Challenges and Considerations in the Evaluation of Bayesian Causal Discovery

Amir Mohammad Karimi Mamaghan [1] **Panagiotis Tigas** [2] **Karl Henrik Johansson** [1 3] **Yarin Gal** [2]
Yashas Annadani [* 4 5] **Stefan Bauer** [* 4 5]

Abstract

Representing uncertainty in causal discovery is a crucial component for experimental design, and more broadly, for safe and reliable causal decision making. Bayesian Causal Discovery (BCD) offers a principled approach to encapsulating this uncertainty. Unlike non-Bayesian causal discovery, which relies on a single estimated causal graph and model parameters for assessment, evaluating BCD presents challenges due to the nature of its inferred quantity – the posterior distribution. As a result, the research community has proposed various metrics to assess the quality of the approximate posterior. However, there is, to date, no consensus on the most suitable metric(s) for evaluation. In this work, we reexamine this question by dissecting various metrics and understanding their limitations. Through extensive empirical evaluation, we find that many existing metrics fail to exhibit a strong correlation with the quality of approximation to the true posterior, especially in scenarios with low sample sizes where BCD is most desirable. We highlight the suitability (or lack thereof) of these metrics under two distinct factors: the identifiability of the underlying causal model and the quantity of available data. Both factors affect the entropy of the true posterior, indicating that the current metrics are less fitting in settings of higher entropy. Our findings underline the importance of a more nuanced evaluation of new methods by taking into account the nature of the true posterior, as well as guide and motivate the development of new evaluation procedures for this challenge.

[*]Equal contribution [1]KTH Royal Institute of Technology [2]OATML, University of Oxford [3]Digital Futures [4]Helmholtz AI [5]TU Munich. Correspondence to: Amir Mohammad Karimi Mamaghan <amkm@kth.se>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Much of the pursuit in acquiring scientific knowledge involves inferring causal relationships within a system of interest and the laws that govern those relationships. Applications in biology, like inferring a gene network and their regulatory mechanisms from gene expression data (Tejada-Lapuerta et al., 2023; Dibaeinia & Sinha, 2020) and protein-signaling networks with single-cell data (Sachs et al., 2005), necessitates a mechanistic understanding of the data generation process. Estimating the causal model in such applications from data, called causal discovery, is an important problem in empirical sciences (Spirtes et al., 2000).

A typical scientific discovery loop for causal understanding involves a scientist first coming up with causal hypotheses based on prior knowledge, and refining these hypotheses based on new evidence obtained through observation and experimentation. In science, a key requirement in light of limited data is that all the plausible hypotheses that explain the data have to be considered before devising an efficient experimentation protocol, as opposed to a single most likely one (Lindley, 1956; Chaloner & Verdinelli, 1995). Bayesian Causal Discovery (BCD) is an elegant framework that fulfills this requirement by quantifying the epistemic uncertainty of the underlying causal model through the Bayesian posterior, which provides a degree of belief of every causal hypothesis proportional to its ability to explain the data (Heckerman et al., 2006; Friedman & Koller, 2003; Chickering et al., 2013). The quantified epistemic uncertainty can be then used to design informative experiments/ interventions to perform (Tong & Koller, 2001; Tigas et al., 2022; Sussex et al., 2022; Lyle et al., 2023) or to estimate causal effects of variables with Bayesian model averaging (Geffner et al., 2022; Emezue et al., 2023).

One of the common frameworks for dealing with questions related to cause and effect is the Structural Causal Model (SCM) with an associated graph indicating the causal relationships between variables (Pearl, 2009; Peters et al., 2017). Under this framework, BCD aims to infer the Bayesian posterior over the graph and the parameters of the SCM. This is a hard problem due to the combinatorial nature of graphs, which renders the posterior intractable for more than 6 variables. Recently, various approximate inference

methods have been introduced that use gradient information allowing to scale to SCMs for which the true posterior is intractable (Annadani et al., 2021; Cundy et al., 2021; Lorch et al., 2021; Nishikawa-Toomey et al., 2022; Deleu et al., 2022; 2023; Hägele et al., 2023; Atanackovic et al., 2023). However, in the absence of the true posterior, evaluation of BCD methods is hard as the inferred quantity is a distribution rather than the most-likely estimate, as in causal discovery. The BCD community so far has relied on proxy metrics[1], many of which are derived from causal discovery evaluation. For instance, a standard metric in causal discovery is the Structural Hamming Distance (SHD) for evaluating the estimated graph, and in BCD, *expected* SHD is used to evaluate the posterior over DAGs. However, such a metric may not be representative of how good the posterior approximation is, and has been briefly discussed in prior work (Deleu et al., 2023; Lorch et al., 2022). Our motivation is that a holistic understanding of the limitations of BCD evaluation is missing. Given that many new BCD algorithms are being proposed, proper understanding of the limitations of the present evaluation metrics is important to make advances in the right direction, especially with regards to applying it to real-world datasets where the amount of samples might be limited.

In this work, we aim to bridge the gap that exists in the understanding of the evaluation of BCD algorithms. In order to do so, we note that the desiderata for an ideal evaluation metric would be to compare the approximated posterior to the *gold standard*, the true posterior. Therefore, we analyze the performance of different BCD methods on all the known metrics for linear additive noise models, for which the true posterior is tractable. This not only helps us to compare how different metrics correlate with evaluating the approximate posterior against the true posterior, but also gives a way to understand some properties of the true posterior, which we shall show, is important in understanding the conditions under which the current metrics are suitable for evaluation or where they may be lacking. Our experimental evaluation with linear additive noise models on 8 different metrics for 5 different algorithms reveals the following aspects:

1. In terms of the relative performance of BCD methods, we find that all the metrics are not correlated to a metric which directly evaluates on the true posterior when the number of samples is low ($n \approx d$ where n is the dataset size and d is the number of variables), indicating that the current metrics are not suitable for evaluation of uncertainty in these settings.

2. With higher number of samples ($n >> d$), the correlation between the current evaluation metrics and the metric on the true posterior significantly improves.

3. Based on a similar correlation analysis, we observe that the current metrics are less suitable for the evaluation of uncertainty when the true model is non-identifiable, as opposed to the identifiable case.

4. Overall, the reliability of existing metrics as evaluation methods is related to the entropy of the true posterior. The true posterior has higher entropy with less data and non-identifiability.

5. Therefore, future algorithms should consider the setting of interest (and the entropy of the true posterior it induces) in deciding whether to use existing metrics or not. In a higher entropy true posterior setting, it would be better to evaluate the posterior on downstream tasks (for example causal effect estimation) where the ground truth is well-defined.

The remaining parts of the paper are organized as follows: Section 2 provides background on causality and Bayesian causal discovery. Section 3 explains all different evaluation metrics for BCD in use and discusses their limitations. Section 4 presents the empirical evaluation of multiple different algorithms on all metrics which highlights the shortcomings of present metrics for BCD evaluation in terms of the quality of the posterior approximation. Section 5 proposes two alternative ways of evaluating BCD models. Finally, Section 6 discusses the limitations and presents the overall conclusion.

2. Background

In this section, we briefly introduce the Structural Causal Model (SCM) formalism under which the problem of causal discovery is defined. We also introduce the problem of Bayesian Causal Discovery under this framework.

2.1. Structural Causal Model

Let $\mathbf{V} = \{1, \ldots, d\}$ be the vertex set of any graph $\boldsymbol{G} = (\mathbf{V}, E)$ and $\mathbf{X} = \{\mathrm{X}_1, \ldots, \mathrm{X}_d\} \subseteq \mathcal{X}$ be the random variables of interest indexed by $\mathbf{V}$. A Structural Causal Model (SCM) consists of a set of equations wherein each variable X_i is assigned a value which is a deterministic function of its direct causes $X_{\mathrm{pa}(i)}$ as well as an exogenous noise variable ϵ_i with a distribution P_{ϵ_i}:

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

f_i's are mechanisms that relate how the direct causes affect the variable X_i. If the structural assignments are assumed to be acyclic, these equations induce a Directed Acyclic Graph (DAG) $\boldsymbol{G} = (\mathbf{V}, E)$ whose vertices correspond to the variables and edges indicate direct causes. A perfect

[1]Unless otherwise specified, metric(s) in this work refers to evaluation method(s) or protocol(s) to assess the goodness of an algorithm.

intervention on any variable X_i corresponds to changing the structural equation of that variable to the desired state (value), $X_i := s_i$, where $s_i \in \mathcal{X}_i$. It is denoted by the do-operator (Pearl, 2009) as $\mathrm{do}(X_i = s_i)$. Under this model, the recursive application of Equation (1) entails a joint distribution $p_{\mathbf{X}}$, such that the Markov factorization holds:

$$p_{\mathbf{X}}(\mathbf{X}) = \prod_{i=1}^{d} p_i(X_i | X_{\mathrm{pa}(i)}) \tag{2}$$

The problem of causal discovery is to estimate the SCM (i.e. the causal graph $\boldsymbol{G}$, parameters of f_i's and ϵ_i's) given N samples from $p_{\mathbf{X}}$. For analysis of different evaluation metrics, we assume that the SCM is causally sufficient, i.e. all the variables are measurable, and the noise variables are mutually independent.

Without further assumptions on the mechanisms and the noise, an SCM is not *identifiable* from observational data, i.e. there could be multiple factorizations that can be consistent with a given joint distribution $p_{\mathbf{X}}$. One of the simplest identifiable setting is a linear Gaussian Additive Noise Model (ANM) with homoscedastic noise (Peters & Bühlmann, 2014):

$$X_i := \boldsymbol{\gamma}_i^T X_{\mathrm{pa}(i)} + \epsilon_i, \;\; \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

where the mechanisms f_i are linear with parameter $\boldsymbol{\gamma}_i \in \mathbb{R}^{|\mathrm{pa}(i)|}$. For notational brevity, henceforth we denote $\boldsymbol{\phi} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d, \sigma^2)$ and all the parameters of interest with $\boldsymbol{\theta} = (\boldsymbol{G}, \boldsymbol{\phi})$. If the noise is heteroscedastic in the above model, under the assumption of faithfulness, it is only identifiable up to an equivalence class over graphs, called Markov Equivalence Class (MEC) (Andersson et al., 1997).

2.2. Bayesian Causal Discovery

Given a dataset $\boldsymbol{D} = \{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)}\}$, DAG $\boldsymbol{G}$ and parameters $\boldsymbol{\phi}$, they induce a unique joint distribution $p(\boldsymbol{D}, \boldsymbol{\phi}, \boldsymbol{G})$ with the prior $p(\boldsymbol{G}, \boldsymbol{\phi})$ and likelihood $p(\boldsymbol{D}|\boldsymbol{G}, \boldsymbol{\phi})$ (Friedman & Koller, 2003). Bayesian causal discovery aims to infer the posterior[2] $p(\boldsymbol{G}, \boldsymbol{\phi}|\boldsymbol{D}) \propto p(\boldsymbol{D}|\boldsymbol{G}, \boldsymbol{\phi})p(\boldsymbol{G}, \boldsymbol{\phi})$. A Bayesian method for causal discovery is preferable to model epistemic uncertainty about the model due to finite data. In addition, with choice of appropriate parameter priors (Geiger & Heckerman, 2002), equivalence classes like MEC can be characterized in the case of non-identifiability. A crucial benefit of posterior inference in causal models is that it is helpful for downstream tasks like experimental design and cause-effect estimation with Bayesian model averaging. However, the true posterior is not tractable for more than 6 variables. The true posterior is given by $p(\boldsymbol{G}, \boldsymbol{\phi}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\boldsymbol{\phi}, \boldsymbol{G})p(\boldsymbol{G}, \boldsymbol{\phi})}{\sum_{\boldsymbol{G}} \int_{\boldsymbol{\phi}} p(\boldsymbol{D}, \boldsymbol{G}, \boldsymbol{\phi})}$. To calculate the true posterior, we need to calculate the summation over G which is infeasible as the number of possible DAGs grows super-exponentially w.r.t. number of variables ($\mathcal{O}(2^{d^2})$). The goal of BCD therefore is to find an approximate posterior $q(\boldsymbol{G}, \boldsymbol{\phi}|\boldsymbol{D})$ that is close to the true posterior.

[2]We refer this as true posterior to emphasize the difference with approximate posterior.

