Authors' responses to reviewers' comments on *Extracting human interpretable structure-property relationships in chemistry using XAI and large language models*

October 14, 2024

Thank you for your constructive feedback on the manuscript. Please see our responses below. Reviewer text is in plain typeface, author replies are in italic typeface and manuscript excerpts are in sans-serif font. Revisions also appear in blue font in the revised draft.

Reviewer 1

Reviewer 1, Comment 1

Since this journal targets readers in the field of chemistry, the title's use of "XAI" may be unclear.

Author Reply:

We appreciate the reviewer's comment. However, we would like to emphasize that the target audience for this publication lies at the intersection of chemistry and AI, where readers are likely to be more familiar with AI terminology than those from a strictly chemistry background. While we acknowledge that some readers may be unfamiliar with the term `XAI`, we have provided a detailed explanation of it within the abstract and manuscript to ensure clarity.

Reviewer 1, Comment 2

Table 2: It would be better to include the number of parameters (in billions). Llama2 comes in 7, 13, and 70 billion parameter versions. Even the smallest 7B model should be around 14GB in size at 16-bit precision, so the 3.8GB mentioned seems unnatural. The authors likely used a quantized model.

Author Reply:

Thank you for this suggestion. We have updated Table 2 to include the number of parameters and level of quantization. We have also added a sentence about the quantization in the main text.

Revised

Open-XpertAI: exploring open-source LLMs:

> The selected LLMs were: Llama2:7b,[1] Mixtral:8x7b-instruct-v0.1-q5_0,[2] Starling-lm:7b-alpha[3] and Phi:2.7b.[4] These open-source LLMs were configured and executed locally utilizing Ollama, a streamlined AI tool designed for local deployment of open-source LLMs. For all LLMs except Mixtral:8x7b-instruct, Q4_0 quantization level was used.

Reviewer 1, Comment 3

At the time of review, state-of-the-art models include Llama3.1, Gemma 2, and Phi3, which significantly outperform the models used in this study. While it may not be necessary to rerun experiments with these models, mentioning the existence of these advanced models at the time of revision would be prudent.

Author Reply:

Thank you for the suggestion. We have updated the manuscript with the following.

Revised
Open-XpertAI: exploring open-source LLMs:

> To investigate the feasibility of integrating open-source LLMs into XpertAI we conducted a brief study. We assessed the performance of4 open-source LLMs that have demonstrated comparable capabilities to GPT-4, focusing on the accuracy of their generated explanations. The selected LLMs were: Llama2:7b, [1] Mixtral:8x7b-instruct-v0.1-q5_0, [2] Starling-lm:7b-alpha [3] and Phi:2.7b. [4] These open-source LLMs were configured and executed locally utilizing Ollama, a streamlined AI tool designed for local deployment of open-source LLMs. For all LLMs except Mixtral:8x7b-instruct, Q4_0 quantization level was used. While we utilized a selection of open-source LLMs available at the time of research, we acknowledge that more advanced models may be available at the time of publication, and future studies could benefit from these enhanced versions.

Reviewer 1, Comment 4

It would be beneficial to discuss the inherent problems of large language models for future perspectives. Are there issues with inference abilities? Or is there a lack of certain chemical knowledge? Does the data that language models are trained on (mainly internet-derived) cover sufficient information for experimental chemists?

Author Reply:

Yes, LLMs often fail to perform satisfactorily in knowledge-intensive, data-sparse domains such as chemistry. We have updated the manuscript to convey the importance of RAG in contrast to using LLM's by themselves.

Revised
Method:

> As seen in the overview of our proposed workflow (Figure 1) LLMs are used to unite the backend modules generating human interpretable explanations. More technically, XpertAI makes use of the retrieval augmented generation (RAG) [5] approach to reliably generate scientific explanations using evidence gathered from the literature. LLMs' inherent knowledge can be limiting in knowledge-intensive, data-sparse fields such as chemistry and materials science. [6] Therefore, LLMs are prone to generate misinformation and hallucinate answers in such cases. RAG approach is commonly used to avoid such limitations in LLMs as it augments the LLM's internal knowledge with external data sources. [7]

Reviewer 1, Comment 5

As the authors point out, the accuracy of RAG (retrieval-augmented generation) is highly dependent on the performance of the search system. A reviewer's concern is the accuracy of the search system itself. For example, if the algorithm searches by semantic similarity, would it be possible to perform appropriate searches without creating a language model familiar with cutting-edge chemistry? I am interested in the authors' thoughts on this.

