nature portfolio

Peer Review File

Exploration and validation of large language models as tools for molecular optimization

Corresponding Author: Dr Gerhard Hessler

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)
This paper describes a significant and ambitious study in which novel molecules were designed using many-shot in-context learning with large language models (LLMs), and then synthesised and evaluated. Of particular note are the differences observed through prompt design approaches (rule-based and non-rule-based) and the successful generation of highly active molecules in the RIPK1 and cathepsin A case studies. These findings strongly suggest the potential for applications in drug discovery and are expected to have a significant impact on the field. However, the presentation and discussion could be improved considerably. To enhance the paper's clarity and persuasiveness, I strongly recommend revising and strengthening the following points.

1)The term 'novel (predicted) active compounds' is inaccurate and misleading. The term 'novel' is currently used to signify compounds outside the internal SAR set, but does not confirm whether they are globally unknown. Furthermore, the distinction between predicted and experimentally validated active compounds is ambiguous. A quantitative assessment of true novelty should be added, such as duplicate verification against external databases and distribution comparisons using metrics such as Tanimoto distance or Jensen–Shannon divergence (JSD). The expression should be revised to be more restrictive and accurate.

2)The main text only shows the generated results, with the rule details described only in the supplementary material. A summary correlating the rules with the generated results should be added to the main text to help readers understand the differences in prompt design intuitively.

3)The supplementary figures (Figure S2: distribution of distances from queries; Figure S3: diversity within the generated set; and Figure S9: comparison of SAR prediction performance) all present crucial results that strengthen the main text's discussion, but are underutilised. These figures should be incorporated into the main text to visually support the key claims (the effectiveness of rule-based approaches and the SAR learning capability of LLMs).

4)The candidate count of 2,714/754 reflects the results after deduplication, removal of PAINS and unnecessary fragments, and physicochemical filtering — though this is not explicitly stated in the main text. A brief description of the process in the text would enhance transparency. Additionally, mentioning the option of formalising PAINS removal into prompt design rules, for example, would clarify the hybrid use of prompts and filters.

5)The paper only uses PCA (ECFP radius 3) for visualising structural similarity, omitting comparisons of physical properties such as molecular weight or logP. However, the prompt design and filters explicitly state constraints such as molecular weight, logP, TPSA, HBD/HBA, and so on. Therefore, the distribution of physical properties such as MW, logP, TPSA and HBD/HBA should be compared using histograms or KDEs with the prompt thresholds overlaid. Furthermore, 2D-KDEs (e.g. MW×logP) can intuitively demonstrate whether the generated molecules are within the project's acceptable range or are venturing into new regions.

6)The synthesizability of generated molecules is not discussed systematically. While the proportion of synthesizable molecules within the entire generation or quantitative evaluations like SAscore is not presented, molecules actually synthesised represent human-filtered selections. If drug discovery applicability is claimed, adding metrics and discussion regarding ease of synthesis is essential.

7)The standard evaluation metrics for generation models are not systematically organised into tables, and the results for validity (valid SMILES rate), uniqueness (proportion after duplicate removal) and diversity are described fragmentarily. Presenting these tables grouped by rule-based or non-rule-based approaches, or by target, would facilitate comparison with existing generation models (VAE, RNN, GAN, etc.) and make the information easier for readers to understand.

Overall, this paper presents important and meaningful research. However, addressing the points for improvement noted here would make the arguments more persuasive and improve the paper's overall completeness.

Reviewer #2

(Remarks to the Author)
In the present manuscript, the authors use a modern LLM with prompting and few shot learning to perform molecular design. Using two targets, early discovery scenarios are explored.

The LLM is used to sample initial molecules for further triage using two prompting strategies. Molecules are compared to the initial seeds.

A convincing computational chemistry strategy is then employed to narrow down the generated molecules and make a sub-selection. Then the virtual molecules are triaged and selected.

In both cases, the LLM could identify higher activity ligands as confirmed by synthesis.

Overall, I think this is an interesting study for the community and should be published after the below minor comments have been addressed.

Comments:
Can you check Structure VI in Scheme 2? I think there is a nitrogen missing.

It could be useful to report Tanimoto similarity of the generated molecules to the most similar seed molecules in the appendix.

I'd suggest to additionally cite https://arxiv.org/abs/2311.07361 which features an early investigation of language based molecular editing and bio-isosteric replacement.

Reviewer #3

(Remarks to the Author)
This paper explores using a large language model (LLM) to generate molecular analogs for drug discovery. The study begins with screening hits for two targets, RIP1K and Cathepsin A. Two different LLM strategies are used to generate a set of analogs from 10 RIP1K and 10 Cathepsin A screening hits. The generated molecules then undergo a series of standard computational chemistry filters. After computational prioritization, 32 RIPK1 inhibitors and 37 Cathepsin A inhibitors were synthesized, resulting in potency increases of 3 to 10 times over the initial “query compounds."