3. On Evaluation of BCD

Evaluating the goodness of posterior approximation $q(\boldsymbol{G}, \boldsymbol{\phi}|\boldsymbol{D})$ in the absence of true posterior $p(\boldsymbol{G}, \boldsymbol{\phi}|\boldsymbol{D})$ requires proxy metrics or downstream task evaluation. The BCD community so far has focused on proxy metrics which are mostly derived from causal discovery evaluation. The current metrics can be classified into two categories: graph-only evaluation metrics and full posterior evaluation metrics.

Graph only evaluation metrics. These metrics aim to evaluate the uncertainty quantified about graphs through the approximate posterior $q(\boldsymbol{G}|\boldsymbol{D})$.

- **$\mathbb{E}$-SHD**: Structural Hamming Distance (SHD) is a measure of number of edges that are to be added, removed, or reversed to get the ground truth graph from the estimated graph. Since we have a posterior distribution $q(\boldsymbol{G} \mid \boldsymbol{D})$ over graphs, the *expected* SHD is measured:

  $$\mathbb{E}\text{-SHD} := \mathbb{E}_{\boldsymbol{G} \sim q(\boldsymbol{G}|\boldsymbol{D})}[\mathrm{SHD}(\boldsymbol{G}, \boldsymbol{G}^{GT})]$$

  where $\boldsymbol{G}^{GT}$ is the ground-truth causal graph.

- **$\mathbb{E}$-CPDAG SHD**: Similar to $\mathbb{E}$-SHD, $\mathbb{E}$-CPDAG SHD measures the expected hamming distance between the Completed partially directed acyclic graph (CPDAG, an equivalence class of DAGs) of the ground truth graph and the CPDAG of the graph sampled from the posterior.

- Threshold Metrics: Area Under Precision Recall Curve (**AUPRC**) and Area Under Receiver Operator Characteristics (**AUROC**) are the two common threshold based metrics. In these evaluation metrics, area under the precision recall curve or ROC curve is measured by thresholding the posterior edge beliefs $q(\boldsymbol{G}_{ij}|\boldsymbol{D})$ and averaging over all possible edges.

These metrics are easy to evaluate and have been widely used in prior works (Lorch et al., 2021; Annadani et al., 2021; Geffner et al., 2022; Deleu et al., 2022; Nishikawa-Toomey et al., 2022; Lorch et al., 2022; Atanackovic et al., 2023). However, all these metrics evaluate samples from the posterior against a single graph (the ground truth) while ignoring the uncertainty due to finite data that makes other graphs plausible hypotheses.

Full posterior evaluation metrics. The other metrics sample from the joint posterior over both $\boldsymbol{G}$ and $\boldsymbol{\phi}$ to evaluate the goodness of the posterior approximation:

- Negative Log-Likelihood: It is the negative Log-Likelihood (**NLL**) of held-out observational samples, computed by sampling the posterior model parameters, i.e. $-\mathbb{E}_{\mathbf{X}\sim p_{\mathbf{X}}(\mathbf{X})}\mathbb{E}_{q(\boldsymbol{G},\boldsymbol{\phi}|\boldsymbol{D})}\log p(\mathbf{X} \mid \boldsymbol{G},\boldsymbol{\phi})$. Unlike in other inference problems like Variational Autoencoders (Kingma & Welling, 2013; Rezende et al., 2014), NLL might not be the most suitable for structure learning because a graph with more edges has lower NLL than the ones with fewer edges.
- Interventional Negative Log-Likelihood: Since a posterior defines a generative model of data, interventional data of unseen interventions can be generated and compared with the ground truth data generative process. Interventional Negative Log-Likelihood (**I-NLL**) averaged over different unseen interventions is defined as: $-\frac{1}{d}\sum_{i=1}^{d}\mathbb{E}_{\mathbf{X}\sim p(\mathbf{X}|\mathrm{do}(X_i))}\mathbb{E}_{q(\boldsymbol{G},\boldsymbol{\phi})}\log p(\mathbf{X} \mid \boldsymbol{G},\boldsymbol{\phi},\mathrm{do}(X_i))$
- Interventional Distance Metrics: Similar to interventional negative log-likelihood, Interventional KL-Divergence (**I-KL**) and Interventional Maximum Mean Discrepancy (**I-MMD**) are metrics which measure the divergence between the unseen interventional distributions between the distribution induced by the generative model and that from the ground truth data generative process, i.e $\frac{1}{d}\sum_{i=1}^{d}\mathrm{D}(p_{\mathbf{X}}(\mathbf{X}|\mathrm{do}(X_i)) \;||\; q_{\mathbf{X}}(\mathbf{X}|\mathrm{do}(X_i))$ where D is either KL-divergence or maximum mean discrepancy (Gretton et al., 2012) and $q_{\mathbf{X}}$ is the data distribution induced by the approximated posterior.

NLL and I-NLL require that likelihood can be evaluated which might not be the case if the SCM is not an ANM. Given that most of the works deal with additive noise models, both these metrics have also been used in prior works (Deleu et al., 2023; Lorch et al., 2021; Annadani et al., 2023; Toth et al., 2022; Deleu et al., 2022; Atanackovic et al., 2023).

Despite the extensive use of these metrics in prior work, it is unclear if they are suitable as proxy metrics. As BCD is a relatively new and emerging field, there is no principled case study yet which has addressed the evaluation problem. In the following section, we address this gap by performing an empirical study specifically with the aim to understand the evaluation metrics better.

4. Experiments and Key Results

In this section, we design and perform a wide set of experiments on BCD methods to establish the suitability of the current evaluation metrics. We restrict our attention to linear additive noise models as most of the existing BCD methods are only applicable to this setting. In addition, true posterior can be computed in this setting in closed form, thus ensuring the drawn conclusions are sound. Model misspecification can be quite hard to deal with causal discovery in general (Montagna et al., 2023). Therefore, we ensure in the experiments that all the methods have the same level of expressivity as the true posterior and have access to data with no model misspecification.

Outline of experiments. We mainly aim to understand the following aspects of the present evaluation metrics: (1) How does true posterior perform on these metrics? (2) Do all metrics correlate in terms of the ranking they induce on different models, and are they correlated with metrics which directly compare with the true posterior? (3) Entropy of the true posterior and how consideration of entropy of the true posterior is important for determining the reliability of the evaluation metrics (4) Downstream tasks that might be suitable for BCD when current metrics are not suitable.

4.1. Experimental Setting

Models. We experiment on the following different BCD models: **BCD Nets** (Cundy et al., 2021), **DIBS** (Lorch et al., 2021) and **VBG** (Nishikawa-Toomey et al., 2022) are methods which perform approximate inference on the graph, the parameters of the linear mechanisms and the variance of the noise variables. BCD Nets performs inference based on node permutation matrices using variational inference, DIBS is a particle-based method with Stein Variational Gradient Descent (SVGD) (Liu & Wang, 2016) as its inference engine and VBG is a VI approach with GFlowNets (Bengio et al., 2021). We also include DAG bootstrap (Friedman et al., 1999) with GIES (Hauser & Bühlmann, 2012; Chickering, 2002) for evaluation though it is not strictly a Bayesian inference method as it has been used extensively as a baseline in prior work. Bootstrap GIES (**BGIES**) performs maximum-likelihood estimate on all the parameters of interest on different datasets bootstrapped from the original dataset, and then weighs each estimate with its unnormalized posterior probability. For comparison, we also include a non-Bayesian method by running DIBS deterministically (setting the number of particles of SVGD to 1), called **DIBS Det**. Details of all the methods, including their hyperparameter search procedure are given in Appendix A.1. When applicable, we also include a version of DIBS that directly uses the BGe score (Geiger & Heckerman, 2002) for likelihood (called **DIBS BGe**).

Synthetic data generation. We test all the methods on synthetic data. This enables us access to ground truth as well as to have control over the SCM that generates the

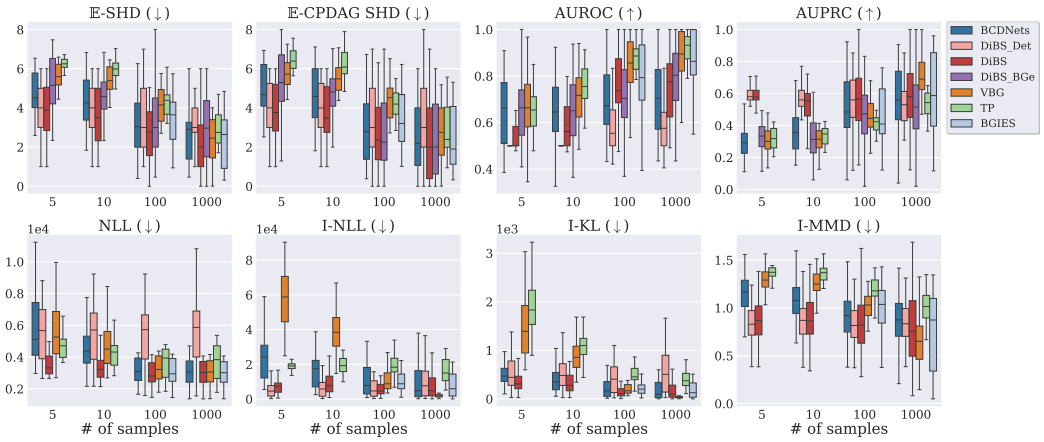

Figure 1. Evaluation of the models on ER1 graphs in the non-identifiable case ($d = 5$). In low sample regimes, true posterior itself is evaluated to be worse on these metrics than their approximations.

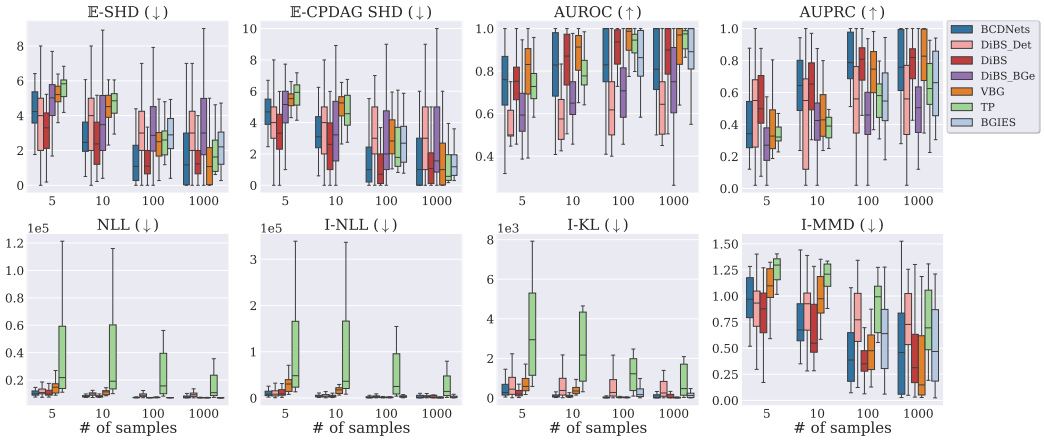

Figure 2. Evaluation of the models on ER1 graphs in the identifiable case ($d = 5$). In low sample regimes, true posterior itself is evaluated to be worse on these metrics than their approximations.

data, thereby ensuring there is no model misspecification. We sample graphs from Erdős–Rényi (ER) (Erdős et al., 1960) and Scale-Free (SF) (Barabási & Albert, 1999) random graph family along with a linear Gaussian ANM. The graphs have expected edge per node of either 1 or 2 (referred to as ER1, ER2, SF1 and SF2). We consider two scenarios for linear ANM: homoscedastic Gaussian (identifiable) and heteroscedastic Gaussian (non-identifiable) (Peters & Bühlmann, 2014). In the first scenario, we set the variance to one, while in the second scenario, the noise variances are sampled from an inverse gamma distribution. The weights are then derived from a multivariate Normal distribution with a mean of 0 and a diagonal covariance matrix corresponding to noise variances. Details of the data generating process is given in Appendix A.2. True posterior can be computed in closed form for both scenarios when $d < 6$. Details of true posterior computation is provided in Appendix C. All the experiments are conducted with 20 different random datasets and 3 random initialization of the model per dataset, resulting in 60 runs for each model.

Metrics for comparison with true posterior. As noted before, the true posterior is the gold standard with which the suitability of the other metrics can be reasonably established. In order to compare the approximate posterior with the true posterior, we use Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) with relevant kernel. More precisely, we compare $q(\boldsymbol{G}|\boldsymbol{D})$ with $p(\boldsymbol{G}|\boldsymbol{D})$ (called **Graph MMD**) using a Hamming Kernel and $q(\boldsymbol{\phi}|\boldsymbol{G}, \boldsymbol{D})$ with $p(\boldsymbol{\phi}|\boldsymbol{G}, \boldsymbol{D})$ using an RBF kernel (called **Params MMD**). We use MMD as it requires only samples from the distribution.

