CELLWRIGHT: MECHANISM-CONDITIONED TRANSPORT FOR VERIFIABLE VIRTUAL CELLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtual cell models promise to forecast cellular responses to unseen perturbations in silico, yet recent audits show that most deep models barely beat mean or linear baselines, and even strong systems answer a narrower question than biologists ask: they regress a control-referenced expression shift, under-model real population heterogeneity, and offer no account of *why* the response occurs. We argue that the right target is the full conditional distribution of perturbed cells *together with* the mechanism that generates it. The obstacle is supervision: mechanisms carry no ground-truth labels at scale, so mechanistic reasoning cannot be trained by imitation and readily hallucinates. CELLWRIGHT recasts this as a verification problem. First, it replaces the latent delta with a *mechanism-conditioned latent Schrödinger bridge*, trained by conditional flow matching over an optimal-transport coupling, that transports an entire control population to the perturbed population and composes velocity fields with an explicit epistatic residual for unseen combinations. Second, an LLM agent emits machine-readable mechanism graphs over the TF→DNA→RNA→protein axis and is optimized by reinforcement learning with verifiable rewards: an energy-based verifier trained by noise-contrastive estimation against hallucinated negatives supplies edge-decomposable rewards, calibrated confidence and a test-time search signal. On held-out perturbations from Norman and Replogle-K562, CELLWRIGHT reduces energy distance by 13.2% and 13.0% over the strongest prior method and improves Top-20 DEG Δ-correlation, while cutting mechanism hallucination from 0.52 to 0.17 and expected calibration error from 0.34 to 0.061 relative to a free-text LLM reasoner. We also report where it does not win: a neural optimal-transport baseline retains the best energy distance on the chemical corpus, the mean baseline retains the best MSE, and the gain on the Systema-corrected score is not significant. Ablations show that distributional transport drives accuracy, while the verifier drives honesty. Our code is available at `https://anonymous.4open.science/r/CellWright-D4B2/`.

1 INTRODUCTION

Predicting how a cell responds to an unseen perturbation is the defining task of the virtual-cell programme, and the last two years have produced a great deal of modelling effort aimed at it. They have also produced an uncomfortable audit. Careful re-evaluations find that deep perturbation-response models frequently fail to beat a mean or a ridge baseline once the evaluation controls for the structure of the metric (Ahlmann-Eltze et al., 2025; Wu et al., 2026; Peidli et al., 2024), and our own protocol reproduces that finding (Sec. 4.2). The usual response is to build a larger model. We think the more productive response is to ask what the models are being asked to predict.

Two things are wrong with the standard target. The first is that it is a point. Almost every system regresses a control-referenced shift Δ, a single vector per perturbation, while a real perturbed population is heterogeneous: it contains subpopulations that respond differently, variance that changes under perturbation, and escaper cells that barely respond. A model trained to emit the conditional mean is by construction optimal for squared error and blind to this, which is also why MSE-ranked leaderboards keep declaring the mean baseline competitive. The second is that it is silent: a predicted Δ says nothing about which transcription factor acted, on which regulatory element, with which sign, and for a biologist choosing the next experiment that account is much of the value.

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.

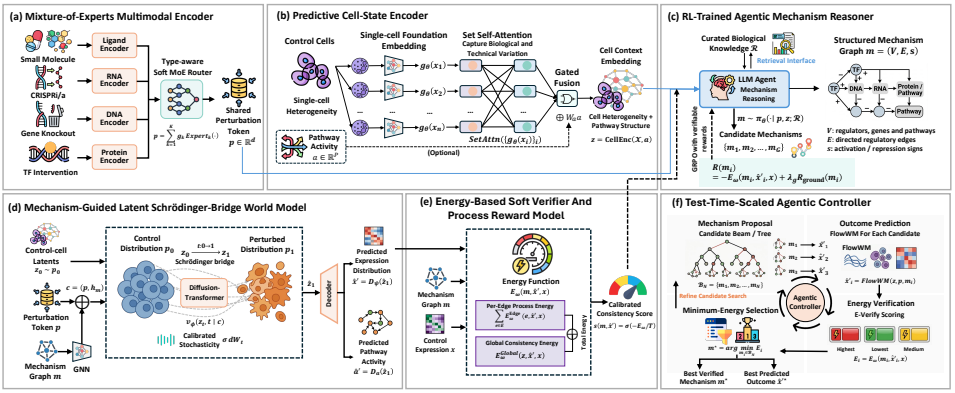


Figure 1: **CELLWRIGHT.** (a) A type-aware soft mixture-of-experts router maps small molecules, guides, knockouts and TF interventions into one shared perturbation token; (b) set self-attention over control cells gives a context embedding that carries single-cell heterogeneity rather than a mean profile; (c) an LLM agent emits a structured mechanism graph over the TF→DNA→RNA→protein axis, trained by GRPO against verifiable rewards; (d) a mechanism-conditioned latent Schrödinger bridge transports the *whole* control distribution to the perturbed one, and an energy verifier factorized into per-edge and global terms scores the (mechanism, outcome) pair, supplying both the reward in (c) and a calibrated confidence; (e) at inference the controller returns the minimum-energy candidate. Specifications are in Appendix A.1.

Targeting the distribution is a modelling problem with known tools. Targeting the mechanism is a supervision problem, and it is the harder one: mechanisms carry no ground-truth labels at any scale, so a mechanism reasoner cannot be trained by imitation, and a language model asked to explain a perturbation will produce fluent and frequently false regulatory claims. Our central move is to treat this as a *verification* problem rather than a labelling problem. If a mechanism graph is machine-readable, and if we can learn a function scoring whether a graph is coherent with the expression outcome it is supposed to explain, then that function is simultaneously a reward for training the reasoner, a calibrated confidence at inference, and a search signal at test time.

CELLWRIGHT is built around these two claims (Figure 1). A mechanism-conditioned latent Schrödinger bridge, trained by conditional flow matching along an optimal-transport coupling, transports an entire control population to the perturbed population, and composes velocity fields with an epistatic residual so that unseen gene pairs are not forced to be additive. An energy-based verifier, trained by noise-contrastive estimation against hallucinated negatives (shuffled signs, swapped targets, mismatched outcomes), factorizes into per-edge terms and supplies edge-decomposable rewards to a GRPO-trained mechanism reasoner. The two parts are complementary in a way our ablations make precise: transport drives accuracy, verification drives honesty, and neither substitutes for the other.

Contributions. We (i) formulate perturbation response as mechanism-conditioned distributional transport, with an epistatic residual letting unseen combinations depart from additivity (Sec. 3.3); (ii) make mechanistic reasoning trainable without mechanism labels by learning an energy verifier from hallucinated negatives and using it as an edge-decomposable RLVR reward (Sec. 3.4); (iii) turn that verifier into test-time compute (Sec. 3.5); and (iv) evaluate on a protocol that includes the metrics on which a distributional model should *lose*, reporting those losses in the main text (Sec. 4.3).

2 RELATED WORK

Perturbation-response prediction and its audits. Task-specific models include autoencoder and disentanglement methods (CPA, scGen, biolord, scVI) (Lotfollahi et al., 2023; 2019; Piran et al., 2024; Lopez et al., 2018), optimal-transport maps (CellOT) (Bunne et al., 2023), and knowledge-graph combination models (GEARS) (Roohani et al., 2024). Single-cell foundation models have also been adapted (Cui et al., 2024; Hao et al., 2024; Theodoris et al., 2023; Rosen et al., 2023; Chen & Zou, 2024; Istrate et al., 2024), with STATE and TxPert defining recent state of the art (Adduri et al., 2026; Kenyon-Dean et al., 2025). Benchmark audits show simple baselines remain competitive under

careful metrics (Ahlmann-Eltze et al., 2025; Wu et al., 2026; Peidli et al., 2024); we therefore include non-learned baselines and report metrics where they win.

Distributional models of perturbation. Flow matching and Schrödinger bridges map control to perturbed populations (Lipman et al., 2023; Tong et al., 2023; De Bortoli et al., 2021), and have recently been applied to single-cell data (Klein et al., 2025; Chi et al., 2026; Yu et al., 2026; Abir et al., 2025; Liang et al., 2025; Palla et al.). Our transport is conditioned on an explicit, verifier-scored mechanism graph rather than only a perturbation embedding, and adds an epistatic residual for composition; the DiT velocity field, classifier-free guidance, and representation alignment follow established practice (Peebles & Xie, 2023; Ho & Salimans, 2022; Yu et al., 2024; Wang et al., 2025).

Verifiable rewards for reasoning. RLVR with group-relative policy optimization trains reasoning given checkable signal (Shao et al., 2024; Guo et al., 2025). Biology lacks such mechanism labels, so we learn a verifier by noise-contrastive estimation (Gutmann & Hyvärinen, 2010) against hallucinated negatives and factorize per edge for decomposable rewards. Concurrent virtual-cell world models target simulation fidelity (Wei et al., 2026); we target fidelity plus a checkable account of why.

3 METHODOLOGY: CELLWRIGHT

CELLWRIGHT is a distribution-level world model for one well-defined cellular subsystem: the regulatory module linking transcription factors, DNA, RNA and proteins. Two choices distinguish it. The latent transition is a conditional flow-matching / Schrödinger-bridge process rather than a point estimate, which retains subpopulation structure, variance and escaper cells. And the mechanism reasoner is an LLM agent trained by reinforcement learning with verifiable rewards, where a learned energy verifier supplies the reward, turning "propose–predict–verify" into a closed loop. Six modules realize this; full specifications, prompts and hyperparameters are in Appendix A.1.

Problem formulation. A basal context is a set of control cells $\mathcal{X} = \{x_i\}_{i=1}^{n}$, $x_i \in \mathbb{R}^G$, with an optional pathway-activity vector $a \in \mathbb{R}^P$. A perturbation π is a small molecule, a CRISPR(i/a) guide, a knockout or a TF intervention. Unlike point predictors, CELLWRIGHT models the full conditional distribution $p(x' \mid \mathcal{X}, \pi)$, from which the mean shift Δ, higher-order statistics, pathway activation $\hat{a}'$ and the population-level state transition follow. It additionally emits a mechanism graph m with a calibrated consistency score $s(m, \hat{x}') \in [0, 1]$.