From the beginning, the authors make misleading claims about the ability of LLMs to “understand” chemistry. According to current literature, there is no evidence to support this. To support their claims, the authors cite several papers that only demonstrate an LLM performing simple tasks like calculating the length of a bond path through a chemical structure. Methods like ChemCrow and ChatMOF are essentially workflow orchestration tools at their core. These can be seen as natural-language versions of existing workflow tools such as PipelinePilot and KNIME.

On page 2, the authors state, “Thus, LLMs can develop an implicit understanding of chemical principles and relationships, enabling them to serve as powerful tools for molecular design.” This claim is fundamentally incorrect. While LLMs may learn the syntax of SMILES, there is no evidence that they understand the underlying chemistry. Statements like these can create false impressions for those outside the field.

The molecules designed by the LLM underwent thorough post-processing, filtering, and evaluation steps, which included traditional machine learning, structural filtering, computational chemistry methods (such as docking and scoring), and assessments of synthetic feasibility using human expertise. To improve synthetic efficiency, the final compounds synthesized consisted of the LLM-designed molecules and an expanded set of compounds created through combinatorial expansion using the building blocks proposed by the LLM. Given these points, it is difficult to determine whether the actual source of the improvements was the LLM or the human operators.

None of the designs appear to have gone through any kind of LLM-driven agentic process. The LLM generated a set of molecules, which were then filtered computationally and prioritized by humans. At best, the LLM served as an idea generator. Based on this, the work described in this manuscript does not seem to advance the current state of the art. Furthermore, the complete disagreement between the LLM scoring and the computational oracle in Figure S12 should cause any reader to pause. From this, it appears that the LLM is merely a random molecule generator. It remains unclear

whether the LLM is doing something that couldn't be achieved with other generative methods like REINVENT or CReM.

It should be noted that the comparison of machine learning models in Figure S9 does not reflect current best practices. Using a 95/5 train-test split is inappropriate. Additionally, proper statistical tests should be employed when comparing model performance.

There has been significant hype surrounding the use of LLMs in science. As practitioners, we must carefully evaluate results and exercise caution when discussing concepts like LLMs "understanding chemistry." I believe this paper leans toward hype and does not demonstrate that LLMs can perform at or above the current state of the art in computational drug discovery. Therefore, I cannot recommend publishing this paper.

Reviewer #4

(Remarks to the Author)
This paper employs large language models for iterative molecular generation and demonstrates successful validation across two target systems. Overall, it presents a well-structured application of molecular design. However, the algorithm and methodology do not exhibit substantial novelty. I recommend the authors address the following points:

1. The structural comparison between the query molecule and the generated molecules requires more in-depth analysis. My primary concern relates to the molecular diversity of the generated compounds and the corresponding expectations for their biological activity. For example, as noted in the paper, pyridine was identified through prompting as a key functional group, and its incorporation as a design rule may help enhance biological activity. However, while this scaffold may theoretically support activity, it also imposes substantial constraints on molecular novelty. Without additional design constraints or diversification strategies, any improvement in activity is likely to remain limited.The structural comparison between the query molecule and the generated molecules requires more in-depth analysis. My primary concern relates to the molecular diversity of the generated compounds and the corresponding expectations for their biological activity. For example, as noted in the paper, pyridine was identified through prompting as a key functional group, and its incorporation as a design rule may help enhance biological activity. However, while this scaffold may theoretically support activity, it also imposes substantial constraints on molecular novelty. Without additional design constraints or diversification strategies, any improvement in activity is likely to remain limited.

2. Clarity and organization of writing.The overall logical flow of the manuscript is somewhat difficult to follow. The Methods section, in particular, would benefit from a workflow diagram to visually clarify the procedural steps and improve readability.

3. Analytical data for synthesized compounds. Mass spectrometry data for the synthesized molecules should be provided. The purity of each compound should also be reported and ensured to exceed 95%.

4. Figure quality. The current figures are of relatively low quality, and several molecular structures are difficult to discern. It is recommended that the figures be revised and rendered at higher resolution to improve clarity.

5. Limitations in model-based binding affinity prediction. The authors employ the Claude 3 Sonnet model both as a language model and as a predictor of molecular binding affinity, using its estimated affinity values to guide iterative molecular generation. This introduces a limitation: because the algorithmic strategy is not fundamentally novel, the work resembles an extension of existing methods. Moreover, despite iterative refinement based on known active compounds, the optimized molecules do not show a linear or clearly incremental improvement in biological activity. This suggests either insufficient feedback-driven correction during generation or inherent limitations in the affinity-prediction approach. The authors should consider evaluating alternative external binding-affinity predictors or incorporating additional molecular pharmacodynamic evaluation methods.

6. Compound diversity analysis. The generated molecules should be evaluated using a compound similarity matrix to assess structural diversity under the constrained generation conditions, rather than limiting the analysis to the small subset of synthesized molecules.

