This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

nature portfolio

Peer Review File

A Folding–Docking–Affinity framework for protein–ligand binding affinity prediction

Corresponding Author: Professor Degui Zhi

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

In recent years, there has been a burst in research exploring structure-free modeling for predicting drug-target interactions, in order to overcome the scarcity of 3D data necessary for model training. Proposed methodology takes advantege of the advancements in the field, such as AlphaFold, and proposes a three-step methodology to tackle this challenge: (1) predicting protein folding, (2) docking the folded protein with a compound, and (3) predicting affinity from the docked structure. While the potential outlined in the manuscript is exciting for the cheminformatics community, the experimental demonstration of the methodology falls short in clarity.

- My major concern is that Davis dataset is quite small to illustrate the performance of the proposed model over others. Additionally, Davis is essentially a Kinase dataset - therefore benchmarking on an additional dataset with wider range of proteins would be very informative.

- Section 2.4. is confusing - the PDB dataset is mentioned to be used in training and validation. Is this also the case for docking-free models?

- Please provide similarity distributions of proteins/ligands in PDB/Davis datasets used as in training and test in Supplementary.

- Section 2.2. needs clarity - for instance, in the first step of protein folding generation, the authors mention that five structures are generated for each protein and only the top ranked are selected. Is the ranking threshold same for every protein? Is the selection criteria have an effect on generated docking poses? e.g. if top-3 ranked structures were to be selected how would the quality of the binding poses change? Similarly, in the second step, for each structure it is mentioned that ten binding poses are generated and affinity is based on top ranked generated poses. Please detail and discuss the ranking and whether these have any downstream impact on affinity prediction task.

- I think one interesting use of generating multiple structures/poses would be in data augmentation. Did authors experiment with this while training with PDB dataset?

- Please provide more details about the dataset size used in ablation study in Section 3.2. Figure 3 is interesting in terms of suggesting that the docking based pre-training strategies might be very helpful. However as the values are quite close to each other (both in RMSE and pearson), I wonder whether this is a statiscally significant indication.

- Please report time - resource dependent complexity of the proposed methodology in comparison to others.

- Please update the Fig 1 such that methods and selection strategies are also included and figure alone is self-explanatory of the methodology.

- Please include include Fig S3 to main manuscript to replace Fig2.

Reviewer #2

(Remarks to the Author)
In this work authors introduce a binding affinity prediction procedure that does not require 3D information about protein and ligand as input and combines protein folding and docking methods. Authors claim that they are the first to construct a general

protein-ligand affinity prediction framework based on a computed explicit binding pose. Besides, authors demonstrate the feasibility of such methods by showing superior performance in a rigorous benchmark experiment in DAVIS.

1. The idea of stacking folding and docking methods for affinity prediction is relevant for the community, but in my opinion saying that the current work proposes a "novel method" is a bit of a stretch. That is said, I believe that the current work can be a good basis for benchmarking more different combinations of existing folding (AlphaFold, ESMFold, RosettaFold, etc), docking (DiffDock, Vina, GLIDE, TANKBind, etc) and structure-based affinity prediction methods resulting in a comprehensive study of this question. In my opinion, presenting these findings in a form of a comparative study will be very helpful and relevant for the community. The current state of the work is however premature for the publication, in my opinion.

2. Authors should take into account that ColabFold and DiffDock training sets can contain data points used in validation of FDA. Then the claim that FDA doesn't require structure or pose data is not properly validated. In my opinion, a comprehensive explanation and analysis of this is necessary given the provided claims.

Minor comments:

Note that DGraphDTA uses PconsC4 to obtain protein contact map. Probably worth mentioning in the text. It also seems that MGraphDTA takes a protein structure as input.

Page 3 line 77: docking -> folding

Page 5 line 186: did you retrain on PDBbind? What splits? How do other methods perform on PDBbind?

Page 6: "ColabFold-DiffDock combination generally achieved better performance." – what does it mean? According to the Table 1 it is not true.

Page 6 lines 206-222: results with RMSD depending of the data availability seem to be expected and do not explain the fact that "ColabFold-DiffDock combination generally achieved better performance."

Page 7 lines 236-239: are these modifications taken into account in the current work?

Page 7 line 252: how can phosphorylation info be taken into account by FDA if you start with the amino acid sequence?

Reviewer #3

(Remarks to the Author)
The paper addresses the question of whether protein folding and docking outcomes predicted using state-of-the-art machine learning-based models can improve the prediction of protein-ligand binding affinity. A framework, FDA, is presented in which machine learning-based systems are used for folding prediction (F) and docking prediction (D), which in turn are used as input to a machine learning-based model for binding affinity prediction (A).

The proposed framework uses available state-of-the-art systems for each subtask. Compared to docking-free binding affinity prediction models using only protein sequences and SMILES strings or molecular graphs of ligands, improved performance on the DAVIS benchmark dataset is reported.