3.1 ENCODERS: SHARED PERTURBATION TOKEN AND SET-LEVEL CELL CONTEXT

MOLENC maps heterogeneous perturbagens into one token $p \in \mathbb{R}^d$ so that drugs, guides, knockouts and TF interventions transfer to one another, using modality-specific encoders and a soft mixture-of-experts router gated by perturbation type τ:

$$p = \sum_{k=1}^{K} g_k\big(\tau,\, e_{\text{prot}}, e_{\text{DNA}}, e_{\text{RNA}}, e_{\text{lig}}\big)\, \textsc{Expert}_k(\cdot), \qquad \textstyle\sum_k g_k = 1, \tag{1}$$

where e_{prot} is a protein-language-model embedding (Lin et al., 2023), e_{DNA} comes from a sequence-to-function trunk (Linder et al., 2025), e_{RNA} from an RNA encoder and e_{lig} from a geometry-aware molecular encoder. Experts are initialized from public checkpoints and LoRA-tuned. A soft router, not a hard branch, lets a modality with few examples borrow from others.

CELLENC produces a context embedding that captures heterogeneity and pathway structure at once, trained with a joint-embedding predictive objective (Assran et al., 2023) so the latent is predictive of held-out cells rather than merely reconstructive:

$$z = \textsc{CellEnc}(\mathcal{X}, a) = \textsc{SetAttn}\big(\{g_\theta(x_i)\}_i\big) \oplus W_a\, a. \tag{2}$$

Operating on a *set* of cells lets the model absorb biological and technical variation (cell cycle, sequencing depth) without distributional assumptions, which is a prerequisite for the transport below.

3.2 MECHREASON: AN RL-TRAINED AGENTIC MECHANISM REASONER

An LLM agent receives (π, z) and a retrieval interface $\mathcal{R}$ over curated knowledge (Ashburner et al., 2000; Gillespie et al., 2022), and emits, after explicit chain of thought, a structured graph $m = (V, E, \mathbf{s}) \sim \pi_\theta(\cdot \mid \pi, z; \mathcal{R}, \textsc{MolEnc})$, where V are implicated regulators, genes and pathways,

E are directed edges along the TF→DNA→RNA→protein axis, and $\mathbf{s} \in \{-1, +1\}^{|E|}$ are activation/repression signs. Because m is machine-readable it can condition the world model and be scored; because the score comes from E-VERIFY rather than human labels, the reasoner improves without annotated explanations. We fine-tune π_θ with GRPO (Shao et al., 2024): for each input the agent samples G candidates, each is scored, and the policy moves toward group-normalized advantages,

$$R(m_i) = -E_\omega\big(m_i, \hat{x}'_i, x\big) + \lambda_g R_{\text{ground}}(m_i), \qquad A_i = \frac{R(m_i) - \text{mean}_j\, R(m_j)}{\text{std}_j\, R(m_j)}, \tag{3}$$

$$\mathcal{J}_{\text{GRPO}} = \mathbb{E}\Big[\tfrac{1}{G}\textstyle\sum_i \min\big(\rho_i A_i,\ \text{clip}(\rho_i, 1{-}\epsilon, 1{+}\epsilon) A_i\big) - \beta D_{\text{KL}}\big(\pi_\theta \| \pi_{\text{ref}}\big)\Big], \quad \rho_i = \tfrac{\pi_\theta(m_i)}{\pi_{\theta_{\text{old}}}(m_i)}. \tag{4}$$

3.3 FLOWWM: A MECHANISM-GUIDED LATENT SCHRÖDINGER BRIDGE

Instead of predicting a latent delta, FLOWWM learns a continuous-time transport carrying the distribution of control latents $z_0 \sim p_0$ to the perturbed distribution $z_1 \sim p_1$, conditioned on the perturbation and the mechanism. With $h_m = \text{GNN}(m)$ and a DiT velocity field v_ϕ (Peebles & Xie, 2023), training is conditional flow matching along a mini-batch optimal-transport coupling Π (Tong et al., 2023; Cuturi, 2013):

$$z_t = (1-t)z_0 + t z_1, \quad c = (p, h_m), \qquad \mathcal{L}_{\text{CFM}} = \mathbb{E}_{t\sim\mathcal{U}[0,1],(z_0,z_1)\sim\Pi}\big\|v_\phi(z_t, t \mid c) - (z_1 - z_0)\big\|_2^2. \tag{5}$$

With the entropic coupling, sampling integrates $\mathrm{d}z_t = v_\phi\,\mathrm{d}t + \sigma\,\mathrm{d}W_t$; starting from real control cells, plus this noise, preserves subpopulation spread and escaper cells rather than a single mean. At inference we start from an *actual* control-cell latent, integrate with classifier-free guidance (Ho & Salimans, 2022) whose scale w is mapped from drug dose to give tunable response intensity, then decode $\hat{x}' = D_\psi(\hat{z}_1)$ and $\hat{a}' = D_a(\hat{z}_1)$. For multiple interventions we compose conditional velocity fields and add an explicit non-additive residual,

$$v_\phi^{\text{multi}}\big(z_t, t \mid \{c_\ell\}\big) = \sum\nolimits_{\ell=1}^{L} v_\phi(z_t, t \mid c_\ell) \ + \ r_\xi\big(z_t, t, \{c_\ell\}\big), \tag{6}$$

which allows a predicted combination to depart from additivity. Because audits report that many combinatorial effects *are* approximately additive (Ahlmann-Eltze et al., 2025), r_ξ is a residual on top of the additive term rather than its replacement, and Sec. 4.5 shows it is inert where it should be.

3.4 E-VERIFY: AN ENERGY-BASED SOFT VERIFIER AND PROCESS REWARD MODEL

E-VERIFY learns $E_\omega(m, x', x)$, low when a mechanism is coherent with the expression outcome it explains and high otherwise. It is trained by noise-contrastive estimation (Gutmann & Hyvärinen, 2010), with index-grounded graphs as positives m^+ and *hallucinated negatives* (shuffled signs, swapped targets or mismatched outcomes), so no human-annotated mechanism labels are needed:

$$\mathcal{L}_{\text{NCE}} = -\log \frac{\exp(-E_\omega(m^+, x', x))}{\exp(-E_\omega(m^+, x', x)) + \sum_j \exp(-E_\omega(m_j^-, x', x))}. \tag{7}$$

We factorize the energy per edge plus a global term, so the reward of Eq. equation 3 decomposes over edges and errors localize, and we convert it to a confidence with a temperature fit on held-out data:

$$E_\omega(m, x', x) = \sum\nolimits_{e\in E} E_\omega^{\text{edge}}(e, x', x) + E_\omega^{\text{glob}}(z, x', x), \qquad s(m, \hat{x}') = \sigma(-E_\omega/T). \tag{8}$$

Two properties of this design are worth stating plainly because they bound what the verifier can establish. E_ω is trained against *synthetic* corruptions, so it certifies coherence with an outcome (observed in training, predicted at inference), not biological truth; and the negatives define what it can detect, so a hallucination of a kind not represented among the corruptions is not guaranteed to be caught. Sec. 4.4 and Appendix A.13 report what it does and does not measure.

3.5 REFINE: VERIFIER-GUIDED SEARCH AT TEST TIME

At inference the controller turns prediction into search whose accuracy depends on the compute budget N: the agent proposes a beam of N candidate mechanisms $\mathcal{B}_N$, FLOWWM predicts an

outcome for each, and the controller returns the candidate of minimum verified energy,

$$m^\star = \arg\min_{m \in \mathcal{B}_N} E_\omega\big(m,\ \textsc{FlowWM}(z, p, m),\ x\big), \qquad \hat{x}'^\star = D_\psi\Big(z + \textstyle\int_0^1 \tilde{v}_\phi(z_t, t \mid p, h_{m^\star})\,\mathrm{d}t\Big). \tag{9}$$

This lets one trade test-time compute for accuracy and lower hallucination without retraining, up to a point, since searching against a frozen verifier eventually over-optimizes it (Sec. 4.6).

3.6 Training

Training is staged: (1) pretrain encoders and the latent autoencoder; (2) train FlowWM with a distribution-level objective; (3) train E-Verify by NCE; (4) RLVR fine-tune MechReason with GRPO against the *frozen* verifier and world model. The supervised stages minimize

$$\mathcal{L} = \lambda_1\mathcal{L}_{\mathrm{CFM}} + \lambda_2\mathcal{L}_{\mathrm{MMD}} + \lambda_3\mathcal{L}_{\mathrm{NCE}} + \lambda_4\mathcal{L}_{\mathrm{REPA}} + \lambda_5\mathcal{L}_{\mathrm{path}} + \lambda_6\mathcal{L}_{\mathrm{cal}}, \tag{10}$$

where $\mathcal{L}_{\mathrm{MMD}}$ is a maximum-mean-discrepancy (equivalently energy-distance (Székely & Rizzo, 2013)) term that aligns predicted and observed populations beyond cell-level correspondence and prevents collapse to the perturbed mean, $\mathcal{L}_{\mathrm{REPA}}$ ties the velocity field's intermediate features to a frozen single-cell embedding (Yu et al., 2024), $\mathcal{L}_{\mathrm{path}}$ supervises pathway activation, and $\mathcal{L}_{\mathrm{cal}}$ regularizes verifier calibration through an expected-calibration-error surrogate. Freezing the verifier during stage 4 is deliberate: a jointly optimized verifier would be free to move its own reward.

4 Experiments

4.1 Setup

Data and splits. We evaluate on public atlases reachable through the scPerturb and PerturBench indexes (Peidli et al., 2024; Wu et al., 2026). Genetic: Norman (Norman et al., 2019) (combinatorial CRISPRa in K562; 287 perturbations, 131 multi-gene) and Replogle (Replogle et al., 2022) (CRISPRi in K562 and RPE1); the genome-wide K562 screen is used for pre-training only, with every held-out test perturbation removed from it. Chemical: sci-Plex (Srivatsan et al., 2020) and a Tahoe-100M subset (Zhang et al., 2025). Regulatory sub-tasks use ENCODE/GTEx/FANTOM5 tracks and splicing / poly(A) QTL benchmarks. Splits are perturbation- or context-disjoint, never cell-disjoint. The default is $70/10/20$; atlases with fewer than 300 perturbations use a larger test fraction so that the held-out set is not degenerate, and Norman is stratified so that the $0/2$-seen compositional bucket retains $n{=}19$ pairs. Per-dataset counts and test fractions are in Appendix A.4, Table 7. Adamson and Dixit are part of the pre-training pool only and carry no reported test numbers.

