We sincerely thank the reviewers for their insightful and constructive comments. First, we address the common questions of **R1,R2,R3:** *Definition of GCN:* We define GCN in eq (4), which not only includes Kipf & Welling [16] $D^{-1/2}AD^{-1/2}$, but also $D^{-1}A$ (e.g. [12] GraphSAGE mean aggregator) among others. Other papers (e.g. JK paper by Xu et al, 2018, p3) also consider these as variants of GCN. We will make this explicit in the final version. *Contribution:* the main contribution of this work is to develop deep theoretical understanding of GCNs, which would inform efficient GCN architectures design. We do not intend to achieve the state-of-the-art graph classification model. We will tone down the emphasis on the model architecture design, as similar work may already exist. *Real-world Experiment:* To verify theoretical results, we needed the ground truth model, which is unknown for real data. We are happy to add real-world experiments in the final version. Next, we address the specific concerns raised by each reviewer below.

R1 *graph attention network (GAT)* multi head attention in GAT concatenated K independent attention mechanism with the *same* propagation rule (PR) while ours utilizes modules with *different* PRs. We will consider incorporating attention mechanism as a future work. *jumping knowledge network* Both JK-network and ours use residuals, but ours uses residual connections for GCN while JK-network is designed for Message Passing Graph Neural Networks. *graph classification setup* We used mean-pooling to aggregate node-level representations, after which a single number is passed to a classification layer. We will clarify this and include discussions with GAT and JK in the final version.

R2 *authors try to use GNN for graph generation* There seem to be some misunderstandings about the paper. We do *not* learn graph generation. Instead, we use GCN to learn graph moments and to classify graphs. *... $D^{-1}A$ has limitations in learning degree of a graph.* We are *not* claiming the limitation of a particular GCN variant (Fig 2 is only an example). Our key message is that unlike fully connected neural networks, GCNs are not universal approximators. Therefore, choosing the right PR (for example, A vs $D^{-1}A$) is crucial in the GCN's ability to learn graph moments. We provide theoretical analysis and offer a solution that can alleviate this issue. Note that $D^{-1}A$ is *not* our definition of GCN (see eq (4)), nor is it used in our theoretical analysis. *... why $D^{-1}A$ is associated with node permutation invariant.* This is a misunderstanding. We are *not* claiming that $D^{-1}A$ is related to node permutation invariance. In fact, any GCN in the form of eq. (4) $F(A,h) = \sigma(f(A)\cdot h\cdot W + b)$ is permutation invariant, regardless of the function $f(A)$ (see sec 2.1 and 2.3). The purpose of Proposition 1 and Theorem 2 is to argue that GCN can be restrictive due to permutation invariance, thus having the right PR, activation and number of layers is crucial in its ability to learn graph moments.

did not see these papers used these three different graph operations It is not explicit but easy to derive. [12] GraphSAGE with MEAN aggregator, averages over $h_i + \sum_{j\in\mathcal{N}_i} h_j$, which is equivalent to the operator $\tilde{D}^{-1}\tilde{A}$ where $\tilde{A} = A + I_N$ is an adjacency with self-loops. Kipf-Welling GCN uses $\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ ([16] eqs (7),(8)). [18] uses the graph module A ([18] eq (2)) *simple variant of original GNN* Our theoretical analysis demonstrates the importance of PR (e.g. using A, GCN cannot learn $\sum_j (D^{-1}A)^p_{ij}$ or vice versa), as well as to have sufficient number of layers with residual connections. Hence, having a modular design with sufficiently many layers and residual connections would be useful for learning graph moments. We have no intention to claim the novelty of our proposed GCN, but only to validate our theoretical findings. In fact, we pointed out the similarities our model with GIN in lines 173-176. *graph generation baselines* As we are *not* generating graphs in this work, our method is unrelated to graph generation baselines.

R3 *better not to claim...expressiveness of GCNs as the first contribution.* While it may be known in the literature, we are not aware of any rigorous theoretical analysis for the exact same setting. We will add more references and tone down the claim in the final version. *bridge the theoretical analysis with the proposed design.* Our theoretical analysis shows the limitations of having a single PR in learning graph moments. It also points out the importance of having sufficient number of layers and residual connections. Our modular design combines these results and arrive at an architecture with multiple PR modules, layers and residual connections. We will improve the writing and include more descriptions in the final version. *compare with some existing GCN designs* Our GCN already includes several existing GCN designs, which we refer as different PRs. We are happy to include other GCN designs for comparison as well. *Could not find the code...* We apologize for the confusion. We will release the code in the updated version.

R4 *lack of related work concerning graph moments.* Graph moments, or "Graph Power" is a concept from graph theory (see the Lin and Skiena (1995)) and has been used extensively in network science. Graph moments encodes topological information of a graph and is closely related to graph coloring and Hamiltonicity. We will include references from graph theory and network science. Note that we are not aware of any other work that learns graph moments using GCNs. *Notation: lack of explicit introduction of notational convenience* Thank you, will fix. *Line 112* yes, for brevity M_p is $M_p(A)$. *an index such as in $f(A)_i$.* Yes, $f(A)_i$ is the ith component. *The moments are matrix valued?* No, our definition eq. (1) is vector valued, (summed over one node index). *residual connections are very strongly stated...* We agree with the reviewer regarding the positioning of our contribution. We will tone down the emphasis on modular design. The restrictions found in our theory show a need for including multiple propagation rules, hence the module.