The paper does not propose new methods for predicting protein-ligand binding affinity. However, given the recent improvements in deep learning-based protein folding and docking methods, a current and relevant research question is addressed. Since experimentally identified 3D structures are not available for most molecules, docking-free models have been developed in the literature so far. Now that deep learning-based advanced 3D structure prediction models are available, it is important to show the utility of these models for predicting protein-ligand binding affinity compared to classical docking-free models.
However, I have the following concerns.

The results reported in the paper are based only on one data set, the DAVIS dataset, which is a kinase dataset. The framework should also be evaluated with other datasets (e.g. KIBA, PDBBing, etc.) to demonstrate the generalizability of the findings.

The proposed framework is presented as a generic framework, which is one of its strengths over KDBNet, a model specific to kinases. However, the experimental results in this paper are based solely on the DAVIS dataset, which is a kinase dataset. The results in the supplementary file show that KDBNet outperforms the system developed in this paper. Therefore, the advantage of the proposed system compared to KDBNet is not clear. It is not clear to what extent KDBNet is specific to kinases and how well the proposed system can be generalized to other types of protein targets. Experiments with more general, non-kinase-specific data sets may provide insight into the generalizability of the proposed framework. The current results do not show this aspect of the developed system.

The results in Fig. 2 show that the docking-free model MGraphDTA performs very similarly to the proposed framework (FDA), except for the Pearson values for the "both new" case. Thus, it seems that the benefit of using (predicted) folding and docking is not very obvious, especially considering that MGraphDTA (and the other docking-free models) is a much simpler and less computationally intensive model compared to the proposed framework.

The benefits of using predicted 3D information should also be demonstrated using other datasets.

Furthermore, the limitations of the proposed framework in terms of computational complexity, especially compared to docking-free models should be discussed. Are there other limitations? For example, it was noted that proteins longer than 3000 amino acids are filtered out due to computational cost. All these limitations should be clearly discussed.

A similar argument applies to the ablation study. The training and test datasets used are very small to reach reliable conclusions. Since the proposed framework does not require known structures, a much larger test dataset with predicted folding and docking can be created. Furthermore, it would be interesting to investigate the case where the training dataset is enlarged by including protein-ligand pairs with predicted folding and docking on top of the experimentally determined ones:

(i) Training set: Experimental
(ii) Training set: Predicted
(iii) Training set: Experimental + Predicted

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
After carefully investigating,
1. Figure 2 with KDBNet included as a benchmark method and KIBA as the additional (Kinase) dataset,
2. Supp. Tables 4-5 reporting resource-time complexity of FDA in comparison to simpler methods,

It's still difficult to observe the benefits of the proposed methodology (FDA) over other simpler and low complexity methods. Currently, the authors claim that their primary objective is to evaluate whether docking structures are essential for accurate binding affinity prediction. This claim is not completely supported by the results in Fig 2, and needs more validation over a diverse dataset (non-Kinase).

I still believe the presented study would be valuable resource for the community but I think the authors could think about shifting the focus of their manuscript. To incorporate one/some of the future works they mention, e.g. either constructing a comprehensive evaluation dataset or investigating the potential of augmentation in training outlined by folding and docking are very interesting and valuable directions that would benefit the community.

Reviewer #2

(Remarks to the Author)
I thank the authors for addressing my comments and providing additional info.
I would like to summarise my final evaluation in four points with short comments.

Importance: high.
I believe that studying the role of the current state-of-the-art protein structure prediction and docking methods in the context of binding affinity prediction problem is very important and relevant for the community.

Novelty: low.
In my opinion, the contribution of this work can be formulated as “benchmarking GIGN on synthetic data”. As I wrote in my initial review, I don’t believe that the claim that this work introduces a novel method is fair.

Methodology: low.
According to the Table S3, 87% and 91% of proteins from DAVIS and KIBA test sets respectively were used for ColabFold training. It means that experimental results do not actually show the ability of FDA to predict affinity for complexes with unknown structural data. This methodological flaw undermines the main idea of the paper. In particular, one of the main questions if "approach enables reliable predictions of binding affinities of protein-ligand pairs without crystallized binding structures" remains untested.

Performance: low
According to the Figure 2, FDA performs on par with other methods like MGraphDTA (taking into account error bars). At the same time, FDA is 27000 times slower than MGraphDTA, and requires 200 times more memory.

Reviewer #3

(Remarks to the Author)
I would like to thank the authors for their revisions. Some of my comments have been addressed, particularly the inclusion of the KIBA dataset in the experiments and the newly added Limitations section, which enhance the overall quality of the paper. These changes support some of the claims and conclusions, making both the contributions and limitations clearer.

