We thank the reviewers for their time and detailed reviews. We are happy to integrate their comments in our revision.

R4 How is background knowledge about the intervention assignment incorporated into the likelihood function? In Sections 3.1 & 3.2, the *only* information about interventions that we make use of is the data itself and the set of targeted variables I_k. The conditionals of targeted variables, parameterized by a separate $\phi^{(k)}$, are learned from scratch. Exploiting domain expertise about the nature of interventions is an interesting direction left as future work.

R4 Invariance in score - "L157 [...] clarify how this is true, or why it is important." Inspection of score (8) and model (7) should make clear that learned conditionals are assumed to be invariant across interventional distributions *in which they are not targeted*. Thus, its maximization will favor graphs for which this invariance holds in the data, which is a central property of causal graphs (see definition of intervention (2) and Peters et al., 2017 Sec 2.1).

R2 Proof sketch and assumptions explanations. This would strengthen the paper and we will add it to the main text.

R4 Perfect interventions. To obtain the score for such interventions, the conditional densities of targeted nodes are removed from the likelihood. This is justified at L171-175, but we will add a formal explanation in the appendix. The essence is that the score (8) separates into two pieces, one of which does not depend on $\mathcal{G}$ and can thus be removed.

R4 Clarification of the $\lambda|\mathcal{G}|$ term. This term is not an acyclicity constraint, it serves to encourage graph sparsity.

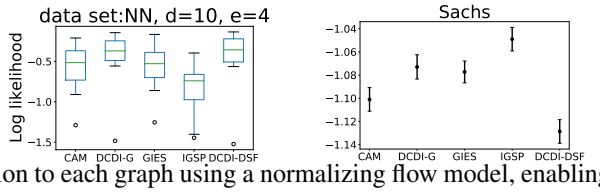
R1 R4 Clarification and justification of the score for unknown targets. In Section 3.3, the phrase "unknown intervention" should be replaced by "unknown targets", which makes more explicit what is actually unknown. In this setting, the interventional targets I_k are unknown and must be learned (except for I_1 which is known to be observational; to be clarified in main text). The approach of Section 3.3 can be seen as a relaxed maximization of the score

$$\mathcal{S}(\mathcal{G},\mathcal{I}) := \max_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X\sim p^{(k)}} \log f^{(k)}(X; M^{\mathcal{G}}, R^{\mathcal{I}}, \phi) - \lambda|\mathcal{G}| - \lambda_R|\mathcal{I}|\,, \tag{1}$$

where $|\mathcal{I}|$ is the total number of targeted nodes in $\mathcal{I}$, $R^{\mathcal{I}} \in \{0,1\}^{K\times d}$ is such that $R^{\mathcal{I}}_{kj} = 1 \iff j \in I_k$ and $f^{(k)}(X; M^{\mathcal{G}}, R^{\mathcal{I}}, \phi)$ is defined in Section 3.3. **If reviewers agree**, we could include the following result to justify (1).

▶ *Extension of Theorem 1 to unknown targets.* Let $\mathcal{I}^*$ be the ground truth intervention family. Under assumptions identical to Theorem 1, we showed that, for $\lambda > 0$ and $\lambda_R > 0$ small enough, if $(\hat{\mathcal{G}}, \hat{\mathcal{I}})$ maximizes score (1), then $\hat{\mathcal{G}} \in \mathcal{I}^*$-MEC$(\mathcal{G}^*)$ and $\hat{\mathcal{I}} = \mathcal{I}^*$. The idea is to add two steps at the beginning of the proof of Theorem 1 showing $\mathcal{I} \neq \mathcal{I}^* \implies \mathcal{S}(\mathcal{G}^*,\mathcal{I}^*) > \mathcal{S}(\mathcal{G},\mathcal{I})$. The argument is similar to Case 5 & 6, except we need to make sure λ and λ_R are small enough (via an argument similar to Case 1). Then, we can resume to the proof of Theorem 1 assuming $\mathcal{I} = \mathcal{I}^*$.

R4 Evaluation on unseen interventional distributions. Such evaluations are uncommon in the closely related causal discovery literature, since some algorithms do not even model distributions. However, we agree that this has scientific value and thus added such an experiment.



Hence, we learn a graph using each method and fit a distribution to each graph using a normalizing flow model, enabling a fair comparison. We report the log likelihood evaluated on an *unseen* intervention. For the NN data set, we report boxplots over 10 graphs. DCDI-G and DCDI-DSF have the best performance like for the structural metrics. For Sachs, we report the log-likelihood and its standard deviation (over data samples). The ordering of the methods is different from the structural metrics: IGSP has the best performance followed by DCDI-G. This will be added to the appendix.

R3 Performance of DCDI-G on the ANM data sets. R3 correctly notes that DCDI-G should outperform DCDI-DSF on ANM data since it has the right inductive bias. We believe that this is not always noticeable in our results due to the relatively high sample size ($n = 10^4$). Hence, we performed a small scale experiment on data sets with a smaller sample size ($n = 10^3$) for perfect interventions. The results support our hypothesis, DCDI-G does outperform DCDI-DSF:

(sparse graph) **DCDI-G** SHD: 5.2 ± 2.5, SID: 20.0 ± 20.8 **DCDI-DSF** SHD: 45.2 ± 12.7, SID: 20.8 ± 9.4
(dense graph) **DCDI-G** SHD: 23.6 ± 9.4, SID: 127.2 ± 37.8 **DCDI-DSF** SHD: 37.1 ± 10.7, SID: 125.9 ± 44.9

R3 Linear DCDI as a sanity check. We trained DCDI as a linear Gaussian model (i.e. no hidden layer in NN). As expected, this version of DCDI obtained competitive results for the linear data set, but poorer results on nonlinear data sets, showing the interest of using high capacity models. We will include this in the appendix.

R2 What contributes to the good performance of DCDI in cases with higher number of average edges? While we do not have a definitive explanation for this, it might be that continuous search has an advantage over discrete greedy search in this setting. This trend has also been noted by Zheng et al. (2018, Section 5.2) (DAGs with NOTEARS).

R1 Additional unknown target methods. While we had seen the list of methods compiled by Mooij et al. (2020), many methods did not have an implementation available or addressed a different setting. Recently (May 2020), the code for Joint Causal Inference (JCI) has been released. We plan to a add a comparison to JCI to the camera-ready version.