We thank the reviewers for their helpful comments and suggestions. Our responses below address the main suggestions for improvement proposed by the reviewers, in particular: clarification of claims, choice of the baseline to compare against (GRISE), and additional experiments that show that our methods work for randomly-generated instances.

Reviewer #1:

Comparison with other methods: We chose to use comparison with GRISE only as a baseline because this algorithm is currently the state-of-the-art for learning for learning undirected graphical models (both theoretically from the sample-complexity point of view [19] and empirically [11]), beating other approaches. Also, while we show in this paper that a particular form of GRISE makes it possible to generalize it to include NN representation of the energy function, it is currently unknown whether for other methods this can be done. This broadens the scope of our contribution. **Broader impact**: We agree with reviewer's suggestion on the broader statement, we will improve it for the camera ready version.

Reviewer # 2:

Non-convexity: Although NN-GRISE is non-convex, we show that the true solution will be at the *global optima* of the loss function if the neural net can represent the true model and if there are enough samples [lines 140-142]. This ensures that the true solution can be obtained with stochastic gradient descent type methods if the neural net is large enough, as shown for many classes of neural network problems at a series of works at previous NeurIPS conferences. We agree that this is an important point and we will highlight it in the final version. We find from our experiments that size of neural nets needed to get to this limit is favorable compared to expanding in the monomial basis for models with higher order interactions and symmetries. **Structure identification claims:** Let us clarify the justification of our claim: We show that if we work with a representative enough neural net, then global optimum is guaranteed to encode the neighborhood information of the model, *even for higher order models*. From the conditional independence property, the output of the NN should not depend on the non-neighbors, and the input regularization biases the training towards the minimum with this structure. This principle is clearly demonstrated in the experiments. Following the reviewer's suggestion, we provide a summary figure with additional experiments for structure learning for pairwise and higher-order models (Fig. 1), where for each point we consider 20 randomized problems with more variables compared to the manuscript.

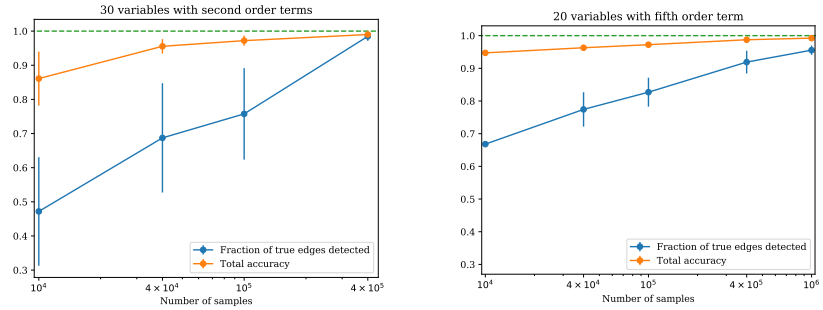

Figure 1: Accuracy of structure learning for models with pairwise couplings, 30 variables (Left), and higher-order interactions, 20 variables (model described in the supplementary section) (Right). "Total accuracy" accounts for both edges and non-edges.

Additional empirical evaluation: Our goal in the paper was to show representative examples that illustrate each of the claims on a concrete example. We agree that providing evidence that the algorithm works for general graphs will strengthen the presentation, which we do with additional experimental results (Fig. 1). For the regularization parameter we used the one for the GRISE algorithm [18,19], as the constant's scaling comes with strong theoretical arguments. Selection of the threshold in practice can be a delicate problem related to the gap to the weakest coupling in the graph, see [11] for a thorough discussion of the setup. Importantly, experiments in this paper show an emergence of such a gap. The threshold for detecting the edges can be chosen either by inspecting the histogram of trained weights or using a standard-deviation based outlier detection method (used in the experiments above). **Discussion of related work:** Our work focuses on undirected graphical models, and not DAGs. Instead of using NNs to directly learn the conditional probabilities, our method uncovers parsimonious basis representations for an unknown class of models. We will make sure to discuss related approaches for continuous variables and their differences with ours in the revised version.

Reviewer #3: Additional experiments: We focus on providing intuitive examples illustrating each of the key points. We agree that a stronger experimental evidence will improve the perception of the narrative. Following the reviewer's recommendation, we provide additional results over the ensemble of networks (see Fig. 1).

Reviewer #4: Presentation: We agree that the presentation of GRISE and problem setting could be done even more accessible for a general reader. We will work on this and discuss lifted inference (including the reference suggested by the reviewer) in the camera-ready version. **Additional experiments:** Applications involving multi-body interactions are numerous, e.g. in natural systems; we use an example of a real problem in quantum computing for an illustration, with a primary goal to showcase our method in a controlled synthetic setting with ground truth. Following the reviewer's recommendation, we produced additional aggregated plots that include confidence intervals and show that our method works on randomized graph instances (see Fig. 1). In the camera-ready version of the paper, we will also clarify that no seed use is needed, all results are obtained through a single run of our algorithm with zero couplings initialization.