However, I still have concerns regarding the question posed in the title: “Is 3D binding pose needed?” What is the answer of the paper to this question? This is a general and important question, while the datasets used are kinase-specific. Furthermore, the results do not convincingly demonstrate that the proposed framework is superior to state-of-the-art docking-free models. For instance, in Figure 2, KDBNet shows better performance. I understand that it uses additional kinase-specific binding pocket information, however, the performance of MGraphDTA and DGraphDTA is also quite similar to that of the proposed framework. Therefore, I believe the claim of "superior performance" is not sufficiently justified by the experiments and results presented, and this assertion appears to be an overstatement.

I believe the paper addresses a relevant and timely research question. I recognize the challenges of compiling a more comprehensive and less biased dataset, as well as conducting more extensive experiments. Nevertheless, I recommend revising the drawn conclusions to align more closely with the results obtained and to avoid overstatements. Additionally, the title should either be revised or the paper should provide a clear answer to the question based on the results.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I appreciate the authors' efforts in addressing the comments. While I understand their intention to incorporate the following points in future work due to the substantial effort required, I have to insist on the importance of these aspects. Without them, the novelty and comprehensiveness of the work does not reach to its full potential.

"I still believe the presented study would be a valuable resource for the community but I think the authors could think about shifting the focus of their manuscript. To incorporate one/some of the future works they mention, e.g. either constructing a comprehensive evaluation dataset or investigating the potential of augmentation in training outlined by folding and docking are very interesting and valuable directions that would benefit the community."

Version 3:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I thank the authors for their effort in incorporating a new section on binding pose augmentation to investigate its impact on predictive performance. This addition enhances the study, providing valuable insights that could inform future research.

Reviewer #1 :

1. In recent years, there has been a burst in research exploring structure-free modeling for predicting drug-target interactions, in order to overcome the scarcity of 3D data necessary for model training. Proposed methodology takes advantege of the advancements in the field, such as AlphaFold, and proposes a three-step methodology to tackle this challenge: (1) predicting protein folding, (2) docking the folded protein with a compound, and (3) predicting affinity from the docked structure. While the potential outlined in the manuscript is exciting for the cheminformatics community, the experimental demonstration of the methodology falls short in clarity.

2. My major concern is that Davis dataset is quite small to illustrate the performance of the proposed model over others. Additionally, Davis is essentially a Kinase dataset - therefore benchmarking on an additional dataset with a wider range of proteins would be very informative.

   We appreciate the reviewer's concern regarding the size and scope of the DAVIS dataset. While it is true that the DAVIS dataset is not large, it has been extensively utilized as a benchmark in numerous protein-ligand affinity prediction studies, demonstrating its relevance and utility in this field.

   To address the concerns about the dataset's size, we have conducted additional experiments using the KIBA dataset, which encompasses a broader range of ligands and features a larger data size. Our results indicate that our method outperforms SOTA docking-free methods on both-new and new-drug split within the KIBA dataset. Detailed results of these experiments can be found in **Section 3.1**.

3. Section 2.4. is confusing - the PDB dataset is mentioned to be used in training and validation. Is this also the case for docking-free models?

   Yes, the splitting of the PDBBind dataset is indeed applied to the docking-free models as well. For the sake of clarity and completeness, we now explicitly state this in **Section 2.4 (line 143-144)**.

4. Please provide similarity distributions of proteins/ligands in PDB/Davis datasets used as in training and test in Supplementary.

   We acknowledge the importance of this similarity analysis. We now mention the result of this analysis in **Section 2.4 (line 148-152)**. In addition, the similarity distributions of proteins and ligands in the PDBbind and Davis-53 datasets are present in **Fig. S3** and **Fig. S4** in the supplementary information.

5. Section 2.2. needs clarity - for instance, in the first step of protein folding generation, the authors mention that five structures are generated for each protein and only the top ranked are selected. Is the ranking threshold same for every protein? Is the selection criteria have an effect on generated docking poses? e.g. if top-3 ranked

structures were to be selected how would the quality of the binding poses change? Similarly, in the second step, for each structure it is mentioned that ten binding poses are generated and affinity is based on top ranked generated poses. Please detail and discuss the ranking and whether these have any downstream impact on affinity prediction task.

In **Section 2.2**, regarding the selection of protein structures, we followed the same ranking strategy that utilized in Alphafold-Multimer. Specifically, we ranked the generated protein structures based on a score, calculated as interface predicted template modeling score (ipTM) weighted at 0.8, and predicted template modeling score (pTM) weighted at 0.2. From these rankings, the top-ranked structure for each protein was selected for subsequent docking steps. This ranking strategy was consistently applied across all proteins, and no fixed threshold score was implemented. we have included the description of the ranking strategy in **Section 2.2 (line 82-87)**.

