Data to Dose: Efficient Synthetic Data Generation with Expert Guidance for Personalized Dosing

H. Satyam Verma[1] **Holly Wiberg**[2] **Shixiang Zhu**[2] **Sridhar R. Tayur**[1]
[1]Tepper School of Business [2]Heinz College
Carnegie Mellon University
hverma@andrew.cmu.edu hwiberg@cmu.edu shixianz@cmu.edu stayur@cmu.edu

Abstract

Effective personalized dosing is constrained by limited data, noisy outcomes, and heterogeneous patient profiles. We introduce GenEx, a hybrid framework that couples preference-based Bayesian optimization with a generative model fine-tuned via expert feedback. Expert pairwise rankings update the Gaussian Process (GP) surrogate and, via an expert–generator agreement test, gate uncertainty-guided synthetic data injection at the GP's highest-variance doses. We provide guarantees of sublinear expected regret under decaying generator bias and validate them in numerical studies. We find that GenEx converges faster and improves decision quality over preference-only and synthetic-only baselines, enabling safer and more effective individualized treatment.

1 Introduction

Current dosing practice is largely protocol-driven, often overlooking patient heterogeneity and leading to avoidable harm and cost from adverse drug events [Hamburg and Collins, 2010, World Health Organization, 2019, Shehab et al., 2016]. The dilemma is most acute in narrow therapeutic windows such as tacrolimus after kidney transplantation, where underdosing risks rejection and overdosing induces toxicity [Lattimore et al., 2024]. In these settings, data are scarce or noisy; rare profiles are underrepresented; and black-box predictors face interpretability and regulatory barriers [Rudin, 2019]. Synthetic augmentation can expand coverage but, unguided, risks implausible or biased samples [Chen et al., 2021, Goncalves et al., 2020]. A practical alternative is to incorporate human expertise: while full numeric outcomes are burdensome to elicit, pairwise preferences provide a lightweight signal that captures nuanced benefit–risk trade-offs [Jang et al., 2022].

We seek a data-to-dose process that learns under scarcity, respects uncertainty, and uses expert signal where it is most informative. This drives the design of GenEx. A Gaussian-process surrogate proposes candidate doses from its posterior; the clinician ranks them; the surrogate updates through a Bradley–Terry likelihood. In parallel, a conditional generator is treated not as a bulk offline synthesizer but as a controllable instrument: a simple agreement test compares the generator's ranking with the expert's on the same candidates. When they align, the generator contributes targeted patient–dose–goal triplets in regions where the surrogate is uncertain; when they do not, the generator is refined using preference supervision while preserving fidelity via replay. This expert–generator feedback loop turns limited preferences into uncertainty-aware exploration and confines synthetic influence to contexts the expert implicitly endorses. The procedure remains offline-friendly for clinical workflows and modular: stronger surrogates or generators can be swapped in without altering the logic of augmentation. Conceptually, the surrogate is no longer a passive estimator but a novel, co-designed agent that shapes its own learning landscape: expert preferences bend the posterior where decisions matter, and aligned synthesis expands support exactly there. This yields two payoffs

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

developed in the paper: a sublinear expected-regret guarantee when synthetic bias decays under supervision, and empirical improvements over expert-only and synthetic-only baselines across data and expert-quality regimes.

2 Problem statement and methodology

Problem setup. Let $\mathbf{x} \in \mathbb{R}^p$ denote patient covariates, $\mathbf{d} \in \mathbb{R}^m$ a dose vector, and $G(\mathbf{x}, \mathbf{d})$ a black-box goal balancing efficacy and toxicity. The personalized optimum is

$$\mathbf{d}^*(\mathbf{x}) = \arg\max_{\mathbf{d}} G(\mathbf{x}, \mathbf{d}).$$

We place a Gaussian process prior on G, yielding posterior mean μ and variance σ^2 once conditioned on data. In parallel, a conditional variational autoencoder (cVAE) serves as the generator, modeling $p_\theta(\mathbf{x}, g \mid \mathbf{z}, \mathbf{d})$ with encoder $q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{d}, g)$ and pretrained via the standard ELBO.

Methodology. The GenEx procedure unfolds in four stages (Fig. 1).

(i) Seed training. A seed dataset $\mathcal{D}_0 = \{(\mathbf{x}_i, \mathbf{d}_i, g_i)\}_{i=1}^{n_0}$ initializes both the GP surrogate and the cVAE generator, ensuring calibrated uncertainty and a faithful latent representation.

(ii) Expert querying. For each refinement patient $\mathbf{x}_t$, the GP proposes L candidates via Thompson sampling. The expert ranks them, inducing $\binom{L}{2}$ pairwise preferences, and the GP updates through a Bradley–Terry likelihood:

$$\mathcal{L}_{\mathrm{BT}}^{\mathrm{GP}} = -\sum_{(\mathbf{d} \succ \mathbf{d}') \in \mathcal{P}_t} \log \sigma\big(G(\mathbf{x}_t, \mathbf{d}) - G(\mathbf{x}_t, \mathbf{d}')\big), \quad \sigma(z) = \tfrac{1}{1+e^{-z}}.$$

(iii) Generator refinement. On the same candidate set, the generator produces scores $\widehat{G}_\phi(\mathbf{x}_t, \mathbf{d})$ and a ranking r_t^{gen}. Agreement with the expert ranking r_t^{exp} is measured by the normalized Kendall distance η_t, smoothed over a window W. If $\bar{\eta}_t > \tau$, the generator is refined using a replayed ELBO plus a margin-ranking loss:

$$\mathcal{L}_{\mathrm{gen}} = \mathcal{L}_{\mathrm{ELBO}}(\phi, \theta; \mathcal{B}) + \alpha \sum_{(\mathbf{d} \succ \mathbf{d}') \in \mathcal{P}_W} \max\Big\{0,\, \delta - \big[\widehat{G}_\phi(\mathbf{x}, \mathbf{d}) - \widehat{G}_\phi(\mathbf{x}, \mathbf{d}')\big]\Big\}.$$

(iv) Synthetic augmentation. When $\bar{\eta}_t \le \tau$, the generator is deemed trustworthy. Sobol-sampled doses are filtered by GP variance, decoded into synthetic triplets $\mathcal{U}$, and incorporated into the GP with a down-weighted likelihood:

$$\mathcal{L}_{\mathrm{GP}}(\mathcal{U}) = -\lambda \sum_{(\tilde{\mathbf{x}}, \mathbf{d}, \tilde{g}) \in \mathcal{U}} \log \mathcal{N}\big(\tilde{g} \,\big|\, \mu(\tilde{\mathbf{x}}, \mathbf{d}), \sigma^2(\tilde{\mathbf{x}}, \mathbf{d})\big).$$

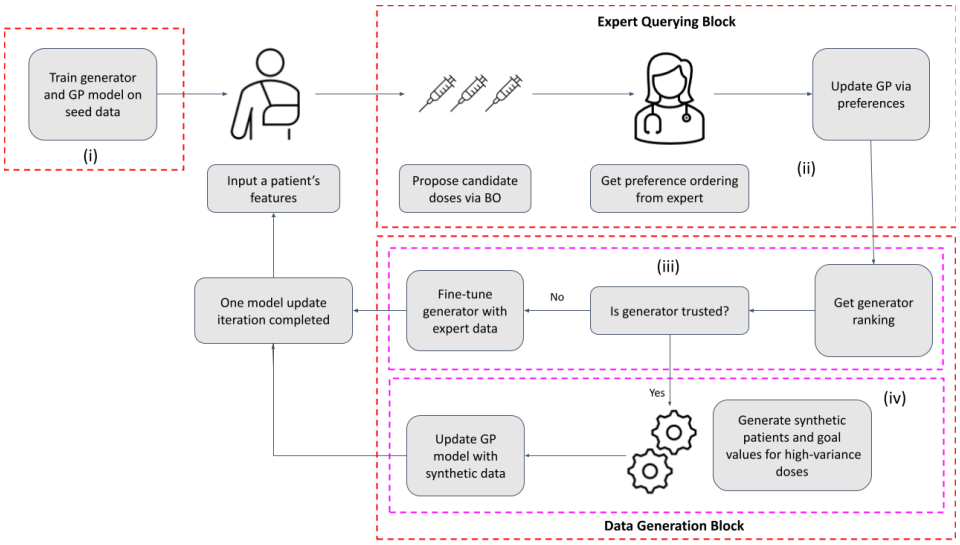
Together, these stages define an iterative loop: expert feedback shapes the surrogate, generator alignment determines whether augmentation or further refinement occurs, and synthetic samples extend the training set exactly where uncertainty is highest. This process converts sparse preferences into informed exploration while constraining synthetic influence to expert-endorsed regions.

