Interventions, Where and How? Bayesian Active Causal Discovery at Scale

Anonymous Author(s)
Affiliation
Address
email

Abstract

Causal discovery from observational and interventional data is challenging due to limited data and non-identifiability which introduces uncertainties in estimating the underlying structural causal model (SCM). Incorporating these uncertainties and selecting optimal experiments (interventions) to perform can help identify the true SCM faster. Existing methods in experimental design for causal discovery from limited data either rely on linear assumptions for the SCM or select only the intervention target. In this paper, we incorporate recent advances in Bayesian causal discovery into the Bayesian optimal experimental design framework which allows for active causal discovery of nonlinear, large SCMs, while selecting both the target and the value to intervene with. We demonstrate the performance of the proposed method on synthetic graphs (Erdos-Rènyi, Scale Free) for both linear and nonlinear SCMs as well as on the *in-silico* single-cell gene regulatory network dataset, DREAM.

1 Introduction

What is the structure of the protein-signaling network derived from a single cell? How do different habits influence the presence of disease? Such questions refer to causal effects in complex systems governed by nonlinear, noisy processes. On most occasions, passive observation of such systems is insufficient to uncover the real cause-effect relationship and costly experimentation is required to disambiguate between competing hypotheses. As such, the design of experiments is of significant interest; an efficient experimentation protocol helps reduce the costs involved in experimentation while aiding the process of producing knowledge through the (closed-loop / policy-driven) scientific method (Fig. 1).

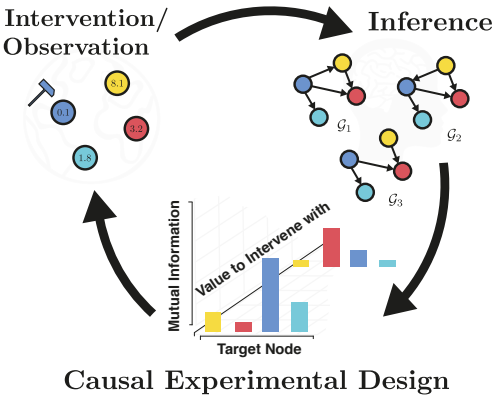

Figure 1: Intervention-Inference-Design loop of Bayesian Optimal Experimental Design for Causal Discovery framework.

In the language of causality [29], the causal relationships are represented qualitatively by a directed acyclic graph (DAG), where the nodes correspond to different variables of the system of study and the edges represent the flow of information between the variables. The abstraction of DAGs allows us to represent the space of possible explanations (hypotheses) for the observations at hand. Representing such hypotheses as Bayesian probabilities (beliefs) allows us to formalize the problem of the scientific method as one of Bayesian inference, where the goal is to estimate the posterior distribution $p(\text{DAGs} \mid \text{Observations})$. A posterior distribution over

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

the DAGs allows us to employ information-theoretic acquisition functions that guide experimentation towards the most informative variables for disambiguating between competing hypotheses. Such design procedures belong to the field of *Bayesian Optimal Experimental Design* [21] for *Causal Discovery* (BOECD) [41, 26].

In the *Bayesian Optimal Experimental Design* (BOED) [21] framework, one seeks the experiment that maximizes the *expected information gain* about some parameter of interest. In *causal discovery*, an experiment takes the form of a causal intervention, and the parameter of interest is the Structural Causal Model (SCM) which consists of a DAG structure and a set of structural equations representing the functional relationships between the variables in the DAG.

An intervention in a causal model refers to the variable (or target) we manipulate and the value (or strength) at which we set the variable. Hence, the design space in the case of learning causal models is the set of all subsets of the intervention targets and the possibly countably infinite set of intervention values of the chosen targets. The intervention value encapsulates important semantics in many causal inference applications. For instance, in medical applications, an intervention can correspond to the administration of different drugs and the intervention value takes the form of a dosage level for each drug. Even though the appropriate choice of this value is crucial for identifying the underlying causal model, existing work on active causal discovery focuses exclusively on selecting the intervention target [1, 7]. There, the intervention value is generally some arbitrary fixed value (like 0) which is suboptimal (see Fig. 2a). Hence, a holistic treatment of selecting the intervention value and the target in the general case of nonlinear causal models has been missing. We present a simple yet effective causal Bayesian experimental design method (`CBED` - pronounced "seabed") to acquire optimal intervention targets and values by performing Bayesian optimization.

Additionally, some settings call for the selection of a batch of interventions. The problem of batched interventions is computationally expensive as it requires evaluating all possible combinations of interventions. We extend `CBED` to the batch setting and propose two different batching strategies for tractable, Bayes optimal acquisition of both intervention targets and values. The first strategy — `Greedy-CBED` — builds up the intervention set greedily. A greedy heuristic is still near-optimal due to submodularity properties of mutual information [19, 1, 17]. The second strategy — `Soft-CBED` — constructs a set of interventions by stochastic sampling from a finite set of candidates, thereby significantly increasing computational efficiency while recovering the original SCM as fast as the greedy strategy. This strategy is well suited for resource-constrained settings.

Throughout this work, we make the following standard assumptions for causal discovery [33]:

Assumption 1 (Causal Sufficiency)**.** *There are no hidden confounders, and all the random variables of interest are observable.*

Assumption 2 (Finite Samples)**.** *There is a finite number of observational/ interventional samples available.*

Assumption 3 (Nonlinear SCM with Additive Noise)**.** *The structural causal model has nonlinear conditional expectations with additive Gaussian noise.*

Assumption 4 (Single Target)**.** *Each intervention is atomic and applied to a single target of the SCM.*

Additionally, we assume that interventions are planned and executed in batches of size $\mathcal{B}$, with a fixed budget of total interventions given by `Number of Batches` $\times \mathcal{B}$. We also assume that there is no model misspecification with respect to the class of conditional expectations and noise distributions. Finally, we are interested in recovering the full graph $\mathbf{G}$ with a small number of batches.