We acknowledge that the selection criteria for protein structures can influence the generated docking poses. Similarly, the selection criteria for ligand poses are expected to impact the final binding affinity predictions. Investigating the effects of varying these selection strategies and optimizing these criteria is indeed crucial and deserves a comprehensive study. However, the primary objective of our manuscript is to evaluate whether docking structures are essential for accurate binding affinity prediction. Including such a study would deviate from the main focus of our research. Moreover, given the substantial computational resources and time required for this additional investigation, we have chosen not to incorporate it into our current revision. Nonetheless, we recognize the importance of this aspect and mention the point in **Section 4.4 (line 381-401)**.

6. I think one interesting use of generating multiple structures/poses would be in data augmentation. Did authors experiment with this while training with PDB dataset?

   The generation of multiple structures/poses for data augmentation is indeed an interesting approach. However, in our study, we did not explore this technique while training with the PDB dataset. Our primary focus was on assessing the impact of ColabFold-generated protein structures and DiffDock-generated ligand poses on downstream binding affinity prediction. Future research could investigate the potential benefits of incorporating multiple structures/poses for data augmentation, as this may enhance model performance. We have noted this potential research direction in our conclusion **(line 437-439)**.

7. Please provide more details about the dataset size used in ablation study in Section 3.2. Figure 3 is interesting in terms of suggesting that the docking based pre-training strategies might be very helpful. However as the values are quite close to each other (both in RMSE and pearson), I wonder whether this is a statistically significant indication.

In the ablation study described in **Section 3.2**, we used a dataset comprising 53 protein-ligand pairs to test our models. We acknowledge the importance of assessing the statistical significance of our findings, given the close values in RMSE and Pearson correlation coefficients. To address this, we conducted additional statistical analyses to determine the significance of the observed differences. These results are included in Table. S2 in supplementary information and were mentioned in **Section 3.2 (line 228-231)**.

8. Please report time - resource dependent complexity of the proposed methodology in comparison to others.

   We recognize the necessity of reporting the time and resource-dependent complexity of the proposed methodology in comparison to other methods. In our revised manuscript, we have included a comprehensive analysis detailing the average runtime and GPU memory usage for our approach and docking-free models. This comparison was made against established methodologies to provide a clear understanding of the efficiency and scalability of our proposed method. This information is included in **Table. S4** and **Table. S5** in the Supplementary Information and is discussed in **Section 4.4 (line 402-418)**.

9. Please update the Fig 1 such that methods and selection strategies are also included and figure alone is self-explanatory of the methodology.

   We have updated **Fig. 1** to include the ColabFold, DiffDock, and GIGN, ensuring that the figure alone provides a self-explanatory overview of the methodology.

10. Please include include Fig S3 to main manuscript to replace Fig 2.

    We agree with the suggestion and have updated **Fig. 2** in the revised manuscript to include the performance of KDBNet for a reference comparison.

Reviewer #2 :

1. In this work authors introduce a binding affinity prediction procedure that does not require 3D information about protein and ligand as input and combines protein folding and docking methods. Authors claim that they are the first to construct a general protein-ligand affinity prediction framework based on a computed explicit binding pose. Besides, authors demonstrate the feasibility of such methods by showing superior performance in a rigorous benchmark experiment in DAVIS.

2. The idea of stacking folding and docking methods for affinity prediction is relevant for the community, but in my opinion saying that the current work proposes a "novel method" is a bit of a stretch. That is said, I believe that the current work can be a good basis for benchmarking more different combinations of existing folding (AlphaFold, ESMFold, RosettaFold, etc), docking (DiffDock, Vina, GLIDE, TANKBind, etc) and structure-based affinity prediction methods resulting in a comprehensive

study of this question. In my opinion, presenting these findings in the form of a comparative study will be very helpful and relevant for the community. The current state of the work is however premature for the publication, in my opinion.

We agree with the argument that the individual steps of our method are not novel. However, the integration of folding and docking methods for affinity prediction, while based on established techniques, introduces a unique approach that leverages the strengths of both processes to improve predictive accuracy. This synergy, combined with our specific implementation and validation strategy, represents an innovative contribution to the field.

We acknowledge that a comprehensive benchmarking study involving multiple folding (e.g., AlphaFold, ESMFold, RosettaFold) and docking (e.g., DiffDock, Vina, GLIDE, TANKBind) methods could provide valuable insights. However, the primary objective of our current work is to introduce and validate our specific method, demonstrating its potential and establishing a foundation for future comparative studies.

We believe that our current findings offer significant value by presenting an approach to binding affinity prediction that can inspire further research and optimization. While a broader comparative study is indeed important, we think that it falls outside the scope of our current work. We propose that such an extensive benchmarking study could be a follow-up project, building upon the foundation laid by our present research. At the same, we also mention this point in **Section 4.4 (line 381-401)**.

