We thank the reviewers for their valuable time and comments. All suggestions will be incorporated in the next revision. Before addressing each of the reviewers' individual comments as best as we can, we want to stress the fact that we do provide extensive and comprehensive experimental results, designed to validate the theoretical results, rather than achieving SOTA performance. We do not deny the importance of pushing the performance of any system to its limit, but we preferred to focus on: (i) proving that identifiability can be achieved in practice by our model; (ii) provide evidence that identifiability can improve certain downstream tasks like the transfer learning and semi-supervised learning examples; (iii) provide evidence on the flexibility of our model, and its capacity to solve both nonlinear ICA and IMCA problems. We also want to bring attention to the fact that identifiability is important in its own right. Whilst improving performance of downstream tasks is indeed an important application of deep models, it would also be hugely beneficial to be able to effectively employ such models to perform principled statistical inference. This latter goal cannot be achieved unless we first achieve basic theoretical results such as identifiability.

Reviewer 1: We thank the reviewer for their time, and for providing a complementary view on the topic. *Definition of identifiability*: As pointed out by the reviewer, we use the definition of identifiability usually found in statistics. Rightly, in the context of probability densities modeled by neural networks, this can be seen as a study of degeneracy of the networks. To avoid confusion, we will add a discussion of how our definition fits within the spectrum of definitions encountered in other fields. *Strong identifiability*: We will clarify that a case where strong identifiability is crucial is causal discovery (Monti et al UAI2019): the linear indeterminacy of weak identifiability will change the causal ordering, making this task impossible. *Semi-supervised learning experiments*: The classification was performed using a logistic regression. We will add this to the manuscript. The reviewer correctly notes that there is a wide range of representation learning algorithms that could have been used in the semi-supervised learning experiments. However, the purpose of this section was to highlight the benefits of identifiability. If we compared the ICE-BeeM approach to such alternative methods we would not know if the performance difference is due to identifiability or something else.

Reviewer 2: We thank the reviewer for their time and suggestions. *Robustness of identifiability*: We find this suggestion by the reviewer fascinating. As mentioned by the reviewer, work in the literature focuses on exact identifiability. Questions like asymptotic variance, performance bounds, and robustness are natural and important directions to follow in the future, to strengthen pre-existing and novel identifiability results. We note that almost no such analysis has been done in the nonlinear ICA literature so far (we are only aware of Sasaki et al, UAI2020). This question certainly merits to be treated as a separate project. Nevertheless, we will add such discussion to the manuscript. *Relation to previous work*: The reviewer is correct about how we relate to previous work. Our conditions are purely "functional" — no assumptions on how the feature distribution are made, contrary to previous work. Our results also extend to the overcomplete case, which has never been covered before. We will add more detail on what was done in previous work to better situate our important contributions. *Strength of conditions*: Condition 2 of Theorem 1 can be replaced by differentiability and "full rankness" of the Jacobian of $\mathbf{g}_\theta$ in just a single point, but this requires the conditioning variable to be continuous. Similarly, Condition 1 of Theorem 1 can be relaxed, and requires the "full rankness" of the Jacobian of $\mathbf{f}_\theta$ to hold only in one point, as we mention in Appendix C.2. In fact, This condition can be scrapped altogether if we relax the definition of the equivalence class in Appendix C.1 to have no conditions on the ranks of matrices $\mathbf{A}_1$ and $\mathbf{A}_2$. This however comes at the expense of a relatively weak, and potentially meaningless, equivalence class. In practice, random initialization of floating point parameters, which are then optimized with stochastic updates (SGD), will result in weights that are almost certainly full rank. We can also encourage this behaviour by adding spectral normalization to the network. *Intuition behind the MLP conditions*: The conditions we present for the MLP example are necessary to satisfy the assumptions of Theorems 1 and 2. They are also necessary to ensure that the learnt representations are not degenerate, since we lose information with low rank matrices. A more detailed discussion of these conditions will be added to the manuscript. The goal behind this example was to translate the functional assumptions of Theorems 1 and 2 into architectural assumptions, and bridge the gap between theoretical models and practical implementations. However, as pointed out by the reviewer, this first iteration might seem artificial, and we intend on improving the assumptions in future work.

Reviewer 3: We thank the reviewer for their thorough read. *On the intuition behind the MCC*: The MCC metric between two representations A and B computes the maximum linear correlations up to any permutation of components. The permutation invariance is required as (similar to linear ICA) we do not have any guarantees on the order of components. We will add more details and some examples to section A.2 to make it easier to grasp. *On the intuition behind fitting ICE-BeeM to IMCA*: The output nonlinearities $\mathbf{H}_l$ play the role of sufficient statistics to the learnt representation $\mathbf{f}_\theta(\mathbf{x})$. Their counterpart in equation (71) is the vector-valued sufficient statistics $\mathbf{T}_i$. We use this trick to ensure that the dot products in equations (71) and (74) happen in the same space, so that we can make conclusions involving square matrices. We agree that this trick is not well explained in the manuscript, and we will amend that.

Reviewer 4: We thank the reviewer for their time and valuable comments. As pointed out above, we decided to dedicate the present manuscript to the theoretical study and a basic empirical validation of the theorems as well as the utility of identifiability; exploring applications of identifiable models in greater depth is an important topic for future work.