We show that our methods, `Greedy-CBED` and `Soft-CBED`, perform better than the state-of-the-art active causal discovery baselines in linear and nonlinear SCM settings. In addition, our approach achieves superior results in the real-world inspired nonlinear dataset, DREAM [13].

2 Background

Notation. Let $\mathbf{V} = \{1, \dots, d\}$ be the vertex set of any DAG $\mathbf{g} = (\mathbf{V}, E)$ and $\mathbf{X_V} = \{\mathrm{X}_1, \dots, \mathrm{X}_d\} \subseteq \mathcal{X}$ be the random variables of interest indexed by $\mathbf{V}$. We have an initial observational dataset $\mathcal{D} = \{\mathbf{x_V}^{(i)}\}_{i=1}^n$ comprised of instances $\mathbf{x_V} \sim P(\mathrm{X}_1 = \mathrm{x}_1, \dots, \mathrm{X}_d = \mathrm{x}_d) = p(\mathrm{x}_1, \dots, \mathrm{x}_d)$.

Causal Bayesian Network. A causal Bayesian network (CBN) is the pair $(\mathbf{g}, P)$ such that for any $\mathbf{W} \subset \mathbf{V}$,

$$P(\mathbf{X_V}|\text{do}(\mathbf{X_W}) = \mathbf{x'_W}) = \prod_{i \in \mathbf{V} \setminus \mathbf{W}} P(X_i \mid X_{\text{pa}_{\mathbf{g}}(i)}) \mathbb{1}(\mathbf{X_W} = \mathbf{x'_W})$$

where $\text{do}(X_{\mathbf{W}})$ represents a hypothetical intervention on the variables $X_{\mathbf{W}}$, $\mathbb{1}(\cdot)$ is an indicator function and $\text{pa}_{\mathbf{g}}(i)$ denotes parents of variable X_i in DAG $\mathbf{g}$. A perfect intervention on any variable X_j completely removes all dependencies with its parents, i.e. $P(X_j \mid X_{\text{pa}_{\mathbf{g}}(j)}) = P(X_j)$ thereby resulting in a mutilated DAG $\mathbf{g}' = (\mathbf{V}, E \backslash (pa_{\mathbf{g}}(j), j))$.

Structural Causal Model. From the data generative mechanism point of view, the DAG $\mathbf{g}$ on $\mathbf{X_V}$ matches a set of *structural equations*:

$$X_i \coloneqq f_i(X_{\text{pa}_{\mathbf{g}}(i)}), \epsilon_i) \quad \forall i \in \mathbf{V} \tag{1}$$

where f_i's are (potentially nonlinear) causal mechanisms that remain invariant when intervening on any variable $X_j \neq X_i$. ϵ_i's are exogenous noise variables with arbitrary distribution that are mutually independent, i.e $\epsilon_i \perp\!\!\!\perp \epsilon_j \forall i \neq j$. (1) represents the conditional distributions in a Causal Bayesian Network and can additionally reveal the effect of interventions if the mechanisms are known. These equations, along with the DAG $\mathbf{g}$, are known as the structural causal model (SCM). Though the mechanisms f can be nonparametric in the general case, we assume that there exists a parametric approximation to these mechanisms with parameters $\boldsymbol{\gamma} \in \Gamma$. In the case of linear SCMs, $\boldsymbol{\gamma}$ corresponds to the weights of the edges in E. In the nonlinear case, they could represent the parameters of a nonlinear function that parameterizes the mean of a Gaussian distribution.

A common form of (1) corresponds to Gaussian additive noise models (ANM)[1]:

$$X_i \coloneqq f_i(X_{\text{pa}_{\mathbf{g}}(i)}; \boldsymbol{\gamma}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2) \tag{2}$$

An ANM is fully specified by a a DAG, $\mathbf{g}$, mechanisms, $f(\cdot; \boldsymbol{\gamma})$, parameterized by $\boldsymbol{\gamma}$, and variances, σ^2. For notational brevity, henceforth we denote $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \sigma^2)$ and all the parameters of the SCM with $\boldsymbol{\phi} = (\mathbf{g}, \boldsymbol{\theta})$.

Bayesian Causal Discovery. A common assumption in causal inference is that causal relations are known qualitatively and can be represented by a DAG. While this qualitative information can be obtained from domain knowledge in some scenarios, it's infeasible in most applications. The goal of causal discovery is to recover the SCM given a dataset $\mathcal{D}$. In general, without further assumptions about the nature of mechanisms f (e.g., linear vs. nonlinear), the true SCM may not be *identifiable* [31] from observational data alone. This non-identifiability is because there could be multiple DAGs (and hence multiple factorizations of $P(\mathbf{X_V})$) which explain the data equally well. Such DAGs are said to be *Markov Equivalent*. Interventions can improve identifiability. In addition to identifiability issues, estimating the functional relationships between nodes using finite data is another source of uncertainty. Bayesian parameter estimation over the unknown SCM provides a principled way to quantify these uncertainties and obtain a posterior distribution over the SCM given observational data. An experimenter can then use the knowledge encoded by the posterior to design informative experiments that efficiently *acquire interventional data to resolve unknown edge orientations and functional uncertainty*.