3. Authors should take into account that ColabFold and DiffDock training sets can contain data points used in validation of FDA. Then the claim that FDA doesn't require structure or pose data is not properly validated. In my opinion, a comprehensive explanation and analysis of this is necessary given the provided claims.

   We acknowledge that the training sets of ColabFold and DiffDock might contain data points that overlap with those used in our validation process. To address this, we conducted a comprehensive protein and ligand similarity analysis between these datasets. The result is shown in **Section 4.4 (line 361-380)** and **Table. S3**. This overlap could potentially influence the performance metrics and lead to an overestimation of our method's predictive capabilities. While some levels of protein overlap and ligand overlap exist, the exact protein-ligand pair overlap with ColabFold and DiffDock train sets were minimal: DAVIS has 4 out of 14,464 pairs, and KIBA has 7 out of 89,957 pairs (**Table. S3**). Therefore, we estimate the impact of these overlaps is probably not too substantial. However, a thorough and objective evaluation with datasets that are completely non-overlap with these training sets is difficult, because it requires carefully designed experiments or even new biochemical measurements.

4. Note that DGraphDTA uses PconsC4 to obtain protein contact map. Probably worth mentioning in the text. It also seems that MGraphDTA takes a protein structure as input.

   We acknowledge that DGraphDTA uses PconsC4 to obtain the protein contact map, and this has been broadly mentioned in the manuscript **(line 21)**. Regarding MGraphDTA, we would like to clarify that it does not take the protein structure as input. Instead, MGraphDTA only requires the protein sequence as its input.

5. Page 3 line 77: docking -> folding

   We acknowledge the mistake and have corrected "docking" to "folding" in the revised manuscript.

6. Page 5 line 186: did you retrain on PDBbind? What splits? How do other methods perform on PDBbind?

   Yes, we retrained our three different GIGN models using the PDBbind dataset for three distinct scenarios: Crystal-Crystal, Crystal-DiffDock, and ColabFold-DiffDock. The details of these splits are provided in **Section 2.4 (line 137-144)**. Additionally, we evaluated the performance of MGraphDTA and GraphDTA under the same conditions, and the results are presented in **Table. 1**.

7. Page 6: “ColabFold-DiffDock combination generally achieved better performance.” – what does it mean? According to the Table 1 it is not true.

   We acknowledge that the statement "ColabFold-DiffDock combination generally achieved better performance" may not accurately reflect the intended meaning. Specifically, the data in **Table. 1** does not consistently support this claim. In **Table. 1**, however, exclusively displays the performance of the three distinct prediction models evaluated on their corresponding datasets. We intended to express that the ColabFold-DiffDock model tends to perform better across the three different test datasets, which can be more clearly observed in **Fig. 3**.

   To enhance clarity, we have revised this statement to: "the ColabFold-DiffDock combination model demonstrated superior performance across three distinct test sets," as found on **line 238-239**.

   In addition, to address Reviewer 1's comment #7, we have re-conducted the ablation study to obtain the p-value. Consequently, the data presented in **Table. 1** and **Figure. 3** may exhibit slight differences.

8. Page 6 lines 206-222: results with RMSD depending of the data availability seem to be expected and do not explain the fact that “ColabFold-DiffDock combination generally achieved better performance.”

   The potential explanations for these results are now discussed in **Section 4.2**.

9. Page 7 lines 236-239: are these modifications taken into account in the current work? In **Section 4.1**, we have indicated that considering these modifications is part of our future work.

10. Page 7 line 252: how can phosphorylation info be taken into account by FDA if you start with the amino acid sequence?

    Currently, our FDA framework cannot consider post-translational modifications such as phosphorylation information. To handle these, our plan involves utilizing the AlphaFold 3 model to generate the initial three-dimensional protein structure with phosphorylation. Subsequently, we will employ biomolecular force fields to further optimize the structure and use molecular dynamics techniques to identify the stable conformation. We have mentioned this plan in **Section 4.1 (line 271-276)**

Reviewer #3 :

1. The paper addresses the question of whether protein folding and docking outcomes predicted using state-of-the-art machine learning-based models can improve the prediction of protein-ligand binding affinity. A framework, FDA, is presented in which machine learning-based systems are used for folding prediction (F) and docking prediction (D), which in turn are used as input to a machine learning-based model for binding affinity prediction (A).

2. The proposed framework uses available state-of-the-art systems for each subtask. Compared to docking-free binding affinity prediction models using only protein sequences and SMILES strings or molecular graphs of ligands, improved performance on the DAVIS benchmark dataset is reported.

3. The paper does not propose new methods for predicting protein-ligand binding affinity. However, given the recent improvements in deep learning-based protein folding and docking methods, a current and relevant research question is addressed. Since experimentally identified 3D structures are not available for most molecules, docking-free models have been developed in the literature so far. Now that deep learning-based advanced 3D structure prediction models are available, it is important to show the utility of these models for predicting protein-ligand binding affinity compared to classical docking-free models.However, I have the following concerns.

