THE VALUE OF MECHANISTIC PRIORS IN SEQUENTIAL DECISION MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Hybrid mechanistic models, physical priors with learned residuals, promise to reduce the data required for good decisions, but have no computable criterion to test this. We characterize the value of mechanistic priors in sequential decision-making within both asymptotic and burn-in regimes. To formalize this, we introduce the *mechanistic information* of a model: the mutual information between the model's recommended policy $\hat{\pi}$ and the true optimal policy π^*, bounded via a centered, occupancy-weighted bias. In the *asymptotic regime* (large N), matched bounds reveal that Bayesian regret scales with the residual entropy H_{mech}, delivering a theoretical sample complexity reduction of $H(\mu)/H_{\mathrm{mech}}$ compared to an uninformed baseline. We further provide a computable pre-trial model certificate. Complementarily, in the clinically relevant *burn-in regime* (small N), we establish a lower bound on the penalty incurred by confidently wrong priors. We demonstrate both the asymptotic and burn-in bounds on an illustrative in-silico 5-fluorouracil (5-FU) chemotherapy plant whose structure follows published FOLFOX pharmacokinetics. The hybrid prior reduces cumulative regret by $1.79\times$ relative to standard body-surface-area dosing and $1.85\times$ relative to an uninformed learner, and remains below both under every calibration bias tested. Finally, we show that priors sensitive to distribution shift can lose half of their mechanistic information under a small distribution shift, motivating physically grounded priors for safety-critical applications.

1 INTRODUCTION

Hybrid mechanistic models, known dynamics paired with learned residuals, are pervasive in scientific machine learning and increasingly central to adaptive medicine. Universal differential equations (Rackauckas et al., 2020), physics-informed neural networks (Karniadakis et al., 2021), grey-box pharmacokinetic models, and the digital twins underlying in silico clinical trials (Eadie et al., 2026) all share one structural premise: physical structure reduces the data required to reach good decisions. The premise is taken for granted, yet no computable, pre-trial criterion tells a decision-maker whether a specific hybrid model will save samples on a specific sequential task. That criterion is the goal of this paper.

A main obstacle is that the quantity routinely reported for a hybrid model, its predictive error on held-out observations, is not the quantity that governs its value for decisions. Predictive accuracy is neither necessary nor sufficient for choosing well: a model may fit accurately and still order two adjacent actions incorrectly, while a model that is badly wrong where no candidate policy operates, or wrong by an offset shared by every action, may still recommend the optimal one. What governs the value of a model is the error it commits where the decision is actually made, measured in a form that can be converted into interactions saved.

Information-theoretic bounds linking prior information to sample complexity exist for other online settings. Russo & Van Roy (2016) show that the Bayesian regret of Thompson sampling on a K-armed bandit scales with the entropy of the optimal arm, and Shufaro et al. (2025) extend this to a regret–information trade-off that quantifies the value of R nats of side information. Both treat the prior as exogenous: R nats are assumed to be available, and the analysis prices them. In a hybrid model the prior is constructed rather than assumed, and the nats it carries are fixed by the mechanistic model's structure, by the data used to calibrate it, and by the error it retains.

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.

Mechanistic models are treated here as priors for sequential learners, and their worth is measured by the information they carry about the optimal policy. We introduce the *mechanistic information* of a model, the number of nats it reveals before any interaction takes place, and establish that the regret still to be incurred is governed by what the model leaves unrevealed. Lower and upper bounds on this regret are derived. Two regimes are covered: the asymptotic one, where the prior's worth appears as a reduction in sample complexity, and the finite-horizon one, where a confident but incorrect prior imposes a cost that must be paid before it can be overturned, the situation that arises whenever a population model is applied to an individual. From this follows a threshold on model error that can be evaluated before experimentation begins, allowing a practitioner to determine in advance whether a candidate mechanistic model is accurate enough to be worth deploying within a hybrid learner. Figure 1 summarizes the setting, and adaptive 5-fluorouracil (5-FU) dosing serves as an illustration throughout.

1.1 Contributions

Mechanistic information and critical bias We define *mechanistic information* $R_{\mathrm{mech}} = I(\pi^*; \hat{\pi})$, the mutual information between the model's recommended policy $\hat{\pi}$ and the optimal policy π^*, and bound it through an occupancy-weighted bias B_μ that charges the model only for errors that affect decision-making. From this bound, we derive the *critical bias* B_μ^{crit}: a closed-form quantity whose inputs are the population spread of the measured parameter and the prior's entropy, both of which can be obtained from reported population-pharmacokinetic studies. This bias gives a *necessary* condition, checkable before any interaction, for a candidate model to reduce sample complexity by a target factor s.

Matched regret bounds Theorems 1 and 2 apply the bounds of Shufaro et al. (2025) and Russo & Van Roy (2016) to the *residual entropy* $H_{\mathrm{mech}} = H(\mu) - R_{\mathrm{mech}}$, giving matching Bayesian-regret bounds (tight up to $\sqrt{\log K}$) and a sample-complexity reduction of $H(\mu)/H_{\mathrm{mech}}$ over an uninformed baseline; the new ingredients are the measurement-channel bound of Proposition 1, the centered occupancy-weighted bias B_μ, the rule-out certificate of Theorem 3, and the burn-in bound of Proposition 2. These bounds allow B_μ^{crit} to turn from an exact information-level threshold to a sample-count target.

Burn-in regime We lower-bound the regret suffered when the prior is concentrated but potentially wrong, the regime of a population-calibrated model deployed for an individual patient. We also show that priors fitted to a population distribution, can lose most of their mechanistic information under small distribution shifts, motivating auditable, physically grounded priors in safety-critical settings.

Paper organization. Section 2 reduces the problem to a policy bandit and introduces mechanistic information; Section 3 proves the asymptotic results and the model-quality certificate; Section 4 treats misspecified priors; Section 5 is the clinical instantiation and simulation.

2 Setting

This section formalizes the reduction from the continuous control problem to a K-armed policy bandit and introduces the central quantity R_{mech}. This frames our problem as repeated selection from a discretized set of dose policies: each round is one treatment cycle and each arm is one dose level held for that cycle.

2.1 Reduction to policy bandit

The mechanistic model $\mathcal{M}$ defines a distribution over the dynamics of the continuous control problem. Given a control signal $u(t)$ and a state trajectory $x(t)$, the cumulative reward is $J(u; \mathcal{M}) = \int_0^T \ell(x(t), u(t))\, dt$. The true unknown dynamics $\mathcal{M}^*$ are drawn from a prior ρ over models. The continuous optimum is $\pi^{\mathrm{opt}} = \arg\max_{u \in \mathcal{U}} J(u; \mathcal{M}^*)$. We discretize the control space into K representative policies, $\Pi = \{\pi_1, \ldots, \pi_K\}$ (e.g. via a finite-difference grid on the infusion rate), reducing our control problem to a K-MAB one. $\pi^* := \arg\max_{\pi \in \Pi} J(\pi; \mathcal{M}^*)$ is the best discrete policy, which is the object the bandit learns. Each policy π_k has mean reward $\bar{r}(\pi_k) = J(\pi_k; \mathcal{M}^*)$, At round t the learner plays $\pi_t \in \Pi$ and observes $r_t = \bar{r}(\pi_t) + \xi_t$, where ξ_t is normally distributed,

with zero-mean and standard deviation $\sigma_\xi > 0$. We make the following assumption on the quality of the discretization.

Assumption 1 (Discretization quality) *The policy π^{opt} belongs to Π, or the best policy in Π has sub-optimality gap at most $\Delta_r/2$ relative to the continuous optimum,*

$$J(\pi^{\text{opt}}; \mathcal{M}^*) - J(\pi^*; \mathcal{M}^*) \leq \frac{\Delta_r}{2}. \tag{1}$$

Before interaction starts, an observation is drawn $Z = h(M^*) + \nu$ where $\nu \sim \mathcal{N}(0, \sigma_\nu^2 I)$ independent of M^*, $\dim Z = d_Z$ and $h : \text{supp}(\rho) \to \mathbb{R}^{d_Z}$ is an invertible map. The calibrated model is the model consistent with the measurement, $\hat{\mathcal{M}} = h^{-1}(Z)$, where h^{-1} is the inverse on decision-relevant equivalence classes guaranteed by Assumption 3. The learner then plays under $\hat{\mathcal{M}}$, which recommends the best policy in Π under the approximate dynamics.

$$\hat{\pi} \;=\; \arg\max_{\pi \in \Pi} J(\pi; \hat{\mathcal{M}}). \tag{2}$$

This policy is computable from the model $\hat{\mathcal{M}}$ alone, without knowing the true dynamics. We remark that $\hat{\pi}$ is not the policy for which the model is most accurate everywhere, but rather it is the policy that performs best under the model's predictions. Even an inaccurate model can produce a useful $\hat{\pi}$ if its prediction errors cancel in the reward integral. We denote by μ the marginal distribution of π^*, when $\mathcal{M}^* \sim \rho$. We write $H(\mu)$ for the Shannon entropy of the prior μ over Π. All logarithms are natural (nats) unless explicitly labeled $\log_2$ (bits). We use the Bayesian regret, which is the mean gap between the optimal and chosen rewards over the randomness present in the prior, π^*, and the algorithm. This metric measures the prior's worth in terms of interactions saved.

Definition 1 (Bayesian regret)

$$\text{BR}^*(N) = \mathbb{E}_{\pi^* \sim \mu} \mathbb{E}\left[\sum_{t=1}^{N} \big(\bar{r}(\pi^*) - \bar{r}(\pi_t) \big) \right].$$

We write $\sigma_h^2 := \text{tr}\, \text{Cov}_{M^* \sim \rho}(h(M^*))$ for the population spread of the measured quantity.

2.2 Occupancy-weighted sensitivity

Assumption 2 (Metric model space) *The model space is metric, with the centered occupancy-weighted reward norm*

$$\|\mathcal{M} - \mathcal{M}'\|_\mu^2 = \sum_{k=1}^{K} \mu(\pi_k) \Big(\big(J(\pi_k; \mathcal{M}) - \bar{J}(\mathcal{M}) \big) - \big(J(\pi_k; \mathcal{M}') - \bar{J}(\mathcal{M}') \big) \Big)^2$$

when $\bar{J}(\mathcal{M}) = \sum_k \mu(\pi_k) J(\pi_k; \mathcal{M})$.

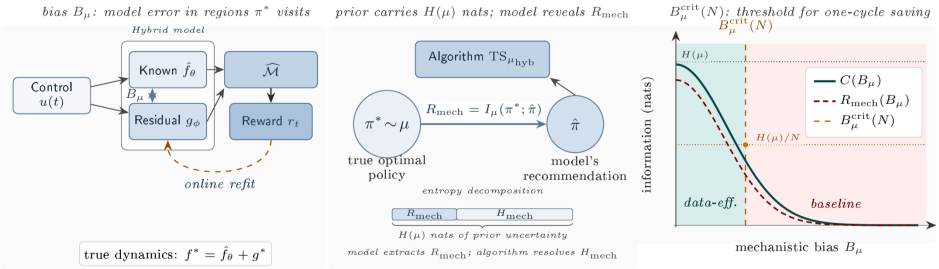

Figure 1: Three layers of the hybrid model learning setting. **(A)** Mechanistic model. **(B)** Information flow from π^* to $\hat{\pi}$ and to the environment. **(C)** Regret-bias diagram: switching regimes at B_μ^{crit}.

Assumption 3 (Identifiability of the decision-relevant model) *There exists $L_h > 0$ such that for all $\mathcal{M}, \mathcal{M}'$ in the support of ρ,*

$$\sqrt{d_Z} L_h \left\| \mathcal{M} - \mathcal{M}' \right\|_\mu \;\leq\; \left\| h(\mathcal{M}) - h(\mathcal{M}') \right\|_2 ,$$

where $\|\cdot\|_\mu$ is the norm of Assumption 2.

Appendix D.1 computes L_h for the 5-FU model; notation is collected in Appendix A. The occupancy weighting charges the model only for errors in regions π^* actually visits. Equation (3) is the expectation of this squared norm between the true and the calibrated model. Concretely,

$$B_\mu^2 = \mathbb{E}_{M^* \sim \rho,\, \nu} \left\| \mathcal{M}^* - h^{-1}(Z) \right\|_\mu^2 \tag{3}$$

This bias is the expected centered, occupancy-weighted squared reward gap between the true and the calibrated model; a common offset of the reward vector scores zero because it does not change the recommendation. A model that is inaccurate in corners of the state space that the policies in Π never reach has a small B_μ, while a model that is slightly wrong exactly where they operate has a large B_μ and transfers almost none. Throughout this paper, we will denote $l_\mu = B_\mu L_h$.

Throughout this paper, we use a running example, for illustration, based on adaptive dosing of 5-fluorouracil (5-FU) in colorectal cancer patients receiving FOLFOX.

Running Example 1 (Adaptive 5-FU dosing) We instantiate the true dynamics $\mathcal{M}^*$ with the population pharmacokinetic model of Tan et al. (2026), a two-compartment ODE with Michaelis–Menten elimination whose state comprises the 5-FU amounts in the central and peripheral compartments. The deployed mechanistic model $\hat{\mathcal{M}}$ is a one-compartment linear-clearance *surrogate* of the Tan model, built from the pre-trial measurement Z of Section 2.1, a single plasma concentration drawn during a body-surface-area–dosed lead-in cycle. The policy class $\Pi = \{\pi_1, \ldots, \pi_8\}$ ($K = 8$) represents eight dose levels spanning 1,600–3,600 mg/m^2, the clinical adjustment range in the Kaldate et al. (2012) dataset. Reward $r_t \in \{0, 1\}$ indicates whether the realised 46-hour AUC falls within the therapeutic window 20–30 mg·h/L.

2.3 Mechanistic Information

Definition 2 (Mechanistic information)

$$R_{\text{mech}} \;=\; I_\mu(\pi^*;\, \hat{\pi}),$$

the mutual information between $\pi^* \sim \mu$ and the mechanistic recommendation $\hat{\pi}$ defined by (2).

R_{mech} measures the number of nats about π^* that the model reveals before a single interaction is performed. When $B_\mu = 0$ the model is exact ($\hat{\pi} = \pi^*$, $R_{\text{mech}} = H(\mu)$); as $B_\mu \to \infty$ the recommendation becomes independent of π^* and $R_{\text{mech}} \to 0$. The residual entropy $H_{\text{mech}} = H(\mu) - R_{\text{mech}}$ quantifies the information the algorithm must discover by interacting with the system. We now introduce an upper bound on the mechanistic information, equivalently a lower bound on the residual entropy.

Proposition 1 (Mechanistic information bounds) *Under Assumptions 2-3,*

$$R_{\text{mech}} \;\leq\; C(B_\mu) \;:=\; \min\Big\{ H(\mu),\; \frac{d_Z}{2} \log\Big(1 + \frac{\sigma_h^2}{L_h^2 B_\mu^2}\Big)\Big\}. \tag{4}$$

The proof is provided in Appendix D. The mechanistic model essentially acts as a noisy Gaussian channel. The model bias contributes noise variance l_μ^2, so the SNR, and with it the information the model can transmit, decreases as B_μ grows. Appendix H.4 provides further details regarding this Proposition.

Running Example 2 (Mechanistic information for 5-FU) The certificate inputs are two constants *derived* from published population-PK statistics (Tan et al. (2026), Appendix H): the between-patient spread of the elimination parameter, $\sigma_h = \sqrt{0.10}$, and the precision of its pre-dose estimate from a three-sample limited-sampling scheme. We get that (see Appendix H)

$$C(B_\mu) = 0.77.$$

Appendix H provides full details regarding this calculation and the parameter's calculation.

3 ASYMPTOTIC REGIME

3.1 LOWER AND UPPER BOUNDS

We now lower-bound the Bayesian regret via the residual entropy H_{mech} and prove a matching upper bound up to a $\sqrt{\log K}$ factor.

Theorem 1 (Hybrid regret lower bound) *Consider the hybrid setting with K policies and N rounds. Under Assumptions 2–3, there exist universal constants $c, N_0 > 0$ such that for any algorithm there exists a prior μ over Π such that for all $N \geq N_0$:*

$$\mathrm{BR}^*(N) \;\geq\; c\sqrt{\frac{KN\,H_{\mathrm{mech}}}{\log K}}, \tag{5}$$

where $H_{\mathrm{mech}} = H(\mu) - R_{\mathrm{mech}} \geq 0$.

Theorem 1 quantifies the regret the model saves: the R_{mech} nats it "spends" before interaction reduce the effective problem from $H(\mu)$ to H_{mech} nats of uncertainty. Thus, the ratio $\sqrt{H(\mu)/H_{\mathrm{mech}}}$ quantifies the reduction in the regret that is obtained by following the model's recommendation. The proof of Theorem 1 follows directly from Proposition 4.1 of Shufaro et al. (2025), and is provided in Appendix E.

Theorem 2 (Thompson Sampling upper bound) *Let μ_{hyb} be the posterior of μ after conditioning on $\hat{\pi}$. By construction, $I(\pi^*; \hat{\pi}) = R_{\mathrm{mech}}$. Let $\mathrm{TS}_{\mathrm{hyb}}$ be Thompson Sampling with this prior. Then there exists a universal constant $C > 0$ such that:*

$$\mathrm{BR}_{\mathrm{TS}_{\mathrm{hyb}}}(N) \;\leq\; C\sqrt{KN\,H_{\mathrm{mech}}}, \tag{6}$$

Theorem 2 shows that Thompson Sampling with μ_{hyb} achieves the lower bound: when the prior is concentrated on $\hat{\pi}$, the algorithm's first prediction identifies $\hat{\pi}$ with probability $\mu_{\mathrm{hyb}}(\hat{\pi})$ (close to 1 when $R_{\mathrm{mech}} \approx \log K$), and subsequent rounds refine or overturn this guess. The proof of Theorem 2 follows from (Russo & Van Roy, 2016) and can be found in Appendix E.

We can define the sample complexity algorithm as the average number of rounds required to achieve mean regret of less than ε, so

$$N(\varepsilon) = \min_{N>0:\, \frac{1}{N}\mathcal{BR}(N) \leq \varepsilon} N$$

Combining Equations (5) and (6), the number of rounds needed to reach mean regret ε, the sample complexity, obeys the asymptotic ratio:

$$\rho(\varepsilon) \;=\; \frac{N_{\text{uninf}}(\varepsilon)}{N_{\text{mech}}(\varepsilon)} \;=\; \Theta\left(\frac{H(\mu)}{H_{\mathrm{mech}}}\right). \tag{7}$$

where the Θ notation omits factors of $\Theta(\log K)$.

Running Example 3 (Lower and upper bounds for 5-FU) We apply Theorems 1 and 2 at the channel ceiling of Running Example 2: a prior that attains its channel ceiling carries $R_{\mathrm{mech}} = 0.77$ nats, i.e. $H_{\mathrm{mech}} = 1.31$ nats. With $K = 8$ and $N = 12$ cycles, $7.77c \leq \mathrm{BR}^*(12) \leq 11.20C$; the ratio $11.20/7.77 = 1.44 = \sqrt{\ln 8}$ is exactly the $\sqrt{\log K}$ slack between the two bounds. The sample-complexity ratio (Eq. (7)) is certified at $\rho = H(\mu)/H_{\mathrm{mech}} = 2.08/1.31 \approx 1.59$ (up to $\log K$ factors): an uninformed learner needs up to $1.59\times$ more mis-dosed cycles to reach the same regret (a ratio of residual entropies; the samples reading inherits the constants of Eq. (7)).