Posterior Distributions over SCMs. The key challenge in obtaining posteriors over SCMs is that the space of DAGs is discrete and superexponential in the number of variables [33]. However, recent techniques based on variational inference [2, 23, 8] provide a tractable and scalable way of approximating this posterior. Given a tractable approximation of the posterior over the parameters $\psi \in \Psi$, variational inference optimizes a lower bound on the (log-) evidence:

$$\log p(\mathcal{D}) \geq \mathcal{L}(\psi \in \Psi) = \mathop{\mathbb{E}}_{q_\psi(\boldsymbol{\phi})} [\log p(\mathcal{D} \mid \boldsymbol{\phi})] - \text{D}_{\text{KL}}(q_\phi(\boldsymbol{\phi}) || p(\boldsymbol{\phi}))$$

The key idea in these techniques is the way the variational family Ψ for SCMs, in particular over DAGs, is parameterized. The variational family for the VCN method [2] is an autoregressive

[1] ANM's can have noise variables that are non-Gaussian as well, but we restrict our exposition to the Gaussian case.

Bernoulli distribution over the adjacency matrix. They further enforce the acyclicity constraint [43] through the prior. BCD-Nets [8] consider a distribution over node orderings through a Boltzmann distribution and perform inference with Gumbel-Sinkhorn [24] operator. DiBS [23] consider latent variables over entries of adjacency matrix and perform inference over these latent variables using SVGD [22]. We demonstrate empirical results in BOECD using the DiBS model in this work because it is easily extendable to nonlinear SCMs.

Bayesian Optimal Experimental Design. *Bayesian Optimal Experimental Design* (BOED) [21, 6] is an information theoretic approach to the problem of selecting the optimal experiment to estimate any parameter θ. For BOED, the *utility* of the experiment ξ is the mutual information (MI) between the observation $\mathbf{y}$ and θ:

$$\begin{aligned}\mathrm{U}_{\text{BOED}}(\xi) \triangleq \mathrm{I}(\mathbf{Y};\theta \mid \xi,\mathcal{D}) &= \mathop{\mathbb{E}}_{p(\mathbf{y}|\theta,\xi)p(\theta)}[\log p(\mathbf{y} \mid \xi,\mathcal{D}) - \log p(\mathbf{y} \mid \theta,\xi,\mathcal{D})] \\ &= \mathop{\mathbb{E}}_{p(\mathbf{y}|\theta,\xi)p(\theta)}[\log p(\theta \mid \xi,\mathcal{D}) - \log p(\theta \mid \mathbf{y},\xi,\mathcal{D})]\end{aligned}$$

The goal of BOED is to select the experiment that maximizes this objective $\xi^* = \arg\max_\xi \mathrm{U}_{\text{BOED}}(\xi)$. A common setting, called *static*, *fixed* or *batch* design, is to optimize multiple designs $\{\xi_1,\dots,\xi_{\mathcal{B}}\}$ at the same time.

3 Method

The true SCM $\tilde{\phi} = (\tilde{\mathbf{g}},\tilde{\boldsymbol{\theta}})$ over random variables $\mathbf{X_V}$ is a matter of fact, but our belief in $\tilde{\phi}$ is uncertain for many reasons. Primarily, it is only possible to learn the DAG $\tilde{\mathbf{g}}$ up to a Markov equivalence class (MEC) from observational data $\mathcal{D}$. Uncertainty also arises from $\mathcal{D}$ from being a finite sample, which we model by introducing the the random variable $\mathbf{\Phi}$. Let $\boldsymbol{\phi} \sim p(\boldsymbol{\phi}|\mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{\phi})p(\boldsymbol{\phi})$ be an instance of the random variable $\mathbf{\Phi}$ that is sampled from our posterior over SCMs after observing the dataset $\mathcal{D}$.

We would like to design an experiment to identify an intervention $\xi \coloneqq \{(j,v)\} \coloneqq \mathrm{do}(\mathrm{X}_j = v)$ that maximizes the information gain about $\mathbf{\Phi}$ after observing the outcome of the intervention $\mathbf{y} \sim P(\mathrm{X}_1 = \mathrm{x}_1,\dots,\mathrm{X}_d = \mathrm{x}_d \mid \mathrm{do}(\mathrm{X}_j = v) = p(\mathbf{y} \mid \xi)$. Here, $\mathbf{y}$ is an instance of the random variable $\mathbf{Y} \subseteq \mathcal{X}$ distributed according to the distribution specified by the mutilated true graph $\tilde{\mathbf{g}}'$ under intervention. Looking at one intervention at a time, one can formalize BOECD as gain in information about $\mathbf{\Phi}$ after observing the outcome of an experiment $\mathbf{y}$. The experiment $\xi \coloneqq \{(j,v)\}$ that maximizes the information gain is the experiment that maximizes the mutual information between $\mathbf{\Phi}$ and $\mathbf{Y}$:

$$\{(j^*,v^*)\} = \arg\max_{j,v} \{\mathrm{I}(\mathbf{Y};\mathbf{\Phi} \mid \{(j,v)\},\mathcal{D})\} \tag{3}$$

The above objective considers taking `arg max` over not just the discrete set of intervention targets $j \in \mathbf{V}$, but also over the uncountable set of intervention values $v \subset \mathcal{X}_j$. While the existing works in BOECD consider only the design of intervention targets to limit the complexity [41, 26, 1], our approach tackles both the problems. We first outline the methodology for a single design and in Section 3.2 demonstrate how to extend this single design to a batch setting.

3.1 Single Design

To maximize the objective in Equation 3, we need to (1) estimate MI for candidate interventions and (2) maximize the estimated MI by optimizing over the domain of intervention value for every candidate interventional target.