4. The results reported in the paper are based only on one data set, the DAVIS dataset, which is a kinase dataset. The framework should also be evaluated with other

datasets (e.g. KIBA, PDBBind, etc.) to demonstrate the generalizability of the findings.

As noted in our response to Reviewer #1, the KIBA dataset is also a kinase dataset, and the PDBBind dataset has a significant overlap with the training data of DiffDock. We acknowledge the reviewer's concern about the need to evaluate our framework with diverse datasets to demonstrate the generalizability.

However, it is a common challenge to find a comprehensive dataset with multiple protein types that is free from overlaps with existing training data and suitable for effectively benchmarking our proposed method. We have mentioned this limitation in **Section 4.4 (line 352-360)** and emphasize the need for future work to develop such data sets and validate our framework using a broader range of datasets.

5. The proposed framework is presented as a generic framework, which is one of its strengths over KDBNet, a model specific to kinases. However, the experimental results in this paper are based solely on the DAVIS dataset, which is a kinase dataset. The results in the supplementary file show that KDBNet outperforms the system developed in this paper. Therefore, the advantage of the proposed system compared to KDBNet is not clear. It is not clear to what extent KDBNet is specific to kinases and how well the proposed system can be generalized to other types of protein targets. Experiments with more general, non-kinase-specific data sets may provide insight into the generalizability of the proposed framework. The current results do not show this aspect of the developed system.

   We completely agree with the reviewer's concern that the specificity of KDBNet to kinases and the potential for our system to generalize across diverse protein targets are critical aspects that require further investigation. To articulate these limitations, we acknowledge the need for experiments with more general, non-kinase-specific datasets. However, finding appropriate datasets that are both diverse and free from significant overlaps with existing training data remains a challenge.

   We now emphasized this limitation in **Section 4.4 (line 352-360)** and outlined our plans for future work to include a broader range of datasets to better demonstrate the generalizability of our proposed framework.

6. The results in Fig. 2 show that the docking-free model MGraphDTA performs very similarly to the proposed framework (FDA), except for the Pearson values for the "both new" case. Thus, it seems that the benefit of using (predicted) folding and docking is not very obvious, especially considering that MGraphDTA (and the other docking-free models) is a much simpler and less computationally intensive model compared to the proposed framework.The benefits of using predicted 3D information should also be demonstrated using other datasets.

   We respectively disagree with the assertion that the benefits of using predicted folding and docking are not obvious. The "both new" case, which is the most challenging and important scenario when benchmarking binding affinity models,

demonstrates significant improvement with our proposed method. Achieving enhanced performance in this case represents a substantial advancement.

Furthermore, in our ablation study, we trained both our proposed model and the MGraphDTA model on the PDBBind dataset and tested them on a subset of the DAVIS dataset. Our FDA method outperformed MGraphDTA and GraphDTA, as shown in **Table 1**, indicating better generalizability of our proposed method across distinct datasets.

Additionally, our proposed framework (FDA) offers greater extensibility. Each component of our framework can be replaced with the most advanced methods available, thereby improving its ultimate predictive performance. Lastly, unlike docking-free methods, our framework can explicitly account for chemical modifications such as methylation, phosphorylation, and acetylation, which expands its potential applications.

7. Furthermore, the limitations of the proposed framework in terms of computational complexity, especially compared to docking-free models should be discussed. Are there other limitations? For example, it was noted that proteins longer than 3000 amino acids are filtered out due to computational cost. All these limitations should be clearly discussed.

   We acknowledge that the computational complexity of our framework, especially when compared to docking-free models such as MGraphDTA, is an important consideration. Our framework integrates folding and docking predictions, which inherently increases computational demands. Additionally, as noted, our framework currently filters out proteins longer than 3000 amino acids due to computational constraints. This is a significant limitation, particularly for studies involving large proteins or protein complexes. We have discussed these limitations in **Section 4.4**.

8. A similar argument applies to the ablation study. The training and test datasets used are very small to reach reliable conclusions. Since the proposed framework does not require known structures, a much larger test dataset with predicted folding and docking can be created. Furthermore, it would be interesting to investigate the case where the training dataset is enlarged by including protein-ligand pairs with predicted folding and docking on top of the experimentally determined ones:

   (i) Training set: Experimental
   (ii) Training set: Predicted
   (iii) Training set: Experimental + Predicted

   There are really good points. For the first concern, we acknowledge that the test datasets used in our ablation study are relatively small, which can limit the reliability of the conclusions drawn. However, given the need to curate datasets with crystallized structures that do not overlap with the training data of DiffDock, we believe the current setting represents a reasonable choice. While it is true that our proposed framework does not require known structures for prediction, the

comparison across three distinct scenarios necessitates the use of crystallized structures to ensure consistency and validity.