Baselines and metrics. We report six non-learned baselines, six task-specific models, six foundation/LLM models and the two world models; Appendix A.5 lists the further systems that are part of the protocol but were not run, so that the coverage of the main tables is unambiguous. Expression accuracy is Pearson of Δ over all genes (PΔ) and over the Top-20 DEGs (PΔ_{20}), plus MSE and MAE; distributional quality is energy distance (E-dist) and Top-20 DEG recall; biological quality is the Differential-Expression Score (DES) and Perturbation Discrimination Score (PDS); robustness is the Systema-corrected score and a DEG-weighted $R^2_w(\Delta)$ that neutralizes the mean-baseline artifact. The mechanism axis (pathway-F1, regulatory sign accuracy, retrieval MRR, hallucination rate and expected calibration error, ECE) is ours, and Appendix A.2 defines each precisely, since these are new and a reader should be able to check what a hallucination is counted as.

Protocol. Every number is the mean over three seeds $\{13, 42, 2026\}$ with the sample standard deviation after $\pm$. Seeds control initialization, mini-batch OT pairing and GRPO sampling; splits are fixed across seeds, so non-learned baselines are deterministic and carry ± 0.000. Where we compare against the strongest baseline we report a Welch t-test over the three seeds (Appendix A.7, Table 10) and mark any comparison that fails at the 95% level with $^{\text{n.s.}}$. Full training configuration, budgets and hardware are in Appendix A.3, Table 6.

4.2 Main results

Table 1 makes two points, and the second is the one we would defend under scrutiny.

Table 1: **Main perturbation-response prediction** on the two genetic atlases. Representative rows; the complete table with all 20 baselines and all four corpora (Replogle-RPE1 and the chemical corpus included) is Appendix A.6, Table 8. Mean ± std over three seeds; best and second are shaded.

	Norman (combo, CRISPRa)			Replogle-K562 (CRISPRi)		
Method	PΔ ↑	$P\Delta_{20}$ ↑	E-dist ↓	PΔ ↑	DES ↑	E-dist ↓
Non-learned						
Perturbed-Mean	0.521 ±0.000	0.308 ±0.000	18.64 ±0.00	0.412 ±0.000	0.412 ±0.000	14.22 ±0.00
Additive	0.607 ±0.000	0.418 ±0.000	15.23 ±0.00	0.431 ±0.000	0.427 ±0.000	13.79 ±0.00
Linear / Ridge	0.592 ±0.000	0.394 ±0.000	16.07 ±0.00	0.446 ±0.000	0.441 ±0.000	13.41 ±0.00
Task-specific						
CPA	0.566 ±0.009	0.357 ±0.012	16.85 ±0.21	0.418 ±0.008	0.408 ±0.011	14.31 ±0.19
CellOT	0.524 ±0.013	0.318 ±0.016	13.87 ±0.24	0.403 ±0.012	0.402 ±0.014	11.62 ±0.21
GEARS	0.628 ±0.009	0.437 ±0.011	14.62 ±0.19	0.471 ±0.008	0.471 ±0.010	12.66 ±0.16
Foundation / LLM						
scGPT	0.603 ±0.008	0.391 ±0.011	15.94 ±0.20	0.458 ±0.008	0.456 ±0.010	13.04 ±0.16
scFoundation	0.617 ±0.008	0.409 ±0.010	15.31 ±0.19	0.469 ±0.007	0.468 ±0.009	12.78 ±0.15
scGenePT	0.612 ±0.009	0.421 ±0.011	15.04 ±0.18	0.464 ±0.008	0.462 ±0.010	12.88 ±0.15
World models						
TxPert	0.679 ±0.007	0.496 ±0.009	12.35 ±0.14	0.521 ±0.006	0.573 ±0.008	10.94 ±0.12
STATE	0.698 ±0.006	0.531 ±0.008	12.43 ±0.14	0.537 ±0.006	0.594 ±0.008	10.81 ±0.11
CELLWRIGHT (ours)	**0.741** ±0.008	**0.581** ±0.010	**10.72** ±0.18	**0.583** ±0.006	**0.668** ±0.008	**9.40** ±0.12

First, the audit premise reproduces on our protocol. CPA sits below the Perturbed-Mean baseline on both K562 DES (0.408 vs. 0.412) and E-dist, and in the full table so do GenePT, Geneformer*, UCE* and kNN; the best foundation model beats Linear/Ridge on Norman PΔ by +0.025. A reader should take this as evidence that our harness is not tuned to flatter deep models, and it is the reason we report non-learned rows everywhere rather than in a footnote.

Second, CELLWRIGHT improves over the strongest *prior* method in each column of Table 1, but the strongest prior method is not always STATE. TxPert holds the best prior E-dist on Norman (12.35, ahead of STATE's 12.43) while STATE holds the best point correlations, so the reductions are 13.2% on Norman *against TxPert* and 13.0% on Replogle-K562 against STATE (8.9% on RPE1). On the chemical corpus the picture changes: CellOT, which optimizes an optimal-transport objective directly on drug-response data, holds the best E-dist (8.04) and **beats CELLWRIGHT** (8.21), a −2.1% result, though the gap is not significant over three seeds ($p = 0.28$, Table 10). We read the CellOT result as informative rather than embarrassing: the chemical corpus is dose-stratified and more heterogeneous than the CRISPRi screens, and a method whose inductive bias is transport does well precisely where distributional shift is largest. That is consistent with our claim about what distributional objectives buy; it also says our mechanism conditioning contributes less when the perturbagen is a small molecule whose regulatory account is weaker.

4.3 METRICS ON WHICH A DISTRIBUTIONAL MODEL SHOULD LOSE

Table 2 is the table we would want a sceptical reviewer to read first. Perturbed-Mean attains the best MSE and MAE; CELLWRIGHT is *fourth* on MSE (behind Perturbed-Mean, STATE and Linear/Ridge) and third on MAE. This is not a defect but a property of squared error, which rewards shrinkage toward the average perturbed profile, and it is precisely the artifact that makes MSE-ranked leaderboards uninformative for this task. The ordering inverts on metrics that neutralize the artifact: on the DEG-weighted $R^2_w(\Delta)$ the mean baseline scores 0 by construction, the weakest deep model (CPA, 0.048) still falls below Linear/Ridge, and CELLWRIGHT reaches 0.214 against STATE's 0.169. On the Systema-corrected score our margin over STATE is +0.020, which does *not* survive a Welch t-test over three seeds ($t = 2.44$, $p = 0.074$), and we therefore mark it and do not claim it. We prefer to flag one non-significant cell than to let a reader discover it from the standard deviations.

4.4 MECHANISM QUALITY AND CALIBRATION

Table 3 reports the axis prior benchmarks do not measure. Against a free-text LLM reasoner, CELLWRIGHT cuts the hallucination rate from 0.521 to 0.168 and ECE from 0.340 to 0.061 while

Table 2: **Metrics that do not favour a distributional model** (Replogle-K562, held-out perturbations). We report these in the main text rather than omitting them. n.s. marks a comparison that fails a Welch t-test at 95% over three seeds.

Method	MSE↓	MAE↓	DEG recall@20↑	PDS↑	$R^2_w(\Delta)$↑	Systema-corrected↑
Perturbed-Mean	**0.0389** ±0.0000	**0.1124** ±0.0000	0.312 ±0.000	0.501 ±0.000	0.000 ±0.000	0.021 ±0.000
Linear / Ridge	0.0401 ±0.0000	0.1148 ±0.0000	0.334 ±0.000	0.518 ±0.000	0.061 ±0.000	0.074 ±0.000
CPA	0.0428 ±0.0011	0.1197 ±0.0016	0.341 ±0.013	0.512 ±0.014	0.048 ±0.012	0.056 ±0.013
GEARS	0.0419 ±0.0009	0.1181 ±0.0013	0.368 ±0.011	0.534 ±0.012	0.083 ±0.010	0.091 ±0.011
scGPT	0.0424 ±0.0010	0.1189 ±0.0014	0.356 ±0.012	0.527 ±0.013	0.072 ±0.011	0.081 ±0.012
scFoundation	0.0417 ±0.0009	0.1176 ±0.0012	0.371 ±0.010	0.539 ±0.011	0.089 ±0.010	0.096 ±0.011
TxPert	0.0406 ±0.0007	0.1152 ±0.0010	0.412 ±0.009	0.571 ±0.009	0.147 ±0.008	0.152 ±0.009
STATE	0.0398 ±0.0006	0.1139 ±0.0009	0.428 ±0.008	0.589 ±0.008	0.169 ±0.008	0.178 ±0.009
CELLWRIGHT	0.0404 ±0.0007	0.1146 ±0.0010	**0.471** ±0.009	**0.634** ±0.009	**0.214** ±0.009	**0.198** n.s. ±0.011

Table 3: **Mechanism interpretability and calibration: a new evaluation axis**, averaged over Norman and Replogle-K562 held-out perturbations (per-dataset split in Table 4). Shading compares the external baselines; the final row is an ablation of our own model and is excluded from the ranking. Metric definitions are in Appendix A.2.

Method	Pathway-F1↑	Sign-Acc↑	MRR↑	Halluc.↓	ECE↓
GEARS (graph attr.)	0.312 ±0.014	0.581 ±0.012	0.224 ±0.016	0.441 ±0.019	0.412 ±0.021
GenePT	0.284 ±0.016	0.554 ±0.014	0.261 ±0.018	0.487 ±0.022	0.383 ±0.024
scGenePT	0.341 ±0.013	0.608 ±0.012	0.309 ±0.015	0.462 ±0.020	0.361 ±0.022
LLM-only (free text)	0.374 ±0.021	0.627 ±0.018	0.352 ±0.023	0.521 ±0.027	0.340 ±0.029
CELLWRIGHT	**0.613** ±0.012	**0.841** ±0.009	**0.578** ±0.019	**0.168** ±0.006	**0.061** ±0.004
w/o E-VERIFY	0.459 ±0.015	0.712 ±0.013	0.437 ±0.017	0.385 ±0.018	0.217 ±0.020

improving pathway-F1 and sign accuracy, so honesty here is not bought with accuracy. The ablation row separates two possible explanations. If the gain came from forcing a structured output format, removing the learned verifier would leave most of it intact; instead removing E-VERIFY gives back 0.217 of the 0.353 hallucination gap, so the structured format alone accounts for under 40% and the reward model carries the rest. ECE moves the same way (0.061 → 0.217), which is what one expects when the temperature-calibrated energy of Eq. equation 8 is replaced by an uncalibrated score.