Author Reply:

Yes, the retrieval component in the RAG method is crucial to the model's performance. While the choice of similarity search algorithms (e.g., max marginal search, vector similarity search) can affect performance, the literature indicates that the impact is not significantly pronounced. These are all well-established approaches. Additionally, the type of embeddings used and the vector database storing the data sources can also influence performance. However, to our knowledge, no systematic investigations into these factors have been conducted. It is important to note that we did not explore these aspects in our study, as they fall outside the scope of our research.

Reviewer 1, Comment 6

Details about how to access the models used and their versions should be clearly stated. Especially for services like ChatGPT, which are updated frequently, including the version and access date would be beneficial.

Author Reply:

We have updated the manuscript with more details.

Revised
Method:

> We leverage on LangChain python package (`https://github.com/langchain-ai/langchain`) Chroma vector database (`https://github.com/chroma-core/chroma`) (the retriever) and, OpenAI's GPT-4[8] language model (the generator) in this workflow. This GPT-4 version (gpt-4-0613 at the time of publication) was trained on data up to September 2021 (`https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4`).

Reviewer 1, Comment 7

The number of samples tested in this study was unclear. For example, in the MOF section, it states "we sampled 3734 MOFs from the CoRE MOF 2019 database," but does this mean that interpretability was tested for all 3734 samples? The SI only listed 2-3 examples for each case. To ensure academic objectivity and statistical significance, it is necessary to describe the number of samples tested and the sampling method. It seemed like cherry-picking for each case study.

Author Reply:

We apologize for any confusion regarding this matter. To clarify; we trained and tested the XGBoost model using an 80-20 split of the sample data. Once the model was trained, the entire sample was utilized to generate the XAI analysis and obtain the "mean" top n features, which were subsequently employed in the RAG method. Our hypothesis was that by averaging the XAI results, the output from XpertAI would be more robust to the data, providing a comprehensive overview of the dataset. Rather than generating explanations for each individual data point, we produced a single explanation for the entire dataset.

Reviewer 1, Comment 8

It might be beneficial to clarify the current level of the system and the level it aims to achieve in the future. For instance, identifying functional groups contributing to toxicity and extracting units that might adversely affect living organisms is a task that trained chemists should be able to perform. As the authors describe, the difficulty in predicting toxicity and its interpretation arises from unexpected phenomena or effects caused by complex interactions of functional groups. What is truly required at the research level is prediction and interpretation at this level. Clarifying the level of the current system (whether it can substitute for the general public, someone with specialized chemical education, or conduct cutting-edge research), and organizing the missing elements and future tasks, seems academically important.

Author Reply:

We have updated the manuscript as follows.

Revised
Conclusion & Outlook:

> Enhancing the performance and efficiency of the RAG model is an ongoing topic of investigation that is beyond the scope of our work. However, we anticipate that improvements in existing tools and methodologies will enhance XpertAI's overall performance. For example, better-performing retrievers and LLMs will undoubtedly improve XpertAI's capabilities. The current version of XpertAI serves as a proof of concept that XAI integrated with LLMs can be a proxy for hypothesis generation in Chemistry. XpertAI mimicks the workflow a scientist would follow to arrive at a hypothesis – following an observation, a hypothesis is generated and supported with literature evidence. Given the growth of tools and methodologies based on LLMs, we aim that XpertAI can become a powerful hypothesis-generating agent in the future.

Reviewer 2

Reviewer 2, Comment 1

As mentioned in the article, feature selection significantly impacts the fit of the surrogate model. Can

authors provide a more efficient method to extract effective features from the literature? Can authors provide a quantitative representation of the correlation between the issues to be resolved and the corresponding descriptor in each case?

