nature portfolio

Peer Review File

Molecular deep learning at the edge of chemical space

Corresponding Author: Professor Francesca Grisoni

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)
Overall a very interesting and solid paper of interest to the field. The introduction of an unfamiliarity metric via the reconstruction loss is an interesting idea to estimate a model's performance. The authors also have done a good job in the analysis showing the benefit of unfamiliarity metrics against some of the cheminformatic metrics and common approaches like embedding distance.

Here are three main concerns/questions:
1) For Figure 3 b vs. c, the author suggests that unfamiliarity is able to clearly differentiate the library molecules from the Test_ID or Test_OOD. However, it's unclear to me why this is the desired outcome (i.e. capture structural difference matters) unless the author can show that the model is indeed performing much much worse for molecules in the library vs. in Test_ID or Test_OOD.
2) For the prospective virtual screening section, it's unclear to me what the author's suggestions are for leveraging the uncertainty and unfamiliarity metric.
a) If the argument that targeting the three quadrants (method A/B/C) will be better than a random sample of top predicted bioactivity alone, I would like to see a better controlled baselines besides citations of typical initial rate ranging from 0.1% and 5%.
b) If the argument is that one of the methods should be better than the other, I would like to understand the intuition a bit even when there is no statistical significance.
3) If the authors believe unfamiliarity and uncertainty are capturing different behavior axes of the model, have you tried to build a hybrid metric that "blends" both unfamiliarity and uncertainty and see if the hybrid will correlate with the model performance better in Table 2?

Nit:
1) For Figure 2 h. is the x axis bins for Test_ID and Test_OOD bin actually aligned? or in other words, -does bin 1 for Test_ID's unfamiliarity metric represent the same range as bin 1 for Test_OOD?
a) If they represent the same range, why do similar unfamiliarity values yield high mol. corre overlap for Test_ID vs. Test_OOD. How data is split should not matter?
b) If they represent a different range, I suggest the author change the x-axis to the actual range or make it into two separate plots to show the trend.

(Remarks on code availability)
The code looks well organized and the data processing, model and analysis codes are detailed enough for reproducing the results. However, I was not able to run the steps in Demo because the data files are not there. For instance:

```
File "/home/user/JointMolecularModel/experiments/0_clean_data.py", line 20, in process_litpcba
with open(f'data/LitPCBA/{dataset_name}/actives.smi', 'r') as file:
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
FileNotFoundError: [Errno 2] No such file or directory: 'data/LitPCBA/ESR1_ant/actives.smi'
```

Reviewer #2

(Remarks to the Author)

This manuscript describes work towards increasing the generalizability of chemical machine learning models, in the context of novel structures. The authors introduce an approach that both trains a model to predict molecular properties, and reconstructs the input molecule. It appears that unfamiliar (poorly-reconstructed) molecules show how far a molecule deviates from the model's "knowledge". It then appears that the proposed unfamiliarity metric has some advantages over the traditional uncertainty estimation approach.

SMILES strings were first encoded into a single factor, which could then be decoded using an RNN. After training this encoder-decoder on a large set of unlabelled structures, the model was fine tuned, with a Bayesian classifier predicting molecular properties and estimating the traditional prediction uncertainty. Later, they screen a commercial compounds library for kinase inhibitors, using joint models trained on bioactivity data for the kinases. After removing compounds too similar to the training set, the top molecules were assayed experimentally, to identify hits.

The authors show correlation of several reliability metrics to the model performance (Table 2), showing that prediction uncertainty and unfamiliarity both correlate to poorer model performance, an unusual case of the dreaded bold table actually showing statistically-meaningful differences! The prediction uncertainty and unfamiliarity themselves show essentially no correlation, as uncertainty focuses only on p(y|x), which is highly interesting and suggests both can separately be metrics worth gathering.

The proposed joint model and unfamiliarity metric seem useful for machine learning applications. Provisionally, they have performed well in this case of bioactivity data, and may be useful for other chemical machine learning models.

The GitHub repository is also organized and well-documented with docstrings in at least some functions, along with comments and type hints. This makes a huge difference for scientists trying to reuse the work and is hugely appreciated!

For Equation 1, it may be "obvious" to some that p(t_i | t_<i) is the probability of t_i given the preceding tokens, but I think it would be worth explicitly stating this for less familiar readers. I have to say some further explanation would be very welcome for the unfamiliarity "providing insight into p(y, x|x)", as mentioned in the 'Unfamiliarity and bioactivity prediction' section! I do not have any further comments for improvement, and believe this manuscript will be of value to the community.