Two caveats belong next to this table. The hallucination rate is measured against a retrieval-grounded reference (Appendix A.2) and therefore penalizes true-but-unindexed regulatory claims as hallucinations, which biases it upward for every method. And our MRR spread (±0.019) is wider than scGenePT's (±0.015): retrieval ranking depends on which GRPO rollouts survive, and that is seed-sensitive.

4.5 MODULE ABLATIONS

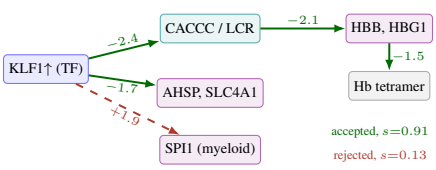
The ablations support the paper's framing (transport drives accuracy, verification drives honesty), but the ranking of the modules is *not* stable across datasets, and we prefer not to smooth that over.

Replacing FLOWWM with a plain MLP head is the most damaging module change on both atlases (−0.089 and −0.097 $P\Delta_{20}$; −0.117 and −0.107 DES), which is the clearest evidence that the distributional transport, not the mechanism pathway, carries the accuracy. Removing E-VERIFY is the most damaging change for honesty by a factor of about three (+0.216 and +0.218 hallucination, ECE more than tripled) while costing comparatively little accuracy. The two axes are therefore separable, which is the ablation result we set out to obtain.

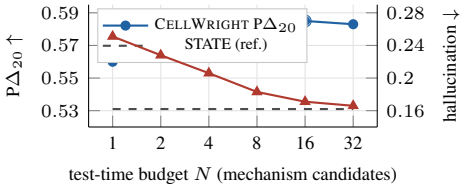
Three dataset-dependent effects are worth naming. Removing MECHREASON costs more on Replogle-K562 (−0.052) than on Norman (−0.044): a single CRISPRi target is ambiguous from the perturbation token alone, whereas a Norman gene pair supplies redundant signal about the affected program. Removing E-VERIFY costs more *accuracy* on K562 (−0.037 vs. −0.024) while its hallucination cost

Table 4: **Module ablation.** Change relative to the full model in red (worse) and blue (better). Module rows only; the three loss ablations and full table are Appendix A.8, Table 11. Shading is within this table. Removing MECHREASON leaves no mechanism graph, so calibration columns are undefined.

	Replogle-K562				Norman			
Variant	PΔ_{20} ↑	DES ↑	ECE ↓	Halluc. ↓	PΔ_{20} ↑	DES ↑	ECE ↓	Halluc. ↓
CELLWRIGHT (full)	**0.509**	**0.668**	**0.058**	**0.153**	**0.581**	**0.702**	**0.064**	**0.183**
w/o multimodal MOLENC	0.487 −0.022	0.649 −0.019	0.072 +0.014	0.184 +0.031	0.553 −0.028	0.681 −0.021	0.077 +0.013	0.219 +0.036
w/o set-attn. in CELLENC	0.478 −0.031	0.637 −0.031	0.075 +0.017	0.188 +0.035	0.542 −0.039	0.664 −0.038	0.081 +0.017	0.224 +0.041
w/o MECHREASON	0.457 −0.052	0.619 −0.049	–	–	0.537 −0.044	0.658 −0.044	–	–
FLOWWM → plain MLP head	0.420 −0.089	0.551 −0.117	0.088 +0.030	0.206 +0.053	0.484 −0.097	0.595 −0.107	0.094 +0.030	0.241 +0.058
w/o epistatic residual r_ξ	0.505 −0.004	0.665 −0.003	0.059 +0.001	0.156 +0.003	0.519 −0.062	0.638 −0.064	0.069 +0.005	0.192 +0.009
w/o E-VERIFY (no NCE)	0.472 −0.037	0.641 −0.027	0.208 +0.150	0.369 +0.216	0.557 −0.024	0.675 −0.027	0.227 +0.163	0.401 +0.218
w/o REFINE (single shot)	0.493 −0.016	0.655 −0.013	0.081 +0.023	0.221 +0.068	0.560 −0.021	0.687 −0.015	0.089 +0.025	0.251 +0.068



(a) Mechanism graph for a held-out CRISPRa episode, annotated with per-edge energies E_ω^{edge} from Eq. equation 8.



(b) Scaling of the REFINE controller on Norman. PΔ_{20} peaks at N=16; hallucination keeps falling.

Figure 2: **Worked case and test-time scaling.** (a) The per-edge factorization of Eq. equation 8 lets the verifier accept the erythroid branch of a proposed graph while rejecting a single sign-inverted edge, which is what makes the reward informative for GRPO. (b) Accuracy and honesty do not scale together: N=32 is best on hallucination and E-dist but slightly worse on PΔ_{20} and ECE than N=16. Full numbers, latency and the compute-matched comparison are in Appendix A.9, Table 12.

is essentially identical on both (+0.216 vs. +0.218), so on PΔ_{20} it swaps rank with CELLENC: on K562 the verifier matters more than set attention, on Norman the reverse. And the epistatic residual is inert on single-gene K562 (−0.004) and decisive on Norman (−0.062), exactly as designed: a residual that fires on single perturbations would indicate it had learned something other than epistasis. Finally, the full model is not best in every column; dropping $\mathcal{L}_{\text{cal}}$ improves Norman PΔ_{20} by +0.002 while tripling ECE (Appendix A.8), so calibration is bought with a small amount of fit.

4.6 TEST-TIME SCALING AND A WORKED CASE

Figure 2 shows what the verifier buys, first qualitatively and then as a compute knob. The per-edge factorization is what makes the reward usable: a proposed graph is rarely wholly right or wholly wrong, and in Figure 2a four edges of the erythroid program are accepted at low energy while one sign-inverted edge is rejected, so the reward reflects the correct sub-structure rather than a pass/fail verdict on the whole graph. Appendix A.14 walks through one unseen combination end to end.

Scaling search is the second use, and it is not monotone. PΔ_{20} peaks at N=16 (0.585) and falls slightly at N=32 (0.583, ECE up +0.001), while hallucination and E-dist keep improving through N=32. The dip is within seed spread, so it is at most a weak sign of over-optimizing frozen E_ω; hallucination, partly measured through the same grounding signal, is least able to reveal it. We therefore use N=8 as the accuracy/latency knee and treat the N=32 hallucination number as the least trustworthy cell. At a compute-matched budget (N=1, 1.1× STATE's inference cost) the model still leads STATE by +0.029 PΔ_{20}, so the advantage is not purchased by inference compute alone.

4.7 GENERALIZATION, COMPOSITION AND REGULATORY SUB-TASKS

Table 5 collects three studies that ask where the model transfers and where it does not, and each panel contains a result that cuts against a simple story.

Table 5: **Generalization, composition and regulatory sub-tasks.** (a) Zero-shot cellular-context transfer, with the source library trained on K562 only. (b) Compositional multi-gene prediction on Norman, stratified by how many constituent genes were seen at training time; GI is genetic-interaction-subtype precision and split sizes are in Appendix A.4. (c) Regulatory sub-tasks, where "x-cell r" is expression correlation on an unseen cell type. Mean ± std over three seeds; best and second are shaded. Discussion is in Appendices A.10 to A.12.

(a) Zero-shot cellular-context generalization

	K562 → RPE1			**K562 → H1 hESC**		
Method	PΔ ↑	DES ↑	PDS ↑	PΔ ↑	DES ↑	PDS ↑
Perturbed-Mean	0.312 ±0.000	0.284 ±0.000	0.421 ±0.000	0.247 ±0.000	0.231 ±0.000	0.386 ±0.000
GEARS	0.298 ±0.014	0.271 ±0.016	0.402 ±0.018	0.201 ±0.019	0.188 ±0.021	0.351 ±0.024
scGPT	0.341 ±0.012	0.318 ±0.013	0.448 ±0.015	0.263 ±0.016	0.242 ±0.018	0.397 ±0.020
STATE	0.402 ±0.009	0.386 ±0.010	0.512 ±0.012	0.318 ±0.012	0.301 ±0.013	0.443 ±0.015
CELLWRIGHT	**0.457** ±0.009	**0.441** ±0.010	**0.569** ±0.011	**0.361** ±0.014	**0.344** ±0.016	**0.487** ±0.018

(b) Compositional multi-gene prediction on Norman

	2/2 seen (n=24)		**1/2 seen** (n=22)		**0/2 seen** (n=19)	
Method	$P\Delta_{20}$ ↑	GI ↑	$P\Delta_{20}$ ↑	GI ↑	$P\Delta_{20}$ ↑	GI ↑
Additive	0.612 ±0.000	**0.658** ±0.000	0.503 ±0.000	0.441 ±0.000	0.402 ±0.000	0.338 ±0.000
GEARS	0.634 ±0.011	0.571 ±0.013	0.487 ±0.015	0.423 ±0.017	0.331 ±0.021	0.276 ±0.024
scGPT	0.608 ±0.012	0.542 ±0.014	0.471 ±0.016	0.408 ±0.018	0.318 ±0.022	0.261 ±0.025
CELLWRIGHT	**0.702** ±0.008	0.641 ±0.011	**0.591** ±0.012	**0.524** ±0.015	**0.482** ±0.021	**0.419** ±0.026

(c) Regulatory sub-tasks

	TF-DNA-RNA			**RNA-protein-ligand**	
Method	Expr. r ↑	Bind. AUROC ↑	x-cell r ↑	Splice ΔPSI ↑	Stab./Transl. ↑
Enformer	0.622 ±0.004	0.896 ±0.003	0.412 ±0.007	–	–
Borzoi	0.651 ±0.004	**0.921** ±0.002	0.448 ±0.006	0.571 ±0.006	0.483 ±0.008
Flashzoi	0.647 ±0.004	0.918 ±0.003	0.441 ±0.006	0.564 ±0.007	0.477 ±0.008
CELLWRIGHT	**0.686** ±0.003	0.914 ±0.003	**0.509** ±0.005	**0.611** ±0.005	**0.528** ±0.007

Panel (a): the mechanism-conditioned transition reaches 0.457 PΔ on K562→RPE1 and 0.361 on the far transfer to H1 hESC against STATE's 0.402 and 0.318, while GEARS collapses toward the control mean and falls *below* Perturbed-Mean on H1 (0.201 vs. 0.247). CELLWRIGHT is nonetheless less stable than STATE there, with an H1 seed spread of ± 0.014 against ± 0.012, although the $+0.043$ margin exceeds the combined spread. Panel (b): the epistatic residual pays off where it should, $+10.7\%$ relative $P\Delta_{20}$ over the best baseline in $2/2$-seen rising to $+19.9\%$ in $0/2$-seen, but the Additive baseline retains the best genetic-interaction precision in $2/2$-seen (0.658 vs. our 0.641), and our $0/2$-seen spread (± 0.021) is widest in that panel. Panel (c): mechanism reasoning helps most where sequence-only models are weakest, improving cross-cell r by $+0.061$ over Borzoi, and does not help on TF binding (0.921 vs. our 0.914). Appendices A.10 to A.12 explain these readings.