3.2 CRITICAL BIAS

The previous section demonstrated how one can use information to effectively quantify regret improvement. However, to be more relevant in clinical settings, we need a quantity that is both relevant to the number of trials we realistically reduce, while also being computable without requiring any additional trials. This section aims to achieve that. We now introduce the critical bias, a computable model certificate that governs the model's ability to reduce sample complexity.

Theorem 3 (Critical bias, rule-out form) *Let $s \geq 1$ be a target sample-complexity ratio. For $s > 1$ define the critical bias,*

$$L_h B_\mu^{\mathrm{crit}}(s) = l_\mu^{\mathrm{crit}}(s) = \frac{\sigma_h}{\sqrt{\exp\left(\frac{2}{d_Z}(1-1/s)H(\mu)\right) - 1}}. \tag{8}$$

Then any model with $B_\mu \geq B_\mu^{\mathrm{crit}}(s)$ suffers regret of at least $c\sqrt{KNH(\mu)/(s \log K)}$ for any $N \geq N_0$ (where N_0 is defined in Theorem 1).

Theorem 3 gives a *necessary* condition, checkable before any interaction, for a candidate model to reduce sample complexity by a target factor s: a model with bias above $l_\mu^{\mathrm{crit}}(s)$ cannot, whatever the algorithm. A model below it may, depending on how close R_{mech} comes to the capacity bound. The threshold is a statement about residual decision entropy; reading it as a factor-s saving in samples goes through Eq. (7) and inherits its constants. The one-cycle-of-N reading corresponds to $s = N/(N-1)$. $l_\mu^{\mathrm{crit}}(s)$ is computable before any trial. Locating a model against it needs that model's l_μ (Limitations). The proof of Theorem 3 is provided in Appendix E.

Remark 1 *The choice $s = N/(N-1)$ is illustrative. Since Theorems 1–2 match only up to constants and a $\sqrt{\log K}$ factor, a one-cycle difference lies below the resolution of the bounds. B_μ^{crit} at this point should therefore be read as an information-level screen: a model above it cannot carry enough mechanistic information to meet the target, while passing it is necessary but not sufficient for any saving in cycles.*

Running Example 4 (Model-quality certificate for 5-FU) With the Tan-sourced constants ($\sigma_h = 0.316$, $L_h = 1$, $d_Z = 1$, $H(\mu) = \ln 8$) and $N = 12$, the critical bias (8) at the one-cycle working point $s = N/(N-1) = 12/11$ (i.e. $(1-1/s)H(\mu) = H(\mu)/12$) is $B_\mu^{\mathrm{crit}} = 0.316/\sqrt{8^{1/6} - 1} \approx 0.49$. This is the working point the clinic can act on: can the model be expected to save one cycle out of N?

4 Misspecified priors in the finite-sample regime

Section 3 characterized regret in the asymptotic regime. We now turn to the finite-sample regime relevant to clinical-trial design.

Proposition 2 (Burn-in lower bound) *Let $K \geq 2$, $\Delta \in (0,1)$ and $\delta \in (0,1)$. Let the prior place mass $1-\epsilon$ on $\hat{\pi}$ (confident but wrong) and $\epsilon/(K-1)$ on each other arm, one of which is π^*, so that the prior log-odds are $L_0 = \log\big((1-\epsilon)(K-1)/\epsilon\big)$. There is a Bernoulli bandit with gap Δ such that any algorithm for which $\Pr(\pi^* = j \mid \mathcal{F}_t) \geq \Pr(\pi^* = k \mid \mathcal{F}_t)$ implies $\Pr(A_{t+1} = j \mid \mathcal{F}_t) \geq \Pr(A_{t+1} = k \mid \mathcal{F}_t)$ then τ_δ satisfies*

$$\mathbb{E}\big[\textit{regret up to } \tau_\delta\big] \;\geq\; \frac{(1-\delta)\,L_0 + 2\log\frac{1}{2.4\delta} - 1}{4\,\log\frac{1+\Delta}{1-\Delta}}. \tag{9}$$

Proposition 2 quantifies the minimum exploration cost that a confidently wrong prior forces. In particular, it grows with the prior log-odds $L_0 = \log[(1-\epsilon)(K-1)/\epsilon]$ and with the identification requirement through $\log(1/(2.4\delta))$, and for small Δ it scales as $1/\Delta$ (the standard rate, since $\log\frac{1+\Delta}{1-\Delta} \approx 2\Delta$). The cost grows logarithmically as the prior becomes more confident ($\epsilon \to 0$) or as the identification confidence requirement tightens ($\delta \to 0$). The proof of Proposition 2 can be found in Appendix F.

Running Example 5 (Burn-in for a miscalibrated 5-FU prior) Using the $K = 8$ setting: the prior places mass $1 - \epsilon = 0.8$ on $\hat{\pi}$ (a confident but, for this patient, wrong recommendation), so $L_0 = \ln\big(0.8 \times 7/0.2\big) = \ln 28 \approx 3.33$; the sub-optimality gap is $\Delta = 0.2$ (approximately one dose step) and the identification confidence is $\delta = 0.01$. Substituting into (9),

$$\text{Burn-in} \;\geq\; \frac{0.99 \times 3.33 + 2\ln\frac{1}{0.024} - 1}{4\,\ln(1.2/0.8)}. \approx 6.02,$$

which is equivalent to ≈ 30 plays of a Δ-suboptimal arm (relaxing δ to 0.10 lowers the bound to ≈ 15 suboptimal plays). The cost of a confidently wrong prior is computable in closed form before

any interaction, from the prior weight, the dose-step gap, and the required identification confidence alone.

The following proposition demonstrates how misspecified priors limit mechanistic information.

Proposition 3 (Misspecified prior information bounds) *Let some policy produce recommendations with mechanistic information $R^{\text{train}}_{\text{mech}}$ on training distribution P_{train}, and assume the recommender is symmetric: $H(\hat{\pi} \mid \pi^* = k)$ is equal for all k (for example, a symmetric confusion channel). Let the deployment distribution P_{test} have KL shift $\Delta_\pi = D_{\text{KL}}(P_{\text{train}} \| P_{\text{test}})$. Denote by $R^{\text{test}}_{\text{mech}}$ the mechanistic information of the recommendations under the test distribution. Then there exist train and test distributions with $\Delta_\pi = \frac{1}{2} R^{\text{train}}_{\text{mech}}$, with uniform $\hat{\pi}$ for which $R^{\text{test}}_{\text{mech}} \leq \frac{1}{2} R^{\text{train}}_{\text{mech}} + \frac{1}{2} \log K$.*

The proof of Proposition 3 is provided in Appendix F. A small shift, undetectable from training data alone, can thus significantly degrade performance — a patient who reacts very differently from what the training literature describes. The mechanistic prior is not immune to shift, since a population model without DPD status as an input is as wrong for a DPD-deficient patient as any population-trained prior, which is the situation Proposition 2 describes. Its advantage is auditability. σ_h and B_μ can be re-estimated on any cohort, so the bound of Proposition 1 can be re-certified after a shift rather than assumed.

Running Example 6 (Misspecified priors for 5-FU) A distribution calibrated on a portion of FOLFOX population data might have $R^{\text{train}}_{\text{mech}} = 1.6$ nats. The calibration data might suffer a shift of $\Delta_\pi \approx 0.5$ nats (rough estimate). By Proposition 3, test distributions with $\Delta_\pi \leq \frac{1}{2} \log K \approx 1.04$ nats exist for which $R^{\text{test}}_{\text{mech}} \leq \frac{1}{2} \log K$. When $P_{\hat{\pi}}$ is uniform, a shift of $\Delta_\pi = \frac{1}{2} R^{\text{train}}_{\text{mech}}$ already halves the information, and a shift of this size can be undetectable from training data alone. ODE priors are often physically bounded, allowing them to provide information while remaining robust to distribution shifts. Calibrated LLM priors, on the other hand, might overfit and fail on parts of the population while appearing to provide more information.

Running Example 7 (End-to-end, in silico) We close the loop on a simulated plant: the world is the two-compartment model of Tan et al. (2026). The deployed prior is the one-compartment surrogate of Running Example 1, fitted on a population slice and inverted once on a lead-in blood draw. Measured on this plant, the draw takes R_{mech} from 0.000 to 0.42 nats, below its certified ceiling $C(B_\mu) = 0.77$, and Thompson sampling with the resulting prior, the algorithm of Theorem 2 observing rewards alone, cuts cumulative regret by $1.8\times$ against both BSA dosing and uninformed TS. Its regret is also comparatively stable: across calibration slices biased in either direction, it degrades by about a tenth at the harshest slice (1.66 to 1.83) while staying below both model-free references, whereas the per-cycle ODE controller varies more than five-fold and ends worse than standard-of-care dosing (Table 7). Appendix Hdescribes the full protocol and constants.

5 Application and Simulation

5.1 Clinical context: the 5-FU individualisation problem

5-Fluorouracil (5-FU) is the backbone of FOLFOX chemotherapy for colorectal cancer, yet only roughly twenty percent of body-surface-area (BSA)-dosed patients reach the therapeutic exposure window (Li et al., 2023; Saif et al., 2009). We instantiate the framework on literature-calibrated 5-FU pharmacokinetic parameters. [1]

Literature-calibrated parameters We calibrate the bandit parameters to published population PK data. The policy class Π is $K = 8$ discrete dose levels spanning the clinical adjustment range, 1,600–3,600 mg/m^2 (Kaldate et al., 2012). Arms above a patient's DPD-adjusted ceiling, known before cycle 1, would be removed from Π first in a clinical instantiation, a gate we do not model. Rewards are normalized: $r_t = 1$ if the AUC falls within 20–30 mg·h/L, $r_t = 0$ otherwise, with Bernoulli noise. The reward is symmetric in under- and over-exposure, a simplification, since clinically overexposure drives grade 3/4 toxicity. Under standard BSA dosing, the target attainment

[1] Code for the experiments is available here: `https://anonymous.4open.science/r/The-Value-of-Mechanistic-Priors-in-Sequential-Decision-Making-F629/`

probability is $p_{\mathrm{BSA}} \approx 0.20$ (Li et al., 2023) and since r_t is Bernoulli, $\sigma_\xi = \sqrt{p(1-p)} \approx 0.40$. The optimal (PK-guided) policy achieves $p_{\mathrm{opt}} \approx 0.85$, an idealized *per-cycle* attainment at the individually optimal dose, where PK-guided *titration* reached the range in 94% of patients only cumulatively (Gamelin et al., 2008). Appendix H sweeps these choices. The certificate inputs and their ceiling ($C \approx 0.77$ nats, $H_{\mathrm{mech}} \geq 1.31$) are those of Running Example 2, and every operating point below respects that ceiling.

5.2 Simulation results

The plant is the two-compartment Michaelis–Menten model of Tan et al. (2026), its per-patient parameters drawn from an illustrative population whose magnitudes are rounded modeling choices rather than estimates of any cohort (Appendix H.2). The deployed model is the one-compartment surrogate of Running Example 1, fitted on a 2,000-patient calibration study and inverted once on a lead-in blood draw. Evaluation uses a disjoint cohort of $n = 10{,}000$, $K = 8$ arms, one unscored lead-in cycle then 11 scored cycles, all policies on the same patients. The hybrid implements Theorem 2 by estimating the posterior. Its prior $\mu_{\mathrm{hyb}} = P(\pi^* \mid \hat{\pi})$ is frozen after the draw, and the binary reward is its only later input (Appendix H).

We compare TS with the hybrid prior ($\mathrm{TS}_{\mathrm{hyb}}$) against three different policies: uninformed TS ($\mathrm{TS}_{\mathrm{uninf}}$), a static BSA baseline, and a policy that sticks with the ODE recommendation (ODE only). We use the first reward as the initial observation (Z). The ODE-only policy uses this information to calibrate its parameters but does not use later observations. Table 1 presents the comparison between these four policies for the calibrated values previously described. For more details, see Appendix H. We also run a full parameter sweep in Appendix H.8. We see that the mechanistic information provides TS a nearly $2\times$ improvement over the uninformed baseline, which is more than the predicted improvement of ≈ 1.12 (using the lower bound ratio $\sqrt{H(\mu)/H_{\mathrm{mech}}}$.

Table 1: The calibrated plant, $n = 10{,}000$, seeds 11/2024: cumulative regret over the 11 scored cycles (mean $\pm$ 95% CI), the comparator/hybrid ratio of means with paired-bootstrap intervals, what each policy reads in a scored cycle and how often it solves the ODE. The lead-in draw is taken from everyone. Per-cycle controller: Table 7.

policy	regret	vs. hybrid	lead-in draw	per-cycle	solves
uninformed TS	3.07 ± 0.02	$1.85\times$ [1.82, 1.88]	not read	binary	0
BSA static	2.97 ± 0.04	$1.79\times$ [1.76, 1.83]	not read	none	0
ODE, one solve (prior only)	1.83 ± 0.04	$1.11\times$ [1.08, 1.13]	read	none	1
hybrid (Thm. 2: TS, ODE prior)	1.66 ± 0.02	$1.00\times$	read	binary	1

Table 2 generates R_{mech} scenarios from the calibrated model using the pre-dose channel and assay precision σ_{assay}, not from model bias. At $R_{\mathrm{mech}} = 0$ the model recommends one arm to everyone, so a learner ignoring the draw is uninformed TS.

Table 2: Mechanistic information generated by the *channel* at the calibrated model ($\gamma = 1$, 11 scored cycles, $n = 10{,}000$), floor $= \sqrt{H(\mu)/H_{\mathrm{mech}}}$, "ODE, once" the one-solve policy of Table 1, uninformed TS 3.07 and BSA 2.96 constant throughout.

σ_{assay}	R_{mech}	$C(B_\mu)$	B_μ	floor	$\mathrm{TS}_{\mathrm{hyb}}$	ODE, once	$\frac{\mathrm{ODEonce}}{\mathrm{TS}_{\mathrm{hyb}}}$	$\frac{\mathrm{TS}_{\mathrm{uninf}}}{\mathrm{TS}_{\mathrm{hyb}}}$	$\frac{\mathrm{BSA}}{\mathrm{TS}_{\mathrm{hyb}}}$
0.50	0.179	0.772	0.268	1.05	1.99 ± 0.02	2.80 ± 0.05	1.41	1.54 [1.53, 1.56]	1.50 [1.47, 1.53]
0.22	0.417	0.929	0.221	1.12	1.66 ± 0.02	1.83 ± 0.04	1.11	1.85 [1.82, 1.88]	1.79 [1.76, 1.83]
0.02	0.631	1.127	0.176	1.20	1.36 ± 0.02	1.20 ± 0.03	0.89	2.26 [2.23, 2.30]	2.19 [2.14, 2.24]

Table 3 re-fits the surrogate's population clearance on slices of the calibration study, moving the deployed model's bias and with it R_{mech}. The hybrid's prior table is fitted on the same slice the deployed model is calibrated on, so the study and the deployment share one population model. Regret is again monotone in R_{mech} and the hybrid stays below both model-free references everywhere, which neither the one solve nor the per-cycle controller does. The floor moves only over 1.08–1.12 while the measured advantage moves over 1.68–1.85: the margin over the floor, not the floor itself, is what the bias costs (nine slices: Table 7, 33: Figure 3). The advantage also grows with the horizon, BSA/hybrid $1.33\times$ at $N = 3$ to $3.45\times$ at $N = 50$, while the floor is horizon-free (Table 9).

Table 3: Regret as the deployed model's calibration is biased (γ = calibrated ÷ true clearance, 11 scored cycles, $n = 10{,}000$): three of the nine canonical slices, floor $= \sqrt{H(\mu)/H_{\mathrm{mech}}}$, uninformed TS 3.068 and BSA 2.975 constant at every slice. All nine slices, with the per-cycle controller: Table 7.

calibration slice	γ	B_μ	R_{mech}	floor	$\mathrm{TS}_{\mathrm{hyb}}$	ODE, once	$\frac{\text{once}}{\mathrm{TS}_{\mathrm{hyb}}}$	$\frac{\mathrm{TS}_{\mathrm{uninf}}}{\mathrm{TS}_{\mathrm{hyb}}}$	$\frac{\mathrm{BSA}}{\mathrm{TS}_{\mathrm{hyb}}}$
slowest-clearing 20%	0.593	0.304	0.306	1.08	1.83 ± 0.02	3.60 ± 0.06	1.97	1.68	1.63
everyone (correct model)	1.000	0.221	0.417	1.12	1.66 ± 0.02	1.83 ± 0.04	1.11	1.85	1.79
fastest-clearing 20%	1.673	0.303	0.303	1.08	1.80 ± 0.02	3.37 ± 0.05	1.87	1.70	1.65

6 Related Work

Regret-information framework. Shufaro et al. (2025) trade off external information against Bayesian regret: R nats of prior information replace $H(\mu)$ by $H(\mu) - R$ in the rate $c\sqrt{KNH(\mu)/\log K}$, demonstrated for LLM priors. We extend this to hybrid mechanistic models: Proposition 1's measurement channel converts the model's bias B_μ into information, via L_h and σ_h. Appendix B situates this work against side-information bandits, hybrid and model-based RL, GP bandits and misspecified Bayesian bandits.

Adaptive 5-FU dosing. PK-guided adaptive 5-FU dosing (Gamelin et al., 2008) and circadian chemotherapy (Lévi et al., 2010) demonstrate clinical benefit. Ernst et al. (2005) introduce tree-based fitted Q iteration and Shortreed et al. (2011) apply RL to sequential treatment decisions; neither addresses continuous dose individualization. We provide the first information-theoretic sample-complexity certificate for such schemes. No policy here is evaluated against observed patient outcomes (Gottesman et al., 2019) (see Appendix B for more details).

7 Discussion

We introduce *mechanistic information* R_{mech} as a pre-trial measure of prior quality for mechanistic-derived priors in Bayesian adaptive control, and from it the *critical bias*, a pre-trial gate on a model's ability to reduce sample complexity. Extending Shufaro et al. (2025) to mechanistic models, we prove regret bounds tight up to $\sqrt{\log K}$ that link R_{mech}, B_μ^{crit} and sample complexity, and quantify the burn-in-regime gain. On a calibrated in-silico 5-FU plant the hybrid prior cuts regret by $1.79\times$ against body-surface-area dosing and $1.85\times$ against an uninformed learner, stays below both under every calibration bias tested, and improves monotonically in the information the pre-dose channel supplies. B_μ^{crit} turns "this model captures the right biology" into a quantitative gate. Finally, priors sensitive to training-distribution shift, LLM priors among them, can lose half their mechanistic information under a shift of $\frac{1}{2}\log K$ nats undetectable from training data alone, an argument for physically-informed priors in safety-critical adaptive control.

Limitations Theorems 1-2 cover only the finite-K problem; a continuous-action extension would close the gap to the original control problem. Our 5-FU evaluation runs on an in-silico plant calibrated to published PK statistics (Tan et al., 2026; Kaldate et al., 2012; Li et al., 2023), not on patient-level outcomes, and B_μ in (3) needs per-arm rewards no clinical dataset contains, so outside simulation it needs a reference model to supply them. The bandit treats arm means as stationary within a patient, which the reported $\sim 20\%$ cycle-to-cycle variability, cumulative toxicity and dose delays violate, so a nonstationary formulation, a toxicity stopping rule and DPD-based arm gating are future work, and the clinical claims require a prospective trial.

