Optimizing Likelihoods via Mutual Information: Bridging Simulation-Based Inference and Bayesian Optimal Experimental Design

Vincent D. Zaballa[1] **Elliot E. Hui**[1]

[1]Department of Biomedical Engineering, University of California, Irvine, Irvine, California, USA

Abstract

Simulation-based inference (SBI) typically depends on expensive simulators, so inference and experimental design must operate under fixed simulation budgets. Bayesian optimal experimental design (BOED) maximizes expected information gain (EIG), but this utility function can hide shortcomings of the underlying inference objects that will be used in downstream inference tasks (posterior estimation). We show that the InfoNCE lower bound on EIG becomes a principled SBI training objective for conditional normalizing flow density model - the workhorse of SBI applications - with an asymptotic likelihood-fitting interpretation that links MI maximization to surrogate-likelihood learning and thereby increases the value of each simulator call for both inference and design. We propose SBI-BOED, a single stochastic-gradient procedure that jointly trains a conditional normalizing-flow likelihood and optimizes designs without requiring simulator differentiability or pathwise gradients to experimental designs. The same MI view also yields simulation active learning under fixed designs via expected predictive information gain (EPIG). With an InfoNCE-λ objective and practical stabilizers, SBI-BOED improves posterior calibration and predictive accuracy over strong BOED baselines on synthetic benchmarks and scientific simulators at matched simulation budgets.

1 INTRODUCTION

Scientific models can be described as simulators that simulate complex, potentially stochastic, interactions that yield observations from the interactions. Additionally, these simulators sometimes have *controllable* inputs that yield a different distribution of observations. For example, a simulator generates an output y based on experimental designs ξ (controllable inputs) and model parameters θ (latent system properties that govern the input-output relationship). Inferring the posterior distribution of model parameters given observed data $p(\theta|y, \xi)$ represents a fundamental challenge in Bayesian statistics and can be viewed as solving an inverse problem for a given simulator Lindley [1972]. In Simulation-Based Inference (SBI), *some* simulators implicitly define a design-dependent likelihood $p(y|\theta, \xi)$ that, combined with a prior over model parameters $p(\theta)$, enables inference of the posterior distribution $p(\theta|y_o, \xi)$ where y_o represents observed data Cranmer et al. [2020]. SBI methods address scenarios where the likelihood is intractable by approximating either 1) the likelihood $p(y|\theta, \xi)$ or 2) the posterior $p(\theta|y, \xi)$ directly using neural density estimators, or, 3) using classifiers to estimate the likelihood-to-evidence ratio, $\frac{p(y,\theta|\xi)}{p(\theta)p(y|\xi)}$. Importantly, simulations can be costly (i.e. *slow*) to collect, introducing a bottleneck in downstream applications of simulators.

Inferring the parameters of the model given observations requires performing experiments that may be costly to perform. For example, in drug development, collecting data is expensive and is partially responsible for the great cost associated with developing new therapies where thousands to millions of designs may be employed *per round* of experimentation in a high-throughput screening campaign [Paul et al., 2010]. Triaging which data points to gather can help reduce the time to develop a new therapy for patients. It is thus imperative to prioritize collection of observed data y_o, using optimal designs $y_o|\xi^*$, to arrive at an accurate and *calibrated* inference of model parameters to make predictions of future responses, such as the likelihood of successful drug treatment.

Towards reducing cost, Bayesian optimal experimental design (BOED) has shown promise as a method for optimizing experiments ξ, even when dealing with potentially high-dimensional design spaces. This can be achieved by employing a likelihood model along with priors for the parameters

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7928–7951.

of interest [Lindley, 1956, Foster et al., 2019, Kleinegesse and Gutmann, 2019]. BOED operates by both *evaluating* and *optimizing* a utility function of the Expected Information Gain (EIG) upon performing a proposed experiment Lindley [1956], Rainforth et al. [2024].

$$I(\xi) \triangleq \mathbb{E}_{p(y|\xi)}\left[H[p(\theta)] - H[p(\theta|y,\xi)]\right]. \tag{1}$$

The EIG can be used as an approximation for the information gained in an experiment with design ξ. The intuition is: which experimental design and outcome would be most surprising given what we assume about the model when conducting the experiment? This would be the optimal experimental design and can be rewritten into the form of calculating the mutual information (MI), where $\mathrm{MI}(\theta; y|\xi) = I(\xi)$, between the observed data and unknown parameters as the ratio of likelihood to marginal likelihood or posterior to prior

$$\begin{aligned}\mathrm{MI}(\theta; y|\xi) &= \mathbb{E}_{p(\theta)p(y|\theta,\xi)}\left[\log \frac{p(y|\theta,\xi)}{p(y|\xi)}\right] \\ &= \mathbb{E}_{p(\theta)p(y|\theta,\xi)}\left[\log \frac{p(\theta|y,\xi)}{p(\theta)}\right].\end{aligned} \tag{2}$$

However, most BOED utility functions require either an explicit likelihood or a differentiable one. Previous BOED work focused on estimating the MI as an objective function within an outer optimizer, such as Bayesian optimization, which results in a nested optimization process [Gutmann et al., 2016, Rainforth et al., 2018, Kleinegesse and Gutmann, 2020, Foster et al., 2019]. Notably, these methods can be combined with implicit models such as those used in SBI, but they can be inefficient, motivating methods to simultaneously optimize the design and MI in a single optimization process [Foster et al., 2020]. However, unified optimization typically depends on an *explicit* likelihood or an implicit likelihood with a *differentiable* simulator [Kleinegesse and Gutmann, 2021, Ivanova et al., 2021], except for *simulation-heavy* reinforcement learning (RL) methods Lim et al. [2022], Blau et al. [2022, 2024], which are often not feasible in SBI settings where generating new simulator samples takes seconds or longer.

Expensive simulators also motivate an active-learning perspective on SBI: under a fixed simulation budget, not all simulator queries are equally valuable. Rather than drawing simulations uniformly, one can prioritize parameter settings whose outcomes are expected to maximally improve the learned likelihood or posterior surrogate. In our approach, this prioritization follows from the same InfoNCE-regularized mutual-information objective used for design optimization, yielding a unified principle for both selecting informative experiments and allocating simulator calls. This further reduces simulation burden by focusing computation on the most informative queries.

We show that SBI and BOED can benefit from a shared information-theoretic perspective. On a standard SBI benchmark, we first study how InfoNCE and its regularized variant affect the training dynamics of amortized likelihood models. We then use the same bound to define the *value* of a simulation, yielding an active-learning strategy for selecting new latent parameters θ under a fixed simulation budget. Next, we apply this objective to BOED with implicit models, enabling design optimization and surrogate training without requiring a differentiable simulator and avoiding the simulation cost of RL-based methods. Finally, we show that SBI-inspired iterative inference can improve cumulative information gain and posterior calibration over multiple rounds of BOED. Overall, our results highlight the value of cross-pollination between SBI and BOED for improving both inference quality and simulation efficiency. We summarize the offline simulator-call comparison in Table 1.

Table 1: Offline simulator calls to learn an amortized design strategy; one call is one draw $y \sim p(y \mid \theta, \xi)$. On SIR (§4.3), SBI-BOED exceeds iDAD accuracy while using $\sim$50$\times$ fewer simulator calls.

	RL-BOED[†]	iDAD	SBI-BOED
Simulator calls	$\sim 10^7$–10^8	$\approx 2.6 \times 10^8$	$\approx 5.1 \times 10^6$

[†]Not evaluated on SIR; Blau et al. report training amortized RL-BOED agents with tens of millions of simulated experiments.

In bridging BOED and SBI, our key contributions are:

- **Regularized MI Bound for Improved Inference**: We show that the $I_{\text{NCE-}\lambda}$ utility improves calibration and predictive accuracy relative to the standard InfoNCE estimator, and we analyze its effect on training dynamics for amortized likelihood models.
- **Active Learning for Simulation-Efficient SBI**: We propose an active-learning strategy for selecting new latent parameters θ under a fixed simulation budget, using the InfoNCE-regularized mutual-information bound to prioritize simulator queries that are expected to be most informative.
- **Practical Optimization for Implicit Likelihoods**: We develop a robust method to jointly optimize experimental designs and train SBI surrogates without requiring a differentiable simulator. To address the limitations of naive gradient-based design optimization, we optimize a design *distribution*, improving performance in non-differentiable settings.
- **Multi-Round BOED with SBI Principles**: We show that incorporating SBI ideas such as repeated rounds of inference improves cumulative information gain and yields better-calibrated posteriors over multiple rounds of BOED.

2 BACKGROUND

2.1 SIMULATION-BASED INFERENCE

SBI addresses Bayesian inference in settings where a simulator g can generate synthetic observations but the likelihood is unavailable or impractical to evaluate. Given a prior $\theta \sim p(\theta)$ and (optionally) a design ξ, the simulator defines a data-generating process $y = g(\theta, \xi, z)$ with randomness z, inducing an implicit likelihood $p(y \mid \theta, \xi)$ and posterior $p(\theta \mid y_o, \xi)$. Modern SBI methods learn amortized surrogates from simulated pairs (θ, y) by training neural density estimators (e.g., normalizing flows) to approximate either the likelihood $p_\phi(y \mid \theta, \xi)$ or the posterior $p_\phi(\theta \mid y, \xi)$ via maximum likelihood [Papamakarios et al., 2021, Lueckmann et al., 2021]. Alternatively, SBI can estimate the likelihood-to-evidence ratio by training a classifier to distinguish samples from the joint $p(\theta, y \mid \xi)$ versus the product of marginals $p(\theta)p(y \mid \xi)$, yielding $\exp f_\phi(\theta, y) \approx \frac{p(y|\theta,\xi)}{p(y|\xi)}$ (and related ratios). The choice between posterior, likelihood, or ratio-based SBI depends on downstream needs, e.g., amortized posteriors are convenient for repeated inference on many observations, while ratio ensembles can improve robustness (e.g., under calibration tests) at additional compute cost [Talts et al., 2020, Hermans et al., 2022].

Neural Likelihood Estimation We can use data from the joint distribution to train a conditional neural density-based likelihood function (NL). If we take a dataset of samples $\{y_n, \theta_n\}_{1:N}$ obtained from a simulator as previously described, we can train a conditional density estimator $p_\phi(y|\theta)$ to model the likelihood by maximizing the total log likelihood of $\sum_n \log p_\phi(y_n|\theta_n)$, which is approximately equivalent to minimizing

$$\mathcal{L}_{\text{NL}}(\phi) = \mathbb{E}_{p(\boldsymbol{\theta})}(D_{\text{KL}}(p(y|\theta)\|p_\phi(y|\theta)) + \text{const}, \quad (3)$$

where the divergence is minimized when $p_\phi(y|\theta) \to p(y|\theta)$.

2.2 NORMALIZING FLOWS

Normalizing flows are flexible, tractable density estimators that model a target distribution via an invertible transformation of a simple base distribution [Papamakarios et al., 2021, Kobyzev et al., 2020]. Let $u \sim p(u)$ (e.g., standard Gaussian) and let $f_\phi : \mathbb{R}^D \to \mathbb{R}^D$ be a bijection, typically a composition of N invertible maps, $f_\phi = f_N \circ \cdots \circ f_1$. The induced density on $y = f_\phi(u)$ follows from change of variables,

$$p_\phi(y) = p(u) \left|\det J_{f_\phi}(u)\right|^{-1},$$

where J_{f_ϕ} is the Jacobian of f_ϕ. Flows are trained by maximum likelihood on samples $\{y_n\}_{n=1}^N$, equivalently minimizing the forward KL divergence $D_{\text{KL}}(p^*(y) \,\|\, p_\phi(y))$. In SBI, flows are commonly used in conditional form by conditioning the transformation (or its parameters) on context such as θ or y, yielding likelihood models $p_\phi(y \mid \theta, \xi)$ or posteriors $p_\phi(\theta \mid y)$ [Durkan et al., 2019]. Because sampling is reparameterized (e.g., $y = f_\phi(u; \theta, \xi)$ with $u \sim p(u)$), flows admit pathwise gradients and are well-suited for gradient-based learning and variational objectives [Rezende and Mohamed, 2015, Mohamed et al., 2020].