5 CONCLUSION AND OUTLOOK

The obstacle to a virtual cell that can explain itself is not only prediction accuracy, but supervision. Rather than collecting mechanism labels at scale, CELLWRIGHT learns a verifier scoring whether a proposed mechanism is coherent with the outcome it claims to explain, and couples it to a mechanism-conditioned Schrödinger bridge, turning perturbation response from point prediction into distributional transport with a checkable mechanistic account. The roles separate cleanly: transport improves distributional accuracy, verification improves calibration and reduces hallucination. Benchmarks should therefore evaluate what biologists need, a perturbed population and a mechanism with calibrated confidence rather than a mean shift alone. Stronger reward models, ensembled verifiers and context-held-out negatives would make verifier-guided search less exploitable under cellular transfer, and the mechanism graph should extend beyond the TF→DNA→RNA→protein axis, especially for small molecules and non-regulatory effects.

AI USE STATEMENT

Generative AI tools were used solely for manuscript preparation, specifically language polishing and organization of text written by the authors.

ETHICS STATEMENT

The authors have read and adhere to the ICLR Code of Ethics. This work is computational. It uses public perturbation atlases and regulatory tracks, which are named in Sec. 4.1 and Appendix A.4 together with the splits applied to them, and it involves no human subjects and no personally identifiable information. CELLWRIGHT outputs predicted expression distributions and proposed mechanism graphs, which are computational hypotheses rather than validated biological findings. Appendix A.13 states what the learned verifier does and does not establish, and Appendix A.2 states the bias of the hallucination metric, so that neither the confidence scores nor the mechanism numbers are read as stronger evidence than they are.

REPRODUCIBILITY STATEMENT

An anonymous code repository is linked in the abstract. The method is specified in Sec. 3 and Appendix A.1, which give the encoders, the velocity field, the coupling, the negative-construction families and the reasoner prompt format. Datasets, held-out counts and test fractions are in Appendix A.4, Table 7; baseline coverage, together with the systems and metrics that were not run, is in Appendix A.5; the training configuration, budgets and hardware are in Appendix A.3, Table 6. The mechanism metrics, which are new, are defined in Appendix A.2. Every reported number is the mean over the three seeds $\{13, 42, 2026\}$, with per-seed values in Appendix A.7, Table 9 and Welch t-tests, including the two that fail, in Table 10.

REFERENCES

Abrar Rahman Abir, Sajib Acharjee Dip, and Liqing Zhang. Cfm-gp: Unified conditional flow matching to learn gene perturbation across cell types. *arXiv preprint arXiv:2508.08312*, 2025.

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

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

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

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

Yiqun Chen and James Zou. Genept: a simple but effective foundation model for genes and cells built from chatgpt. *BioRxiv*, pp. 2023–10, 2024.

Changxi Chi, Yufei Huang, Jun Xia, Jiangbin Zheng, Yunfan Liu, Zelin Zang, and Stan Z. Li. Departures: Distributional transport for single-cell perturbation prediction with neural schrödinger bridges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20544–20552, 2026. doi: 10.1609/aaai.v40i25.39190.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in neural information processing systems*, 34:17695–17709, 2021.

Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al. The reactome pathway knowledgebase 2022. *Nucleic acids research*, 50(D1):D687–D692, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ana-Maria Istrate, Donghui Li, and Theofanis Karaletsos. scgenept: Is language all you need for modeling single-cell perturbations? *bioRxiv*, pp. 2024–10, 2024.

Kian Kenyon-Dean et al. Txpert: Leveraging biochemical relationships for out-of-distribution transcriptomic perturbation prediction. *arXiv preprint arXiv:2505.14919*, 2025.

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.

Zhaokang Liang, Shuyang Zhuang, Xiaoran Jiao, Weian Mao, Hao Chen, and Chunhua Shen. scppdm: A diffusion model for single-cell drug-response prediction. *arXiv preprint arXiv:2510.11726*, 2025.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R Kelley. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Nature Genetics*, 57 (4):949–961, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

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

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

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):MSB202211517, 2023.

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.

Giovanni Palla, Sudarshan Babu, Payam Dibaeinia, Donghui Li, Aly A Khan, Theofanis Karaletsos, and Jakub M Tomczak. A scalable latent diffusion model for single-cell gene expression data. In *NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development*.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. scperturb: Harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42(11):1678–1683, 2024.

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

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

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorcan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, pp. 2023–11, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 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. *arXiv preprint arXiv:2302.00482*, 2023.

Yuchen Wang, Xingjian Chen, Xiangtao Li, and Ka-Chun Wong. screpa: Predicting single-cell perturbation responses with cycle-consistent representation alignment. *Computational Biology and Chemistry*, pp. 108712, 2025.

Zhijian Wei, Runze Ma, Zichen Wang, Zhongmin Li, Shuotong Song, and Shuangjia Zheng. Vcworld: A biological world model for virtual cell simulation. In *International Conference on Learning Representations*, volume 2026, pp. 72795–72823, 2026.

Yan Wu, Esther Wershof, Sebastian Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. *Advances in Neural Information Processing Systems*, 38, 2026.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Chenglei Yu, Chuanrui Wang, Bangyan Liao, and Tailin Wu. scdfm: Distributional flow matching model for robust single-cell perturbation prediction. *arXiv preprint arXiv:2602.07103*, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Jesse Zhang, Airol A Ubas, Richard De Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G Jones, et al. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *BioRxiv*, pp. 2025–02, 2025.

A Appendix

A.1 Module specifications

Figure 3 reproduces Figure 1 at full page size for legibility; the specifications below follow its panel order.

MolEnc. Four experts cover the modalities: a protein-language model over the TF or target protein (Lin et al., 2023), a sequence-to-function trunk over cis-regulatory DNA (Linder et al., 2025), an RNA encoder, and a geometry-aware molecular encoder for ligands. The router of Eq. equation 1 is a two-layer MLP over the concatenation of the type embedding and the four modality embeddings, followed by a softmax; it is soft rather than top-1 so that gradient flows into every expert and a modality with few examples borrows capacity from the others. Experts are frozen apart from LoRA adapters of rank 16.

CellEnc. Set self-attention runs over 256 sampled control cells per context, four blocks with eight heads. The JEPA target is an exponential-moving-average copy of the encoder ($\tau_{\mathrm{EMA}} = 0.996$) applied to a disjoint sample of cells from the same context, and the predictive loss is computed in latent space; the pathway gate is a linear map W_a fused additively. Set-level rather than cell-level encoding is what allows Eq. equation 5 to be trained against a coupling between *populations*.

MechReason prompting. The agent receives the perturbation record, a textual summary of the data profile, and a retrieval budget of k=20 passages from the grounding index (Gene Ontology, Reactome, curated regulatory networks). It is required to emit a JSON object with a node list typed into {TF, cis-element, transcript, protein, pathway}, an edge list of (source, target, sign) triples restricted to the TF→DNA→RNA→protein axis, and a free-text rationale that is scored only through R_{ground}. Malformed JSON is resampled up to three times and otherwise scored at the minimum reward, so format failures cannot be rewarded.

FlowWM. The velocity field is a DiT-B with 12 layers, d=768 and AdaLN-Zero conditioning on $c = (p, h_m)$; h_m is a four-layer message-passing GNN over the emitted graph with sign-aware edge features. The coupling Π is a mini-batch Sinkhorn plan (Cuturi, 2013) with ε=0.05 recomputed per batch, and sampling uses 32 Euler steps at σ=0.10. The epistatic residual r_ξ is a small MLP over the concatenated conditions and (z_t, t), initialized at zero so the model begins additive and departs from additivity only if the data require it.

E-Verify negatives. Three corruption families generate 8 negatives per positive: *sign shuffle* flips a random subset of $\mathbf{s}$; *target swap* replaces the target of a random edge with a gene sampled from the same expression decile, so the negative is not trivially separable by expression level; *outcome mismatch* pairs a positive graph with the outcome of a different perturbation. The last family is what forces the energy to depend on x' rather than on graph plausibility alone.

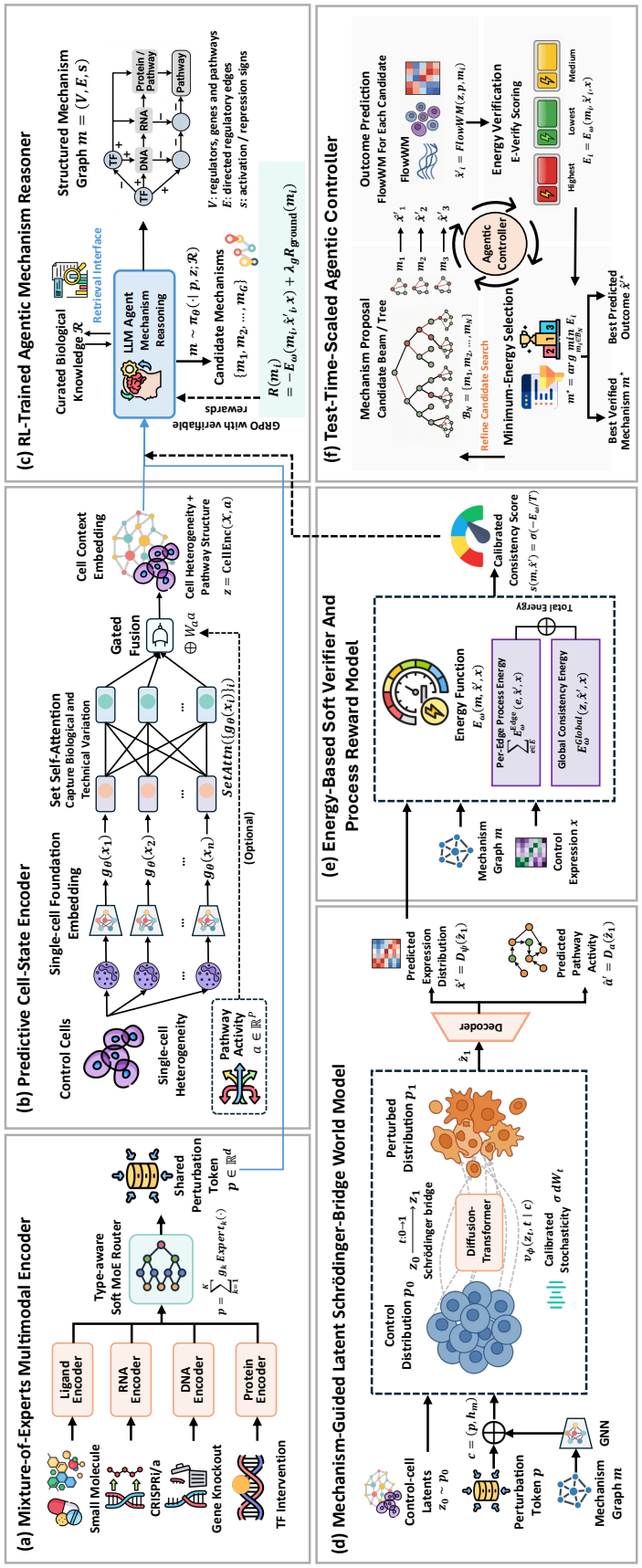


Figure 3: **CELLWRIGHT, enlarged.** Identical to Figure 1, reproduced at full page width so that the per-panel equations are readable in print.