Estimating the MI. As mutual information is intractable, there are various ways to estimate it depending on whether we can sample from the posterior and whether the likelihood can be evaluated [10, 34, 15]. Since the models we consider allow both posterior sampling and likelihood evaluation, it suffices to obtain an estimator which requires only likelihood evaluation and Monte Carlo approximations of the expectations. To do so, we derive an estimator similar to Bayesian Active Learning by Disagreement (BALD) [15], which considers MI as a difference of conditional entropies over the outcomes $\mathbf{Y}$:

$$\begin{aligned}&\mathrm{I}(\mathbf{Y};\mathbf{\Phi} \mid \{(j,v)\},\mathcal{D}) = \mathrm{H}(\mathbf{Y} \mid \{(j,v)\},\mathcal{D}) - \mathrm{H}(\mathbf{Y} \mid \mathbf{\Phi},\{(j,v)\},\mathcal{D}) \\ &= -\mathop{\mathbb{E}}_{p(\mathbf{y}|\{(j,v)\},\mathcal{D})}\left[\log\left(\mathop{\mathbb{E}}_{p(\boldsymbol{\phi}|\mathcal{D})}[p(\mathbf{y} \mid \boldsymbol{\phi},\{(j,v)\})]\right)\right] + \mathop{\mathbb{E}}_{p(\boldsymbol{\phi}|\mathcal{D})}\left[\mathop{\mathbb{E}}_{p(\mathbf{y}|\boldsymbol{\phi},\{(j,v)\})}[\log(p(\mathbf{y} \mid \boldsymbol{\phi},\{(j,v)\}))]\right]\end{aligned} \tag{4}$$

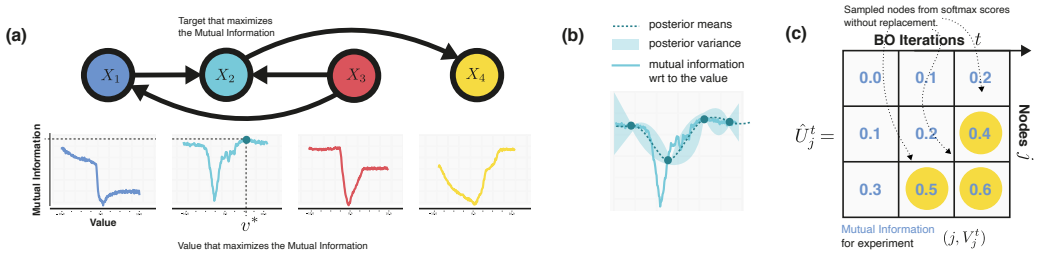

Figure 2: **(a)** Each graph shows how the **mutual information** (MI) (y-axis) changes for intervening on that node (plot color matching the node color) with different **values** (x-axis). The SCM in this example is a nonlinear SCM with Additive Gaussian noise. We can see that by intervening on node X_2 with the value v^*, the mutual information gets maximized. **(b)** The posterior distribution of a GP on the Mutual Information function as a response to different intervention values after four Bayesian Optimization (BO) steps. **(c)** For each t iteration of the BO algorithm and each node j, we get a utility function evaluation $\hat{U}_j^t$ (the utility being the MI in our case). Then we sample without replacement proportionally to the scores to prepare a batch (3.2).

where $\mathrm{H}(\cdot)$ is the entropy. See Appendix B.1 for the derivation. A Monte Carlo estimator of the above equation can be used as an approximation (Appendix B.2). Equation (4) has an intuitive interpretation. It assigns high mutual information to interventions that the model disagrees the most regarding the outcome. We denote the MI for a single design as $\mathcal{I}(\{(j, v)\}) \coloneqq \mathrm{I}(\mathbf{Y}; \mathbf{\Phi} \mid \{(j, v)\}, \mathcal{D})$.

Selecting the Intervention Value. As shown in (3), maximizing the objective is achieved not only by selecting the intervention target but also by setting the appropriate intervention value. Although optimizing the intervention target is tractable (discrete and finite number of nodes to select from), selecting the value to intervene is usually intractable since they are continuous. For any given target node j, MI is a nonlinear function over $v \in \mathcal{X}_j$ (See Fig 2) and hence solving with gradient ascent techniques only yields a local maximum. Given that MI is expensive to evaluate, we treat MI for a given target node j as a black-box function and obtain its maximum using Bayesian Optimization (BO) [20, 44, 25]. BO seeks to find the maximum of this function $max_{v \in \mathcal{X}_j} \mathcal{I}(\{(j, v)\})$ over the entire set $\mathcal{X}_j$ with as few evaluations as possible. See appendix E for details.

BO typically proceeds by placing a Gaussian Process (GP) [35] prior on the function $\mathcal{I}(\{j, \cdot\})$ and obtain the posterior over this function with the queried points $\mathbf{v}^* = \{v^{(1)*}, \ldots, v^{(T)*}\}$. Let the value of the mutual information queried at each optimization step t be $\hat{\mathrm{U}}_j^t = \mathcal{I}(\{(j, v^{(t)*}\})$. The posterior *predictive* of a point $v^{(t+1)}$ can be obtained in closed form as a Gaussian with mean $\boldsymbol{\mu}_{\boldsymbol{j}}^{(\boldsymbol{t+1})}(v)$ and variance $\boldsymbol{\sigma}_{\boldsymbol{j}}^{(\boldsymbol{t+1})}(v)$. Subscript j signifies the fact that we maintain different μ and σ per intervention target and superscript t represents the BO step. Querying proceeds by having an acquisition function defined on this posterior, which suggests the next point to query. For BO, we use an acquisition function called as the Upper Confidence Bound (UCB) [39] which suggests the next point to query by trading-off *exploration* and *exploitation*:

$$v_j^{(t+1)*} = \arg\max_v \boldsymbol{\mu}_{\boldsymbol{j}}^{\boldsymbol{t}}(v) + \sqrt{\beta_j^{t+1}} \boldsymbol{\sigma}_{\boldsymbol{j}}^{\boldsymbol{t}}(v)$$

We run GP-UCB independently on every candidate intervention target $j = \{1, \ldots, d\}$ by querying points within a fixed domain $[-k, k] \subset \mathbb{R}$. Note that the domain can be chosen based on the application, for example, if we must constrain dosage levels within a fixed range. Each GP is one-dimensional in our setup; hence a few evaluations of UCB are sufficient to get a good value maxima candidate. Further, GP-UCB for each candidate target is parallelizable, making it efficient. We finally select the design with the highest MI across the candidate intervention targets.

