We would like to thank all reviewers for their insightful comments and suggestions, and provide replies below.

Reviewer 1

Settings and normalization. The tensors T in equations (2)-(4) may not be normalized. However, in this work we always consider models for probability mass functions of the form T/Z_T, where Z_T is the normalization factor (which can be computed efficiently). Additionally, the A's in the definitions are arbitrary tensors containing the free parameters of the model. They do not have to originate from quantum circuits, but for any quantum circuit one can define A's such that it is equivalent to a BM. To enhance the clarity of the manuscript, we will add a paragraph at the beginning of Section 2 expanding on the above explanations and detailing the settings and requirements on the tensor networks.

Link between Figure 3 and Proposition 1. We will add the following explanation of the relationship between the expressive power, the ranks, Figure 3 and Proposition 1: *For a given rank, there is a set of non-negative tensors that can be exactly represented by a given tensor network, and as the rank increases, this set grows. These sets are represented in Fig. 3 for the case in which the ranks of the different tensor networks are equal. When one set is included in another, it means that for every non-negative tensor, the rank of one of the tensor-network factorizations is always greater than or equal to the rank of the other factorization. The inclusion relationships between these sets can therefore be characterized in terms of inequalities between the ranks, as detailed in Proposition 1.*

Intuition behind propositions. Given the above explanation, Proposition 3 can be intuitively understood to ask by how much one needs to increase the rank of a tensor network such that the set of non-negative tensors it can represent includes the set corresponding to another tensor network. We will expand on this in the text and include intuition for the remaining propositions. In particular, the separations between $\text{MPS}_{\mathbb{R}_{\geq 0}}$ and BM arise from the difference in ranks between probability distributions and square-roots of probability distributions, and the separation between real or complex BM comes from the combination of real and imaginary parts through the modulus squared (see also the reply to Reviewer 2). The growth rate of these separations is lower-bounded in Propositions 4-7. An upper bound is not available, as there could be other distributions providing larger separations than the ones we have found.

Reviewer 2

On the role of complex numbers. We agree that the separation between real and complex BM comes from the use of the modulus of complex numbers in these specific networks. As pointed out, "using real BMs outputting vectors [...] would result in the same benefits": This is correct and precisely why a real LPS of purification dimension 2 includes a complex BM. This fact (shown in Table 2) will be highlighted in the text to provide a correct interpretation for this result. We acknowledge that an expressivity advantage due to complex numbers cannot be extrapolated to general cases such as neural networks. This limitation will be included in the paper.

Relationship with Sum-Product Networks and Arithmetic Circuits. We would like to thank the reviewer for providing these references and we will include a paragraph on these relationships and previously obtained results.

Numerical experiments. We agree that the numerics do not demonstrate the practical advantage of LPS in real-world problems, but rather provide evidence that the theoretical results hold also for distributions that have not been fine-tuned. We aim to investigate their performance on real-world datasets in future work, which might require further research, for example on the use of these tensor networks with continuous variables. In order to provide a comparison, we will add an indication on Fig. 5 of the accuracy of the optimal Bayesian network without hidden variables, where the network graph is learned from the data. This includes simple autoregressive models and avoids hyper-parameter and architecture tuning. It reaches a negative log-likelihood of 5.8, 13.4, 10.4, 9.9, 8.7 and 6.0 on datasets (a)-(f) respectively.

Reviewer 3

Generalization performance. We agree that generalization performance is a very important topic, and that relating generalization, either heuristically or analytically, to quantities such as the rank of these models would be highly desirable. As our analysis is focused on expressive power, evaluation of the models on training sets is useful for validating our theoretical results in practical settings. In order to investigate generalization performance we will add a plot in the supplementary material showing the test set accuracy of these models with respect to the rank. On the biofam dataset the lowest negative log-likelihood on the test set attained by an LPS is 6.4, while for an HMM it is 7.4.

Usefulness for machine learning algorithms. The tensor networks we consider are a class of probabilistic models which admit efficient learning, inference and sampling algorithms, and can therefore be used for the same ML tasks as HMMs while having some expressivity advantages. Indeed, it remains unclear whether this method can lead to state-of-the-art performances, but our theoretical results show that this is worth investigating in the future. Non-negative tensor factorizations are also used in diverse areas of ML such as recommendation systems or signal processing, and the factorizations we introduce may be useful in this context. Moreover, our results and techniques can be straightforwardly generalized to other tensor networks and interpreted as a general comparison between different strategies for ensuring non-negativity of a tensor factorization. We will add a paragraph expanding upon this in the paper.