We thank all the reviewers for their valuable and positive feedback. The specific questions of each reviewer are addressed below. All minor comments (like typos and notations) pointed out by reviewers will be addressed.

Response to Reviewer #1:

There are error bars in one plot, but no mention I saw of what they indicate. Sorry for the missing specifications. The error bars in Figure 2 correspond to the variance from the 5-fold cross-validation results.

Response to Reviewer #2:

Since the embeddings are created by predicting the attributes of other nodes in a path ... We have two levels of embedding: vertex and graph levels. The vertex embedding employs the CBoW-like pipeline, which predicts the attributes of nodes from their neighborhoods in the graph (not just the other nodes in a path). The graph embedding will then assemble the node embeddings along each path. This is not an end-to-end model, and these two embedding stages are separate from each other.

... it is not very clear whether these methods are broadly applicable (apart from the molecule domain) or if there are any conditions under which they may not work well. First of all, the molecule property prediction/drug discovery is an important application domain. It is a perfect scenario for graph-level prediction tasks with comparatively rich data samples, and a lot of work has been exploring this field. Thus we believe it is significant to study in this field, which is also supported by the other reviewers. Second, there are various challenging tasks and datasets in this domain, which requires thorough studies. Therefore, we have explicitly focused on molecules (e.g., pointed out in the title) and conducted extensive analysis and experiments. Finally, though the other domains are not the focus of this paper, in principle, our method can be applied there. A thorough study will require another set of extensive analysis and experiments, which we leave as future work, as pointed out in the Conclusion section.

The baselines also look weak. The baselines here includes both the classic machine learning and the state-of-the-art graph neural network methods. As also confirmed by Reviewer #3, we conducted rich experiments on the benchmark datasets (including 60 tasks), and the baselines methods are the state-of-the-art ones. They are robust as experimentally tested in the recent few years.

About appendix. After double-checking, we do have the appendix attached in the supplementary material (in "N_Gram_Graph_Paper.pdf"). Reviewer #3's comments can also help verify this.

Response to Reviewer #3:

About the semi-supervised setting. We agree that this is a very good point. Here are several reasons that we did not test the semi-supervised learning in this paper. 1) We would like to provide a thorough study in the basic setting of supervised learning, which already requires extensive experiments. 2) All the baselines are designed in the supervised setting, and to show that N-Gram Graph is a good representation, we want to first compare in the same setting. 3) In the semi-supervised learning setting, usually the number of labelled data is much smaller than the total data size. This is rarely seen in the virtual screening scenario unless we are introducing some extra data points from different datasets (or different data distribution). We explored this a little bit in Table 3. So to sum up, applying N-Gram Graph in the semi-supervised learning setting is very natural and can further exploit the power of our method. We leave this as an exciting and promising future direction.

Why N-Gram Graph is better than other graph neural networks. 1) As pointed out by our theoretical analysis, the N-gram graph embedding has very strong representation and prediction power. It preserves the count statistics of the graph. Intuitively, such statistics are important for predicting the molecule properties. 2) N-Gram Graph utilizes some very simple operations to get the embeddings, yet all of them are by design ordering-invariant (both in the vertex-level and graph-level). This inductive bias is important for graph level prediction tasks and leads to advantages over the other general GNNs. 3) The other GNNs require highly non-trivial optimization which can prevent fully exploit their power, while our method is much simpler. 4) Our method is unsupervised so the representations can be used by different learning models, including those that are great in extracting useful information for the tasks, such as XGB.