.474	yes	4	0
R06	3.220 (0.033)	7.613 (0.099)	0.750	1.565	yes	5	0
R07	3.200 (0.033)	7.394 (0.086)	0.745	1.513	yes	4	0
R08	3.161 (0.021)	7.383 (0.094)	0.727	1.516	yes	4	0

families have separate nominal 95% coverage, not a joint guarantee, conditional on fixed cells. A03's fresh global/target upper bounds are $0.749\%/1.518\%$.

E.8 Earlier CD4_0 audits and fresh pricing

The original audit tests 60 candidates across four checkpoints on the same 342 CD4_0 cells. Its reference budget δ_{SE} is one donor-aware standard error of baseline risk, fixed before search. No sign reverses: checkpoint median effects are 5.30–6.16 control SD above zero, and the closest cells need 0.06–0.49 SD to cross against at most 0.006 SD of tested movement. This is not a certificate that the family cannot change a sign. Nor does any candidate pass the full numerical gate: Euler-1000 round-trip relative RMSE is 0.001015–0.001044, above 10^{-3}, despite score refinement below 0.0035 SD and direct-field endpoint errors 0.000544–0.000639. Original quality diagnostics had no equivalence margin and do not qualify the later follow-up.

Of the 60 candidates, 24 pass the original single-sample price rule and 51 have nondegenerate bounds. Maximum price-qualified RMS is 0.068, 0.116, 0.126 and 0.129 SD across checkpoints. The largest mean drop, 0.038 SD, has upper bound $3.93\,\delta_{\mathrm{SE}}$ and fails; the largest passing drop is 0.028 SD. Complete checkpoint, numerical and quality tables remain in the supplementary source and result records rather than being repeated here.

A follow-up on seed 1337 forms C from 30 finite-difference edits on 128 search cells. The individual direction, M ray and best of six random directions yield twelve frozen candidates at four budgets. They reuse the 342-cell panel; this is not a second preregistered witness. Pricing uses the archived 128 cells and 128 disjoint cells with four fresh bridge draws each, sharing the eight donors. Five candidates pass both samples and both Heun grids (Table 6). The largest passing RMS is 0.100 SD, maximum 0.388 SD and mean change -0.0072 SD; 100 of 342 cells move over 0.1 SD, with donor RMS 0.079–0.119 SD. Its maximum bound is $0.878\,\delta_{\mathrm{SE}}$. These are price-only qualifications, not passage of the full gate.

At each sample/grid, CR1 donor-cluster Student-t bounds use tail $0.025/12$. Requiring four bounds to pass is a conjunction, not simultaneous coverage of all 48 bounds. The original sign audit used family 120, hence its larger upper bound for an unchanged vector. Eight candidates pass archived pricing but only five pass fresh pricing. The 0.134 SD candidate's fresh mean is $0.773\,\delta_{\mathrm{SE}}$ and upper bound $1.552\,\delta_{\mathrm{SE}}$: loss of qualification, not a demonstrated violation. An all-one response has unavailable donor-based population bounds, not a zero-width interval.

The individual/random pair at search budgets δ_{SE} and $2\delta_{\mathrm{SE}}$ has archived prices within $0.012\,\delta_{\mathrm{SE}}$ and $H = 0.0316$ versus 0.0212. The paired donor-cluster difference interval covers zero, as does the comparison with the M ray. Selection by nearby price estimates makes these descriptive post-selection comparisons, not equivalence-certified costs or evidence of search superiority.

For comparison, the four unrotated editors share cell IDs, order, donors, preprocessing, PCA and score scale, but each refits encoder and flow for $200 + 200$ epochs. Their learned contexts may

Table 6: All twelve individual-search candidates. Price UB and New UB are the maximum upper bounds across Heun-64/128 in the archived and fresh pricing samples; bold denotes passage on that sample. Query summaries use Euler-1000 with Euler-2000 refinement. Search budgets are not confirmed prices.

