Response to reviewer comments:
In the following document we address the comments and questions raised by the different reviewers. For each reviewer, we first rewrite their comment in full, then write a point-by-point response to their questions. The reviewer's comments are written in Blue, and our reply are in black with the changes to the main text in Red.

Program Committee:

> AI4Mat-NeurIPS-2024 Program Committee believes that the feedback from the reviewers and the recommendation by the Area Chair pointed to some considerable issues with the submission in its current form. We believe that the issues can be addressed between now and the camera-ready deadline and as such decided to accept the submission conditional on fixing the concrete points outlined below. Please post a comment here acknowledging your intent to address the below. Following the camera-ready deadline, the Program Committee will review the draft in detail to make sure things have been properly addressed. We recommend that you post an additional comment outlining the changes made between now and the camera-ready deadline.
> **Requested Changes:**
> Please motivate the presented problem settings in terms of their relevant to materials design.

Response to the program Committee Comments:

We thank the program committee for their comments and for their decision to consider this work for publication. For the requested changes:

> Please motivate the presented problem settings in terms of their relevant to materials design:

We have added a few sentences in the main text to address this question.
For Problem 1 and 2, where we only used a synthetic problem, we have added the following sentence:

> *"The first two synthetic problems, are relevant to cases where the design space is continuous such as reaction conditions [\citenum{shields_bayesian_2021}] and thin film deposition conditions [\citenum{gencoa_ltd_automation_2022}]."*

We have also added a section in the appendix to introduce the function in more detail.
Problem 3, and 4 are directly related to materials design as they relate to both an experimental problem, and the second purely computational, but with properties relevant for organic photovoltaic applications.

Reviewer 1:

Strengths

- Well-written; writing style is clear and easy to follow
- I like the creation of low-fidelity datasets -- how transferable are they to real world settings? Seems like Problem 4 is the closest?
- Multi-fidelity results are interesting although would be nice to see on more chemistry problems (maybe different modalities for example), unclear how problems 1 and 2 generalize

Weakness/Questions

- What is the RKHS function?
- how was low fidelity data computed for Problem 3?
- I think more can be done to extend the analysis -- would be interesting to extend Problem 4 to more than 2 scales incl. e.g. DFT

Minor points

- only need to expand abbreviations once (e.g. "multi-fidelity Bayesian optimization" in conclusion)

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Rating: 3: Marginally above acceptance threshold

Response to reviewer 1:

We thank the reviewer for their thoughtful questions and suggestions; we have incorporated them into the manuscript and describe the changes here for clarity.

- I like the creation of low-fidelity datasets -- how transferable are they to real world settings? Seems like Problem 4 is the closest?

Within the multi-fidelity setting that we explore in this work, measurements at lower fidelities are considered to be noisier observations of the true, underlying objective function. Indeed, this is analogous to real-world settings, where often the difference between high and low fidelity measurements is often the rigor of the assumptions that are involved in the respective methods. Both Problems 3 and 4 feature tasks that fall under this category.

Indeed, the fidelities that are included in Problem 3 feature two different types of simulations to assess the target property that is being optimized. The main difference between the high and low fidelity simulations in this case study is the assumptions that are made, the larger the assumptions, the less expensive (in terms of time) the simulation. Problem 4 is the closest to a real-world application of MFBO. Yet, we stress that the method used to construct the low-fidelity datasets for Problems 1 and 2 is only really advantageous in the synthetic problem settings – as generating the low fidelity dataset in this way requires that the high-fidelity datapoint can be accessed. If this were the case, there wouldn't be a need for lower fidelity measurements.

- Multi-fidelity results are interesting although would be nice to see on more chemistry problems (maybe different modalities for example), unclear how problems 1 and 2 generalize

We agree; Problems 1 and 2 are included because they are standard benchmarking functions used to assess BO algorithm performance. Including these helped us to better understand how the underlying algorithm was operating, decoupling and simplifying the problem by removing the chemistry is critical to gaining a deeper understanding and intuition for how MFBO algorithms work.
We are considering including more chemistry problems as the next step.

- What is the RKHS function?

The RKHS function is a standard benchmarking function for optimization algorithms. This presents a complex, high-dimensional parameter space to search, while still being analytically tractable for conventional methods. This is powerful for comparison studies like the one presented in this work.

See Assael (2019) for the definition of the function, which can be described using two Squared Exponential kernels. The known solution is located at the point

x=0.892, f(x)=5.734

We agree that this is ambiguous within the manuscript, and have included a section in the Appendix to address this:

A.3 RKHS function

- How was low fidelity data computed for Problem 3?