3 Results

We assess GenEx from two angles. First, we formalize how alignment-gated augmentation interacts with Bayesian optimization by deriving a regret bound that isolates the effect of synthetic bias. Second, we test GenEx in a controlled dosing environment that captures patient heterogeneity, nonlinear responses, and label scarcity, varying data availability and expert quality.

Expected cumulative regret. Let $\tilde{G}_t(\mathbf{x}, \mathbf{d}) = G(\mathbf{x}, \mathbf{d}) + b_t(\mathbf{x}, \mathbf{d})$ where b_t encodes bias due to synthetic data. Assume: (i) G and b_t lie in the RKHS with bounded norms and b_t's norm decreases with more expert queries, (ii) pointwise bias is bounded as $|b_t| \le B(N_t)$ with $B(N_t) \downarrow 0$, (iii) observations have Gaussian noise, (iv) the input domain is compact with bounded kernel variance, and (v) the maximum information gain γ_T is finite. Under these conditions, the following holds.

Figure 1: Overview of the GenEx framework.



The process consists of four stages: (i) seed training of GP and cVAE; (ii) expert querying with GP-proposed candidates and preference-based GP update; (iii) generator refinement when rankings diverge from expert; (iv) synthetic data generation in uncertain regions when the generator is trusted.

Theorem 1 (Expected Cumulative Regret)**.** *If* $\mathbf{d}_t$ *is drawn by Thompson sampling from the GP posterior, then*

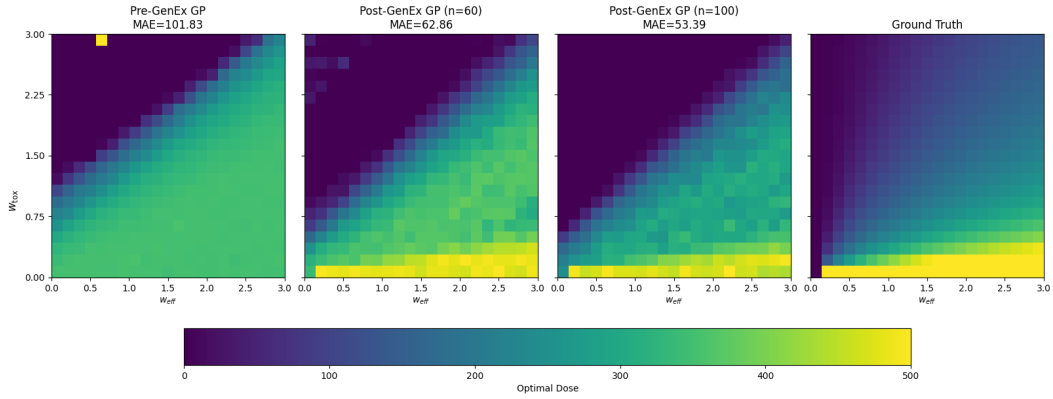
$$\mathbb{E}[R_T] \le 2\,\beta_T^{1/2}\sqrt{T\gamma_T} + 2\sum_{t=1}^{T} B(N_{t-1}), \qquad \beta_T = \max_{1\le t\le T} 2C_k^2 B_{\tilde{G}}(N_{t-1})^2\big(\gamma_{t-1} + 1 + 3\log T\big),$$

with $C_k = \sup_{(\mathbf{x},\mathbf{d})} \sqrt{k((\mathbf{x},\mathbf{d}),(\mathbf{x},\mathbf{d}))}$*. If* $\gamma_T = o(T)$ *and* $\sum_t B(N_{t-1}) < \infty$*, then* $\mathbb{E}[R_T]/T \to 0$*.*

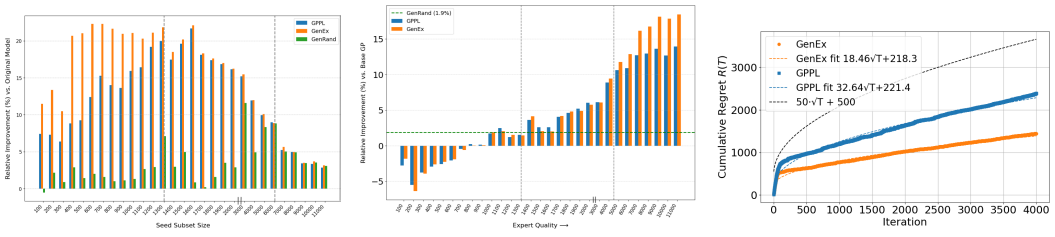
Regret separates into two components: an uncertainty term of order $\mathcal{O}(\sqrt{T\gamma_T})$ and an additive synthetic-bias term $\sum_t B(N_{t-1})$. As expert supervision drives $B(N_t)\downarrow 0$, the bias term vanishes, leaving the standard Bayesian optimization rate. Under the usual kernel condition $\gamma_T = o(T)$, the bound implies sublinear cumulative regret and hence no-regret convergence, i.e., $\mathbb{E}[R_T]/T \to 0$. Intuitively, GenEx benefits from early alignment-gated augmentation to accelerate exploration, while long-run behavior remains controlled and asymptotically unbiased.

Empirical evaluation. We use a clinically motivated simulator with ten covariates and a scalar dose. The ground-truth objective $G(\mathbf{x}, \mathbf{d})$ trades sigmoidal efficacy against quadratic toxicity. A GP surrogate and cVAE generator are first fit on a seed set; refinement then proceeds via expert rankings and agreement-gated augmentation. Unless noted: cVAE refinement weight α=1 (losses normalized to a common scale), trust window W=5 with threshold τ=0.10. A perfect expert ranks candidates using the true G. An imperfect expert is a GP trained on labeled triplets of varying size (a proxy for clinician experience). Baselines: (i) Base GP with no refinement, (ii) GPPL using a Bradley–Terry preference likelihood with no synthetic data, (iii) GenRand performing uncertainty-guided cVAE augmentation with no preferences. A multi-drug extension and additional diagnostics are provided in the full technical paper [Verma et al., 2025].

Panel (a): learned dosing surfaces. Starting from n_0=700 and refining with a perfect expert for $T \in \{60, 100\}$ rounds, we visualize the optimal dose over a grid of utility weights $(w_{\text{eff}}, w_{\text{tox}})$ for a held-out patient. GenEx progressively calibrates the surface: MAE improves from 101.83 (pre-GenEx) to 62.86 (60 rounds) and 53.39 (100 rounds). The post-GenEx maps recover the expected frontier, with lower doses where toxicity dominates and higher doses as the efficacy weight rises, while eliminating spurious extremes seen in the seed-only model. **Panel (b): perfect-expert comparison across data regimes.** With T=60 and L=5, we compare GenEx to GPPL (preferences only) and GenRand (augmentation only) across nested seed sizes n_0 from 100 to 10,000. Three



(a) Optimal-dose heatmaps over $(w_{\text{eff}}, w_{\text{tox}})$ for a held-out patient: pre-GenEx, post-GenEx (60), post-GenEx (100), and ground truth. MAE $101.83 \to 62.86 \to 53.39$.