Direction	Budget (δ_{SE})	Price UB (δ_{SE})	New UB (δ_{SE})	RMS (SD)	Max (SD)	> 0.1 (of 342)	$\Pr\{D > 0\}$
Individual (h)	0.25	**0.484**	**0.878**	0.1003	0.388	100	100.0%
	0.50	**0.861**	1.552	0.1344	0.487	152	100.0%
	1.00	1.590	2.837	0.1778	0.685	194	100.0%
	2.00	3.017	5.319	0.2234	0.882	221	100.0%
Proportion (M ray)	0.25	**0.222**	**0.495**	0.0666	0.253	47	100.0%
	0.50	**0.400**	**0.890**	0.0923	0.335	96	100.0%
	1.00	**0.756**	1.642	0.1256	0.429	150	100.0%
	2.00	1.496	3.095	0.1674	0.522	198	100.0%
Random (best of 6)	0.25	**0.243**	**0.250**	0.0554	0.204	28	100.0%
	0.50	**0.462**	**0.536**	0.0780	0.291	63	100.0%
	1.00	**0.891**	1.139	0.1081	0.402	122	100.0%
	2.00	1.743	2.387	0.1455	0.503	153	100.0%

Table 7: Refitting and within-checkpoint sensitivity on the older Kang panel. RMS uses control SD; ρ is Spearman correlation. The six pairs share four models and are not independent replications. The last row is the seed-1337 Individual$(0.25\delta_{\mathrm{SE}})$ alternative. Refits need not have equal risks, quality or density paths.

Contrast	RMS (SD)	Centered RMS (SD)	ρ_S	Top-68 overlap	Top-12 overlap
Seeds 13/1337	1.5361	1.3236	0.588	33	3
Seeds 13/42	1.1891	1.1883	0.614	33	4
Seeds 13/69	1.5274	1.4890	0.543	25	1
Seeds 1337/42	1.5909	1.3617	0.553	34	4
Seeds 1337/69	1.6358	1.5757	0.525	27	1
Seeds 42/69	1.4102	1.3572	0.588	36	1
1337: Individual(0.25)	0.1003	0.1000	0.998	65	12

therefore differ. Archive row zero is the stimulation edit at $\beta = 0$, not the factual input. This is the older CD4_0 panel, separate from the CD4_1 consensus analysis.

Refit RMS is 1.189–1.636 SD, about 12–16 times the rotation's 0.1003 SD. Refits share one to four top-twelve cells; the rotation retains all twelve and 65 of the top 68, with $\rho = 0.998$. Euler-2000 preserves every top-twelve count. This establishes neither a larger sensitivity signal nor a runtime advantage for the audit.

After the failed numerical gate, seed-13 Euler refinements from 1,000 to 8,000 steps reduce round-trip RMSE from 0.00104364 to 0.00013060, consistent with first-order error; recorded qualification is unchanged. CPU replay reproduces original price means within 3×10^{-10} and uncertainty summaries within 1.1×10^{-9}. Follow-up per-cell means recover donor-cluster intervals but not per-draw SEs. Original calibration contributions and generated endpoints were not saved, so baseline risk and quality still require regenerated outputs.

E.9 CUI PROTOCOL AND DIAGNOSTIC RESULTS

The Cui data (Cui et al., 2024) contain $110{,}378$ cells and $31{,}053$ genes. All 100 CD4 PBS cells are excluded from fitting and pricing (40 search, 60 query confirmation), as are all 100 CD4 TGF-β1 target cells. The remaining metadata-stratified split has $66{,}098$ training, $16{,}891$ calibration, $14{,}185$ search-risk and $13{,}004$ confirmation-risk cells. Released `sample|rep` labels are descriptive strata, not established independent or paired mice.

Table 8: Three Cui alternatives on 60 cells, Euler-1000. Prices are means (SE), Heun-64, in 10^{-6} native units. The maximum approximate upper bound across grids is in ppm of calibration risk 0.3543405, not a price estimate or a true-price ordering. Each row represents four identical vectors; Euler-2000 retention is unchanged.