2.3 BAYESIAN OPTIMAL EXPERIMENTAL DESIGN

InfoNCE Bound Following from Equation 2, Ivanova et al. [2021] proposed a lower bound of the MI based on the InfoNCE bound using an implicit likelihood with a differentiable simulator. They trained a design policy network π_ψ that proposed designs based on the history of design-observation pairs, $h_{t-1} = \{(\xi_i, y_i)\}_{i=1:t-1}$ and a critic network U_ϕ that encapsulates the true likelihood when it maximizes the lower bound of MI. We adjust the bound to reflect the myopic experimental design setting (ignoring the design policy) and reformulate the bound as

$$\mathcal{L}_{\text{NCE}}(\xi, \phi; L) \coloneqq \mathbb{E}\left[\log \frac{\exp(U_\phi(y, \xi, \theta_0))}{\frac{1}{L+1}\sum_{i=0}^{L} \exp(U_\phi(y, \xi, \theta_i))}\right], \quad (4)$$

where the expectation is over $p(\theta_0)p(y|\theta_0, \xi)p(\theta_{1:L})$, ξ is the proposed design, θ_0 is the original parameter that generated data y, generated from the differentiable simulator $p(y|\theta_0, \xi)$, L is the number of contrastive samples, and U is a "critic" function such that $U : (y, \xi) \times \Theta \to \mathbb{R}$. This bound has low variance but is upper-bounded by $\log(L+1)$, potentially leading to large bias with insufficient contrastive samples. Additionally, previous work required a differentiable simulator to take gradients with respect to the inputs of the simulator, which may not be available in SBI settings.

2.4 ACTIVE LEARNING IN SBI

SBI is often bottlenecked by the number of simulator evaluations, motivating *active* strategies that prioritize which simulator queries to run. Sequential SBI methods such as SNPE/SNL already improve sample efficiency by adapting a proposal over parameters toward regions that better explain the observation, effectively steering simulations toward higher-posterior-density regions. However, proposal updates typically do not explicitly optimize the utility of individual simulator queries and may still allocate budget to redundant or low-value simulations. Bayesian active learning formalizes query selection through an acquisition function (often information-theoretic) that scores candidate parameters by their expected utility under the current model. A closely related recent approach, Active Sequential Neural Posterior Estimation (ASNPE) Griesemer et al. [2024], explicitly introduces an acquisition step into the SNPE loop by ranking candidate simulator parameters according to epistemic

uncertainty in the current neural density estimator (approximated via Bayesian NDE machinery such as MC-dropout), and simulating only the top-ranked candidates. In our setting, this perspective motivates using prediction-oriented information-gain criteria (EPIG) to prioritize simulator calls when designs are fixed, yielding an active-learning view of SBI that complements BOED.

3 SBI-BOED

We begin by noting that the mutual-information objectives used in BOED also provide a unifying view of several SBI procedures. In our setting, this shared MI perspective has two consequences. First, when experimental designs ξ are optimized, maximizing an InfoNCE-based lower bound yields a practical objective for jointly training an amortized likelihood model and selecting informative designs, even for closed-box simulators. Second, when designs are fixed, the same learned likelihood can be used to prioritize simulator queries over latent parameters θ, giving an active-learning strategy for allocating a limited simulation budget. We develop these two perspectives in turn: first through the connection between MI maximization and surrogate likelihood learning, and then through EPIG-based active learning of simulations.

3.1 BRIDGING SBI AND BOED

We take inspiration from previous SBI and BOED methods to allow optimization of designs with respect to closed-box simulators that are modeled using normalizing flows. We start by noting how the loss function of contrastive ratio estimation (CRE) lower bounds Equation (4)

$$\begin{aligned}\log \frac{e^{g_\phi(\theta_0,y)}}{\frac{1}{L}\sum_{\ell=1}^{L} e^{g_\phi(\theta_\ell,y)}} &\le \log \frac{e^{g_\phi(\theta_0,y)}}{\frac{1}{1+L}\sum_{\ell=0}^{L} e^{g_\phi(\theta_\ell,y)}} \\ &= \log \frac{p_\phi(y|\theta_0,\xi)}{\frac{1}{1+L}\sum_{\ell=0}^{L} p_\phi(y|\theta_l,\xi)},\end{aligned} \tag{5}$$

where L is the number of contrastive samples, which is K in CRE, and g_ϕ is a discriminative classifier, which holds for a single batch of data and constant experimental design, i.e. when ξ is constant. We use a neural density estimator to create a Likelihood-Free version of Equation (4). We now have a MI lower bound

$$\mathcal{L}_{\text{NCE}}(\xi,\phi;L) \coloneqq \mathbb{E}\left[\log \frac{p_\phi(y|\theta_0,\xi)}{\frac{1}{1+L}\sum_{\ell=0}^{L} p_\phi(y|\theta_l,\xi)}\right] \tag{6}$$

where the expectation is over $p(\theta_0)p(y|\theta_0,\xi)p(\theta_{1:L})$. We can now simultaneously optimize designs and parameters of a neural density estimator. If we are to use a normalizing flow instead of a classifier as $\exp g_\phi(y,\theta,\xi) = p_\phi(y|\theta,\xi)$, then the lower bound of the MI holds since normalizing flows are normalized probability distribution functions. The result is an amortized likelihood at a potentially optimal experimental design. We discuss more SBI methods and their connection to MI optimization and BOED in Appendix D. Given this connection, we state our main proposition.

Proposition 3.1. *In the limit as the number of contrastive samples $L \to \infty$, maximizing the lower bound of the Mutual Information (MI) between parameters θ and observations y in a Simulation-Based Inference (SBI) setting is equivalent, up to constants independent of ϕ, to minimizing the Kullback-Leibler (KL) divergence, $D_{KL}(p(y|\boldsymbol{\theta})||p_\phi(y|\boldsymbol{\theta}))$, between the likelihood $p(y|\boldsymbol{\theta})$ and its approximation $p_\phi(y|\boldsymbol{\theta})$, together with the corresponding marginal-likelihood approximation term, given as*

$$\begin{aligned}\max_\phi I_\phi(\theta;y) \Leftrightarrow \min_\phi \big[\, &\mathbb{E}_{p(\theta)}\, D_{KL}(p(y|\theta)||p_\phi(y|\theta)) \\ &+ \mathbb{E}_{p_\phi(y)}\, D_{KL}(p(y)||p_\phi(y))\big],\end{aligned} \tag{7}$$

where $I(\theta;y)$ is the mutual information between θ and y, and $I_\phi(\theta;y)$ is its approximation under parameter ϕ.

Details and full proof of this proposition are in Appendix A.1. This proposition characterizes the likelihood-learning interpretation of the MI objective. The same objective also has a complementary fixed-design interpretation: when ξ is held constant, it can be used to score the value of simulator queries over θ, yielding an active-learning strategy for simulation-efficient SBI.

3.2 SBI ACTIVE LEARNING VIA EPIG

While SBI-BOED focuses on selecting designs ξ to maximize expected information gain, the same MI-based objective also induces an *active learning* strategy over simulator parameters when designs are fixed (or when one wishes to prioritize which simulator queries to collect). Concretely, if ξ is held constant, each simulator call corresponds to querying a parameter value θ and receiving an observation $y \sim p(y \mid \theta,\xi)$. Under a learned conditional likelihood $p_\phi(y \mid \theta,\xi)$, we can score candidate queries by the *expected predictive information gain* (EPIG), which measures how strongly joint predictions under shared model uncertainty deviate from the product of marginals Smith et al. [2024, 2023]. Intuitively, EPIG prioritizes θ values for which the current likelihood model exhibits high epistemic uncertainty, yielding simulator calls that are expected to be most informative for improving the model.

To approximate epistemic uncertainty without requiring an explicit Bayesian neural network, we use MC-dropout to form a set of stochastic likelihood "particles" $\{\phi^m\}_{m=1}^M$. This is motivated by prior work showing that fixed-dropout-mask ensembles provide a practical mechanism for epistemic uncertainty estimation in normalizing flows [Berry

and Meger, 2023]. For a candidate query θ and a set of target parameters $\theta^\star$ selected as the top-K contrastive negatives under the InfoNCE objective, we estimate

$$\mathrm{EPIG}(\theta) \triangleq \mathbb{E}\left[\log \frac{p_\phi(y, y^\star \mid \theta, \theta^\star, \xi)}{p_\phi(y \mid \theta, \xi)\, p_\phi(y^\star \mid \theta^\star, \xi)}\right]. \quad (8)$$

where the expectation is over $p(\theta^\star)\, p_\phi(y, y^\star \mid \theta, \theta^\star, \xi)$ and the joint predictive $p_\phi(y, y^\star \mid \theta, \theta^\star, \xi)$ is sampled by using the *same* dropout particle $\phi^{(m)}$ for both y and $y^\star$, capturing shared epistemic variation. We then acquire new simulator data by selecting θ with maximal EPIG and adding the resulting (θ, y) pairs to the training set. In this way, the same likelihood model used for BOED can be leveraged to make simulator calls more effective even in the absence of design optimization.

3.3 OPTIMIZING SBI-BOED

Regularization Stability of the density estimator is a challenge when optimizing the MI lower bound due to data distribution shift as a result of changing $p(y|\theta, \xi)$ when optimizing ξ. To address this, we added a regularization term, λ, to help stabilize the training of the density estimator during design optimization as

$$\mathcal{L}_{\text{NCE-}\lambda}(\xi, \phi; L) \coloneqq \mathbb{E}\left[\log \frac{p_\phi(y|\theta_0, \xi)}{\frac{1}{1+L}\sum_{\ell=0}^{L} p_\phi(y|\theta_l, \xi)} + \lambda \cdot \log p_\phi(y|\theta_0, \xi)\right]. \quad (9)$$

where the expectation is over $p(\theta_0)p(y|\theta_0, \xi)p(\theta_{1:L})$. This regularization parameter gives more importance to accuracy of likelihood predictions when training where we provide theoretical analysis in Appendix B.1.

Regularized Mutual Information Estimator We integrate the regularization term λ into a mutual information estimator and propose the InfoNCE-λ objective

$$I_{\text{NCE-}\lambda}(\phi, \lambda) \coloneqq \mathbb{E}_{p(\theta, y|\xi)}\left[\log \frac{p_\phi(y|\theta, \xi)^{1+\lambda}}{\frac{1}{1+L}\sum_{\ell=0}^{L} p_\phi(y|\theta_l, \xi)}\right]. \quad (10)$$

This objective reweights the learned likelihood term through the exponent $1 + \lambda$, allowing us to control the tradeoff between likelihood fitting and the resulting EIG estimate. In particular, varying λ changes how strongly the objective emphasizes the surrogate likelihood relative to the marginal term. We provide a theoretical analysis of how λ affects likelihood prediction accuracy and EIG values in Appendix B.1.

Design Distributions A drawback of gradient-based methods is when there are sparse rewards, such as when the signal to noise ratio is, or approaches, zero. This results in designs struggling to optimize by gradient descent because optimizers may not be able to handle starting, or residing, in areas with no information. We demonstrate this failing in an ablation study in Section 4.3 and take inspiration from RL's use of a replay buffer [Lin, 1992, Mnih et al., 2013] by imposing a parameterized distribution of designs and optimizing parameters of that distribution to return more information. When using design distributions, we compute the EIG for each individual observation y_i associated with design ξ_i and corresponding to the nominal parameter set θ_0. In our usage, EIG_i refers to the theoretical information gain expected from a single sample realization under design ξ_i, based on the current parameterization of the likelihood. It quantifies the information gain computed from a possible outcome as

$$\mathrm{EIG}_i(\psi, \phi, L, \lambda) = \mathbb{E}\left[\log \frac{p_\phi(y_i|\theta_0, \xi_i)^{1+\lambda}}{\frac{1}{1+L}\sum_{\ell=0}^{L} p_\phi(y_i|\theta_l, \xi_i)}\right], \quad (11)$$

with expectation over $p_\psi(\xi)p(\theta_0)p(y|\theta_0, \xi)p(\theta_{1:L})$ and where the parameter of the design distribution ψ is optimized rather than the designs themselves.