3.2 Batch Design

In many applications, it is desirable to select the most informative *set* of interventions instead of a single intervention at a time. Take, for example, a biologist entering a wet lab with a script of experiments to execute. Batching experiments removes the bottleneck of waiting for an experiment to finish and get analyzed until executing the next one. Given a budget per batch $\mathcal{B}$ which

Algorithm 1: `Greedy-CBED`

Input : $\mathcal{E}$ environment, N initial observational samples, $\mathcal{B}$ batch Size, d number of nodes

▷ Initialize set of experiments Ξ to empty

$\Xi \leftarrow \varnothing$

for $n = 1 \dots \mathcal{B}$ **do**

  for $j = 1 \dots d$ **do**

    ▷ Select optimal intervention value per node j using GP-UCB

    $V_j \leftarrow \arg\max_v \mathcal{I}(\Xi \cup \{(j, v)\})$

    $U_j \leftarrow \mathcal{I}(\Xi \cup \{(j, V_j)\})$

  $j^* \leftarrow \arg\max_j U_j$

  $v^* \leftarrow V_{j^*}$

  $\Xi \leftarrow \Xi \cup \{(j^*, v^*)\}$

return Ξ

Algorithm 2: `Soft-CBED`

Input : $\mathcal{E}$ environment, N initial observational samples, $\mathcal{B}$ batch Size, d number of nodes, ζ softmax temperature

for $j = 1 \dots d$ **do**

  ▷ Select candidate intervention values per node j using GP-UCB

  Initialize μ_j^0 and σ_j^0

  for $t = 1 \dots T$ **do**

    $V_j^t \leftarrow$ $\arg\max_v \mu_j^{t-1}(v) + \sqrt{\beta^t}\sigma_j^{t-1}(v)$

    $\hat{U}_j^t \leftarrow \mathcal{I}(\{(j, V_j^t)\})$

    Update the GP to obtain μ_j^t and σ_j^t

$\{(t_i, j_i)\}_{i \in \{1,\dots,\mathcal{B}\}} \leftarrow \mathcal{B}$ samples *without* replacement $\propto \exp(\hat{U}_j^t/\zeta)$

$\Xi \leftarrow \{(j_i, V_{j_i}^{t_i})\}_{i \in \{1,\dots,\mathcal{B}\}}$

return Ξ

denotes the number of experiments in a batch, the problem of selecting the batch then becomes $\arg\max_{\Xi} \mathrm{I}(\mathbf{Y}; \mathbf{\Phi} \mid \Xi, \mathcal{D})$, such that $\texttt{cardinality}(\Xi) = \mathcal{B}$, where Ξ is a set of interventions $\bigcup_{i=1}^{\mathcal{B}} (j_i, v_i)$ and $\mathbf{Y}$ denotes the random variable for the outcomes of the interventions of the batch. We denote the MI for a batch design as $\mathcal{I}(\Xi) \coloneqq \mathrm{I}(\mathbf{Y}; \mathbf{\Phi} \mid \Xi, \mathcal{D})$.

Greedy Algorithm. Computing the optimal solution $\mathcal{I}(\Xi^*)$ is computationally infeasible. However, as the conditional mutual information is *submodular* and *non-decreasing* (see Appendix B.4 for proof), we can derive a simple greedy algorithm (Algorithm 1) that can achieve at least a $(1 - 1/e) \approx 0.64$ approximation of the optimal solution [19, 27]. We denote this strategy as `Greedy-CBED`.

Soft Top-K. Although the greedy algorithm is tractable, it requires $O(\mathcal{B}d)$ instances of GP-UCB. Kirsch *et al.* [18] show that a soft top-k selection strategy performs similarly to the greedy algorithm, reducing the computation requirements to $O(d)$ runs of GP-UCB. To achieve this, we construct a finite set of candidate intervention target-value pairs by keeping all the T evaluations of GP-UCB for each node $j = \{1, \dots, d\}$. Therefore, for d nodes, we our candidate set is comprised of $d \times T$ experiments. We score each experiment in this candidate set using the MI estimate. We then sample *without replacement* $\mathcal{B}$ times proportionally to the *softmax* of the MI scores (Algorithm 2). We denote this strategy as `Soft-CBED`.

3.3 Comparison with existing active causal discovery methods

We outline how our approach compares with two main existing active causal discovery methods.

ABCD [1]. The estimator of MI used in ABCD is based on weighted importance sampling. However, for the specific choice of the importance sampling weights used in ABCD, their MI estimator ends up with the same approximation as in our method (see Appendix B.5). Nevertheless, ABCD does not select intervention values but suboptimally sets them to a fixed value. In addition, our proposed `Soft-CBED` is a faster and more efficient batch strategy, especially when values also have to be acquired. From this perspective, our approach is an extension of ABCD with nonlinear assumptions, value acquisition, and a soft top-k batching strategy.

AIT [38]. AIT is an F-score-based intervention target acquisition strategy. Although it is not a BOECD method, we prove here that it can be viewed as a Monte Carlo estimate of the approximation to MI when the outcomes $\mathbf{Y}$ are Gaussian. Nevertheless, AIT does not select intervention values like ABCD and does not have a batch strategy.