(b) Perfect expert: improvement across nested n_0 (T=60, L=5).

(c) Imperfect experts: sensitivity to expert quality (n_0=700, L=10).

(d) Regret over T=4000 rounds (n_0=700, L=5).

Figure 2: GenEx at a glance. Top: (a) recovery and calibration of the dosing surface. Bottom: (b) behavior across data regimes with a perfect expert, (c) robustness to expert quality, and (d) learning dynamics via regret.

regimes emerge. Scarce data ($\lesssim$1,300): GenEx leads, since early preferences steer the generator and aligned synthesis adds coverage where uncertainty is largest; GPPL helps but explores less, and GenRand lags due to weak pretraining. Mid-range (about 1,300 to 6,000): GenEx $\approx$ GPPL, because once the surrogate prior is sensible, preferences alone deliver most gains and augmentation has diminishing returns. High data ($\gtrsim$6,000): methods converge as both GP and cVAE are already well calibrated. **Panel (c): robustness to expert quality.** Fixing n_0=700 and using L=10 to stress ranking, we vary expert quality by training imperfect experts on 100 to 10,000 labeled triplets (60 rounds). We observe three phases again. Weak experts (up to about 1,300): GPPL and GenEx can underperform the base model, making performance even worse than the seed model if the expert surrogate is trained on less points, while GenRand remains stable by ignoring supervision. Moderate experts (about 1,400 to 5,000): GPPL and GenEx both surpass GenRand as preferences become informative. Strong experts (at least about 5,000): GenEx outperforms GPPL using the same number of queries by turning reliable rankings into targeted, high-uncertainty synthesis that amplifies the expert signal, thus squeezing out more information from the expert feedback while the GPPL begins to plateau. **Panel (d): cumulative regret.** With n_0=700, a perfect expert, L=5, and T=4000, we plot $R(T)$. A fit $a\sqrt{T}+b$ gives GenEx $(18.36,\ 218.3)$ vs. GPPL $(32.64,\ 221.4)$, implying a lower asymptotic rate and curve for GenEx. The gap is consistent with sublinear growth and no visible bias accumulation under agreement-gated synthesis.

4 Conclusion

We address the challenge of learning dosing policies from scarce, preference-only feedback by introducing GenEx, which augments GP surrogates with generator-based synthesis gated by expert alignment. Under standard smoothness and bounded bias, the method achieves sublinear regret. Simulations show that GenEx sharpens dosing surfaces, outperforms expert-only and synthetic-only baselines in low-data regimes, and maintains competitive performance as data scale.

Acknowledgments

We thank Amit D. Tevar, MD, FACS (Professor of Surgery and Director, Kidney and Pancreas Transplant Program, Thomas E. Starzl Transplantation Institute, UPMC) for valuable discussions that helped guide our simulated data design.

Funding. This work was supported in part by the Center for Intelligent Business (CIB), Tepper School of Business, Generative AI Fellows Program (2024).

References

Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6): 493–497, 2021.

Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. Generation and evaluation of synthetic patient data. *BMC medical research methodology*, 20:1–40, 2020.

Margaret A. Hamburg and Francis S. Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

Ikbeom Jang, Garrison Danley, Ken Chang, and Jayashree Kalpathy-Cramer. Decreasing annotation burden of pairwise comparisons with human-in-the-loop sorting: Application in medical image artifact rating. *arXiv preprint arXiv:2202.04823*, 2022.

Sherene Lattimore, Anastasia Chambers, Isabella Angeli-Pahim, Abhishek Shrestha, Benjamin O Eke, Ariel Pomputius, Carma Bylund, Megan E Gregory, and Ali Zarrinpar. Impact of intrapatient immunosuppression variability in liver transplantation outcomes: A systematic review and meta-analysis. *Transplantation Direct*, 10(9):e1700, 2024.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Nadine Shehab, Maribeth C Lovegrove, Andrew I Geller, Kathleen O Rose, Nina J Weidle, and Daniel S Budnitz. Us emergency department visits for outpatient adverse drug events, 2013-2014. *Jama*, 316(20):2115–2125, 2016.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. 2010. ISBN 9781605589077.

H Satyam Verma, Holly Wiberg, Shixiang Zhu, and Sridhar R Tayur. Data to dose: Efficient synthetic data generation with expert guidance for personalized dosing. *SSRN Electronic Journal*, 2025. URL `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5400895`. Available at SSRN: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5400895`.

World Health Organization. Medication safety in polypharmacy: A who technical report. 2019.

A Simulated Dataset Generation

Our simulated dataset is inspired by a real-world collaboration on dose personalization in kidney transplantation, but it is not specific to that domain. Rather, it abstracts and generalizes the core challenges of clinical dosing—nonlinear response dynamics, inter-patient variability, and trade-offs between efficacy and toxicity—across broader therapeutic settings. While tacrolimus motivates the structure, the magnitudes and transformations below are synthetic and chosen to support controlled experiments; they are not intended to mirror clinical dosing practice.

Feature Construction. We construct patient features to imitate clinical heterogeneity. Although column names reference familiar biomarkers, they are not tied to a specific disease. Continuous variables:

- **BMI** (body mass index; proxy for body size/adiposity affecting distribution and dosing) $\sim \mathcal{U}(18, 40)$
- **Age** (years; age-related physiology and toxicity risk) $\sim \mathcal{U}(18, 90)$
- **Creatinine** (renal function marker; higher indicates reduced clearance) $\sim \mathcal{U}(0.5, 2.0)$
- **WBC** (baseline immune/inflammation proxy) $\sim \mathcal{U}(4.0, 11.0)$
- **Bilirubin** (hepatic excretory function marker) $\sim \mathcal{U}(0.1, 3.0)$
- **Liver_Enzymes** (ALT/AST surrogate; hepatocellular stress) $\sim \mathcal{U}(10, 120)$
- **Height** (meters; used with BMI to approximate body mass) $\sim \mathcal{U}(1.5, 1.9)$

Categorical/ordinal variables:

- **SmokingStatus** $\in \{0, 1, 2\}$ (never, former, current) with probabilities $(0.4, 0.3, 0.3)$
- **ComorbidityScore** $\in \{0, \ldots, 10\}$ (integer; overall burden of chronic conditions)

Let $\mathbf{x}$ collect these features.

Dose Assignment. We simulate a single-drug regimen, so $\mathbf{d}$ reduces to a scalar dose $d \in \mathbb{R}_+$ (units: mg). A latent *Size–Health Index* combines size and health status:

$$S(\mathbf{x}) = \text{BMI} \cdot \text{Height}^2 + 0.1 \cdot (\text{Age} - 50) - 0.05 \cdot \text{ComorbidityScore}.$$

This index is a synthetic, non-clinical heuristic introduced solely to add an extra layer of heterogeneity and complexity to the simulated dose–response mapping. The nominal (noise-free) dose applies a nonlinear transformation with saturation,

$$\bar{d} = 4 \cdot S(\mathbf{x}) + \frac{50}{1 + \exp\big(-0.05 \cdot (S(\mathbf{x}) - 70)\big)},$$

and the realized dose adds zero-mean Gaussian noise scaled to $\bar{d}$:

$$d = \bar{d} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\big(0,\, 0.05 \cdot \bar{d}\big).$$

Efficacy function. Efficacy is denoted $E(\mathbf{x}, d)$ and follows a Hill form with a mild logarithmic correction:

$$E(\mathbf{x}, d) = \frac{E_{\max}\, d^{\,h}}{\text{EC}_{50}^{h} + d^{\,h}} + \log\big(1 + d/100\big) + \epsilon_1, \quad \epsilon_1 \sim \mathcal{N}(0, 1).$$

The parameters $E_{\max}, h$, and EC_{50} may vary by patient (e.g., as smooth functions of BMI, Age, WBC, and Creatinine) to introduce heterogeneity in dose–response curves.