We used a truncated Normal distribution for the design parameters, defined as $p_\psi(\xi) = \mathcal{N}_{\text{trunc}}(\mu_\xi|\sigma_n^2)$, where $\mathcal{N}_{\text{trunc}}$ denotes a Normal distribution truncated within the bounds $[\alpha, \beta]$, and α and β are the lower and upper bounds, respectively, and σ that uses a decays schedule depending on the training round, n. Whenever the rewards are sparse, if the design distribution is initialized with sufficient support then it will contain an EIG with significantly more reward that can be updated with gradient descent. In our implementation, we used the reparameterization trick and a standard deviation schedule for σ that decreased according to an exponential schedule $\sigma_n = \sigma_{\text{end}} + (\sigma_{\text{start}} - \sigma_{\text{end}}) \cdot e^{-\frac{n*\rho}{N}}$, where

Algorithm 1 SBI-BOED

Require: Simulator $p(y|\theta, \xi)$, Estimator $p_\phi(y|\theta, \xi)$, Prior $p(\theta)$, Number of experimental designs T, Number of BOED training steps N
Return: Optimal designs $\{\xi_i^* \mid i = 1, \ldots, T\}$, Approximate likelihood $p_\phi(y|\theta, \xi_T^*)$
Initialize dataset $\mathcal{D} = \{\}$
Initialize $\hat{p}_0(\theta) \leftarrow p(\theta)$
for $t = 1, \ldots, T$ **do**
  Sample $\theta_{0:L} \sim \hat{p}_{t-1}(\theta)$
  for $n = 1, \ldots, N$ **do**
    Sample $\xi \sim \mathcal{N}_{\text{trunc}}(\mu_\xi \mid \sigma_n^2)$
    Simulate $y \sim p(y|\theta, \xi)$
    Estimate $\nabla\mathcal{L}_{\text{NCE}-\lambda}(\xi, \phi; L)$ via Equation (9)
    Update μ_ξ and ϕ using ∇_ξ and ∇_ϕ
    Checkpoint $\xi^* = \xi$ **if** $EIG_\xi > EIG_{\xi^*}$
  end for
  Observe y_o using ξ^* in an experiment
  Add (ξ^*, y_0) to dataset $\mathcal{D}$
  Set $\hat{p}_t(\theta) \propto \left(\prod_{(\xi_i, y_i)\in\mathcal{D}} p_\phi(y_i \mid \theta, \xi_i)\right) p(\theta)$
end for

σ_n is the new standard deviation for the Normal distribution that parameterizes ξ in training round n, σ_{start} and σ_{end} are hyperparameters for the initial and final variances respectively, N is the total number of rounds, and ρ is a decay rate hyperparameter.

Design Checkpoints Checkpoints are used in supervised learning to save parameters that achieved low validation error [Vaswani et al., 2017, Devlin et al., 2019]. Fujimoto et al. [2023] proposed the use of checkpoints in RL applications to save a policy that obtains a high reward during training to help improve performance at test time. Even with a distribution of designs, we found designs falling into a local minima during optimization, such as in Appendix D.3. We used design checkpoints, illustrated in Figure 3, to mitigate the risk of gradient-based designs falling into local minima that do not contain sufficient support to encapsulate the global minimum loss.

Posterior Inference The likelihood trained on the optimized design can return approximate posterior samples by sampling $\hat{p}(\theta|y,\xi) \propto p_\phi(y|\theta,\xi)p(\theta)$. Should the likelihood be trained on i.i.d. data, then the joint factorizes to product likelihood $p_\theta(y_{1,\dots,N}|\theta,\xi_{1,\dots,N}) = \prod_{i=0}^{N} p(y_i|\theta,\xi_i)$ which can be used in Markov chain Monte Carlo (MCMC) to draw posterior samples.

4 EXPERIMENTS

We evaluate the InfoNCE-λ objective and its application to SBI-BOED across four settings. We first consider a fixed-design SBI benchmark (Two Moons), which isolates the optimization dynamics of the InfoNCE-λ objective and enables the study of *simulation active learning* via EPIG when ξ is held constant (§3.2).

We then evaluate SBI-BOED on three design optimization tasks: (ii) a noisy linear model probing high-dimensional MI lower-bound optimization and the effect of λ regularization; and (iii–iv) two sequential BOED problems, one time-dependent (SIR) and one i.i.d. (BMP). For BOED baselines, we compare against MINEBED-BO [Kleinegesse and Gutmann, 2020] and, when the simulator is differentiable, iDAD [Ivanova et al., 2021] and MINEBED [Kleinegesse and Gutmann, 2021]. We also include random and Sobol design SBI baselines to separate the value of design optimization from the likelihood objective (Tables 4 and 5).

For the BOED experiments (ii–iv), we report EIG as the primary information-gain metric and assess posterior quality via visual comparisons to NUTS samples and calibration metrics such as L-C2ST (with SBC in the appendix) [Linhart et al., 2024]. Full experimental details are provided in Appendix E.

4.1 EVALUATION OF MI ON TWO MOONS

We first study how well our amortized generative model performs in a *fixed-design* (non-BOED) setting to isolate the behavior of the InfoNCE-λ objective. This experiment highlights the tradeoffs between regularization in Equation (9) and the number of contrastive samples L in Equation (4). As noted by Miller et al. [2024], Glaser et al. [2022], aggressively optimizing MI between y and θ can degrade likelihood fidelity, as reflected in the validation log probability $\mathbb{E}\,\log p_\phi(y \mid \theta)$. We therefore sweep λ and L and evaluate both information-gain proxies (EIG / MI lower bound) and likelihood quality (validation log probability). The resulting sweep is shown in Figure 1: increasing L improves the MI lower bound and generally improves likelihood validation metrics, while λ controls the tradeoff between emphasizing the MI objective and likelihood accuracy. We note that, with our chosen parameterization, optimizing Equation (4) does not eliminate the known mode collapse behavior in Two Moons likelihood-based models [Greenberg et al., 2019]; an example is shown in Appendix F.

InfoNCE-λ active learning In addition to this diagnostic sweep, Two Moons provides a simple setting to evaluate *active learning of simulations* when ξ is fixed: rather than drawing simulator queries $\theta \sim p(\theta)$ i.i.d., we select θ using EPIG computed from the current learned likelihood with MC-dropout (§3.2). We track calibration over sequential rounds using C2ST trajectories and summarize final local calibration using the L-C2ST statistic. Across this sweep, λ has the most pronounced effect on calibration: smaller λ yields better C2ST trajectories and lower final L-C2ST values (better calibration), while larger λ tends to worsen local calibration (Fig. 5). The contrastive top-K has a less systematic, non-monotone effect and primarily acts as a variance-control knob; increasing K is relatively inexpensive (increase bank size) but needs to be balanced with the number of θ samples. Finally, we compare EPIG-based querying to an i.i.d. (random) θ baseline and observe that EPIG yields more efficient improvement in calibration over rounds under the same simulation budget (details in Appendix F.2).

4.2 NOISY LINEAR MODEL

In our first BOED task, we evaluate how SBI-BOED performs on the EIG metric with increasing design dimensions. We follow Kleinegesse and Gutmann [2020] and evaluate optimal designs on a noisy linear model where a response variable y has a linear relationship with experimental designs ξ, which is determined by values of the model parameters $\boldsymbol{\theta} = [\theta_0, \theta_1]$, which model the offset and gradient. We would like to optimize the value of D measurements to estimate the posterior of $\boldsymbol{\theta}$, and so create a design vector $\boldsymbol{\xi} = [\xi_1, \dots, \xi_D]^\mathsf{T}$. Each design, ξ_i returns a measurement

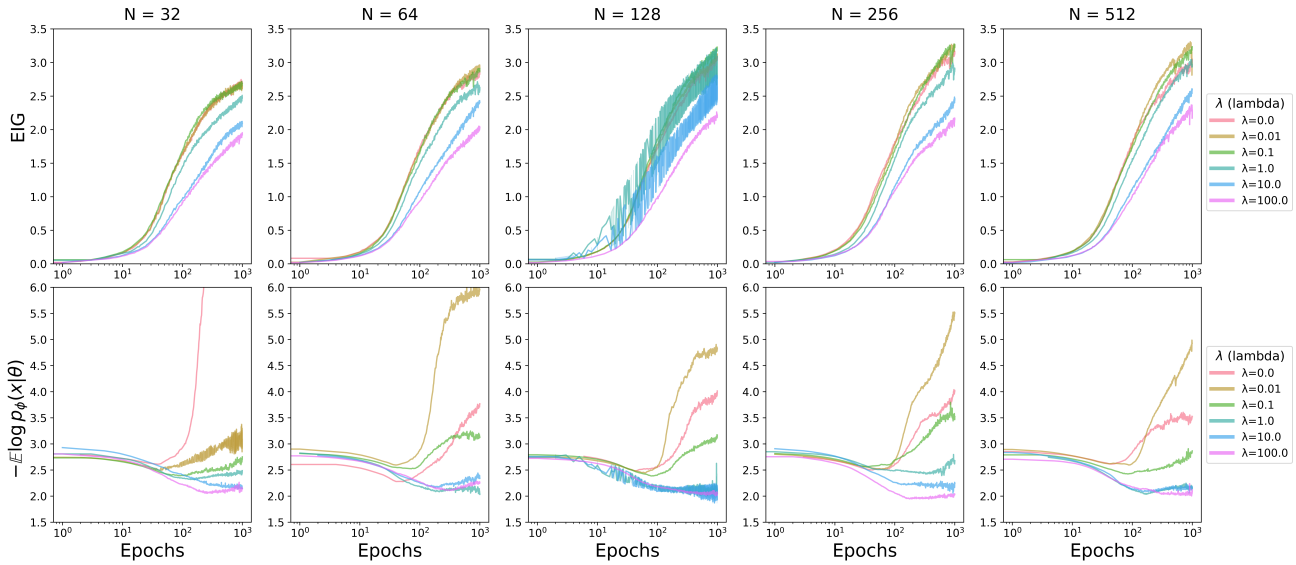

Figure 1: Comparison on the Two Moons task of the EIG and the validation loss $-\mathbb{E}\log p_\phi(y|\theta)$ across varying number of contrastive samples ($L = N - 1$) and λ regularization. Increasing number of contrastive samples improves the information lower bound and likelihood validation metrics. The λ parameter helps improve the likelihood accuracy at the expense of MI.

y_i, which results in the data vector $\boldsymbol{y} = [y_1, \ldots, y_D]^\mathsf{T}$. We use a Gaussian noise source $\mathcal{N}(\epsilon; 0, 1)$ and Gamma noise source $\Gamma(\nu; 2, 2)$. The model is then

$$\boldsymbol{y} = \theta_0 \mathbf{1} + \theta_1 * \boldsymbol{\xi} + \boldsymbol{\epsilon} + \boldsymbol{\nu}, \tag{12}$$

where $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_D]^\mathsf{T}$ and $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_D]^\mathsf{T}$ are i.i.d. samples. We used a prior distribution on model parameters as $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; 0, 3^2)$. We evaluate SBI-BOED, *purely using design gradients and no distribution over designs*, to examine how changing the λ regularization parameter in Equation (9) influences the resulting MI bound and design optimization stability with increasing design dimension.

For all design dimensions, we randomly initialize designs $\xi \in [-10, 10]$. For SBI-BOED, we chose $N = 10$, the number of batch samples $\boldsymbol{y} \sim p(\boldsymbol{y}|\theta_0, \boldsymbol{\xi})$, and $L = 50$ contrastive samples. We used a neural spline flow with training details in Appendix E. We show the plots of the EIG in Figure 7, where we can see that SBI-BOED optimizes a lower bound of MI, which increases for higher design dimensions. This corresponds with intuition that gathering more data results in more information. Similar to the two moons task, we see how the choice of λ influences the stability of MI estimation for SBI-BOED in high design dimensions where there seems to be a tradeoff with choice of regularization and MI estimation. This may be due to regularization limiting gradient updates of designs that lead to out of distribution data distributions, but is balanced by the stability in the training objective.

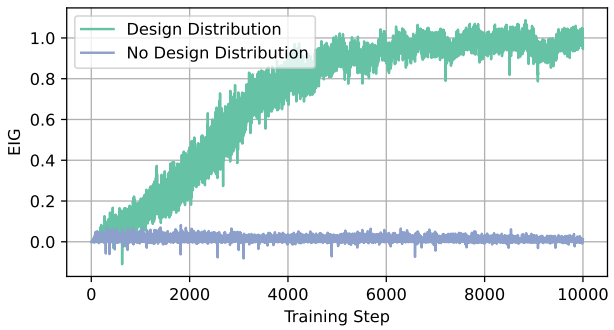

Figure 2: Training (higher is better) without a design distribution in the first round of optimization for the SIR model fails to find designs with high rewards.