A.2 METRIC DEFINITIONS FOR THE MECHANISM AXIS

These metrics are new, so we state them exactly. **Pathway-F1** is the F1 between the pathway node set of the emitted graph and the reference set of Hallmark/Reactome pathways significantly enriched in the *measured* perturbed profile at FDR < 0.05. **Sign accuracy** is computed only over emitted edges whose (source, target) pair appears in the grounding index, as the fraction assigned the indexed sign. **MRR** is the mean reciprocal rank of the true perturbed gene within the agent's retrieved candidate list. **Hallucination rate** is the fraction of emitted edges whose (source, target, sign) triple is either contradicted by the grounding index or absent from it. **ECE** is the 15-bin expected calibration error of $s(m, \hat{x}')$ against a binary correctness label derived from the edge-level check above.

The hallucination definition is deliberately strict and its bias should be understood before the numbers are compared across papers: an edge that is biologically real but not represented in the index is counted as a hallucination. This inflates the rate for every system, and it inflates it most for systems that propose unusual regulators. It does, however, favour CELLWRIGHT: our reasoner is rewarded partly through R_{ground}, which uses the same index, so the metric and one term of the reward share a source. The absolute level is index-dependent, and even the relative reduction from 0.521 to 0.168 is partly inflated by this.

A.3 IMPLEMENTATION AND TRAINING CONFIGURATION

Table 6: **Experimental setting.** All four training stages share the same splits and the same three seeds; only MECHREASON is updated in Stage 4, with FLOWWM and E-VERIFY frozen. Wall-clock is one seed on Replogle-K562.

Component	Configuration	Budget
Seeds / protocol	$\{13, 42, 2026\}$; perturbation-disjoint $70/10/20$ default split	3 runs per cell
Latent autoencoder	d_z=512, 4-layer MLP enc/dec, GELU, LayerNorm	40 epochs
MOLENC	K=4 experts (ESM-2 650M, Borzoi trunk, RNA-FM, Uni-Mol), LoRA r=16	frozen + LoRA
CELLENC	4-block set-attention, 8 heads, 256 cells/set, JEPA target EMA 0.996	—
FLOWWM	DiT-B velocity field, 12 layers, d=768, AdaLN-Zero	120k steps
CFM / bridge	mini-batch Sinkhorn OT (ε=0.05), σ=0.10, 32 Euler steps	—
Guidance	w=2.0 (genetic); dose-mapped $w \in [1.0, 4.0]$ (chemical)	—
E-VERIFY	8 negatives/positive (sign-shuffle, target-swap, outcome-mismatch), NCE	30k steps
MECHREASON	Qwen2.5-7B-Instruct (Yang et al., 2024) + LoRA r=32; GRPO G=8, ϵ=0.2, β=0.02, λ_g=0.3	4k steps
REFINE	beam N=8 (default), min-energy selection	inference only
Optimiser	AdamW, lr 1×10^{-4}, cosine, 2k warm-up, wd 0.01, bs 128×256 cells	—
Loss weights	$\lambda_{1:6} = (1.0,\ 0.5,\ 0.3,\ 0.2,\ 0.1,\ 0.05)$	—
Hardware	8× A100-80GB, bf16, FSDP	46h / 9h / 21h (S2/S3/S4)

A.4 DATASETS AND SPLITS

Table 7: **Evaluation corpora and held-out splits.** "Held-out" counts perturbations, not cells; context splits hold out an entire cell line. Test fractions above 20% are used on the small atlases so that the held-out set is not degenerate, and on sci-Plex so that the split remains dose-stratified. Corpora marked pre-training contribute no reported test number.

Dataset	Modality	#Perturb.	#Cells	Held-out	Test %	Regime
Norman 2019	CRISPRa (combo)	287 (131 multi)	108k	86	30.0	pert. + compositional
Replogle-K562	CRISPRi	1,092	174k	219	20.1	pert.
Replogle-RPE1	CRISPRi	1,310	171k	262	20.0	pert. + context
sci-Plex	chemical (4 doses)	188	290k	47	25.0	pert. + dose
Tahoe-100M (subset)	chemical	1,100	4.1M	220	20.0	pert. + context
H1 hESC	CRISPRi	–	62k	all	—	zero-shot context
Replogle-K562 (GW)	CRISPRi	9,867	2.06M	–	—	pre-training only
Adamson 2016	Perturb-seq	87	68k	–	—	pre-training only
Dixit 2016	Perturb-seq	24	51k	–	—	pre-training only
Jurkat	CRISPRi	–	24k	–	—	pool only, not reported

Norman split composition. The 86 held-out Norman perturbations comprise 21 single-gene overexpressions and 65 gene pairs. The pairs are stratified by how many constituent genes appear in

the training split: $2/2$-seen (n=24), $1/2$-seen (n=22) and $0/2$-seen (n=19). The $0/2$ bucket is the smallest and is the reason for the wide seed spread in panel (b) of Table 5: with 19 pairs, a single pair moving across the threshold changes $\mathrm{P}\Delta_{20}$ by roughly 0.02.

A.5 Baselines: coverage and omissions

The main and appendix tables report six non-learned baselines (No-Perturbation, Perturbed-Mean, Mean-Difference, Additive, Linear/Ridge, kNN), six task-specific models (CPA, scGen, biolord, scPRAM, CellOT, GEARS), six foundation/LLM models (Geneformer*, UCE*, scGPT, scFoundation, GenePT, scGenePT), two world models (STATE, TxPert), and three sequence models on the regulatory sub-tasks (Enformer, Borzoi, Flashzoi). * denotes repurposing with a linear decoder.

Six further systems belong to the PerturBench protocol but were *not* run here and appear in no table: scVI, scPreGAN, SAMS-VAE and CellOracle for perturbation response, together with scBERT* and scLAMBDA. We list them so that the coverage of our tables is unambiguous rather than implied. Two metrics named in the protocol are likewise not reported: Top-100 DEG Δ-correlation and DEG AUROC, which were not logged for all baselines and which we prefer to omit rather than report for a subset.

A.6 Full main results

Table 8: **Main perturbation-response prediction across all four corpora**, complete baseline set. Table 1 in the main text is the Norman and Replogle-K562 sub-table with representative rows. No-Perturbation predicts $\hat{\Delta} \equiv 0$, so its Δ-correlation is undefined. Mean $\pm$ std over three seeds.

