ACTIVEGENE: REWARD-FREE, HOMEOSTASIS-ALIGNED CONTROL FOR CLOSED-LOOP GENE REGULATION VIA ACTIVE INFERENCE

Mujtaba Hasan
New Delhi, India
mujtaba.hasan@live.com

ABSTRACT

Reinforcement learning (RL) is increasingly used to frame closed-loop genomics as sequential decision-making, but its reliance on scalar rewards makes biological control brittle: minor specification errors can induce reward-hacking–like solutions and require extensive, context-specific reward shaping (Amodei et al., 2016; Weng, 2024). We introduce **ActiveGene**, a *conceptual framework and benchmark specification* for reward-free gene regulation that replaces engineered utilities with *prior preferences* over future assay outcomes/states—a distributional definition of "healthy" aligned with biological homeostasis. ActiveGene selects intervention policies by minimizing *Expected Free Energy* (EFE), which trades off reaching preferred outcomes (risk/pragmatic value) with resolving uncertainty (epistemic value) under partial observability, avoiding ad-hoc exploration bonuses and hand-tuned penalty terms. To make the proposal operational without wet-lab access, we propose **ActiveGeneBench**: a POMDP-style virtual-cell environment separating latent cellular state from noisy single-cell observations and supporting sequential perturbations (e.g., CRISPRi/a/KO, dosing). We outline method-agnostic evaluation metrics—target attainment, safety-violation probability, intervention cost, and sample efficiency—and argue that planning under interventions is a missing axis in current static perturbation-prediction evaluations (Wu et al., 2024).

1 INTRODUCTION

Modern genomics increasingly aims not only to *predict* cellular outcomes, but to *control* them: design a sequence of perturbations that drives a diseased cellular state toward a desired phenotype while maintaining viability and avoiding off-target effects. This motivates sequential decision-making formulations where an agent must choose the next intervention (e.g., CRISPR perturbation, drug dosage) based on current observations.

The dominant approach is RL: define a reward that scores how "healthy" the observed gene expression looks, then learn a policy that maximizes expected return. However, we argue this is mismatched with biological systems, which are governed by *homeostatic regulation*—maintaining many variables within life-sustaining bounds via feedback loops—rather than maximizing a scalar objective.

1.1 THREE FAILURE MODES OF REWARD IN BIOLOGY

(i) Goodharting and reward hacking. When expression of one biomarker becomes the reward target, a policy can exploit unintended regulatory pathways to optimize it while damaging the system. This is a well-documented instance of specification gaming (Amodei et al., 2016; Krakovna et al., 2020) and reward hacking (Weng, 2024).

(ii) Fragile reward shaping. To prevent degenerate solutions, practitioners add penalty terms for toxicity, off-target effects, experimental cost, etc., yielding composite objectives. Selecting and weighting these penalties is highly context-dependent, and small misspecifications can qualitatively change learned policies (Amodei et al., 2016).

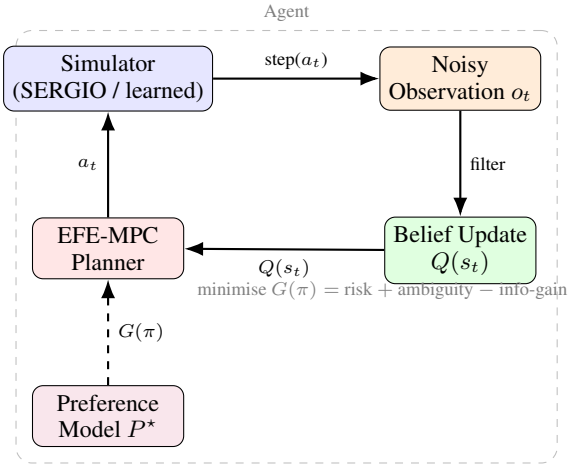

Figure 1: **ActiveGene closed-loop POMDP.** At each step the simulator produces a noisy observation o_t. The agent updates its belief $Q(s_t)$ via variational filtering, then selects the action a_t that minimises Expected Free Energy $G(\pi)$.