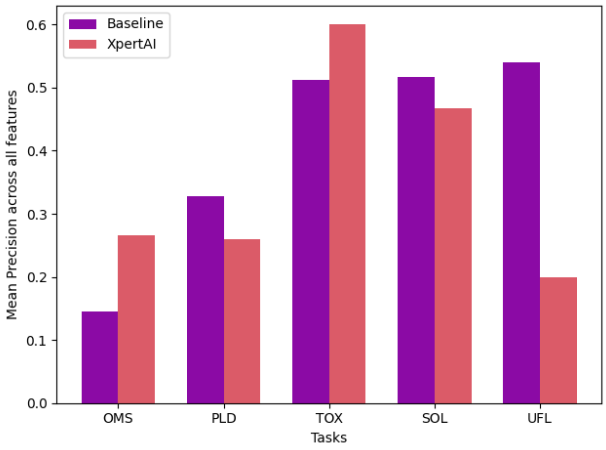
Author Reply:

We appreciate the reviewer's questions and we have been experimenting along these lines as well. We have updated the XpertAI prompts to look for synonymous features if labels are not found verbatim in the literature. We agree that a more quantitative analysis was required to analyze the precision of the generated hypotheses (please see figure 5). We have updated the manuscript to reflect these changes. The newer explanations are updated in Appendix A.

Revised

Evaluation:

> We further evaluated the precision of the generated hypotheses by XpertAI and compared with GPT-4o explanations. For each task and each feature that is identified in the XpertAI explanation, we extracted each proposed hypothesis and labeled the correlation of the feature with the property under study (OMS, PLD etc.) as `positive`, `negative`, or `unclear`. An example of an unclear correlation is "Metal ions with favorable redox properties might be more likely to form stable open metal sites." We used the following formula to compute the precision.
>
> $$\text{precision} = \frac{1}{N_{features} \times N_{runs}} \sum_{i}^{N_{features}} \left| (1) \times n_{i,positive} + (-1) \times n_{i,negative} + (0) \times n_{i,unclear} \right|$$
>
> Here $N_{features}$ is the total number of unique features listed in the explanations and $N_{runs} = 5$. assigned arbitrary weights of +1, -1 and 0 to calculate precision $\in 0, 1$. These weights allow us to assess whether the generated explanations align or diverge at the feature level. The results shown in Figure 5 reflect this analysis. XpertAI either outperforms or is comparable to the baseline at the feature level, except for case study 5, our negative example. In instances where XpertAI scores lower, the majority of per-feature correlations were classified as `unclear`, reflecting the absence of explicit correlations in the literature. This suggests that XpertAI avoids generating speculative or unfounded conclusions. Conversely, while the GPT-4o baseline model demonstrated more consistency in its claims across the five runs, there is no assurance that these claims or correlations are free from hallucinations.



Mean precision of hypotheses generated (↑). XpertAI-generated hypotheses are compared with baseline GPT-4o. Equation (1) was used to compute the precision $\in 0, 1$.

Reviewer 2, Comment 2

The quantity and the quality of input literature have a great impact on the interpretability and directly determine the ability of XAI interpretability. Is it possible to use the advantages of LLMs to autonomously find more relevant literatures and extract effective text information for analysis?

Author Reply:

Yes absolutely. This is a great suggestion. In the current version of XperAI, it already allows automatic scraping of arXiv database given keywords. While this may not contain peer-reviewed articles, we demonstrate the capabilities of XpertAI in an automated setting with this work. We aim to give XpertAI access to more open-source literature databases in future updates. The main reason for asking a user to upload a paper dataset is that a user (an expert) in a subject may already have access to high-quality, peer-reviewed publications that are specific to the task at hand. Additionally, this avoids expensive API calls in the backend.

Reviewer 2, Comment 3

Can authors use specific experiments to prove the correctness of these explanations?

Author Reply:

As human judgment is considered to be the gold standard, we mainly relied on our human expert evaluations to measure the correctness of the explanations. In the current updates, we manually evaluated precision (see answer above) and citation accuracy (table 1).

Reviewer 2, Comment 4

When faced with insufficient data, is it possible to improve the accuracy of model training by generating some hypothetical data?

Author Reply:

Not necessarily, as data engineering in chemistry is a multi-faceted, task-specific problem. While data augmentation is useful in some tasks, it needs to be addressed carefully. For example, in the case of upper flammability limit of organic compounds, we find it impossible to generate hypothetical data that reflects ground truth behavior. This will deteriorate the model performance and generate misinformation. Hence, we did not investigate on this front.

