nature portfolio

Peer Review File

Optimizing Drug Design by Merging Generative AI With a Physics-Based Active Learning Framework

Corresponding Author: Professor Victor Guallar

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

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The paper by Filella-Merce reports an interesting generative modeling approach making use of an active learning cycle. The paper prospectively synthesized and tested generated molecules, which is very refreshing to see in this kind of work. I believe the work is publishable in Communications Chemistry after few minor points to address:

- In the Introduction, the authors mentioned the advantages of VAE over GAN. While I agree with their choice and justification, a brief discussion and comparison with other generative frameworks (eg diffusion models) would be much appreciated especially because some of those models are regularly employed in drug discovery
- Active learning in hit finding is an established technique; while I appreciate the generative angle that the authors used (which differs from most active learning works), elaborating on existing applications could help to put the work more in context
- In the ABFE section for CDK2, the authors mentioned that some molecules were predicted actives and some inactives - what does this mean? Did the authors use a specific threshold? Or were inactives molecules displaying a positive ABFE value?
- The authors used BioSolveIT InfiniSee to look for molecules similar to their generated structures both for CDK2 and KRAS. They mentioned that the closest molecule(s) was(were) 0.8 similar to their compounds (or at least this is how I interpreted it); this means that they actually found very close molecules, and I think it is fair to elaborate on this - for example by showing a Tanimoto map or plotting the 2d structures of the closest pairs?

Reviewer #2

(Remarks to the Author)
This paper describes a machine learning approach to generate novel inhibitors of CDK2 and KRAS. The process is divided in two stages: the first filters molecules based on drug-likeness, diversity and synthesizability while the second filters molecules based on docking scores. These stages are performed iteratively to find the best candidates for experimental testing. In the end, authors claim to have found 8 active molecules for CDK2 and 4 molecules predicted as active for KRAS, but these have not been tested.

Major comments:

- The definition of “active” molecules seems a bit awkward to me. The authors label molecules with an IC50 value of less than 100uM as active, which appears quite permissive and perhaps unrealistic. It is true that the threshold between active and inactive can be arbitrary and varies depending on the target, assay, and other factors. Generally, a threshold of 10uM is common for hit identification and it typically reduces to 1uM for well-established targets like kinase inhibitors such as CDK2. A standard lead optimization campaign for kinase inhibitors aims to discover compounds in the low nanomolar range (less than 100nM). Assuming this work would be equivalent to a hit identification stage and use 1uM as the threshold, then only 1 out 9 molecules showed in vitro activity for CDK2.

- It is challenging to discern if the computational methods applied in this paper provide a good estimate for experimental values in the case of KRAS. The authors used a set of known KRAS inhibitors consisting of 73 molecules. There is no

demonstration of whether the docking Glide score or PELE binding free energy correlates with the experimental values of these molecules (as presented for CDK2 in Supplementary Fig2). Even though these methodologies work for CDK2, they are highly dependent on the target. ABFE was evaluated on just 4 compounds, but the results are not presented. Moreover, the authors indicate that there's no correlation between the Glide score and ABFE, which questions the effectiveness of Glide scores in selecting compounds in the outer loop. Without concrete evidence of how accurately ABFE predicts bioactivity in KRAS, I believe it's insufficient to declare the four chosen molecules as active without an experimental value.

Minor comments:

- Docking scores are generally an estimate of how well a molecule fits into the binding pocket, but this doesn't necessarily imply bioactivity. It's not unusual for docking scores to not align with experimental values or with more precise binding prediction methods, as we've seen in the case of KRAS. Authors should be cautious when referring to docking scores as binding energy predictions or target engagement.
- It's suggested to avoid referring to untested experimentally molecules as "active molecules". For instance, in the last sentence of the introduction, "we identified 4 active molecules for KRAS" could be better expressed as "we identified 4 molecules with potential activity against KRAS".
- The authors refer to their method as an "active learning framework". However, in my view, this doesn't seem to be genuine active learning as it doesn't involve experimental labeling of new data at any point. Instead, it resembles the self-distillation approach used in AlphaFold [1] for training with unlabeled data.
- Figure 2B clearly shows that the number of validated molecules decreases with each cycle. This might be due to the constraint of molecular diversity, but it's hard to tell from the current figure. It would be useful to display the change in each property with each cycle.
- Supplementary Figure 2A illustrates the correlation between Glide score and experimental values for the set of known molecules. It indicates a correlation coefficient of -0.56, which could be supplemented with an error measurement like MSE or RMSE. Based on the present data, Glide scores appear to moderately correlate with experimental values and may not be suitable for detecting molecules that fit in the binding site.
- Figure 4C presents the structures of the molecules selected for experimental validation. It could be beneficial to include the structure of the nearest neighbor in the initial set as part of the supplementary material to back up the claim that these differ from the initial set.
- Figure 6B and 6C do not align with the provided legend. They appear to be switched.