(Remarks on code availability)
Reviewed to some extent – I did not run the code, but it did seem much better documented than the vast majority of academic code, including installation instructions and many comments throughout.

Reviewer #3

(Remarks to the Author)
This manuscript addresses the problem of out-of-distribution (OOD) generalization in molecular property prediction. The authors propose a joint modeling approach and a new evaluation metric called unfamiliarity to examine the distributional differences between training and testing data. The topic is important; however, the core ideas and methodology appear incrementally novel, and the results do not convincingly support the stated goals. Several issues exist regarding data selection, experimental design, and result interpretation, as outlined below.

1. The autoencoder model in JMM was pre-trained on a molecular reconstruction task using unlabeled data. Considering that training on larger and more diverse datasets can inherently improve model generalization and reduce OOD issues, it would be more appropriate to train on a much larger dataset, such as the Enamine REAL Database (~10 billion compounds), rather than being limited to the ChEMBL dataset (~1.2 million compounds, suggested in the paper).

2. Since unfamiliarity is defined as the logarithm of the reconstruction loss, the quality of the autoencoder's training has a decisive impact on this metric. The authors should report the specific training and testing performance of the autoencoder to ensure that unfamiliarity is reliable and meaningful.

3. Since unfamiliarity and uncertainty are later used as ranking criteria for compound selection, the role of the prediction head in JMM for property prediction becomes unclear. The necessity of including this prediction component should be better justified.

4. As shown in Figure 2d, the classification performance of the JMM model does not significantly differ from other baselines, showing no clear advantage. The observation that the "no decoder" condition does not affect performance requires more detailed discussion, since the decoder has a direct and substantial influence on the unfamiliarity values.

5. Regarding the claimed advantage of unfamiliarity over uncertainty in distinguishing distribution shifts, Figures 3b and 3c show that both metrics exhibit statistically significant differences between in-distribution and out-of-distribution test sets. The visual differences alone do not convincingly demonstrate the superiority of unfamiliarity in this context.

Based on the above reasons, I believe this work requires substantial revisions and improvements before it can fully achieve its intended objectives.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I appreciate the author response and clarifiaction. They are very helpful and resolved most of my questions, except pieces around the experimental screening (Fig 4).

My understanding from author's response in 2)a is that:
i) low unfamilarity and high uncertainty (method A) performs the best
ii) low uncertainty (method B/C) are not very helpful.

Based on my understanding, here are my queston:
1) what's the intuition for method A doing better? based on author's response in 3) and writing of the manuscript, shouldn't low unfamilairy + low uncertainty (method B) performs the best?
2) I do not think author can conclude that low unfamiliarity is more informative than low uncertainty just by comparing method A vs. method B/C. Instead, author should compare the difference between low vs high unfamiliarity and low (B+C) vs. high uncertainty (A+D, a new method). If we see no difference in low vs. high uncertainty and significant difference in low (A+B) vs. high (C+D) unfamiliarity, I think it would support author's claim better.
If author cannot run more experiment, they should make the claim by comparing A+B vs. B+C. I do not think comparing A vs. B+C is sound.

I understand doing new experiments would be costly, but I do want to make it clear that lack of clear win in a prospective screening setup does limit the impact of the work, especially for a high impact journal.

(Remarks on code availability)
I appeciate the authors' effort in organizating the code and data.

Reviewer #2

(Remarks to the Author)
I have read through the revised manuscript and responses to reviewers, and have no further comments. This will make for an interesting publication!

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I would like to thank the authors for their detailed reply. All my concerns have been addressed in the rebuttal, and I have no further comments.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I would like to thank the authors for their detailed reply. All my concerns have been addressed in the rebuttal, and I have no further comments.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
Thanks to all authors for conscientious reply to the reviews. All my comments have been addressed so I have no further comment to make.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I have carefully reviewed the authors' latest rebuttal and the revised manuscript. The authors have effectively addressed the concerns, and I have no further comments.

(Remarks on code availability)

Reviewer 1

Overall, a very interesting and solid paper of interest to the field. The introduction of an unfamiliarity metric via the reconstruction loss is an interesting idea to estimate a model's performance. The authors also have done a good job in the analysis showing the benefit of unfamiliarity metrics against some of the cheminformatic metrics and common approaches like embedding distance.

>> We thank the reviewer for the positive feedback on the relevance of our study and the scientific soundness of our analysis.

