nature portfolio

Peer Review File

Chemist-aligned retrosynthesis by ensembling diverse inductive bias models

Corresponding Author: Dr Marwin Segler

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
In this manuscript, the authors present RetroChimera, a novel single-step retrosynthesis prediction method that integrates two components to achieve superior performance. Specifically, one component is a molecule-editing approach, where the authors propose a new module, NeuralLoc, which encodes templates and products independently. The other component is a de-novo generation method, based on a modified version of R-SMILES. To integrate these two modules, the authors introduce a “scoring function” that fuses results based on their overlaps. Experimental results demonstrate that compared with existing methods, RetroChimera exhibits clear advantages in low-data and out-of-distribution scenarios (Fig. 3), and is able to generate complex synthetic routes that chemists often find difficult to conceive but prefer once discovered (Fig. 4, 6). Overall, while the methodological development may not represent a fundamental breakthrough, the framework significantly improves model scalability, which will likely be valuable to the community. However, the biological validation of predictions is relatively weak, and several major concerns need to be addressed:
1. Ease of use. In the provided examples, the authors only show workflows for single-step and multi-step predictions. The internal computational process appears largely opaque and uncontrollable for chemists without an AI background. For instance, users seem unable to customize or modify intermediate steps. I recommend that the authors consider developing a human–computer interaction platform to enhance the practical utility of their tool. Existing platforms such as SYNTHIA, ASKCOS, and Reaxy Predictive Retrosynthesis could serve as useful references.
2. Framework scalability. As an ensemble learning framework, RetroChimera should be highly extensible to accommodate future methodological updates. The authors are encouraged to design interfaces that support the integration of additional models. Furthermore, the claim that the framework can incorporate reaction databases and other external resources is intriguing and potentially impactful. A more detailed explanation of how this integration could be practically realized would be valuable.
3. Validity of predictions. The authors state that RetroChimera is able to predict uncommon reactions. Have these predictions been experimentally validated, or are they simply deemed plausible from a chemical perspective?
4. Choice of models. As an ensemble framework, incorporating more models might further improve performance. A natural question arises: can the proposed scoring function adaptively filter out noise introduced by poorly performing models, such as low-confidence or evidently incorrect predictions? Although this study integrates two models—one template-based and one de novo—the authors should provide theoretical justification and quantitative evidence of the complementarity between the two, as well as assess potential redundancy. For example, are there cases where specific model types perform better for certain reaction categories?
5. Comparison with recent methods. The authors should compare RetroChimera with more recently developed large-scale pretrained retrosynthesis models that report strong performance, such as RXNGraphormer (Nature Machine Intelligence, 2025) and RSGPT (Nature Communications, 2025). In addition, other representative deep learning methods like RetroExplainer (Nature Communications, 2023) should be included for a more comprehensive evaluation.
6. Model interpretability. It would strengthen the work if the authors could provide insights into why the model selects a particular step (e.g., which bond is broken/formed, or which substructure is chosen as the reaction center). Beyond interpretability, such analysis may reveal novel patterns. For instance, RetroChimera performs well in low-data and out-of-distribution scenarios—does this arise from the discovery of higher-order “meta-rules” or from leveraging complementary heuristics, thereby creating a kind of “shortcut” in retrosynthetic reasoning?

7.Multi-step planning and pathway search. The authors integrate RetroChimera into syntheseus for multi-step planning. Compared with other available methods, what are the specific advantages of syntheseus? A discussion of this choice would provide important context.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
This paper presents an ensembling approach for retrosynthesis prediction, which fuses two types of retrosynthesis prediction methods---molecule editing and de novo generation---by learning to rank the importance of the predictions from the two types approaches. Experimental results show very competitive performance on both public data sets and an internal dataset from a major pharmaceutical company for both frequent and rare reaction types.

Ensembling two different types of approaches for retrosynthesis prediction is an interesting idea, which has different inductive biases, and combining them leads to a performant and robust approach, especially on rare reaction types. Although the reviewer appreciates the competitive performance of the proposed approach, the novelty of the proposed approach is very limited as ensembling has been widely studied in machine learning literature.

