We thank all four reviewers for the thoughtful comments. First, we want to point out that the novelty of the proposed **connectivity persistence** is the first metric to look at interaction strength from the topological perspective. Second, unlike existing methods, the proposed **PID** is the first interaction detection algorithm to conduct both global (4.1, 4.2) and local (4.3) interaction detection without the need to train additional interpretable model [2]. We also proved PID is insensitive to weight perturbation in models and verified its superior performance in detecting complex interactions (4.1) and in engineering features which greatly boosted the performance of models in real-world tasks (4.2). We believe this is an important contribution for bringing topological properties and interpretability to interaction detection. Below, we address the reviewers' comments individually:

Discussion about NID and PID (R1: Q1.1). To extract interactions, PID considers connectivity of the entire NN. In contrast, NID leverages weights beyond the first hidden layer to obtain the maximum gradient magnitude of the hidden units in the first hidden layer, loosing some information encoded in latter layers in the process. Hence, the similar results of NID and PID are likely because the latter layers played lesser roles in this specific setting. However, we remark PID constantly outperformed NID with various settings, as shown in Appendix E.3, Figure 8, 9, and 10.

Discussion about AG (R1: Q1.2). We remark that the results of AG is adapted from NID [7], which attributes AG's performance difference in F_5, F_6, F_8 to "limitations on the model capacity of AG, which is tree-based".

Image Experiment (R1: Q2). CNNs are indeed trained to classify images. As it is standard to build saliency maps to evaluate how CNNs make decisions, we aggregate interaction strengths of interacting pixels detected by PID to get the importance of each pixel on the image (line 321). We remark that a key difference between interaction detection and explainable CV (e.g., GradCAM) is that the latter does not consider interactions between pixels because it does not have access to Hessian matrix. In contrast to PID, explainable CV cannot give strength between any group of pixels. For ImageNet, our PID has a $2^{224\times224} \approx 10^{1021609}$ search space (the search space for MNIST is 10^{236}), which is intractable. To illustrate the search space's magnitude, the search space of AlphaGO is 10^{360} [6].

Tasks other than classifications (R1: Q3) We will include the discussion in the revised manuscript.

Inadequate broader impact (R2). The main application of global interaction detection is knowledge discovery. Therefore, PID can help us discover the combined effects of drugs on human body. For example, by utilizing PID on patients' records, we might find using Phenelzine togther with Fluoxetine has a strong interaction effect towards serotonin syndrome. Thus, PID has great potential in helping the development of new therapies for saving lives.

PID assumes well-aligned data & nonlinear operations block paths (R3: Q1, Q2). We remark PID is agnostic to input alignment. See 4.1 and 4.2 for global interaction detection on tabular (well-aligned) data and 4.3 for local interaction detection on image (not well-aligned) data. Appendix D addresses how to adapt PID for local interaction detection by incorporating nonlinear operation (ReLU).

Mainly considers edges with large weights (R3: Q3). We remark that PID extracts interaction based on persistence, not large weight (Section 3). In addition, [1] can only capture "pairwise interaction effects", not all interactions.

MLP cannot fit some complex function (R3: Q4). According to the Universal Approximation Theorem, MLP (with ReLU) can fit any continuous function. Appendix E.1 shows $\exp(\cdot)$ is considered in F_3, F_4, F_5, F_6, and trigonometric functions are considered in F_6, F_8, F_{10}. Appendix E.3 shows the test error of trained MLPs is very low.

Limited improvement with real world data (R3: Q5). For tabular data, we remark an improvement around 0.001 in AUC on these datasets is considered SOTA [3]. In addition, 4.2 shows interactions detected by PID are useful to real-world tasks. Since human-found interactions have been setting the standard in the industry, showing PID finds interactions matching those found by humans is meaningful.

Introduction for image task (R3: Q6). For introduction to the image task, please see R1: Q2. We also remark it is conventional to evaluate interaction detection task on image data qualitatively [2].

Difficult notations (R3: Q7) We apologize and will modify our mathematical notations in the revised manuscript.

Persistent homology (R4: Q1). We are aware that some nice properties have been lost, such as Excision Theorem does not hold. However, as NNs contain only 1-simplex, many of these properties degenerate to the field of graph theory and become easier to evaluate whether they are useful for interaction detection. Due to the page limit, here we only list our high-level idea. From the graph theory perspective, the proposed filtration process is equivalent to building maximum spanning trees (MSTs) of NNs using Kruskal algorithm. The proposed persistence of feature groups is the gap length between MSTs of two sub-networks. There are many papers discussing about the relationship between MSTs and persistent homology, and we could easily extend their results [4, 5]. We will add it in the revised manuscript and discuss about the limitations accordingly. By extending the Barcode from persistent homology and $\langle \phi = \lambda \rangle$-connection from size theory, we derived a topology-motivated algorithm to efficiently detect interaction (Lemma 1) with stability guarantee (Theorem 1).

Unclear whether there is high persistent feature sets in one network, but not the other (R4: Q2). The proof that this situation only happens if the perturbation magnitude δ is greater than a threshold relating to persistence will be added to the revised manuscript. We remark that the stability theorem in the paper is customized for Algorithm 1, which is not comparable to the stability theorem of topological features in persistent homology. Also, we clarify that the results in Appendix C, Table 3 actually took this situation into account. Namely, if an interaction $\mathcal{I}$ is only detected in f but not in g, we treat $\rho_g(\mathcal{I}) = 0$ to get perturbation results.

Unique contribution of interactions detected by PID (R4: Q3). We will add comparisons in the revised manuscript.

[1] Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. In*arXiv preprint arXiv:1802.03888*.

[2] Tsang, M., Cheng, D., Liu, H., ... & Liu, Y. (2020). Feature Interaction Interpretability: A Case for Explaining Ad-Recommendation Systems via Neural Interaction Detection. In *ICLR* 2020

[3] Khurana, U., Samulowitz, H., & Turaga, D. (2017). Feature engineering for predictive modeling using reinforcement learning. In *arXiv preprint arXiv:1709.07150.*

[4] Steele, J. M. (1988). Growth rates of Euclidean minimal spanning trees with power weighted edges. In *The Annals of Probability, 1767-1787.*

[5] Robins, V., ...& Bradley, E. (2000). Computational topology at multiple resolutions: foundations and applications to fractals and dynamics (*Doctoral dissertation, University of Colorado*).

[6] Silver, D., Huang, A., Maddison, C. J., Guez, ...& Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. In *nature*, 529(7587), 484-489.).

[7] Tsang, M., Cheng, D., & Liu, Y. (2017). Detecting statistical interactions from neural network weights. *arXiv preprint arXiv:1705.04977.*