1- https://www.nature.com/articles/s41586-021-03819-2

Reviewer #3

(Remarks to the Author)
The manuscript „Optimizing Drug Design by Merging Generative AI With a Physics-Based Active Learning Framework" presents a novel generative AI workflow that combines a variational autoencoder (VAE) with two nested active learning cycles and docking as well as physics-based estimation of binding affinity. The approach is applied to the design of CDK2 inhibitors and kRAS inhibitors. CDK2 inhibitors are experimentally confirmed.
The manuscript is written well and clearly describes the technology. The integration of generative AI and physics-based approaches is a meaningful approach for using structural knowledge. The authors present the successful design of CDK2 inhibitors based on this approach with 8 molecules showing biological activity.
The described approach would be especially helpful in the low-data regime where only little ligand information is available. kRAS inhibitors have been chosen to illustrate the capability of the technology, but binding affinity of suggested molecules have only been validated by ABFE calculations. Therefore, it is crucial, that validation data for the ABFE predictions are included at least in the supplementary data. Correlations show for a subset of training data on CDK2 appear moderate. Also explain how the displayed subset was selected.
For generated molecules, standard analysis should be included such as validity, uniqueness and novelty. The supplementary material should also include the distribution on the target properties like QED and SA.
Please comment on the progress of learning during the different learning cycles on aspects like novelty and docking score. The distributions of docking score show some shift towards higher score (fig.3 and 5), but that appears to be moderate. How does active learning influences this learning progress?

Some more experimental details are needed for the following topics (could be partially added to the supplementary material):
• How is the active learning done? There are details on given in the manuscript.
• What is the ATP concentration in the assay? This helps to better understand the strength of binding of the identified ligands.
• Give more clarity on the difference between temporal and permanent data in the design phase.
• How is the training data for fine-tuning exactly selected. Please give more details.
• HTVS is only sparsely described, but is a key contributor for training data.
The approach merits publication due to its novel workflow and experimental validation for CDK2 inhibitors. However, additional details on KRAS validation and technical implementation would significantly strengthen the manuscript and increase its value to the scientific community. The work represents an important contribution to AI-driven drug design, particularly for targets with limited available data.

Minor aspects:
It would be helpful for the reader to directly show clinical molecules of the two targets for comparison.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors addressed the comments very efficiently. I believe the manuscript can now be accepted for publication.

Reviewer #3

(Remarks to the Author)
The authors have addressed comments carefully. Revisions have significantly enhanced the clarity and scientific rigor of the manuscript.
The combining generative AI with physics-based active learning represents an exciting contribution to the field of drug discovery, particularly given the experimental validation of the CDK2 inhibitors and the demonstration of the method's utility in low-data scenarios.

I am satisfied with all the changes made and believe the manuscript is now suitable for publication.

Reviewer #1

The paper by Filella-Merce reports an interesting generative modeling approach making use of an active learning cycle. The paper prospectively synthesized and tested generated molecules, which is very refreshing to see in this kind of work. I **believe the work is publishable in Communications Chemistry** after few minor points to address:

We thank Reviewer 1 for their thoughtful and positive evaluation of our work. We appreciate their recognition of the practical implementation of our GM workflow. Below, we address each of their suggestions and questions point by point.

- In the Introduction, the authors mentioned the advantages of VAE over GAN. While I agree with their choice and justification, a brief discussion and comparison with other generative frameworks (eg diffusion models) would be much appreciated especially because some of those models are regularly employed in drug discovery

We agree that providing a broader contextualization of generative frameworks is valuable. In response, we have added a new paragraph (now paragraph 3 of the Introduction) where we briefly discuss and cite the main architectures used for molecular generation, highlighting their key strengths and limitations. Among these, we also included diffusion models, which, at the time of our initial submission to arXiv, were only beginning to emerge as a potential alternative in this field. That said, and as further elaborated in the revised paragraph 4 of the Introduction, we would still have chosen a VAE as our generator. This choice is primarily based on the reasons outlined there, particularly its seamless integration with AL, which serves as the guiding tool and core of our method.