4.3 SIR MODEL

We next evaluate the performance of SBI-BOED with different amounts of regularization on a real-world implicit likelihood model that has a differentiable simulator as a comparison to alternative methods. We use the Susceptible, Infected, or Recovered (SIR) epidemiology model specified in Ivanova et al. [2021]. The SIR model represents a fixed population that has three groups: susceptible, infected, and recovered. Individuals transition from susceptible to infected with a parameter β and from infected to recovered with parameter γ. Therefore, the model parameters are $\theta := [\beta, \gamma]$ and the design space Ξ, consists of a time κ to measure individuals to infer parameters θ. We measured the EIG, L-C2ST, and median distance from the observed data point to samples generated by the final posterior distribution with ground truth parameters $\theta = [0.7399, 0.0924]$. Table 2 summarizes the results. We find that iDAD and differentiable

Table 2: SIR (Section 4.3) and BMP (Section 4.4) results: comparison of EIG, L-C2ST, and median distance (Med.) for two tasks with $T = 2$ experimental design rounds for the SIR model and $T = 3$ for the BMP model. For the InfoNCE-λ baselines, we report the best λ by EIG for each task and design policy. We report mean and standard error over 3 experiments using different seeds. On BMP, wall-clock time was $2,235 \pm 9$ s for SBI-BOED and $7,095 \pm 2,519$ s for MINEBED-BO under the matched-budget sweep.

Method	**SIR (T=2)**			**BMP (T=3)**		
	EIG (↑)	L-C2ST† (↓)	Med. (↓)	EIG (↑)	L-C2ST† (↓)	Med. (↓)
Random + NLE	0.02 ± 0.02	0.01 ± 0.01	78.97 ± 33.23	0.02 ± 0.01	0.01 ± 0.01	11.49 ± 5.33
Random + InfoNCE-λ ($\lambda = 0.01$)	0.93 ± 0.36	0.01 ± 0.01	77.82 ± 33.09	0.46 ± 0.07	0.01 ± 0.01	10.86 ± 5.72
MINEBED(-BO)	$\mathbf{2.69 \pm 0.03}$	0.07 ± 0.05	71.26 ± 5.66	$\mathbf{14.23 \pm 0.60}$	0.25 ± 0.00	0.82 ± 0.06
iDAD (InfoNCE)	2.67 ± 0.02	0.10 ± 0.04	71.65 ± 2.78	-	-	-
SBI-BOED ($\lambda = 1$)	1.01 ± 0.03	0.05 ± 0.01	47.99 ± 2.62	10.39 ± 0.01	0.01 ± 0.01	0.61 ± 0.01
SBI-BOED ($\lambda = 0.1$)	1.47 ± 0.23	0.03 ± 0.01	$\mathbf{46.85 \pm 3.21}$	10.38 ± 0.01	0.01 ± 0.01	$\mathbf{0.60 \pm 0.01}$
SBI-BOED ($\lambda = 0.01$)	1.63 ± 0.23	0.04 ± 0.02	52.85 ± 0.70	10.39 ± 0.01	0.01 ± 0.01	0.61 ± 0.01

† L-C2ST values smaller than 0.01 are displayed as 0.01 for readability.

MINEBED achieve the best information gain, likely thanks to using differentiability of the simulator, but perform worse than SBI-BOED on L-C2ST and median distance metrics. This may indicate the learned critic is not as accurate as the amortized likelihood trained in SBI-BOED. Among SBI-BOED methods, we see that more regularization generally leads to improved median distance, and all perform roughly the same on calibration. Qualitative predictive, posterior, and SBC diagnostics for a representative SIR posterior are shown in Figure 9. Thus, improved information gain *does not* necessarily correlate with improved downstream prediction, as measured by the median distance metric. This highlights an important caveat to BOED methods that we should not rely on a single metric and holistically assess the resulting inference model in terms of accuracy and calibration before making decisions (performing experiments).

Ablation Study We compare the performance of SBI-BOED with and without a design distribution when maximizing the EIG in Figure 2 in the first round of design optimization. The prior used in the SIR model creates a challenge for myopic and local design optimization by introducing regions with little EIG signal, such as the flatter part of the SIR curve shown in Appendix G. The design distribution overcomes this challenge by querying diverse sets of designs with better gradients to optimize Equation (9).

4.4 BONE MORPHOGENETIC PROTEIN MODEL

The Bone Morphogenetic Protein (BMP) pathway is important in developmental and disease processes. A mass action kinetics model was proposed for the BMP pathway [Antebi et al., 2017]. The one-step model proposed by Su et al. [2022] models type I (A) and type II (B) receptors coupling with a ligand (L) to form a trimer complex (T) with equilibrium affinity K in a single step as $A + B + L \overset{K}{\rightleftharpoons} T$. The trimeric complex then phosphorylates SMAD protein to send a downstream gene expression signal, S, with a certain efficiency, ε as $\varepsilon T = S$. Steady-state gene expression signals can be simulated using convex optimization in a closed-box optimization process. Thus, we would like to infer the parameters K and ε given data, S. In the experimental design context, we would like to optimize the ligand concentration, $\boldsymbol{L}^*$, used in an experiment so to gain the most information about the parameters of this model of a biological signaling pathway. The model parameters are $\theta := [K, \varepsilon]$ and we use ground truth parameters $\theta = [0.85, 0.85]$. We show qualitative predictive, posterior, and SBC diagnostics for BMP in Figure 10. We evaluated a 1D design dimension for the ligand concentration on a uniform range from 10^{-3}ng/mL to 10^{3}ng/mL. We compare against the same benchmarks except the iDAD algorithm, which cannot work on this non-differentiable simulator. We train the EIG for 500 steps in each design optimization round because of the expensive simulation time required in each round (about 15 seconds). All SBI-BOED design algorithms perform well in this setting, partially due to the bias in larger concentrations containing more information. Under the matched-budget MINEBED-BO comparison, MINEBED-BO obtains higher estimated EIG than SBI-BOED (14.23 ± 0.60 vs. 10.39 ± 0.01), but SBI-BOED achieves better posterior metrics, with lower L-C2ST (0.01 ± 0.01 vs. 0.25 ± 0.00) and lower median distance (0.60 ± 0.01 vs. 0.82 ± 0.06). Thus, the model with the best EIG does not correspond to the most-calibrated model nor the best accuracy predictions.

5 DISCUSSION

This work connects BOED-style mutual-information maximization with likelihood learning in SBI by showing that an InfoNCE-based MI lower bound with a normalized conditional density model contains a likelihood-fitting term and a finite-contrastive marginal-likelihood approximation while optimizing for informative queries. Empirically, our Two Moons study illustrates the central tradeoff induced by the InfoNCE-λ objective: regularization and contrastive sampling affect not only MI estimates (EIG) but also likelihood fidelity and posterior calibration, and EPIG-based simulation active learning can improve calibration over sequential rounds under a fixed simulation budget. In sequential BOED benchmarks, we further observe that designs with higher EIG do not necessarily yield better downstream inference, such as posterior calibration and predictive accuracy diverging from information-gain objectives, and highlighting the need to evaluate BOED methods beyond EIG alone.

On the optimization side, we introduced practical stabilizers for closed-box simulators and purely generative surrogates, including optimizing a distribution over designs to mitigate sparse-reward landscapes and design checkpointing to avoid local optima during gradient-based search. Across differentiable and non-differentiable simulators, SBI-BOED achieves competitive or improved information gain while consistently improving calibration and predictive metrics relative to strong BOED baselines, supporting the view that simulation efficiency should be measured by the quality of the resulting posterior, not solely by utility values.

Finally, while we instantiate SBI-BOED with conditional normalizing flows, the approach extends to any likelihood-based or likelihood-bounded generative model that supports evaluating (or lower-bounding) $\log p_\phi(y \mid \theta, \xi)$. This suggests natural extensions to more expressive generative families such as diffusion models and flow-matching [Ho et al., 2020, Song and Ermon, 2019, Lipman et al., 2023, Ben-Hamu et al., 2024], which may better handle high-dimensional observations and offer additional flexibility for design optimization through alternative parameterizations of the data and noise spaces. However, these approaches add complexity over our simple approach that should be a starting point for scientific simulators suited to likelihood-based SBI modeling.

References

Yaron E Antebi, James M Linton, Heidi Klumpe, Bogdan Bintu, Mengsha Gong, Christina Su, Reed McCardell, and Michael B Elowitz. Combinatorial signal perception in the bmp pathway. *Cell*, 170(6):1184–1196, 2017.

Heli Ben-Hamu, Omri Puny, Itai Gat, Brian Karrer, Uriel Singer, and Yaron Lipman. D-flow: Differentiating through flows for controlled generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 3462–3483. PMLR, 21–27 Jul 2024.

Lucas Berry and David Meger. Normalizing flow ensembles for rich aleatoric and epistemic uncertainty modeling, 2023. URL https://arxiv.org/abs/2302.01312.

Christopher M. Bishop. Mixture density networks. Working Paper 4288, Aston University, 1994.

Tom Blau, Edwin V. Bonilla, Iadine Chades, and Amir Dezfouli. Optimizing sequential experimental design with deep reinforcement learning, 2022. URL https://arxiv.org/abs/2202.00821.

Tom Blau, Iadine Chades, Amir Dezfouli, Daniel Steinberg, and Edwin V. Bonilla. Statistically efficient bayesian sequential experiment design via reinforcement learning with cross-entropy estimators, 2024. URL https://arxiv.org/abs/2305.18435.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, June 2019.

Robert M Dirks, Justin S Bois, Joseph M Schaeffer, Erik Winfree, and Niles A Pierce. Thermodynamic analysis of interacting nucleic acid strands. *SIAM review*, 49(1): 65–88, 2007.

Junpeng Dong, Casper Jacobsen, Mohamed Khalloufi, Mohammad Akram, Wei Liu, Karthik Duraisamy, and Xun Huan. Variational bayesian optimal experimental design with normalizing flows. *Computer Methods in Applied Mechanics and Engineering*, 433:117457, 2025. doi: 10.1016/j.cma.2024.117457. URL https://doi.org/10.1016/j.cma.2024.117457.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Conor Durkan, Iain Murray, and George Papamakarios. On contrastive learning for likelihood-free inference. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2771–2781. PMLR, 13–18 Jul 2020.

Adam Foster, Martin Jankowiak, Eli Bingham, Paul Horsfall, Yee Whye Teh, Tom Rainforth, and Noah Goodman. Variational Bayesian optimal experimental design. *Advances in Neural Information Processing Systems*, March 2019.

Adam Foster, Martin Jankowiak, Matthew O'Meara, Yee Whye Teh, and Tom Rainforth. A unified stochastic gradient approach to designing bayesian-optimal experiments. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2959–2969. PMLR, 26–28 Aug 2020.

Scott Fujimoto, Wei-Di Chang, Edward J. Smith, Shixiang Shane Gu, Doina Precup, and David Meger. For sale: State-action representation learning for deep reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

Pierre Glaser, Michael Arbel, Samo Hromadka, Arnaud Doucet, and Arthur Gretton. Maximum likelihood learning of unnormalized models for simulation-based inference. *arXiv preprint arXiv:2210.14756*, 2022.

David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, pages 2404–2414. PMLR, 2019.

Sam Griesemer, Defu Cao, Zijun Cui, Carolina Osorio, and Yan Liu. Active sequential posterior estimation for sample-efficient simulation-based inference, 2024. URL `https://arxiv.org/abs/2412.05590`.

Michael U. Gutmann, Jukka Cor, and er. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(125):1–47, 2016. URL `http://jmlr.org/papers/v17/15-017.html`.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, Volodimir Begy, and Gilles Louppe. A crisis in simulation-based inference? beware, your posterior approximations can be unfaithful. *Transactions on Machine Learning Research*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Desi R. Ivanova, Adam Foster, Steven Kleinegesse, Michael U. Gutmann, and Tom Rainforth. Implicit deep adaptive design: Policy-based experimental design without likelihoods. In *Advances in Neural Information Processing Systems*, volume 34, pages 25785–25798, 2021.

Steven Kleinegesse and Michael U Gutmann. Efficient bayesian experimental design for implicit models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 476–485. PMLR, 2019.

Steven Kleinegesse and Michael U. Gutmann. Bayesian experimental design for implicit models by mutual information neural estimation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 5316–5326, 2020. arXiv: 2002.08129 Publication Title: arXiv.

Steven Kleinegesse and Michael U. Gutmann. Gradient-based Bayesian Experimental Design for Implicit Models using Mutual Information Lower Bounds. May 2021. arXiv: 2105.04379.

Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, August 2020. ISSN 0162-8828.

Vincent Lim, Ellen Novoseller, Jeffrey Ichnowski, Huang Huang, and Ken Goldberg. Policy-based bayesian experimental design for non-differentiable implicit models, 2022.

Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8:293–321, 1992.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

Dennis V Lindley. *Bayesian statistics, a review*, volume 2. SIAM, 1972.

Julia Linhart, Alexandre Gramfort, and Pedro L. C. Rodrigues. L-c2st: Local diagnostics for posterior approximations in simulation-based inference. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 2023.

Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in neural information processing systems*, 30, 2017.

Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pages 343–351. PMLR, 2021.

Benjamin Kurt Miller, Marco Federici, Christoph Weniger, and Patrick Forré. Simulation-based inference with the generalized kullback-leibler divergence. In *1st Workshop on the Synergy of Scientific and Machine Learning Modeling at ICML2023*, 2023.

Benjamin Kurt Miller, Christoph Weniger, and Patrick Forré. Contrastive Neural Ratio Estimation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2024.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*, 2013.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21 (132):1–62, 2020.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, November 2010. ISSN 1557-9654. doi: 10.1109/tit.2010.2068870.

Rafael Orozco, Felix J. Herrmann, and Peng Chen. Probabilistic bayesian optimal experimental design using conditional normalizing flows, 2024. URL https://arxiv.org/abs/2402.18337.

George Papamakarios and Iain Murray. Fast ϵ-free inference of simulation models with bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, volume 29, pages 1028–1036, 2016.

George Papamakarios, David C. Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature reviews Drug discovery*, 9(3):203–214, 2010.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.

Tom Rainforth, Robert Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting monte carlo estimators. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4267–4276, 2018.

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

Wenlong Shen, Junpeng Dong, and Xun Huan. Variational sequential optimal experimental design using reinforcement learning. *Computer Methods in Applied Mechanics and Engineering*, 444:118068, 2025. doi: 10.1016/j.cma.2025.118068. URL https://doi.org/10.1016/j.cma.2025.118068.

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning, 2023. URL https://arxiv.org/abs/2304.08151.

Freddie Bickford Smith, Adam Foster, and Tom Rainforth. Making better use of unlabelled data in bayesian active learning. In *International conference on artificial intelligence and statistics*, pages 847–855. PMLR, 2024.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Christina J Su, Arvind Murugan, James M Linton, Akshay Yeluri, Justin Bois, Heidi Klumpe, Matthew A Langley, Yaron E Antebi, and Michael B Elowitz. Ligand-receptor promiscuity enables cellular addressing. *Cell systems*, 13 (5):408–425, 2022.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

A MUTUAL INFORMATION-BASED LIKELIHOOD OPTIMIZATION DERIVATION & PROOFS

We provide a derivation and proof of how optimizing a lower bound of the mutual information is the same as minimizing the KL divergence in (3). We also discuss alternative BOED bounds and how adding a distribution of designs to the MI approximation keeps it a valid lower bound.

A.1 DERIVATION OF OPTIMIZATION OF THE MUTUAL INFORMATION

We investigate how optimizing the mutual information within the SBI-BOED loss framework leads to an implicit optimization of the likelihood. Following the approach outlined by Miller et al. [2024], we optimize our approximation to the true likelihood by minimizing the KL divergence:

$$D_{\mathrm{KL}}(p(y|\theta)||p_\phi(y|\theta)). \tag{13}$$

Within the SBI framework, we draw samples of parameters θ from the prior $p(\theta)$ and of data y conditioned on these parameters from the likelihood $p(y|\theta)$. This sampling enables us to approximate the expected KL divergence across the parameter space:

$$\mathbb{E}_{p(\theta)}[D_{\mathrm{KL}}(p(y|\theta)||p_\phi(y|\theta))]. \tag{14}$$

Now we express the KL divergence as the expectation of the log ratio of probabilities:

$$\begin{aligned}
\mathbb{E}_{p(\theta)}[D_{\mathrm{KL}}(p(y|\theta)||p_\phi(y|\theta))] &= \mathbb{E}_{p(\theta,y)}\left[\log \frac{p(y|\theta)}{p_\phi(y|\theta)}\right] && (15)\\
&= \mathbb{E}_{p(\theta,y)}\left[\log \frac{p(y|\theta)}{p(y)}\frac{p(y)}{p_\phi(y|\theta)}\right] && (16)\\
&= I(\theta;y) + \mathbb{E}_{p(\theta,y)}\left[\log \frac{p(y)}{p_\phi(y|\theta)}\right] && (17)\\
&= I(\theta;y) + \mathbb{E}_{p(y)}\left[\log p(y)\right] - \mathbb{E}_{p(\theta,y)}\left[\log p_\phi(y|\theta)\right]. && (18)
\end{aligned}$$

Since the KL divergence is always non-negative, the optimization process aims to find the parameters ϕ that minimize this expected divergence, implicitly maximizing the mutual information between parameters.

We now base our proof of Proposition 3.1 as follows:

Proof. Let us consider the MI between random variables Θ and Y, where $y \sim Y$ and $\theta \sim \Theta$, then we have

$$I(\theta;y) = \mathbb{E}_{p(\theta,y)}\left[\log \frac{p(y|\theta)}{p(y)}\right]. \tag{19}$$

In the SBI setting, we would like to approximate the likelihood $p_\phi(y|\theta)$. We can approximate the marginal likelihood by the Strong Law of Large Numbers using $\mathbb{E}_{p(\theta)}\left[p(y|\theta)\right] = \int_\theta p(y|\theta)p(\theta)d\theta \approx \frac{1}{L}\sum_\ell^L p(y|\theta_\ell)$. The bound gets tighter as $L \to \infty$ and as the likelihood better-approximates the true likelihood. Assuming we are optimizing the parameters ϕ to maximize this objective, then we have

$$\hat{\phi} = \arg\max_{\phi} \mathbb{E}_{p(\theta,y)} \left[\log \frac{p_\phi(y|\theta)}{\hat{p}(y)} \right] \tag{20}$$

$$= \arg\min_{\phi} \left(I(\theta; y) - \mathbb{E}_{p(\theta,y)} \left[\log \frac{p_\phi(y|\theta)}{\hat{p}(y)} \right] \right) \tag{21}$$

$$= \arg\min_{\phi} \mathbb{E}_{p(\theta,y)} \left[\log \frac{p(y|\theta)}{p(y)} \frac{\hat{p}(y)}{p_\phi(y|\theta)} \right] \tag{22}$$

$$\approx \arg\min_{\phi} \left(\mathbb{E}_{p(\theta)} D_{KL}(p(y|\theta)||p_\phi(y|\theta)) - \mathbb{E}_{\hat{p}(y)} D_{KL}(\hat{p}(y)||p(y)) \right) \tag{23}$$

$$= \arg\min_{\phi} \left(\mathbb{E}_{p(\theta)} D_{KL}(p(y|\theta)||p_\phi(y|\theta)) + \mathbb{E}_{\hat{p}(y)} \left[\log \hat{p}(y) \right] - \text{const.} \right) \tag{24}$$

where the marginal likelihood is approximated as $\hat{p}(y) = \frac{1}{L} \sum_{i=1}^{L} p_\phi(y|\theta_i), \quad \theta_i \sim p(\theta)$. Thus, as $L \to \infty$, maximizing the InfoNCE bound from Equation (4) is asymptotically equivalent to optimizing the likelihood objective in Equation (3), with a finite-L marginal-likelihood approximation term whose accuracy depends on the number of contrastive samples L.

□

Remark. In the BOED setting, the MI is simply conditional on a design, ξ, as $I(\theta; y|\xi)$ which allows for gradient-based optimization by a pathwise gradient estimator as detailed in §2.2. An interesting insight to this derivation is that InfoNCE bound used to maximize a lower bound of MI may be biased by the reverse KL of the marginal likelihood to models that better represent the data. This may surface in BOED with designs preferring models that do not cover enough of the potential parameter spaces that explain the data.

B THEORETICAL ANALYSIS OF SBI-BOED COMPONENTS

We provide more theoretical analysis of the impact of the λ parameter of $I_{\text{NCE-}\lambda}$ on predictive accuracy, estimating the EIG, and gradients of EIG and normalizing flow model parameters, ϕ.

B.1 ANALYSIS OF THE λ REGULARIZATION PARAMETER

We experimentally demonstrated in §4.1 that the λ regularization parameter in SBI-BOED influences both the likelihood's validation accuracy and the EIG bound. We theoretically analyze both scenarios as well as λ influence on optimization gradients in the following sections.

Influence of λ on likelihood prediction accuracy Starting from Equation (22) of the previous section, including the λ regularization parameter results in

$$\hat{\phi} = \arg\min_{\phi} \mathbb{E}_{p(\theta,y)} \left[\log \frac{p(y|\theta)}{p(y)} \frac{\hat{p}(y)}{p_\phi(y|\theta)} \right] + \lambda \, \mathbb{E}_{p(\theta,y)} \log[p_\phi(y|\theta)] \tag{25}$$

$$= \arg\min_{\phi} \mathbb{E}_{p(\theta,y)} \left[\log \frac{p(y|\theta)}{p(y)} \frac{\hat{p}(y)}{p_\phi(y|\theta)} p_\phi(y|\theta)^\lambda \right]. \tag{26}$$

We now focus on the contribution of the added $p_\phi(y|\theta)^\lambda$ term grouped with the likelihood's KL divergence from Equation (24). Reformulating this term, we express the λ-regularized KL divergence as

$$D_{\text{KL}}(p(y|\theta)\|p_\phi(y|\theta)^{1-\lambda}) = \mathbb{E}_{p(y|\theta)} \left[\log \frac{p(y|\theta)}{p_\phi(y|\theta)^{1-\lambda}} \right] \tag{27}$$

$$= \mathbb{E}_{p(y|\theta)} \left[\log p(y|\theta) - (1-\lambda) \log p_\phi(y|\theta) \right]. \tag{28}$$

This modified divergence adjusts the weighting of the log-likelihood term $\log p_\phi(y|\theta)$ based on λ. We show how λ influences optimization, omitting the trivial case when $\lambda = 0$:

- For $\lambda > 0$: The term $(1 - \lambda)$ reduces the weight of the approximate likelihood, ensuring broad coverage of the parameter space while emphasizing accuracy.

- For $\lambda < 0$: The term $(1 - \lambda) > 1$ amplifies the weight of the approximate likelihood, leading to mode-seeking behavior that prioritizes high-probability regions at the expense of the tails.

Influence of λ on the EIG We start from Equation (10) and rephrase it here using the shortened marginal likelihood notation $\hat{p}(y|\xi)$ for convenience

$$\mathbb{E}_{p(\theta,y|\xi)}\left[\log \frac{p_\phi(y|\theta,\xi)^{1+\lambda}}{\hat{p}(y|\xi)}\right]. \tag{29}$$

Expanding the log ratio, we can separate the contributions of the likelihood and the marginal likelihood

$$\mathbb{E}_{p(\theta,y|\xi)}\left[\log \frac{p_\phi(y|\theta,\xi)^{1+\lambda}}{\hat{p}(y|\xi)}\right] = \mathbb{E}_{p(\theta,y|\xi)}\left[(1+\lambda)\log p_\phi(y|\theta,\xi) - \log \hat{p}(y|\xi)\right]. \tag{30}$$

The parameter λ influences the scaling of the likelihood term $(1 + \lambda) \log p_\phi(y|\theta, \xi)$ and indirectly affects the marginal likelihood $\hat{p}(y|\xi)$ through its dependence on $p_\phi(y|\theta, \xi)$. The EIG linearly depends on the λ value. Positive λ decrease the EIG estimate while negative values increase the EIG at the expense of likelihood approximation, as previously discussed. The decrease in EIG with increasing λ aligns with empirical observations (Figure 1).

C RELATED WORK

In BOED, we focus on the setting of myopic gradient-based experimental design. Previously, Foster et al. [2019] proposed various likelihood-free information bounds that could work in the SBI setting based on variational inference estimators but did not make use of a normalized generative model like a normalizing flow. This is problematic for use in sequential SBI methods that make use of a normalized likelihood or posterior functions. Kleinegesse and Gutmann [2020] developed MINEBED to simultaneously optimize experimental designs and a critic that could draw samples from the posterior, but relied on a differentiable simulator or using Bayesian optimization. Ivanova et al. [2021] extended both previous works and developed policy-based experimental designs for non-myopic experimental designs, but whose critics also relied on differentiable simulators - simulators whose inputs can be connected to a differentiable computation graph, which may not be available for scientific simulators. An RL approach [Lim et al., 2022] optimized a critic without requiring a differentiable simulator but relies on computationally expensive RL algorithms and may not be feasible for practitioners with scientific simulators that can take a significant amount of time to simulate. This is exemplified in our biological experiment where each round of simulation takes about ten seconds.

