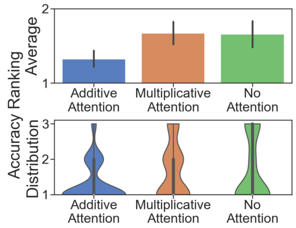
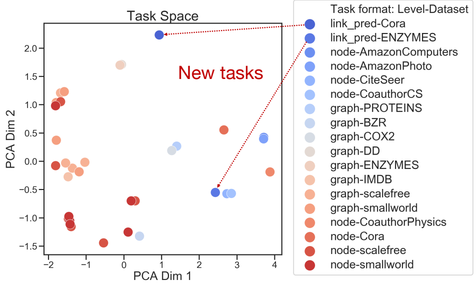
We thank the reviewers for their constructive feedback. All reviewers point out that our paper presents *the first* systematic approach to study GNN designs, *the first* quantitative analysis for GNN task similarity, and offers rigorous findings via novel evaluation techniques. With 1000+ new GNN papers each year, we hope our framework can greatly facilitate the design and evaluation of GNNs. Reviewers ask for clarifications and new experiments, which we answer below:

1 Lack of theoretical analysis (R2 R3) We thank R2 and R3 for raising that our paper lacks theoretical analysis. Indeed, our paper focuses on empirical understandings of GNN design: the novelty of our systematic framework and valuable findings are acknowledged by all the reviewers. Here we emphasize that our framework provides a solid tool that can *verify and inspire theoretical findings*. For instance, the GIN paper shows the nice theoretical result that SUM aggregation is more expressive than MEAN and MAX; however, their evaluation can be improved, *e.g.*, only compare on a fixed GNN design (5-layer, 64-dim, etc.) on a few graph classification tasks. In contrast, our framework samples hundreds of models from 10M possible model-task combinations, *with every design dimensions controlled except the aggregation function*, which is the first comprehensive and rigorous evaluation that verifies SUM is indeed empirically successful (Fig 3). Similarly, our framework provides *rigorous evidence* to other theoretical results *in the context of GNN*, *e.g.*, BN helps neural network training, skip connections avoid the problem of vanishing gradients. More interestingly, our paper makes the novel discovery that PRELU activation significantly improves GNN performance. We think this finding suggests the uniqueness of GNN optimization landscape, and hope it can inspire theoretical works towards the open question of improving GNN optimization. We will add these new discussions to the revised paper.

2 Additional design dimensions (R1 R2 R4). We thank reviewers for suggesting other design dimensions to explore. We defined a general design space including intra-layer design, inter-layer design and learning configurations; however, we were not able to cover all aspects, and especially thank R4's appreciation for our efforts. We wish to present a systematic framework which can inspire researchers to propose and understand new design dimensions – reviewers' constructive suggestions in fact illustrate the importance of such a framework. Based on these suggestions, we run new experiments. **New results for attention (R1 R2 R4).** We compare GNNs without attention, using additive attention or multiplicative attention using the same approach that we produce Fig 3. The results show that using additive attention is favorable than multiplicative attention and no attention. This is consistent with the choice of GAT where additive attention is used. **New results for link prediction (R4).** Following R4's suggestion, we additionally include link prediction tasks on Cora and ENZYMES to the task space. The best architecture we found for Cora is "(1, 8, 3, skipsum, mean, 400)", for ENZYMES is "(1, 6, 2, skipcat, max, 400)" (*c.f.*, Fig 1(c) in our paper). Interestingly, by visualizing the task embeddings via the proposed task similarity metric, we find link prediction on Cora is different from other tasks, while link prediction on ENZYMES is similar to some node classification tasks. We will include these new results in the revised version.



New design dimension



3 More comparisons 1) With standard architectures (R1). We thank R1 for asking the comparison with standard GNN architectures. We emphasize that the goal of our paper is not pursuing STOA performance, but presenting a systematic approach for GNN design. In fact, our systematic approach can be used to determine the hyperparameters of existing architectures. Following R1's suggestion, we implement standard GCNs with message passing layers $\{2, 4, 6, 8\}$, while *keeping all the other optimal hyper-parameters we discovered* in line 268. The best model in our design space is better than the best GCN model in 24 out of 32 tasks. Note that we defined a simple GNN design space. Our new results show that adding attention further improve the performance. We will include these new results. **2) With NAS approaches (R4).** Our framework is orthogonal to NAS approach: we focus on designing and evaluating a search space, while NAS approaches focus on finding the best model from a given search space. Unfortunately, applying Auto-GNN on the large ogbg-molhiv dataset requires training 2000+ models which is beyond our computing resources.

4 Related work (**R2**). We thank R2 for pointing out other powerful GNNs and will cite them in the revised version.

5 Clarifications. **Q(R2):** "Issue of multiple hypothesis testing" **A:** We thank R2 for pointing out the issue. We resample experiments for each design dimension in Fig 3 so this is less of concern. Nevertheless, we run one-way ANOVA with Bonferroni correction (p-value 0.05). 8 (without correction) and 7 (after correction) out of the 12 design dimensions have significant findings. **Q(R2):** "Use v's own embedding in message passing" **A:** The SKIP-SUM design choice that we use is equivalent to what R2 suggests. **Q(R3):** "Experiment-driven task similarity" **A:** We agree with R3 that our approach can be improved; however, how to define the "real" similarity between tasks is still an open question. We are the first who introduce the notion of task similarity to the GNN community, and we provide strong evidence that the proposed task similarity is useful (Fig 5). **Q(R2 R4):** "In the range of common GNNs" **A:** R2 and R4 are correct that the models we consider are common GNNs, thus will fail in expressiveness tasks. Designing more powerful GNNs is still an open domain which cannot be summarized into a design space yet, therefore we do not include in our paper.