Declaration of AI usage. In this work, we used generative AI tools for designing or providing feedback on research methodology or experiments, finding related work, drafting sections of the paper, and implementing methods. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, or propose or refine hypotheses. The rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for drafting parts of the paper, creating or editing software code, and identifying relevant literature. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39.1–39.26. JMLR Workshop and Conference Proceedings, 2012.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

M. G. Azar, R. Munos, and H. J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 844–853. JMLR.org, 2017.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

Ashley L. Eadie, Holly Fernandez Lynch, Naomi Scheinerman, and Ravi B. Parikh. The arrival of digital twins and in silico trials in drug development. *Nature Medicine*, 32(6):1967–1971, 2026.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

L. Fang, W. Xin, H. Ding, et al. Pharmacokinetically guided algorithm of 5-fluorouracil dosing: a meta-analysis. *Scientific Reports*, 6:25913, 2016.

Erick Gamelin, Remy Delva, Jacques Jacob, Yacine Merrouche, Jean Luc Raoul, Denis Pezet, Etienne Dorval, Gilles Piot, Alain Morel, and Michele Boisdron-Celle. Individual fluorouracil dose adjustment based on pharmacokinetic follow-up compared with conventional dosage: Results of a multicenter randomized trial of patients with metastatic colorectal cancer. *Journal of Clinical Oncology*, 26(13):2099–2105, 2008. doi: 10.1200/JCO.2007.13.3934. URL https://ascopubs.org/doi/abs/10.1200/JCO.2007.13.3934. PMID: 18445839.

Divyansh Garg, Skanda Vaidyanath, Kuno Kim, Jiaming Song, and Stefano Ermon. LISA: Learning interpretable skill abstractions from language. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=XZhipvOUBB.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 100–108, Bejing, China, 2014. PMLR. URL https://proceedings.mlr.press/v32/gopalan14.html.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, 2019. doi: 10.1038/s41591-018-0310-5.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT 2010 - The 23rd Conference on Learning Theory*, pp. 67–79, 01 2010.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor,

Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as i can, not as i say: Grounding language in robotic affordances. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=bdHkMjBJG_w.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pp. 2137–2143. PMLR, 2020.

Rajesh R. Kaldate, A. Haregewoin, C. E. Grier, S. A. Hamilton, and Howard L. McLeod. Modeling the 5-fluorouracil area under the curve versus dose relationship to develop a pharmacokinetic dosing algorithm for colorectal cancer patients receiving FOLFOX6. *The Oncologist*, 17(3): 296–302, 2012.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Shinji Kobuchi and Yukako Ito. Application of pharmacometrics of 5-fluorouracil to personalized medicine: A tool for predicting pharmacokinetic–pharmacodynamic/toxicodynamic responses. *Anticancer Research*, 40(12):6585–6597, 2020. doi: 10.21873/anticanres.14683.

Branislav Kveton, Mikhail Konobeev, Manzil Zaheer, Chih-Wei Hsu, Martin Mladenov, Craig Boutilier, and Csaba Szepesvari. Meta-thompson sampling. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5884–5893. PMLR, 2021. URL https://proceedings.mlr.press/v139/kveton21a.html.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. ISSN 0196-8858. doi: https://doi.org/10.1016/0196-8858(85)90002-8. URL https://www.sciencedirect.com/science/article/pii/0196885885900028.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Alessandro Lazaric and Mohammad Ghavamzadeh. Bayesian multi-task reinforcement learning. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 599–606, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Francis A. Lévi, A. Okyar, S. Dulong, P. F. Innominato, and J. Clairambault. Circadian timing in cancer treatments. *Annual Review of Pharmacology and Toxicology*, 50:377–421, 2010.

M. Li, S. Mindt, A. Lück, U. Hutzschenreuter, M. Kollendt, et al. Drug monitoring detects under- and overdosing in patients receiving 5-fluorouracil-containing chemotherapy: results of a prospective, multicenter German observational study. *ESMO Open*, 8(2):101201, 2023.

Che-Yu Liu and Lihong Li. On the prior sensitivity of Thompson sampling. In *Proceedings of the 27th International Conference on Algorithmic Learning Theory (ALT)*, pp. 321–336. Springer, 2016. doi: 10.1007/978-3-319-46379-7_22.

Katarzyna Morawska, Françoise Goirand, Laurine Marceau, Madeline Devaux, Adèle Cueff, Aurélie Bertaut, Julie Vincent, Leila Bengrine-Lefevre, François Ghiringhelli, and Antonin Schmitt. 5-FU therapeutic drug monitoring as a valuable option to reduce toxicity in patients with gastrointestinal cancer. *Oncotarget*, 9(14):11559–11571, 2018.

Felix Mueller, Barbara Büchel, Dieter Köberle, Sarah Schuürch, Barbara Pfister, Stephan Krähenbühl, Tanja K Froehlich, Carlo R Largiader, and Markus Joerger. Gender-specific elimination of continuous-infusional 5-fluorouracil in patients with gastrointestinal malignancies: results from a prospective population pharmacokinetic study. *Cancer Chemotherapy and Pharmacology*, 71(2): 361–370, 2013.

Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*, 2020.

M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2018.10.045. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/24541007.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–99, 2018.

M. Wasif Saif, A. Choma, S. J. Salamone, and E. Chu. Pharmacokinetically guided dose adjustment of 5-fluorouracil: a rational approach to improving therapeutic outcomes. *JNCI: Journal of the National Cancer Institute*, 101(22):1543–1552, 2009.

Susan M. Shortreed, Eric Laber, Daniel J. Lizotte, T. Scott Stroup, Joelle Pineau, and Susan A. Murphy. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Mach. Learn.*, 84(1–2):109–136, July 2011. ISSN 0885-6125. doi: 10.1007/s10994-010-5229-0. URL https://doi.org/10.1007/s10994-010-5229-0.

Itai Shufaro, Nadav Merlis, Nir Weinberger, and Shie Mannor. On bits and bandits: Quantifying the regret-information trade-off. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0oWGVvC6oq.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 1015–1022, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Zhiyuan Tan et al. Advancing precision dosing of 5-fu: Population pk model development, limited sampling strategies, and fit-for-use application. *Clinical Pharmacokinetics*, 2026. doi: 10.1007/s40262-026-01645-1.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 12 1933. ISSN 0006-3444. doi: 10.1093/biomet/25.3-4.285. URL https://doi.org/10.1093/biomet/25.3-4.285.

Anna D Wagner, Axel Grothey, Thierry Andre, Jesse G Dixon, Norman Wolmark, Daniel G Haller, Carmen J Allegra, Aimery De Gramont, Eric VanCutsem, Steven R Alberts, et al. Sex and adverse events of adjuvant chemotherapy in colon cancer: an analysis of 34 640 patients in the accent database. *JNCI: Journal of the National Cancer Institute*, 113(4):400–407, 2021.

Yuan Yin, Vincent LE GUEN, Jérémie DONA, Emmanuel de Bezenac, Ibrahim Ayed, Nicolas THOME, and patrick gallinari. Augmenting physical models with deep networks for complex dynamics forecasting. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=kmG8vRXTFv.

A NOTATION AND CONVENTIONS

The following table summarises the symbols used throughout the paper, in order of first appearance. Full formal definitions are given in the corresponding sections.

Table 4: Symbols used throughout the paper, in order of first appearance.

Symbol	Meaning
Problem primitives (§2)	
K	Number of policies (bandit arms); e.g. $K = 8$ discrete dose levels
N	Number of interaction rounds (e.g. chemotherapy cycles)
T	Horizon of the continuous-time control problem
$x(t) \in \mathbb{R}^n$	State trajectory
$u(\cdot) \in \mathcal{U}$	Continuous control input ($\mathcal{U}$ compact)
$\ell(x, u)$	Instantaneous (stage) reward
$J(u; \mathcal{M})$	Cumulative reward $\int_0^T \ell(x(t), u(t))\, dt$ under dynamics $\mathcal{M}$
$\mathcal{M}$	Mechanistic model (a distribution over dynamics)
M^*	Unknown true dynamics, drawn from the prior ρ
ρ	Prior over dynamics models
$\hat{\mathcal{M}}$	Calibrated (deployed) mechanistic model, $\hat{\mathcal{M}} = h^{-1}(Z)$
$r_t = \bar{r}(\pi_t) + \xi_t$	Observed reward at round t
$\bar{r}(\pi) = J(\pi; M^*)$	Mean reward of policy π
ξ_t, σ_ξ	Zero-mean reward observation noise and its standard deviation
Δ_r	Discretisation sub-optimality gap (Assumption 1)
The pre-trial measurement (§2.1)	
$Z = h(M^*) + \nu$	Pre-trial measurement (e.g. one lead-in blood draw)
$h(\cdot)$	Measurement map from models to $\mathbb{R}^{d_Z}$
$\nu \sim \mathcal{N}(0, \sigma_\nu^2 I)$	Measurement noise, independent of M^*
σ_ν	Measurement noise standard deviation per dimension
d_Z	Measurement dimension
$\sigma_h^2 = \operatorname{tr} \operatorname{Cov}_{M^* \sim \rho}(h(M^*))$	Population spread of the noise-free measurement
L_h	Identifiability constant of Assumption 3
$h_{\min}$, $g_{\max}$	Scalar-case constants in the L_h appendix ($L_h = h_{\min}/g_{\max}$)
$J_h(\theta)$	Measurement Jacobian (H is reserved for entropy)
Policies and priors (§2.1–2.2)	
$\Pi = \{\pi_1, \ldots, \pi_K\}$	Finite policy class
π^{opt}	Continuous-control optimum $\arg\max_u J(u; M^*)$
π^*	Best discrete policy (unknown; $\pi^* \sim \mu$)
$\hat{\pi}$	Model-recommended policy: $\arg\max_{\pi \in \Pi} J(\pi; \hat{\mathcal{M}})$
μ	Prior on Π (occupancy of the optimal policy)
$H(\mu)$	Shannon entropy of μ (in nats)
μ_{hyb}	Hybrid prior on Π (posterior of μ given $\hat{\pi}$; Theorem 2)
$\mu(\pi_k)$	Occupancy weight of arm k (the single weight symbol of Eq. 3)
$\lVert\cdot\rVert_\mu$	Centred, occupancy-weighted reward norm of Eq. 3
B_μ	Centred occupancy-weighted mechanistic bias of $\hat{\mathcal{M}}$ vs M^*
Information quantities (§2.3, §3)	
$I(X; Y)$	Mutual information
$D_{\text{KL}}(P \Vert Q)$	Kullback–Leibler divergence
$\text{kl}(p, q)$	Binary KL divergence
R_{mech}	Mechanistic information $I_\mu(\pi^*; \hat{\pi})$ (Definition 2)
$H_{\text{mech}} = H(\mu) - R_{\text{mech}}$	Residual entropy
C_ν	Estimator-free channel bound on R_{mech} (Proposition 1)
$C(B_\mu)$	Bias-form channel bound on R_{mech} (Proposition 1)
$s \geq 1$	Target sample-complexity ratio in Theorem 3

Symbol	Meaning
$B_\mu^{\mathrm{crit}}(s)$	Critical bias at target s (Theorem 3); one-cycle point $s = N/(N-1)$
Finite-sample / misspecification (§4)	
ϵ	Prior weight off the recommended arm in the burn-in setting
δ	Identification confidence threshold (δ-PAC)
$L_0 = \log \frac{(1-\epsilon)(K-1)}{\epsilon}$	Prior log-odds of the recommended arm
Δ	Sub-optimality gap in the burn-in construction
Λ_t	Posterior log-likelihood ratio after t rounds (Proposition 2 proof)
τ_δ	δ-PAC stopping time
$R_{\mathrm{mech}}^{\mathrm{train}}$, $R_{\mathrm{mech}}^{\mathrm{test}}$	Mechanistic information on train / test distribution
$P_{\mathrm{train}}, P_{\mathrm{test}}$	Joint distributions over $(\pi^*, \hat{\pi})$ on train / test
$\Delta_\pi = D_{\mathrm{KL}}(P_{\mathrm{train}} \Vert P_{\mathrm{test}})$	Distribution shift (Proposition 3)
Performance and constants	
$\mathrm{BR}^*(N)$	Bayesian regret over N rounds (Definition 1)
$\rho_{\mathrm{sc}} = H(\mu)/H_{\mathrm{mech}}$	Asymptotic sample-complexity ratio (Eq. 7)
ε	Mean-regret tolerance in the sample-complexity definition
c, C, N_0	Universal constants in regret lower / upper bounds
$\Theta(\cdot), O(\cdot), o(\cdot)$	Standard asymptotic notation
Clinical / experimental parameters (§5, calibration appendix)	
n	Number of in-silico patients (evaluation / calibration-study cohorts)
p_{BSA}	Target attainment probability under BSA dosing (≈ 0.20)
p_{opt}	Target attainment probability under PK-guided dosing (≈ 0.85)
C_{ss}	Steady-state plasma concentration
R^2	Coefficient of determination (Kaldate calibration)
$\Delta\mathrm{AUC}, \Delta\mathrm{dose}$	Increments in plasma AUC and dose level
Clinical acronyms	
5-FU	5-Fluorouracil (cytotoxic chemotherapy agent)
FOLFOX, FOLFOX6	Folinic acid + 5-FU + Oxaliplatin combination regimens
AIO, FUFOX, LV5FU2	Alternative 5-FU-containing regimens
BSA	Body Surface Area (standard dosing convention)
AUC	Area Under the Curve (plasma 5-FU concentration over T)
PK / PD	Pharmacokinetic / Pharmacodynamic
DPD	Dihydropyrimidine Dehydrogenase (clearance enzyme)
DPYD	Gene encoding DPD
UH2/U	Dihydrouracil to uracil ratio (DPD-phenotype marker)
ORR	Objective Response Rate
ODPM	On-Demand PK Monitoring (Capitain et al. adjustment protocol)
CV	Coefficient of Variation (intra-patient AUC CV $\approx 20\%$)
ACCENT	Adjuvant Colon Cancer ENd Points colorectal-trials database
Convention	
$\log, \ln$	Natural logarithm (nats) unless explicitly labelled $\log_2$ (bits)

Note on the symbol H. H denotes Shannon entropy throughout; the measurement Jacobian of the L_h appendix is written J_h, and the Hamiltonian of the optimal-control preliminaries in Appendix D is written $\mathcal{H}$ where it appears; the meaning is unambiguous from context.

B Extended Related Work

Bandits with side information. Prior work uses external information to reduce exploration: Bayesian transfer (Lazaric & Ghavamzadeh, 2010), meta-TS with task-distribution priors (Kve-

ton et al., 2021), and TS in structured action spaces (Gopalan et al., 2014). We derive the prior quantitatively from the mechanistic model's bias rather than assuming an abstract task prior.

Hybrid models and model-based RL. Neural ODEs (Chen et al., 2018), UDEs (Rackauckas et al., 2020), and PINNs (Karniadakis et al., 2021; Raissi et al., 2019) combine ODE structure with learned residuals, targeting expressive fitting rather than decision sample complexity. Yin et al. (2021) address identifiability. Sample-complexity results for model-based RL, LQR (Dean et al., 2020), linear MDPs (Jin et al., 2020), tabular MDPs (Azar et al., 2013), assume uniform model classes without structured physical priors. Our identifiability constant L_h quantifies the inductive bias of a physical model, and R_{mech} gives the resulting information-theoretic gain, orthogonal to both lines of work.

Gaussian process bandits. The measurement channel of Section 2.1 plays the role that the information gain γ_T plays in GP bandits (Srinivas et al., 2010; Chowdhury & Gopalan, 2017): it bounds the information available before interaction. Srinivas et al. (2010) establish $\tilde{O}(\sqrt{NT\gamma_T})$ regret for GP-UCB; Chowdhury & Gopalan (2017) tighten these bounds for kernelised bandits.

Reinforcement learning for treatment, and its evaluation critiques. A parallel literature warns that RL-for-healthcare results are routinely overstated when policies are evaluated off-policy on observational data: Gottesman et al. (2019) catalogue the failure modes (confounding, small effective sample sizes at the recommended actions, reward misspecification). Our results are deliberately upstream of this debate: every number in Section 5.2 is an on-policy simulation from a declared generative model, and no policy here is evaluated against observed patient outcomes. The certificate of Theorem 3 is a statement about the deployed model's bias, not a clinical-effectiveness claim; establishing the latter requires a prospective trial (Limitations).

Misspecified Bayesian bandits. The burn-in proposition (Proposition 2) is related to the literature on bandits with misspecified priors. Honda & Takemura (2010) give an asymptotically optimal algorithm for bandits with bounded-support rewards; Liu & Li (2016) study prior misspecification in Thompson Sampling. Our result differs in giving an explicit *lower bound* on the burn-in cost in terms of the prior weight $1-\epsilon$ and the identification confidence δ, derived from a martingale argument on the posterior log-odds together with the change-of-measure inequality of Kaufmann et al. (2016).

Regret lower bounds for stochastic bandits. The $\Omega(\sqrt{KN})$ minimax lower bound for stochastic bandits is due to Lai & Robbins (1985) and Auer et al. (2002); the Bayesian $\Omega(\sqrt{KNH(\mu)/\log K})$ lower bound follows from Fano's inequality applied to the prior (Shufaro et al., 2025; Lattimore & Szepesvári, 2020). Theorem 1 specializes these to the hybrid setting by replacing $H(\mu)$ with the residual H_{mech}. Russo et al. (2018); Lattimore & Szepesvári (2020) provide comprehensive treatments of information-theoretic lower bounds in bandit settings.

Bayesian bandits with informative priors. Thompson Sampling (Thompson, 1933) exploits the prior to reduce exploration. Russo & Van Roy (2016) and Agrawal & Goyal (2012) establish its near-optimal regret. Russo & Roy (2014) characterizes the value of informative priors in the Bayesian regret framework. Theorem 2 recovers the standard TS bound with H_{mech} replacing $H(\mu)$, showing that the mechanistic ODE acts as a computationally tractable source of prior information.

LLMs as decision-making priors. The LLM proposition (Proposition 3) is motivated by recent work using LLMs as in-context planners (Ichter et al., 2022) and as priors for reinforcement learning (Garg et al., 2022). Shufaro et al. (2025) quantify the regret benefit of LLM information in question answering. Our impossibility result shows that LLM priors (among others) can lose half their mechanistic information from a KL shift of just $\frac{1}{2}\log K$ nats, making them unreliable as the sole prior in safety-critical settings.

C Extended Discussion and Future Directions

ODE priors versus LLM priors. Proposition 3 provides a formal argument for preferring physically-grounded ODE priors over LLM priors in safety-critical settings. An LLM prior with

$R^{\text{train}}_{\text{mech}}$ nats of mechanistic information on its training population can lose half its information from a small KL shift, which is potentially undetectable from training data alone. This distinction matters most in personalized medicine, where the test population (an individual patient) always differs from the training population (a cohort).

The mechanistic information quantity. $R_{\text{mech}} = I(\pi^*; \hat{\pi})$ is a mutual information between two policy-valued variables, so a direct estimate needs samples of π^*: the dose that *would have been* optimal for each patient. Observed (dose, response) pairs do not contain that, since the administered dose is not the optimal one and the responses at the other doses are missing. A standard mutual-information estimator applies only once a reference model has supplied π^* per patient, which is what the in-silico tier of Appendix H does. What is available without such a model is the ceiling C_ν of Proposition 1, which needs only σ_h and σ_ν; this asymmetry is why the paper certifies through the ceiling rather than through a direct estimate of R_{mech}. Running Example 2 illustrates the worst-case channel bound certified from published population constants ($R_{\text{mech}} \leq 0.77$ nats). Running Example 7 measures the realized value on the in-silico plant (0.42 nats against its ceiling $C(B^{\text{crit}}_\mu) = 0.77$). In practice this gives a pre-trial gate, with the same asymmetry: $B^{\text{crit}}_\mu(s)$ follows from σ_h, L_h and $H(\mu)$, which a population-PK study reports, so the threshold itself is available before any trial; but the comparison of a particular model against it needs that model's B_μ, and hence the reference model discussed above. If $B_\mu / B^{\text{crit}}_\mu \ll 1$ the ODE prior is worth using, and if B_μ exceeds the one-cycle working point $B^{\text{crit}}_\mu(s{=}N/(N{-}1))$ the ODE cannot be expected to save even one cycle of the trial and may not justify the complexity of the hybrid model. What routine pre-FOLFOX calibration studies deliver directly is the channel side, σ_h and σ_ν, and therefore C_ν and B^{crit}_μ; the bias side takes the additional modelling step set out in Appendix H.