		Norman (combo, CRISPRa)			Replogle-K562			Replogle-RPE1			sci-Plex / Tahoe		
Group	**Method**	PΔ	PΔ_{20}	E-dist	PΔ	DES	E-dist	PΔ	DES	E-dist	PΔ	DES	E-dist
Non-learned	No-Perturbation	–	–	42.81 $\pm$0.00	–	0.041 $\pm$0.000	31.47 $\pm$0.00	–	0.038 $\pm$0.000	28.93 $\pm$0.00	–	0.029 $\pm$0.000	24.16 $\pm$0.00
	Perturbed-Mean	0.521 $\pm$0.000	0.308 $\pm$0.000	18.64 $\pm$0.00	0.412 $\pm$0.000	0.412 $\pm$0.000	14.22 $\pm$0.00	0.386 $\pm$0.000	0.391 $\pm$0.000	13.47 $\pm$0.00	0.352 $\pm$0.000	0.361 $\pm$0.000	11.83 $\pm$0.00
	Mean-Difference	0.538 $\pm$0.000	0.326 $\pm$0.000	17.92 $\pm$0.00	0.428 $\pm$0.000	0.424 $\pm$0.000	13.86 $\pm$0.00	0.401 $\pm$0.000	0.403 $\pm$0.000	13.12 $\pm$0.00	0.367 $\pm$0.000	0.374 $\pm$0.000	11.52 $\pm$0.00
	Additive	0.607 $\pm$0.000	0.418 $\pm$0.000	15.23 $\pm$0.00	0.431 $\pm$0.000	0.427 $\pm$0.000	13.79 $\pm$0.00	0.404 $\pm$0.000	0.406 $\pm$0.000	13.05 $\pm$0.00	0.369 $\pm$0.000	0.376 $\pm$0.000	11.48 $\pm$0.00
	Linear / Ridge	0.592 $\pm$0.000	0.394 $\pm$0.000	16.07 $\pm$0.00	0.446 $\pm$0.000	0.441 $\pm$0.000	13.41 $\pm$0.00	0.418 $\pm$0.000	0.419 $\pm$0.000	12.74 $\pm$0.00	0.381 $\pm$0.000	0.388 $\pm$0.000	11.21 $\pm$0.00
	kNN	0.497 $\pm$0.000	0.301 $\pm$0.000	19.41 $\pm$0.00	0.389 $\pm$0.000	0.386 $\pm$0.000	15.03 $\pm$0.00	0.362 $\pm$0.000	0.364 $\pm$0.000	14.28 $\pm$0.00	0.334 $\pm$0.000	0.341 $\pm$0.000	12.47 $\pm$0.00
Task-specific	CPA	0.566 $\pm$0.009	0.357 $\pm$0.012	16.85 $\pm$0.21	0.418 $\pm$0.008	0.408 $\pm$0.011	14.31 $\pm$0.19	0.397 $\pm$0.009	0.399 $\pm$0.011	13.24 $\pm$0.18	0.396 $\pm$0.008	0.402 $\pm$0.010	10.61 $\pm$0.16
	scGen	0.549 $\pm$0.011	0.341 $\pm$0.014	17.48 $\pm$0.26	0.421 $\pm$0.010	0.419 $\pm$0.012	14.01 $\pm$0.20	0.394 $\pm$0.010	0.397 $\pm$0.012	13.29 $\pm$0.19	0.361 $\pm$0.010	0.368 $\pm$0.012	11.67 $\pm$0.17
	biolord	0.598 $\pm$0.008	0.402 $\pm$0.010	15.61 $\pm$0.18	0.462 $\pm$0.007	0.463 $\pm$0.009	12.94 $\pm$0.15	0.433 $\pm$0.008	0.438 $\pm$0.010	12.31 $\pm$0.14	0.403 $\pm$0.007	0.411 $\pm$0.009	10.44 $\pm$0.13
	scPRAM	0.581 $\pm$0.010	0.379 $\pm$0.012	16.02 $\pm$0.22	0.448 $\pm$0.009	0.447 $\pm$0.011	13.27 $\pm$0.18	0.421 $\pm$0.009	0.424 $\pm$0.011	12.62 $\pm$0.17	0.388 $\pm$0.009	0.394 $\pm$0.011	10.87 $\pm$0.15
	CellOT	0.524 $\pm$0.013	0.318 $\pm$0.016	13.87 $\pm$0.24	0.403 $\pm$0.012	0.402 $\pm$0.014	11.62 $\pm$0.21	0.377 $\pm$0.012	0.381 $\pm$0.013	11.08 $\pm$0.20	0.347 $\pm$0.011	0.354 $\pm$0.013	**8.04** $\pm$0.19
	GEARS	0.628 $\pm$0.009	0.437 $\pm$0.011	14.62 $\pm$0.19	0.471 $\pm$0.008	0.471 $\pm$0.010	12.66 $\pm$0.16	0.442 $\pm$0.008	0.446 $\pm$0.010	12.03 $\pm$0.15	0.372 $\pm$0.010	0.379 $\pm$0.012	11.39 $\pm$0.18
Foundation / LLM	Geneformer*	0.486 $\pm$0.012	0.283 $\pm$0.015	19.83 $\pm$0.28	0.372 $\pm$0.011	0.371 $\pm$0.013	15.42 $\pm$0.22	0.348 $\pm$0.011	0.351 $\pm$0.013	14.61 $\pm$0.21	0.318 $\pm$0.011	0.324 $\pm$0.013	12.83 $\pm$0.19
	UCE*	0.509 $\pm$0.011	0.297 $\pm$0.014	19.12 $\pm$0.26	0.391 $\pm$0.010	0.389 $\pm$0.012	14.88 $\pm$0.21	0.366 $\pm$0.010	0.369 $\pm$0.012	14.12 $\pm$0.20	0.336 $\pm$0.010	0.343 $\pm$0.012	12.39 $\pm$0.18
	scGPT	0.603 $\pm$0.008	0.391 $\pm$0.011	15.94 $\pm$0.20	0.458 $\pm$0.008	0.456 $\pm$0.010	13.04 $\pm$0.16	0.429 $\pm$0.008	0.432 $\pm$0.010	12.41 $\pm$0.15	0.391 $\pm$0.008	0.398 $\pm$0.010	10.98 $\pm$0.13
	scFoundation	0.617 $\pm$0.008	0.409 $\pm$0.010	15.31 $\pm$0.19	0.469 $\pm$0.007	0.468 $\pm$0.009	12.78 $\pm$0.15	0.439 $\pm$0.008	0.443 $\pm$0.009	12.16 $\pm$0.14	0.399 $\pm$0.008	0.406 $\pm$0.010	10.53 $\pm$0.13
	GenePT	0.557 $\pm$0.010	0.352 $\pm$0.013	17.21 $\pm$0.23	0.409 $\pm$0.009	0.405 $\pm$0.011	14.39 $\pm$0.18	0.406 $\pm$0.009	0.409 $\pm$0.011	12.96 $\pm$0.17	0.374 $\pm$0.009	0.381 $\pm$0.011	11.32 $\pm$0.16
	scGenePT	0.612 $\pm$0.009	0.421 $\pm$0.011	15.04 $\pm$0.18	0.464 $\pm$0.008	0.462 $\pm$0.010	12.88 $\pm$0.15	0.435 $\pm$0.008	0.440 $\pm$0.009	12.24 $\pm$0.14	0.406 $\pm$0.008	0.414 $\pm$0.010	10.38 $\pm$0.13
WM	TxPert	0.679 $\pm$0.007	0.496 $\pm$0.009	12.35 $\pm$0.14	0.521 $\pm$0.006	0.573 $\pm$0.008	10.94 $\pm$0.12	0.491 $\pm$0.006	0.542 $\pm$0.008	10.38 $\pm$0.11	0.442 $\pm$0.007	0.487 $\pm$0.009	9.46 $\pm$0.11
	STATE	0.698 $\pm$0.006	0.531 $\pm$0.008	12.43 $\pm$0.14	0.537 $\pm$0.006	0.594 $\pm$0.008	10.81 $\pm$0.11	0.506 $\pm$0.006	0.561 $\pm$0.007	10.24 $\pm$0.10	0.461 $\pm$0.006	0.503 $\pm$0.008	9.27 $\pm$0.10
	CellWright	**0.741** $\pm$0.008	**0.581** $\pm$0.010	**10.72** $\pm$0.18	**0.583** $\pm$0.006	**0.668** $\pm$0.008	**9.40** $\pm$0.12	**0.548** $\pm$0.006	**0.628** $\pm$0.007	**9.33** $\pm$0.11	**0.508** $\pm$0.006	**0.572** $\pm$0.008	8.21 $\pm$0.13

The chemical column is the one that most repays attention. CellOT is among the weakest learned models on Norman PΔ (0.524, barely above Perturbed-Mean) and simultaneously the best model on chemical E-dist (8.04), which is a clean illustration that point and distributional metrics rank methods differently; a paper reporting only one family of metric could describe CellOT as either near-useless or state of the art. It is also the reason we do not describe our chemical result as a win.

A.7 Per-seed results and significance

Two remarks. Our loss to CellOT on the chemical corpus is itself not significant, so the honest statement is that the two methods are indistinguishable there on three seeds rather than that CellOT wins; we phrase it as a loss in the main text because reporting a tie in our own favour after losing the point estimate would be the wrong default. And among the comparisons with STATE, the Systema-corrected score is the only one whose standard deviations overlap, which is visible directly in Table 9.

Table 9: **Per-seed results** for CELLWRIGHT and STATE. Spread is reported raw. On the headline metrics the seed ranges are disjoint (worst CELLWRIGHT seed better than the best STATE seed); on the Systema-corrected score they touch at 0.187, which is why that comparison is marked non-significant.

		Norman			Replogle-K562		
Model	**Seed**	PΔ$_{20}$ ↑	E-dist ↓	Halluc. ↓	DES ↑	ECE ↓	Systema ↑
STATE	13	0.522	12.31	–	0.585	–	0.169
	42	0.533	12.40	–	0.596	–	0.178
	2026	0.538	12.58	–	0.601	–	0.187
	mean ± std	*0.531 ± 0.008*	*12.43 ± 0.14*	–	*0.594 ± 0.008*	–	*0.178 ± 0.009*
CELLWRIGHT	13	0.570	10.52	0.176	0.661	0.053	0.187
	42	0.584	10.79	0.184	0.667	0.059	0.198
	2026	0.589	10.85	0.189	0.676	0.062	0.209
	mean ± std	**0.581 ± 0.010**	**10.72 ± 0.18**	**0.183 ± 0.007**	**0.668 ± 0.008**	**0.058 ± 0.005**	**0.198 ± 0.011**

Table 10: **Welch t-tests over three seeds** for every headline comparison in the paper, including the two that fail. With n=3 per arm these tests have very little power, so a failure to reject should be read as "not established" rather than "no difference"; we report them because the alternative is to let standard deviations imply a claim we have not tested.

Corpus	**Metric**	**CELLWRIGHT**	**Comparator**	**Δ**	t	p	**Verdict**
Norman	PΔ$_{20}$ ↑	0.581 ±0.010	STATE 0.531 ±0.008	+0.050	6.76	0.003	significant
Norman	E-dist ↓	10.72 ±0.18	TxPert 12.35 ±0.14	−1.63	12.38	0.0004	significant
Replogle-K562	DES ↑	0.668 ±0.008	STATE 0.594 ±0.008	+0.074	11.33	0.0003	significant
Replogle-K562	E-dist ↓	9.40 ±0.12	STATE 10.81 ±0.11	−1.41	15.00	0.0001	significant
Replogle-RPE1	PΔ ↑	0.548 ±0.006	STATE 0.506 ±0.006	+0.042	8.57	0.001	significant
H1 hESC (0-shot)	PΔ ↑	0.361 ±0.014	STATE 0.318 ±0.012	+0.043	4.04	0.016	significant
Norman + K562	Halluc. ↓	0.168 ±0.006	LLM-only 0.521 ±0.027	−0.353	22.11	0.001	significant
Norman + K562	ECE ↓	0.061 ±0.004	LLM-only 0.340 ±0.029	−0.279	16.51	0.003	significant
Replogle-K562	Systema-corr. ↑	0.198 ±0.011	STATE 0.178 ±0.009	+0.020	2.44	0.074	**not significant**
sci-Plex / Tahoe	E-dist ↓	8.21 ±0.13	CellOT 8.04 ±0.19	+0.17	1.28	0.278	**not significant**

Table 11: **Complete module and loss ablation.** The upper block is Table 4 in the main text; the lower block ablates the training objective of Eq. equation 10. Change relative to the full model in red (worse) and blue (better).