Reviewer 3

Reviewer 3, Comment 1

This could be a misdirection in the application of XpertAI; in the Supplementary Information (SI), Case Study 1 of XpertAI Explanation is referenced. In the discussion of density, the original text of the cited article does not highlight the influence of solid density on OMS. Specifically, the third point mentions solid density in relation to hydrogen storage capacity, not OMS. Compared to XpertAI, ChatGPT, which does not utilize any RAG technology, provides more accurate information.

The same misdirection is present in Case Study 2. In your manuscript: "The Volume Per Atom can influence the pore limiting diameter in MOFs as it provides an indication of the size of the atoms that make up the MOF. Larger atoms may result in larger pores, while smaller atoms may result in smaller pores (Haldoupis et al., 2010)." In Haldoupis's research, it is the MOFs capable of accommodating larger guest molecules that have larger pore sizes, not the size of the MOF atoms. Similarly: "Symmetry Function G1 is a measure of the symmetry of the MOF structure. MOFs with higher symmetry may have more uniform pore sizes, while those with lower symmetry may have more varied pore sizes (Choudhuri & Truhlar, 2020)." The article does not mention Symmetry Function.

The pore size and OMS of MOFs are relatively simple properties, yet XpertAI did not excel in the test. Might it be worth considering changing the case study example to focus on explaining the properties of small molecules?

Author Reply:

Reply We sincerely appreciate the reviewer's feedback regarding the citations, and we agree that the initial XpertAI explanations could be vague and lacked effective cross-referencing. In this round of revisions, we have upadated the XpertAI workflow by incorporating improved prompt engineering and utilizing a more advanced model –GPT-4o. The explanations have been regenerated, the results updated, and a thorough cross-reference analysis was conducted. The new explanations demonstrate significant improvements in cross-referencing, clearly identifying when supporting evidence from the literature is not found. Updated explanations are now in Appendix A and the manuscript is updated with these results. We have also updated Table 1 to reflect the accuracy of citations. See an example update below.

> Following a similar approach in case study 2, we used the same MOF dataset but with pore-limiting diameters as the label. Unlike case study 1, this is a regression-type problem. We uploaded a literature dataset with 24 journal articles to support XpertAI explanations. The pore size characterized by the pore limiting diameter is an important property in MOFs that can control charge transfer and direct air capture.[9] According to XpertAI, key factors influencing the pore-limiting diameter include volume per atom, symmetry function G, and unoccupied energy levels at the conduction band. For instance, XpertAI hypothesizes that "Symmetry Function G1 may impact the pore limiting diameter by influencing the spatial arrangement of atoms, potentially affecting the uniformity and size of the pores". It continues to state that "an explicit relationship between Symmetry Function G1 and the pore limiting diameter was not found in the given documents. However, the documents discuss the geometric properties of MOFs, which are inherently related to symmetry[10]". This highlights XpertAI's capability to produce insightful and plausible explanations while maintaining scientific rigor, avoiding speculative conclusions when supporting evidence is absent. We prompted XpertAI to go beyond merely identifying the most relevant molecular features associated with the target property. It also hypothesizes potential structure-property relationships and offers scientific reasoning, drawing on insights from the provided literature, emulating the approach a human scientist would take. For a comprehensive textual explanation, please refer to the Supporting Information (Appendix A).

Reviewer 3, Comment 2

The manuscript employs XGBoost and Random Forest algorithms to construct machine learning models. We are also interested in the features used in constructing the machine learning models, in addition to the most important features. This indicates which features XpertAI's explanations were derived from.

Author Reply:

Reply In this analysis, we had to select molecular features which were human-interpretable. We selected commonly used molecular features such as MACCS, or keys in this study. Each case study lists the features or featurization approach used. For example, to convert CIF files to crystal features, we used the CrystalFeatures tool[11]

Reviewer 3, Comment 3

Why were 3734 MOFs selected from CoREMOF when there are over 20,000 MOF materials in the CoREMOF database? What is the rationale for this selection?