7. Selection criteria for synthesis. Given the large number of generated candidates, the manuscript should clarify how the final molecules were selected for synthesis. Were criteria such as synthetic accessibility, ADMET predictions, or drug-likeness metrics applied? Incorporating such rational selection principles would strengthen the study.

Version 1:

Reviewer comments:

Reviewer #3

(Remarks to the Author)
The authors have appropriately addressed my comments and those of the other reviewers. This paper is suitable for publication in its current form.

Reviewer #4

(Remarks to the Author)
my questions are almost answered. I think it could be accepted in this version.

Reviewer #5

(Remarks to the Author)
The work presented by Grebner et al., seeks to fill a gap in knowledge regarding small-molecule drug discovery and the utilization of Large Language Models (LLMs). The work is timely and unique as it seeks to address key aspects of how LLMs may impact decision making, especially regarding generative AI for compound structures. The drug discovery community is highly interested in such work and continually seeking for ways to utilize LLMs to generate new directions for projects that are capable of directing for new compound design (instead of simply being reminded of old, bad ideas). The case studies of RIPK1 and Cathepsin A are good examples to highlight the nature behind these techniques and methodologies. My only question that would be useful to include in the revision would be related to conventional structure-guided design. Especially for RIPK1, where X-ray structures are known, how are results from the LLM generated molecules different or similar to those that may be arrived at through rational structure-guided design? Do the LLMs provide unique perspectives that ordinarily may be missed? Can similar points be said for Cathepsin A? The answer to this general question would give important perspective to how LLMs are able to add to drug discovery beyond what would be considered typical "human" reasoning.

Version 2:

Reviewer comments:

Reviewer #5

(Remarks to the Author)
The authors replies to the comments are sufficient and the manuscript it ready for publication.

Reply to reviewer's comments: "Large Language Models as Tools for Molecular Optimization: Experimental Validation in Two Case Studies"

We thank the reviewers for their comprehensive analysis of our manuscript. We appreciate the time they have dedicated to reviewing our work. We have made substantial revisions that significantly strengthen the manuscript and provide a point-by-point response to each concern raised, outlining specific revisions and analysis we have taken to address these questions. We believe that the manuscript provides an accurate analysis of the potential of LLM methods in molecular design for drug discovery and is to our knowledge the **first prospective experimental validation** of LLM-generated molecules in real drug discovery projects, with synthesis and biological testing of compounds across two targets.

Reviewer #1 (Remarks to the Author):

This paper describes a significant and ambitious study in which novel molecules were designed using many-shot in-context learning with large language models (LLMs), and then synthesized and evaluated. Of particular note are the differences observed through prompt design approaches (rule-based and non-rule-based) and the successful generation of highly active molecules in the RIPK1 and cathepsin A case studies. These findings strongly suggest the potential for applications in drug discovery and are expected to have a significant impact on the field. However, the presentation and discussion could be improved considerably. To enhance the paper's clarity and persuasiveness, I strongly recommend revising and strengthening the following points.

1)The term 'novel (predicted) active compounds' is inaccurate and misleading. The term 'novel' is currently used to signify compounds outside the internal SAR set, but does not confirm whether they are globally unknown. Furthermore, the distinction between predicted and experimentally validated active compounds is ambiguous. A quantitative assessment of true novelty should be added, such as duplicate verification against external databases and distribution comparisons using metrics such as Tanimoto distance or Jensen–Shannon divergence (JSD). The expression should be revised to be more restrictive and accurate.

We clarified the discussion and extended our analysis respectively. To illustrate the novelty of generated molecules, we compared all synthesized molecules against the ChEMBL36 database. None of the compounds show similarity above 0.8 (Morgan fingerprints radius 2 with 2048 bits) and only one compound shows similarity above 0.7. We added figures for illustration to the supporting information and clarified the usage of the term "novel" in the manuscript.

2)The main text only shows the generated results, with the rule details described only in the supplementary material. A summary correlating the rules with the generated results should be added to the main text to help readers understand the differences in prompt design intuitively.

The method section "Skeleton of main prompts" now contains information about the used prompts for the different design scenarios. We also added extended descriptions of the rules

and settings of the prompt in the main text and added additional full prompts to the SI.

3)The supplementary figures (Figure S2: distribution of distances from queries; Figure S3: diversity within the generated set; and Figure S9: comparison of SAR prediction performance) all present crucial results that strengthen the main text's discussion, but are underutilised. These figures should be incorporated into the main text to visually support the key claims (the effectiveness of rule-based approaches and the SAR learning capability of LLMs).

Thanks a lot for the comments. We moved Figure S2 and S3 to the main text and extended the discussion. As the main purpose of the study was to evaluate compound design and chemical space sampling of LLMs and, additionally, the number of figures is already high, we believe that Figure S9 is better suited in the SI.