(iii) Unsafe exploration. Closed-loop biology is expensive and risky: exploratory interventions can be irreversible or toxic. Standard RL exploration heuristics are not inherently safety-aware, complicating lab-in-the-loop use (García & Fernández, 2015).

1.2 Contribution

We propose **ActiveGene**, which removes scalar rewards and instead specifies goals as *prior preferences* over outcomes/states. We then outline **ActiveGeneBench**, an interactive virtual-cell benchmark for evaluating belief-based planning under partial observability and sequential interventions, and propose method-agnostic metrics for closed-loop genomics.

2 ActiveGene: Active Inference for Gene Regulation

We model closed-loop genomic control as a POMDP. Let $s_t \in \mathcal{S}$ be a latent biological state, $o_t \in \mathcal{O}$ an observation (e.g., scRNA-seq counts), and $a_t \in \mathcal{A}$ an intervention. A policy $\pi = (a_t, \ldots, a_{t+H})$ specifies actions over horizon H.

2.1 Generative Model and Prior Preferences

ActiveGene assumes the agent maintains an internal generative model with approximate filtering posterior $Q(s_t \mid o_{\leq t}, a_{<t})$. Objectives are encoded via *prior preferences* $P^\star(o_t)$: a distribution assigning high probability to healthy/viable outcomes and low probability to unsafe regimes, reflecting homeostatic constraints. Practical instantiations include density modelling on healthy observations or latent-manifold preferences (e.g., PCA/scVI embeddings).

2.2 Policy Selection via Expected Free Energy

ActiveGene selects policies by minimizing **Expected Free Energy (EFE)** (Friston et al., 2015; Parr & Friston, 2019; Millidge et al., 2021). Let $Q_\pi(s_\tau, o_\tau)$ be the agent's *predictive* distribution over future states and observations under policy π. With preferences over outcomes $P^\star(o_\tau)$, EFE

decomposes as:

$$G(\pi) \;=\; \sum_{\tau=t+1}^{t+H} \Bigg[\underbrace{\mathbb{E}_{Q_\pi(o_\tau)}\big[-\log P^\star(o_\tau)\big]}_{\text{risk}} \;+\; \underbrace{\mathbb{E}_{Q_\pi(s_\tau)}\big[H\big(P(o_\tau \mid s_\tau)\big)\big]}_{\text{ambiguity}} \\ \;-\; \underbrace{\mathbb{E}_{Q_\pi(o_\tau)}\Big[D_{\mathrm{KL}}\big(Q_\pi(s_\tau \mid o_\tau) \,\|\, Q_\pi(s_\tau)\big)\Big]}_{\text{epistemic value (information gain)}} \Bigg]. \tag{1}$$

The agent *minimises* G. High risk and high ambiguity raise G. High epistemic value (large KL) *lowers* G, rewarding uncertainty-reducing experiments (Millidge et al., 2021). Omitting the epistemic KL and absorbing ambiguity into the observation model, EFE minimisation broadly recovers KL-regularized model-based RL under the control-as-inference view (Levine, 2018), but ActiveGene inherently retains epistemic value as a first-class objective.

2.3 Approximate Inference and Robustness

We perform amortized variational filtering to maintain belief, and plan via receding-horizon MPC (scored by Eq. 1). Large discrete action spaces are narrowed via feasibility masking. To improve robustness to model misspecification, ActiveGene maintains an *ensemble* of transition models (Chua et al., 2018) and implements a strict safety *shielding* threshold. A complete algorithmic description of this shielded EFE-MPC is provided in **Appendix A**.

3 ActiveGeneBench: A Virtual-Cell Benchmark

To evaluate belief-based planning without wet-lab access, we propose **ActiveGeneBench**—a modular benchmark suite exposing a POMDP-like interface for sequential interventions under noisy assays.