4.2. Key Results

Evaluation on current metrics. We first evaluate all the methods on the metrics outlined in Section 3 to give a representative idea of the performance seen and reported in

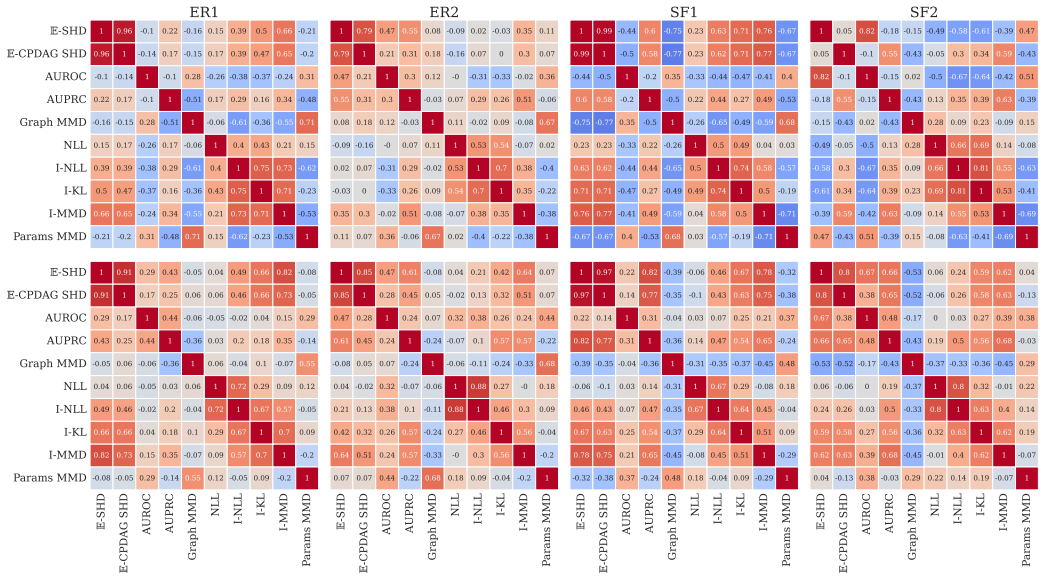

Figure 3. Spearman's rank correlation coefficient between evaluation metrics with 5 samples ($d = 5$). The first and the second rows correspond to the non-identifiable and identifiable cases, respectively. All the graph-based metrics are not correlated with the Graph MMD. Params MMD is not correlated with any of the other metrics. Graph MMD and Params MMD are metrics that evaluate against the true posterior.

prior work. This would serve as a useful context for what is currently being evaluated in the literature. In addition, we also include the performance that true posterior achieves on these metrics. Figure 1 presents results for ER graph for non-identifiable setting and Figure 2 for the identifiable setting. It is interesting to note that when the number of samples is smaller ($d = n = 5$), the true posterior itself performs significantly worse on all metrics, including that of some of the BCD algorithms that are approximating the true posterior. For most applications, especially in biology, $n \approx d$ is a fairly common setting. In fact, many of the algorithms in BCD benchmark on synthetic datasets with the number of samples less than 100 (and in many cases just 50 samples, with d ranging from 10 to 50). As the number of samples increases, the methods perform better on these metrics. However, the relative performance of all the methods, especially the true posterior, does not increase much when the number of samples is increased from 100 to 1000 (see Appendix B for a simple example illustrating this point). This is consistent across different random graph models as well (Appendix E.2). As prior works mostly evaluate on higher dimensional cases where the true posterior is not tractable, this issue of worse performance of true posterior on these metrics has not been demonstrated before. At least preliminarily, this calls into question the suitability of the current evaluation metrics.

Evaluation on true posterior. For comparison, we also present results that evaluate on true posterior with Graph MMD and Params MMD. Figures 5 and 6 presents results for ER1 graphs. The evaluation indicates that the models considered do not estimate either the graph posterior or the parameter posterior well for $d = 5$, as the MMD is greater than 0 for both cases. Similar observation can be made for other graph types (Appendix E.4).

Rank correlation between metrics. In order to further understand the suitability of current metrics, we analyze the Spearman's rank correlation coefficient between different metrics (Spearman, 1961). We rank different methods based on their performance in each metric and measure Spearman's rank correlation coefficient between rankings induced by each metric. It ranges between -1 to 1 – a coefficient of 1 would correspond to perfect correlation and -1 to inverse correlation. In other words, if Spearman's correlation between two metrics is -1, the model that is evaluated as the best under one metric would be evaluated by the other metric as the worst. With Spearman's rank correlation, we aim to analyze the following two questions: (1) Are the different proxy metrics correlated? and (2) More importantly, how correlated are the proxy metrics to the metrics that compare with the true posterior, i.e Graph MMD and Params MMD? Figure 3 presents the result for non-identifiable scenario on a dataset with $d = 5$, $n = 5$. Several interesting conclusions can be drawn. Firstly, the graph-based proxy metrics are not correlated (for example $\mathbb{E}$-SHD and AUPRC), while the interventional-based metrics I-NLL, I-KL, and I-MMD are largely correlated. The correlation is higher in denser graphs. However, it is interesting to note that the interventional metrics do not correlate with NLL. Though the community has

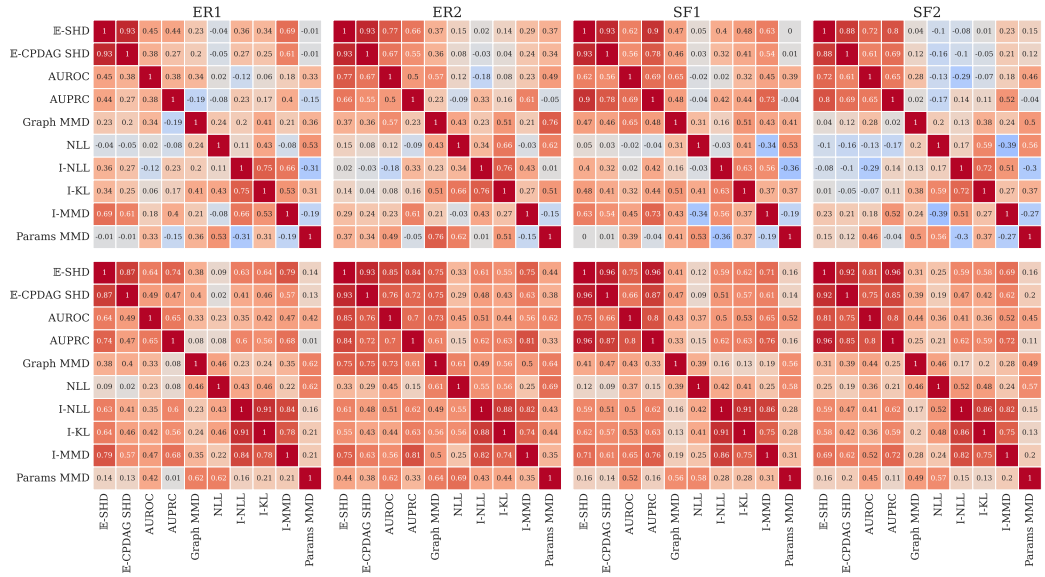

Figure 4. Spearman's rank correlation coefficient between evaluation metrics with 100 samples ($d = 5$). The first and the second rows correspond to the non-identifiable and identifiable cases, respectively. All the graph-based metrics are correlated with each other and also the Graph MMD. Params MMD is also correlated with other metrics. Graph MMD and Params MMD are metrics that evaluate against the true posterior.

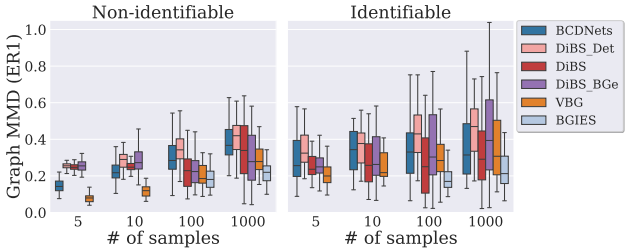

Figure 5. Graph MMD of the models on ER1 graphs ($d = 5$).

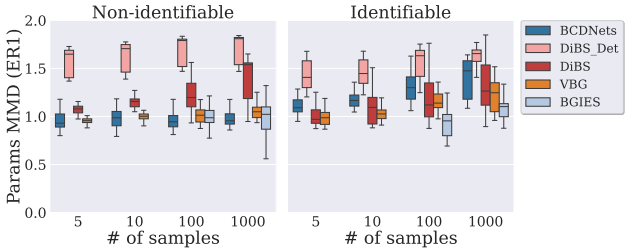

Figure 6. Params MMD of the models on ER1 graphs ($d = 5$).

relied on NLL as a reasonable metric, it is sensitive to measurement errors and scale of the data (Lorch et al., 2022; Reisach et al., 2021). Secondly, all the graph-based metrics have very little to no correlation with graph MMD, and the Params MMD is not correlated with other metrics.

In order to further understand if the same pattern exists in other settings, we examine the Spearman's rank correlation coefficient for the identifiable setting (Figure 3 bottom row). We observe a very interesting pattern. Unlike in the unidentifiable case, the graph-based metrics are more correlated, and the interventional metrics are correlated with each other and also with NLL. However, the graph-based proxy metrics are **not** well correlated with Graph MMD, although the level of correlation is slightly higher than the non-identifiable case. Similarly, Params MMD is not correlated with other metrics. This indicates that, while the metrics are usually correlated between each other in terms of ranking the models, the ranking that they induce would be different from the rankings induced by comparison with the true posterior when the number of samples is less.

Correlation between metrics for large datasets. In order to see if the same correlation pattern persists for a higher number of samples, we plot Spearman's correlation coefficient for $N = 100$ (Figure 4). We observe that the correlation between Graph MMD and graph-based proxy metrics is higher than before, with the identifiable scenario having a much higher correlation than the non-identifiable one. A similar observation can be made for Params MMD. It is reasonable to expect based on this result that the current proxy metrics are viable for evaluation of BCD algorithms with more samples and an identifiable underlying SCM.

A similar observation when $N = \{10, 1000\}$ (Figures 21 and 22) reveals that the proxy metrics are not correlated with the gold-standard metrics in practical settings with less data and non-identifiability, where being Bayesian about causal discovery is supposed to be advantageous. This calls

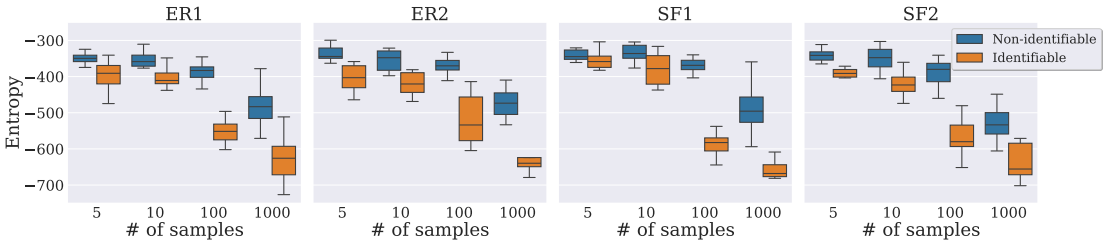

Figure 7. Entropy of the true posterior with different number of training samples ($d = 5$). Entropy decreases as the number of samples increases. Entropy also decreases with identifiability.

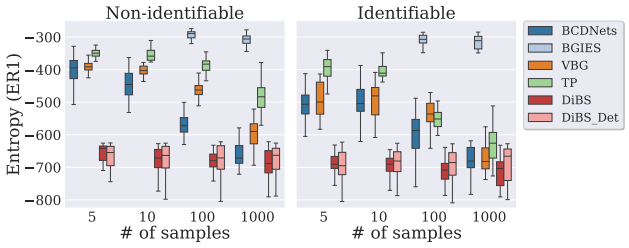

Figure 8. Entropy of the models on ER1 graphs with different number of training samples ($d = 5$). Entropy of most models except DIBS decreases with more data samples.