Toxicity function. Toxicity is $T(\mathbf{x}, d)$, comprising a baseline term and a quadratic dose-dependent term modulated by BMI, Age, and a random subpopulation label $g \in \{0, 1, 2\}$:

$$\begin{aligned} T_{\text{base}}(\mathbf{x}) = {} & 5.0 + 0.8\,(\text{Creatinine} - 1.0)^2 + 0.2\,\log\big(1 + \text{Bilirubin}\big) \\ & + 0.05\,(\text{Liver\_Enzymes} - 40) + 0.3\,\text{SmokingStatus} + 0.1\,\text{ComorbidityScore}. \end{aligned}$$

The dose-dependent term uses one of three parameter sets depending on g, introduced solely to create additional heterogeneity in the simulated toxicity response:

$$\begin{aligned} g = 0: &\quad c_{\text{tox}} = 1.5 \times 10^{-4},\ \kappa = 0.03,\ \mu = 30,\ \lambda = 0.005, \\ g = 1: &\quad c_{\text{tox}} = 2.0 \times 10^{-4},\ \kappa = 0.04,\ \mu = 28,\ \lambda = 0.006, \\ g = 2: &\quad c_{\text{tox}} = 1.0 \times 10^{-4},\ \kappa = 0.02,\ \mu = 32,\ \lambda = 0.008. \end{aligned}$$

Thus

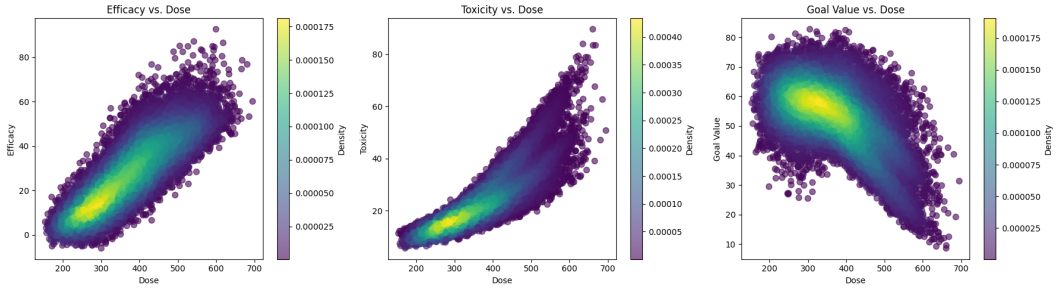
$$T(\mathbf{x}, d) = T_{\text{base}}(\mathbf{x}) + \frac{c_{\text{tox}}\, d^2}{1 + \exp\big(-\kappa\,(\text{BMI} - \mu)\big)}\,\big(1 + \lambda\,\text{Age}\big) + \epsilon_2, \quad \epsilon_2 \sim \mathcal{N}(0, 0.5).$$

Goal function. The scalar outcome is $g = G(\mathbf{x}, d)$, mapping benefit–risk to $[0, 100]$:

$$G(\mathbf{x}, d) = \frac{100}{1 + \exp\Big(-\big[w_{\text{eff}} \cdot \log\big(E(\mathbf{x}, d) + 10\big) - w_{\text{tox}} \cdot \sqrt{|T(\mathbf{x}, d)| + 1}\big]\Big)},$$

with default weights $w_{\text{eff}} = 1.0$ and $w_{\text{tox}} = 0.7$.

Figure 3: Dose–outcome relations across all patients.



Left: $E(\mathbf{x}, d)$ increases with d and saturates. **Center:** $T(\mathbf{x}, d)$ grows superlinearly, especially above $\sim$400 mg. **Right:** $g = G(\mathbf{x}, d)$ peaks at intermediate d. Plots are colored by data density.

Interpretation (Fig. 3). The left panel shows that $E(\mathbf{x}, d)$ increases with dose and then saturates, consistent with the Hill-form pharmacodynamics. Most observations cluster around mid-dose values, reflecting common efficacy patterns across patients, while sparser regions reveal variability in less-frequent dosing regimes. The center panel shows $T(\mathbf{x}, d)$ rising sharply with dose, especially above $\sim$400 mg, capturing the convex risk of adverse effects at higher exposure; moderate toxicity dominates common dose levels, but extreme values appear in low-density regions, indicating a natural upper dosing limit. The right panel displays $g = G(\mathbf{x}, d)$, which peaks at intermediate d; the density gradient highlights where optimal outcomes concentrate, with both under- and over-dosing associated with lower frequency and reduced scores.

Interpretation (Fig. 4). Each curve traces $G(\mathbf{x}, d)$ for a randomly sampled patient as d varies, with vertical dotted lines marking the patient-specific optimizer. The heterogeneity in curve shapes (from smooth, symmetric peaks to flatter or skewed profiles) and in the locations of their optima underscores the need for individualized dosing: a population-average dose is unlikely to perform well across all individuals.

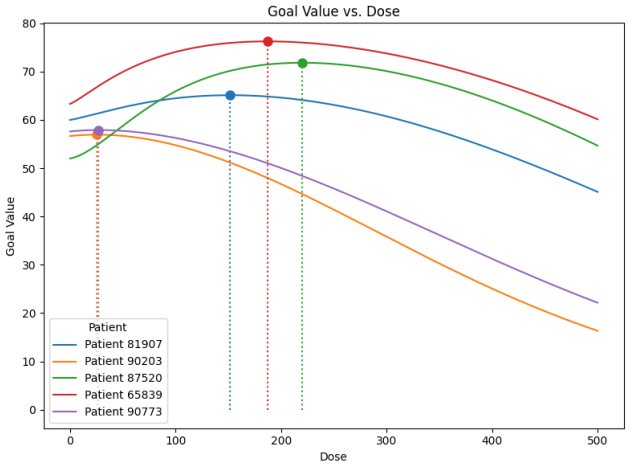
B Sparse Variational Inference for the GP Surrogate

We expand on the sparse variational Gaussian Process (GP) approximation used for the surrogate over the personalized objective $G(\mathbf{x}, \mathbf{d})$. Recall that we place a GP prior

$$G(\cdot) \sim \mathcal{GP}\big(m_\theta,\ k_\theta\big), \qquad k_\theta\big((\mathbf{x}, \mathbf{d}), (\mathbf{x}', \mathbf{d}')\big) \text{ positive-definite},$$

with a constant mean function m_θ and kernel k_θ. To scale inference, we adopt a sparse variational formulation with inducing variables.

Figure 4: Trajectories of $G(\mathbf{x}, d)$ for five patients.



Each line shows G versus d; the vertical dotted line marks the dose maximizing $G(\mathbf{x}, d)$ for that patient, highlighting individualized optima.

The GP surrogate must update cheaply at every round t while preserving calibrated uncertainty for Thompson sampling, expert ranking, and filtering of synthetic samples. Exact GP inference scales cubically in the number of labeled inputs and is therefore impractical as n_t grows. At the same time, our observation model alternates between scalar outcomes on the seed set $\mathcal{D}_0$ and a Bradley–Terry preference likelihood during refinement, which favors minibatched, stochastic optimization. A sparse variational formulation with inducing variables satisfies these requirements: it reduces training from cubic cost to $\mathcal{O}(n_t\, M_{\text{ind}}^2)$ while retaining coherent posterior mean/variance $\big(\mu(\mathbf{x}, \mathbf{d}), \sigma^2(\mathbf{x}, \mathbf{d})\big)$ and supports noisy, comparison-based updates via the ELBO.