(Remarks on code availability)

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
I thank the authors for the substantial revision. The authors have broadened model comparisons beyond prior baselines, provided evidence for why the ensemble performs well, and conducted a large-scale chemist-led study of hallucination frequency. The new filter combining feasibility and forward prediction, along with the updated search experiments on filtered and adversarially selected targets, further strengthen the work. The following minor suggestions remain:

1. The manuscript describes NeuralLoc as predicting template classes and atom-level localization scores, but the actual process of converting these predictions into reactant structures is hard to follow. How exactly do the predicted reaction centers constrain template matching. How are atom correspondences mapped between template and product. What graph edits are performed, and how are multiple template applications handled beyond ranking. The paper would be much clearer with a step-by-step workflow or at least a concrete example showing one full inference pass.

2. The multi-step evaluation focuses heavily on solve rate and route diversity, but these don't tell me whether a chemist would actually want to run the proposed route. Things like total yield, step count, and whether the reactions are known to work at scale matter in practice. The feasibility filter and expert checks on individual steps are useful, but they don't add up to a picture of route-level viability. Adding estimated overall yields or route success probabilities would give a better sense of whether RetroChimera's outputs are practically useful, not just computationally correct.

3. The paper argues that ensembling diverse models is key to RetroChimera's performance, but only two models are included in the ensemble. To support this claim, it would help to test additional methods with different architectural assumptions. For example, RetroExplainer adopts a molecular assembly formulation for retrosynthesis, using energy-based decision curves to identify reaction centers and plan synthesis routes. The authors mention in the response letter that they couldn't reproduce RetroExplainer's results due to checkpoint issues, but the GitHub repo (https://github.com/GeorgeZhang-dev/RetroExplainer) has since been updated with working checkpoints.As shown in the updated comparative performance, RetroExplainer is the leading method amongst the existing editing methods. Considering to including it in the proposed ensemble model is necessary. I'd suggest including RetroExplainer as both a standalone baseline and as an additional ensemble member in their proposed model. Also, some baselines appear in the USPTO-50K results but are missing from USPTO-FULL and Pistachio. The comparison would be more complete if all methods were evaluated on all benchmarks.

(Remarks on code availability)

Blue = reviewer comments

Black = our response

Dear Reviewers,

Thank you for your feedback on our manuscript!

In the last six months, we invested considerable effort into updating the paper, addressing all reviewer comments, as well as generally improving many of the lines of experiments.

On a high level, changes to the manuscript can be divided into five groups:

- We have added comparisons to several recent retrosynthesis AI models, including large language models/reasoning models in addition to previous baselines
- As advised by Referee #1, we analysed the strengths of RetroChimera's submodels, and found they excel at predicting different reaction types (Figure 3cd, Extended Data Figure 12). This provides strong evidence behind why RetroChimera performs so well.
- We scaled up analysis by expert chemists: on top of the 600 entry pairwise comparison from the previous version of the manuscript, we added a separate 6000 entry study of individual model outputs, mapping out hallucination frequency and issues for different models (Figure 3f (right), Extended Data Figure 14). This is the first in-depth chemist-led study of model behaviour on that scale, taking stock of the last decade of research in the field. We anticipate this analysis will shape AI model development in the area for the coming years.
- Using data collected for the above, we extended analyses of using filtering models, and built a more advanced filter utilizing a combination of feasibility and forward filtering (Figure 4b).
- We updated the search experiments, which now include both filtered search experiments where we use the aforementioned filter model (Figure 4c), as well as experiments on very difficult targets that were adversarially selected to expose model issues (Figure 4d). We also included a very detailed analysis of the latter from the chemistry perspective (Extended Data Tables 3-5, Extended Data Figures 15-24).

We believe that the revision constitutes a significant upgrade. Below we address specific points raised by both reviewers.