Regarding the second suggestion, we agree that investigating the impact of training set would be highly valuable. Exploring these scenarios could provide deeper insights into the generalizability and robustness of our framework. We have mentioned it in our conclusion **(line 437-439)** for future research to enhance the comprehensiveness of our study.

Reviewer #1 (Remarks to the Author):

After carefully investigating,
1. Figure 2 with KDBNet included as a benchmark method and KIBA as the additional (Kinase) dataset,
2. Supp. Tables 4-5 reporting resource-time complexity of FDA in comparison to simpler methods,

It's still difficult to observe the benefits of the proposed methodology (FDA) over other simpler and low complexity methods. Currently, the authors claim that their primary objective is to evaluate whether docking structures are essential for accurate binding affinity prediction. This claim is not completely supported by the results in Fig 2, and needs more validation over a diverse dataset (non-Kinase).

We appreciate the reviewer's concerns and acknowledge that the mixed results in Fig 2 do not fully support the argument that protein-ligand binding structures are essential for accurate binding affinity prediction. Therefore, we revised our conclusion to “Our experiments show that the proposed framework generally exhibits performance comparable to SOTA docking-free models. Hence, the docking-based models are potentially promising, but the question of whether directly modeling 3D binding poses is necessary for accurate binding affinity prediction remains unresolved.” in **line 12-15, 69-71, 214, 271-281, 357-358, and 428-431**.

Additionally, we acknowledge our evaluation is limited to the kinase dataset. Although most existing evaluations rely on these datasets, we recognize the importance of validating our framework on datasets that include a broader range of protein types. While curating such diverse datasets presents a significant challenge, we plan to address this limitation in our future work, which has been mentioned in **line 357-365**.

I still believe the presented study would be a valuable resource for the community but I think the authors could think about shifting the focus of their manuscript. To incorporate one/some of the future works they mention, e.g. either constructing a comprehensive evaluation dataset or investigating the potential of augmentation in training outlined by folding and docking are very interesting and valuable directions that would benefit the community.

We agree that the proposed directions, such as constructing a comprehensive evaluation dataset and investigating the role of data augmentation through folding and docking, are very valuable. However, we believe that our current focus is on evaluating the feasibility and benefits of the proposed framework. In this context, our study aims to address an important gap in the current literature by demonstrating the potential of using predicted protein-ligand structures for binding affinity prediction. We expect that this study could serve as a starting point for research into incorporating binding structures to improve the accuracy of binding affinity prediction.

Reviewer #2 (Remarks to the Author):

I thank the authors for addressing my comments and providing additional info.
I would like to summarize my final evaluation in four points with short comments.

Importance: high.
I believe that studying the role of the current state-of-the-art protein structure prediction and docking methods in the context of binding affinity prediction problem is very important and relevant for the community.

Novelty: low.
In my opinion, the contribution of this work can be formulated as “benchmarking GIGN on synthetic data”. As I wrote in my initial review, I don’t believe that the claim that this work introduces a novel method is fair.

We do not agree with the accusation that our work is "benchmarking GIGN on synthetic data". Benchmarking on synthetic data typically means using computer simulations to generate some synthetic ground truth data, and then evaluate methods over the synthetic ground truth. This is different from the way that we evaluate our and others' methods. We and others used the data sets DAVIS and KIBA, which are derived from real experimental measurements. Although we used AlphaFold and DiffDock as intermediate steps of our methods, these are not treated as “ground truth”. What ColabFold and DiffDock does here is to approximate or reconstruct the ground truth given the available data.

In one of our experiments, we evaluated the accuracy of AlphaFold and DiffDock by comparing the ground truth derived from Protein Data Bank (PDB). However, this experiment was only a post-hoc diagnostic to help us understand factors contributing to the final accuracy of our FDA framework. This experiment is to point out ways of potential future improvements of FDA, rather than to evaluate GIGN.

We acknowledge that the individual components of our method are not entirely novel. However, the integration of folding and docking methods for affinity prediction, while based on established techniques, introduces an approach that serves as a starting point for integrating binding structures for more accurate binding affinity prediction. We are probably one of the first who stitch folding, docking, and affinity prediction together into a unified pipeline.

Methodology: low.
According to the Table S3, 87% and 91% of proteins from DAVIS and KIBA test sets respectively were used for ColabFold training. It means that experimental results do not actually show the ability of FDA to predict affinity for complexes with unknown structural data. This methodological flaw undermines the main idea of the paper. In particular, one of the main questions if "approach enables reliable predictions of binding affinities of protein-ligand pairs without crystallized binding structures" remains untested.

