We thank all three reviewers for their time and feedback. Below we have done our best to respond to the major concerns.

[R1] Score-based learning. We'd like to emphasize that our paper focuses on *score-based learning*, and all the claims made in the paper are specifically in regards to score-based learning. See L70–74, and in particular L73–74 where we acknowledge that papers such as [21] and [23] achieve impressive results. These are newer methods, and do not shed much light on score-based methods, which are very popular in practice and have a long history. Part of the purpose of this paper is to (a) Shed light on the pros and cons of score-based methods vs. newer approaches and (b) To provide a comprehensive, rigorous mathematical framework through which to analyze score-based methods. These contributions have been acknowledged by R2 and R3, and we will be happy to revise the introduction and abstract accordingly to make this distinction more clear.

[R1] Practicality. Please see L33–37, L43–48, and L294–301, where this is discussed explicitly. We have provided 13 total references for both exact and approximate algorithms for solving this problem. For example, this estimator can be computed approximately using very fast coordinate descent methods for problems with 1,000s of nodes (ref [1]). Exact computation, as acknowledged at L45–46, is of course NP-hard. Furthermore, the reduction in sample complexity from $s^4 \log p$ to $s \log p$ is substantial (see L30–34), and raises the important question of whether or not there exists a polynomial-time estimator that can achieve $s \log p$ sample complexity or better. Our work provides important theoretical justification for this inquiry (in addition to the existing body of work cited above).

[R1] Assumptions. Please see the discussion above on score-based learning. Our analysis is the first *score-based method* to avoid faithfulness, and we propose a novel proof to do this. This is important since existing proofs in score-based learning *crucially* and *fundamentally* rely on faithfulness in order to learn a consistent structure (refs [5, 40]; see also L74–88). Regrettably, at L6–7 this is not made clear, and we will be happy to clarify this in the revision.

[R1, R3] Real data example. As pointed out at L43–44, our approach has been extensively studied in applications. Please see refs [1, 26, 51, 54, 68, 76] for real data examples, including a cytometry dataset [1], gene regulatory networks [26], S&P stock data [68], as well as real data examples from the BN repository [1, 51].

[R2] Faithfulness. We appreciate the reviewer's concern regarding our claims regarding faithfulness, and will be happy to modify and clarify these claims. It is true that the connection between strong faithfulness and the beta-min condition is not well-understood; see for example Uhler et al. (2013, AOS, p. 25): "*...a thorough analysis of the 'permutation beta-min' condition and a comparison to the strong-faithfulness condition more generally is quite challenging and remains an interesting open problem.*"

[R2] Related work. We will be happy to expand our discussions and comparisons to related work, esp. recent work such as [21, 23, 52, 53, 73, 67], which provides an interesting contrast to the score-based method presented in our work.

[R3] Min-trace DAGs. We thank the reviewer for raising the important practical question of how to interpret min-trace DAGs and what their relevance is in practice. As mentioned at L43–44, the estimator $\widehat{B}$ is very popular and used frequently in practice (e.g. refs [1, 26, 51, 54, 68, 76]), despite our having little theoretical understanding of this. In fact, prior to our work, it was an open question what the behavior of $\widehat{B}$ is. For example, does it converge, and if so, to what? We note that even the former question is surprisingly tricky; please see the discussion at L211–216. One of the main contributions of this work is to address these questions: Yes, $\widehat{B}$ converges, and we can in fact pinpoint what it converges to. As this reviewer acknowledges, proving this is nontrivial and required developing novel analytical techniques that can be applied more broadly. The importance of this result lies not in the fact that we might be interested in min-trace DAGs, but perhaps that we might not be! This is the reason why we mention "caution" at L50. Whether or not one would be interested in a min-trace (or equivariance) DAG depends on the application.

We regret that this important discussion was omitted, and will be happy to make these issues more explicit, both in the introduction and in the conclusion, in the final version. We will also add some intuition on what min-trace means in practice, namely the DAG that minimizes the out-of-sample predictive loss, which has obvious practical interest.

[R3] Real data example. Please see related discussion for R1 above.

[R3] Gap condition. Since the problem considered here is at least as hard as p separate regression problems, requiring $\text{gap}(\Sigma) = o(1)$ amounts to a very restrictive condition, and should not be expected in this setting. A similar argument would apply to the ℓ_2-estimation error as measured in Frobenius norm: We do not and should not expect $\|\widehat{B} - \widetilde{B}_{\min}\|_F$ to vanish asymptotically—instead, we expect this to scale appropriately with p. A more reasonable ask is that the "average" gap per node vanishes, i.e. $\text{gap}(\Sigma) = o(p)$, which can hold in high-dimensional setting where $p = O(e^n)$, as discussed at L204–206. In order to obtain strictly weaker results (e.g. **not** including structure consistency for $p > n$), a gap of the same order was also assumed in [35] (Lemma 19) and [62] (Condition 5.1).

Of course, as the reviewer notes, under very strong conditions, these quantities may vanish in the high-dimensional regime $p \to \infty$, and we are happy to add this discussion in the final version in order to clarify this point.