We thank all reviewers for thoughtful feedback! We reply separately to each reviewer.

Reviewer #1: We would like to point out some of the paper's main contributions, not fully recognized in the review.

- We would like to correct that the paper is *not* devoted to Gaussian BNs: two out of the three main contributions concern any BN models with efficient-to-compute local scores (cf. Questions Q1 and Q2, Sections 3 and 4).
- We actually claim several contributions with both conceptual and technical novelty; we will better highlight these in the next version of the paper.
  - One example is the definition of the maximum coverage problem (two variants), along with the associated (exponential-time) algorithm to solve the problem optimally.
  - Another example is our algorithm for sampling DAGs conditionally on a root-partition (Sections 3.4 and A.3). Not only do we introduce a novel and somewhat generic idea of reusing space by sampling component-wise (parents for fixed node) as opposed to object-wise (whole DAGs), but we also present a technique for sampling weighted subsets of a ground set using inclusion–exclusion. As far as we know, these ideas are novel and may well have applications also beyond the context of DAG sampling.

  Since the established full Bayesian framework gives us a principled machine learning approach, the essential challenges concern the computational tasks. Accordingly, our main innovations are algorithmic.
- We would like to correct that our algorithm for sampling DAGs is *not* "classical" (cf. previous item). There generally are no compelled edges, just the constraint of selecting at least one parent from the previous part.

Reviewer #2: We would like to clarify possible minor misunderstandings and further justify our chosen approach.

- Our approach can handle discrete and continuous data sets, and we use both extensively in our experiments.
- We would like to stress that we are not seeking a single optimal DAG or even MEC. Instead, we take a full Bayesian approach and perform inference based on the posterior distribution over all DAGs (in practice, a sample from the posterior). By averaging over DAGs (belonging to different MECs), we can properly account for the uncertainty related to the graph structure in any subsequent inference tasks. This is not the case when using non-Bayesian structure learning methods, such as PC and GES, which return a single DAG or MEC.
- Furthermore, in practice, the Bayesian approach has been shown to outperform non-Bayesian methods in accuracy in causal inference tasks [29, 1, 19]. We confirm this finding in our experiments on estimating linear causal effects (Figure 3(b–d)), comparing against both PC- and GES-based methods.
- Thus the explicit reasons for our aim here are methodological and related to superior accuracy performance. We will better motivate our choice of approach and highlight its advantages in contrast to PC and GES in the next version of the paper. Thank you for pointing out this need.

Reviewer #3: Thank you for your insightful comments and questions (suggesting grounds for higher confidence)!

- In Section A.4 (Supplement), we discuss how allowing parents outside the candidate set is implemented in our method, comparing it to the approach of Kuipers et al. [17].
- We will add mixing plots illustrating the convergence of the chains to the next version.
- The number K is only relevant to Gadget and BiDAG. Both methods determine an appropriate value of K based on the data set and computation time allowed. The improved accuracy of Gadget in comparison to BiDAG in Figure 3(d) is, in part, due to Gadget being able to use a larger K, namely $K = 15$. We will examine this observation further in the next version.
- Our intention was not to explicitly claim our method can handle hundreds of variables, although the theory (bounds in Table 1 linear in n) and simulations suggest so. We will change the wording in Concluding remarks.
- Thanks for pointing out the typos. These will be fixed in the next version of the paper.

Reviewer #4: Thank you for appreciating our ambitious Bayesian approach.

- The moves and the associated proposal probabilities are indeed described in detail in the papers presenting partition MCMC [16] and BiDAG [17].
- We did not analyze how much the performance of the sampler could be improved by tuning the selection probabilities among the moves, as our choices (equal probabilities) provided good performance already. This question warrants more careful analysis in future work.
- Thanks for the pointer on the naming conflict! We will find a new name for the next version of the paper.