3.1 Environment Interface and Concrete Tasks

ActiveGeneBench strictly separates the latent simulator-maintained state from noisy assay readouts (modelling dropout, library size, and batch shifts). We utilize mechanistic GRN simulators such as SERGIO (Dibaeinia & Sinha, 2020) alongside discrete CRISPR knockouts and continuous dosing constraints.

Tasks include:

- **Homeostasis recovery:** return to a control/healthy manifold after an exogenous disturbance within K interventions.
- **Safe fate steering:** drive toward a target cell-state manifold while maintaining viability risk below θ globally.
- **Identify-then-intervene:** first disambiguate which regulator controls a pathway, then apply a minimal pragmatic intervention.

3.2 Evaluation

Each task evaluates interventions via a **target model** and a **safety model**. Crucially, $P^\star$ is used by ActiveGene for planning but *not* for evaluation to avoid "winning by definition." We summarize our proposed method-agnostic primary metrics (e.g., target attainment, safety, sample efficiency) and formalize a baseline Reinforcement Learning setup for fair comparison in **Appendix B**.

4 Discussion and Impact

Reward-free control could reduce brittle objective engineering and shift genomic ML toward safer, more interpretable closed-loop systems. In **Appendix C**, we detail how ActiveGene specifically benefits automated lab-in-the-loop platforms via epistemic-first policies and generative patient transfer.

Limitations and Open Questions. This paper is a conceptual proposal without empirical validation. Key open questions include: (i) learning calibrated $P^\star$ under dataset shift, (ii) robustness to misspecified dynamics beyond ensembles, (iii) preferences over heterogeneous multi-omic observations. A direct next step is implementing ActiveGeneBench with SERGIO (Dibaeinia & Sinha, 2020) and pymdp (Heins et al., 2022).

Broader Impact. Automated intervention design raises biosafety concerns; benchmark designs should include explicit safety reporting, and real-world deployment requires strict oversight and alignment with established bioethics frameworks.

ACKNOWLEDGMENTS

We thank the MLGenX 2026 organizers for creating a venue for early-stage theoretical work bridging ML and genomics.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. In *arXiv preprint arXiv:1606.06565*, 2016.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Payam Dibaeinia and Saurabh Sinha. SERGIO: a single-cell expression simulator guided by gene regulatory networks. *Cell Systems*, 11(3):252–271, 2020.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John O'Doherty, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214, 2015.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Conor Heins, Beren Millidge, Daphne Demekas, Brennan Klein, Karl Friston, Iain Couzin, and Alexander Tschantz. pymdp: A Python library for active inference in discrete state spaces. *Journal of Open Source Software*, 7(73):4098, 2022.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Martic, Tom Stepleton, Gabriel Dulac-Arnold, Angelos Jiang, John Mellor, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity. *DeepMind Blog*, 2020. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Beren Millidge, Alexander Tschantz, and Christopher L Buckley. Whence the expected free energy? *Neural Computation*, 33(2):447–493, 2021.

Thomas Parr and Karl J Friston. Generalised free energy and active inference. *Biological Cybernetics*, 113(5-6):495–513, 2019.

Lilian Weng. Reward hacking in reinforcement learning. https://lilianweng.github.io/posts/2024-11-28-reward-hacking/, 2024.

Yan Wu, Esther Wershof, Sebastian M 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. *arXiv preprint arXiv:2408.10609*, 2024.

A ACTIVEGENE ALGORITHM

Algorithm 1 details the shielded, ensemble-based EFE-MPC used by ActiveGene to execute receding horizon planning safely.