4)The candidate count of 2,714/754 reflects the results after deduplication, removal of PAINS and unnecessary fragments, and physicochemical filtering — though this is not explicitly stated in the main text. A brief description of the process in the text would enhance transparency. Additionally, mentioning the option of formalising PAINS removal into prompt design rules, for example, would clarify the hybrid use of prompts and filters.

We extended and added more details to the descriptions of the steps as well as referencing them directly in the main text. The idea of formalizing filters such as PAINS directly in the prompts is very interesting and should be explored in future studies. We added a comment in the text and in the conclusion section.

5)The paper only uses PCA (ECFP radius 3) for visualising structural similarity, omitting comparisons of physical properties such as molecular weight or logP. However, the prompt design and filters explicitly state constraints such as molecular weight, logP, TPSA, HBD/HBA, and so on. Therefore, the distribution of physical properties such as MW, logP, TPSA and HBD/HBA should be compared using histograms or KDEs with the prompt thresholds overlaid. Furthermore, 2D-KDEs (e.g. MW×logP) can intuitively demonstrate whether the generated molecules are within the project's acceptable range or are venturing into new regions.

We added additional analysis plots to the supporting information using physicochemical properties (MW and log) as well as commonly used metrics such as QED and SASCore. The new Figures S8 and S12 illustrate that a similar property range is occupied for design molecules.

6)The synthesizability of generated molecules is not discussed systematically. While the proportion of synthesizable molecules within the entire generation or quantitative evaluations like SAscore is not presented, molecules actually synthesized represent human-filtered selections. If drug discovery applicability is claimed, adding metrics and discussion regarding ease of synthesis is essential.

Thanks a lot for the comment. A more thorough analysis of the distribution of QED and SAScore helps to judge the quality of the generated molecules. We added a plot of histograms and distributions for generated as well as selected molecules for RIPK1 and CatA (Supporting Figure S8). The overall distribution overlaps very well with the distribution of the starting points of the designs. The analysis demonstrates that LLM-generated molecules exhibit reasonable SAScore distributions, with mean values of ~7±0.5 for both RIPK1 and CatA. The reference compounds cover a very similar range in both cases. We have also added corresponding discussion in the main text (page 14) highlighting that "the overall distribution of synthetic accessibility scores closely resembles that of the starting points,

demonstrating that the LLM generates molecules within synthetically accessible chemical space." This systematic evaluation confirms that our LLM-based approach produces molecules suitable for practical drug discovery applications.

7)The standard evaluation metrics for generation models are not systematically organised into tables, and the results for validity (valid SMILES rate), uniqueness (proportion after duplicate removal) and diversity are described fragmentarily. Presenting these tables grouped by rule-based or non-rule-based approaches, or by target, would facilitate comparison with existing generation models (VAE, RNN, GAN, etc.) and make the information easier for readers to understand.

Thanks a lot for the comment. Information about typical evaluation parameters was already included in a formatted way in the supporting information. To make it clearer and easier to read, we moved Table S1 to the main text and included it in the discussion more clearly.

Overall, this paper presents important and meaningful research. However, addressing the points for improvement noted here would make the arguments more persuasive and improve the paper's overall completeness.

Reviewer #2 (Remarks to the Author):

In the present manuscript, the authors use a modern LLM with prompting and few shot learning to perform molecular design. Using two targets, early discovery scenarios are explored.

The LLM is used to sample initial molecules for further triage using two prompting strategies. Molecules are compared to the initial seeds.

A convincing computational chemistry strategy is then employed to narrow down the generated molecules and make a sub-selection. Then the virtual molecules are triaged and selected.

In both cases, the LLM could identify higher activity ligands as confirmed by synthesis.

Overall, I think this is an interesting study for the community and should be published after the below minor comments have been addressed.

Comments:
Can you check Structure VI in Scheme 2? I think there is a nitrogen missing.

Thanks a lot for identifying this error. The structure is corrected.

It could be useful to report Tanimoto similarity of the generated molecules to the most similar seed molecules in the appendix.

Thanks for the comment. We added supporting figures for the distribution of Tanimoto similarity values to the entire ChEMBL36 database to discuss the novelty of generated molecules. In addition, a we added the distribution the highest similarity of designed compounds to each query compound.

I'd suggest to additionally cite https://arxiv.org/abs/2311.07361 which features an early investigation of language based molecular editing and bio-isosteric replacement.
We added the reference to our introduction.

Reviewer #3 (Remarks to the Author):

This paper explores using a large language model (LLM) to generate molecular analogs for drug discovery. The study begins with screening hits for two targets, RIP1K and Cathepsin A. Two different LLM strategies are used to generate a set of analogs from 10 RIP1K and 10 Cathepsin A screening hits. The generated molecules then undergo a series of standard computational chemistry filters. After computational prioritization, 32 RIPK1 inhibitors and 37 Cathepsin A inhibitors were synthesized, resulting in potency increases of 3 to 10 times over the initial "query compounds."