1) For Figure 3 b vs. c, the author suggests that unfamiliarity is able to clearly differentiate the library molecules from the Test_ID or Test_OOD. However, it's unclear to me why this is the desired outcome (i.e. capture structural difference matters) unless the author can show that the model is indeed performing much much worse for molecules in the library vs. in Test_ID or Test_OOD.

>> The desired outcome is to detect when molecules lie far outside the training distribution, because this is where the model is expected to perform poorly. In the manuscript, we show that:

- While unfamiliarity correlates with molecular similarity (Table 1), it provides a better estimate of performance than similarity alone (Table 2).
- The higher the unfamiliarity, the lower the performance, across metrics and experiments (Supporting Fig. 4).

Although measuring the performance on the entire screening library would be unfeasible (as it would require substantial experimental efforts), we observe that several of those molecules are markedly more distant ('unfamiliar') from the training data than the test molecules. Hence, we can expect a more pronounced performance drop on screening molecules that have higher unfamiliarity values compared to the those in the test sets. This underscores the value of unfamiliarity in revealing differences that are not apparent when considering uncertainty or similarity alone. We now incorporated this reasoning more explicitly in the revised paper, as follows:

Line 222: "*The screening molecules showed a lower structural overlap with the training sets than the test$_{OOD}$ molecules (Fig. 3a, Supplementary Fig. S6a). Based on previous results (Fig. 2d), this indicates that we can also expect a corresponding performance drop on the screening molecules.*"

2) For the prospective virtual screening section, it's unclear to me what the author's suggestions are for leveraging the uncertainty and unfamiliarity metric.

>> Thank you for the opportunity to clarify. See our point-by-point answers below.

   a) If the argument that targeting the three quadrants (method A/B/C) will be better than a random sample of top predicted bioactivity alone, I would like to see a better controlled baselines besides citations of typical initial rate ranging from 0.1% and 5%.

   >> Our argument is not that one should target the three quadrants all together, but rather that some quadrants will be more favorable than others depending on the intended goals, as explained below.

b) If the argument is that one of the methods should be better than the other, I would like to understand the intuition a bit even when there is no statistical significance.

>> Statistical significance is inherently difficult to establish in small-scale experimental studies, and a large-scale prospective evaluation is currently unfeasible. However, we do see interesting trends. As visible from the table below (Table R1), considering low unfamiliarity scores and high uncertainty (method A) yields the highest number of hits across both protein targets. In contrast, methods with low prediction uncertainty (methods B and C) were less informative. We hope that broader adoption of our unfamiliarity metric will enable statistically robust conclusions. In the revised version of the manuscript (L292-296), we discuss these observations further.

Table R1. Summary of experimental results across targets and methods. The best result per metric is highlighted in boldface.

Target	Metric	Method A Unfam. ↓ Uncert. ↑	Method B Unfam. ↓ Uncert. ↓	Method C Unfam. ↑ Uncert. ↓
PIM1	Hit rate	**30%**	**10%**	10%
	Most active molecule (IC_{50})	**0.9±1.0 µM**	7.5 µM	**0.9±0.5 µM**
	Second most active molecule (IC_{50})	**1.5±0.4 µM**	NA	NA
CDK1	Hit rate	**20%**	0%	0%
	Most active molecule (IC_{50})	**2.2±0.7 µM**	NA	NA
	Second most active molecule (IC_{50})	**3.4 µM**	NA	NA

3) If the authors believe unfamiliarity and uncertainty are capturing different behavior axes of the model, have you tried to build a hybrid metric that "blends" both unfamiliarity and uncertainty and see if the hybrid will correlate with the model performance better in Table 2?

>> We appreciate this interesting suggestion. Fig. 3 already explores this idea: by jointly visualizing uncertainty and unfamiliarity, we show that they capture largely independent aspects of model behavior and that combining them yields a more structured view of when predictions are reliable. Developing a new composite score is beyond the scope of the present work, which focuses on characterizing unfamiliarity itself rather than proposing additional reliability metrics on top of it.

4) For Figure 2 h. is the x axis bins for Test_ID and Test_OOD bin actually aligned? or in other words, does bin 1 for Test_ID's unfamiliarity metric represent the same range as bin 1 for Test_OOD?

>> Yes, unfamiliarity values were binned per dataset, meaning that for every target, Test_ID and Test_OOD bins are aligned. We made this more explicit in the revised manuscript and in the figure caption.

a) If they represent the same range, why do similar unfamiliarity values yield high mol. core overlap for Test_ID vs. Test_OOD. How data is split should not matter?