Additional Open directions. The bound treats R_{mech} as known. Folding the cost of its estimation from calibration data into the regret would give an end-to-end guarantee. When data from multiple patients are available, the effective identifiability constant L_h can be improved by pooling occupancy measures. Formalizing this would extend the framework to population-level adaptive dosing.

D Additional details for Section 2

D.1 The identifiability constant L_h

Beyond the population quantities σ_h, σ_ν and d_Z, the only property of the mechanistic model that enters Proposition 1 and the rule-out half of Theorem 3 is the constant L_h of Assumption 3,

$$L_h \, \|\mathcal{M} - \mathcal{M}'\| \;\leq\; \|h(\mathcal{M}) - h(\mathcal{M}')\| \qquad \text{for all } \mathcal{M}, \mathcal{M}' \in \operatorname{supp} \rho, \tag{10}$$

where $\|\cdot\|_\mu$ is the reward-space norm of (3). This appendix explains what L_h measures, shows that it is computable for parametric models from the sensitivity equations already used in pharmacokinetic calibration, and gives its value for the 5-FU model of Section 5. It also records the case in which the constant is one, which is the form used in the running examples.

What L_h measures. The left-hand side of (10) is a distance between the decisions two models lead to: it vanishes when the two reward vectors agree up to a common offset, so that $\hat{\pi}$ is the same under both, and it is large when the models disagree about which dose is best. The right-hand side is the distance between the two measurements. Inequality (10) therefore states that patients whose optimal decisions differ produce measurements that differ proportionally: the pre-dose measurement resolves whatever the dose decision depends on. It is an identifiability condition on the decision-relevant part of the model. When $L_h \to 0$, the reward vector can change without any change in Z, the measurement carries no information about the decision, and the ceiling of Proposition 1 is vacuous, however precise the assay.

Only the lower constant is needed. The proof of Proposition 1 converts the reward-space precision B_μ into a lower bound on the measurement noise, $\mathbb{E}\left\|h(\hat{\mathcal{M}}) - h(\mathcal{M}^*)\right\|^2 \geq L_h^2 B_\mu^2$, and a lower bound on the noise is what a capacity ceiling requires. An upper Lipschitz constant of h would bound the noise from above and could not enter an upper bound on information.

Invertibility is a consequence. Assumption 3 uses the model recovered from a noiseless measurement, $\hat{\mathcal{M}} = h^{-1}(Z)$. This requires no separate hypothesis: if $h(\mathcal{M}) = h(\mathcal{M}')$ then by (10) $\|\mathcal{M} - \mathcal{M}'\|_\mu = 0$, so the two models have the same centered reward vector and the same recommendation. Thus h is injective on decision-relevant equivalence classes of models, which is the only sense of invertibility the recommendation needs.

Computation for parametric models. Let the model be parametrized by $\theta \in \Theta \subseteq \mathbb{R}^p$, with reward vector $r(\theta) = (J(\pi_1;\theta), \dots, J(\pi_K;\theta))$ and measurement map $h : \Theta \to \mathbb{R}^{d_Z}$. Write $\bar{r}(\theta)$ for the centered, $\sqrt{\mu}$-weighted reward vector, so that $\|\theta - \theta'\|_\mu = \|\bar{r}(\theta) - \bar{r}(\theta')\|_2$. Integrating the forward sensitivity equations $\partial_t(\partial x/\partial\theta) = f_x\,(\partial x/\partial\theta) + f_\theta$ alongside the model under each grid policy, as is routine in pharmacokinetic software, yields the Jacobians

$$G(\theta) = \partial_\theta \bar{r}(\theta) \in \mathbb{R}^{K\times p}, \qquad J_h(\theta) = \partial_\theta h(\theta) \in \mathbb{R}^{d_Z \times p}.$$

(H denotes Shannon entropy throughout; the measurement Jacobian is written J_h.) To first order $\|\theta - \theta'\|_\mu \approx \|G(\theta)(\theta - \theta')\|$ and $|h(\theta) - h(\theta')| \approx |J_h(\theta)(\theta - \theta')|$, so the linearized constant is

$$L_h = \inf_{\theta \in \operatorname{supp}\rho} \inf_{v \neq 0} \frac{|J_h(\theta)\, v|}{\|G(\theta)\, v\|}. \tag{11}$$

The infimum is positive only if $\ker J_h(\theta) \subseteq \ker G(\theta)$ for every θ: every parameter direction that changes the reward vector must also change the measurement, which forces $\operatorname{rank} G(\theta) \le d_Z$. This is the precise sense in which the dimension of the pre-dose measurement limits what a mechanistic model can transfer, and it is why d_Z appears in Proposition 1. When more parameters affect the decision than the measurement resolves, one restricts θ to the identified directions and treats the remaining inter-patient variation as a Z-independent error e, which enters the margin condition and not the ceiling.

For a single patient-specific parameter the constant is exact rather than linearised.

Lemma 1 (Identifiability constant, scalar parameter) *Let $p = 1$ and Θ an interval. Suppose h and r are C^1 on Θ with*

$$|h'(\theta)| \;\ge\; h_{\min} > 0 \qquad \textit{and} \qquad \|\partial_\theta \bar{r}(\theta)\|_2 \;\le\; g_{\max} \qquad \textit{for all } \theta \in \Theta.$$

Then Assumption 3 holds with $L_h = h_{\min}/g_{\max}$.

Proof For $\theta, \theta' \in \Theta$ the mean value theorem gives $|h(\theta) - h(\theta')| \ge h_{\min}|\theta - \theta'|$. Writing $\bar{r}(\theta) - \bar{r}(\theta') = \int_0^1 \partial_\theta \bar{r}\big(\theta' + s(\theta - \theta')\big)\, ds\ (\theta - \theta')$ and applying the triangle inequality under the integral, $\|\theta - \theta'\|_\mu \le g_{\max}|\theta - \theta'|$. Hence $L_h \|\theta - \theta'\|_\mu \le h_{\min}|\theta - \theta'| \le |h(\theta) - h(\theta')|$. ■

Instantiation for 5-FU. The model of Section 5 has a single patient-specific parameter, the log-clearance $\theta = \log \mathrm{CL}$, and the pre-dose measurement is a clearance estimate in log units, $Z = \theta^* + \nu$, so h is the identity and $h_{\min} = 1$. For a linear pharmacokinetic model the cycle exposure at dose D_k is $\mathrm{AUC}_k = D_k e^{-\theta}$, and the mean reward of π_k is the probability that this exposure, multiplied by the intra-patient cycle-to-cycle factor, falls in the therapeutic window; each $J(\pi_k;\theta)$ is a difference of two cumulative distribution functions of the log cycle-to-cycle factor, evaluated at $\log 30 - \log D_k + \theta$ and $\log 20 - \log D_k + \theta$. Its derivative in θ is a difference of two densities, so $g_{\max}$ is bounded by $\sqrt{2}$ times the peak density of the log cycle-to-cycle factor, and is computed exactly by integrating the model over the dose grid for θ in the support of the population prior. By Lemma 1, $L_h = 1/g_{\max}$ in reward units; in parameter units, which the running examples use, the constant is one and the certificate compares the precision of a single clearance estimate with the inter-patient spread of log-clearance.

D.2 Proof of Proposition 1

Proof Since $\hat{\pi}$ is a function of Z and π^* is a function of M^*, the data-processing inequality gives $R_{\mathrm{mech}} = I(\pi^*;\hat{\pi}) \le I(M^*;Z)$. The channel $Z = h(M^*) + \nu$ has additive Gaussian noise $\nu \sim \mathcal{N}(0, \sigma_\nu^2 I)$, so for any input distribution with $\mathbb{E}\|h(M^*)\|^2 = \sigma_h^2$,

$$I(M^*;Z) \;\le\; \frac{d_Z}{2}\log\left(1 + \frac{\sigma_h^2}{\mathbb{E}\|\nu\|^2/d_Z}\right) \;=\; \frac{d_Z}{2}\log\left(1 + \frac{\sigma_h^2}{\sigma_\nu^2}\right)$$

(a Gaussian input maximizes mutual information under a power constraint. The d_Z prefactor uses concavity of the logarithm across coordinates). Substituting $\hat{\mathcal{M}} = h^{-1}(Z)$ we have $h(\hat{\mathcal{M}}) - h(M^*) = \nu$, and Assumption 3 gives $\frac{\mathbb{E}\|\nu\|^2}{d_Z} \geq L_h^2 B_\mu^2$; substituting this lower bound on the noise power yields (4), and since it can only enlarge the ratio, $I(M^*; Z) \leq C(B_\mu)$. We finish the proof by noting that $R_{\text{mech}} \leq H(\mu)$. ■

Remark 2 *We note that along the proof we have used a relaxation of $\mathbb{E}[\|\nu\|^2] \geq L_h^2 B_\mu^2$. One might consider directly substituting σ_ν, avoiding further relaxation. However, σ_ν is a measure of the spread of an auxiliary noise and is not a quantity that is actively observed or can be derived from observables. The model bias, B_μ, on the other hand, can be measured directly from a calibration study, and L_h is a model-dependent constant which can be calculated using h alone.*

Remark 3 *We denote $C_\nu = \frac{d_Z}{2} \log\left(1 + \frac{\sigma_h^2}{\mathbb{E}\|\nu\|^2/d_Z}\right)$. This quantity will be used later on in the rest of the Appendix.*

E Proofs for Section 3

E.1 Proof of Theorem 1

We begin by introducing Proposition 4.1 of Shufaro et al. (2025).

Proposition 4 (Proposition 4.1 (Shufaro et al., 2025)) *For any agent, the regret K-MAB problem is lower bounded by $c\sqrt{KH(\pi^*)T/\log K}$, under the assumption $D_{\text{KL}}(p_t \| P^{\pi^*}) > c' > 0$, where p_t is the distribution of the agent's action at time t and P^{π^*} is the action law that plays π^* with probability one. The condition holds while the algorithm has not yet concentrated on π^*, which is the regime in which the bound is informative; we assume it throughout.*

Proof We begin by stating that the mechanistic recommendation $\hat{\pi}$ is computed before round 1 and revealed to the algorithm as a free side-channel. By the mutual information chain rule,

$$H(\pi^* \mid \hat{\pi}) = H(\mu) - I_\mu(\pi^*; \hat{\pi}) = H(\mu) - R_{\text{mech}} = H_{\text{mech}}.$$

Thus, the algorithm's *posterior* prior on π^* after observing $\hat{\pi}$ has *expected* entropy H_{mech}. The lower bound follows directly by applying Proposition 4. The vacuous case $H_{\text{mech}} \leq 0$ means the ODE uniquely identifies π^*, which results in a policy with 0 regret. ■

E.2 Proof of Theorem 2

Proof By Propositions 1 and 3 of Russo & Van Roy (2016), Thompson Sampling with any prior ν over K arms achieves $\text{BR}^*(N; \text{TS}) \leq C\sqrt{KN\,H(\nu)}$. Substituting $\nu = \mu_{\text{hyb}}$ and $H(\nu) = H(\mu_{\text{hyb}})$ yields

$$\text{BR}^*(N; \text{TS}_{\text{hyb}}) \;\leq\; C\sqrt{KN\,H(\mu_{\text{hyb}})}.$$

The prior μ_{hyb} is the posterior after conditioning on $\hat{\pi}$ and is random. Taking expectations over $\hat{\pi}$ and using concavity of the square root, $\mathbb{E}[\sqrt{H(\mu_{\text{hyb}})}] \leq \sqrt{\mathbb{E}[H(\mu_{\text{hyb}})]} = \sqrt{H_{\text{mech}}}$, since $H_{\text{mech}} = H(\pi^* \mid \hat{\pi})$ is the expected conditional entropy. Substituting gives (6). ■

E.3 Proof of Theorem 3

Proof Fix the convention $s \geq 1$: the target sample-complexity ratio, $H_{\text{mech}} = H(\mu)/s$. A model that reduces sample complexity by the factor s must satisfy $H_{\text{mech}} \leq H(\mu)/s$, i.e. $R_{\text{mech}} \geq \left(1 - \frac{1}{s}\right) H(\mu)$. By the upper bound of Proposition 1, this is impossible whenever

$$\frac{d_Z}{2} \log\left(1 + \frac{\sigma_h^2}{L_h^2 B_\mu^2}\right) \leq \left(1 - \frac{1}{s}\right) H(\mu),$$

i.e. whenever $C(B_\mu) \le (1-1/s)H(\mu)$. Solving for the bias at which equality holds:

$$\log\left(1+\frac{\sigma_h^2}{L_h^2 B_\mu^2}\right) \le \frac{2}{d_Z}\left(1-\frac{1}{s}\right)H(\mu)$$
$$\frac{\sigma_h^2}{L_h^2 B_\mu^2} \le \exp\left(\frac{2}{d_Z}\left(1-\frac{1}{s}\right)H(\mu)\right)-1$$
$$B_\mu \ge \frac{\sigma_h/L_h}{\sqrt{\exp\left(\frac{2}{d_Z}(1-1/s)H(\mu)\right)-1}} := B_\mu^{\mathrm{crit}}(s).$$

Thus $B_\mu \ge B_\mu^{\mathrm{crit}}(s)$ implies $R_{\mathrm{mech}} \le (1-1/s)H(\mu)$ and $H_{\mathrm{mech}} \ge H(\mu)/s$; substituting into Theorem 1 gives $\mathrm{BR}^*(N) \ge c\sqrt{KNH_{\mathrm{mech}}/\log K} \ge c\sqrt{KNH(\mu)/(s\log K)}$, which concludes the proof. ∎

F Proofs for Section 4

F.1 Proof of Proposition 2

Proof We define by $\mathcal{F}_t = \sigma\left(\{r_s, a_s\}_{s=1}^t\right)$ a filtration which contains the entire play history up to time t. Let Λ_t denote the log-likelihood ratio of $\nu_{\hat\pi}$ against ν_{π^*} accumulated over the first t rounds,

$$\Lambda_t = \sum_{s\le t} \log \frac{p_{\nu_{\hat\pi}}(r_s \mid a_s)}{p_{\nu_{\pi^*}}(r_s \mid a_s)}.$$

The two instances differ only on arms $\hat\pi$ and π^*, so only plays of those arms change Λ_t. The action probabilities are common to all instances and cancel in the posterior, as do the prior masses on the remaining instances, so the posterior log-odds between the two are exactly

$$\log \frac{\Pr(\nu_{\hat\pi} \mid \mathcal{F}_t)}{\Pr(\nu_{\pi^*} \mid \mathcal{F}_t)} = L_0 + \Lambda_t. \tag{12}$$

Under ν_{π^*}, a play of $\hat\pi$ yields a reward with mean $\frac{1}{2}-\frac{\Delta}{2}$ and decreases Λ by $d(\Delta) = \mathrm{kl}\big(\frac{1}{2}-\frac{\Delta}{2}, \frac{1}{2}+\frac{\Delta}{2}\big) = \Delta\log\frac{1+\Delta}{1-\Delta}$ in expectation. A play of π^* yields mean $\frac{1}{2}+\frac{\Delta}{2}$ and, by symmetry, also decreases Λ by $d(\Delta)$ in expectation. Writing $T_t := N_{\hat\pi}(t) + N_{\pi^*}(t)$ for the number of contested plays, $M_t := \Lambda_t + d(\Delta)\,T_t$ is therefore a martingale with increments bounded by $\log\frac{1+\Delta}{1-\Delta}$. Optional stopping applies at any stopping time whose contested-play count has finite expectation; when that expectation is infinite, the corresponding bound holds trivially.

We will also need a bound on upward excursions of Λ. Under ν_{π^*}, e^{Λ_t} is the likelihood ratio of $\nu_{\hat\pi}$ against ν_{π^*} on $\mathcal{F}_t$, hence a nonnegative martingale with mean one, and Ville's inequality gives $\Pr(\sup_t \Lambda_t \ge x) \le e^{-x}$ for all $x \ge 0$. Consequently

$$\mathbb{E}\Big[\sup_t \Lambda_t^+\Big] \;\le\; \int_0^\infty e^{-x}\,dx \;=\; 1. \tag{13}$$

Let $\tau' := \inf\{t : L_0 + \Lambda_t \le \log\frac{\delta}{1-\delta}\}$ be the first round at which the posterior odds of $\hat\pi$ against π^* fall to $\delta/(1-\delta)$. On the event that the algorithm outputs π^*, which has probability at least $1-\delta$, its posterior on ν_{π^*} at τ_δ is at least $1-\delta$, so by (12) $\tau' \le \tau_\delta$ and $\Lambda_{\tau'\wedge\tau_\delta} = \Lambda_{\tau'} \le -L_0 - \log\frac{1-\delta}{\delta}$. On the complementary event, $\Lambda_{\tau'\wedge\tau_\delta} \le \sup_t \Lambda_t^+$. Optional stopping at $\tau'\wedge\tau_\delta$ together with (13) gives

$$d(\Delta)\,\mathbb{E}[T_{\tau_\delta}] \;\ge\; d(\Delta)\,\mathbb{E}[T_{\tau'\wedge\tau_\delta}] \;=\; -\mathbb{E}[\Lambda_{\tau'\wedge\tau_\delta}] \;\ge\; (1-\delta)\Big(L_0 + \log\tfrac{1-\delta}{\delta}\Big) - 1.$$

(i), prior term. Let $\tau := \inf\{t : L_0 + \Lambda_t < 0\}$ be the first round at which the posterior prefers π^* to $\hat\pi$. The posterior cannot reach $1-\delta$ on π^* without first preferring it, so $\tau \le \tau_\delta$ on the good event, where $\Lambda_\tau < -L_0$; on the complementary event $\Lambda_{\tau\wedge\tau_\delta} \le \sup_t \Lambda_t^+$. Optional stopping at $\tau\wedge\tau_\delta$ and (13) give $d(\Delta)\,\mathbb{E}[T_{\tau\wedge\tau_\delta}] \ge (1-\delta)L_0 - 1$. For $t < \tau$ the posterior odds (12) are non-negative, so by posterior-monotonicity the algorithm is at least as likely to play $\hat\pi$ as π^* in round $t+1$; since $\{t < \tau\wedge\tau_\delta\}$ is

$\mathcal{F}_t$-measurable, summing these conditional probabilities yields $\mathbb{E}[N_{\hat{\pi}}(\tau \wedge \tau_\delta)] \geq \mathbb{E}[N_{\pi^*}(\tau \wedge \tau_\delta)]$. Hence $2\, d(\Delta)\, \mathbb{E}[N_{\hat{\pi}}(\tau \wedge \tau_\delta)] \geq (1-\delta)L_0 - 1$, and $N_{\hat{\pi}}(\tau_\delta) \geq N_{\hat{\pi}}(\tau \wedge \tau_\delta)$ pathwise.