From the beginning, the authors make misleading claims about the ability of LLMs to "understand" chemistry. According to current literature, there is no evidence to support this. To support their claims, the authors cite several papers that only demonstrate an LLM performing simple tasks like calculating the length of a bond path through a chemical structure. Methods like ChemCrow and ChatMOF are essentially workflow orchestration tools at their core. These can be seen as natural-language versions of existing workflow tools such as PipelinePilot and KNIME.

On page 2, the authors state, "Thus, LLMs can develop an implicit understanding of chemical principles and relationships, enabling them to serve as powerful tools for molecular design." This claim is fundamentally incorrect. While LLMs may learn the syntax of SMILES, there is no evidence that they understand the underlying chemistry. Statements like these can create false impressions for those outside the field.

Thanks a lot for the comment and for highlighting this important issue. We agree that terms such as "understanding" or "chemical intuition" can potentially be misleading to non-expert readers. Therefore, we have systematically revised the text to replace such terminology with a more accurate description of the LLM capabilities. We believe these revisions now accurately represent LLM capabilities without overstatement, and we appreciate the reviewer's call for scientific rigor in this emerging field.

The molecules designed by the LLM underwent thorough post-processing, filtering, and evaluation steps, which included traditional machine learning, structural filtering, computational chemistry methods (such as docking and scoring), and assessments of synthetic feasibility using human expertise. To improve synthetic efficiency, the final compounds synthesized consisted of the LLM-designed molecules and an expanded set of compounds created through combinatorial expansion using the building blocks proposed by the LLM. Given these points, it is difficult to determine whether the actual source of the improvements was the LLM or the human operators.

We respectfully disagree with the reviewer's assessment and would like to clarify the distinct methodologies employed for each target system.

For Cathepsin A, the reviewer's characterization is inaccurate. Of the 95 compounds initially suggested by the LLM, the selection process involved **only assessment of synthetic feasibility** with available building blocks—no additional molecular modeling filters, docking, or computational scoring were applied. Several of those (as mentioned in the paper) were

found to be synthesized already and confirmed biological activity (not part of the training and query data sets, hence unknown to the LLM). The final 10 LLM-designed molecules were synthesized and tested directly, achieving a 100% hit rate with 60% showing IC50 < 1 μM. In contrast, the combinatorial expansion using the same building blocks yielded only 22% of compounds with IC50 < 1 μM. This large difference provides evidence that the LLM's molecular generation capability extends well beyond random sampling.

For RIPK1, we employed a comprehensive molecular modeling post-processing workflow that represents standard practice in computational drug discovery. We emphasize that we do not position LLMs as a standalone solution to drug design challenges, but rather as a powerful augmentation to human expertise. The key innovation lies in enabling medicinal and computational chemists to interact with molecular generation through natural language—producing valid, reasonable, and biologically active compounds through intuitive, instruction-based guidance.

Regarding our post-processing workflow: The application of multi-stage filtering—including physics-based methods (docking, MM-GBSA scoring), structural filters (PAINS, unwanted fragments), physicochemical property filters, and expert assessment—represents established best practice in contemporary drug discovery, routinely applied to molecules from any generative source, whether virtual screening, traditional models (VAE, RNN, GAN), fragment-based design, or medicinal chemist proposals.

Rather than diminishing the LLM contribution, this workflow demonstrates seamless integration into existing drug discovery infrastructure. The key advancement is not replacement of established methods, but the introduction of an intuitive natural language interface that enables SAR-guided chemical space exploration, a capability unavailable in traditional generative approaches such as REINVENT, which require specialized computational expertise to operate effectively. Looking ahead, agentic frameworks would further extend these capabilities by incorporating additional non-LLM engines for molecular evaluation.

None of the designs appear to have gone through any kind of LLM-driven agentic process. The LLM generated a set of molecules, which were then filtered computationally and prioritized by humans. At best, the LLM served as an idea generator. Based on this, the work described in this manuscript does not seem to advance the current state of the art. Furthermore, the complete disagreement between the LLM scoring and the computational oracle in Figure S12 should cause any reader to pause. From this, it appears that the LLM is merely a random molecule generator. It remains unclear whether the LLM is doing something that couldn’t be achieved with other generative methods like REINVENT or CReM.

We appreciate the reviewer's observation regarding agentic processes. While we acknowledge that our study does not implement a fully autonomous LLM-driven agentic system, we respectfully clarify that this was not our primary objective.

Our research specifically aimed to address a fundamental question in the field: **Can LLMs effectively generate novel, synthetically accessible, and biologically active molecules for real-world drug discovery applications? And how can we steer compound design to a desired chemical spaces.** This question required experimental validation through synthesis and biological testing, a critical step that has been notably absent in many LLM-based molecular design literature to date.