- Active learning in hit finding is an established technique; while I appreciate the generative angle that the authors used (which differs from most active learning works), elaborating on existing applications could help to put the work more in context

We agree that expanding on existing applications of AL in drug discovery will help to more clearly contextualize the contribution of our work. Building on this suggestion, we have added a new paragraph (now paragraph 5 of the Introduction) to better situate AL within the field. Additionally, we have revised paragraph 6 to emphasize how our approach introduces a novel generative perspective to AL, distinct from the more common use of AL to efficiently navigate fixed molecular libraries without the need for exhaustive evaluation.

- In the ABFE section for CDK2, the authors mentioned that some molecules were predicted actives and some inactives - what does this mean? Did the authors use a specific threshold? Or were inactives molecules displaying a positive ABFE value?

To clarify this point and ensure a more objective presentation of our results, we have revised the manuscript to remove references to “active” or “inactive” molecules. Instead of defining a potentially arbitrary threshold to distinguish between active and inactive, we now report only the IC_{50} and predicted K_d values for all evaluated molecules. This change has been applied consistently throughout the manuscript.

- The authors used BioSolveIT InfiniSee to look for molecules similar to their generated structures both for CDK2 and KRAS. They mentioned that the closest molecule(s) was(were) 0.8 similar to their compounds (or at least this is how I interpreted it); this means that they actually found very close molecules, and I think it is fair to elaborate on this - for example by showing a Tanimoto map or plotting the 2d structures of the closest pairs?

We agree that further clarification on this point was needed. To address this, we performed a more in-depth analysis by expanding the analog search to the larger Enamine Real Space (76B, instead of the previous 5B) and provided detailed information on the closest analogs found for each compound in two new figures (Supplementary Figures 6 and 7) and two new tables (Supplementary Tables 7 and 8). Based on this analysis, we revised the final paragraphs of the CDK2 and KRAS results subsections to highlight key findings. Specifically, we emphasized that none of the generated compounds were present in the chemical space, and for most, no analogs with similarity above 0.5 were found. In the few cases where analogs with similarity between 0.5 and 0.72 were identified, they shared similar but not identical scaffolds with the query compounds. These results further support the novelty of the generated molecules.

Reviewer #2

This paper describes a machine learning approach to generate novel inhibitors of CDK2 and KRAS. The process is divided in two stages: the first filters molecules based on drug-likeness, diversity and synthesizability while the second filters molecules based on docking scores. These stages are performed iteratively to find the best candidates for experimental testing. In the end, authors claim to have found 8 active molecules for CDK2 and 4 molecules predicted as active for KRAS, but these have not been tested.

We would like to thank the reviewer for their valuable comments and suggestions. Each point is addressed individually in the responses below.

Major comments:

- The definition of “active” molecules seems a bit awkward to me. The authors label molecules with an IC50 value of less than 100uM as active, which appears quite permissive and perhaps unrealistic. It is true that the threshold between active and inactive can be arbitrary and varies depending on the target, assay, and other factors. Generally, a threshold of 10uM is common for hit identification and it typically reduces to 1uM for well-established targets like kinase inhibitors such as CDK2. A standard lead optimization campaign for kinase inhibitors aims to discover compounds in the low nanomolar range (less than 100nM). Assuming this work would be equivalent to a hit identification stage and use 1uM as the threshold, then only 1 out 9 molecules showed in vitro activity for CDK2.

To address the reviewer’s concern, we have revised the manuscript to adopt a more objective tone throughout. References to compounds as ‘active’ or ‘inactive’ have been removed, and we now present only the measured IC_{50} or predicted K_d values of the evaluated molecules. These changes have been highlighted in red and blue across the abstract, introduction, results, and discussion sections.

That said, we appreciate the reviewer's point regarding the IC_{50} thresholds used to define active or inactive molecules, which indeed depend on the target and the stage of the drug discovery process. As this work is positioned within the hit identification phase, with the primary goal of identifying novel chemical scaffolds rather than optimizing potency, the identification of molecules with scaffolds structurally distinct from known CDK2 inhibitors and IC_{50} values below 30 µM represents, in our view, promising hits worthy of further investigation.

