We thank the reviewers for their time and valuable feedback. Overall, we are glad that the reviewers found OGB to be important to advance the field of graph ML. Below, we clarify a number of important points raised by the reviewers: **(1) originality of datasets**, **(2) datasets selection criteria**, **(3) core contribution**, **(4) reproducibility**, and **(5) others**.

(1) Originality of datasets. R4 and R5 raised a critical concern that many of our datasets are simply reproduction of existing ones. This is not true for two reasons: First, nearly all (except `ogbn-products` and `ogbg-mol*`) datasets are in fact constructed by ourselves from public raw data, together with domain experts. Thus, the majority of our OGB datasets (12 out of 15) are completely new and original (note that `ogbg-ppa` and `ogbn-proteins` are also original and different from the existing PPI graph benchmark. Details in L643–L653 in Appendix). Second, although the graphs of `ogbn-products` and `ogbg-mol*` are indeed defined by existing works (which we will emphasize more in our final version), we identified and resolved some important problems around data splitting. Specifically, for `ogbn-products`, existing work [15] did not define a validation set and used a random split with the large training proportion (90%), yielding almost no generalization gap (L167–L169). Even for MOLECULENET, we noticed a serious problem that the scaffold split is not standardized (*e.g.*, "scaffold split" used in [92] is different from [36,40] as well as recent arXiv:2007.02835), because how different scaffolds are put into different splits is quite arbitrary (L689–L693 in Appendix). We think resolving these issues and standardizing the evaluation procedure is important. We will clearly mention this contribution in our final version.

(2) Datasets selection criteria. R4 questioned whether we have clear criteria to select the 15 datasets, similarly to those in OpenML-CC18. Indeed, we do have such criteria: we ensure the diversity of task categories, scales (as defined in L63–L69), and domains, as illustrated in Figure 1 of the paper. Consequently, our 15 datasets are indeed diverse, as shown in the green cells in the right table, resulting in diversity in graph structure (Section 2). OGB is an on-going community-driven effort, and we are constructing additional datasets to fill in the grey blanks so that OGB datasets are maximally diverse.

Node property prediction (`ogbn-`)

Domain	Nature	Society	Information
Small		`arxiv`	
Medium	`proteins`	`products`	`mag`
Large		`papers100M`	

Link property prediction (`ogbl-`)

Domain	Nature	Society	Information
Small	`ddi`	`collab`	`biokg`
Medium	`ppa`	`citation`	`wikikg`
Large			

Graph property prediction (`ogbg-`)

Domain	Nature	Society	Information
Small	`molhiv`		
Medium	`molpcba / ppa`		`code`
Large			

We would also like to clarify fundamental differences between OGB and OpenML-CC18. First, OGB actually *constructs* most of the datasets from scratch, while the OpenML selected existing datasets. Second, OGB extensively evaluates and tests the datasets to make sure they are appropriate, well-behaved and robust. Third, OGB deliberately provides a small selected set of challenging large-scale datasets (around 5 for each task category), as opposed to OpenML-CC18 that provides many small-scale datasets (72 datasets with 500 to 100K samples). The benefit of the former benchmark (OGB) is to allow the community to really focus on challenging problems and easily compare different models on a few benchmarks (similar to ImageNet, CIFAR10, and SQuAD), which is why we did not include many existing small-scale datasets, such as Cora and CiteSeer (extensive discussion on existing datasets is provided in Appendix A, due to space constraint). Our design principle is in contrast to the OpenML-CC18 and the TU datasets [43] (a collection of more than 50 small-scale non-original graph classification datasets); which inevitably causes different works to evaluate models on different subsets of the datasets, making it hard to compare performance across papers.

(3) Core contribution As R4 nicely pointed out, OGB indeed has contributions to many directions, but as our paper title suggests, our main focus is on introducing and defining a set of realistic, large-scale, and challenging datasets and tasks, many of which are our original ones. We also perform extensive baseline experiments and provide easy-to-use code (as done by *e.g.*, MS-COCO and the OpenML), with the goal of analyzing how existing models perform on our newly-introduced tasks and making our new datasets easily accessible to users (in the same spirit as the OpenML).

As suggested by R4, in the final version, we will cite the OpenML and OpenML-CC18 and carefully discuss the above (1)–(3), clarifying our exact contribution.

(4) Reproducibility and code. R4 raised an important concern about reproducibility, which we agree is crucial for OGB. To address this, we have provided to our Area Chair the link to our anonymized Github repository, which contains all of our package scripts and baseline code (note that external links are not allowed to be included in our response here). Regarding the package description, it is provided with example code snippets in Appendices E.1 and E.2, due to space constraints. The data loading and training performance are the same as PyG and DGL, which is highly efficient but the exact number depends on the dataset sizes (*e.g.*, loading a processed medium-scale dataset takes about 5 seconds).

(5.1) Sales ranking split of `ogbn-products`. R4 and R5 raise concerns about the split used in `ogbn-products`. We did try different split ratios, and selected the current one to ensure the training nodes are not too skewed (as pointed out by R4) in the sense that the class balance is almost the same across training/validation/test splits. Also, we think 10% for training is an appropriate number to simulate a realistic transductive semi-supervised learning setting.

Method	Additional Features	Virtual Node	AP (%) Training	AP (%) Validation	AP (%) Test
GCN	✗	✓	36.25±0.71	23.88±0.22	22.91±0.37
	✓	✗	28.04±0.58	20.59±0.33	20.20±0.24
	✓	✓	38.25±0.50	24.95±0.42	24.24±0.34
GIN	✗	✓	45.70±0.61	27.54±0.25	26.61±0.17
	✓	✗	37.05±0.31	23.05±0.27	22.66±0.28
	✓	✓	46.96±0.57	**27.98**±0.25	**27.03**±0.23

(5.2) PRC-AUC of `ogbg-pcba`. We thank R4 for pointing out the issue with PRC-AUC. We now use the suggested Average Precision (AP), and observed the same trend (see Table above). We will update this in the final version.