Theorem 3.1. *Let* $\mathbf{Y}$ *be a Gaussian random variable. Then the discrepancy score of Scherrer et al.[38] is a Monte Carlo estimate of an approximation to mutual information (Eq.* (4)*). See Appendix B.6 for proof.*

4 Related Work

Early efforts of using *Bayesian Optimal Experimental Design for Causal Discovery* (BOECD) can be found in the works of [26] and [41]. However, these approaches deal with simple settings like limiting the graphs to topologically ordered structures, intervening sequentially, linear models, and discrete variables.

In [7] and [28], BOECD was applied for learning biological networks structure. More recently, ABCD framework [1] extended the work of [26] and [41] in the setting where interventions can be applied in batches with continuous variables. To achieve this, they (approximately) solve the submodular problem of maximizing the batched mutual information between interventions (experiments), outcomes, and observational data, given a DAG. DAG hypotheses are sampled using *DAG-bootstrap* [11]. Our work differs from ABCD in a few ways: we work with both linear and nonlinear SCMs by using state-of-the-art posterior models over DAGs [23], we apply BO to select the value to intervene with, but we also prepare the batch using *soft*BALD [18] which is significantly faster than the greedy approximation of ABCD method.

In [42] the authors proposed the use of *Gaussian Processes* to model the posterior over DAGs and then use BO to identify the value to intervene with, however, this method was not shown to be scalable for larger than bivariate graphs since they rely on multi-dimensional Gaussian Processes for modeling the conditional distributions.

A new body of work has emerged in the field of differentiable causal discovery, where the problem of finding the structure, usually from observational data, is solved with gradient ascent and functional approximators, like neural networks [43, 16, 5]. In recent works [8, 23, 2], the authors proposed a variational approximation of the posterior over the DAGs which allowed for modeling a distribution rather than a point estimate of the DAG that best explains the observational data $\mathcal{D}$. Such work can be used to replace *DAG-bootstrap* [11], allowing for the modeling of posterior distributions with greater support.

Besides the BOECD-based approaches, a few active causal learning works have been proposed [14, 12, 38]. Active ICP [12] uses ICP [32] for causal learning while using an active policy to select the target, however, this work is not applicable in the setting where the full graph needs to be recovered. Closer to our proposal belongs AIT [38], which uses a neural network-based posterior model over the graphs but evaluates the F-score to select the interventions.

5 Experiments

We evaluate the performance of our method on synthetic and real-world causal experimental design problems and a range of baselines. We aim to investigate the following aspects empirically: (1) competitiveness of the overall proposed strategies of `Greedy-CBED` and `Soft-CBED` at scale (50 nodes) on synthetic datasets; (2) performance of the value acquisition strategy based on GP-UCB; and (3) performance of the proposed approach on a real-world inspired dataset.

5.1 Acquisition Functions

Random: Random baseline acquires interventional targets at random. **AIT / *soft*AIT**: active intervention targeting (AIT) [38] uses an f-score based acquisition strategy to select the intervention targets. See appendix B.6 for more details. Since the original proposed approach does not consider a batch setting, we introduce a variant that augments AIT with the proposed soft batching, as described in section 3.2. **CBED / GreedyCBED / SoftCBED** : These are the Monte Carlo estimates of MI, as described in section 3. `CBED` selects a single intervention (target and value) that maximizes the MI and this intervention is applied for the whole batch. In `Greedy-CBED` , the batch is built up in a greedy fashion selecting the target, value pairs one at a time (Algorithm 1). `Soft-CBED` is sampling (target, value) pairs proportionally to the MI scores to select a batch, as described in section 3.2 and Algorithm 2.

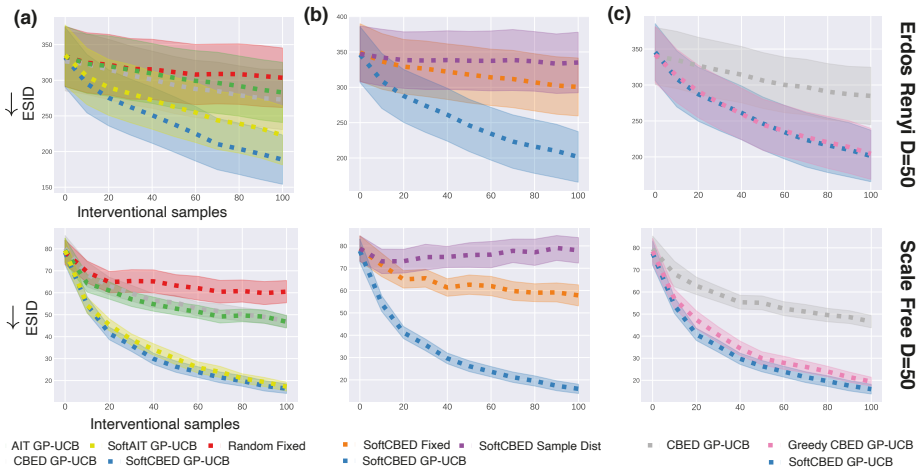

Figure 3: Results on the $\mathbb{E}$**-SID** $\downarrow$ metric (100 seeds, with standard error of the mean shaded) for 50 variables involving nonlinear functional relationships and additive Gaussian noise. (a) We show that `Soft-CBED` with GP-UCB value selection strategy significantly outperforms the baselines. (b) We isolate the effect of the value selection strategy. We show that intervening with a fixed value and sampling from the support of data both perform worse than having an optimizer like GP-UCB. (c) we compare non-batch (`CBED`) vs batch-based acquisition functions (`Greedy-CBED`, `Soft-CBED`). As we can see, `Soft-CBED` performs as well as `Greedy-CBED`. For all experiments, we use the DiBS [23] posterior model.