Recent posterior- and variational-BOED methods also use normalizing flows or amortized posteriors to optimize design objectives [Dong et al., 2025, Orozco et al., 2024, Shen et al., 2025]. These approaches are complementary to SBI-BOED, but they largely target a different operating regime. Dong et al. [2025] can handle implicit likelihoods through sampling, but simultaneous design optimization relies on gradients through the design path or an equivalent differentiable construction. Orozco et al. [2024] similarly uses conditional normalizing flows with differentiable simulator/design structure. Shen et al. [2025] addresses sequential design with reinforcement learning, but shifts the cost to large numbers of simulated rollouts. In contrast, our evaluation focuses on expensive closed-box simulators where simulator/design gradients are unavailable, simulator calls are budget-limiting, and the learned likelihood should remain reusable for posterior inference after design optimization.

Recent work connecting SBI methods to MI-based optimization studied how to stably train a discriminative and generative SBI model [Miller et al., 2023], and applying a nonparametric function to the selection of data points to reduce variance of the resulting MI estimate [Glaser et al., 2022]. These methods take a complementary approach to ours but require a generative and discriminative model whereas we only require a generative model. Additionally, we study the utility of MI-based optimization of SBI models in BOED.

D IMPLEMENTATION OF SBI-BOED IN POSTERIOR ESTIMATION, RATIO ESTIMATION, AND CHECKPOINTING

D.1 APPLYING THE INFONCE BOUND TO NEURAL POSTERIOR ESTIMATION

We demonstrate the theoretical basis for optimizing a Neural Posterior Estimation (NPE) network and experimental designs, which learns a surrogate posterior conditioned on simulated data, $p_\phi(\theta|y,\xi)$.

Neural Posterior Estimation Previous methods for directly estimating the posterior by maximum likelihood estimation struggled with bias Papamakarios and Murray [2016] or variance Lueckmann et al. [2017]. An alternative method was developed by Greenberg et al. [2019] to recover the unbiased posterior,

$$p(\theta|y) \approx \frac{p_\phi(\theta|y)/p(\theta)}{Z_\phi(y)} \tag{31}$$

by calculating the normalizing constant $Z_\phi = \int p_\phi(\theta|y)/p(\theta)d\theta$. They then minimize $\mathcal{L}_{\text{NP}}(\phi) = \mathbb{E}_{p(y|\theta)p(\theta)}[-\log p_\phi(\theta|y)]$ to return an amortized posterior. Except for mixture density networks [Bishop, 1994], the normalizing constant is difficult to approximate. Greenberg et al. [2019] alternatively proposed to approximate the intractable normalizing constant by replacing the integral with a summation term that takes draws of the parameters from a proposal set Θ such that the posterior is approximately:

$$p(\theta|y) \approx \frac{p_\phi(\theta|y)/p(\theta)}{\sum_{\theta'\in\Theta} p_\phi(\theta'|y)/p(\theta')}. \tag{32}$$

Applying NPE in BOED We can use this in experimental design by again using the reparameterization trick to pass gradients back to the conditional ξ inputs to the approximate posterior. In some cases, it may be more desirable to directly infer a posterior distribution instead of a likelihood. For example, if the data, y_o, to train a flow is computationally infeasible but whose latent parameters, θ, is sufficiently small for use in a normalizing flow. Since the posterior is proportional to the likelihood, $p(\theta|y) \propto p(y|\theta)p(\theta)$, we can replace the likelihood in Equation (6) with a posterior

$$\mathcal{L}_{\text{NCE-NPE}}(\xi,\phi,L) \coloneqq \mathbb{E}_{p(\theta_0)p(y|\theta_0,\xi)p(\theta_{1:L})}\frac{p_\phi(\theta_0|y,\xi)/p(\theta_0)}{\frac{1}{1+L}\sum_{\ell=0}^{L} p_\phi(\theta_\ell|y,\xi)/p(\theta_\ell)}, \tag{33}$$

which is a biased lower bound of the MI by a factor of the prior $p(\theta)$, but whose gradient can still be used to train a density estimator and optimize experimental designs. This case should be used when it is desirable to have an amortized posterior and the absolute EIG does not matter.

D.2 THE NWJ BOED BOUND & CONTRASTIVE RATIO ESTIMATION

We demonstrate the theoretical basis for optimizing Neural Ratio Estimation (NRE) network and experimental designs, which learns a surrogate ratio estimator conditioned on experimental designs, $g_\phi(y,\theta|\xi)$.

Neural Ratio Estimation Classifiers can be used to approximate the likelihood-to-evidence ratio such that $g_\phi(\theta,y) \approx \log\frac{p(y|\theta)}{p(y)} + c(y)$, where $c(y)$ is a bias term introduced from approximating the ratio, and which can be minimized [Durkan et al., 2020, Hastie et al., 2009]. The classifier can be trained by minimizing the loss

$$\mathcal{L}_{\text{CRE}}(\phi) = -\frac{1}{B}\sum_{b=1}^{B}\log\frac{\exp(g_\phi(\theta^{(b)},y^{(b)}))}{\sum_{k=1}^{K}\exp(g_\phi(\theta^{(k)},y^{(b)}))} \tag{34}$$

over B batches of contrasting parameters. Given the use of contrastive samples, Durkan et al. [2020] called this contrastive ratio estimation (CRE) and noted the similarity between CRE and the Noise Contrastive Estimation (InfoNCE) MI lower bound proposed by Poole et al. [2019].

Applying NRE in BOED We present another relevant MI bounds to this paper, the NWJ Nguyen et al. [2010] bound has been used in BOED in Kleinegesse and Gutmann [2021], Ivanova et al. [2021]. Adjusting the bound from Ivanova et al. [2021],

$$\mathcal{L}_{\text{NWJ}}(\xi,\phi) \coloneqq \mathbb{E}_{p(\theta)p(y|\theta,\xi)}\left[g_\phi(y,\theta)\right] - e^{-1}\,\mathbb{E}_{p(\theta)p(y|\xi)}\left[\exp(g_\phi(y,\theta))\right], \tag{35}$$

where g_ϕ is a classifier that returns the probability that y belongs to θ. This function has lower bias than the InfoNCE bound but higher variance Poole et al. [2019], Song and Ermon [2020]. In practice, this can be calculated by drawing N samples from the joint distribution and shuffling to return $N(N-1)$ marginal samples. Miller et al. [2024] noted the connection between Equation (34) and the NWJ bound, and gave a tighter bound on the NWJ-based bound using their SBI-based CRE method. Their results hint at using a bound on the MI to optimize a likelihood-to-evidence ratio, and saw increasing EIG (lower bound of MI) with increasing number of contrastive samples. This is evident from Equation (35). While the NWJ bound may have less bias than the InfoNCE bound, it has higher variance that grows exponentially with the value of the true MI Song and Ermon [2020]. In the SBI setting, choosing when to use the InfoNCE or the NWJ bound will depend on what type of density estimator is required for the scientific task and ease of drawing samples from the joint distribution (simulator efficiency).

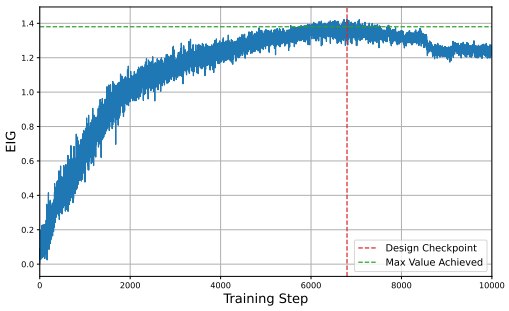

Figure 3: Design checkpointing stores the design ξ^* with the highest EIG during training.

D.3 DESIGN CHECKPOINTS

Since designs and model parameters calculate the "reward" in the form of the EIG, it is possible that the EIG may fall into a local optima by the end of training. This may be because of decaying learning rates of the design gradients or decreasing area searched by the design distribution. We found checkpointing mitigated this issue.

E EXPERIMENT METHODOLOGICAL DETAILS

For all experiments, we use a neural spline flow (NSF) normalizing flow. We specify the different parameterizations in Table 3. We used the `ReduceLROnPlateau` function to reduce the learning rate for the SIR and BMP experiments, whereas we used a constant learning rate for the linear experiment. We use the same learning rate for all flow parameters, ϕ. Notably, for the iterated experimental design, we keep the previous round's normalizing flow parameters, ϕ_{t-1} to use to return the subsequent round's posterior. We set the Adam optimizer β_2 to be 0.95 to mitigate large jumps in ξ that might destabilize training. All code is available online.

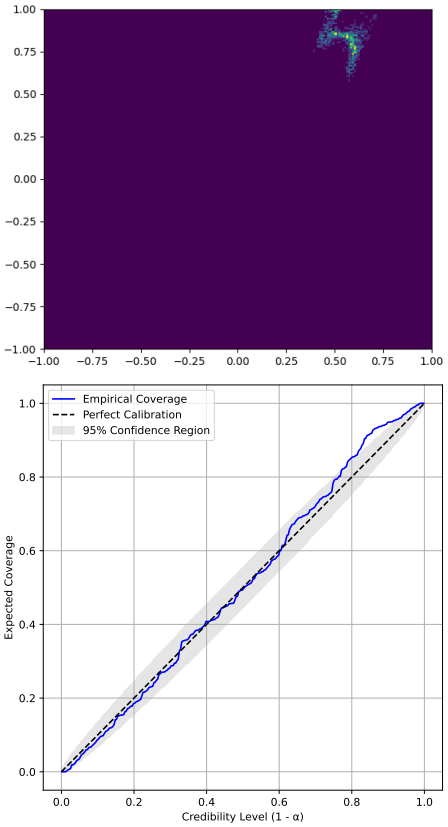

Figure 4: (Top) Posterior of the two moons experiments with mode collapse. (Bottom) Simulation-Based Calibration (SBC) of the posterior distribution for two moons.

SIR EXPERIMENT DETAILS

We follow the implementation of Ivanova et al. [2021] for the SIR model. We solve an SDE describing the process using the Euler-Maruyama method and discretize the domain $\Delta\tau = 10^{-2}$. The total population is fixed at $N = 500$. For the model parameters β and γ, we use log-normal priors such that $p(\beta) = \text{Lognorm}(0.50, 0.50^2)$ and $p(\gamma) = \text{Lognorm}(0.10, 0.50^2)$. Since solving the SDE is time-consuming, we pre-simulate data on a time grid in each round and access the relevant data regions during training.

BMP EXPERIMENT DETAILS

The BMP signaling pathway can be described by mass action kinetics of proteins binding to one another and conservation laws to describe the process of a downstream genetic expression signal reaching a steady-state based on receptors available and ligands in a cell's environment. The one-step model of BMP signaling was originally proposed by Su et al. [2022]. While the model is described by an ODE in Antebi et al. [2017], its steady-state signal is solved by convex optimization Dirks et al. [2007] as a closed-box solver.

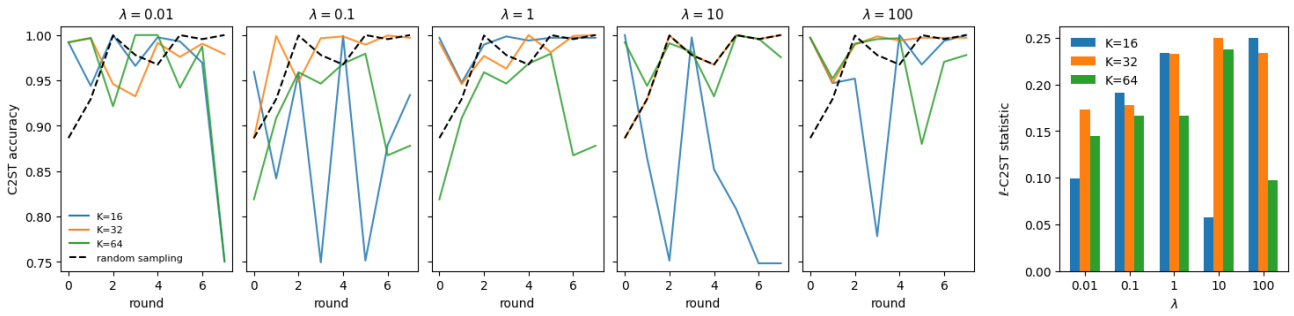

Figure 5: *Left:* C2ST (lower is better) trajectories over sequential rounds for a sweep of regularization strengths $\lambda \in \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$ (one panel per λ); curves correspond to the size of the θ^* set $K \in \{16, 32, 64\}$. *Right:* Final ℓ-C2ST statistic for the same sweep, grouped by K. Overall, larger λ tends to increase the final ℓ-C2ST statistic (worse calibration), while varying K yields non-monotone changes, consistent with K primarily affecting the variance of the epistemic approximation rather than the dominant calibration–regularization tradeoff.