The objective for problem 3 is to maximize the Xe/Kr selectivity of covalent-organic frameworks. There are several ways to assess this target property – in the dataset that our work uses (originally presented by Gantzler et al), the low fidelity approach is to calculate the Xe/Kr selectivity using Henry's law and relies on the dilute approximation in the binary grand-canonical ensemble. This is in contrast to the high-fidelity method, which involves a Markov chain Monte Carlo simulation in the binary grand-canonical ensemble. One of the major advantages of the Problem 3 formulation is that the low and high-fidelity data are available through the original publication.
We added the sentence below to clarify this:

The lower-fidelity data was created using Henry's law (as opposed to Markov chain Monte Carlo simulation in the binary grand-canonical ensemble as used for the high-fidelity case) and found to have a correlation 0.97 to the higher-fidelity data, and we assigned a relative cost of 0.2.

- I think more can be done to extend the analysis -- would be interesting to extend Problem 4 to more than 2 scales incl. e.g. DFT

We thank the reviewer for the suggestion, it would indeed be interesting to include different level of theories, similar to the work in https://www.nature.com/articles/s41524-022-00947-9 .
We are considering this as a follow up study; the current work was aimed at assessing when using multiple fidelities may be advantageous, thus we focus on cases featuring only two fidelities.

Reviewer 2:

> The authors provide an exploration of multi-fidelity BO's application in a few study cases and presents insights into its limitations.
> While the authors expected multi-fidelity BO to outperform single-fidelity BO, their tests, on synthetic problems and chemical datasets, suggest that multi-fidelity BO's performance depends heavily on data correlation and cost. I'd also add a two other factors: (1) since the exact parameters used for surrogate model GPs in BO's weren't provided, I would consider the hyper parameter settings for GPs also affect performance. Using default hyper parameters for GP might not suite every dataset. (2) The tested acquisition functions for multi-fidelity BO are fairly generic, and I don't think they are state-of-the-art or have consensus on being robust for chemistry related questions.
> The results also show that GP-EI is a powerful combination for single-fidelity BO and generally speaking, applies to a broader range of problems. In the meantime, as mentioned by the authors, multi-fidelity BO may be limited by its (over-)reliance on high correlation as well as low-cost assumptions. It's clear that the proposed acquisition function needs more testing, and this work, in general, lacks more testing across more diverse chemistry and material science domains.
> Clarity:3 Quality:3 Impact:3
> **Confidence:** 5: The reviewer is absolutely certain that the evaluation is correct and very familiar with the relevant literature
> **Rating:** 3: Marginally above acceptance threshold

Response to reviewer 2:

We thank the reviewer for their consideration, suggestions and thoughts. We agree with the major points that are raised, and address these here, as well as in the manuscript. We highlight the changes made to manuscript here for clarity.

> (1) since the exact parameters used for surrogate model GPs in BO's weren't provided, I would consider the hyper parameter settings for GPs also affect performance. Using default hyper parameters for GP might not suit every dataset.

We have addressed the lack of clarity in the exact methods used in the BO algorithm formulation and include additional description in both the main text and Appendix; all changes are highlighted here for simplicity.
Within the main text, we include a reference to the detailed description of the SFBO and MFBO algorithm frameworks:
"The choice of the GP kernel is important to the optimization performance; Appendix **\textbf{\textit{INSERT}}** *includes a detailed description of the SFBO and MFBO algorithm frameworks."*

We also agree that the influence of GP hyperparameters is not negligible and is not addressed within the manuscript. This is now addressed and raised as a future consideration in the Conclusions and Future Outlooks section; exact changes to the manuscript are provided here for clarity.

"This suggests the acquisition function should be carefully adopted for each case explored. We not that we did not explicitly consider the role of GP hyperparameters in this work; indeed, additional consideration is required here, as the GP hyperparameters are known to impact the optimization performance."

> (2) The tested acquisition functions for multi-fidelity BO are fairly generic, and I don't think they are state-of-the-art or have consensus on being robust for chemistry related questions.

We agree that the tested acquisition functions for MFBO are not state-of-the-art or have consensus on being robust for chemistry related questions. Indeed, there is not a wide consensus on acquisition functions for chemistry-related questions. One of the motivations for our work was to highlight the need for additional consideration in and study of the MFBO case. While a necessary contribution to the field, solutions and discussions of state-of-the-art approaches were outside the context of the study that we performed.

> The results also show that GP-EI is a powerful combination for single-fidelity BO and generally speaking, applies to a broader range of problems. In the meantime, as mentioned by the authors, multi-fidelity BO may be limited by its (over-)reliance on high correlation as well as low-cost assumptions. It's clear that the proposed acquisition function needs more testing, and this work, in general, lacks more testing across more diverse chemistry and material science domains.