Inducing variables and variational family. Let $\mathcal{D}_t$ denote the labeled data available by round t (including the seed set $\mathcal{D}_0$ and any subsequent labels/preferences), and let

$$\mathbf{g} = \big[G(\mathbf{x}_1, \mathbf{d}_1), \ldots, G(\mathbf{x}_{n_t}, \mathbf{d}_{n_t})\big]^\top$$

collect the latent function evaluations at the n_t unique training inputs seen up to round t. Introduce $M_{\text{ind}} \ll n_t$ inducing inputs

$$\mathbf{Z} = \{(\tilde{\mathbf{x}}_j, \tilde{\mathbf{d}}_j)\}_{j=1}^{M_{\text{ind}}}, \qquad \mathbf{u} = G(\mathbf{Z}),$$

and define a Gaussian variational distribution over the inducing outputs, $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$ with $\mathbf{S} = \mathbf{L}\mathbf{L}^\top$ to ensure positive definiteness. The variational posterior over $\mathbf{g}$ is obtained by marginalizing:

$$q(\mathbf{g}) = \int p(\mathbf{g} \mid \mathbf{u})\, q(\mathbf{u})\, d\mathbf{u}.$$

Objective and likelihood. We learn $(\mathbf{m}, \mathbf{S})$, the inducing locations $\mathbf{Z}$, and kernel hyperparameters θ by maximizing the evidence lower bound (ELBO)

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(\mathbf{g})}\big[\log p(\mathcal{D}_t \mid \mathbf{g})\big] - \text{KL}\big[q(\mathbf{u}) \,\|\, p(\mathbf{u})\big].$$

The observation model $p(\mathcal{D}_t \mid \mathbf{g})$ matches the main text: (i) when pretraining on the seed set $\mathcal{D}_0 = \{(\mathbf{x}_i, \mathbf{d}_i, g_i)\}_{i=1}^{n_0}$, we use a Gaussian likelihood on the scalar outcomes g_i; (ii) during expert-guided refinement, we use the Bradley–Terry preference likelihood from Sec. 2, where the expected log-likelihood term corresponds to the negative of $\mathcal{L}_{\text{BT}}^{\text{GP}}$:

$$\mathbb{E}_{q(\mathbf{g})}\left[\sum_{(\mathbf{d} \succ \mathbf{d}') \in \mathcal{P}_t} \log \sigma\big(G(\mathbf{x}_t, \mathbf{d}) - G(\mathbf{x}_t, \mathbf{d}')\big)\right].$$

The KL term is analytic; the expectation is estimated with Monte Carlo. We optimize with stochastic gradients over mini-batches of comparisons.

Complexity and predictions. With M_{ind} inducing points, training scales as $\mathcal{O}(n_t\, M_{\text{ind}}^2)$. From the variational posterior, we read off the GP predictive mean and variance at any $(\mathbf{x}, \mathbf{d})$, denoted $\mu(\mathbf{x}, \mathbf{d})$ and $\sigma^2(\mathbf{x}, \mathbf{d})$, which are used for Thompson sampling, expert querying, and filtering synthetic samples as described in the main text.

C Point Predictions and Ranking Permutations: Implementation Details

We use the cVAE specified in the main paper; this appendix details how point scores are computed for ranking and restates the agreement metric used for gating.

Generator and point predictions. The generator used in this framework is trained to reconstruct both the patient profile and the goal value given a dose. During training, the decoder learns to map latent representations and conditioning inputs to full triplets $(\mathbf{x}, \mathbf{d}, g)$, making joint generation of patient features and outcomes a natural output of the model.

When computing the generator's ranking over candidate doses for a given patient, however, the corresponding profile $\mathbf{x}_t$ is assumed to be known and is supplied explicitly as input. In this setting, only the predicted goal value is retained; the generated patient features are ignored, as they serve no downstream purpose. The encoder remains functional even when goal values are not observed, treating them as latent and integrating over the associated uncertainty when forming the posterior over $\mathbf{z}$.

Given a latent representation $\mathbf{z}$ drawn from this posterior, the model defines a point estimate of the goal via the decoder's conditional mean. The resulting function

$$\widehat{G}_\phi(\mathbf{x}, \mathbf{d}) = \mathbb{E}_{\mathbf{z}\sim q_\phi}\left[g \mid \mathbf{x}, \mathbf{d}\right]$$

serves as the generator's internal surrogate for the true goal and plays an analogous role to the GP surrogate G introduced earlier. While the full decoder output includes both $\hat{\mathbf{x}}$ and $\hat{g}$, only $\hat{g}$ is retained for ranking and comparison during the refinement process.

Finally, to get the generator's ranking for the L Thompson-sampled doses $\mathcal{C}_t = \{\mathbf{d}_t^{(1)}, \ldots, \mathbf{d}_t^{(L)}\}$ proposed for patient $\mathbf{x}_t$, we compute

$$\hat{g}_t^{(i)} \;=\; \underbrace{\mathbb{E}_{\mathbf{z}\sim q_\phi}\left[g \mid \mathbf{x}_t, \mathbf{d}_t^{(i)}\right]}_{\widehat{G}_\phi(\mathbf{x}_t, \mathbf{d}_t^{(i)})}, \qquad \forall i = 1, \ldots, L.$$

Ranking permutations. Define the generator's ordering as a permutation $r_t^{\text{gen}} : \{1, \ldots, L\} \to \{1, \ldots, L\}$ where $r_t^{\text{gen}}(i)$ is the *rank position* of candidate i (1 = best). Equivalently, $r_t^{\text{gen}}(i) < r_t^{\text{gen}}(j)$ if and only if $\hat{g}_t^{(i)} \geq \hat{g}_t^{(j)}$. We interpret this as a pairwise preference relation $\mathbf{d}_t^{(i)} \succ_{\text{gen}} \mathbf{d}_t^{(j)} \iff r_t^{\text{gen}}(i) < r_t^{\text{gen}}(j)$. Only the order induced by the scores matters; any monotone transform of $\{\hat{g}_t^{(i)}\}_{i=1}^L$ leaves r_t^{gen} unchanged. When score ties occur in practice, they are broken deterministically (e.g., by smaller predictive variance or a fixed index order) so r_t^{gen} is a valid permutation. Let r_t^{exp} denote the expert's permutation over the same L candidates from Stage (ii), with the analogous relation $\mathbf{d}_t^{(i)} \succ_{\text{exp}} \mathbf{d}_t^{(j)} \iff r_t^{\text{exp}}(i) < r_t^{\text{exp}}(j)$.

Agreement metric. Alignment is tested every round t. The misalignment score is the normalized Kendall rank distance

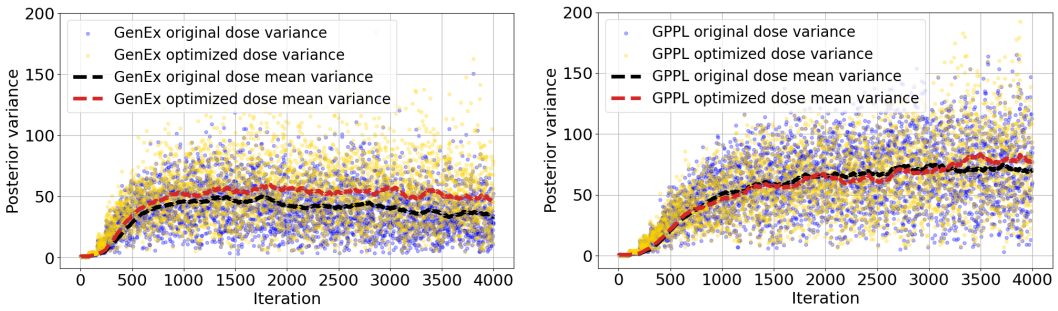
$$\eta_t = \frac{1}{\binom{L}{2}} \sum_{i<j} \mathbb{I}[r_t^{\text{gen}}(i) > r_t^{\text{gen}}(j) \wedge r_t^{\text{exp}}(i) < r_t^{\text{exp}}(j)]\,.$$

It counts the fraction of discordant pairs (i, j) for which the two permutations disagree on relative order (0 = complete agreement; 1 = all pairs inverted). We smooth this over a trailing window of W patients:

$$\bar{\eta}_t = \frac{1}{W} \sum_{s=t-W+1}^{t} \eta_s.$$

The windowed average reduces sensitivity to occasional noisy inversions while remaining responsive to recent behavior.