(ii), certification term. This is the change-of-measure argument of (Kaufmann et al., 2016). Consider the alternative Bernoulli bandit that raises the mean of $\hat{\pi}$ alone to $\frac{1}{2} + \frac{\Delta}{2} + \eta$, making $\hat{\pi}$ optimal while leaving every other arm unchanged. A δ-PAC algorithm outputs π^* with probability at least $1-\delta$ under ν_{π^*} and at most δ under the alternative, so the data-processing inequality for the stopped trajectories yields $\mathbb{E}[N_{\hat{\pi}}(\tau_\delta)]\, \mathrm{kl}\big(\frac{1}{2} - \frac{\Delta}{2}, \frac{1}{2} + \frac{\Delta}{2} + \eta\big) \geq \mathrm{kl}(\delta, 1-\delta)$. Letting $\eta \downarrow 0$ and using $\mathrm{kl}(\delta, 1-\delta) \geq \log \frac{1}{2.4\delta}$ gives the second term. Both terms bound the same quantity, so their maximum does; the single-logarithm form follows from $\max\{a, b\} \geq \frac{1}{2}(a+b)$. The regret is then lower bounded by

$$\Delta\, \mathbb{E}[N_{\hat{\pi}}(\tau_\delta)] \;\geq\; \frac{\Delta}{d(\Delta)} \cdot \frac{1}{2}\Big(\log \frac{1}{2.4\delta} + \frac{(1-\delta)L_0 - 1}{2} \Big) = \frac{(1-\delta)\log \frac{(1-\epsilon)(K-1)}{\epsilon} + 2\log \frac{1}{2.4\delta} - 1}{4 \log \frac{1+\Delta}{1-\Delta}},$$

which is (9). ∎

F.2 Proof of Proposition 3

Proof Let P, Q be the joint distributions of $(\pi^*, \hat{\pi})$ under the training and test distributions, respectively, with common marginal $P_{\pi^*} = Q_{\pi^*} = \mu$ uniform on Π, and let $\Delta_\pi = D_{\mathrm{KL}}(P\|Q)$. Construct the test distribution as follows. Let $S \subset \Pi$ with $|S| = \lfloor K/2 \rfloor$, and set $Q_{\hat{\pi}|\pi^*}(\cdot) = P_{\hat{\pi}|\pi^*}(\cdot)$ for $\pi^* \in S$ and $Q_{\hat{\pi}|\pi^*}(\cdot) = \mathrm{Uniform}(\Pi)$ for $\pi^* \notin S$. Since the marginals of π^* agree, the chain rule gives $D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_\mu[D_{\mathrm{KL}}(P_{\hat{\pi}|\pi^*}\|Q_{\hat{\pi}|\pi^*})]$. For $\pi^* \in S$ the term is zero; for $\pi^* \notin S$ it is $D_{\mathrm{KL}}(P_{\hat{\pi}|\pi^*=\cdot}\|\mathrm{Uniform}) = \log K - H_P(\hat{\pi} \mid \pi^* = \cdot)$, the entropy of the point-conditional. Averaging over $\pi^* \notin S$ under uniform μ, and writing $H_P(\hat{\pi} \mid \pi^*) = \mathbb{E}_{\pi^*}[H(P_{\hat{\pi}|\pi^*=\cdot})]$ for the average conditional entropy,

$$D_{\mathrm{KL}}(P\|Q) \;=\; \frac{|\Pi \setminus S|}{K}\Big(\log K - H_P(\hat{\pi} \mid \pi^*) \Big) \;\leq\; \tfrac{1}{2} \log K,$$

where for odd K the factor $|\Pi \setminus S|/K = \lceil K/2 \rceil / K$ replaces $\frac{1}{2}$ throughout and does not affect the conclusion. Under Q, since for $\pi^* \in S$ the conditional equals $P_{\hat{\pi}|\pi^*}$ and for $\pi^* \notin S$ it equals $\mathrm{Uniform}(\Pi)$,

$$H_Q(\hat{\pi} \mid \pi^*) = \tfrac{1}{2} H_P(\hat{\pi} \mid \pi^*) + \tfrac{1}{2} \log K.$$

Since $H_Q(\hat{\pi}) \leq \log K$,

$$\begin{aligned} I_Q(\pi^*; \hat{\pi}) &= H_Q(\hat{\pi}) - H_Q(\hat{\pi} \mid \pi^*) \\ &\leq \log K - \big(\tfrac{1}{2} H_P(\hat{\pi} \mid \pi^*) + \tfrac{1}{2} \log K \big) \\ &= \tfrac{1}{2}\big(\log K - H_P(\hat{\pi} \mid \pi^*) \big) \\ &\leq \tfrac{1}{2} \log K. \end{aligned}$$

The shift is $D_{\mathrm{KL}}(P\|Q) = \frac{1}{2}(\log K - H_P(\hat{\pi} \mid \pi^*)) = \frac{1}{2} R^{\mathrm{train}}_{\mathrm{mech}} + \frac{1}{2}(\log K - H_P(\hat{\pi})) \leq \frac{1}{2} \log K$, since $H_P(\hat{\pi}) \geq 0$. When $P_{\hat{\pi}}$ is uniform, $H_P(\hat{\pi}) = \log K$, so $D_{\mathrm{KL}}(P\|Q) = \frac{1}{2} R^{\mathrm{train}}_{\mathrm{mech}}$ and $R^{\mathrm{test}}_{\mathrm{mech}} \leq \frac{1}{2}(\log K - H_P(\hat{\pi} \mid \pi^*)) = \frac{1}{2} R^{\mathrm{train}}_{\mathrm{mech}} \leq \frac{1}{2} \log K$. ∎

G Additional Experiment Details

Real-data motivation. In a prospective 37-site German cohort of 434 patients (Li et al., 2023), only 20.3% of BSA-dosed patients reached the therapeutic target AUC range of 20–30 mg·h/L, while 60.6% were underdosed and 19.1% overdosed. In the BSA arm of the randomized trial of Gamelin et al. (2008), only 4 of 49 assessable patients had plasma levels in the therapeutic range. The consequences are large: the Gamelin et al. (2008) randomized trial reported an objective response rate of 33.7% under PK-guided dosing versus 18.3% under BSA dosing ($P = .004$) in an intent-to-treat analysis of 208 patients, with median overall survival of 22 versus 16 months, a difference that did not reach significance ($P = .08$), and significantly fewer toxic events ($P = .003$) (Gamelin et al., 2008). A meta-analysis of four prospective trials ($N = 504$) confirmed 50–200% response rate improvement and $\approx 60\%$ reduction in grade 3/4 toxicity (Fang et al., 2016).

The learning problem. The key clinical bottleneck is the number of cycles required for the algorithm to converge to a patient's optimal dose. Kaldate et al. (2012) characterize the dose-AUC relationship in 187 FOLFOX6 patients (307 cycle-pair observations) as $\Delta\text{AUC} = 0.021 \times \Delta\text{dose mg/m}^2$ ($R^2 = 0.51$, $n = 307$) (Kaldate et al., 2012). The moderate R^2 reflects the large individual variability in 5-FU clearance, approximately 100-fold in plasma 5-FU levels under BSA-based dosing (Kobuchi & Ito, 2020), driven by DPD genotype, sex, age, and hepatic function. The intra-individual cycle-to-cycle variability is $\approx 20\%$ (Li et al., 2023), and the 10 mg·h/L width of the target window is chosen to accommodate it (Kaldate et al., 2012).

Sex differential. Females clear 5-FU $\approx 21\%$ slower (men have 26% higher clearance), producing a higher AUC under standard BSA dosing (22 vs. 18 mg·h/L) (Mueller et al., 2013). Clinical toxicity studies report disproportionately larger downstream effects on patient safety, driven by the asymmetric consequences of overdosing (Wagner et al., 2021). The algorithm individualizes rather than stratifying by sex, and sex enters only as a covariate of the clearance model if it is included there.

Burn-in. Proposition 2 gives a formal lower bound on the cost of a confidently-wrong prior. For the $K = 8$ setting with $\epsilon = 0.2$ and $\delta = 0.01$, the expression in (9) evaluates to ≈ 6.02 units of per-cycle reward (see the Running Example on burn-in for the working computation): an identification cost, incurred before the wrong recommendation is overturned, that is equivalent to $\approx$ 30 plays of a Δ-suboptimal arm.

Compute. The experiments were done on the author's personal machine using CPU only and do not require any special computing, and a full reproduction of every table and the in-silico example completes in well under an hour.

H CALIBRATION OF THE 5-FU EXAMPLE

This appendix instantiates the certificate twice, at two different levels of trust. The *literature tier* (Section H.1) evaluates Proposition 1 and Theorem 3 on two constants taken directly from a published population-PK study: the computation Running Examples 2–4 perform, and the one a clinician could repeat today with no simulation at all. The *in-silico tier* builds the full loop of Running Example 7: the true dynamics are the two-compartment Michaelis–Menten model of Tan et al. (2026), the deployed mechanistic model is the one-compartment linear-clearance surrogate, and every constant, including the realized R_{mech} that no published table can provide, is measured on that pair and checked against its certified ceiling.

The in-silico tier is laid out as follows. Section H.2 states the plant's stochastic model and the Bernoulli reward, with the provenance of each noise magnitude. Section H.3 measures the pre-trial measurement channel (σ_h, L_h, B_μ). Section H.4 substitutes those into Proposition 1 and Theorem 3 and checks the whole chain. Section H.5 isolates the channel in a scalar end-to-end system (Table 5). Section H.6 documents the deployed protocol and every algorithm exactly as run. Section H.7 compares the five policies (Tables 1 and 6) and their behaviour as the deployed model's bias grows (Table 7, Figures 2 and 3). Section H.8 generates R_{mech} from the measurement channel and scales the horizon (Tables 8 and 9, Figure 4). Section H.9 sweeps every choice made along the way (Tables 10–12, Figure 5). Section H.10 varies each design choice one at a time (Table 13, Figures 6 and 7).

How to read the provenance of a number. Every constant in this paper falls into exactly one of five classes, and we label it as such wherever it is introduced. *Quoted*: read off a table or sentence of a cited study. *Derived*: obtained by arithmetic on quoted numbers. The steps and the assumptions each step makes are stated, never only the result. *Convention*: a choice of units, discretisation or normalisation that we make and that no source dictates. *Identification*: an equality between two quantities that holds only under explicitly stated conditions, and that is not a general property of the framework. *Measured in-silico*: computed on the declared plant of Section H.3 with fixed seeds, reproducible from the repository. The distinction matters because only quoted and derived quantities are available to a reader with no simulation, and only measured ones can involve R_{mech} or B_μ, which require knowledge of the true dynamics.

H.1 LITERATURE TIER: THE CERTIFICATE FROM PUBLISHED CONSTANTS

Population spread $\sigma_h = 0.316$ (derived from one quoted number). Tan et al. (2026) (their Table 2) report the inter-individual variance of the eliminating parameter $V_{\max}$ as $\Omega^2 = 0.10$ on the log scale (relative standard error 11%): that variance is *quoted*. Turning it into σ_h takes two further steps, both ours. (i) The population-PK convention $V_{\max} = \mathrm{TV} \cdot e^{\eta}$, $\eta \sim \mathcal{N}(0, \Omega^2)$, makes Ω a standard deviation on the log scale, so $\sigma_h = \sqrt{0.10} = 0.316$. This is consistent with the 31% coefficient of variation the same study reports for the covariate model. (ii) The channel of Section 2.1 carries the *decision-relevant* parameter of the deployed one-compartment surrogate, and we identify that direction with $V_{\max}$. This is a modelling choice, not a finding of the cited study. The same study also reports 12% inter-occasion variability, which is deliberately excluded from σ_h because it is within-patient and is carried instead by the cycle-to-cycle reward noise of Section 5.

Pre-dose estimate precision $\sigma_\nu = 0.165$ (derived, four assumptions). This constant is *not* quoted anywhere: it is the end of a four-step chain, and we state the chain because each step is contestable. Tan et al. (2026), Section 3.4, report a mean absolute prediction error of 2.76 mg·h/L for the best three-sample limited-sampling scheme under continuous infusion (their $T = 0.8, 2, 5$ h, note that their "MAPE" is defined in absolute units, not as a percentage). The reference exposure is the median $\mathrm{AUC}_{0-\infty}$ of 21 mg·h/L for the 2,400 mg/m^2 continuous-infusion regimen, from their dosing simulation (their Section 3.5 and Fig. 5, 90% prediction interval $[12, 36]$). Then

$$\sigma_\nu \;=\; \frac{2.76}{21} \times \sqrt{\pi/2} \;=\; 0.131 \times 1.253 \;=\; 0.165.$$

The four assumptions, and what each costs: (1) a mean error is normalised by a *median* exposure of a single regimen, mixing two summary statistics. Using the 2,800 mg/m^2 panel's median of 25 instead gives $\sigma_\nu = 0.138$ and $C = 0.91$ nats. (2) The resulting relative error is treated as a log-scale error, exact only for small errors. (3) The conversion from mean absolute error to standard deviation assumes a zero-mean Gaussian error, $\mathbb{E}|X| = \sigma\sqrt{2/\pi}$. A limited-sampling estimator may be biased, and omitting the $\sqrt{\pi/2}$ factor altogether gives $\sigma_\nu = 0.131$ and $C = 0.96$ nats. (4) An error in estimated AUC is identified one-to-one with an error in the eliminating parameter, exact for linear pharmacokinetics but only approximate under the Michaelis–Menten elimination of the plant. Taken together these assumptions move the headline ceiling by roughly 0.1–0.2 nats, which is the honest precision of the literature tier. The certificate's qualitative verdicts are unchanged across that range. In this setting $l_\mu = \sigma_\nu$. Identification under (i)–(iii), $\hat{M} = h^{-1}(Z)$ makes the model's measurement-space error the draw's error: $C_\nu = C(l_\mu) = 0.77, l_\mu^{\mathrm{crit}} = 0.49, l_\mu = 0.165 < 0.49$.

H.2 THE PLANT'S STOCHASTIC MODEL AND THE REWARD

This subsection states the generative model of the in-silico tier in full: what is drawn, from which distribution, on which published grounds, and how the Bernoulli reward of Section 5 arises from it. Every choice is either quoted from the literature or labelled as an assumption.

Patients. A patient is a pair $(\mathrm{BSA}, V_{\max})$. Body surface area is a clipped normal (mean 1.90 m^2, sd 0.25, range $[1.4, 2.6]$, Tan et al., 2026, Table 1), and the eliminating parameter follows the published covariate model the functional form of $V_{\max} = \mathrm{TV}V_{\max}\,(\mathrm{BSA}/1.94)^{1.14}\,e^{\eta}$ with $\eta \sim \mathcal{N}(0, \Omega^2)$ (Tan et al., 2026, Table S4): the log-normal random effect is the standard population-PK convention and the form the source itself fits. The plant draws $\eta \sim \mathcal{N}(0, \Omega^2)$ with $\Omega^2 = 0.14$, the total between-patient variance the cohort inversion returns, rather than the 0.10 quoted for the eliminating parameter alone: the plant's log-clearance also carries the residual of the BSA regression. The literature tier keeps the quoted value. The true dynamics are Tan's two-compartment Michaelis–Menten ODE, integrated once per (patient, arm) pair. The therapeutic window $[20, 30]$ mg·h/L and the 1,600–3,600 mg/m^2 adjustment range are Kaldate et al. (2012)'s.

The reward is Bernoulli by construction. In cycle t on arm k the exposure the patient actually receives is $A_t = \mathrm{AUC}_k \cdot e^{\varepsilon_t}$, $\varepsilon_t \sim \mathcal{N}(0, \sigma_{\mathrm{cyc}}^2)$ drawn independently per cycle, and the reward is the indicator $r_t = \mathbf{1}\{A_t \in [20, 30]\}$. Its success probability is therefore the exact closed form

$$p_k \;=\; \Phi\Big(\frac{\ln(30/\mathrm{AUC}_k)}{\sigma_{\mathrm{cyc}}}\Big) - \Phi\Big(\frac{\ln(20/\mathrm{AUC}_k)}{\sigma_{\mathrm{cyc}}}\Big),$$

which is what the certificate constants and the regret are computed from. This is not a second model layered on the simulator: it is the exact mean of the indicator the simulator draws, so the reward drawn and the regret charged are consistent by construction, and p_k is never itself sampled (an estimated p_k would flip the arg-max π^* for a few percent of patients and corrupt every information quantity). On this plant the realised attainments are $p_{\mathrm{BSA}} = 0.353$ and $p_{\mathrm{opt}} = 0.624$ against the published 0.20 and 0.85 of Section 5. The plant is kinder to standard dosing and harsher to the ideal policy than the literature-calibrated attainments of Section 5, and the two are not two estimates of one number.

Three variance channels, and which of them the reward sees. Between-patient variation (BSA, $V_{\max}$) determines each patient's AUC_k vector and hence π^*. Its spread across the cohort is the entropy $H(\mu)$ a learner must resolve. Within-patient between-cycle variation (σ_{cyc}) moves the exposure the patient actually receives, so it, and only it, randomises the reward. Assay error (σ_{assay}) moves the number the laboratory reports, corrupting the model's clearance estimate while leaving the realised exposure, and therefore the reward, untouched. The reward integrates the biological channel alone because therapeutic benefit depends on the exposure received, not on the reading taken from it. Both channels corrupt the lead-in measurement, which is why $\sigma_\nu = \sqrt{\sigma_{\mathrm{cyc}}^2 + \sigma_{\mathrm{assay}}^2} = 0.297$ enters the certificate while σ_{cyc} alone enters p_k.

Why the within-patient factor is multiplicative log-normal. The reward's randomness is what population PK calls inter-occasion variability (IOV), and the plant's own source models exactly that quantity: Tan et al. (2026) test the first-versus-later-cycle difference "by adding IOV on top of the IIV of $V_{\max}$" and estimate it at 12%. Their IIV enters multiplicatively as e^{η} (the covariate model above), so IOV enters the same way and the induced perturbation of exposure is log-normal. the simulator's e^{ε} copies the source's own convention. The *proportional* residual-error model of the 5-FU literature ($y = f \cdot (1 + \epsilon)$, 25.5% in Mueller et al., 2013, 30–34% in Tan et al., 2026) describes a different quantity: error on the *observed concentration* at the sampling time, which is the concept behind σ_{assay}, not behind σ_{cyc}. The choice was measured rather than assumed: replacing the reward's link by $\mathrm{AUC}_k(1 + \varepsilon)$ moves the aggregates negligibly (p_{opt} $0.624 \to 0.625$, $H(\mu)$ $2.0316 \to 2.0302$ at $\sigma = 0.20$, on the evaluation cohort) while changing the arg-max arm for 10.5% of patients. The two forms converge as σ shrinks (97% agreement at $\sigma = 0.12$), as $e^{\varepsilon} \approx 1 + \varepsilon$ requires.

Magnitudes, with provenance. $\sigma_{\mathrm{cyc}} = 0.20$ is quoted: the "intraindividual variability of $\sim$20%" of Li et al. (2023). It is the top of the published range, and two independent sources corroborate the scale: Tan et al. (2026) fit an IOV of 12% (and judge it negligible), and Morawska et al. (2018) measure a 17% intra-individual coefficient of variation of the AUC/dose ratio under repeated cycles. The assay-side magnitude $\sigma_{\mathrm{assay}} = 0.22$ is an assumed, rounded order-of-magnitude choice, of plausible size next to the proportional residual errors the 5-FU literature reports on observed concentrations (25.5% in Mueller et al., 2013, 30–34% in Tan et al., 2026). It is a modelling choice, not an estimate of any dataset, and the split between the biological and the assay channel is an assumption, and a load-bearing one: moving variance from the assay channel to the biological channel lowers every policy's attainable reward while leaving σ_ν unchanged. The magnitude choice is consequential for the absolute attainment level: at $\sigma_{\mathrm{cyc}} = 0.12$ (Tan's own IOV) the best-arm attainment would be $p_{\mathrm{opt}} = 0.797$ against 0.624 at the quoted 0.20, while $H(\mu)$, and with it the entire certificate chain, is unmoved (2.032 either way). The simulation keeps the quoted, conservative 0.20.