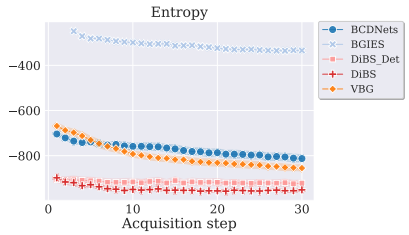

Figure 9. Changes of the entropy on ER1 graphs in an incremental setting in the non-identifiable case ($d = 10$). We start with 10 observational samples and at each step, we add 5 interventional samples and retrain the models. For all the models, as we give them more interventional samples, the entropy does not decrease substantially.

into question the suitability of the current proxy metrics in these settings.

Entropy of true posterior. From the rank correlation, it is clear that the proxy metrics are only reliable with a high number of data samples and also depend on the nature of SCM, i.e. identifiability. We note that both of these aspects are connected to the entropy of the true posterior. In fact, it is reasonable to expect that the entropy of the true posterior decreases as the number of samples is increased. If an SCM is non-identifiable, all the graphs within the MEC, which could be exponentially many, have a high probability, thereby making the posterior more entropic. We empirically demonstrate this on the true posterior corresponding to different settings. We use an approximator of entropy which only requires samples from the distribution (Kozachenko & Leonenko, 1987). Details are given in Appendix D. Figure 7 illustrates the entropy of true posterior under various settings. It can be seen that entropy decreases with higher samples and identifiability. Since the proxy metrics are usually derived from causal discovery, they do not reflect the quality of approximation when there are many graphs (and corresponding parameters) with high posterior probability. Therefore, it is reasonable to conclude that the current metrics are not suitable where BCD is most desirable – higher entropy settings of the true posterior.

Entropy of models. Using the same entropy estimator, we also examine the entropy of BCD models. In particular, we are interested in the following two aspects: 1) How entropic are the BCD algorithms in comparison to true posterior? 2) Does the entropy decrease as more observational and interventional data is given? Our goal is not to decide which method is the best but to understand if the methods respond to additional data to reduce their entropy. Figure 8 presents the results for ER1 graphs. Most of the methods have entropies the same as that of the true posterior, except DIBS, which always gives very low entropy solutions. Similar behaviour is seen in other graph types as well (Appendix E.4). When interventional data is given and the model is updated at each step, similar to an experimental design loop (Tigas et al., 2022), the reduction in entropy is very good for VBG and BCD Nets while it does not necessarily decrease for DIBS and BGIES (Figure 9).

Effect of prior. An important factor in obtaining a good estimate of the true posterior is the choice of the prior over the graphs and parameters of the model, i.e. $p(\boldsymbol{G}, \boldsymbol{\phi})$. Apart from DIBS, all methods do not use an informative prior. DIBS leverages the knowledge of the underlying data generative process to design priors that match that are informative. While a more extensive study is required to understand the performance of various methods due to the choice of the prior, we do notice that for SF graphs, the solution of DIBS is completely dominated by the prior in low data regimes. While this is the intended behavior in a Bayesian setting,

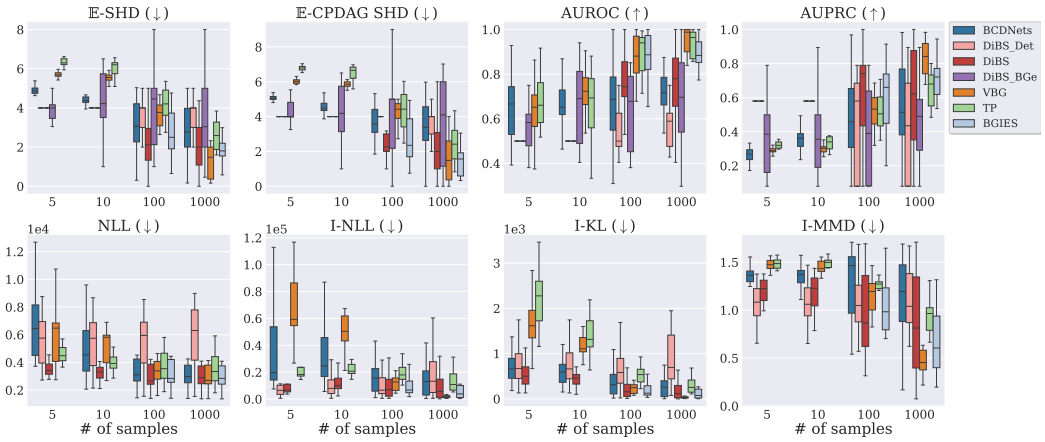

Figure 10. Evaluation of the models on SF1 graphs in the non-identifiable case ($d = 5$). In low sample regimes, true posterior itself is evaluated to be worse on these metrics than their approximations.

the solution of DIBS is very low entropy. In fact, we found that it samples a graph with no edges in low data regimes (Figures 10 and 19). However, for ER graphs, the prior is less dominant than for SF graphs and it leads to reasonable samples with DIBS.

5. Possible Alternative Evaluation Procedures

Although our study mainly focuses on identifying the potential issues in the evaluation metrics for BCD, we suggest two possible alternate way of evaluating BCD algorithms by considering the empirical results obtained in Section 4.

5.1. Experimental Design

As seen in Figure 7, after acquiring enough (interventional) data, the true posterior will have less entropy. Therefore, one possible way to evaluate the BCD algorithms is to evaluate it downstream, for instance, by performing experimental design to acquire enough interventional data and then evaluating with the proxy metrics when they are more suitable. The task of choosing the intervention that results in the highest expected reduction in entropy is concerned with Bayesian optimal experimental design (Lindley, 1956; Chaloner & Verdinelli, 1995; Foster et al., 2019), a downstream task of Bayesian Causal Discovery. Many specific experimental design procedures exist for BCD (Tigas et al., 2022; 2023; Agrawal et al., 2019; Zhang et al., 2023; Toth et al., 2022) that can be used to collect data and perform evaluation.

5.2. Causal Effect Estimation

In some applications, proxy metrics either require access to the underlying ground truth graph or other parameters thereof, which might not be available. In such a case, experimental design as a downstream evaluation tool might not be applicable. An alternative evaluation procedure in such a case, therefore, is the downstream task of causal effect estimation. Causal effect estimation is the task of estimating the state of a variable that is part of the causal model when the system is subject to interventions. This method has been thoroughly studied in Emezue et al. (2023) and has shown to be useful in identifiable cases with few data samples.

6. Discussion and Conclusion

In this work, we demonstrate the shortcomings of the present evaluation metrics for Bayesian Causal Discovery with an extensive empirical study. Our key result is that when the true posterior has high entropy - which is the case with less data and non-identifiability, current evaluation metrics do not lead to the same ranking of different BCD models compared to metrics that involve the true posterior. Therefore, evaluation of BCD should be considered carefully in settings with limited data and identifiability of SCM. This challenge of evaluating BCD in these settings could potentially be overcome by evaluation in downstream tasks: for example, causal effect estimation or Bayesian experimental design to acquire interventional samples, after which the true posterior has less entropy which enables reliable evaluation with current metrics. While our study sheds a light on evaluation procedures and their shortcomings in BCD, our study is limited to causally sufficient linear additive noise models. As the field of Bayesian Causal Discovery progresses in terms of posterior approximation in settings where these assumptions do not hold, a similar analysis as presented in this work might be necessary for such settings.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded

by Knut and Alice Wallenberg Foundation, and the computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

Impact Statement

This work is concerned with proper evaluation of Bayesian causal discovery algorithms which highly benefits the research community. The authors do not foresee negative societal impact of this work beyond what is brought about by general advances in machine learning.

References

Agrawal, R., Squires, C., Yang, K., Shanmugam, K., and Uhler, C. Abcd-strategy: Budgeted experimental design for targeted causal structure discovery. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3400–3409. PMLR, 2019.

Andersson, S. A., Madigan, D., and Perlman, M. D. A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

Annadani, Y., Rothfuss, J., Lacoste, A., Scherrer, N., Goyal, A., Bengio, Y., and Bauer, S. Variational causal networks: Approximate bayesian inference over causal structures. *arXiv preprint arXiv:2106.07635*, 2021.

Annadani, Y., Pawlowski, N., Jennings, J., Bauer, S., Zhang, C., and Gong, W. Bayesdag: Gradient-based posterior inference for causal discovery. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Atanackovic, L., Tong, A., Wang, B., Lee, L. J., Bengio, Y., and Hartford, J. Dyngfn: Towards bayesian inference of gene regulatory networks with gflownets. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Barabási, A.-L. and Albert, R. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

Bengio, E., Jain, M., Korablyov, M., Precup, D., and Bengio, Y. Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, 34:27381–27394, 2021.

Chaloner, K. and Verdinelli, I. Bayesian experimental design: A review. *Statistical science*, pp. 273–304, 1995.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of machine learning research*, 3 (Nov):507–554, 2002.

Chickering, D. M., Heckerman, D., and Meek, C. A bayesian approach to learning bayesian networks with local structure. *arXiv preprint arXiv:1302.1528*, 2013.

Cho, H., Berger, B., and Peng, J. Reconstructing causal biological networks through active learning. *PloS one*, 11 (3):e0150611, 2016.

Cundy, C., Grover, A., and Ermon, S. Bcd nets: Scalable variational approaches for bayesian causal discovery. *Advances in Neural Information Processing Systems*, 34: 7095–7110, 2021.

Deleu, T., Góis, A., Emezue, C., Rankawat, M., Lacoste-Julien, S., Bauer, S., and Bengio, Y. Bayesian structure learning with generative flow networks. In *Uncertainty in Artificial Intelligence*, pp. 518–528. PMLR, 2022.

Deleu, T., Nishikawa-Toomey, M., Subramanian, J., Malkin, N., Charlin, L., and Bengio, Y. Joint bayesian inference of graphical structure and parameters with a single generative flow network. *arXiv preprint arXiv:2305.19366*, 2023.

Dibaeinia, P. and Sinha, S. Sergio: A single-cell expression simulator guided by gene regulatory networks. *Cell systems*, 2020. URL https://api.semanticscholar.org/CorpusID:221467051.

Emezue, C. C., Drouin, A., Deleu, T., Bauer, S., and Bengio, Y. Benchmarking bayesian causal discovery methods for downstream treatment effect estimation. *arXiv preprint arXiv:2307.04988*, 2023.

Erdős, P., Rényi, A., et al. On the evolution of random graphs. *Publ. math. inst. hung. acad. sci*, 5(1):17–60, 1960.

Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Teh, Y. W., Rainforth, T., and Goodman, N. Variational bayesian optimal experimental design. *Advances in Neural Information Processing Systems*, 32, 2019.

Friedman, N. and Koller, D. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine learning*, 50:95–125, 2003.

Friedman, N., Goldszmidt, M., and Wyner, A. Data analysis with bayesian networks: A bootstrap approach. *arXiv preprint arXiv:1301.6695*, 1999.

Geffner, T., Antoran, J., Foster, A., Gong, W., Ma, C., Kiciman, E., Sharma, A., Lamb, A., Kukla, M., Pawlowski, N., et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.

Geiger, D. and Heckerman, D. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30 (5):1412–1440, 2002.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Hägele, A., Rothfuss, J., Lorch, L., Somnath, V. R., Schölkopf, B., and Krause, A. Bacadi: Bayesian causal discovery with unknown interventions. In *International Conference on Artificial Intelligence and Statistics*, pp. 1411–1436. PMLR, 2023.

Hauser, A. and Bühlmann, P. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464, 2012.

Heckerman, D., Meek, C., and Cooper, G. A bayesian approach to causal discovery. *Innovations in Machine Learning: Theory and Applications*, pp. 1–28, 2006.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kozachenko, L. F. and Leonenko, N. N. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

Kuipers, J., Moffa, G., and Heckerman, D. Addendum on the scoring of gaussian directed acyclic graphical models. 2014.

Lindley, D. V. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27 (4):986–1005, 1956.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Lombardi, D. and Pant, S. Nonparametric k-nearest-neighbor entropy estimator. *Physical Review E*, 93(1): 013310, 2016.

Lorch, L., Rothfuss, J., Schölkopf, B., and Krause, A. Dibs: Differentiable bayesian structure learning. *Advances in Neural Information Processing Systems*, 34:24111–24123, 2021.

Lorch, L., Sussex, S., Rothfuss, J., Krause, A., and Schölkopf, B. Amortized inference for causal structure learning. *Advances in Neural Information Processing Systems*, 35:13104–13118, 2022.