- It is challenging to discern if the computational methods applied in this paper provide a good estimate for experimental values in the case of KRAS. The authors used a set of known KRAS inhibitors consisting of 73 molecules. There is no demonstration of whether the docking Glide score or PELE binding free energy correlates with the experimental values of these molecules (as presented for CDK2 in Supplementary Fig2). Even though these methodologies work for CDK2, they are highly dependent on the target. ABFE was evaluated on just 4 compounds, but the results are not presented. Moreover, the authors indicate that there's no correlation between the Glide score and ABFE, which questions the effectiveness of Glide scores in selecting compounds in the outer loop. Without concrete evidence of how accurately ABFE predicts bioactivity in KRAS, I believe it's insufficient to declare the four chosen molecules as active without an experimental value.

Unfortunately, unlike CDK2, we could not validate the predictive power of Glide or PELE for KRAS due to limited experimental data. For $KRAS^{G12D}$, only ~35 molecules with reported IC_{50} values were available after excluding $KRAS^{G12C}$ covalent inhibitors (with warheads removed), compounds with K_d instead of IC_{50}, and crystal structures lacking affinity data. These molecules show a narrow IC_{50} range and belong mainly to a few closely related congeneric series developed via structure-based design, limiting both chemical diversity and affinity variability. In our experience, robust validation of molecular modeling methods requires a sufficiently large and diverse dataset with broad affinity variation. Given these constraints, a correlation analysis for KRAS, like that shown for CDK2 in Supplementary Figure 2, was not feasible. We have now clarified this limitation in the Methods subsection 'Glide docking'.

We thank the reviewer for noting the omission of ABFE benchmarking. The ABFE results for the KRAS and CDK2 compounds used as benchmarks are included in Suppl. Tables 6 and 3, respectively, but were not previously mentioned in the text. We have now added a direct reference to these results in the 'ABFE simulations' paragraph of the CDK2 and KRAS Results section and in the Methods subsection 'Absolute binding free energy simulations'.

Regarding the lack of correlation between Glide and ABFE scores in KRAS, we would like to point out that Glide is only used in our workflow as an initial filtering step within an active learning cycle, aimed at guiding molecular generation. Selected candidates are then evaluated with more rigorous *in silico* evaluations, PELE and ABFE, which, despite being orthogonal, show consistent correlation.

Finally, we have followed the reviewer's recommendation and revised the manuscript to replace any reference to active compounds for KRAS, instead describing them as 'molecules with potential activity against KRAS'.

Minor comments:

- Docking scores are generally an estimate of how well a molecule fits into the binding pocket, but this doesn't necessarily imply bioactivity. It's not unusual for docking scores to not align with experimental values or with more precise binding prediction methods, as we've seen in the case of KRAS. Authors should be cautious when referring to docking scores as binding energy predictions or target engagement.

We fully agree that docking scores are only predictions and do not directly imply bioactivity or accurate binding affinity. To avoid any potential confusion for the reader, we have revised the manuscript to clarify that docking is used to guide molecular generation toward chemical regions expected to exhibit improved target engagement (or higher predicted affinity) and to facilitate the generation of virtual hits. These changes have been highlighted in red and blue across the text.

Furthermore, we would like to emphasize that in our workflow, docking is used solely as an initial filtering tool to guide molecular generation through active learning and to select an initial set of candidates, which are subsequently subjected to more rigorous *in silico* and *in vitro* validations.

- It's suggested to avoid referring to untested experimentally molecules as "active molecules". For instance, in the last sentence of the introduction, "we identified 4 active molecules for KRAS" could be better expressed as "we identified 4 molecules with potential activity against KRAS".

We agree with the reviewer's suggestion, and in line with their first major comment, we have removed references to 'active' or 'inactive' molecules throughout the manuscript. Instead, we focus on reporting objective IC_{50} and predicted K_d values. In instances where it is necessary to refer to molecules with low predicted K_d values from ABFE simulations, we have followed the reviewer's recommendation and describe these as "molecules with potential activity against KRAS." These changes have been highlighted in red and blue across the text.

- The authors refer to their method as an "active learning framework". However, in my view, this doesn't seem to be genuine active learning as it doesn't involve experimental labeling of new data at any point. Instead, it resembles the self-distillation approach used in AlphaFold [1] for training with unlabeled data.

Active learning is a commonly used machine learning technique that iteratively refines model predictions/generations through a feedback loop. While this feedback can involve experimental labeling, in many drug discovery endeavors, particularly in early stages, computational evaluations are used as a practical alternative. Thus, our framework aligns with the broader definition of active learning, where iterative selection and evaluation help guide the model. To clarify this, we have expanded the introduction to include a paragraph (5th) presenting active learning in drug discovery, including both experimental and computational feedback strategies.