H.3 In-silico tier: the measurement channel

Every constant in this subsection is measured on the calibration-study cohort ($n_{\mathrm{study}} = 2{,}000$, seed 2024), never on the evaluation cohort, so the certificate is a function of study data alone.

The plant and the pre-trial measurement Z. The true dynamics are the two-compartment Michaelis–Menten model of Tan et al. (2026). The deployed mechanistic model is a one-compartment linear-clearance surrogate. Section 2.1 posits $Z = h(M^*) + \nu$ with $\dim Z = d_Z$. On the plant this is the lead-in blood test: cycle 1 administers the body-surface-area dose (2,457 mg/m^2 on the grid) and draws one plasma concentration at 40 h, on the infusion plateau. Thus $h(M^*)$ is the noise-free log plateau concentration, ν collects the biological cycle-to-cycle and assay variation ($\sigma_\nu = 0.297$ in log units, derived), and $d_Z = 1$.

Measurement spread $\sigma_h = 0.393$**.** $\sigma_h^2 = \operatorname{tr}\operatorname{Cov}(h(M^*))$ is the variance of the noise-free log plateau concentration across the patient population, the signal the channel carries. Measured directly on the study cohort.

Identifiability constant $L_h = 0.765$**.** Following Appendix D.1, the plant has a single patient-specific direction (the clearance scale, perturbed through $V_{\max}$ at fixed body surface area), h is the identity on the log-concentration measurement ($h_{\min} = 1$ exactly), and $L_h = 1/g_{\max}$ with $g_{\max} = \sup_i \|\sqrt{\mu} \odot (\partial J/\partial z - c\mathbf{1})\|_2 = 1.308$, the sample maximum over the 2,000 study patients, the largest centred, occupancy-weighted sensitivity of the per-arm reward vector to the measurement, the same norm that defines B_μ in (3), which is what the proof of Proposition 1 requires. (Robustness: under the sup norm that dominates it, the measured constant is $L_h = 0.547$, giving the looser $C(B_\mu) = 1.23$ nats.) The estimate is stable to the finite-difference step: Table 12.

Model bias $B_\mu = 0.221$**, and why the definition is centred.** Equation (3) evaluated on the deployed model, the one-compartment surrogate inverted once on the lead-in draw, against the plant's per-arm in-window probabilities: centred by the occupancy-weighted mean, occupancy-weighted, and pooled across patients by RMS. Without the centring the value is 0.250, giving $C(B_\mu) = 0.83$ nats in place of 0.93. Both are reported here because an earlier version of (3) was uncentred and stated for a single realized pair $(\mathcal{M}^*, \hat{\mathcal{M}})$ rather than in expectation, so a reader comparing versions should see what the change is worth. The centred, expectation form is primary, for three reasons. First, it is the only part of the reward vector the results depend on: the recommendation is an $\arg\max$ and the regret is a difference of arm means, and both are invariant to adding a constant to every arm, so a model that is uniformly optimistic across all doses makes the same recommendation and pays the same regret while an uncentred B_μ would charge it without bound. Second, it is the norm in which L_h is constructed (Appendix D.1), so the chain $\mathbb{E}\|h(\hat{\mathcal{M}}) - h(\mathcal{M}^*)\|^2 \geq L_h^2 B_\mu^2$ used in Proposition 1 holds in one norm throughout. In the uncentred norm two models differing by a pure offset would be at positive distance while their measurements coincide, and no positive L_h could satisfy Assumption 3. Third, it is conservative: centring can only shrink B_μ, hence only widen the ceiling $C(B_\mu)$ and weaken the claim, which is the safe direction for a certificate. On this plant the choice changes no verdict: both values sit well below the one-cycle threshold of Theorem 3.

Prior entropy and mechanistic information. $H(\mu) = 2.034$ nats, the plug-in entropy of the measured optimal-arm occupancy (the uniform reference is $\ln 8 = 2.079$). The deployed prior carries $R_{\text{mech}} = 0.417$ nats, measured as the plug-in mutual information between π^* and the model's recommendation, a shuffled-estimator null gives 0.012 nats. Before the lead-in draw the recommendation is constant across patients and $R_{\text{mech}} = 0.000$ exactly: the single blood test is what makes the ODE a prior.

H.4 Composite certificate at the measured values

Substituting into the two bounds of Proposition 1 and the threshold (8):

$$C(B_\mu) = \tfrac{1}{2}\ln\left(1 + \frac{0.393^2}{0.765^2 \times 0.221^2}\right) = 0.929 \text{ nats} \qquad \text{(bias form: the certificate)}, \tag{14}$$

$$R_{\text{mech}} = 0.417 \;\leq\; C(B_\mu) = 0.929 \qquad \text{(implied by } R_{\text{mech}} \leq C_\nu\text{)}, \tag{15}$$

$$H_{\text{mech}} \geq H(\mu) - C(B_\mu) = 1.10 \text{ nats} \quad \text{(estimator-free: 1.53, measured: 1.62)}. \tag{16}$$

H.5 End-to-end scalar channel ablation

Between the literature tier (constants only) and the plant (full ODE pair) sits the smallest end-to-end system that exercises every part of the theory: a scalar channel plus the dose grid, with the literature constants. A patient is a log-clearance deviation $\theta^* \sim \mathcal{N}(0, \sigma_h^2)$. The pre-dose estimate is $Z = \theta^* + \nu$, $\nu \sim \mathcal{N}(0, \sigma_\nu^2)$. The model with parameter Z recommends its best arm. The realized $I(\pi^*; \hat{\pi})$ is computed by Monte Carlo over 400,000 patients and Thompson sampling runs with the induced prior $\mu_{\text{hyb}} = P(\pi^* \mid \hat{\pi})$, the posterior over the best arm being approximated by reweighting a 128-sample Monte-Carlo estimate of the best arm under per-arm Beta posteriors. Result: the realized information is $R_{\text{mech}} = 0.63$ nats against its ceiling $C(B_\mu) = C_\nu = 0.77$ ($L_h = 1$ on this plant, and the channel bound holds, with the slack coming from the arg-max discretization), and the induced prior yields the regret ordering of Table 5: the induced prior beats uninformed TS by $2.06\times$ and fixed BSA dosing

by $1.86\times$. On this *linear* plant the model is exactly specified, so the whole gap is the value of the channel. The model-misspecified regime is what the next subsections measure on the saturable plant.

policy	regret over $N = 12$ cycles	observation it uses
uninformed TS	3.20 ± 0.03	binary in-window reward
fixed BSA dose	2.89 ± 0.06	none
$\mathrm{TS}_{\mathrm{hyb}}$ (induced μ_{hyb})	1.55 ± 0.02	prior + binary reward

Table 5: Scalar channel ablation: mean $\pm$ 95% CI over $m = 6{,}000$ patients, seed 11. Realized $R_{\mathrm{mech}} = 0.63 \leq C_\nu = 0.77$ nats.

H.6 The deployed protocol and the algorithms, exactly as run

This subsection records the full pipeline.

Cohort discipline. Two disjoint cohorts are drawn from the population of Section H.2: a *calibration study* of 2,000 patients (the size of a realistic dose-ranging study, and everything the hybrid knows before treating anyone is a function of this cohort alone) and an *evaluation* cohort of 10,000 patients on which every regret is scored. The study-phase construction takes the evaluation cohort as no argument, which makes "no leakage" a checkable property of the code rather than an assertion.

Model calibration. The deployed one-compartment model's population clearance is fitted on the study cohort: each study patient's apparent clearance is recovered from their noise-free plateau concentration at the BSA dose, the BSA covariate is regressed out at the exponent $b = 1.149$ (the plant's own $d \log \mathrm{CL}/d \log \mathrm{BSA}$ at the typical patient, computed from Tan's constants), and the typical value is the geometric mean over the enrolled slice. Fitted on the full cohort the calibration is correct ($\gamma = 1$). The miscalibration sweep re-fits it on the slowest- or fastest-clearing fraction, which biases the whole clearance scale by a known γ that multiplies every subsequent inversion, so no consultation of the model escapes it.

The offline study (what the hybrid knows in advance). The study cohort runs the identical lead-in protocol (below), giving each study patient a model recommendation $\hat{\pi}$. Two objects are then estimated against the study patients' known optimal arms π^*, and they are all the hybrid ever receives: (i) the prior table $\mu_{\mathrm{hyb}}(j \mid i) = P(\pi^* = j \mid \hat{\pi} = i)$, estimated with add-one (Laplace) smoothing so no (recommendation, truth) pair sits at probability zero, and (ii) the reward-shape matrix $Q[k, j] = P(r = 1 \mid \text{arm } j \text{ played}, \pi^* = k)$, estimated from *one binary outcome per (patient, arm)*, never from the plant's exact probabilities (a dose-ranging study records the arm administered, and the outcomes at the other arms are counterfactuals the simulator supplies, an idealisation stated here as one). Rows with 20 or fewer patients fall back to the cohort mean row, and entries are clipped to $[0.02, 0.98]$. Row k of Q is the dose–response curve of a patient whose optimum is arm k, so a failure at arm j is evidence about how far the truth is from j: the structure that lets a short horizon identify one of K arms. Because the table conditions the model's recommendation on ground truth, the model contributes its information and not its error: a systematic model error is a systematic off-diagonal pattern that row-normalising cancels, and when the model is useless the table collapses to the cohort marginal. This is what makes $H(\pi^* \mid \hat{\pi}) = H(\mu) - R_{\mathrm{mech}}$ hold by construction, the hypothesis of Theorem 2. The cost of this construction is stated as the assumption it is: the study patients' optimal arms must be known, i.e. a prior dose-ranging study established them.

The lead-in (cycle 1, identical for every policy). Every patient receives the body-surface-area dose (2,457 mg/m^2 on the grid) and gives one plasma sample at 40 h, on the infusion plateau. The model is inverted on that single reading: $\widehat{\mathrm{CL}} = \gamma R / c_{\mathrm{obs}}$ at the infusion rate R, read as a plateau. Before the draw, the model's estimate is the covariate-scaled population value $\mathrm{CL}_{\mathrm{pop}}(\mathrm{BSA}/1.94)^{1.149}$. Because dose is per square metre while apparent clearance scales as $\mathrm{BSA}^{1.149}$, this pre-draw model recommends the same arm to every patient and $R_{\mathrm{mech}} = 0.000$ exactly. The lead-in cycle is not scored, its regret (0.270 per patient) is identical for every policy, and the model is never consulted for its dose.

The recommendation rule. Given a clearance estimate, the model predicts $\widehat{\mathrm{AUC}}_k = d_k \cdot \mathrm{BSA}/\widehat{\mathrm{CL}}$, converts it to a predicted in-window probability by the same closed form as Section H.2, and recommends the arg-max. Nothing assumes the reward is symmetric: a different reward would

change the recommendation through the predicted probabilities and through nothing else. One degenerate case needs a rule: a patient the model places entirely outside the window has a probability row that underflows to exact zeros, and those rows fall back to the arm whose predicted exposure is closest to the window as an interval.

The five policies over the eleven scored cycles. All five face the same patients and the same pre-drawn noise streams. The paired comparisons are paired by construction. In each cycle a policy chooses an arm, regret $p_{\pi^*} - p_{\text{arm}}$ accrues in expected reward (Definition 1), the cycle's biological factor e^{ε_t} realises the exposure, the binary reward $r_t = \mathbf{1}\{A_t \in [20, 30]\}$ is observed, and the policy updates:

- *BSA static*: the standard-of-care arm in every cycle, reading nothing.
- *Uninformed TS*: per-arm Beta–Bernoulli Thompson sampling from $\mathrm{Beta}(1 + \text{successes}, 1 + \text{failures})$, reading the binary reward. Posterior draws use inverse-CDF sampling on the pre-drawn uniforms so the random streams of the policies never desynchronise.
- *ODE, one solve*: the mechanistic model is solved once, on the lead-in draw, and its recommendation $\hat{\pi}$ is played in every scored cycle. No reward is read, no further concentration is read and the ODE is never re-solved. It therefore holds exactly the information the hybrid's prior holds, the same draw, the same solve and the same $\hat{\pi}$, and it does not learn, which is what makes it the comparator that prices the reward feedback rather than the prior.
- *Mechanistic (per-cycle controller)*: plays the model's current recommendation, then refits the clearance on a fresh draw of the cycle's observed concentration (c_{obs}, which carries the assay error): the running estimate is the mean of all inverted log-clearances, so the lead-in and every scored cycle enter with equal weight. Twelve ODE solves per course.
- *Hybrid (Theorem 2)*: Thompson sampling over the *latent best arm*. Its prior is the row of the study table selected by the cycle-1 recommendation, log-posterior $\log \mu_{\text{hyb}}(\cdot \mid \hat{\pi}) + \sum_t \log Q[\cdot, a_t]^{r_t}(1 - Q[\cdot, a_t])^{1-r_t}$, and each cycle the arm is *sampled* from that posterior (probability matching), never arg-maxed. The model is solved once, at the lead-in, the prior is frozen thereafter and the concentration reading is never consulted again. The per-round observation is the binary reward alone, the paper's formal observation model.

One idealisation is explicit: the reward is computed from the true realised exposure, while the concentration the mechanistic policy refits on carries the assay error, so the binary readout is noise-free where the continuous one is not. This favours the reward-reading policies and is the same idealisation the paper's reward model makes.

H.7 Policy comparison on the plant

Table 1 of the main text reports four of the five policies on the calibrated plant of Running Example 7, and Table 6 reports all five, the per-cycle controller included. This section documents how they were produced. The protocol is identical for all: cycle 1 is the shared BSA dose with one blood test (unscored, lead-in cost 0.27 per patient), then 11 scored cycles. The hybrid implements the policy of 2, with an estimated, smoothed prior table and a Q-likelihood: Thompson sampling over the latent best arm with prior $\mu_{\text{hyb}} = P(\pi^* \mid \hat{\pi})$, frozen after the single lead-in draw, observing the binary reward alone, the paper's formal observation model. It reads the lead-in draw once and solves the ODE once for the whole course, and thereafter only the binary reward. The one solve reads the same draw and performs the same single solve and then reads nothing at all, while the per-cycle controller reads the continuous concentration and re-solves the model in every cycle.

Three readings. First, what the prior alone is worth. The one solve consumes the lead-in draw and one ODE solve and nothing else, and it already pays 1.83 against 2.97 for BSA and 3.07 for the uninformed learner, $1.62\times$ and $1.67\times$. That is the mechanistic recommendation priced on its own, with no learning of any kind, and it is most of the paper's headline gain. The $R_{\text{mech}} = 0$ point of the channel sweep below (Section H.8) is the same comparator seen from the other side: before the lead-in draw the model recommends the same arm to every patient, so a learner that ignores the draw carries no mechanistic information at all.

Second, what the reward feedback is worth. The hybrid pays 1.66 against the one solve's 1.83, a ratio of 1.11 $[1.08, 1.13]$, and that gap is the whole contribution of Theorem 2's posterior update at this

policy	regret	vs. hybrid	lead-in draw	per-cycle readout	solves
uninformed TS	3.07 ± 0.02	$1.85\,[1.82, 1.88]$	not read	binary	0
BSA static	2.97 ± 0.04	$1.79\,[1.76, 1.83]$	not read	none	0
ODE, one solve	1.83 ± 0.04	$1.11\,[1.08, 1.13]$	read	none	1
hybrid (Thm. 2)	1.66 ± 0.02	1.00	read	binary	1
mechanistic, per cycle	0.73 ± 0.02	$0.44\,[0.43, 0.45]$	read	continuous	12

Table 6: All five policies at the correctly calibrated model ($\gamma = 1$, 11 scored cycles, $n = 10{,}000$ evaluation patients, seeds 11/2024): cumulative regret (mean $\pm$ 95% CI), the comparator/hybrid ratio of means with paired-bootstrap 95% intervals, and the information each policy consumes. The main text's Table 1 carries the first four rows, the per-cycle controller appearing only here and in Table 7. The unscored lead-in cycle costs 0.27 for everyone and is excluded from the regret column.

operating point, since the two policies enter cycle 1 with identical information. The regret-ratio floor Theorem 1 predicts at the measured $H_{\mathrm{mech}} = 1.62$ is $\sqrt{H(\mu)/H_{\mathrm{mech}}} = 1.12$, and the observed 1.85 against the uninformed learner exceeds it, as a minimax floor should be exceeded. The hybrid holds two objects the one solve does not, the study table $P(\pi^* \mid \hat{\pi})$ and the shape matrix Q, and both require a dose-ranging study whose patients' optimal arms are known, which is a real cost and is stated as an assumption in Section H.6.

Third, the per-cycle controller, which is why it stays in this appendix. At the calibrated model it pays 0.73, $0.44\times$ the hybrid, but it reads a continuous concentration and re-solves the ODE in all twelve cycles, an information budget the theory does not assume and a clinic would have to fund, and under bias it is the most fragile of the five (Table 7).

The two-compartment world versus the one-compartment model, under calibration bias. The deployed surrogate differs from the world in two ways. The *topology* gap, dropping the peripheral compartment, is worth under 1% of exposure (measured by switching the plant's integrator to one compartment). The real misspecification is the *linearisation*: the surrogate's dose–exposure line is exact at the calibration dose and drifts as DPD saturation bends the true curve, and re-fitting the surrogate's population clearance on a slice of the cohort tilts the whole line. Figure 2 shows that behavioral distance directly (the model−world AUC error across the dose range, one line per calibration slice), and it maps monotonically onto the measured centred bias, $B_\mu = 0.22 \to 0.30$.

Figure 3 and Table 7 show what that distance costs. The hybrid's prior table is fitted on the same slice the deployed model is calibrated on, so the dose-ranging study and the deployment describe one population model. Under that discipline the hybrid's regret grows from 1.66 at the correct model to 1.83 at the harshest slice, about a tenth, because the prior enters once, so once the rewards accumulate they take over and the initial misdirection is not re-asserted. At every slice it remains below uninformed TS (3.07) and BSA (2.97).

Both mechanistic policies degrade further and faster, and they separate the two ways a biased model can be paid for. The one solve, which carries the drifted recommendation into every cycle but never re-estimates it, moves $1.83 \to 3.60$ over the sweep. It crosses BSA between the slowest-clearing 40% and 35% and the fastest-clearing 35% and 30% slices and uninformed TS between the slowest 35% and 30% and the fastest 30% and 25%, so a model good enough to beat standard of care at $\gamma = 1$ stops being so once B_μ passes roughly 0.28, while the hybrid built on the same recommendation never crosses either reference. The per-cycle controller, which re-asserts the wrong model twelve times per course, degrades roughly linearly in B_μ and ends $1.8\times$ worse than the hybrid at the harshest slice. Its fragility decomposes: most of it is already in the one solve, that is, in the biased model itself, and only the remainder is the cost of re-asserting it. The two cross between the 50% and 45% slices, on both sides. At the 50% slices and wider the per-cycle controller is still the better of the two (2.56 against 2.61 slow, 2.39 against 2.45 fast), since its fresh readings partly correct the population term, and from the 45% slices inward it is the worse (2.83 against 2.74 slow, 2.63 against 2.57 fast), since the readings are inverted through a model whose whole clearance scale is wrong. Information transferred once, as a prior, degrades gracefully. A model re-asserted every cycle re-asserts the drift, and a model played open-loop keeps the drift it started with.