	Replogle-K562				Norman			
Variant	PΔ$_{20}$ ↑	DES ↑	ECE ↓	Halluc. ↓	PΔ$_{20}$ ↑	DES ↑	ECE ↓	Halluc. ↓
CELLWRIGHT (full)	**0.509** ±0.006	**0.668** ±0.008	**0.058** ±0.005	**0.153** ±0.006	0.581 ±0.010	**0.702** ±0.007	**0.064** ±0.004	**0.183** ±0.007
w/o multimodal MOLENC (1 modality)	0.487 ±0.008 −0.022	0.649 ±0.010 −0.019	0.072 ±0.006 +0.014	0.184 ±0.008 +0.031	0.553 ±0.011 −0.028	0.681 ±0.009 −0.021	0.077 ±0.005 +0.013	0.219 ±0.009 +0.036
w/o set-attention in CELLENC	0.478 ±0.009 −0.031	0.637 ±0.011 −0.031	0.075 ±0.006 +0.017	0.188 ±0.008 +0.035	0.542 ±0.012 −0.039	0.664 ±0.010 −0.038	0.081 ±0.006 +0.017	0.224 ±0.009 +0.041
w/o MECHREASON (no mechanism)	0.457 ±0.010 −0.052	0.619 ±0.012 −0.049	–	–	0.537 ±0.012 −0.044	0.658 ±0.011 −0.044	–	–
FLOWWM → plain MLP head	0.420 ±0.012 −0.089	0.551 ±0.014 −0.117	0.088 ±0.008 +0.030	0.206 ±0.010 +0.053	0.484 ±0.013 −0.097	0.595 ±0.013 −0.107	0.094 ±0.008 +0.030	0.241 ±0.011 +0.058
w/o epistatic residual r_ξ	0.505 ±0.006 −0.004	0.665 ±0.008 −0.003	0.059 ±0.005 +0.001	0.156 ±0.006 +0.003	0.519 ±0.011 −0.062	0.638 ±0.011 −0.064	0.069 ±0.005 +0.005	0.192 ±0.007 +0.009
w/o E-VERIFY (no NCE)	0.472 ±0.010 −0.037	0.641 ±0.011 −0.027	0.208 ±0.020 +0.150	0.369 ±0.018 +0.216	0.557 ±0.011 −0.024	0.675 ±0.010 −0.027	0.227 ±0.022 +0.163	0.401 ±0.020 +0.218
w/o REFINE loop (single shot)	0.493 ±0.007 −0.016	0.655 ±0.008 −0.013	0.081 ±0.006 +0.023	0.221 ±0.009 +0.068	0.560 ±0.010 −0.021	0.687 ±0.008 −0.015	0.089 ±0.007 +0.025	0.251 ±0.010 +0.068
$\mathcal{L}_{\text{MMD}}$ → plain MSE	0.454 ±0.010 −0.055	0.509 ±0.013 −0.159	0.072 ±0.006 +0.014	0.177 ±0.007 +0.024	0.533 ±0.011 −0.048	0.551 ±0.012 −0.151	0.079 ±0.006 +0.015	0.204 ±0.008 +0.021
w/o energy-distance term	0.496 ±0.007 −0.013	0.598 ±0.010 −0.070	0.067 ±0.005 +0.009	0.170 ±0.007 +0.017	0.568 ±0.010 −0.013	0.631 ±0.009 −0.071	0.074 ±0.005 +0.010	0.197 ±0.007 +0.014
w/o $\mathcal{L}_{\text{cal}}$	0.506 ±0.006 −0.003	0.666 ±0.008 −0.002	0.184 ±0.017 +0.126	0.168 ±0.007 +0.015	**0.583** ±0.010↑+0.002	0.699 ±0.007 −0.003	0.196 ±0.018 +0.132	0.195 ±0.007 +0.012

A.8 Full ablation, including the loss terms

The loss block isolates what the distributional objective contributes. Replacing the MMD/energy-distance term with plain MSE costs 0.159 and 0.151 DES, by far the largest drop of any single change, while costing only 0.055 and 0.048 $P\Delta_{20}$. That asymmetry is the quantitative form of the argument in Sec. 4.3: a squared-error objective keeps the overall shift roughly right but loses which genes are differentially expressed. Dropping the calibration term is nearly free in accuracy (indeed $+0.002$ on Norman $P\Delta_{20}$, the one cell where the full model is not best) and triples ECE, so calibration is a cheap purchase but not a free one.

A.9 Test-time scaling

Table 12: **Test-time scaling of the REFINE controller** on Norman (held-out combinations); Figure 2b plots the first and third columns. $N=1$ reduces exactly to the "w/o REFINE" ablation row of Table 11. Latency is per perturbation on one A100.

Budget N	$P\Delta_{20}\uparrow$	E-dist$\downarrow$	Halluc.$\downarrow$	ECE$\downarrow$	Latency (s)$\downarrow$	vs. STATE
STATE (ref.)	0.531 $\pm$0.008	12.43 $\pm$0.14	–	–	0.38	1.0×
1	0.560 $\pm$0.010	11.38 $\pm$0.19	0.251 $\pm$0.010	0.089 $\pm$0.007	0.41	1.1×
2	0.568 $\pm$0.010	11.12 $\pm$0.18	0.228 $\pm$0.009	0.080 $\pm$0.006	0.73	1.9×
4	0.575 $\pm$0.010	10.91 $\pm$0.18	0.206 $\pm$0.008	0.072 $\pm$0.005	1.36	3.6×
8	0.581 $\pm$0.010	10.72 $\pm$0.18	0.183 $\pm$0.007	0.064 $\pm$0.004	2.62	6.9×
16	**0.585** $\pm$0.009	<u>10.61</u> $\pm$0.17	<u>0.171</u> $\pm$0.006	**0.060** $\pm$0.004	5.14	13.5×
32	<u>0.583</u> $\pm$0.009	**10.58** $\pm$0.17	**0.166** $\pm$0.006	<u>0.061</u> $\pm$0.004	10.21	26.9×

A.10 Zero-shot cellular-context generalization

Panel (a) of Table 5 reports this study; the seed spread is the informative detail. CELLWRIGHT is less stable than STATE on the far transfer (±0.014 PΔ against ±0.012), which we attribute to the architecture: on an unseen cell type the GRPO-trained reasoner proposes graphs outside the distribution the verifier was fit on, so its energies, and therefore which candidate REFINE selects, vary more across seeds. A verifier trained with context-held-out negatives would be the natural fix, and we did not do it.

A.11 Compositional multi-gene prediction

Two results in panel (b) of Table 5 cut against a simple story. The Additive baseline retains the best genetic-interaction precision in the $2/2$-seen regime (0.658 vs. our 0.641): when both genes are well observed, most held-out pairs fall in the additive subtype and a learned residual only adds variance. And GEARS, which depends on a gene–gene knowledge graph, is beaten by Additive as soon as one gene is unseen (0.487 vs. 0.503) and badly beaten when neither is (0.331 vs. 0.402), which suggests its graph prior transfers less well than the additivity prior. Our own $0/2$-seen spread (±0.021 $P\Delta_{20}$) is our widest $P\Delta_{20}$ spread in that panel, as expected from $n=19$ held-out pairs; scGPT's is comparable (±0.022), so we do not read our spread as a distinctive instability.

A.12 Regulatory sub-tasks

Panel (c) of Table 5 reports these results. Coupling the sequence trunk with mechanism reasoning helps most where sequence-only models are weakest: unseen cell-type transfer improves by $+0.061$ r over Borzoi ($+13.6\%$ relative), consistent with the claim that a mechanism graph supplies *trans*-acting context a *cis*-only trunk cannot see. It does *not* help on TF binding, where Borzoi remains best (0.921 vs. our 0.914): binding is dominated by local sequence motifs, a trans-context prior adds nothing, and conditioning slightly dilutes the trunk's signal. We regard this negative result as supporting rather than undermining the interpretation, since a mechanism prior that improved motif-level binding would be difficult to explain.

A.13 WHAT THE VERIFIER DOES AND DOES NOT ESTABLISH

Three properties bound the interpretation of every verifier-based number in this paper. *The verifier is trained on synthetic corruptions*, so it learns to separate a coherent (mechanism, outcome) pair from sign-shuffled, target-swapped and outcome-mismatched pairs; coherence with the observed outcome is not the same as biological correctness, and a mechanism that is wrong in a way no corruption family represents is not guaranteed to receive high energy. *The reward and one metric share a grounding index*: R_{ground} and the hallucination rate both consult the same curated resources, so the absolute hallucination level should be compared across methods within this paper and not across papers with different indexes. *Search against a frozen verifier is exploitable*: the non-monotonicity at $N{=}32$ in Table 12, where $\text{P}\Delta_{20}$ and ECE dip within seed spread while hallucination keeps improving, is consistent with this, and it places a practical ceiling on test-time scaling that an adaptive or ensembled verifier would need to raise.

The constructive reading is that a learned verifier is sufficient to make mechanistic reasoning trainable without mechanism labels, which is the claim we make, and insufficient to certify a mechanism as true, which is a claim we do not make.

A.14 CASE STUDY

We follow one held-out Norman pair, CBL + CNN1, through the whole pipeline. The pair comes from the $0/2$-seen bucket, so neither gene was seen in training, and it is one of the synergistic erythroid pairs reported in the original screen (Norman et al., 2019). We chose it because its interaction is strongly non-additive, not because CELLWRIGHT does well on it. Its episode-level $\text{P}\Delta_{20}$ (0.547) ranks 6th of the 19 pairs in the bucket, above the bucket mean of 0.482 (Table 5).

With the default budget $N{=}8$, the REFINE beam includes three kinds of candidate. Two share an erythroid core (CBL $\rightarrow$ GATA1/KLF1 $\rightarrow$ LCR $\rightarrow$ HBG1, HBZ): m_1 adds one edge from CNN1 into this module, and m_2 leaves CNN1 unassigned. A third candidate, m_4, proposes a myeloid SPI1/CEBPA program with inverted erythroid signs. The verifier assigns energies of -6.8, -6.5 and $+2.3$, and this order matches accuracy: 0.547, 0.553 and 0.304 $\text{P}\Delta_{20}$. m_4 is rejected through its per-edge terms, since three of its five edges carry inverted signs, which is the error that the sign-shuffle negatives target. The selected graph also captures the synergy. The two single perturbations raise HBG1 by 0.9 and 0.4 log-fold change, so the additive expectation is 1.3, against a measured pair effect of 2.9. CELLWRIGHT predicts 2.6, while the variant without r_ξ predicts 1.4 and loses 0.116 $\text{P}\Delta_{20}$ on this episode.

The case also shows the limit of the verifier. REFINE returns m_1, yet m_2 is slightly more accurate. The only difference between them is the CNN1 edge, which is absent from the grounding index and is therefore counted as a hallucination. The verifier scores this edge near-neutral (-0.3, exactly the energy gap between the two graphs) because it barely changes coherence with the outcome, and an extra plausible but unsupported edge belongs to none of the corruption families. On this edge, then, the verifier certifies coherence but not truth (Appendix A.13). The more cautious graph m_2 would be the more useful output for choosing the next experiment.