- Figure 2B clearly shows that the number of validated molecules decreases with each cycle. This might be due to the constraint of molecular diversity, but it's hard to tell from the current figure. It would be useful to display the change in each property with each cycle.

We have extended Figure 2B to include all outer AL cycles in Supplementary Figures 8B (CDK2) and 9B (KRAS). In addition, A and C-E panels display the evolution of all inner AL cycle metrics used for molecule acceptance: validity, QED, SA, and maximum Tanimoto similarity to the temporal-specific training set. Notably, the decreasing trend in accepted molecules suggested by Figure 2B (which only reflects inner AL cycles of the first outer AL cycle) does not hold consistently across all outer AL cycles. Instead, the number of accepted molecules by inner cycles appears to be primarily influenced by validity, which fluctuates due to the stochastic nature of the VAE, and by molecular diversity, as reflected in the Tanimoto similarity filtering. SA is consistently satisfied, and while QED may also contribute to rejection, its impact appears to be less significant than validity and similarity.

- Supplementary Figure 2A illustrates the correlation between Glide score and experimental values for the set of known molecules. It indicates a correlation coefficient of -0.56, which could be supplemented with an error measurement like MSE or RMSE. Based on the present data, Glide scores appear to moderately correlate with experimental values and may not be suitable for detecting molecules that fit in the binding site.

We agree that Glide gscore predictive power is limited and should be utilized with caution. Accordingly, in our study, docking is not used as a definitive predictor of binding affinity, but rather as an initial prioritization tool within an active learning framework. Its role in the generative workflow is to iteratively guide molecular generation toward compounds with increasingly lower Glide gscores, thereby enriching the pool of virtual hits that advance to more rigorous *in silico* and *in vitro* validation steps.

Our objective is to enrich the low-scoring region, where the likelihood of identifying true positives is higher, rather than to achieve a strong correlation with experimental binding data. Finally, as the reviewer pointed out, docking scores are metrics that estimate how well a molecule fits into the binding pocket and can be interpreted as rough predictors of affinity. However, they are not intended to directly predict IC_{50} values. For this reason, we did not include RMSE or MSE calculations, as we believe that, in this context, correlation provides a more informative assessment of the relationship between docking scores and experimental data.

- Figure 4C presents the structures of the molecules selected for experimental validation. It could be beneficial to include the structure of the nearest neighbor in the initial set as part of the supplementary material to back up the claim that these differ from the initial set.

In response to the reviewer's suggestion, we have included in Figure 4B the Tanimoto similarity between each synthesized molecule and its nearest neighbor in the initial-specific training set (values also provided in Suppl. Table 3). All similarities are below 0.3, supporting the claim that the synthesized compounds are structurally distinct from the initial-specific set. To further highlight the novelty of the synthesized molecules, we included the 2D structure of their nearest neighbors from Enamine REAL Space (76 billion molecules) in Suppl. Figure 6 and expanded on this analysis in the final paragraph of the CDK2 results section. Similarly, for KRAS, we have added the similarity to the nearest neighbor in the initial-specific training set in Figure 6B, and we present the corresponding novelty analysis against Enamine REAL Space in Suppl. Figure 7 and the final paragraph of the KRAS results section.

- Figure 6B and 6C do not align with the provided legend. They appear to be switched.

We would like to thank the reviewer for the observation. Figure 6 has been corrected to ensure consistency with the legend.

Reviewer #3

The manuscript „Optimizing Drug Design by Merging Generative AI With a Physics-Based Active Learning Framework" presents a novel generative AI workflow that combines a variational autoencoder (VAE) with two nested active learning cycles and docking as well as physics-based estimation of binding affinity. The approach is applied to the design of CDK2 inhibitors and kRAS inhibitors. CDK2 inhibitors are experimentally confirmed. The manuscript is written well and clearly describes the technology. The integration of generative AI and physics-based approaches is a meaningful approach for using structural knowledge. The authors present the successful design of CDK2 inhibitors based on this approach with 8 molecules showing biological activity.

We would like to thank the reviewer for their considerate comments and positive evaluation of our work. We appreciate the suggestions and will address each point in detail in the responses below.

The described approach would be especially helpful in the low-data regime where only little ligand information is available. kRAS inhibitors have been chosen to illustrate the capability of the technology, but binding affinity of suggested molecules have only been validated by ABFE calculations. Therefore, it is crucial, that validation data for the ABFE predictions are included at least in the supplementary data. Correlations show for a subset of training data on CDK2 appear moderate. Also explain how the displayed subset was selected.