>> What the reviewer observes is what Fig. 2h shows: regardless of data splitting, there is a direct relationship between the similarity to the training set and the unfamiliarity score. We clarified this point more explicitly in the revised manuscript (L175).

b) If they represent a different range, I suggest the author change the x-axis to the actual range or make it into two separate plots to show the trend.

>> Referring to bins allows us to have a more robust statistics across datasets (which might have different distributions of unfamiliarity) on how unfamiliarity relates to structural similarity. Complementary to Fig. 2h, we show the "unbinned" unfamiliarity values in Fig. 2g, encompassing all datasets together in relation to structural similarity to the training set.

Remarks on code availability.

The code looks well organized and the data processing, model and analysis codes are detailed enough for reproducing the results. However, I was not able to run the steps in Demo because the data files are not there. For instance:

```
File "/home/user/JointMolecularModel/experiments/0_clean_data.py", line 20,
in process_litpcba
with open(f'data/LitPCBA/{dataset_name}/actives.smi', 'r') as file:
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
FileNotFoundError: [Errno 2] No such file or directory:
'data/LitPCBA/ESR1_ant/actives.smi'
```

>> Thank you for the positive comments on the code and repository. The code the reviewer tried to run indeed requires data that is supplied separately in an Zenodo repository. We have now included better instructions in the code repository to clarify this and provide the data accordingly. Likewise, we now also include the code and data to replicate all our main figures and the statistical tests accompanying them.

Reviewer 2

The authors show correlation of several reliability metrics to the model performance (Table 2), showing that prediction uncertainty and unfamiliarity both correlate to poorer model performance, an unusual case of the dreaded bold table actually showing statistically-meaningful differences! The prediction uncertainty and unfamiliarity themselves show essentially no correlation, as uncertainty focuses only on p(y|x), which is highly interesting and suggests both can separately be metrics worth gathering.

The proposed joint model and unfamiliarity metric seem useful for machine learning applications. Provisionally, they have performed well in this case of bioactivity data, and may be useful for other chemical machine learning models.
>> We thank the reviewer for thoroughly analyzing our approach and for the recognition of its relevance in the molecular sciences.

The GitHub repository is also organized and well-documented with docstrings in at least some functions, along with comments and type hints. This makes a huge difference for scientists trying to reuse the work and is hugely appreciated!
>> We are glad to see that our efforts in making a usable repository are appreciated by this reviewer.

1) Add some further explanation around unfamiliarity for Eq 1. “providing insight into p(y, x|x)”, as mentioned in the ‘Unfamiliarity and bioactivity prediction’ section!

>> Thank you for the suggestion. We added additional clarifications on unfamiliarity to the revised manuscript (L96-97).

Remarks on code availability: Reviewed to some extent – I did not run the code, but it did seem much better documented than the vast majority of academic code, including installation instructions and many comments throughout.
>> Thank you for appreciating our efforts in providing code that is easy to understand, use and expand upon.

Reviewer 3

This manuscript addresses the problem of out-of-distribution (OOD) generalization in molecular property prediction. The authors propose a joint modeling approach and a new evaluation metric called unfamiliarity to examine the distributional differences between training and testing data. The topic is important; however, the core ideas and methodology appear incrementally novel, and the results do not convincingly support the stated goals. Several issues exist regarding data selection, experimental design, and result interpretation, as outlined below.

>> Thank you for recognizing the relevance of the topic we are addressing. In what follows, we have carefully addressed your comments on a point-by-point basis to better highlight the rigor of our analysis, as well as the novelty and efficiency of the proposed architecture.

1) The autoencoder model in JMM was pre-trained on a molecular reconstruction task using unlabeled data. Considering that training on larger and more diverse datasets can inherently improve model generalization and reduce OOD issues, it would be more appropriate to train on a much larger dataset, such as the Enamine REAL Database (~10 billion compounds), rather than being limited to the ChEMBL dataset (~1.2 million compounds, suggested in the paper).

   >> Existing evidence on SMILES generative models indicates that scaling to extremely large datasets does not yield proportional gains (Skinnider et al. 2021. Chemical language models enable navigation in sparsely populated chemical space. *Nat. Mach. Intell.* 3, 1112). In fact, the performance in learning the "SMILES grammar" plateaus after ~50,000-100,000 molecules. Our autoencoder was pretrained on ~1.2 million ChEMBL structures – well beyond where diminishing returns begin – so we do not expect substantially different behavior from pretraining on multi-billion-compound collections. We added this clarification to the manuscript (L87).