Search direction	Retained (/12)	RMS (SD)	Mean (SE) (10^{-6})	Max UCB (ppm)
Individual (H)	5	0.067	-1.151 (0.849)	3.7
Response fraction	5	0.042	2.771 (2.188)	25.7
Random (best of 6)	4	0.069	-4.815 (3.699)	16.6

Table 9: Cui edit quality in native PCA coordinates, with 60 cells per group; lower is closer to target. rMMD uses the released normalization (denominator 0.03893084, kernel $1/128$). No equivalence margin was prespecified.

	SWD		Released-formula rMMD	
Output	Euler-1000	Euler-2000	Euler-1000	Euler-2000
Unedited factual	0.158640		0.140428	
Baseline edit	0.164928	0.165004	0.189123	0.189385
C03 (Random)	0.168254	0.168334	0.205577	0.205866

Counts are normalized to training median library size 4,114 and logged once. Training-only Seurat HVGs (3,000) plus 40 program genes feed a training-only 128-PC reconstruction. The model follows the shared architecture and schedule (Appendix E.5), with 94 known and 128 unknown concepts. Seed 13 is fitted locally and frozen.

Scores average 53 mouse Hallmark TGF-β signaling genes in the frozen reconstruction (MSigDB v2026.1.Mm; MSigDB Team, 2026b; Liberzon et al., 2015). One SD is 0.04908 log-expression units from 240 training T-cell PBS controls, not held-out CD4 controls. Projection RMSE on confirmation cells is 0.554 SD.

Editing sets PBS/TGF-β1 to $[0, 1]$, fixes other concepts and uses deterministic inversion without guidance or noise. The floating TGF-β gate acts at both endpoints of rank-six rotations (15 generators, norm cap 4). Search selects an individual direction, an M ray and the best of six fixed random directions, using separate bridge draws and full risk. Budgets $\{0.25, 0.5, 1, 2\}\delta_{\mathrm{C}}$ all reach the cap: twelve frozen records reduce to three vectors, not a resolved frontier.

Calibration uses 1,024 iid rows and eight bridge draws each: baseline 0.3543405, reference $\delta_{\mathrm{C}} = 0.003543405$. Risk is uniform on eligible held-out contexts, not CD4-specific. Independent pricing draws 256 of the 13,004 rows with replacement and four fresh bridges per draw, shared across Heun-64/128. Index-level SEs give approximate Student-t upper bounds (255 degrees of freedom, tail $0.05/24$ for twelve records and two grids), excluding lower bounds. Uncertainty is fixed-pool Monte Carlo error, not population inference. The denominator is total risk, not excess risk over an estimated optimal conditional risk, which we do not estimate.

For Individual, response-fraction and random directions, linear prices are $(-1.513, 2.381, -6.389) \times 10^{-6}$ and quadratic prices $(0.362, 0.391, 1.574) \times 10^{-6}$ at Heun-64. One sampled index contributes 90.9%, 83.2% and 96.9% of the quadratic totals; the largest four contribute 97.6%, 97.8% and 99.5%. Such concentration limits confidence in the bounds without proving undercoverage. Actual concept gates were not saved; labels cannot recover them. Near-zero means neither establish zero risk nor price the excluded CD4 PBS/TGF-β1 subgroup. Grid changes below 3.3×10^{-9} do not certify ppm accuracy, since numerical error can be shared.

The top-20% list ranks signed increments, breaking ties by cell ID. Minimum price-qualified retention selects C03; larger H and fixed order break ties. Lists agree across grids, but the M ray's positive count changes from seven to eight.

Both the baseline and C03 worsen both target distances (Table 9). Thus all three alternatives pass only the approximate global-price rule, not joint application qualification.

Euler-1000/2000 round-trip relative RMSE is $0.0013395/0.0006704$; maximum score refinement is 0.0021543 SD (baseline) and 0.0018735 SD (alternatives). Refinement neither replaces primary results nor transfers Kang's gates to Cui.