Figure 5: Posterior variance: original dose vs. optimized dose.



(a) (b)

Panels (a) and (b) compare GenEx and GPPL in terms of posterior variance at original versus optimized doses. GenEx (a) exhibits consistently elevated variance at optimized doses, reflecting active exploration. GPPL (b) shows nearly overlapping traces, indicating conservative, low-risk behavior.

D Posterior Variance Dynamics

We report posterior variance results for GenEx and GPPL over refinement iterations as a supplementary diagnostic to the regret analysis in the main text (Sec. 3). While not our core performance metric, variance offers insight into each algorithm's exploration–exploitation balance and confidence calibration. Zero variance can indicate overconfidence and under-exploration, whereas arbitrarily high variance may reflect uncritical or random search. Ideally, variance should be maintained at moderate levels and directed toward informative regions of the dose space.

Figure 5 shows the posterior variance at the original and model-optimized doses for both GenEx and GPPL. The *original* dose refers to the heuristic used to populate the "Dose" column in our simulated dataset—these are fixed, pre-specified prescription values that exist in the database prior to any refinement. This heuristic is the same one described in detail in Appendix A. The *optimized* dose denotes the model-selected alternative aimed at maximizing predicted response at refinement time. Thus, variance at the original dose captures the model's uncertainty about a typical starting prescription, while variance at the optimized dose captures its uncertainty about its own proposed improvement. Each curve includes a running mean to highlight long-run behavior.

The rising variance trend observed in both models is a natural consequence of exposing the surrogate to increasingly diverse patient covariates. Each new iteration introduces an independent sample from the population, prompting the GP to make predictions in previously unseen regions of the covariate–dose manifold. This affects both the optimized doses (through the model's search) and the original doses (since these too may lie in regions less represented in the seed data), producing higher uncertainty in both spaces.

By jointly examining Figure 5(a) and (b), we observe that GenEx maintains lower variance than GPPL at original doses, indicating stronger confidence in familiar regions of the input space. At the same time, GenEx exhibits controlled but distinct variance spikes at optimized doses, particularly early in training, signaling active exploration into high-uncertainty regions in pursuit of improved outcomes. GPPL, by contrast, shows flatter variance curves with little distinction between original and optimized doses, suggesting a more conservative exploration policy that avoids uncertain areas.

Over time, variance growth stabilizes in both models, as shown by the running-mean curves. This saturation reflects the model's increasing exposure to a representative patient population. Importantly, GenEx reaches this stabilization with significantly tighter overall variance, suggesting more sample-efficient learning and more effective coverage of the decision space.

E Multidimensional Synthetic Environment

To complement the scalar dosing setting in the main text, we also evaluate GenEx on a higher-dimensional dosing testbed. We retain a 10-dimensional patient representation and now consider

3-dimensional dosing configurations. We model the ground truth objective function as a highly nonlinear and nonconvex function that deliberately embeds multiple layers of heterogeneity, providing a rich testbed for evaluating how well algorithms can adapt to context variation, local nonconvexities, and shifting optima. The ground-truth objective function is

$$G(\mathbf{x}, \mathbf{d}) \;=\; s(\mathbf{x}) - \tfrac{1}{2}\,(\mathbf{d} - \mu(\mathbf{x}))^\top A\,(\mathbf{d} - \mu(\mathbf{x})) + \epsilon \cdot \mathrm{ripple}(\mathbf{x}, \mathbf{d}),$$

with $\mathbf{x} \in [-1, 1]^{10}$ and $\mathbf{d} \in [0, 1]^3$.

The components of G are defined as follows. Let $\sigma(z) = \frac{1}{1+e^{-z}}$. With matrices $W \in \mathbb{R}^{3\times 10}$, $M \in \mathbb{R}^{3\times 3}$, $N \in \mathbb{R}^{3\times 10}$, a bias vector $\mathbf{b} \in \mathbb{R}^3$, and $L \in \mathbb{R}^{3\times 3}$ drawn once from zero-mean Gaussian distributions with the specified scales, we set

$$\mu(\mathbf{x}) = 0.2 + 0.6\,\sigma(W\mathbf{x} + \mathbf{b}) \;\in\; [0.2, 0.8]^3,$$

$$s(\mathbf{x}) = 0.2 \sum_{i=1}^{4} \sin(x_i) \;-\; 0.05 \sum_{j=1}^{10} x_j^2,$$

$$\mathrm{ripple}(\mathbf{x}, \mathbf{d}) = \sum_{k=1}^{3} \sin\Big(2\pi\,[\,(\mathbf{d}M^\top + \mathbf{x}N^\top)_k\,]\Big),$$

$$A = LL^\top + 1.5\,I_3 \quad \text{(symmetric positive definite)},$$

$$\epsilon = 0.02.$$

The distributions are $W \sim \mathcal{N}(0,\, 1/\sqrt{p})$, $\mathbf{b} \sim \mathcal{N}(0,\, 0.5)$, $L \sim \mathcal{N}(0,\, 0.5)$, $M \sim \mathcal{N}(0,\, 0.4)$, and $N \sim \mathcal{N}(0,\, 0.4)$, all elementwise.

This construction deliberately embeds multiple layers of heterogeneity. The term $s(\mathbf{x})$ is a bounded context-only component, combining sinusoidal features with quadratic penalties so that different regions of the context space yield systematically different baselines. The mapping $\mu(\mathbf{x})$ produces a context-dependent shift in decision space, so the best decision region varies across $\mathbf{x}$. The quadratic form with positive definite A enforces curvature around $\mu(\mathbf{x})$, while the ripple term adds small, high-frequency oscillations in both $\mathbf{x}$ and $\mathbf{d}$, introducing local nonconvexity. The surrogate must therefore capture broad contextual heterogeneity, adapt to context-specific shifts, and remain robust to local perturbations, making this testbed richer and more challenging than the scalar setting.

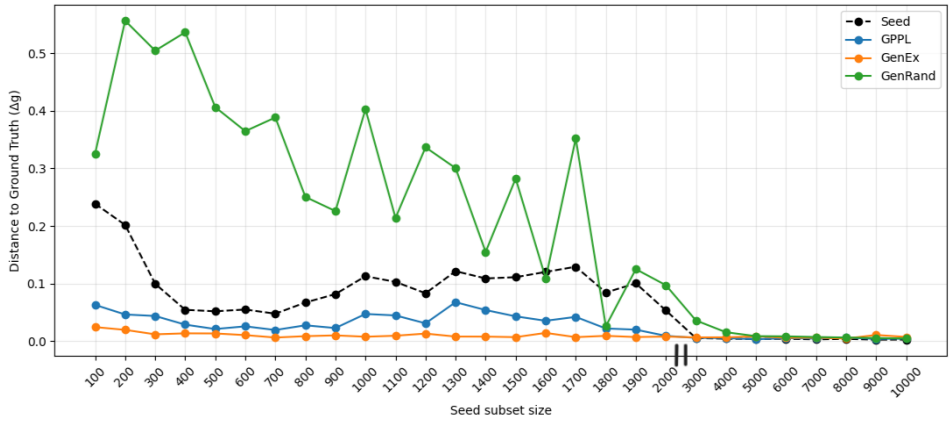
We compare GenEx, GPPL, and GenRand under a setting with a perfect expert. For each context $\mathbf{x}$ in the refinement pool, the algorithms generate $L = 5$ candidate decisions using Thompson sampling from the GP surrogate, constructed over a Sobol grid of size 4096. The refinement loop runs for 50 iterations, and performance is then assessed on 100 independent test contexts. The ground-truth upper bound is approximated by grid search on the same Sobol set. All hyperparameters governing refinement and generation, such as the trust threshold τ, preference margin, generator latent spread, and ELBO weighting, are dynamic, adapting to both data volume and trust rather than being fixed constants, following standard hyperparameter tuning practices. Specifically, the trust threshold τ is annealed based on the rolling misalignment rate, the ELBO weight decays with cumulative synthetic data to prevent over-reliance, and the generator latent spread narrows as more refinement iterations build confidence. These mechanisms ensure that exploration is cautious when data are scarce and becomes more assertive as confidence grows.