Several key aspects distinguish our work from "merely a random molecule generator":

1. Unlike most computational studies that remain in silico, we synthesized and biologically tested compounds across two targets, demonstrating that LLM-generated molecules translate to real-world activity. The high hit rates (14/16 for RIPK1 pyrrolidine series and 10/10 for CatA) provide compelling evidence of non-random design capability.

2. Our results demonstrate that LLMs can effectively transfer SAR knowledge between chemical series—a sophisticated capability requiring pattern recognition across molecular structures. For CatA, the LLM successfully applied SAR from biaryl compounds to optimize pyrazolone-based molecules, despite these occupying distinct regions of chemical space (as shown in Figure 1). We further extended our retrospective analysis to two additional targets showing that classification tasks were significantly improved by providing context data in all cases.

3. We systematically demonstrated how different prompting strategies influence the exploration of chemical space (Figures 2-4), showing that LLMs can be guided to specific regions through natural language instructions—a capability fundamentally different from random generation.

4. The conversational interface enables medicinal chemists to express design constraints in human language rather than through complex programming or parameter tuning. This represents a significant advancement in accessibility and usability compared to existing tools like REINVENT or CReM, which require specialized computational expertise.

We have extended our analysis to further substantiate these points, adding comprehensive distributions of molecular properties (SAScore, QED, MW, logP) and similarity comparisons to ChEMBL36, starting points, and reference data confirming that the LLM generates molecules with drug-like properties and novel structures, which are not simply recapitulations of known compounds.

While we agree that integration into a fully agentic framework represents an exciting future direction (which we discuss in our conclusion), the current work provides the essential experimental foundation demonstrating that LLM-generated molecules can indeed yield active compounds—a prerequisite for any more advanced agentic implementation.

We thank the reviewer for raising this important point regarding the apparent disagreement between LLM predictions and computational oracle predictions in Figure S12. However, we believe this observation requires careful contextualization within the broader experimental outcomes. In addition, we extended our analysis and provided further validation data.

First, we must distinguish between prediction performance and generation capability. While Figure S12 shows moderate correlation between LLM and oracle predictions for CatA, the experimental validation demonstrates that the LLM's molecular generation capability is highly effective: all 10 LLM-designed CatA molecules were active (100% hit rate), with 60% showing IC50 < 1 μM. These experimental results provide direct evidence that the LLM successfully captured and applied SAR patterns, regardless of the quantitative prediction accuracy.

Second, the CatA dataset presents inherent modeling challenges that affect all computational methods, not just LLMs. As shown in Supporting Figures S1, S2, S4 and S16 (previously S12), classical descriptor-based machine learning models exhibit similar modest performance on this dataset, which features a narrow dynamic activity range and limited structural diversity. The fact that both LLM and traditional ML methods struggle with

quantitative prediction, yet the LLM still generates highly active molecules, suggests that the generation process leverages qualitative SAR patterns extracted from the context examples.

Third, we have now extended our retrospective analysis with further examples to demonstrate SAR capturing capabilities of the LLM. In four datasets (CatA, MMP8, RIPK1, GSK3β), the LLM classified active and inactive molecules with and without context data. In all cases, the performance significantly increased to above 95% correct classifications (see new supporting figure S3).

Fourth, the oracle ensemble approach is a deliberate methodological choice that leverages complementary molecular representations. LLMs process SMILES strings and learn sequential patterns, while classical ML uses explicit molecular descriptors (fingerprints, physicochemical properties, Mol2Vec embeddings). These fundamentally different representations capture distinct aspects of molecular structure. The disagreement between methods is not a weakness but an opportunity: consensus predictions from diverse representations provide more robust candidate selection than any single method alone. This ensemble strategy is well-established in machine learning and drug discovery.

The ultimate validation remains experimental. The high hit rates across both targets (87.5% for RIPK1 pyrrolidines, 100% for CatA) demonstrate that our combined LLM generation and ensemble selection workflow successfully identifies active compounds. The moderate correlation in Figure S16 (previously S12) does not undermine these experimental outcomes—rather, it highlights that molecular generation and activity prediction are distinct tasks, and the LLM excels at the former even when the latter is challenging. Supporting Figure S3 provides a more nuanced view of LLM capabilities across dataset characteristics.

It should be noted that the comparison of machine learning models in Figure S9 does not reflect current best practices. Using a 95/5 train-test split is inappropriate. Additionally, proper statistical tests should be employed when comparing model performance.

We thank the reviewer for this comment. We agree that a 95/5 split would be inappropriate for typical machine learning model evaluation. However, the in-context learning framework operates differently from traditional ML approaches, which necessitates a different evaluation strategy.

In traditional ML, models undergo iterative parameter optimization during training, making train-test separation critical to avoid overfitting. In contrast, LLMs in our in-context learning approach have frozen parameters and perform inference based solely on the examples provided in the prompt context. The LLM cannot "memorize" training data in the conventional sense—it must extract patterns from the provided examples in real-time during a single forward pass.