Lyle, C., Mehrjou, A., Notin, P., Jesson, A., Bauer, S., Gal, Y., and Schwab, P. Discobax discovery of optimal intervention sets in genomic experiment design. In *International Conference on Machine Learning*, pp. 23170–23189. PMLR, 2023.

Montagna, F., Mastakouri, A. A., Eulig, E., Noceti, N., Rosasco, L., Janzing, D., Aragam, B., and Locatello, F. Assumption violations in causal discovery and the robustness of score matching. *arXiv preprint arXiv:2310.13387*, 2023.

Nishikawa-Toomey, M., Deleu, T., Subramanian, J., Bengio, Y., and Charlin, L. Bayesian learning of causal structure and mechanisms with gflownets and variational bayes. *arXiv preprint arXiv:2211.02763*, 2022.

Pearl, J. *Causality*. Cambridge university press, 2009.

Peters, J. and Bühlmann, P. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Reisach, A., Seiler, C., and Weichwald, S. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Spearman, C. The proof and measurement of association between two things. 1961.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, prediction, and search*. MIT press, 2000.

Sussex, S., Makarova, A., and Krause, A. Model-based causal bayesian optimization. *arXiv preprint arXiv:2211.10257*, 2022.

Tejada-Lapuerta, A., Bertin, P., Bauer, S., Aliee, H., Bengio, Y., and Theis, F. J. Causal machine learning for single-cell genomics. *arXiv preprint arXiv:2310.14935*, 2023.

Tigas, P., Annadani, Y., Jesson, A., Schölkopf, B., Gal, Y., and Bauer, S. Interventions, where and how? experimental design for causal models at scale. *Advances in Neural Information Processing Systems*, 2022.

Tigas, P., Annadani, Y., Ivanova, D. R., Jesson, A., Gal, Y., Foster, A., and Bauer, S. Differentiable multi-target causal bayesian experimental design. In *International Conference on Machine Learning*, pp. 34263–34279. PMLR, 2023.

Tong, S. and Koller, D. Active learning for structure in bayesian networks. In *International joint conference on artificial intelligence*, volume 17, pp. 863–869. Lawrence Erlbaum Associates ltd, 2001.

Toth, C., Lorch, L., Knoll, C., Krause, A., Pernkopf, F., Peharz, R., and Von Kügelgen, J. Active bayesian causal inference. *Advances in Neural Information Processing Systems*, 35:16261–16275, 2022.

Zhang, Z., Li, C., Chen, X., and Xie, X. Bayesian active causal discovery with multi-fidelity experiments. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

A. Experimental Details

A.1. Models

This section provides a casual overview of the models featured in the study, along with details regarding their implementation and the choices made for hyperparameters.

BCD-Nets. BCD-Nets (Cundy et al., 2021) is a bayesian posterior approximation method designed to model linear-Gaussian SCMs. It decomposes the weighted adjacency matrix W of the linear SCM into a permutation matrix P and a strictly lower-triangular matrix L, i.e. $W = PLP^T$. It uses variational inference to learn the posterior distribution $q_\phi(P, L, \Sigma)$ over the SCM parameters by maximizing the Evidence Lower Bound (ELBO) w.r.t. variational parameters ϕ. For the implementation, we utilize the public implementation of BCD-Nets [1] with the same hyperparameters as in the original paper except for the number of training steps which we change to 20k steps.

DIBS. DIBS (Lorch et al., 2021) is a fully differentiable bayesian posterior approximation method suitable to model both linear and non-linear SCMs. It proposes to transfer the posterior inference into the latent space of a probabilistic graph representation and assumes there is a latent variable Z that models the generative process of the underlying causal graph. They factorize the joint distribution $p(Z, G, \Theta, D)$ in a way that allows for joint posterior inference of both the graph structure and the conditional distribution parameters. To be more precise, $p(Z, \boldsymbol{G}, \boldsymbol{\phi}, \boldsymbol{D}) = p(Z)p(\boldsymbol{G} \mid Z)p(\boldsymbol{\phi} \mid \boldsymbol{G})p(\boldsymbol{D} \mid \boldsymbol{G}, \boldsymbol{\phi})$. They apply the gradient-based SVGD algorithm (Liu & Wang, 2016) for sampling. In this work, we utilize 3 different versions of DiBS+: the linear version (we refer to as *DIBS*), a deterministic variant of *DiBS* in which we have only 1 particle in the model (referred to as *DIBS-Det*), and a marginal version (we refer to as *DIBS-Bge*) where the marginal posterior over the graphs, i.e. $p(G|D)$, is computed using the Bayesian Gaussian Equivalent (*BGe*) marginal likelihood. Also, we use the implementation of Tigas et al. (2022)[2] and change it to learn the noise variances together with other parameters. For all experiments, we set the σ_z, α, γ_z, and γ_θ to 0.5, 0.02, 5, and 500, respectively, use 50 particles, and run the model for 20k iterations. We use the default values for other hyperparameters.

VBG. VBG (Nishikawa-Toomey et al., 2022) is another bayesian posterior approximation model suitable designed to model linear-Gaussian SCMs. It extends the *DAG-GFlowNet* (Deleu et al., 2022) to not only learn the graph structure but also the parameters of a linear Gaussian model between the variables in the DAG. In order to model the posterior distribution over the parameters, it utilizes *GFlowNets* (Bengio et al., 2021). We use the public implementation of VBG [3] and use its default hyperparameters for all experiments. As VBG assumes fixed noise variances, we experimented with various values for the noise variance and determined that 0.1 yields the best results in our settings.

DAG Bootstrap. DAG Bootstrap (Friedman et al., 1999) is a non-parametric model that performs model averaging by bootstrapping the data to yield a collection of synthetic datasets. Each dataset is then utilized to learn an individual graph and its associated causal mechanisms, employing the score-based GIES algorithm (Chickering, 2002; Hauser & Bühlmann, 2012). The ensemble of distinct single graphs approximates the posterior by assigning weights to each graph based on its unnormalized posterior probability. For our experiments, we employ the implementation of Tigas et al. (2022)[2], and use 100 bootstraps.

A.2. Synethetic Dataset Details

In this study, we adopt Erdős–Rényi (ER) (Erdős et al., 1960) and Scale-Free (SF) (Barabási & Albert, 1999) graphs as the underlying graph structures for all experiments, utilizing a linear structural equation model (SEM) with 5 and 10 nodes. We generated the graph by setting an expected edge of 1 or 2 for each node. For the SCM weights and noises, we considered two scenarios. In the first scenario, we introduce Gaussian noise with equal variances across all nodes, with the variance value set to 1, and sample the weights of the SCM from independent normal distributions with the mean and the variance set to 0 and 2 respectively. In this scenario, the underlying causal model will be identifiable. In the second scenario, we explore a non-equal variance case where the noise variances are sampled from an inverse gamma distribution with α and β set to 4 and 0.5, respectively. The parameters of the inverse gamma distribution are chosen to restrict the noise variances to a low value, preventing the generation of data with high levels of noise. The weights are then derived from a Multivariate Normal

[1] https://github.com/ermongroup/BCD-Nets
[2] https://github.com/yannadani/cbed
[3] https://github.com/mizunt1/vbg

distribution with a mean of 0 and a diagonal covariance matrix corresponding to noise variances. Data was then sampled using ancestral sampling, and different numbers of training samples ($N = \{5, 10, 100, 1000\}$ were generated for different experiments.

B. Limitations of Graph Metrics: Example

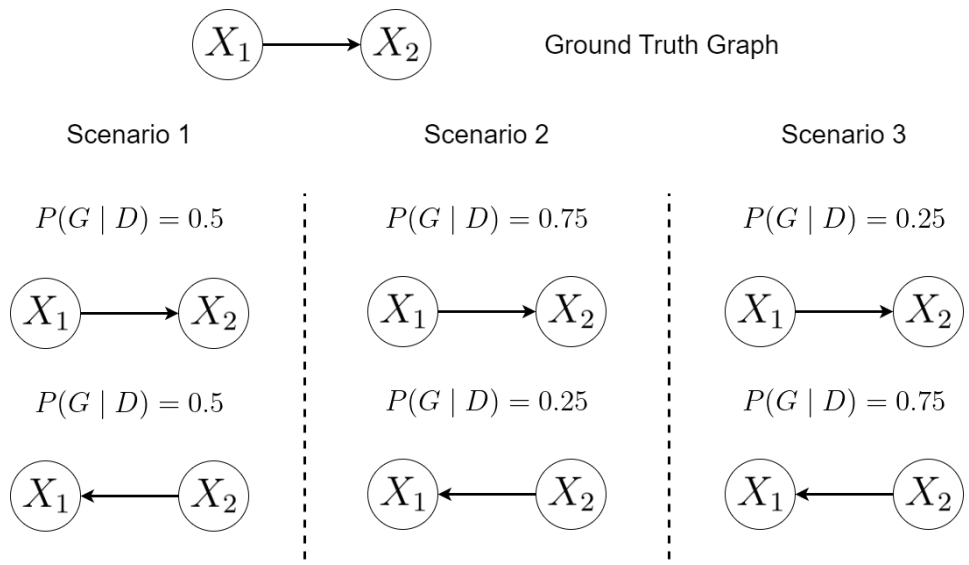

Figure 11. A simple example showing the shortcomings of the graph-based metrics in evaluating posterior distributions. Each scenario corresponds to a posterior distribution over the possible graphs. In the first scenario, $\mathbb{E}$-SHD is 0.5, $\mathbb{E}$-CPDAG SHD is 0, AUROC is 0.66, and AUPRC is 0.5375. In the second scenario, $\mathbb{E}$-SHD is 0.25, $\mathbb{E}$-CPDAG SHD is 0, AUROC is 1, and AUPRC is 0.775. In the third scenario, $\mathbb{E}$-SHD is 0.75, $\mathbb{E}$-CPDAG SHD is 0, AUROC is 0.66, and AUPRC is 0.2625. Suppose if the model was non-identifiable, then our posterior would correspond to scenario 1 even with lots of samples. However, we don't necessarily get the best performance as evaluated by these metrics. Rather, approximate inference method might lead to solutions similar to other scenarios (for example, scenario 2) which are evaluated to be better than the true posterior (scenario one).

C. True Posterior Computation

Note that for an ANM, the likelihood can be evaluated through the noise variable, which we assume to be Gaussian (Geffner et al., 2022). Therefore, $p(\boldsymbol{D} \mid \boldsymbol{G}, \boldsymbol{\phi}) = \prod_{j=1}^{N} \prod_{i=1}^{d} \mathcal{N}(\boldsymbol{\gamma}_i^T X_{\text{pa}(i)}^{(j)}, \sigma_i^2)$

C.1. Parameter Posterior

We follow the posterior computation from (Cho et al., 2016). More precisely, $\sigma^2 \sim$ Inv-Gamma(α, β) and $\boldsymbol{\phi}_i \sim \mathcal{N}(\mu_i, \sigma^2(\Lambda_i)^{-1})$. Let $\mathbf{X}_{\text{pa}(i)} \in \mathbb{R}^{N \times |\text{pa}|}$ be the matrix of parents for variable i and $\mathbf{X}_i \in \mathbb{R}^N$ be the vector of samples of variable i. The posterior over parameters has the same form with parameters for a given graph:

$$
\begin{aligned}
\Lambda_i' &:= \mathbf{X}_{\text{pa}(i)}^T \mathbf{X}_{\text{pa}(i)} + \Lambda_i \\
\mu_i' &:= (\Lambda_i')^{-1}(\Lambda_i \mu_i + \mathbf{X}_{\text{pa}(i)} \mathbf{X}_i) \\
\alpha' &:= \alpha + \frac{N}{2} \\
\beta' &:= \beta + \frac{1}{2}(\mathbf{X}_i^T \mathbf{X}_i + \mu_i^T \Lambda_i \mu_i - (\mu_i')^T \Lambda_i' \mu_i')
\end{aligned}
$$

In this work, we set $\alpha = 4$, $\beta = 0.5$ and $\Lambda = \mathbb{I}$.