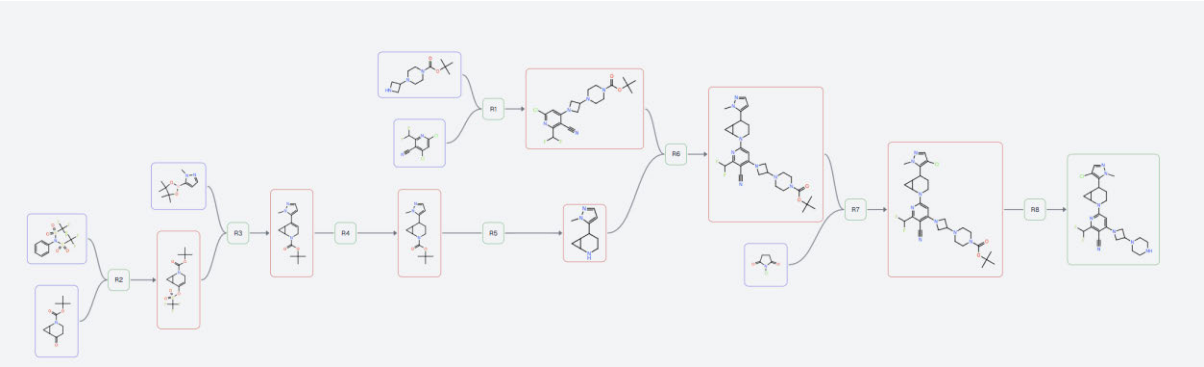
Referee #1

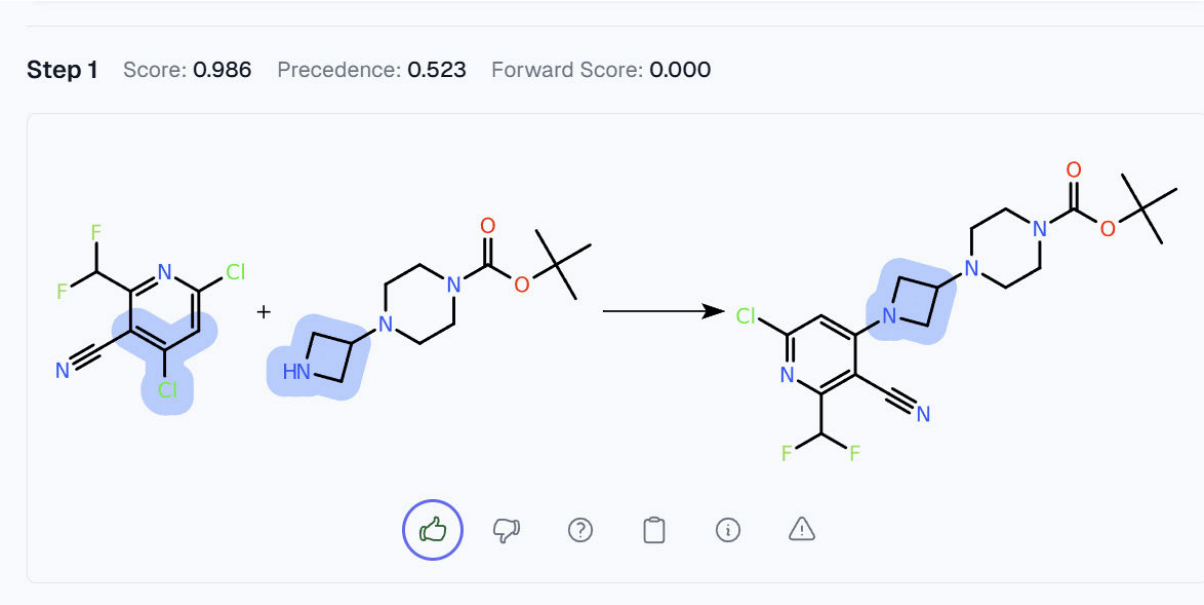
1. Ease of use. In the provided examples, the authors only show workflows for single-step and multi-step predictions. The internal computational process appears largely opaque and uncontrollable for chemists without an AI background.