Author Reply:

We selected an initial smaller sample of 4000 MOFs in this work to minimize compute and filtered based on input validity after featurization. Even with only, 3734 sample, we see that the trained XGBOOST model's accuracy is above 80%. We have updated the manuscript with to highlight this.

Revised

Method:

> In case study 1, we sampled 4000 MOFs from the CoRE MOF 2019 database[12] that contained labels for the presence of open metal sites and pore-limiting diameter. After input validation and the featurization step, we ended up with 3734 structures. These crystal structures obtained as CIF files were then featurized using the CrystalFeatures tool.[11]

Reviewer 3, Comment 4

The author used Claude to evaluate the model's responses, but did not provide Claude with the correct answers. Relying solely on the knowledge base of the large model itself is not rigorous.

Author Reply:

In our study, we utilized Claude solely to compare 3 types of explanations: 1) XpertAI explanation (GPT-4 + XAI), 2) GPT-4 alone, and 3) XAI alone. Claude was prompted to assess these explanations based on their relevance and interpretability. It is important to note that the accuracy here refers to Claude's inherent understanding of the subject, similar to the GPT-4-based explanations. We acknowledge that, like any other LLM, Claude may not be able to rigorously evaluate the accuracy of the explanations. For this reason, we incorporated human validation into the study, considering human evaluation to be the gold standard for assessing the quality of the explanations.

Reviewer 3, Comment 5

In the realm of open-source large language models, Bai et al.'s study (`https://pubs.acs.org/doi/10.1021/acs.jcim.4c00065`) has already tested various open-source large language models in the field of MOFs. Consider incorporating GLM series and Llama models that performed well in this study. Furthermore, the tested Llama model has been updated to version 3.1. As a frontier in research, it should endeavor to use the most recent model.

Author Reply:

Reply We appreciate the reviewer's suggestion regarding the use of open-source LLMs. Our primary objective in incorporating open-source models was to demonstrate that XpertAI can be seamlessly integrated with them if needed. The tests were conducted as a proof of concept. While we recognize that newer and more advanced open-source LLMs are being released at an accelerated pace, we chose not to conduct experiments with the latest models in order to stay focused on the main objectives of the study and to complete the revision process in a timely manner. To clarify this, we have updated the manuscript accordingly to indicate this "cutoff" in our experimentation.

Revised

Open-XpertAI: exploring open-source LLMs:

> For all LLMs except Mixtral:8x7b-instruct, Q4_0 quantization level was used. While we utilized a selection of open-source LLMs available at the time of research, we acknowledge that more advanced models may be available at the time of publication, and future studies could benefit from these enhanced versions. Detailed performance metrics of each LLM against benchmark datasets can be found in the respective references.

References

[1] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., arXiv preprint arXiv:2307.09288, 2023.

[2] M. AI, Mixtral of Experts, 2023, `https://mistral.ai/news/mixtral-of-experts/`.

[3] B. Zhu, E. Frick, T. Wu, H. Zhu and J. Jiao, Starling-7B: Improving LLM Helpfulness & Harmlessness with RLAIF, 2023.

[4] Microsoft, Microsoft Phi-2, 2023, `https://ai.azure.com/explore/models/microsoft-phi-2/version/4/registry/azureml-msr?reloadCount=1`.

[5] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun and H. Wang, arXiv preprint arXiv:2312.10997, 2023.

[6] W. Xu, S. Agrawal, E. Briakou, M. J. Martindale and M. Carpuat, Transactions of the Association for Computational Linguistics, 2023, **11**, 546–564.

[7] K. Sawarkar, A. Mangal and S. R. Solanki, arXiv preprint arXiv:2404.07220, 2024.

[8] OpenAI, ArXiv, 2023.

[9] M. Cai, Q. Loague and A. J. Morris, The journal of physical chemistry letters, 2020, **11**, 702–709.

[10] E. Haldoupis, S. Nair and D. S. Sholl, Journal of the American Chemical Society, 2010, **132**, 7528–7539.

[11] S. A. Tawfik and S. P. Russo, Journal of Cheminformatics, 2022, **14**, 1–11.

[12] Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp et al., Journal of Chemical & Engineering Data, 2019, **64**, 5985–5998.