C.2. Graph Posterior

The marginal likelihood function $p(\boldsymbol{D} \mid \boldsymbol{G})$ can also be obtained in closed form, through which the graph posterior $p(\boldsymbol{G} \mid \boldsymbol{D})$ can be derived by enumerating all the possible graphs. For the identifiable case, the marginal likelihood is given by:

$$p(\boldsymbol{D} \mid \boldsymbol{G}) = (2\pi)^{Nd} \cdot \frac{(\beta)^{d\alpha}}{(\beta')^{d\alpha'}} \cdot \frac{\Gamma(\alpha')^d}{\Gamma(\alpha)^d} \prod_{i=1}^{d} \sqrt{\frac{\det(\Lambda_i)}{\det(\Lambda_i')}}$$

If the posterior has to ensure that all the graphs within the MEC have the same probability given large number of samples, it can be ensured with the BGe score (Geiger & Heckerman, 2002). The marginal likelihood is given in (Kuipers et al., 2014) (Equation 6), and we use the implementation of (Lorch et al., 2021). Note that BGe score assumes that the parameter priors are sampled from a Gaussian-Wishart distribution, instead of Gaussian-Inverse Gamma. Although strictly this assumption is violated in our data generative process, the computation of $p(\boldsymbol{G}|\boldsymbol{D})$ is still valid.

D. Entropy Estimator

For any random variable $\mathbf{Y} \in \mathbb{R}^p$, the Kozachenko-Leonenko estimate of the entropy $\mathrm{H}(\mathbf{Y})$, with N iid samples from $p_{\mathbf{Y}}$ is given by (Kozachenko & Leonenko, 1987):

$$\hat{\mathrm{H}}_{\mathrm{KL}}(\mathbf{Y}) = \psi(N) - \psi(n) + \log(c_p) + \frac{p}{N} \sum_{i=1}^{N} \log(\epsilon(i)) \tag{3}$$

where $\epsilon(i)$ is the distance of the ith sample to its nth nearest neighbor, $c_p = \frac{\pi^{\frac{p}{2}}}{\Gamma(1+\frac{p}{2})}$, $\psi(\cdot)$ is the digamma function and $\Gamma(\cdot)$ is the Gamma function. As $\mathbf{Y}$ corresponds to parameters of the causal model and the causal graph in our case, we measure $\hat{\mathrm{H}}_{\mathrm{KL}}(\mathbf{Y})$ of the distance between likelihoods of the samples induced by the posterior estimates, as that would reflect the information geometry of the approximate posterior better than the parameters themselves. More precisely, we measure the Kozachenko-Leonenko estimate of the entropy on between distances of likelihoods of held-out data as measured by a kernel.

$$\hat{\mathrm{H}}(\boldsymbol{G}, \boldsymbol{\phi}) \approx \hat{\mathrm{H}}_{\mathrm{KL}} \left[\mathop{\mathbb{E}}_{p_{\mathbf{X}}} \mathop{\mathbb{E}}_{\boldsymbol{G}', \boldsymbol{\phi}' \sim q} \left[k(\log p(\mathbf{X} \mid \boldsymbol{G}, \boldsymbol{\phi}), \log p(\mathbf{X} \mid \boldsymbol{G}', \boldsymbol{\phi}')) \right] \right] \tag{4}$$

where $k(\cdot, \cdot)$ is the RBF kernel. We use the implementation provided by (Lombardi & Pant, 2016).

E. Additional Results

In this section, we report additional results and show the evaluation of models on different graph types.

E.1. Effect of Data Normalization.

Recently, it has been shown that synthetic data might induce *varsortability* bias, i.e. causal discovery algorithms take advantage of increasing marginal variance across the causal graph from root to leaf (Reisach et al., 2021). In order to account for this, we run all the methods wherein the marginal variance of each variable is roughly 1, and plot rank correlation (Figures 12 to 15). We observe that a similar pattern of correlation holds as before when the variables had different scales.

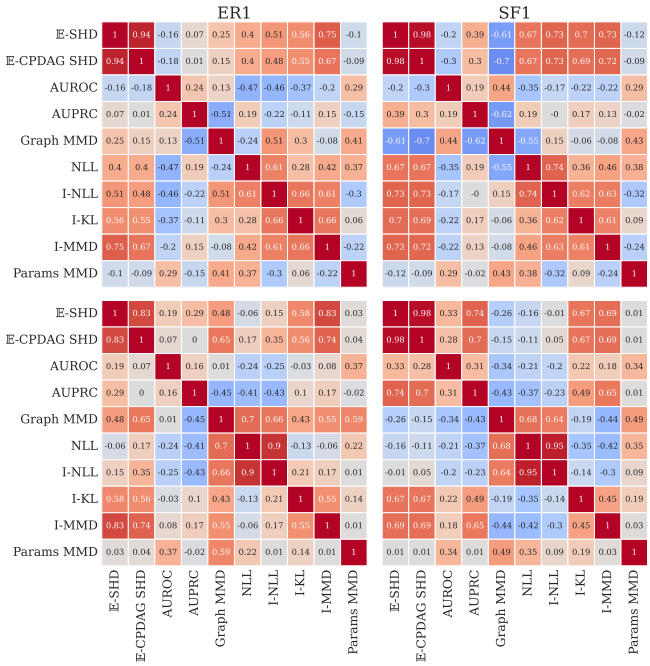

Figure 12. Spearman's rank correlation coefficient between evaluation metrics with 5 normalized training samples ($d = 5$). The first and the second rows correspond to the non-identifiable and identifiable cases, respectively. Similar to the unnormalized case, all the graph-based metrics are not correlated with the Graph MMD, and Params MMD is not correlated with any of the other metrics. Graph MMD and Params MMD are metrics that evaluate against the true posterior.

E.2. Additional Results: Evaluation on Metrics

Here we report additional results. Figures 16 to 20 show the performance of models on different graph types in both non-identifiable and identifiable cases.

E.3. Additional Results: Correlation Between Metrics

Figures 21 and 22 show the Spearman's rank correlation coefficient between evaluation metrics on 5-node graphs with 10 and 1000 samples, respectively.

E.4. Additional Results: Entropy and Comparison with True Posterior

Figure 23 shows the entropy of the models on 5-node graphs with different graph types. Figures 24 and 25 show the Graph MMD and Params MMD of the models on 5-node graphs with different types.

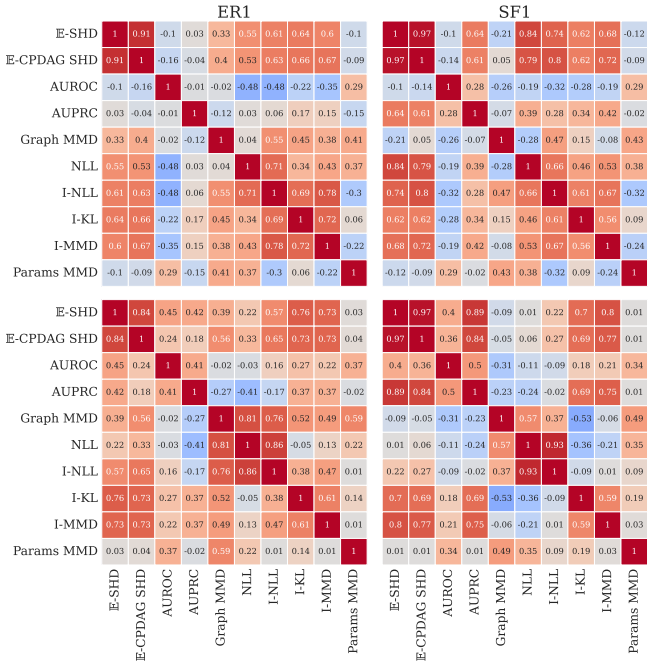

Figure 13. Spearman's rank correlation coefficient between evaluation metrics with 10 normalized training samples ($d = 5$). The first and the second rows correspond to the non-identifiable and identifiable cases, respectively. The same pattern as in the non-normalized case is observed.

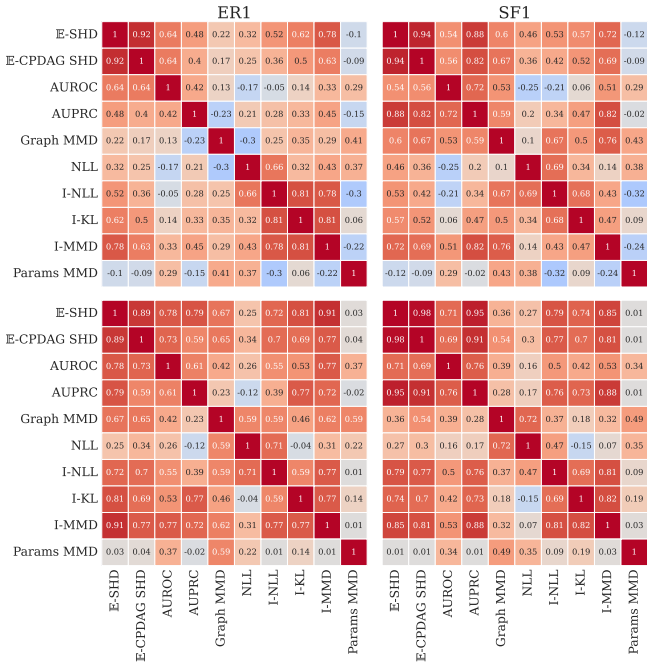

Figure 14. Spearman's rank correlation coefficient between evaluation metrics with 100 normalized training samples ($d = 5$). The first and the second rows correspond to the non-identifiable and identifiable cases, respectively. The same pattern as in the non-normalized case is observed.

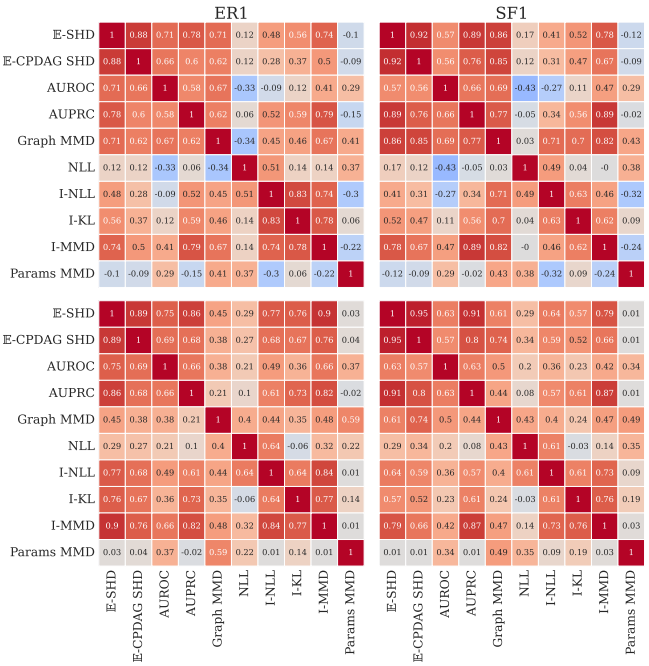

Figure 15. Spearman's rank correlation ceofficient between evaluation metrics with 1000 normalized training samples ($d = 5$). The first and the second rows correspond to the non-identifiable and identifiable cases, respectively. Similar to the unnormalized case, All the graph-based metrics are correlated with each other and also the Graph MMD. Params MMD is also correlated with other metrics. Graph MMD and Params MMD are metrics that evaluate against the true posterior.

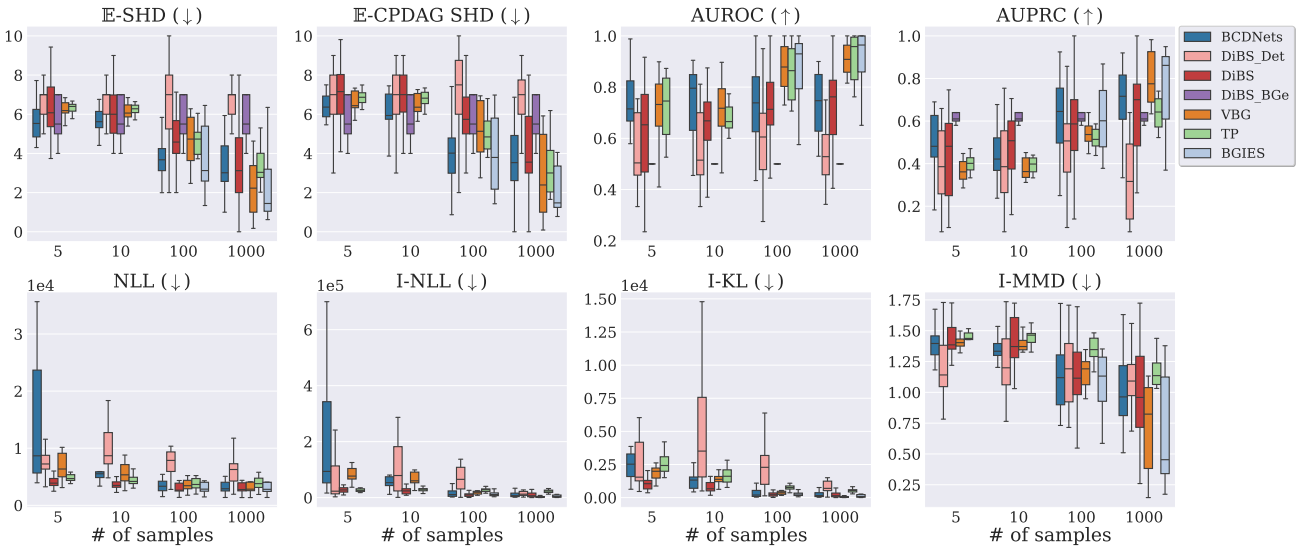

Figure 16. Evaluation of the models on ER2 graphs in the non-identifiable case ($d = 5$). In low sample regimes, true posterior itself is evaluated to be worse on these metrics than their approximations.