F ADDITIONAL TWO MOONS RESULTS

F.1 MI OPTIMIZATION AND MODE COLLAPSE

We analyze the mode collapse of the likelihood-based two moons posterior prediction related to the MI optimization. We also show a Simulation-Based Calibration (SBC) Talts et al. [2020], Hermans et al. [2022] curve for the two moons plot. For SBC we average over the parameters and compare the average posterior parameter values to the average prior values.

F.2 ACTIVE LEARNING OF SIMULATIONS IN SBI ADDITIONAL RESULTS

Figure 5 provides a closer look at calibration diagnostics for the Two Moons task under a sweep of the objective regularization strength λ and the top-K contrastive samples collected in θ^* used to approximate model uncertainty. Across the sweep, λ has the most pronounced effect on posterior calibration: increasing λ generally increases the final ℓ-C2ST statistic, indicating worse local calibration. This suggests that, in this setting, regularization of the MI objective primarily trades off information-seeking behavior against fidelity of the learned likelihood and calibrated posterior geometry.

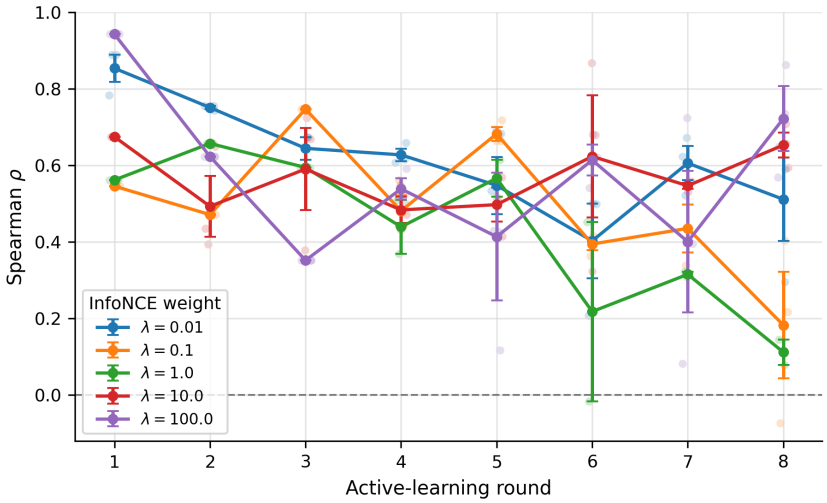

Figure 6: MC-dropout diagnostic for EPIG-based active learning on Two Moons. For each InfoNCE regularization value λ, we report the Spearman rank correlation between MC-dropout disagreement and held-out deterministic NLL over active-learning rounds; faint points show individual diagnostic runs and error bars denote standard error across the available top-K settings. Positive correlations indicate that the dropout-based acquisition signal tends to rank simulator queries with higher held-out likelihood error as more uncertain.

Table 3: Training hyperparameters.

	Linear	SIR	BMP
Batch Size	10	256	128
Number of Contrastive Samples	50	255	127
Number of Gradient Steps	10000	10000	500
ϕ & ξ Learning Rate	1×10^{-3}	1×10^{-3}	1×10^{-3}
Annealing Rate	NA	0.8	0.8
Final Learning Rate	NA	1×10^{-4}	1×10^{-4}
Gradient Clipping Threshold	NA	5	5
Hidden Layer Size	128	64	64
Number of Hidden Layers	4	2	2
Number of Flow layers (bijectors)	5	5	4
Number of bins for NSF	4	4	4

In contrast, the effect of K is less systematic. While larger K can improve calibration in some regimes, it does not yield a monotone improvement across all λ values, consistent with K acting mainly as a variance-control knob for the epistemic approximation rather than changing the dominant behavior induced by λ. Importantly, increasing K is computationally inexpensive relative to model training, since it only increases the number of target samples to keep track of when used in the acquisition/diagnostic computations. Consequently, when compute allows, one can typically increase K without materially altering the qualitative calibration–regularization trends governed by λ.

To diagnose whether MC-dropout provides a useful acquisition signal, we compare dropout disagreement against held-out likelihood error across active-learning rounds. Figure 6 shows that the dropout score is positively correlated with held-out deterministic NLL across the sweep, supporting its use as a practical uncertainty heuristic rather than a principled posterior over flow parameters.

Mutual information optimization does not avoid mode collapse Mode collapse in the two moons problem (Figure 4) is a common issue when using a likelihood-based flow model. While we initially considered analyzing this through the lens of mutual information and its interpretation as the expected log ratio of the posterior to prior:

$$I(\theta; y) = \mathbb{E}_{p(\theta,y)}\left[\log \frac{p(y|\theta)}{p(y)}\right] = \mathbb{E}_{p(\theta,y)}\left[\log \frac{p(\theta|y)}{p(\theta)}\right],$$

our analysis in Appendix A.1 revealed that the root cause of mode collapse likely stems from deficiencies in the marginal likelihood approximation. There are two primary sources of error in the marginal likelihood estimation: the use of an approximate likelihood $p_\phi(y|\theta)$ and the reliance on finite samples to estimate the expectation. These approximations can introduce biases that may contribute to the observed mode collapse. Specifically, the InfoNCE bound used to maximize a lower bound of MI may be biased towards modes that better represent the observed data but potentially underrepresent the full range of parameter spaces that could explain the data.

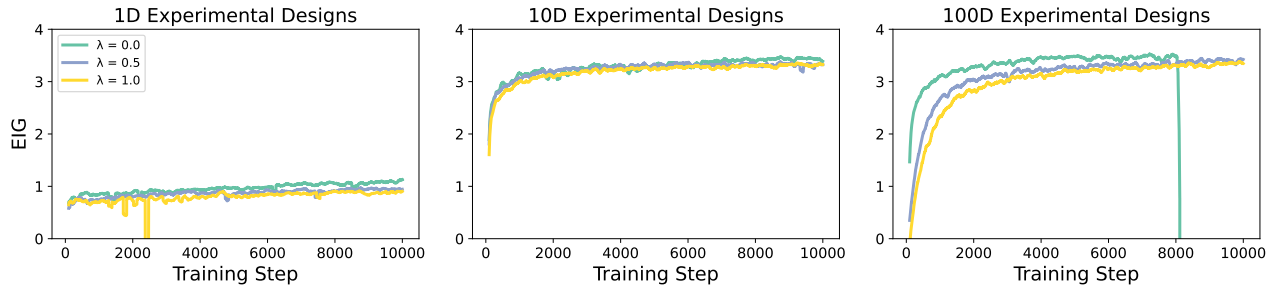

Figure 7: Comparison of the EIG across design dimensions, type of BOED, and λ regularization for the noisy linear model examining the moving average over 10 different random seed initializations. For the single design dimension, all SBI-BOED regularization levels are generally similar, with the non-regularized version being the closest to the optimal MI bound. In the higher-dimension design cases, SBI-BOED increases its EIG with more designs. In the 100-dimensional design case, we see the benefit of using λ regularization to stabilize the training of a design-dependent normalizing flow in high-dimensional input space at the cost of slightly lower EIG.

Indeed, Foster et al. [2019] suggest simultaneously optimizing a posterior distribution at the same time as a likelihood via Likelihood-Free Adaptive Contrastive Estimation (LF-ACE). This approximates the marginal likelihood with a root sample from the prior $p(\theta_0)$, and samples from an approximate posterior $\theta_i \sim q(\theta|y)$. This is essentially using samples from a posterior as importance sample estimates to help address the mode collapse of the approximate marginal likelihood. While theoretically sound, we attempted this using a normalizing flow to approximate and sample from the posterior. We found that training both the likelihood and posterior while optimizing designs to be unstable. Future work may address this with pretraining of one or both of the density estimators to improve stability of estimation.

Recent SBI work on accurate mutual information approximation [Miller et al., 2023, Glaser et al., 2022] while training amortized inference networks typically relies on using a generative model and critic in a form of importance sampling similar to LF-ACE. By addressing these issues in the marginal likelihood estimation, we may be able to develop more robust methods that avoid mode collapse in the two moons problem and related scenarios.

G EXPANDED BOED RESULTS

Linear Model We show the effect of varying design dimensions and λ regularization on the EIG metric in Figure 7. Optimizing a likelihood while optimizing experimental designs can struggle with high-dimensional designs but this is addressed with our regularization parameter. The corresponding 100-dimensional posterior in Figure 8 concentrates near the true parameters relative to the broad prior.

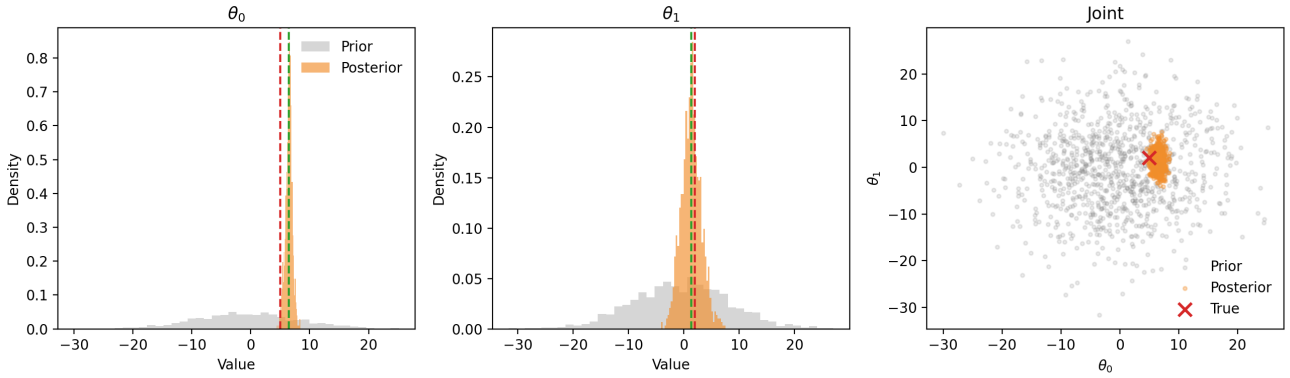

Figure 8: Posterior diagnostic for a representative 100-dimensional noisy linear design run. Marginal histograms for θ_0 and θ_1 and the joint posterior compare prior samples against posterior samples from the learned likelihood after optimizing the high-dimensional design; dashed lines and the red marker denote the ground-truth parameter values. The posterior contracts around the true parameters despite the 100-dimensional design vector.

Improving Posterior Estimation For both the SIR and BMP model, we only use a single round of inference for each round of BOED. All SBI algorithms can be refined by sequential application of drawing posterior samples conditioned on the observed data which will then help improve the quality of the likelihood Papamakarios et al. [2019]. We forego this refinement step in our study in favor of examining how our method works with different settings of our regularization parameter. Depending on the simulator, this step can be computationally expensive but performing a calibration analysis of the likelihood can help to determine whether it is worth performing refinement steps.

H APPLYING SEQUENTIAL SBI IN BOED

Until now we have focused on comparing SBI-BOED to comparable BOED methods without using a key advantage of SBI methods: refinement of a likelihood, or any inference object, with observed data before designing the next experiment. Our experimental results from Table 2 show we still outperform comparable methods in accuracy and calibration, but our calibration in Table 2 indicates there is indeed room for improvement. We demonstrate the benefits of applying sequential SBI methods to refine the likelihood after every round, resulting in a calibrated posterior. This is a key benefit of SBI-BOED over alternative BOED methods that do not allow practitioners to refine their posterior distribution before designing a subsequent experiment; Algorithm 2 summarizes this refinement procedure. We see in Figure 11 a clear improvement of the calibration of the BMP model after the last experimental design using SBI refinement.

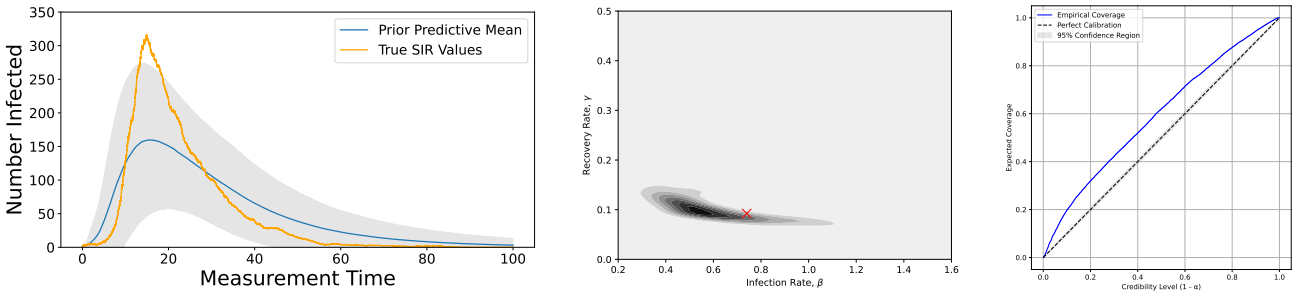

Figure 9: (*Left*) Predictive SIR trajectories with the observed trajectory superimposed. (*Middle*) Approximate posterior for the SIR model with true parameters marked in red. (*Right*) SBC expected coverage for the SIR posterior.