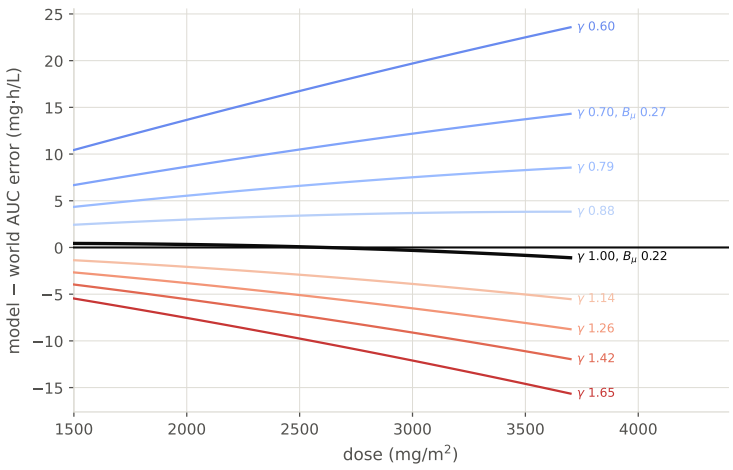

Figure 2: How far the deployed one-compartment model is from the two-compartment world: model−world AUC prediction error across the dose range for the median-clearance patient, one line per calibration slice, labelled with γ (calibrated ÷ true clearance) and the measured centred B_μ. Under-estimated clearance ($\gamma < 1$, blue) reads too much exposure and pushes doses down, while over-estimated clearance (red) does the reverse. The correctly calibrated model (black) is exact at the calibration dose and drifts only with saturation.

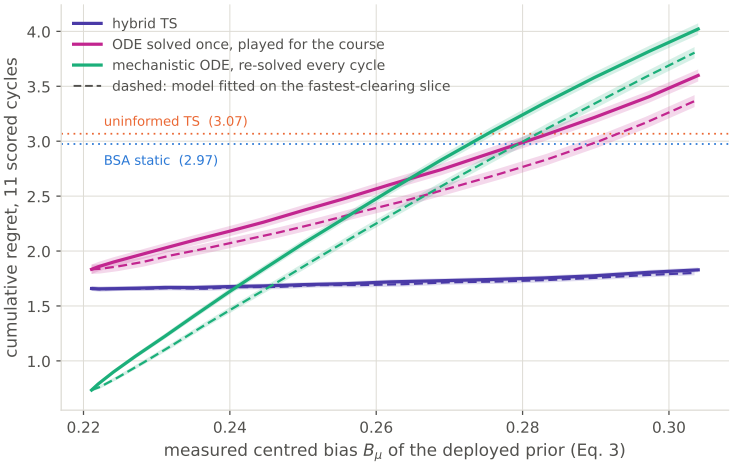

Figure 3: What the model's bias costs each deployment: cumulative regret over the 11 scored cycles as a function of the measured centred B_μ: mean curves with shaded 95% bands over the dense calibration sweep (33 slices). Solid: model fitted on the slowest-clearing slice. Dashed: on the fastest. For the hybrid the two directions nearly coincide (they differ by at most 0.03 at matched slices), so B_μ, not the direction of the miscalibration, is what prices its regret, while the per-cycle controller separates further, by up to 0.235. Three model-based curves are shown. The hybrid (Theorem 2: TS with the ODE prior, one draw, one solve, prior table fitted on the same slice as the deployed model, rises by about a tenth across the sweep and stays below both model-free references (dotted, constant). The ODE solved once on the same draw, playing $\hat{\pi}$ in every cycle with no feedback, rises from 1.83 to 3.60 and crosses both references near $B_\mu \approx 0.28$. The per-cycle controller degrades linearly, is the best of the three near $\gamma = 1$ and the worst of the three from the 45% slices inward, where it also passes the one solve.

calibration slice	B_μ	R_{mech}	floor	$\mathrm{TS}_{\mathrm{hyb}}$	ODE, once	$\frac{\mathrm{once}}{\mathrm{TS}_{\mathrm{hyb}}}$	per-cycle ODE	$\frac{\mathrm{TS}_{\mathrm{uninf}}}{\mathrm{TS}_{\mathrm{hyb}}}$
slowest 20%	0.304	0.306	1.08	1.83 ± 0.02	3.60 ± 0.06	$1.97\,[1.94, 2.00]$	4.02 ± 0.05	1.68
slowest 40%	0.276	0.364	1.10	1.74 ± 0.02	2.90 ± 0.05	$1.66\,[1.63, 1.69]$	3.09 ± 0.04	1.76
slowest 60%	0.250	0.396	1.11	1.69 ± 0.02	2.37 ± 0.05	$1.40\,[1.37, 1.42]$	2.07 ± 0.03	1.81
slowest 80%	0.231	0.410	1.12	1.67 ± 0.02	2.03 ± 0.05	$1.21\,[1.19, 1.24]$	1.22 ± 0.02	1.84
everyone ($\gamma = 1$)	0.221	0.417	1.12	1.66 ± 0.02	1.83 ± 0.04	$1.11\,[1.08, 1.13]$	0.73 ± 0.02	1.85
fastest 80%	0.232	0.408	1.12	1.66 ± 0.02	1.96 ± 0.04	$1.18\,[1.16, 1.20]$	1.14 ± 0.02	1.85
fastest 60%	0.252	0.387	1.11	1.69 ± 0.02	2.26 ± 0.05	$1.34\,[1.31, 1.36]$	1.94 ± 0.03	1.82
fastest 40%	0.277	0.356	1.10	1.72 ± 0.02	2.70 ± 0.05	$1.57\,[1.54, 1.60]$	2.88 ± 0.04	1.78
fastest 20%	0.303	0.303	1.08	1.80 ± 0.02	3.37 ± 0.05	$1.87\,[1.83, 1.90]$	3.81 ± 0.05	1.70

Table 7: The nine canonical calibration slices in full: the hybrid (Theorem 2's algorithm, one draw, one solve) against the ODE solved once on the same draw, the per-cycle mechanistic controller (twelve of each) and the two model-free references, with the certificate quantities measured in each slice. Regret over 11 scored cycles (mean $\pm$ 95% CI, $n = 10{,}000$ evaluation patients). The once/hybrid column is a ratio of means with paired-bootstrap 95% intervals, and the per-cycle controller's ratio against the hybrid runs from $0.44\times$ at $\gamma = 1$ to $2.20\times$ at the slowest 20% slice. The floor is Theorem 1's $\sqrt{H(\mu)/H_{\mathrm{mech}}}$ with $H_{\mathrm{mech}} = H(\mu) - R_{\mathrm{mech}}$, $H(\mu) = 2.034$ nats. The slices' γ values are those of Table 3 and Figure 2. Model-free references, constant at every bias: uninformed TS 3.068 and BSA 2.975, so BSA/hybrid is 0.970 times the uninformed ratio in every row. The hybrid's prior table is fitted on the same slice as the deployed model. Table 3 of the main text reports three of these rows, and Figure 3 shows the full 33-slice sweep.

H.8 Channel-design scenarios and horizon scaling

Regret against R_{mech} generated by the channel. Table 8 is the full version of Table 2: R_{mech} scenarios produced at the calibrated model by the pre-dose measurement channel, the assay precision σ_{assay} over 0.02–0.50 at one draw and the number of pooled lead-in draws $d_Z \in \{1, 2, 3\}$ at the quoted $\sigma_{\mathrm{assay}} = 0.22$. This is scenario construction over the channel, not a sensitivity of a calibrated constant and not model bias: the deployed model is the correctly calibrated one in every row, and the miscalibration axis is Table 7's. The $R_{\mathrm{mech}} = 0$ row is the *uninformed* reference: before the lead-in draw the model's recommendation is constant across patients and $R_{\mathrm{mech}} = 0.000$ exactly, so a learner that does not read the draw is uninformed Thompson sampling, and the hybrid's regret in that row is that reference's 3.07. Figure 4 plots the same sweep. The certificate holds in all eleven rows ($R_{\mathrm{mech}} \le C(B_\mu)$, $B_\mu < B_\mu^{\mathrm{crit}}$), with the tightest margin $3.02\times$ over $B_\mu^{\mathrm{crit}} = 0.810$ at one draw. Precision saturates below $\sigma_{\mathrm{assay}} \approx 0.05$ because the residual $\sigma_{\mathrm{cyc}} = 0.20$ is patient biology: the maximum a single draw carries on this plant is 0.631 nats. Pooling draws does move the ceiling, to 0.706 nats at three draws, and the literature ceiling of 0.77 nats (Running Example 2, a three-sample scheme with $\sigma_\nu = 0.165$) is not attained on this plant.

This sweep is also where the hybrid loses, and the table says so plainly. The one solve falls faster in R_{mech} than the hybrid does, from 2.80 at $\sigma_{\mathrm{assay}} = 0.50$ to 1.20 at 0.02 against the hybrid's 1.99 to 1.36, because a sharper channel improves the recommendation it plays directly while the hybrid still spends cycles sampling around that recommendation. The two are statistically tied at $\sigma_{\mathrm{assay}} = 0.15$ (1.53 each, ratio $1.00\,[0.98, 1.02]$), the one solve is ahead at 0.10 and below (0.94 at 0.10, 0.90 at 0.05) and ahead at two and three pooled draws (0.88 and 0.77). So the crossing lies between $\sigma_{\mathrm{assay}} = 0.10$ and the quoted operating point 0.22, that is between $R_{\mathrm{mech}} = 0.56$ and 0.42 nats, and the paper's operating point sits just on the hybrid's side of it. The reading is not that the posterior update is worthless at high R_{mech} but that probability matching keeps paying for exploration whose value shrinks as the prior concentrates, which is the finite-N face of Theorem 2's asymptotic statement. Under bias (Table 7) the ordering never reverses, so the hybrid's advantage is an advantage against drift, not against information and short horizons.

Horizon scaling at the calibrated model. Table 9 varies the number of scored cycles at the correctly calibrated model with everything else held fixed. The advantage grows with the horizon, from $1.33\times$ over BSA at $N = 3$ to $3.45\times$ at $N = 50$, while the certificate does not move: $H_{\mathrm{mech}} = 1.617$ nats and the floor 1.12 in every row. The two are not in tension, because the floor is a statement about sample complexity and not about the ratio of N-cycle regrets. The uninformed learner must resolve eight arms from one bit per cycle, so it keeps accruing regret where the hybrid's prior has already

σ_{assay}	d_Z	σ_ν	R_{mech}	$C(B_\mu)$	B_μ	floor	TS_{hyb}	ODE, once	$\frac{\text{once}}{\text{TS}_{\text{hyb}}}$	$\frac{\text{TS}_{\text{uninf}}}{\text{TS}_{\text{hyb}}}$
—	—	—	0	—	—	1.00	3.07 ± 0.02	—	—	1.00
0.50	1	0.539	0.179	0.772	0.268	1.05	1.99 ± 0.02	2.80 ± 0.05	1.41 [1.38, 1.43]	1.54 [1.53, 1.56]
0.40	1	0.447	0.246	0.805	0.257	1.07	1.90 ± 0.02	2.53 ± 0.05	1.33 [1.31, 1.35]	1.61 [1.59, 1.63]
0.30	1	0.361	0.330	0.862	0.240	1.09	1.78 ± 0.02	2.17 ± 0.05	1.22 [1.20, 1.24]	1.72 [1.70, 1.74]
0.22	1	0.297	0.417	0.929	0.221	1.12	1.66 ± 0.02	1.83 ± 0.04	1.11 [1.08, 1.13]	1.85 [1.82, 1.88]
0.15	1	0.250	0.508	1.005	0.202	1.15	1.53 ± 0.02	1.53 ± 0.04	1.00 [0.98, 1.02]	2.01 [1.98, 2.04]
0.10	1	0.224	0.564	1.063	0.189	1.18	1.45 ± 0.02	1.35 ± 0.04	0.94 [0.92, 0.96]	2.12 [2.09, 2.15]
0.05	1	0.206	0.614	1.110	0.180	1.20	1.38 ± 0.02	1.24 ± 0.03	0.90 [0.88, 0.92]	2.23 [2.19, 2.26]
0.02	1	0.201	0.631	1.127	0.176	1.20	1.36 ± 0.02	1.20 ± 0.03	0.89 [0.87, 0.90]	2.26 [2.23, 2.30]

Table 8: Regret against mechanistic information generated by the measurement channel at the calibrated model ($\gamma = 1$, 11 scored cycles, $n = 10{,}000$ evaluation patients, regret mean $\pm$ 95% CI, ratios of means with paired-bootstrap 95% intervals). Upper block: the σ_{assay} axis at one draw, with $\sigma_\nu = \sqrt{\sigma_{\text{cyc}}^2 + \sigma_{\text{assay}}^2}$ and $\sigma_{\text{cyc}} = 0.20$. Lower block: the pooled-draw axis at the quoted $\sigma_{\text{assay}} = 0.22$. The floor is Theorem 1's $\sqrt{H(\mu)/H_{\text{mech}}}$, $H(\mu) = 2.034$ nats. The model-free references are constant throughout, uninformed TS 3.07 and BSA 2.97. First row: the pre-draw model, which recommends one arm to everyone, so $R_{\text{mech}} = 0.000$ and the hybrid *is* uninformed TS. † per-draw ceiling $C(B_\mu)/d_Z$. The raw values 2.179 and 3.639 carry the d_Z prefactor.

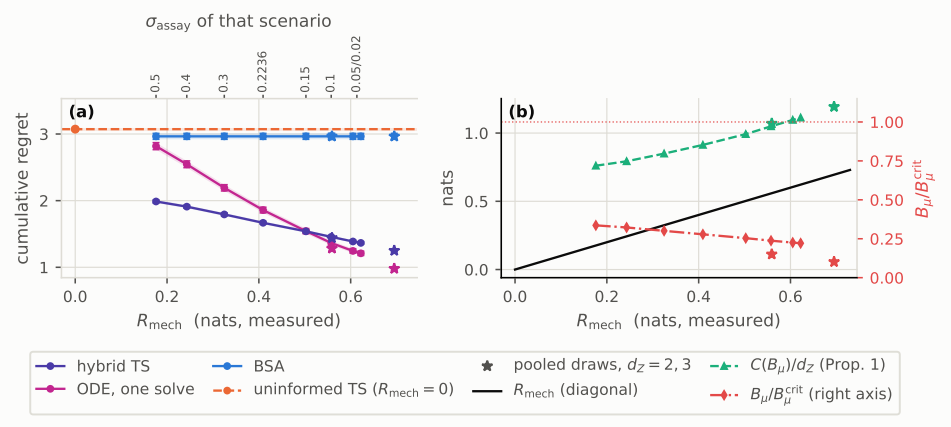

Figure 4: The channel sweep of Table 8: cumulative regret over the 11 scored cycles against the measured R_{mech} that the pre-dose channel supplies, at the correctly calibrated model. The two model-free references do not read the draw and are flat, while the model-based policies fall monotonically in R_{mech}, the one solve steeply enough to pass the hybrid between $R_{\text{mech}} = 0.42$ and 0.56 nats.

localised the arm, and BSA, which never adapts, accrues linearly. Only $N \leq 12$ is a clinical course, and longer horizons are reported for the scaling and have no per-patient reading.

This axis is also where the hybrid loses, and the table states it. The one solve accrues linearly at 0.167 per cycle, since it plays one fixed arm, so it is *below* the hybrid at the short horizons: 0.50 against 0.61 at $N = 3$ (ratio $0.82\,[0.81, 0.84]$) and 1.00 against 1.08 at $N = 6$ (0.93), with the cumulative crossover at $N = 8$ and the hybrid ahead by $1.11\times$ at the reported $N = 11$, $1.39\times$ at 20 and $2.13\times$ at 50. The mechanism is Theorem 2's sampling step. The hybrid draws its arm from $\mu_{\mathrm{hyb}}(\cdot \mid \hat{\pi})$ rather than playing that row's mode, so it plays $\hat{\pi}$ itself with probability 0.360 in cycle 1 and pays 0.225 in that cycle against the one solve's 0.167. It is ahead on a per-cycle basis only from cycle 5, by which point the posterior has concentrated on a better arm than $\hat{\pi}$ and stays there. A clinic with three cycles to spend should therefore play the model and not the posterior, which is the finite-N caveat that Proposition 2 and this table together draw.

N	$\mathrm{TS}_{\mathrm{hyb}}$	$\mathrm{TS}_{\mathrm{uninf}}$	BSA	ODE, once	$\frac{\mathrm{once}}{\mathrm{TS}_{\mathrm{hyb}}}$	$\frac{\mathrm{TS}_{\mathrm{uninf}}}{\mathrm{TS}_{\mathrm{hyb}}}$
3	0.61 ± 0.01	0.96 ± 0.01	0.81 ± 0.01	0.50 ± 0.01	$0.82\,[0.81, 0.84]$	$1.57\,[1.54, 1.60]$
6	1.08 ± 0.01	1.81 ± 0.01	1.62 ± 0.02	1.00 ± 0.02	$0.93\,[0.91, 0.94]$	$1.67\,[1.65, 1.70]$
11	1.66 ± 0.02	3.07 ± 0.02	2.97 ± 0.04	1.83 ± 0.04	$1.11\,[1.08, 1.13]$	$1.85\,[1.82, 1.88]$
20	2.40 ± 0.03	4.98 ± 0.03	5.41 ± 0.08	3.33 ± 0.08	$1.39\,[1.36, 1.42]$	$2.08\,[2.05, 2.11]$
50	3.92 ± 0.05	9.63 ± 0.06	13.52 ± 0.20	8.34 ± 0.19	$2.13\,[2.08, 2.18]$	$2.46\,[2.42, 2.49]$

Table 9: Scored-cycle scaling at the calibrated model ($\gamma = 1$): cumulative regret (mean $\pm$ 95% CI over $n = 10{,}000$ evaluation patients) and ratios of means with paired-bootstrap 95% intervals. The certificate is horizon-free: $H_{\mathrm{mech}} = 1.617$ nats and Theorem 1's floor 1.12 in every row. The one solve is linear in N at 0.167 per cycle and is below the hybrid at $N = 3$ and $N = 6$, the cumulative crossover falling at $N = 8$. The per-cycle controller is omitted here and reported in Table 6. The biased counterpart of this axis is the N block of Table 13.

H.9 Sensitivity analysis

Every choice made above is swept with the remaining configuration held fixed. Figure 5 shows the certificate across working points and under miscalibration.

Miscalibration (the axis that matters clinically). The surrogate's population clearance is re-fitted on a slice of the calibration study (the slowest- or fastest-clearing $80/60/40/20\%$) and the whole protocol re-run (Table 10). The measured bias crosses the one-cycle threshold nowhere. The measured R_{mech} falls exactly as the shrinking $C(B_\mu)$ predicts, and the chain $R_{\mathrm{mech}} \leq C_\nu \leq C(B_\mu)$ holds in every slice. Note that C_ν is constant by construction (it depends only on the channel, not on the deployed model), so the bias form $C(B_\mu)$ is the bound that tracks miscalibration.

calibration slice	γ	B_μ	$C(B_\mu)$	R_{mech}	chain	$B_\mu < B_\mu^{\mathrm{crit}\,\text{1-cycle}}$ (not ruled out)
slowest-clearing 20%	0.593	0.304	0.676	0.306	yes	yes
slowest-clearing 40%	0.702	0.276	0.750	0.364	yes	yes
slowest-clearing 60%	0.792	0.250	0.827	0.396	yes	yes
slowest-clearing 80%	0.879	0.231	0.893	0.410	yes	yes
everyone (correct model)	1.000	0.221	0.929	0.417	yes	yes
fastest-clearing 80%	1.140	0.232	0.889	0.408	yes	yes
fastest-clearing 60%	1.266	0.252	0.821	0.387	yes	yes
fastest-clearing 40%	1.418	0.277	0.747	0.356	yes	yes
fastest-clearing 20%	1.673	0.303	0.677	0.303	yes	yes

Table 10: The certificate under miscalibration: γ = calibrated/true clearance, B_μ is the centred bias of Eq. (3), $C_\nu = 0.506$ throughout, and the threshold is $B_\mu^{\mathrm{crit}\,\text{1-cycle}} = 0.810$ ($s = 12/11$, the full 12-cycle course: one lead-in and eleven scored).