Figure 6 shows that GenEx most effectively reduces the absolute gap to the ground truth across all seed subset sizes. The contrast is sharpest at small to medium subsets, where limited training data make refinement especially valuable. In this regime, the advantage of combining expert preferences with generative augmentation is clear: GenEx consistently pulls the surrogate closer to the ground truth, whereas GPPL delivers moderate gains and GenRand often degrades performance.

As the seed grows larger, all methods converge toward the ground truth and the absolute distances shrink. This is expected, since a well-trained baseline GP already captures much of the structure of $G(\mathbf{x}, \mathbf{d})$. Still, GenEx maintains a consistent margin over GPPL, showing that targeted synthetic augmentation provides incremental value even when data are abundant. GPPL saturates more quickly, reflecting its dependence on direct preference updates without the additional leverage of synthesized observations.

Compared to these consistent patterns, GenRand behaves quite differently. At small subsets, its distance to the ground truth often increases relative to the seed, indicating that unguided synthetic data

Figure 6: Distance to ground truth objective across seed subset sizes.



Lower values indicate closer alignment with the true function. GenEx maintains the smallest distance across subset sizes; GPPL improves over the seed baseline but less strongly. GenRand is unstable at small subsets and often diverges, indicating that unstructured augmentation can be counterproductive. The curves flatten beyond larger seed subsets, consistent with diminishing returns once the base surrogate is adequately trained.

Table 1: Summary metrics relative to ground truth.

Method	**Relative gain (%)**	**Ground-truth fraction (%)**
GenEx	26.2	86.7
GPPL	19.1	63.3
GenRand	-58.8	-194.8

add noise and bias the surrogate cannot easily correct. Even at larger seeds, where its performance stabilizes, GenRand fails to yield systematic gains. This instability underscores that random augmentation is not a neutral baseline: it can actively worsen calibration, particularly when supervision is scarce.

Table 1 reports summary statistics aggregated across all seed subset sizes. Relative gain is the percentage improvement over the seed baseline, while ground-truth fraction is the percentage of the seed–truth gap closed. These metrics complement the trends in Fig. 6. GenEx closes nearly 87% of the seed–ground-truth gap and achieves the largest overall gains, while GPPL closes about 63%. GenRand performs negatively, often moving further from the ground truth, reinforcing the need for structured preference guidance and quality-controlled generative augmentation. Together with Fig. 6, these results demonstrate that GenEx provides the strongest and most stable improvements in the higher-dimensional setting, particularly in data-scarce regimes where refinement is most impactful.

Importantly, these results extend the main paper's findings beyond the scalar case. Here, the decision vector has three components, creating a higher-dimensional optimization problem. The qualitative ranking of methods remains consistent, demonstrating that the framework's benefits are not limited to simplified, one-dimensional settings. This shows that GenEx can scale effectively as the decision space grows, whereas methods without principled augmentation struggle to maintain improvements. The richer heterogeneity and local nonconvexity of this function make the task substantially harder than the scalar case, underscoring that GenEx's advantage persists even in more complex environments.

F Proofs

F.1 Model Assumptions

We model GenEx as learning a composite function $\tilde{G}_t(\mathbf{x}, \mathbf{d}) = G(\mathbf{x}, \mathbf{d}) + b_t(\mathbf{x}, \mathbf{d})$, where G is the true goal function and b_t captures bias introduced by synthetic data, i.e., synthetic patients and goal values. We assume that $G(\mathbf{x}, \cdot)$ lies in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ with norm at

most B_G (A1), a standard condition in GP theory that ensures smoothness and avoids arbitrarily sharp curvature. The bias term $b_t(\mathbf{x}, \cdot)$ also lies in $\mathcal{H}_k$, with norm bounded by a non-increasing function $B_b(N_t)$ (A2), where N_t denotes the number of expert queries observed up to round t, reflecting that expert feedback incrementally corrects the generator. Consequently, the composite function satisfies $\|\tilde{G}_t\|_{\mathcal{H}_k} \leq B_G + B_b(N_t)$ (A3), which we denote as $B_{\tilde{G}}(N_t)$.

We further assume that the pointwise bias is bounded as $|b_t(\mathbf{x}, \mathbf{d})| \leq B(N_t)$ (A4), where $B(\cdot)$ decays with more queries, reflecting that synthetic errors shrink as expert input accumulates. Observations are modeled with i.i.d. Gaussian noise $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ (A5), a classical setup that allows analytical posterior updates. We also adopt several standard Bayesian optimization assumptions: the goal function is bounded, $0 \leq G(\mathbf{x}, \mathbf{d}) \leq G_{\max}$ (A6), and the kernel k is continuous, symmetric, and satisfies $k((\mathbf{x}, \mathbf{d}), (\mathbf{x}, \mathbf{d})) \leq K$ (A7), ensuring bounded variance. The joint input domain $\mathcal{X} \times \mathcal{D}$ is compact (In practice every synthetic covariate $\tilde{\mathbf{x}}$ is generated by first sampling a dose in the empirical range of the seed data and then decoding $(\tilde{\mathbf{x}}, \tilde{g})$ with the cVAE; hence all coordinates of $\tilde{\mathbf{x}}$ remain inside the original min–max bounds, so the compact-domain assumption continues to hold) (A8). Finally, the maximum information gain from any T observations is finite, i.e. $\gamma_T = \max_{A \subseteq \mathcal{X} \times \mathcal{D} : |A| = T} \log\det\left(\mathbf{I} + \sigma^{-2}\mathbf{K}_A\right)/2 < \infty$ (A9). Together, these assumptions support our theoretical analysis by bounding the composite function's complexity and controlling the impact of synthetic bias on learning performance.

F.2 Proof of Theorem 1

Goal. Bound the expected cumulative regret $R_T = \sum_{t=1}^{T} r_t$ where $r_t = G(\mathbf{x}_t, \mathbf{d}^*(\mathbf{x}_t)) - G(\mathbf{x}_t, \mathbf{d}_t)$.

Step 1: one–step bound on the conditional mean regret. At round t apply Theorem 2 with confidence parameter $\delta/(2T)$ to each of the *two* doses $\mathbf{d}_t$ and $\mathbf{d}^*(\mathbf{x}_t)$. With probability at least $1 - \delta/T$ for each dose,

$$\left|G - \mu_{t-1}\right| \;\leq\; \beta_{t-1}^{1/2}\sigma_{t-1} + B(N_{t-1}).$$

Choosing $\delta = 1/T$ and union-bounding over the $2T$ concentration events (two points per round) contributes an extra $\log T$ term, which we absorb into the definition of β_T. Conditioning on the past $\mathcal{F}_{t-1}$ and using the union bound,

$$\mathbb{E}\left[r_t \mid \mathcal{F}_{t-1}\right] \;\leq\; 2\beta_{t-1}^{1/2}\sigma_{t-1}(\mathbf{x}_t, \mathbf{d}_t) \;+\; 2B(N_{t-1}).$$

Step 2: control the sum of posterior standard deviations. Recall from Step 1 that

$$\mathbb{E}[R_T] \leq 2\sqrt{\beta_T}\sum_{t=1}^{T}\mathbb{E}\big[\sigma_{t-1}(\mathbf{x}_t, \mathbf{d}_t)\big] + 2\sum_{t=1}^{T}B(N_{t-1}).$$

Since at time t, $\sigma_{t-1}(\mathbf{x}_t, \mathbf{d}_t)$ is $\mathcal{F}_{t-1}$-measurable once $\mathbf{d}_t$ is drawn, we can bound

$$\sum_{t=1}^{T}\sigma_{t-1}(\mathbf{x}_t, \mathbf{d}_t) \leq \sqrt{T}\Big(\sum_{t=1}^{T}\sigma_{t-1}^2(\mathbf{x}_t, \mathbf{d}_t)\Big)^{1/2} \leq \sqrt{T\,\gamma_T},$$

where the last line is Lemma F.4.