We have built an initial system for querying the model, which allows to inspect unfiltered outputs from RetroChimera, [redacted]

The demo, for reviewing purposes only, [redacted]

Below you can see some screenshots from the app we have been building for inspecting and rating the predictions.





2. Framework scalability. As an ensemble learning framework, RetroChimera should be highly extensible to accommodate future methodological updates. The authors are encouraged to design interfaces that support the integration of additional models.

Furthermore, the claim that the framework can incorporate reaction databases and other external resources is intriguing and potentially impactful. A more detailed explanation of how this integration could be practically realized would be valuable.

Integrating more models into the ensemble is facilitated via syntheseus, which defines a clear model interface and a script to save model outputs. As seen in the RetroChimera codebase (retrochimera/retrochimera/cli/optimize_ensembles.py), we do not hard-code the choice of models; any models built on top of syntheseus can be used. Also refer to Extended Data Figure 9, where we ensemble many different model pairs, and Extended Data Figure 11, where we show results ensembling up to 10 models (the later figure we extended as part of this revision – in the original submission we only had results ensembling large sets of models on USPTO, whereas now we also have results on Pistachio). Together, this shows our framework is versatile and extensible.

Building on top of syntheseus enables RetroChimera (and our ensembling framework in general) to plug into a rich existing ecosystem of retrosynthesis methods developed on top of it over the last few years. For further context, see our response to point 7.

Integrating database search could be done by treating the search results as a "pseudo-model". If one believes the database search is always more reliable than a learned prediction, the ensembling weight for such pseudo-model could be hard-coded to a large value. Otherwise, it could be tuned jointly with all other weights, given a "source of truth" that is higher quality than the lookup database itself.

3. Validity of predictions. The authors state that RetroChimera is able to predict uncommon reactions. Have these predictions been experimentally validated, or are they simply deemed plausible from a chemical perspective?

We would argue that expert plausibility assessment is already a strong signal, and one that can be scaled to a much larger volume than possible for experimental validation, which depends also on many other factors that are unrelated to model quality. In the revised manuscript, we have extended this direction: on top of the pairwise comparison study that involved ~600 entries from experts, we also run a study where experts are shown a single prediction and mark it as plausible or hallucinated, where we gathered over 6000 entries (Figure 3f (right)). We also ran expert investigations on predictions for highly difficult products (Figure 4d). Together, this constitutes an expert evaluation on unprecedented scale, which we argue gives more insight that e.g. running lab syntheses for 10 targets.

4. Choice of models. As an ensemble framework, incorporating more models might further improve performance. A natural question arises: can the proposed scoring function adaptively filter out noise introduced by poorly performing models, such as low-confidence or evidently incorrect predictions? Although this study integrates two models—one template-based and one de novo—the authors should provide

theoretical justification and quantitative evidence of the complementarity between the two, as well as assess potential redundancy. For example, are there cases where specific model types perform better for certain reaction categories?

We thank the reviewer for the great suggestion – in the revised manuscript, we include anew analysis of model accuracies split by reaction type (Figure 3cd). We find that, indeed, the template-based model and the de novo model excel at different reaction types. RetroChimera recovers the strengths of both models, despite not knowing a priori which submodel is the one it should "trust more" for a particular input, because that would depend on the reaction type of the ground-truth answer. This is strong evidence that the models we include into RetroChimera are indeed complementary.

Regarding noise from poorly-performing models: such models should be downweighted in ensembling, and this is what we see empirically in our study of pairwise model combinations in Extended Data Figure 9. Equally, we find that even ensembling a relatively weak model with a stronger one does improve results over the stronger model by itself.

5. Comparison with recent methods. The authors should compare RetroChimera with more recently developed large-scale pretrained retrosynthesis models that report strong performance, such as RXNGraphormer (Nature Machine Intelligence, 2025) and RSGPT (Nature Communications, 2025). In addition, other representative deep learning methods like RetroExplainer (Nature Communications, 2023) should be included for a more comprehensive evaluation.

