We thank all four reviewers for their time and feedback. We will implement major changes, including redoing the simulations with a new evaluation metric: **[R2] Not beating Ind Mult (and [R1] methodology).** This was a major concern for us too. The issue was the specific evaluation metric we used following Inouye et. al. [12]. Their idea was to aggregate all the pairwise MMD (i.e., MMD between bi-variate marginals) to form a histogram. The issue is that pairwise dependence might not be strong for many pairs in real data. However, the overall distribution is still far from a product of multinomials due to higher-order dependencies. Ideally, one should compute MMD between full joint distributions of the "learned model" and the "test data". This however is a single number and we liked the idea of aggregation which gives a more robust metric. It occurred to us that we can retain this nice feature, while measuring higher-order dependencies much better, by looking at all the $\binom{d}{d-2}$ marginals instead of $\binom{d}{2}$. That is, assuming $d = 10$, we plan to measure the maximum discrepancy of the two distributions on moments of the form $\mathbb{E}f(X_{i_1}, X_{i_2}, \ldots, X_{i_8})$ for all f in the unit ball of the RKHS and all $\{i_1, \ldots, i_8\} \subset \{1, \ldots, 10\}$. We refer to this as *pair-complement* MMD. This is in contrast to our current approach of looking only at moments of the form $\mathbb{E}f(X_{i_1}, X_{i_2})$. If the RKHS is rich enough, functions of the form $f(X_{i_1}, X_{i_2}, \ldots, X_{i_8})$ already include those of the form $f(X_{i_1}, X_{i_2})$ by being constant in the extra arguments. Below we have provided some figures with this new metric which clearly shows that POIS and bootstrap (as well as Copula Mult) significantly beat Ind Mult. **[R2] Copula Mult.** Thanks. Indeed, this is a natural choice and will be added to the simulations. As the sample figures show, it generally performs very well, and POIS is quite competitive with it. **[R2] Poisson models.** We agree, though they are not totally unreasonable when λ is small, since Poisson concentrates ($\approx \lambda + O(\sqrt{\lambda})$). We will elaborate more in the paper. **[R2] Inverse Spearman corr.** We understand your concern and try to find a better way to compare (including with the graphical structure of Copula Mult.) **[R2] Why Coord. Dec. beats Cond. Like.?** Not clear if this is the case, esp. in light of the new metric.

[R1,R2,R4] Inference results rather than MMD; qualitative/intuitive comparisons; interpretation. It is difficult to evaluate unsupervised approaches on inference results because of the lack of a ground truth, and since different models estimate different parameters. MMD provides an objective measure of how close the learned model is to the empirical distribution of the test data, and can be uniformly applied to all methods. That being said, we plan to provide more detailed qualitative comparisons based on the estimated correlations, and whether the results agree with domain knowledge. In fact, our original motivation for writing this paper was the interesting correlations predicted by POIS in the toxicity data, which intuitively made sense based on co-occurrence of symptoms. We will elaborate more in the revision. We will also add some simulated data and compare with the ground truth. **[R1] Scalability beyond $d = 50$.** We plan to investigate scalability more thoroughly in the revision. **[R1] Sampling parameter $m = 1000$.** We are sampling from the learned model where there is no limitation. We will clarify more in the paper. **[R1,R2] Notation.** Thanks for your suggestions. We will simplify and clarify the notation and technical arguments. **[R3] Single data set.** We are looking at two different types of toxicity data sets (PRO and toxicity) as well as multiple rating and count data sets. We will try to expand more. **[R3] Class Correlation.** Our point here is that lumping together the correlations among nonzero levels/classes is a good approximation in some applications. It may not be as severe as it may seem. We agree that one can add more complexity (at the expense of interpretation and possible over-fitting). **[R4]** PIM = Probabilistic Index Model; other suggestion? **[R1,R2,R3,R4]** We will incorporate as much of the other suggestions as possible.