2) Since unfamiliarity is defined as the logarithm of the reconstruction loss, the quality of the autoencoder's training has a decisive impact on this metric. The authors should report the specific training and testing performance of the autoencoder to ensure that unfamiliarity is reliable and meaningful.

   >> We thank the reviewer for this remark. We understand the concern: if the autoencoder were of poor quality or had collapsed, reconstruction losses would be uniformly high, and unfamiliarity would indeed be uninformative. In practice, however, unfamiliarity depends on *relative* reconstruction behavior rather than absolute reconstruction accuracy. As long as the autoencoder learns a meaningful chemical manifold, the reconstruction loss remains a reliable indicator of deviation from that manifold. This is exactly what we observe in practice: the reconstruction losses are consistently lower for training and in-distribution molecules than for OOD molecules (Fig. 2f), demonstrating that the model has not collapsed and that unfamiliarity captures distribution shift as intended.

3) Since unfamiliarity and uncertainty are later used as ranking criteria for compound selection, the role of the prediction head in JMM for property prediction becomes unclear. The necessity of including this prediction component should be better justified.

   >> The prediction head has a key role in our pipeline, as it serves to (a) predict bioactivity, (b) provide a measure of uncertainty over these predictions, and (c) link the model reconstruction

to the task being performed. Without this part, unfamiliarity would not capture information on the predictive task to be performed, and hence it would not serve for OOD detection. In the prospective study, we used the predicted bioactivity alongside uncertainty and unfamiliarity to select molecules. We now mention this aspect more explicitly in the revised manuscript (257-263) and in Figure 4.

4) As shown in Figure 2d, the classification performance of the JMM model does not significantly differ from other baselines, showing no clear advantage. The observation that the "no decoder" condition does not affect performance requires more detailed discussion, since the decoder has a direct and substantial influence on the unfamiliarity values.
>> What the reviewer observes is exactly the point we want to show. The decoder in the JMM does not serve to improve performance, but to enable the estimation of unfamiliarity. Hence, it is crucial that this addition does not lead to a detriment in performance compared to existing approaches. Figure 2d shows that (a) the JMM matches the performance of established baselines, and (b) removing the decoder yields nearly identical results. This is a positive, and important, outcome: the decoder provides additional information (unfamiliarity), without imposing a performance penalty on the classifier. In the revised version of the manuscript, we have added a clarification to make this rationale more explicit (L158-160).

5) Regarding the claimed advantage of unfamiliarity over uncertainty in distinguishing distribution shifts, Figures 3b and 3c show that both metrics exhibit statistically significant differences between in-distribution and out-of-distribution test sets. The visual differences alone do not convincingly demonstrate the superiority of unfamiliarity in this context.
>> Indeed, both unfamiliarity and uncertainty show statistically significant differences between in-distribution and out-of-distribution molecules. However, in the context of distribution-shift detection, statistical significance is not the main criterion, but the effect size is. We now report the effect size (D) of the distribution shift with the Kolmogorov-Smirnov statistic: unfamiliarity (Library: D=0.99897, Test_OOD: D=0.36773) vs. uncertainty (Library: D=0.18071, Test_OOD: D=0.15546). These numbers indicate (a) a small difference in the distributions of uncertainty values (D≈0.18) and (b) almost no overlap (D≈1.00) in the case of unfamiliarity distributions. Even extremely small shifts in uncertainty become significant, but the distributions remain practically indistinguishable (Fig. 3b). The dramatic quantitative shift observed for unfamiliarity indicates that it is substantially more informative for detecting distribution shifts than uncertainty, despite both metrics reach statistical significance. We now explicitly mention this in the revised manuscript and report the effect sizes (L226-230). Thank you for this important observation.

Reviewer 1

I appreciate the author response and clarification. They are very helpful and resolved most of my questions, except pieces around the experimental screening (Fig 4).

My understanding from author's response in 2)a is that:

i) low unfamiliarity and high uncertainty (method A) performs the best

ii) low uncertainty (method B/C) are not very helpful.

Based on my understanding, here are my question:

1) what's the intuition for method A doing better? based on author's response in 3) and writing of the manuscript, shouldn't low unfamiliarly + low uncertainty (method B) performs the best?