We took a closer look at the methods mentioned by the reviewer. We refer to them one-by-one below:

RSGPT

The RSGPT paper quotes very impressive results due to large-scale pretraining. To properly compared to their model, we downloaded the released code and weights.

First, it is worth noting that the released code does not run out-of-the-box due to environment issues, missing config files, and paths pointing to local (unreleased) files. This has been raised by the community in GitHub Issues, but we found the authors' responses not sufficiently clear. We were able to run the model after several hours of engineer time; this alone is a barrier to real-life adoption.

To compare, we focused on the RSGPT checkpoint trained on USPTO-FULL, as this is the largest and most diverse dataset studied in this work (also, the results on USPTO-50K and USPTO-MIT used reaction templates extracted from USPTO-FULL in pretraining, and thus leak information to the model, whereas the result on USPTO-FULL is free from this caveat). RSGPT utilized a version of USPTO-FULL released by the authors of the RetroXpert model (referred to as USPTO-FULL-RX below), which is different from the one we use (referred to as USPTO-FULL-RC). Therefore,

we ran a comparison of the two models on the train and test sets of USPTO-FULL-RX, test set of USPTO-FULL-RC, and our time-split Pistachio test set based on reactions published in 2024; for each of these datasets we subsampled to the first 3200 reactions to understand rough performance trends (note that even with this relatively small sample, the standard-error-of-the-mean should be no larger than 50%/sqrt(3200) ≈ 1%, so deviations larger than a few % are statistically significant).

We observe the following top-1 accuracies:

Dataset	RetroChimera (USPTO-FULL)	RSGPT (USPTO-FULL)
USPTO-FULL-RX (Train)	74.3%	22.3%
USPTO-FULL-RX (Test)	72.2%	76.0%
USPTO-FULL-RC (Test)	52.5%	28.4%
Pistachio-2024 (Test)	40.3%	18.2%

The results we see from RetroChimera are exactly as expected:

- accuracy on USPTO-FULL-RC matches the one we report in Extended Data Table 2 (modulo a small deviation due to subsampling)
- accuracy on both folds of USPTO-FULL-RX is inflated because the random data split is different from the one used in USPTO-FULL-RC (thus approximately 80% of any fold of the former is in the training set of the latter)
- accuracy on Pistachio-2024 is reasonable, but lower than achieved by RetroChimera trained on Pistachio directly

The results from RSGPT we cannot explain: we see excellent performance on USPTO-FULL-RX test set, and poor on anything else. It is possible that we have not run the model correctly (after all, we needed to fill in some missing details to run it), but then results should be low overall. This also cannot be explained by overfitting to USPTO-FULL-RX as then results on the train set would be good. As things stand, we have no hypothesis that would explain the empirical results in the table above.

In the end, the closest proxy for real-world use is Pistachio-2024, as that is cleaner than USPTO-FULL, and also time-split from it. We see low results from RSGPT there as well, so released checkpoints appear not practically useful as they stand.

Finally, even if RSGPT's result could be reproduced, it leverages large-scale pretraining, which we view as a promising direction completely orthogonal to our work. We foresee that in the future RetroChimera could be further improved by pretraining each submodel on synthetic data (which itself could come from RetroChimera 1, as opposed to brute-force template application as in RSGPT, allowing the current model to slingshot its next version towards even stronger performance).

RetroExplainer

Similarly to RSGPT, we made significant effort trying to run the model and reproduce the results. We found the provided checkpoints don't load out-of-the-box, because the leaving group library is missing; regenerating it from scratch using the provided code results in shape mismatch with respect to the checkpoint file. Therefore, for this model we also could not verify its performance.

We did include RetroExplainer in Extended Data Table 1 for reference, even though this goes against a philosophy we applied to *most* baseline comparisons in our work, which is to compare to results that were fairly reproduced through syntheseus to make sure the setup is consistent.

RXNGraphormer

We compared the results quoted in the paper to our, and found they are reported to be much lower than those of RetroChimera.

RetroDFM-R

Finally, we have added a comparison to RetroDFM, a very recent reasoning language model-type approach, by taking the numbers reported by the authors, in Extended Data Table 2.