5.2 Value Selection Strategies

Fixed: This value selection strategy assumes setting the value of the intervention to a fixed value. In the experiments, we fixed the value to 0. **Sample-Dist**: This value selection strategy samples from the support of the observational data. **GP-UCB**: This strategy uses the proposed GP-UCB Bayesian optimization strategy to select the value that maximizes MI.

5.3 Tasks

Synthetic Graphs. In this setting, we generate Erdős–Rényi [9] (ER) and Scale-Free (SF) graphs [3] of size 20 and 50. For linear SCMs, we sample the edge weights γ uniformly at random. For the nonlinear SCM, we parameterize each variable to be a Gaussian whose mean is a nonlinear function of its parents. We model the nonlinear function with a neural network. In all settings, we set noise variance $\sigma^2 = 0.1$. For both types of graphs, we set the expected number of edges per vertex to 1. We provide more details about the experiments in appendix D.1.

Single-Cell Protein-Signalling Network. The DREAM family of benchmarks [13] are designed to evaluate causal discovery algorithms of the regulatory networks of a single cell. A set of ODEs and SDEs generates the dataset, simulating the reverse-engineered networks of single cells. We use `GeneNetWeaver` [37] to simulate the steady-state *wind-type* expression and single-gene *knockouts*. Refer to appendix D.2 for the exact settings.

6 Results

For each of the acquisition objectives and datasets, we present the mean and standard error of the expected structural hamming distance $\mathbb{E}$**-SHD**, expected structural interventional distance $\mathbb{E}$**-SID** [30], area under the receiver operating characteristic curve **AUROC** and area under the precision-recall curve **AUPRC**. In Appendix H we describe the metrics in detail. We evaluate these metrics as a function of the number of acquired interventional samples (or experiments), which helps quantitatively compare different acquisition strategies. Apart from $\mathbb{E}$**-SID**, we relegate results with other metrics to the appendix I.

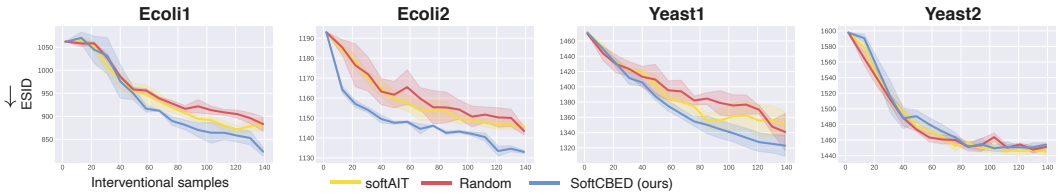

Figure 4: Comparison of acquisition functions on DREAM dataset, for 50 dimensions and batch size 10 on $\mathbb{E}$**-SID**↓ metric (6 seeds, with standard error of the mean).

On the synthetic graphs (Figure 3(a)), we can see that for ER and SF graphs with $50D$ variables and nonlinear functional relationships, the proposed approach based on soft top-k to select a batch with GP-UCB outperforms all the baselines in terms of the $\mathbb{E}$**-SID** metric. On the other hand, AIT alone does not converge to the ground truth graph fast even after combining with the proposed value acquisition, but when further augmented with the proposed soft strategy, the *soft*AIT recovers the ground truth causal graph upto 4 times faster and performs competitively to `Soft-CBED`. We observe similar performance across other metrics as well. In addition, we found this trend to hold for $20D$ variables and linear models. Full results are presented in the appendix I.

Table 1: Performance comparison between different value selection and batch strategies for `CBED`. Experiments are performed using an AMD EPYC 7662 64-Core CPU and Tesla V100 GPU.

Strategy		
Value	**Batch**	**Runtime(s)**
Fixed	Greedy	32.56
	Soft	6.42
GP-UCB	Greedy	284.98
	Soft	24.17

Next, we examine the importance of having a value selection strategy for active causal discovery. We use the MI estimator in Equation 4; moreover, we test the proposed GP-UCB with two heuristics - the fixed value strategy and sampling values from the support. As we can see in Figure 3(b), selecting the value using GP-UCB clearly benefits the causal discovery process. We expect this finding as the mutual information is not constant with respect to the intervened value. To make this point clear, we demonstrate in the appendix G the influence of the value in a simple two variables graph. In addition, we note that naively sampling from the support of the observed dataset performs worse than fixing the value to 0. We hypothesize that this is due to lower epistemic uncertainty in the high density regions of the support, hinting that these regions might be less informative.

In order to further understand how the soft batch strategy compares with other batch selection strategies, we compare the results of `Soft-CBED` with `Greedy-CBED` and `CBED`. We observe (Figure 3(c)) that `Greedy-CBED` and `Soft-CBED` give very similar results overall. While `Greedy-CBED` is optimal under certain conditions [17], `Soft-CBED` remains competitive and has the advantage that the batch can be selected in a one-shot manner. This is also evident from the runtime performance of both these batching strategies in Table 1. Both these batch selection strategies perform significantly better than selecting one intervention target/value pair, and executing them $\mathcal{B}$ times (`CBED`).

Finally, on the DREAM task, we see that our method outperforms *soft*AIT and random baselines on the $\mathbb{E}$**-SID** metric (see Figure 4). In these experiments, since the intervention is emulating the gene knockout setting, we only use the fixed value strategy, with a value of 0.0. Although random baseline still remains a competitive choice, in certain settings, `Soft-CBED` objective is significantly better (`Ecoli1`, `Ecoli2` datasets).