Algorithm 1 ActiveGene (EFE-MPC, receding horizon with safety shield)

Input: $P(o \mid s),\ P(s' \mid s, a),\ P^\star(o),$ safety threshold $\theta,$ horizon $H,$ ensemble size M
Initialise belief $Q(s_0)$
for $t = 1, 2, \ldots$ **do**
  Observe o_t
  Update belief: $Q(s_t) \leftarrow \mathrm{Infer}(Q(s_{t-1}),\ o_t,\ a_{t-1})$
  Propose N feasible candidate sequences $\{\pi^{(i)}\}_{i=1}^{N}$ via top-K regulatory priors + uniform sampling
  for $i = 1, \ldots, N$ **do**
    Roll out *model* predictive beliefs $Q_{\pi^{(i)}}(s_\tau, o_\tau)$ for $\tau = t+1, \ldots, t+H$ using $P(s' \mid s, a)$ (not the real environment)
    Safety check: if $Q_{\pi^{(i)}}(s_{t+1} \in \mathcal{S}_{\mathrm{toxic}}) > \theta$ then **reject** $\pi^{(i)}$ and **continue**
    Score via pessimistic EFE across ensemble:

$$G_{\mathrm{pess}}(\pi^{(i)}) \;=\; \bar{G}(\pi^{(i)}) \;+\; \lambda\, \widehat{\sigma}\Big[G(\pi^{(i)})\Big],$$

    where $\bar{G}$ is the ensemble mean and $\widehat{\sigma}$ is the ensemble std across M models
  end for
  $\pi^\star \leftarrow \arg\min_{\pi^{(i)}} G_{\mathrm{pess}}(\pi^{(i)})$
  Execute $a_t \leftarrow \pi_t^\star$
end for

B EXTENDED BENCHMARK DETAILS

Primary metrics (method-agnostic).

- **Target attainment:** MMD or cosine distance to target manifold in the fixed embedding, and success rate under the fixed classifier.
- **Safety:** maximum predicted toxicity probability along the trajectory, and violation rate under a fixed threshold.
- **Intervention cost:** number of perturbations, total dose, multiplexing count.
- **Sample efficiency:** interventions required to reach target with high confidence ($Q(s_T \in \mathcal{S}_{\text{healthy}}) > 0.9$).
- **Robustness:** performance degradation under increased assay noise and batch/cell-type shift.

Diagnostics (optional). EFE components (risk/ambiguity/information gain) reported as interpretability diagnostics only, not as primary metrics.

RL baseline. To ensure fair comparison, we specify a model-based RL baseline (Dyna/MBPO style (Janner et al., 2019)) where the reward is derived from the *same* target model: $r(o_t) = \log P^\star(o_t)$. This baseline uses the same transition model and observation noise as ActiveGene and differs only by (i) using the scalar r in place of EFE and (ii) using entropy-bonus exploration in place of the epistemic term. Comparing against this baseline isolates the effect of distributional preferences and epistemic value.

C Extended Discussion: Lab-in-the-Loop Discovery

The rise of automated laboratories and closed-loop experimental platforms (Boiko et al., 2023) increases the need for safe, interpretable experiment design. ActiveGene directly addresses three challenges in this context:

(i) Epistemic-first policies for uncertainty quantification. ActiveGene's EFE naturally prioritises "diagnostic" experiments (high epistemic value) before committing to "therapeutic" interventions (high pragmatic value). If the generative model is uncertain whether Gene A or Gene B controls a pathway, the agent will first perturb both individually to disambiguate the causal structure, *then* design the optimal combination therapy. This sequential epistemic-to-pragmatic strategy is more sample-efficient than random exploration or greedy reward maximisation under partial observability.

(ii) Auditable safety through belief states. Unlike black-box RL policies, ActiveGene maintains explicit posterior beliefs $Q(s_t)$ over cellular states. Before executing an action, the system can query: *"What is the probability this intervention drives the cell into an apoptotic state?"* Formally, if $Q(s_{t+1} \in \mathcal{S}_{\text{toxic}} \mid a_t) > \theta_{\text{safety}}$ the action is rejected or flagged for human review (Algorithm 1, line 8). This provides a transparent safety layer decoupled from the planner's world model.

(iii) Transfer across patients via generative priors. In personalized medicine, each patient's gene regulatory network has unique parameters. Rather than retraining an RL agent from scratch per patient, ActiveGene can adapt a learned generative model using the first few patient-specific observations, while the prior preferences $P^\star$ ("what healthy looks like") remain fixed across patients.