>> The outcome of the prospective screening, together with the distributional analyses in Fig 3a-c, indicates that the model can seem 'highly confident' while operating outside the training support. This implies that under substantial distribution shift, uncertainty may lose discriminative power for identifying unreliable predictions. This is a well-known phenomenon, previously observed in literature (e.g., Tossou et al. 2024, *J. Chem. Inf. Model.* 64; Ovadia et al. 2019, *NeurIPS*; Scalia et al. 2020, *J. Chem. Inf. Model.* 60). On the contrary, unfamiliarity seems to retain its utility as a performance indicator, since it was explicitly designed to capture out-of-distribution predictions. We believe that this is the reason why method A outperforms method B.

2) I do not think author can conclude that low unfamiliarity is more informative than low uncertainty just by comparing method A vs. method B/C. Instead, author should compare the difference between low vs high unfamiliarity and low (B+C) vs. high uncertainty (A+D, a new method). If we see no difference in low vs. high uncertainty and significant difference in low (A+B) vs. high (C+D) unfamiliarity, I think it would support author's claim better.
If author cannot run more experiment, they should make the claim by comparing A+B vs. B+C. I do not think comparing A vs. B+C is sound.

I understand doing new experiments would be costly, but I do want to make it clear that lack of clear win in a prospective screening setup does limit the impact of the work, especially for a high impact journal.

>> We appreciate the reviewer's suggestion of a full 2x2 comparison between uncertainty (low/high) and unfamiliarity (low/high). We agree that such a factorial design would be required to estimate main effects cleanly and to make a definitive statement that "unfamiliarity is more informative than uncertainty" in general.

We believe that such analysis would be valuable *if and only if* statistical significance is achieved. To address this point rigorously, we performed a prospective power analysis. Using standard power calculations for two-sample comparisons of binomial proportions (Fleiss et al., 2003, Statistical Methods for Rates and Proportions), and assuming baseline hit rates typical for such assays (≈10%-15%; Macarron et al., 2011, *Nat. Rev. Drug Discov.* 10), we estimate that detecting

a modest but scientifically meaningful improvement in hit rate with 80% power at α = 0.05 would require on the order of 200–500 independent molecules per condition, totaling approximately 800-2,000 additional molecules per target (with full IC_{50} determination with replicates; with a cost of approx. 100-300 euros/molecule). These experiments are experimentally and logistically prohibitive in terms of material consumption, assay time, and cost, and go well beyond the scope of the current study, and of traditional experimental validations of machine learning approaches (which are sparse, and rarely include more than the strategy being proposed, unlike our study).

That said, our prospective results do support important points:

1. In this strongly out-of-distribution screening setting, prioritizing the low-uncertainty optimum (methods B and C) did not yield a clear prospective advantage. Importantly, Fig. 3a-c shows that screening-library compounds can appear low-uncertainty despite being clearly shifted in unfamiliarity, indicating that low uncertainty is not a reliable competence signal under this shift.
2. Among the two low-unfamiliarity strategies, imposing a low-uncertainty objective (method B) did not improve outcomes relative to method A (indeed, most low-micromolar hits originated from method A, albeit in a small sample). Given the evidence that uncertainty is not a reliable performance signal under the screening-library shift, the most conservative interpretation is that remaining in low-unfamiliarity regions is important, whereas optimizing for low uncertainty is not beneficial in this OOD regime and may bias selection toward overconfident regions. A full decomposition of uncertainty-unfamiliarity interactions would indeed require the missing high-uncertainty/high-unfamiliarity quadrant, but the data support this restricted conclusion without requiring that "high uncertainty" is itself causative.

Regarding the comment on impact: we agree that the prospective set is not powered to establish a statistically robust ranking of selection strategies; however, the intent of this section is to demonstrate that the proposed approach can identify novel bioactive compounds in a real screening context. In that sense, even a limited prospective validation provides nontrivial evidence that the method is not purely retrospective and strengthens the practical relevance of the work.

We have revised the Prospective virtual screening section to clarify that this experiment is primarily a feasibility validation rather than a powered comparison of strategies (line 307-308), to explicitly note that a full 2×2 factorial design would be required to estimate main effects (line 308), and to further nuance the corresponding claims accordingly (line 309-314).

Remarks on code availability

I appreciate the authors' effort in organizing the code and data.
>> Thank you!

Reviewer 2

I have read through the revised manuscript and responses to reviewers, and have no further comments. This will make for an interesting publication!
>> Thank you!

Reviewer 3

I would like to thank the authors for their detailed reply. All my concerns have been addressed in the rebuttal, and I have no further comments.
>> Thank you!