Given the limited dataset size (296 compounds for CatA), maximizing the number of examples available to the LLM was essential for effective pattern recognition. We systematically evaluated three data split scenarios: leave-one-out, 95/5, and 80/20. Importantly, all three approaches yielded qualitatively similar results, demonstrating that our findings are robust across different data partitioning strategies. The 95/5 split was selected as it provided the optimal balance between context richness and evaluation rigor for this specific application.

There has been significant hype surrounding the use of LLMs in science. As

practitioners, we must carefully evaluate results and exercise caution when discussing concepts like LLMs "understanding chemistry." I believe this paper leans toward hype and does not demonstrate that LLMs can perform at or above the current state of the art in computational drug discovery. Therefore, I cannot recommend publishing this paper.

We appreciate the reviewer's call for scientific rigor. We have carefully considered this feedback and revised our manuscript accordingly.

We have systematically revised the manuscript to remove anthropomorphic language such as "understanding chemistry" or "chemical intuition." These phrases have been replaced with precise descriptions of LLM capabilities, including "operate on chemical data," "identify patterns in structure-activity relationships," and "process molecular representations." We thank the reviewer for highlighting this important issue, and we believe the revised manuscript now accurately represents LLM capabilities without overstatement.

Regarding state-of-the-art and contribution: We would like to offer some context for evaluating our work. While numerous publications describe computational methods for molecule generation, prospective experimental validation remains relatively rare in the field. Many generative model studies report primarily computational metrics (validity, uniqueness, novelty) without extensive synthesis or biological testing.

In this context, our study provides:

- Synthesis and biological testing of designed compounds across two therapeutically relevant targets from different protein classes
- High experimental hit rates
- Demonstrated potency improvements of up to 100-fold over starting compounds
- Evidence of SAR transfer across distinct chemical scaffolds

We believe this level of prospective validation represents a meaningful contribution to the field, as experimental confirmation of generative design methods remains an important benchmark for practical applicability.

Beyond the experimental validation, our work highlights several capabilities of LLM contributions that complement existing generative methods:

1. **Natural language interface:** Medicinal chemists can guide molecular generation through intuitive instructions rather than parameter tuning or specialized programming, potentially broadening access to generative design tools.
2. **SAR transfer between scaffolds:** Our prospective applications demonstrate that the LLM successfully applied SAR information from one series to another as illustrated in particular for CatA in Figure 1.
3. **Interpretable reasoning:** The LLM can provide chemical rationale for proposed modifications (Supporting Table S3), offering a degree of transparency in the design process and allowing them to judge the proposed design suggestions.
4. **Flexible constraint specification:** Design rules can be expressed and modified in natural language without retraining, allowing for iterative refinement of design criteria.

We demonstrate through experimental validation that LLMs can serve as a useful addition to the drug discovery toolkit when appropriately integrated into established workflows. We believe this balanced perspective, supported by prospective experimental data, offers value to the community.

We have carefully revised the manuscript to ensure all claims are appropriately qualified and grounded in experimental evidence. We hope these clarifications address the reviewer's concerns, and we remain open to further suggestions for improving the scientific rigor of our presentation

Reviewer #4 (Remarks to the Author):

This paper employs large language models for iterative molecular generation and demonstrates successful validation across two target systems. Overall, it presents a well-structured application of molecular design. However, the algorithm and methodology do not exhibit substantial novelty. I recommend the authors address the following points:

1. The structural comparison between the query molecule and the generated molecules requires more in-depth analysis. My primary concern relates to the molecular diversity of the generated compounds and the corresponding expectations for their biological activity. For example, as noted in the paper, pyridine was identified through prompting as a key functional group, and its incorporation as a design rule may help enhance biological activity. However, while this scaffold may theoretically support activity, it also imposes substantial constraints on molecular novelty. Without additional design constraints or diversification strategies, any improvement in activity is likely to remain limited.

We thank the reviewer for this comment. Therefore, we have added a comparison of synthesized compounds versus all compounds in ChEMBL to demonstrate chemical novelty. We further moved Figure S2 and S3 from the supporting information to the main text. In this section, we compare the similarity of newly generated molecules as well as internal similarity of generated ensembles for different prompting strategies (non-rule and rule based) for CatA.

We agree with the reviewer that adding constraints for a substructure (i.e. core) will limit the design space. This is, however, sometimes desirable, in particular when optimizing a new or existing core motif. Demonstrating the different prompting strategies will enable users and designers to move and adapt the compound generation to any specific use case. We extended the discussion about this fact in the main text to highlight that both scenarios are important. We wanted to illustrate the key feature of guiding the design with human readable conversational input.

2. Clarity and organization of writing. The overall logical flow of the manuscript is somewhat difficult to follow. The Methods section, in particular, would benefit from a workflow diagram to visually clarify the procedural steps and improve readability.

Thanks a lot. We carefully reviewed the manuscripts and clarified the writing. In our view, it is now much more streamlined and easy to follow.