We thank the reviewer for noting the omission of ABFE validation data. The ABFE results for the KRAS and CDK2 compounds used as benchmarks are included in Suppl. Tables 6 and 3, respectively, but were not previously mentioned in the text. We have now corrected this by adding a direct mention in the 'ABFE simulations' paragraph of the CDK2 and KRAS Results section and in the Results subsection 'Absolute binding free energy simulations'. Additionally, we would like to clarify that the correlation between ABFE predictions and experimental data in CDK2 was assessed for the synthesized compounds.

For generated molecules, standard analysis should be included such as validity, uniqueness and novelty. The supplementary material should also include the distribution on the target properties like QED and SA.

In response to the reviewer's suggestion, we have added Supplementary Figures 8 and 9 for CDK2 and KRAS, respectively. These figures present the evolution of validity, uniqueness, and novelty of the generated molecules across inner AL cycles. They also show the progression of the AL properties, average QED, SA, and maximum Tanimoto similarity to molecules in the temporal-specific training set, for the valid molecules. References to these supplementary figures have now been added to the corresponding sections of the main text, highlighted in red and blue.

Please comment on the progress of learning during the different learning cycles on aspects like novelty and docking score. The distributions of docking score show some shift towards higher score (fig.3 and 5), but that appears to be moderate. How does active learning influences this learning progress

In response to the previous comment, we have added Supplementary Figures 8 and 9 for CDK2 and KRAS, respectively, showing the progress of learning across AL cycles with generative and AL metrics. As expected, the use of Tanimoto filtering maintains high novelty among unique generated molecules, supporting effective exploration of new chemical space. Regarding the Glide gscore distributions, while the observed shift might be seen as moderate, there is a clear enrichment of molecules with Glide gscores below –8.0 kcal/mol, driven by the outer AL cycle thresholding, as discussed in the main text. Although a more aggressive approach (e.g., gradually tightening the threshold) could have pushed scores lower, our strategy focused on enriching and selecting from already promising regions and compensating with more extensive *in silico* and *in vitro* validation to support candidate prioritization.

Some more experimental details are needed for the following topics (could be partially added to the supplementary material):
• How is the active learning done? There are details on given in the manuscript.

To better clarify how active learning guides the generation process in our workflow, we have added a dedicated sub-subsection titled 'Nested AL' within the 'Generative Model Workflow' subsection of the Methods.

• What is the ATP concentration in the assay? This helps to better understand the strength of binding of the identified ligands.

The ATP concentration used in the assay was 150 µM. This information, along with additional experimental details, has been added to the Methods subsection 'Inhibitory Assays' to provide greater clarity on assay conditions and to help contextualize the binding strength of the identified ligands.

• Give more clarity on the difference between temporal and permanent data in the design phase.

To better clarify the differences between temporal and permanent specific sets, we have added two dedicated sub-subsections titled 'Temporal-Specific Set' and ‘Permanent-Specific Set’ within the 'Generative Model Workflow' subsection of the Methods.

• How is the training data for fine-tuning exactly selected. Please give more details.

We have expanded on how the training data for fine-tuning is selected in the newly added sub-subsection "Nested AL" within the "Generative Model Workflow" subsection of the Methods.

• HTVS is only sparsely described, but is a key contributor for training data.

In response to the reviewer's suggestion, we would like to clarify that the details of the HTVS used to construct the $KRAS^{G12D}$ unknown initial-specific training set are provided in the Methods subsection '$KRAS^{G12D}$ initial specific training sets.' We have revised this subsection to improve clarity and provide more comprehensive information about the virtual screening procedure, compound filtering criteria, and selection process.

The approach merits publication due to its novel workflow and experimental validation for CDK2 inhibitors. However, additional details on KRAS validation and technical implementation would significantly strengthen the manuscript and increase its value to the scientific community. The work represents an important contribution to AI-driven drug design, particularly for targets with limited available data.

Minor aspects:
It would be helpful for the reader to directly show clinical molecules of the two targets for comparison.
In response to the reviewer's point, we have included the 2D structure of CDK2 inhibitors currently in clinical trials in a new Supplementary Figure 11, which is now referenced in the CDK2 Results subsection. For $KRAS^{G12D}$, the only clinical molecule to date is MRTX1133 (to the best of our knowledge). While its structure was already shown in Supplementary Figure 5, we have now also included it in Figure 6E and referred to it explicitly in the KRAS results subsection to facilitate direct comparison.