6. Model interpretability. It would strengthen the work if the authors could provide insights into why the model selects a particular step (e.g., which bond is broken/formed, or which substructure is chosen as the reaction center). Beyond interpretability, such analysis may reveal novel patterns. For instance, RetroChimera performs well in low-data and out-of-distribution scenarios—does this arise from the discovery of higher-order "meta-rules" or from leveraging complementary heuristics, thereby creating a kind of "shortcut" in retrosynthetic reasoning?

In the newly added analysis, we found RetroChimera performs well directly because different models have varying capability of recalling different reaction types (see point 4). Therefore, for rare reactions, we believe results suggest that RetroChimera's submodels already partly have that ability, but it is inconsistent across different chemistries, and ensembling can resurface and strengthen these capabilities.

Regarding interpretability, in the revision we also added in-depth analysis of routes constructed by RetroChimera and other models (Extended Data Table 5, Extended Data Figures 15-24).

7.Multi-step planning and pathway search. The authors integrate RetroChimera into syntheseus for multi-step planning. Compared with other available methods, what are the specific advantages of syntheseus? A discussion of this choice would provide important context.

We designed syntheseus to allow plug-and-play experimentation, where single-step models can be arbitrarily combined with various search algorithms. Syntheseus has

clean reimplementations of the commonly used search algorithms (Retro* and MCTS variants). It has been picked up by the research community, with many papers and models building on it, and we intend to continue to develop and maintain the package going forward. Therefore, it makes sense to embed RetroChimera into that ecosystem. We are also now working towards making RetroChimera the default single-step model in syntheseus.

Referee #2

1. Ensembling two different types of approaches for retrosynthesis prediction is an interesting idea, which has different inductive biases, and combining them leads to a performant and robust approach, especially on rare reaction types. Although the reviewer appreciates the competitive performance of the proposed approach, the novelty of the proposed approach is very limited as ensembling has been widely studied in machine learning literature.

Thank you for acknowledging the strong performance of our model, but we want to emphasize that we are not just doing benchmark maximization in this work. Please also see our response to Reviewer 1. Whilst we agree that ensembling is an established concept in general ML, it has not yet been studied with great success in the chemical synthesis prediction domain.

Ensembling is also only part of our ML contribution. In particular:

1) We introduce the first Graph Neural Network (GNN) model able to process reaction templates. This requires a new GNN architecture that is able to process logical formulae on nodes (as in reaction SMARTS), as well as a novel localization component of the model that is able to predict the reaction centers. This model alone achieves state of the art in the edit/template-based category.
2) We drastically improved the performance of Transformer models on this task, using an upgraded attention mechanism and novel beam search. This model also achieves state of the art in its category (template-free/copy-based).
3) We provide an in-depth analysis of the benchmarking practices of the past years of AI synthesis planning and reveal critical shortcomings of past state of the art works, which we rectify with our expert-led analysis.

Since our initial submission, we also significantly strengthened both the ML experiments as well as the chemist-driven analysis of the models, with more than 6000 additional reactions manually annotated. This allows additional insight into the models at unprecedented depths, and showcases the (often orthogonal) failure

modes in particular of the baseline models. It also highlights the challenge of the unchecked application of search solve rate alone, which most prior work has been doing. We anticipate our analyses will shape model development and benchmarking in this important area for years to come and will allow the community to build on our results. Finally, we have added additional evaluation results using in-house data from a second pharmaceutical company.

Dear Reviewers,

Please find our response below:

Referee #1 (Remarks to the Author):

I thank the authors for the substantial revision. The authors have broadened model comparisons beyond prior baselines, provided evidence for why the ensemble performs well, and conducted a large-scale chemist-led study of hallucination frequency. The new filter combining feasibility and forward prediction, along with the updated search experiments on filtered and adversarially selected targets, further strengthen the work. The following minor suggestions remain

We want to thank the reviewer for a constructive assessment of our work and the additional suggestions, which have helped to improve the paper.