3. Analytical data for synthesized compounds. Mass spectrometry data for the synthesized molecules should be provided. The purity of each compound should also be reported and ensured to exceed 95%.

Additional data for the analytical results were added to the experimental section.

4. Figure quality. The current figures are of relatively low quality, and several molecular structures are difficult to discern. It is recommended that the figures be revised and rendered at higher resolution to improve clarity.

Thanks a lot. Figures have been revised and updated

5. Limitations in model-based binding affinity prediction. The authors employ the Claude 3 Sonnet model both as a language model and as a predictor of molecular binding affinity, using its estimated affinity values to guide iterative molecular generation. This introduces a limitation: because the algorithmic strategy is not fundamentally novel, the work resembles an extension of existing methods. Moreover, despite iterative refinement based on known active compounds, the optimized molecules do not show a linear or clearly incremental improvement in biological activity. This suggests either insufficient feedback-driven correction during generation or inherent limitations in the affinity-prediction approach. The authors should consider evaluating alternative external binding-affinity predictors or incorporating additional molecular pharmacodynamic evaluation methods.

We thank the reviewer for the comment. This is a very important and valid suggestion. As we outline in the publication, an agentic framework where the LLM compound generation would be integrated with a more sophisticated affinity prediction would be a future application scenario. However, the intention of the present study was to demonstrate for the first time that molecules designed with LLMs would actually translate into valid, reasonable, and biologically active compounds. The key takeaway of our study is that scientists can interact with a design engine using human language and modify the expected outcome using human readable and easy configurable prompts. We agree that improved binding affinity predictors for compounds design is crucial for success of any in silico drug design. Our study enables the usage of LLMs in drug design tasks, while agentic frameworks would allow to include more elaborate methods directly.

6. Compound diversity analysis. The generated molecules should be evaluated using a compound similarity matrix to assess structural diversity under the constrained generation conditions, rather than limiting the analysis to the small subset of synthesized molecules.

We moved figure S2 and S3 from the SI to the main text illustrating the similarity of generated molecules towards the queries as well as internal diversity of generated molecules for several repetitions of CatA designs. We further added an analysis of different metrics for the generated sets of RIPK1 and CatA (QED and SAScore) as well as the similarity of these molecules to the query compounds.

7. Selection criteria for synthesis. Given the large number of generated candidates, the manuscript should clarify how the final molecules were selected for synthesis. Were

criteria such as synthetic accessibility, ADMET predictions, or drug-likeness metrics applied? Incorporating such rational selection principles would strengthen the study.

Thanks a lot for the comment. We have extended this section and now provide more details on the selection procedure and properties considered. In addition, we added an analysis and histograms of metrics such as QED and SAScore on the generated as well as selected molecules.

We thank all reviewers for their valuable comments, which helped to improve the quality of the manuscript.

Reviewer 5: *"The work presented by Grebner et al., seeks to fill a gap in knowledge regarding small-molecule drug discovery and the utilization of Large Language Models (LLMs)... My only question that would be useful to include in the revision would be related to conventional structure-guided design. Especially for RIPK1, where X-ray structures are known, how are results from the LLM generated molecules different or similar to those that may be arrived at through rational structure-guided design? Do the LLMs provide unique perspectives that ordinarily may be missed? Can similar points be said for Cathepsin A? The answer to this general question would give important perspective to how LLMs are able to add to drug discovery beyond what would be considered typical 'human' reasoning."*

We thank the reviewer for this insightful question, which touches on a central aspect of the practical value of LLM-based design. We have addressed this comment and believe it has meaningfully strengthened the manuscript.

We have added a dedicated paragraph in the RIPK1 section discussing the fundamental differences between LLM-based and conventional structure-guided design. These two approaches operate on distinct information sources and are best understood as complementary rather than competing strategies. LLM-based design leverages two-dimensional structure-activity relationship (SAR) data derived from existing ligands, while structure-based design exploits three-dimensional protein structural information from X-ray crystallography. Importantly, structure-based design is inherently constrained to the chemical space defined by available crystal structures — a subset that represents only a fraction of biologically relevant molecular space.

To illustrate this distinction concretely, we performed a principal component analysis (PCA) comparing the chemical space covered by 119 LLM-generated pyrrolidine derivatives against 24 known type III RIPK1 binders from the Protein Data Bank (May 2026). As shown in the new Supporting Figure S17, the LLM-designed compounds occupy substantially different regions of chemical space: with the exception of one LLM design related to 3–4 public structures, all remaining proposals — including the six synthesized compounds — fall in regions not covered by the known crystal structure-derived binders.

Taken together, these findings highlight the potential of LLM-based design to complement traditional structure-based approaches by expanding the accessible chemical space beyond what is defined by existing X-ray structures, while maintaining the essential requirements for biological activity. As noted in the revised manuscript, the parallel sampling of chemical space through diverse design strategies is increasingly recognized as an important paradigm in drug discovery.