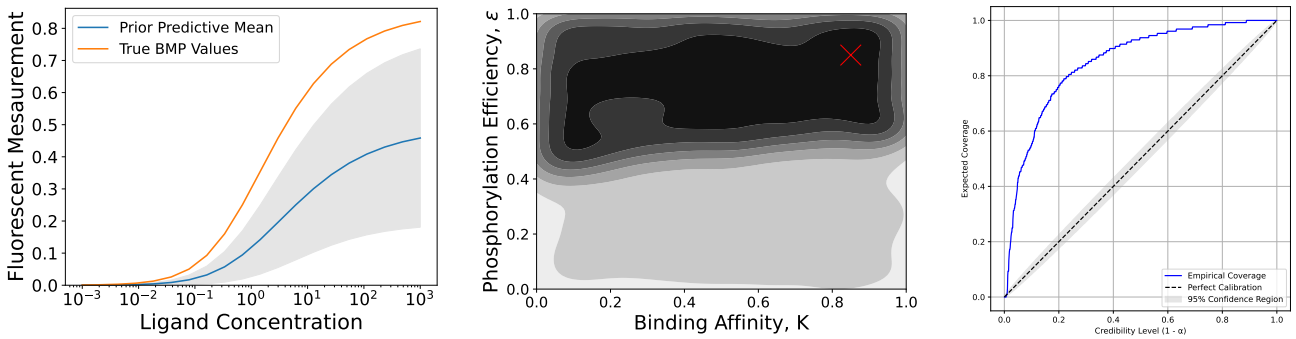

Figure 10: (*Left*) Prior predictive distributions for the BMP model with the true evaluation curve superimposed. (*Middle*) Approximate posterior for the BMP model with true parameters marked in red. (*Right*) SBC expected coverage indicating a conservative posterior, which is supported by the posterior in question.

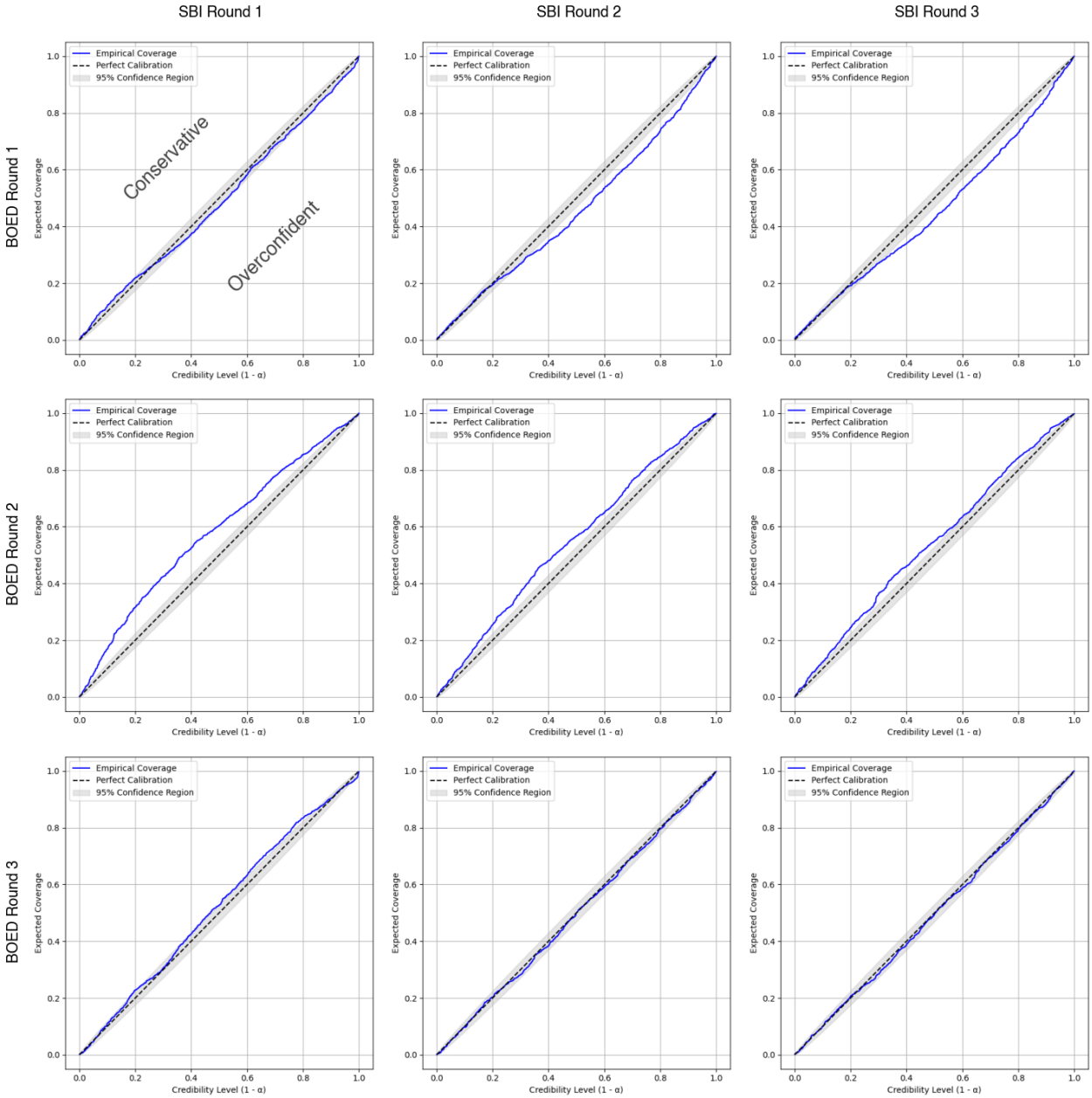

Figure 11: SBI refinement on the BMP model over 3 experimental design (BOED) rounds where the likelihood is refined after each experimental design round for three rounds of SBI. We show the SBC charts where coverage tending to the upper left indicates a conservative posterior while those to the bottom right indicate overconfident ones, and the gray area represents the 95% confidence region. For the first round of BOED we see that refinement of the likelihood object is not necessary and performing subsequent rounds of SBI might make it overconfident. For the second BOED round, we see the initial posterior is conservative and that after two rounds of SBI becomes almost calibrated, hinting that another round of SBI might make it calibrated. The final BOED round starts slightly conservative but is calibrated after two SBI rounds. This demonstrates the modularity and flexibility in improving posterior inference that SBI methods offer in between BOED rounds.

Algorithm 2 SBI-BOED with likelihood refinement

Require: Simulator $p(y|\theta, \xi)$, estimator $p_\phi(y|\theta, \xi)$, prior $p(\theta)$, number of experimental designs T, number of BOED training steps N, refinement rounds R, simulations per refinement round M
Return: Optimal designs $\{\xi_i^* \mid i = 1, \ldots, T\}$ and refined likelihood $p_\phi(y|\theta, \xi)$
Initialize observed dataset $\mathcal{D}_{\text{obs}} = \{\}$
Initialize $\hat{p}_0(\theta) \leftarrow p(\theta)$
for $t = 1, \ldots, T$ **do**
  Sample $\theta_{0:L} \sim \hat{p}_{t-1}(\theta)$
  for $n = 1, \ldots, N$ **do**
    Sample $\xi \sim \mathcal{N}_{\text{trunc}}(\mu_\xi \mid \sigma_n^2)$
    Simulate $y \sim p(y|\theta, \xi)$
    Estimate $\nabla \mathcal{L}_{\text{NCE}-\lambda}(\xi, \phi; L)$ via Equation (9)
    Update μ_ξ and ϕ using ∇_ξ and ∇_ϕ
    Checkpoint $\xi^* = \xi$ **if** $EIG_\xi > EIG_{\xi^*}$
  end for
  Observe y_t using ξ^* in an experiment
  Add (ξ^*, y_t) to dataset $\mathcal{D}_{\text{obs}}$
  Set $\hat{p}_{t,0}(\theta) \propto \left(\prod_{(\xi_i, y_i) \in \mathcal{D}_{\text{obs}}} p_\phi(y_i \mid \theta, \xi_i)\right) p(\theta)$
  Initialize refinement dataset $\mathcal{D}_{\text{sim}} = \{\}$
  for $r = 1, \ldots, R$ **do**
    for $m = 1, \ldots, M$ **do**
      Sample $\theta_m \sim \hat{p}_{t,r-1}(\theta)$
      Sample $\tilde{\xi}_m$ from $\{\xi_i^*\}_{i=1}^t$
      Simulate $\tilde{y}_m \sim p(y|\theta_m, \tilde{\xi}_m)$
      Add $(\theta_m, \tilde{\xi}_m, \tilde{y}_m)$ to $\mathcal{D}_{\text{sim}}$
    end for
    Refine $p_\phi(y|\theta, \xi)$ on $\mathcal{D}_{\text{sim}}$
    Set $\hat{p}_{t,r}(\theta) \propto \left(\prod_{(\xi_i, y_i) \in \mathcal{D}_{\text{obs}}} p_\phi(y_i \mid \theta, \xi_i)\right) p(\theta)$
  end for
  Set $\hat{p}_t(\theta) \leftarrow \hat{p}_{t,R}(\theta)$
end for

Table 4: SIR results comparing NLE and InfoNCE-λ objectives under random and Sobol design policies. We report mean $\pm$ standard error over 3 random seeds. Higher EIG is better, while lower L-C2ST and median distance are better.

Objective	**Random design**			**Sobol design**		
	EIG ($\uparrow$)	L-C2ST ($\downarrow$)	Med. ($\downarrow$)	EIG ($\uparrow$)	L-C2ST ($\downarrow$)	Med. ($\downarrow$)
NLE	0.024 ± 0.022	0.0120 ± 0.0037	78.97 ± 33.23	0.0033 ± 0.0013	$\mathbf{0.0015 \pm 0.0005}$	69.04 ± 55.42
InfoNCE-λ ($\lambda = 0.01$)	$\mathbf{0.932 \pm 0.359}$	0.0053 ± 0.0042	77.82 ± 33.09	0.848 ± 0.246	0.0048 ± 0.0023	65.64 ± 52.06
InfoNCE-λ ($\lambda = 0.1$)	0.713 ± 0.422	0.0066 ± 0.0047	79.29 ± 33.42	$\mathbf{0.871 \pm 0.289}$	0.0049 ± 0.0017	$\mathbf{59.92 \pm 46.40}$
InfoNCE-λ ($\lambda = 1$)	0.355 ± 0.283	$\mathbf{0.0041 \pm 0.0015}$	$\mathbf{77.68 \pm 32.59}$	0.304 ± 0.146	0.0060 ± 0.0004	68.08 ± 54.40

Table 5: BMP baseline sweep results comparing NLE and InfoNCE-λ objectives under random and Sobol design policies. We report mean $\pm$ standard error over 3 random seeds. L-C2ST values smaller than 0.01 are displayed as 0.01 for readability.

Objective	**Random design**			**Sobol design**		
	EIG ($\uparrow$)	L-C2ST ($\downarrow$)	Med. ($\downarrow$)	EIG ($\uparrow$)	L-C2ST ($\downarrow$)	Med. ($\downarrow$)
NLE	0.02 ± 0.01	0.01 ± 0.01	11.49 ± 5.33	0.03 ± 0.02	0.01 ± 0.01	4.06 ± 1.15
InfoNCE-λ ($\lambda = 0.01$)	$\mathbf{0.46 \pm 0.07}$	0.01 ± 0.01	10.86 ± 5.72	$\mathbf{0.61 \pm 0.36}$	0.01 ± 0.01	14.57 ± 6.59
InfoNCE-λ ($\lambda = 0.1$)	0.46 ± 0.10	0.01 ± 0.01	14.58 ± 5.14	0.57 ± 0.20	0.01 ± 0.01	4.59 ± 2.67
InfoNCE-λ ($\lambda = 1$)	0.38 ± 0.17	0.01 ± 0.01	7.73 ± 5.25	0.43 ± 0.20	0.01 ± 0.01	13.52 ± 4.24