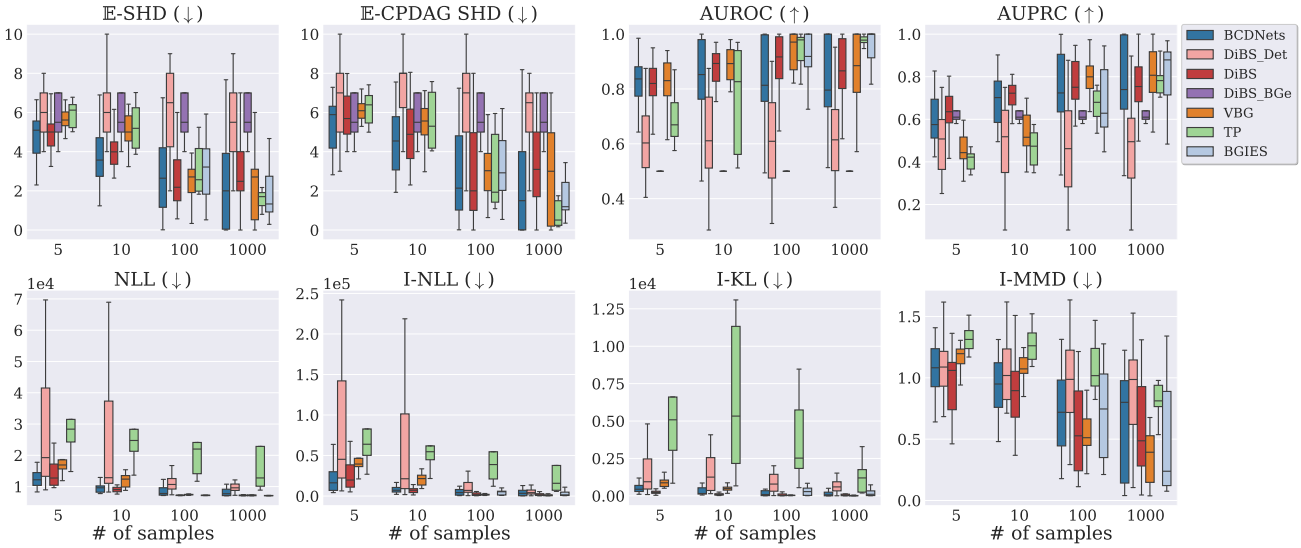

Figure 17. Evaluation of the models on ER2 graphs in the identifiable case ($d = 5$). In low sample regimes, true posterior itself is evaluated to be worse on these metrics than their approximations.

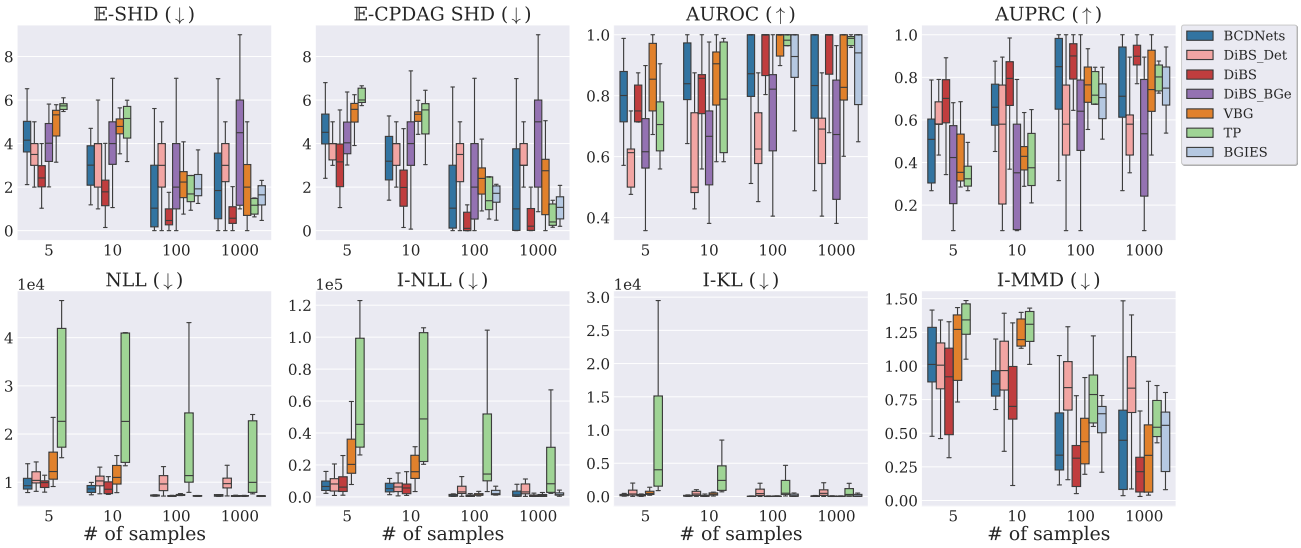

Figure 18. Evaluation of the models on SF1 graphs in the identifiable case ($d = 5$). In low sample regimes, true posterior itself is evaluated to be worse on these metrics than their approximations.

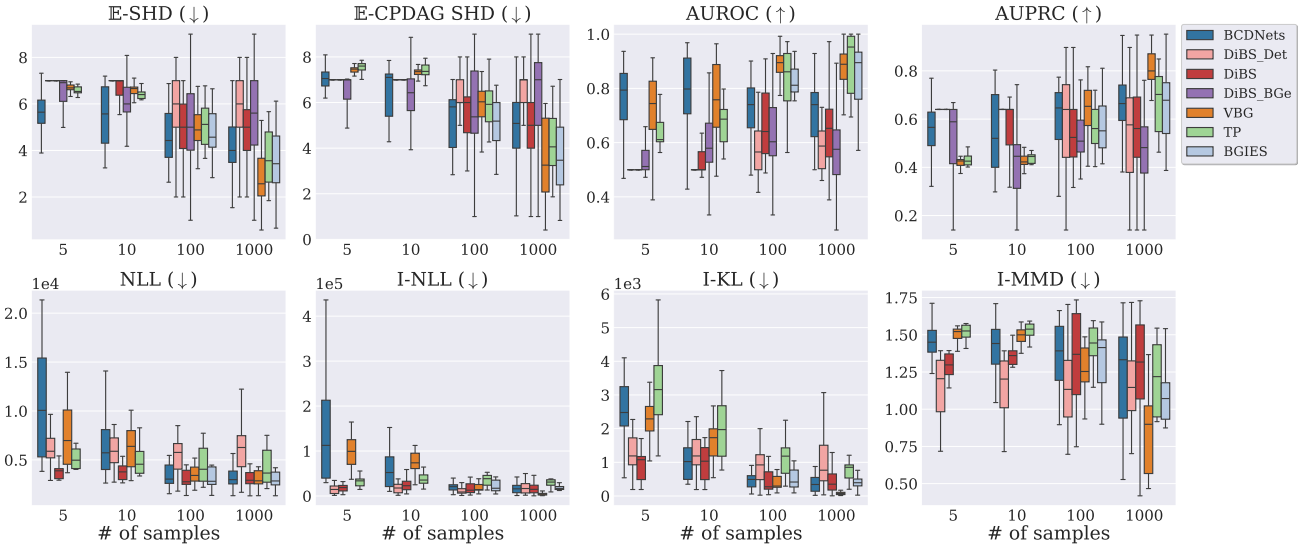

Figure 19. Evaluation of the models on SF2 graphs in the non-identifiable case ($d = 5$). In low sample regimes, true posterior itself is evaluated to be worse on these metrics than their approximations.

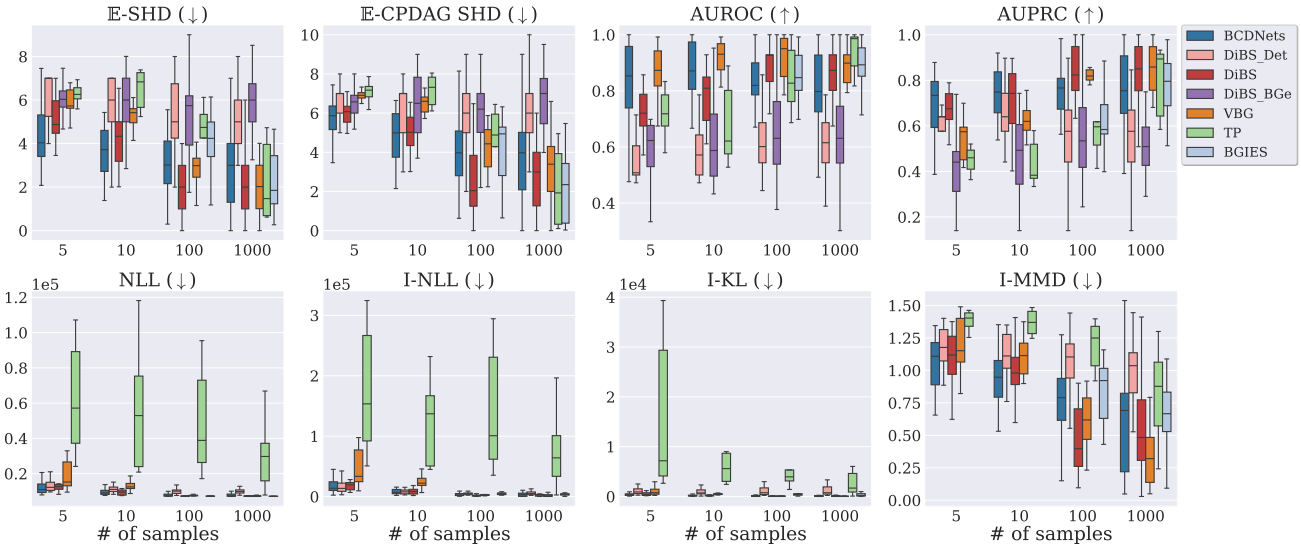

Figure 20. Evaluation of the models on SF2 graphs in the identifiable case ($d = 5$). In low sample regimes, true posterior itself is evaluated to be worse on these metrics than their approximations.