Step 3: aggregate over t. Summing the one–step bound of Step 1 and substituting the inequality above,

$$\mathbb{E}[R_T] \;\leq\; 2\beta_T^{1/2}\sqrt{T\gamma_T} \;+\; 2\sum_{t=1}^{T}B(N_{t-1}),$$

with $\beta_T \;=\; \max_{1 \leq t \leq T} 2C_k^2 B_{\tilde{G}}(N_{t-1})^2\big(\gamma_{t-1} + 1 + 3\log T\big)$.

Step 4: conclude no-regret. If $\gamma_T = o(T)$ and $\sum_t B(N_{t-1}) < \infty$, divide by T and let $T \to \infty$ to obtain $\mathbb{E}[R_T]/T \to 0$. □

F.3 Proof of Theorem 2

Theorem 2 (Adjusted Confidence Interval)**.** *Fix $\delta \in (0,1)$ and let σ^2 be the observation–noise variance from (A5). Under Assumptions (A1)–(A9), with probability at least $1-\delta$,*

$$\big|G(\mathbf{x},\mathbf{d}) - \mu_{t-1}(\mathbf{x},\mathbf{d})\big| \le \beta_{t-1}^{1/2}\sigma_{t-1}(\mathbf{x},\mathbf{d}) \;+\; B(N_{t-1}), \qquad \forall (\mathbf{x},\mathbf{d}) \in \mathcal{X} \times \mathcal{D},$$

where $\sigma_{t-1}(\mathbf{x},\mathbf{d})$ is the GP posterior standard deviation *at round $t-1$ and*

$$\beta_{t-1} = 2C_k^2 B_{\tilde{G}}(N_{t-1})^2\big(\gamma_{t-1} + 1 + \log\tfrac{2}{\delta}\big), \qquad C_k = \sup_{(\mathbf{x},\mathbf{d})} \sqrt{k\left((\mathbf{x},\mathbf{d}),(\mathbf{x},\mathbf{d})\right)} \le \sqrt{K}.$$

Here γ_{t-1} is the maximum information gain defined in (A9).

Goal. Show that, with probability at least $1-\delta$,

$$\big|G(\mathbf{x},\mathbf{d}) - \mu_{t-1}(\mathbf{x},\mathbf{d})\big| \le \beta_{t-1}^{1/2}\sigma_{t-1}(\mathbf{x},\mathbf{d}) + B(N_{t-1}) \quad \forall (\mathbf{x},\mathbf{d}).$$

Step 1: Decompose the error. Recall the definition $\tilde{G}_{t-1} = G + b_{t-1}$ with b_{t-1} the bias from synthetic data. Add–subtract $\tilde{G}_{t-1}$:

$$\underbrace{G - \mu_{t-1}}_{\text{total error}} = \underbrace{\tilde{G}_{t-1} - \mu_{t-1}}_{\text{estimation error}} - \underbrace{b_{t-1}}_{\text{bias}}. \tag{1}$$

Step 2: Bound the estimation error. By (A3), $\|\tilde{G}_{t-1}\|_{\mathcal{H}_k} \le B_{\tilde{G}}(N_{t-1})$. Hence Lemma F.4 (with $B = B_{\tilde{G}}(N_{t-1})$) gives

$$\big|\tilde{G}_{t-1}(\mathbf{x},\mathbf{d}) - \mu_{t-1}(\mathbf{x},\mathbf{d})\big| \le C_k B_{\tilde{G}}(N_{t-1})\sigma_{t-1}(\mathbf{x},\mathbf{d})\sqrt{2\big(\gamma_{t-1} + 1 + \log\tfrac{2}{\delta}\big)} \quad (\forall (\mathbf{x},\mathbf{d})). \tag{2}$$

Step 3: Bound the bias. Assumption (A4) states $|b_{t-1}(\mathbf{x},\mathbf{d})| \le B(N_{t-1})$ for all inputs. Therefore

$$|\text{bias}| \le B(N_{t-1}). \tag{3}$$

Step 4: Combine. Insert (2) and (3) into (1) and use the triangle inequality:

$$|G - \mu_{t-1}| \le C_k B_{\tilde{G}}(N_{t-1})\sigma_{t-1}\sqrt{2(\gamma_{t-1} + 1 + \log\tfrac{2}{\delta})} + B(N_{t-1}).$$

Define $\beta_{t-1} = 2C_k^2 B_{\tilde{G}}(N_{t-1})^2(\gamma_{t-1} + 1 + \log\frac{2}{\delta})$ so that the first term becomes $\beta_{t-1}^{1/2}\sigma_{t-1}$. The desired inequality follows. □

F.4 Auxiliary Lemmas

[GP Concentration with Information Gain] Let $f \in \mathcal{H}_k$ with $\|f\|_{\mathcal{H}_k} \le B$ and set $C_k = \sup_{(\mathbf{x},\mathbf{d})} \sqrt{k\big((\mathbf{x},\mathbf{d}),(\mathbf{x},\mathbf{d})\big)}$. After n noisy observations with noise variance σ^2, the GP posterior (μ_n, σ_n) satisfies, for any $\delta \in (0,1)$,

$$\big|f(\mathbf{x},\mathbf{d}) - \mu_n(\mathbf{x},\mathbf{d})\big| \le C_k B\sigma_n(\mathbf{x},\mathbf{d})\sqrt{2\big(\gamma_n + 1 + \log\tfrac{2}{\delta}\big)} \quad \forall (\mathbf{x},\mathbf{d}).$$

Proof. Direct application of the uniform RKHS-concentration inequality (see Appendix Srinivas et al. [2010]), using (A7) to upper-bound the evaluation functional. □

[Adjusted Variance Bound] Let γ_T be the adjusted information gain after T observations. Then for any sequence $\{(\mathbf{x}_t,\mathbf{d}_t)\}_{t=1}^T$,

$$\sum_{t=1}^{T} \sigma_{t-1}^2(\mathbf{x}_t,\mathbf{d}_t) \le \gamma_T.$$

Proof. For the observed sequence $\{(\mathbf{x}_t,\mathbf{d}_t)\}_{t=1}^T$, the mutual information satisfies

$$I(\mathbf{f};\mathbf{y}) = \tfrac{1}{2}\sum_{t=1}^{T} \log\big(1 + \sigma^{-2}\,\sigma_{t-1}^2(\mathbf{x}_t,\mathbf{d}_t)\big) \le \tfrac{1}{2}\sum_{t=1}^{T} \frac{\sigma_{t-1}^2(\mathbf{x}_t,\mathbf{d}_t)}{\sigma^2} \le \tfrac{1}{\sigma^2}\, I(\mathbf{f};\mathbf{y}) \le \tfrac{1}{\sigma^2}\,\gamma_T.$$

Rearranging (and absorbing the factor $1/\sigma^2$ into the definition of γ_T if desired) gives

$$\sum_{t=1}^{T} \sigma_{t-1}^2(\mathbf{x}_t, \mathbf{d}_t) \leq \gamma_T.$$

See Srinivas et al. [2010] for the detailed derivation. □