All the points raised by the reviewer were already acknowledged in our previous revision, specifically in **line 366-385**. However, the reviewer failed to point out the fact that only 0.027% and 0.008% of the protein-ligand pairs in DAVIS and KIBA, respectively, have experimentally determined binding pose represented in the DiffDock's training set. Since the main goal of our evaluation is to answer the question "Is 3D binding pose needed", the large percentage of proteins represented in the ColabFold training set is not of primary concern.

We admit that the definition of the both-new, new-protein, and sequence-identity categories might be a bit loose for FDA, as the proteins might have been already included in the ColabFold training set. Therefore, the performance of FDA on these categories might be somewhat overestimated. Given that most proteins are already represented in the training set of ColabFold/AlphaFold, and given the hefty cost of training a protein folding model from scratch, it is very challenging for anyone to make a "clean" evaluation to answer this question. All these points are made in our main text (**line 366-385**).

Performance: low
According to the Figure 2, FDA performs on par with other methods like MGraphDTA (taking into account error bars). At the same time, FDA is 27000 times slower than MGraphDTA, and requires 200 times more memory.

We acknowledge that the computational complexity of our proposed approach is considerably higher than that of docking-free methods, while the improvement in prediction performance is not as pronounced. However, the primary focus of this manuscript is to evaluate the feasibility and potential benefits of our framework in bridging the gap between state-of-the-art protein structure prediction and binding structure prediction in the context of binding affinity prediction. We believe that future optimizations of our pipeline could help reduce the inherent computational demands, making the approach more efficient without sacrificing its potential to provide deeper structural insights into protein-ligand interactions, mentioned in **line 407-423**.

Reviewer #3 (Remarks to the Author):

I would like to thank the authors for their revisions. Some of my comments have been addressed, particularly the inclusion of the KIBA dataset in the experiments and the newly added Limitations section, which enhance the overall quality of the paper. These changes support some of the claims and conclusions, making both the contributions and limitations clearer.

However, I still have concerns regarding the question posed in the title: “Is 3D binding pose needed?” What is the answer of the paper to this question? This is a general and important question, while the datasets used are kinase-specific. Furthermore, the results do not convincingly demonstrate that the proposed framework is superior to state-of-the-art docking-free models. For instance, in Figure 2, KDBNet shows better performance. I understand that it uses additional kinase-specific binding pocket information, however, the performance of MGraphDTA and DGraphDTA is also quite similar to that of the proposed framework. Therefore, I believe the claim of "superior performance" is not sufficiently justified by the experiments and results presented, and this assertion appears to be an overstatement.

We acknowledge that the mixed results in Figure 2 do not fully answer the proposed question. Besides, we agree that our claims regarding the performance of the proposed method may have been overstated. Therefore, we revised our conclusion to “Our experiments show that the proposed framework generally exhibits performance comparable to SOTA docking-free models. Hence, the docking-based models are potentially promising, but the question of whether directly modeling 3D binding poses is necessary for accurate binding affinity prediction remains unresolved.” in **line 12-15, 69-71, 214, 271-281, 357-358, and 428-431**.

I believe the paper addresses a relevant and timely research question. I recognize the challenges of compiling a more comprehensive and less biased dataset, as well as conducting more extensive experiments. Nevertheless, I recommend revising the drawn conclusions to align more closely with the results obtained and to avoid overstatements. Additionally, the title should either be revised or the paper should provide a clear answer to the question based on the results.

We agree with the suggestion and have revised the conclusion to avoid overstatements regarding our FDA performance, specifically in **line 12-15, 69-71, 214, 271-281, 357-358, and 428-431**. As for the direct answer to the question in the title, we added our final statement: The docking-based models are potentially promising, but the necessity of directly modeling 3D binding poses for accurate binding affinity prediction remains unresolved.

Reviewer #1 (Remarks to the Author):

I appreciate the authors' efforts in addressing the comments. While I understand their intention to incorporate the following points in future work due to the substantial effort required, I have to insist on the importance of these aspects. Without them, the novelty and comprehensiveness of the work does not reach its full potential.

"I still believe the presented study would be a valuable resource for the community but I think the authors could think about shifting the focus of their manuscript. To incorporate one/some of the future works they mention, e.g. either constructing a comprehensive evaluation dataset or investigating the potential of augmentation in training outlined by folding and docking are very interesting and valuable directions that would benefit the community."

Thanks for the reviewer's valuable feedback on our work. We fully agree with the reviewer's perspective and insistence, which we believe enhances the impact of our manuscript. In response, we have incorporated one aspect of our future work—exploring the potential of binding pose augmentation—into the manuscript. Specifically, we conducted experiments evaluating different binding pose augmentation scenarios on the DAVIS and KIBA datasets across four distinct data-split tasks. Our results demonstrate that augmenting binding poses, generated during the folding and docking process, improves predictive performance. However, we also observe a plateau effect as the number of binding poses increases. These findings have been included in Section 3.3 and Figure 4.