1. The manuscript describes NeuralLoc as predicting template classes and atom-level localization scores, but the actual process of converting these predictions into reactant structures is hard to follow. How exactly do the predicted reaction centers constrain template matching. How are atom correspondences mapped between template and product. What graph edits are performed, and how are multiple template applications handled beyond ranking. The paper would be much clearer with a step-by-step workflow or at least a concrete example showing one full inference pass.

We have substantially updated Figure 2 to make the overall flow clearer, and have further provided 3 additional, concretely worked examples for the ensemble aggregation works in the extended figures and supporting information. These shows explicitly how 1) the templates are preselected through classification, 2) how the templates are localized within the target (via the target-template assignment matrix, indicated by atoms maps, 3) how the classifcation and localization score are combined to rank the final reactants. The ensemble figures them show how the the predictions are then aggregated within the ensemble.

2. The multi-step evaluation focuses heavily on solve rate and route diversity, but these don't tell me whether a chemist would actually want to run the proposed route. Things like total yield, step count, and whether the reactions are known to work at scale matter in practice. The feasibility filter and expert checks on individual steps are useful, but they don't add up to a picture of route-level viability. Adding estimated overall yields or route success probabilities would give a better sense of whether RetroChimera's

outputs are practically useful, not just computationally correct.

We agree that yield prediction is important, however, it is an unsolved topic in the field under active research, and for yield prediction we would refer to existing other work (e.g. by Schwaller or Madzhidov). Our search framework Syntheseus (as well as other search frameworks, such as AIZynthfinder) can incorporate any numerical hard numbers like number of steps, reagent cost and availability, tree convergence, as well as estimates, like yield or reaction feasibility or scale-up potential into the route scoring. The end user chemist then needs to decide what criteria they want to incorporate.

Route-level viability assessment is done in the new qualitative study in the paper, see Extended Data Tables 3 and 4, which demonstrate that when evaluated via stricter route level metrics, where a single failed step invalidates a whole route, the commonly used baselines perform much weaker than our new Ensemble.

3. The paper argues that ensembling diverse models is key to RetroChimera's performance, but only two models are included in the ensemble. To support this claim, it would help to test additional methods with different architectural assumptions. For example, RetroExplainer adopts a molecular assembly formulation for retrosynthesis, using energy-based decision curves to identify reaction centers and plan synthesis routes. The authors mention in the response letter that they couldn't reproduce RetroExplainer's results due to checkpoint issues, but the GitHub repo (https://github.com/GeorgeZhang-dev/RetroExplainer) has since been updated with working checkpoints.As shown in the updated comparative performance, RetroExplainer is the leading method amongst the existing editing methods. Considering to including it in the proposed ensemble model is necessary. I'd suggest including RetroExplainer as both a standalone baseline and as an additional ensemble member in their proposed model. Also, some baselines appear in the USPTO-50K results but are missing from USPTO-FULL and Pistachio. The comparison would be more complete if all methods were evaluated on all benchmarks.

Please note that on USPTO-50k we have demonstrated ensembling of up to 10 models in one Ensemble, and on Pistachio, ensembles of up to 5 models (see Extended Data Figure 6).

We agree that RetroExplainer is a powerful model; it was already cited in the earlier version of our manuscript. Thank you for providing the pointer to the updated checkpoint for RetroExplainer, which was made public only after our resubmission. We also note that the RetroExplainer model checkpoint was trained on a different version of USPTO-Full (see discussion on this dataset in the method section), which makes a 1:1 comparison difficult. Retraining the model at this point would require significant additional computational and experimental cost. We thus only include models in that

table that have been tested on the exact USPTO-FULL split we have been using, which is taken from our reference baseline RSMILES.

In our present work, our main goal is to establish the principled concept of ensembling synthesis planning models. Given that we have demonstrated on several datasets and several baseline models that our ensembling approach benefits from adding more models, we anticipate that adding strong models like RetroExplainer will almost certainly improve the overall Ensemble. However, we leave this exploration for future work.