7 Summary and Conclusions

This paper studies the problem of efficiently selecting the Bayes optimal experiments to discover causal models. Our proposed framework simultaneously answers the questions of *where* and *how* to intervene in a batched setting. We present a Bayesian optimization strategy to acquire interventional targets and values. Further, we propose two different batching strategies: one based on greedy selection and the other based on soft top-k selection. The proposed methodology for selecting intervention target-value pairs in a batched setting provides superior performance over the state-of-the-art for causal models up to $50D$ variables. We validate this using synthetic datasets and using real-world inspired datasets of single-cell regulatory networks, showing the potential impact on areas like biology and other experimental sciences.

References

[1] Raj Agrawal, Chandler Squires, Karren Yang, Karthikeyan Shanmugam, and Caroline Uhler. Abcd-strategy: Budgeted experimental design for targeted causal structure discovery. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3400–3409. PMLR, 2019.

[2] Yashas Annadani, Jonas Rothfuss, Alexandre Lacoste, Nino Scherrer, Anirudh Goyal, Yoshua Bengio, and Stefan Bauer. Variational causal networks: Approximate bayesian inference over causal structures. *arXiv preprint arXiv:2106.07635*, 2021.

[3] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[4] Vladimir Batagelj and Ulrik Brandes. Efficient generation of large random networks. *Physical Review E*, 71(3):036113, 2005.

[5] Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *arXiv preprint arXiv:2007.01754*, 2020.

[6] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[7] Hyunghoon Cho, Bonnie Berger, and Jian Peng. Reconstructing causal biological networks through active learning. *PloS one*, 11(3):e0150611, 2016.

[8] Chris Cundy, Aditya Grover, and Stefano Ermon. Bcd nets: Scalable variational approaches for bayesian causal discovery. *Advances in Neural Information Processing Systems*, 34, 2021.

[9] P Erdős and A Rényi. On random graphs i. publicationes mathematicae (debrecen). 1959.

[10] Adam Foster, Martin Jankowiak, Matthew O'Meara, Yee Whye Teh, and Tom Rainforth. A unified stochastic gradient approach to designing bayesian-optimal experiments. In *International Conference on Artificial Intelligence and Statistics*, pages 2959–2969. PMLR, 2020.

[11] Nir Friedman, Moises Goldszmidt, and Abraham Wyner. Data analysis with bayesian networks: A bootstrap approach. *arXiv preprint arXiv:1301.6695*, 2013.

[12] Juan L Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. *arXiv preprint arXiv:2006.05690*, 2020.

[13] Alex Greenfield, Aviv Madar, Harry Ostrer, and Richard Bonneau. Dream4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PloS one*, 5 (10):e13397, 2010.

[14] Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(Nov):2523–2547, 2008.

[15] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[16] Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

[17] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32:7026–7037, 2019.

[18] Andreas Kirsch, Sebastian Farquhar, and Yarin Gal. A simple baseline for batch active learning with stochastic acquisition functions. *arXiv preprint arXiv:2106.12059*, 2021.

[19] Andreas Krause and Carlos E Guestrin. Near-optimal nonmyopic value of information in graphical models. *arXiv preprint arXiv:1207.1394*, 2012.

[20] Harold J Kushner. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. 1964.

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

[22] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *arXiv preprint arXiv:1608.04471*, 2016.

[23] Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable bayesian structure learning. *arXiv preprint arXiv:2105.11839*, 2021.

[24] Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. *arXiv preprint arXiv:1802.08665*, 2018.

[25] Jonas Močkus. On bayesian methods for seeking the extremum. In *Optimization techniques IFIP technical conference*, pages 400–404. Springer, 1975.

[26] Kevin P Murphy. Active learning of causal bayes net structure. 2001.

[27] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

[28] Robert Osazuwa Ness, Karen Sachs, Parag Mallick, and Olga Vitek. A bayesian active learning experimental design for inferring signaling networks. In *International Conference on Research in Computational Molecular Biology*, pages 134–156. Springer, 2017.

[29] Judea Pearl. *Causality*. Cambridge university press, 2009.

[30] Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799, 2015.

[31] Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*, 2012.

[32] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

[33] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[34] 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.

[35] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[36] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

[37] Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16): 2263–2270, 2011.

[38] Nino Scherrer, Olexa Bilaniuk, Yashas Annadani, Anirudh Goyal, Patrick Schwab, Bernhard Schölkopf, Michael C Mozer, Yoshua Bengio, Stefan Bauer, and Nan Rosemary Ke. Learning neural causal models with active interventions. *arXiv preprint arXiv:2109.02429*, 2021.

[39] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias W Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *ICML*, 2010.

[40] Scott Sussex, Andreas Krause, and Caroline Uhler. Near-optimal multi-perturbation experimental design for causal structure learning. *arXiv preprint arXiv:2105.14024*, 2021.

[41] Simon Tong and Daphne Koller. Active learning for structure in bayesian networks. In *International joint conference on artificial intelligence*, volume 17, pages 863–869. Citeseer, 2001.

[42] Julius von Kügelgen, Paul K Rubenstein, Bernhard Schölkopf, and Adrian Weller. Optimal experimental design via bayesian optimization: active causal structure learning for gaussian process networks. *arXiv preprint arXiv:1910.03962*, 2019.

[43] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *arXiv preprint arXiv:1803.01422*, 2018.

[44] AG Zhilinskas. Single-step bayesian search method for an extremum of functions of a single variable. *Cybernetics*, 11(1):160–166, 1975.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See end of Section 1. Our limitations arise from the fact that the proposed methodology and conclusion hold when the assumptions laid out are satisfied.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Negative societal impact is discussed in appendix A and will be added to the extra page of the camera ready after the reviewing process.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The assumptions are laid out in Introduction as well as in Theorem 3.1.
   (b) Did you include complete proofs of all theoretical results? [Yes] The proofs are in Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix D.1, D.2 and H
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 3 and 4.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please check appendix K for details.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] Please check appendix L for details.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Data are simulated.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] Data are simulated.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]