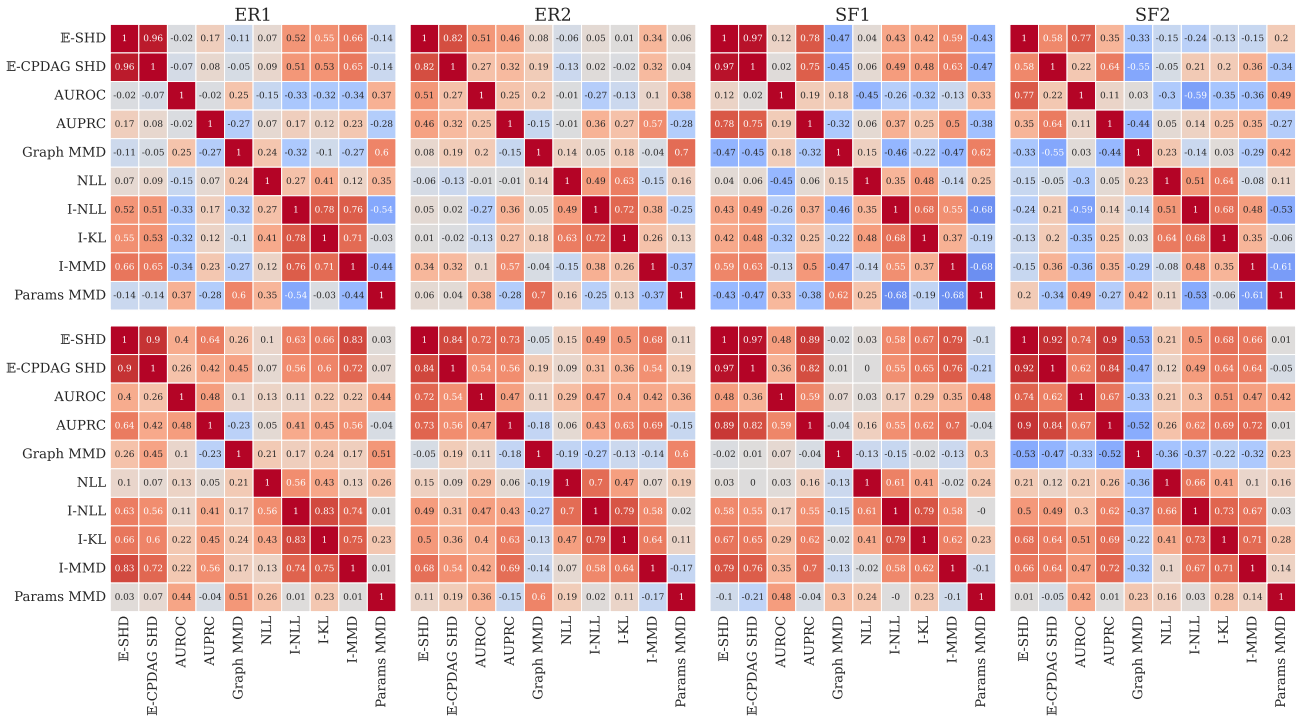

Figure 21. Spearman's rank correlation coefficient between evaluation metrics with 10 training samples ($d = 5$). The first and the second rows correspond to the non-identifiable and identifiable cases, respectively. All the graph-based metrics are still not correlated with the Graph MMD, and Params MMD is still not correlated with any of the other metrics. Graph MMD and Params MMD are metrics that evaluate against the true posterior.

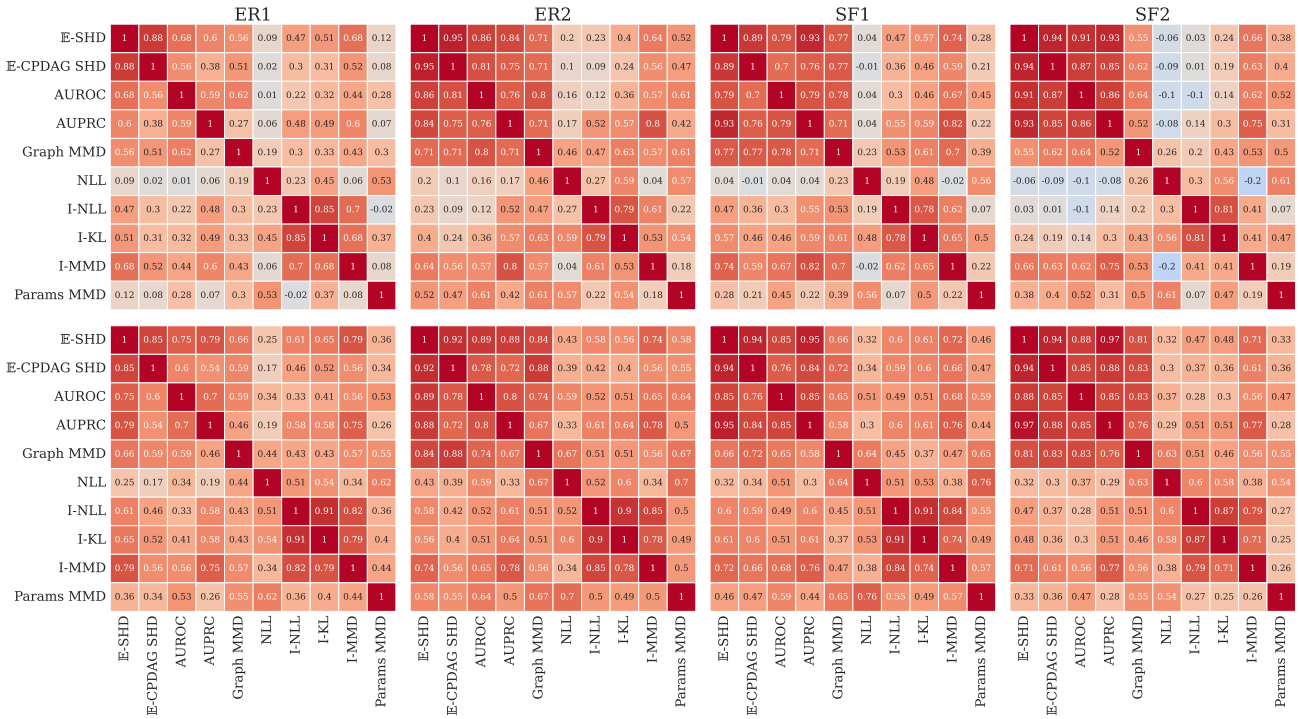

Figure 22. Spearman's rank correlation coefficient between evaluation metrics with 1000 training samples ($d = 5$). The first and the second rows correspond to the non-identifiable and identifiable cases, respectively. All the graph-based metrics are starting to correlate with each other and also the Graph MMD. Params MMD is also start to correlate with other metrics. Graph MMD and Params MMD are metrics that evaluate against the true posterior.

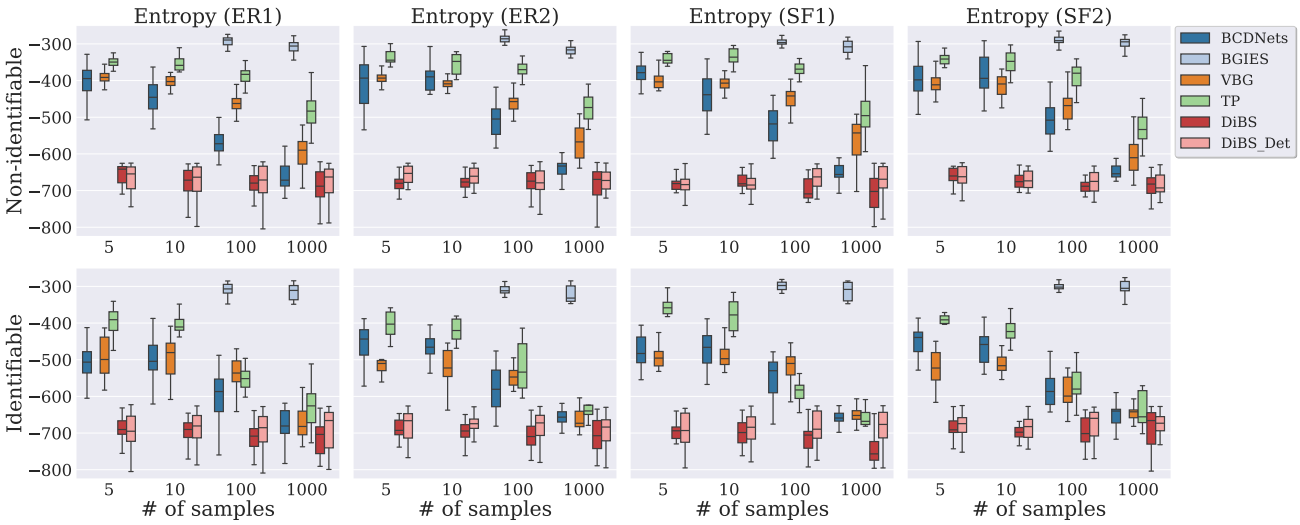

Figure 23. Entropy of the models on different graph types ($d = 5$).

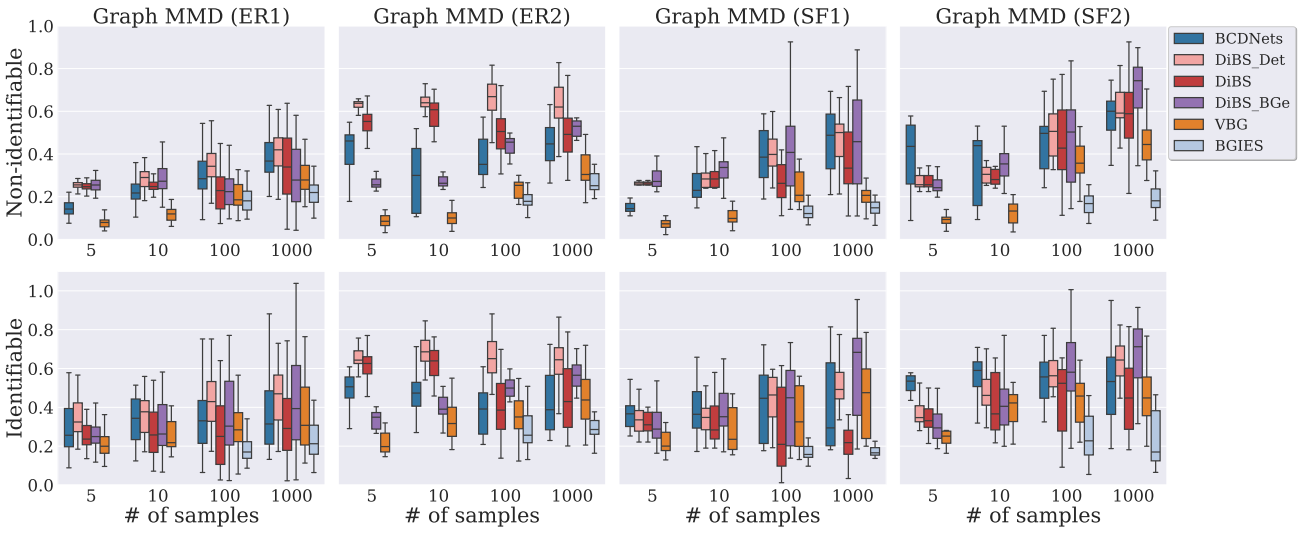

Figure 24. Graph MMD of the models on different graph types ($d = 5$).

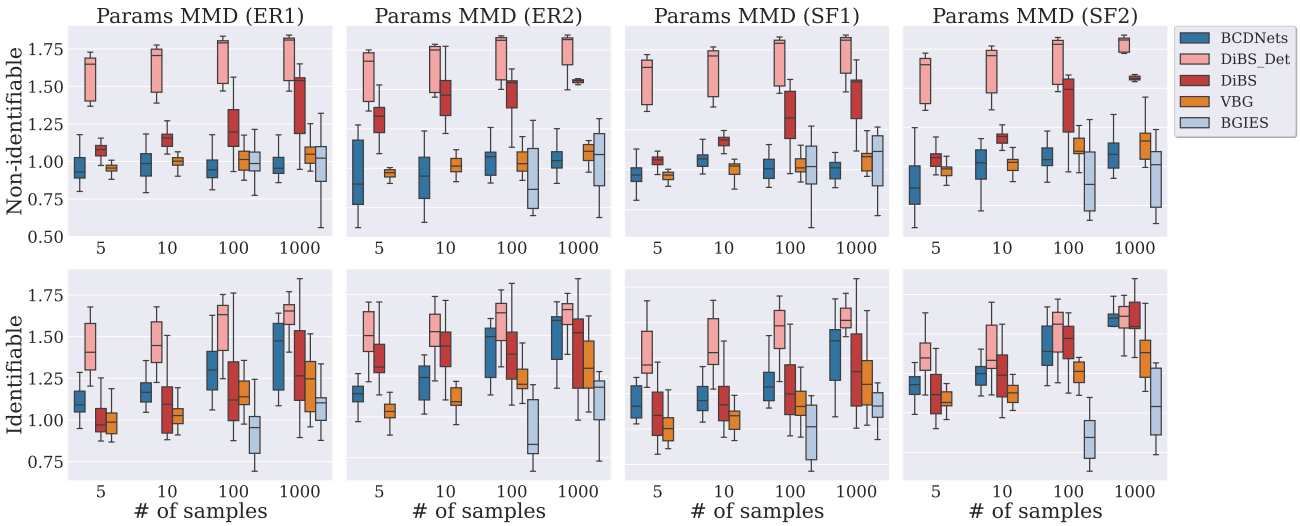

Figure 25. Params MMD of the models on different graph types ($d = 5$).