Discretization K. Table 11: σ_h, L_h, B_μ and $C(B_\mu)$ are essentially flat in K. The one-cycle threshold tightens slowly as the measured occupancy entropy $H(\mu)$ grows. The classification at the

working point is unchanged across $K \in \{4, \ldots, 16\}$, and the same flatness holds on a drifted model (Table 13).

K	$H(\mu)$	L_h	B_μ	$C(B_\mu)$	$B_\mu^{\text{crit 1-cycle}}$
4	1.374	0.737	0.223	0.947	1.046
6	1.767	0.763	0.221	0.929	0.880
8	2.034	0.765	0.221	0.929	0.810
10	2.228	0.769	0.221	0.925	0.764
12	2.385	0.767	0.221	0.923	0.732
16	2.624	0.767	0.221	0.927	0.692

Table 11: Sensitivity to the discretization choice ($s = 12/11$ for the one-cycle point).

Estimator stability. Table 12: the slope underlying L_h is insensitive to the finite-difference step, and every constant is stable across independent calibration-study draws.

perturbation step	L_h	$C(B_\mu)$
0.005	0.7640	0.930
0.010	0.7642	0.930
0.020	0.7647	0.929
0.040	0.7659	0.928

seed	σ_h	L_h	B_μ	R_{mech}	$C(B_\mu)$
2024	0.3935	0.7647	0.2211	0.417	0.929
2025	0.3989	0.7643	0.2136	0.442	0.971
2047	0.4181	0.7625	0.2229	0.428	0.977
2101	0.4092	0.7735	0.2276	0.407	0.928

Table 12: Left: finite-difference step for the L_h slope. Right: independent calibration-study draws ($n = 2{,}000$ each).

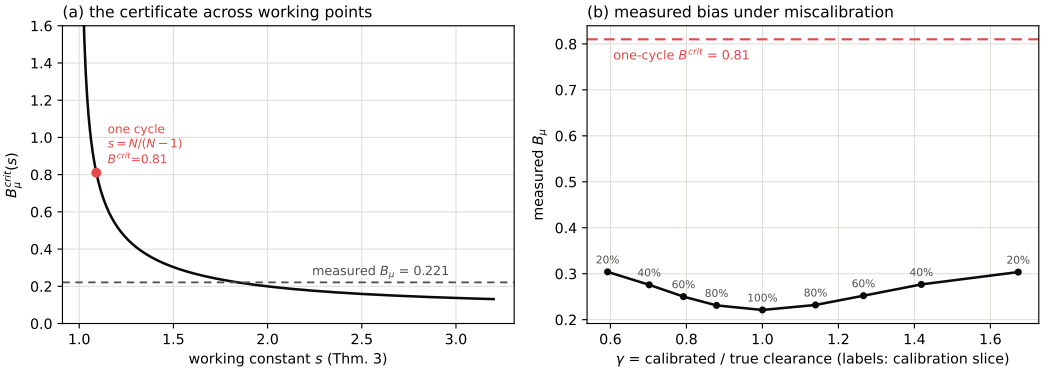

Figure 5: (a) Theorem 3's critical bias across the working constant s, with the one-cycle working point marked against the measured B_μ (grey dashed). (b) Measured B_μ under miscalibration (point labels: the calibration slice), against the one-cycle threshold.

H.10 Design-axis sensitivity of the certificate and the regret ordering

Protocol. Sections H.7–H.9 report one deployment. This subsection asks which of the trial designer's choices that deployment depends on, varying each *one at a time* with everything else pinned. Four plant axes are swept: the number of dose arms $K \in \{4, 6, 8, 12, 16\}$ on the fixed 1,600–3,600 mg/m^2 range, the scored horizon $N \in \{3, 6, 11, 20, 50\}$, the calibration-study size $n_{\text{study}} \in \{200, 500, 1000, 2000, 5000\}$ at three study seeds, and the number of lead-in cycles $\in \{1, 2, 3\}$: 28 configurations, each a complete re-run of the three-cohort protocol of Section H.6 with $n = 10{,}000$ evaluation patients, evaluation seed 11 and study seed 2024. The literature-calibrated constants σ_{assay}, σ_{cyc} and $\Omega^2(V_{\max})$ are *not* swept: they are the plant's declared population magnitudes (quoted where a published source exists, assumed otherwise).

The operating point is a *biased* deployment, and deliberately so. The mechanistic model is fitted on the slowest-clearing 40% of the calibration study ($\gamma \approx 0.702$, the slowest-clearing 40% slice of

Table 7), the regime in which a population model must be trusted despite drift. The hybrid's prior table $P(\pi^* \mid \hat{\pi})$ is fitted on the same slice the deployed model is calibrated on, so the study and the deployment share one population model. At $\gamma = 1$ the per-cycle mechanistic controller wins on every axis (Table 6). That is not the regime the paper argues about. Every certificate quantity below (R_{mech}, B_μ, $C(B_\mu)$, B_μ^{crit}, C_ν) is the deployed, biased model's, measured on the *study* cohort, so it is checkable before an evaluation patient is treated. The five regret columns and the ratios are the disjoint evaluation cohort. Intervals are 95% confidence intervals on the mean over that cohort (paired-bootstrap for ratios). On the n_{study} axis the reported band is the min–max over the three study seeds, with each run's own interval as the cap. Nothing is smoothed: every row is one Monte-Carlo run.

One convention must be stated because the lead-in axis exercises it. The implementation carries the prefactor $d_Z/2$ in $C(B_\mu)$ and divides Theorem 3's target by d_Z, both linear in d_Z, while σ_h and σ_ν keep their single-draw values. Both $C(B_\mu)$ and B_μ^{crit} therefore grow with the number of lead-in draws for reasons of accounting rather than of information, and the comparable quantity across those rows is the per-draw ceiling $C(B_\mu)/d_Z$: 0.750, 0.722, 0.713 nats. Nothing at $d_Z > 1$, certificate or robustness line, should be read without re-deriving the noise constants for the pooled draw.

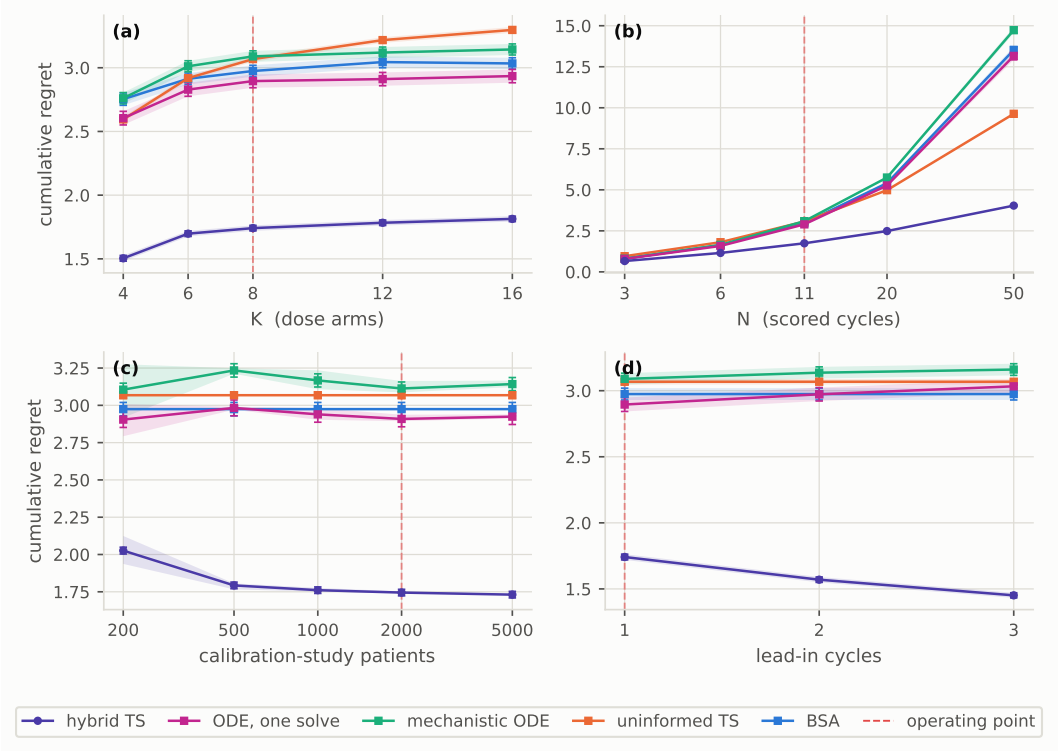

Figure 6: Cumulative regret of the five policies as each design choice is varied alone around the operating point: **(a)** dose arms K on the fixed 1,600–3,600 mg/m^2 range, **(b)** scored horizon N, **(c)** calibration-study size, **(d)** lead-in cycles. Everything not named in a panel is held at the operating point ($n = 10{,}000$ evaluation patients, $n_{\text{study}} = 2{,}000$, 11 scored cycles, seeds 11/2024). The literature-calibrated plant constants are never varied. The deployed mechanistic model is *biased*, its population clearance fitted on the slowest-clearing 40% of the study ($\gamma = 0.702$), and the hybrid's prior table is fitted on that same slice. Bands and capped bars are 95% confidence intervals on the mean, and on (c), which runs three study seeds per size, the band is the min–max over seeds. The dashed rule marks the operating point. The hybrid has the lowest regret in all 28 configurations, the ODE solved once on the drifted model is above it in all 28 and below the per-cycle controller in all 28, and the per-cycle controller, which re-asserts the drifted model every cycle, sits at or above BSA almost everywhere.

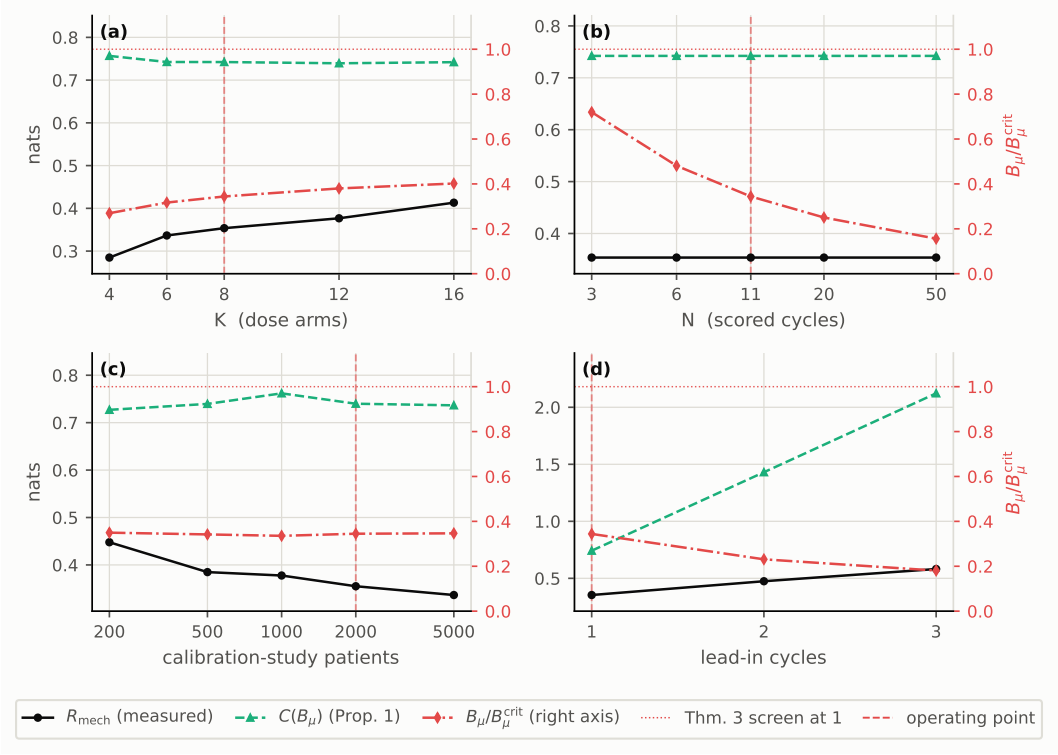

Figure 7: The pre-trial certificate along the same four axes. Left axis, in nats: the measured mechanistic information $R_{\text{mech}} = I(\pi^*; \hat{\pi})$ and Proposition 1's bias-form ceiling $C(B_\mu) = \frac{d_Z}{2}\log\big(1 + \sigma_h^2/(L_h^2 B_\mu^2)\big)$. The certificate is that the former lies below the latter. Right axis: the measured bias normalised by Theorem 3's threshold, $B_\mu/B_\mu^{\text{crit}}$, against the dotted screen at 1. The theorem is a rule-out, so passing is necessary and not sufficient. All quantities are the deployed (biased) model's, measured on the study cohort, hence checkable before any evaluation patient is treated. Both statements hold in all configurations. The screen is tightest at $N = 3$ (margin $1.40\times$), where saving one cycle in three is the most demanding target a model can be set. Panel (d) alone exercises the d_Z convention: $C(B_\mu)$ and B_μ^{crit} carry a factor linear in d_Z while σ_h and σ_ν stay single-draw, so the rise of the ceiling there is that scaling. Per draw it is 0.750, 0.722, 0.713 nats.

axis	value	R_{mech}	$C(B_\mu)$	B_μ	B_μ^{crit}	margin	hybrid	once	uninf.	BSA	mech.
K (arms)	4	0.290	0.765	0.279	1.046	3.75	1.504	2.604	2.595	2.754	2.761
	6	0.343	0.750	0.276	0.879	3.19	1.698	2.829	2.919	2.913	3.012
	8	0.364	0.750	0.276	0.810	2.94	1.741	2.895	3.068	2.975	3.089
	12	0.386	0.747	0.275	0.732	2.66	1.783	2.911	3.217	3.045	3.119
	16	0.417	0.749	0.275	0.692	2.52	1.814	2.935	3.297	3.034	3.144
N (cycles)	3	0.364	0.750	0.276	0.387	**1.40**	0.661	0.790	0.955	0.811	0.804
	6	0.364	0.750	0.276	0.580	2.10	1.153	1.579	1.807	1.623	1.644
	11	0.364	0.750	0.276	0.810	2.94	1.741	2.895	3.068	2.975	3.089
	20	0.364	0.750	0.276	1.113	4.04	2.487	5.264	4.980	5.408	5.748
	50	0.364	0.750	0.276	1.786	6.47	4.038	13.161	9.632	13.521	14.733
n_{study}	200	0.416	0.746	0.287	0.838	2.92	2.017	2.912	3.068	2.975	3.122
	500	0.370	0.766	0.287	0.869	3.03	1.806	2.971	3.068	2.975	3.212
	1000	0.373	0.743	0.285	0.830	2.91	1.752	2.934	3.068	2.975	3.158
	2000	0.364	0.750	0.276	0.810	2.94	1.741	2.895	3.068	2.975	3.089
	5000	0.343	0.741	0.283	0.822	2.90	1.728	2.912	3.068	2.975	3.122
lead-in	**1**	0.364	0.750	0.276	0.810	2.94	1.741	2.895	3.068	2.975	3.089
	2	0.479	1.444†	0.286	1.250†	4.37	1.570	2.973	3.068	2.975	3.136
	3	0.585	2.140†	0.289	1.613†	5.58	1.451	3.033	3.068	2.975	3.160

Table 13: Design-axis sweep at the biased operating point ($\gamma = 0.702$, bold = the operating point, repeated once per axis). R_{mech}, $C(B_\mu)$, B_μ and B_μ^{crit} are measured on the study cohort in nats, margin $= B_\mu^{\mathrm{crit}}/B_\mu$, and the five regret columns (hybrid, ODE solved once, uninformed TS, BSA, per-cycle mechanistic) are cumulative regret over the scored cycles on the disjoint evaluation cohort ($n = 10{,}000$), 95% intervals in the repository output. The certificate holds in every row: $B_\mu < B_\mu^{\mathrm{crit}}$ and $R_{\mathrm{mech}} \le C(B_\mu)$ in $28/28$, as does $R_{\mathrm{mech}} \le C_\nu$. The n_{study} rows are study seed 2024. The other two seeds (1729, 8675) are in the repository output and agree to about 0.05 from $n_{\mathrm{study}} = 500$ upward. † raw values at $d_Z =$ lead-in. Per draw the ceiling is 0.750, 0.722, 0.713. Theorem 1's floor $\sqrt{H(\mu)/H_{\mathrm{mech}}}$ is exceeded in every row, by $1.25\times$–$2.16\times$.

Findings. The hybrid has the lowest regret in all 28 configurations (BSA/hybrid $1.23\times$–$3.35\times$, uninformed/hybrid $1.44\times$–$2.39\times$, mechanistic/hybrid $1.22\times$–$3.65\times$), and under bias the ordering of Table 6 reverses: the per-cycle mechanistic controller is at or above BSA in 26 of 28 configurations, the exceptions being $N = 3$ and one $n_{\text{study}} = 200$ seed, each within 0.1 of BSA, against $0.44\times$ the hybrid's regret at $\gamma = 1$. The ODE solved once on the same drifted model is never below the hybrid here, in 0 of 28 configurations, at $1.19\times$–$3.26\times$ its regret with the tightest ratio at $N = 3$, and it is below the per-cycle controller in 28 of 28: the short-horizon reversal of Table 9 does not survive the bias, because at $N = 3$ the open-loop policy plays a drifted arm three times while the hybrid's first sampled arm is drawn from a table that already accounts for how often the model's recommendation is right. Bias and horizon push in opposite directions on the one comparison the main text flags as close. The N axis shows why: nothing about the model depends on how long the course is scored, so R_{mech}, $C(B_\mu)$ and B_μ are identical in all five rows and only the target moves, yet the controller grows from 0.804 to 14.73 while the hybrid grows to 4.04. A wrong prior is paid for once. A model re-asserted every cycle re-asserts the drift. The certificate holds throughout, and it is tightest exactly where the theory says it should be: at $N = 3$ the one-cycle target $s = N_{\text{course}}/(N_{\text{course}} - 1)$ is at its most demanding and the margin falls to $1.40\times$, the closest this sweep comes to a screen failure. B_μ and $C(B_\mu)$ are near-constant across the whole sweep (0.270–0.288, 0.702–0.791), reproducing Table 11's K-flatness on a drifted model: the measurement channel is a property of the model and the assay, not of the design. Two axes carry a caution. At $n_{\text{study}} = 200$ the plug-in mutual-information estimator inflates R_{mech} by roughly 0.1 nat over the value the large studies converge to (0.416, 0.458, 0.494 against ≈ 0.35), the order of the Miller–Madow correction $(K - 1)^2/2n \approx 0.12$ nats at $n = 200$, $K = 8$, and hybrid regret degrades to 1.94–2.12: a 200-patient conditional table is about 25 patients per arm. On the lead-in axis, extra draws help under bias, as they do at the calibrated point ($1.741 \to 1.451$ against $1.66 \to 1.24$), even though B_μ rises rather than falls, $0.276 \to 0.289$: the information the extra draws add outweight the bias they sharpen. Finally, Theorem 1's floor is exceeded in every configuration, by $1.25\times$ to $2.16\times$, and is met nowhere, as a minimax floor should be.