We agree that additional testing across more diverse chemistry and material science domains is necessary for the field. One of the core contributions of this work is to highlight areas where additional study and consideration is needed when formulating a BO problem within chemistry. One of the challenges associated with this task is the limited available datasets for appropriate benchmarking and testing.

Reviewer 3:

Originality: The idea of using MBFO for chemical design seems original. However, possible reiteration can be found for MBFO use cases in different problems within the chemical design and beyond.
Quality: Authors have put an honest effort into providing enough work to justify their belief in MBFO, with correct data and method.
Clarity: Authors have clarity in their presentation.
Significance: The significance of this paper is questionable. The authors do provide important analysis using different problems to show the case for MBFO; however, it lacks any significant contribution.
Review Tools Track:
paper does not qualify for the tool
Review Findings Track:
Clarity: The paper's finding makes use case for MBFO, with clear results and comparison. However, more work is required to make a broad case beyond case study comparison
Quality: The authors seem to have done a good job; however, a broader view of the problem is necessary.
Confidence: 3: The reviewer is fairly confident that the evaluation is correct
Rating: 2: Ok but not good enough - rejection

Response to reviewer 3:

We thank the reviewers for their thoughtful comments. We would like to highlight that this work is intended to identify areas of consideration and improvement for MFBO problem formulations in chemistry. We agree that we do not offer a concrete solution to the highlighted challenges in the field.

Reviewer 4:

Clarity: The paper explains the technical problem well, but some sections are a bit unclear. The motivation for the four proposed problems is missing (what are the relations and difference between them? what do we gain from solving each problem?). It might be hard for readers who aren't familiar with MFBO to follow everything.
Quality: Some technical content is described in detail, some other content is not. More motivation for the experiments is needed. (See below for more comments.)
Impact: The technical work would be useful to other researchers and there are lessons to be learned from this work. Namely, identifying regimes where MFBO outperforms SFBO.

1. Abstract

- Line 6: "We specifically analyze the conditions under which lower-fidelity data can enhance performance compared to single-fidelity problem formulations." Consider rewriting given that the results you show are not general, but are specific to the synthetic problems you study.
- Line 8-9: Not clear which are the two key challenges as three problems are mentioned here.
- Line 9: Can add an adverb in the final sentence. Eg. "Finally, we discuss…"

2. Introduction

- Line 18: Consider replacing "Often this challenge" with "Ofter, the challenge of search in chemical space…" or something else that improves clarity.
- Line 26-27: Consider adding references.
- Line 30-31: Consider saying "including but not limited to" or adding "…, among others" at the end of the sentence. Examples of other BO applications in chemistry include https://doi.org/10.1038/s41586-021-03213-y, https://doi.org/10.1126/science.adk9227, etc.
- Line 37-38: Please add references for each molecular encoding for consistency.
- Line 32-54: Consider rewriting these paragraph for clarity. It is mention here that there are "several limitations" of BO. However, both the second and third points mentioned (encoding to machine readable format and ability to handle mixed domains) are not limitations. Also, the introduction of the kernel is abrupt and unclear.
- Line 55: Consider replacing *"*As chemists, we often have access" to a sentence that includes non-chemist readers.
- Line 61: Fix spacing in "(MFBO)"
- Line 64-65: This sentence is unclear.
- Line 71: Unclear statement: "it is possible to optimize the objective function more economically."
- Line 76: Consider replacing "the algorithm." by "MFBO."

3. Experimental methods

- Line 86: "We did not include Knowledge Gradient[22] acquisition function here, as preliminary 87 tests showed the computational cost of the approach was

prohibitive." It is unclear what the Knowledge Gradient acquisition function is and what made it computationally expensive. Consider removing.

- Line 145: "Relative Improvement" is used before it is defined. Not really clear from the text. Better to put a formula.

4. Results and discussion

- Consider relabeling problems from a number to their true description. The number label does not quantify difficulty and when referencing problems later in the text, this labeling is confusing (Eg. captions of Fig 3 and appendix Figures).

5. Conclusions and Outlook

- Line 167: "These initial results were surprising" — How come it is more surprising a BO algorithm is better when using more data (despite lower quality data?). This should be explained in more detail.

6. Appendix

- Figs 4-7: Captions of all these figures are unclear.
- Consider unifying histograms from Figures 4-7 into a 2x2 subplot.

7. General comments

- Citation style is not consistent throughout the work. This should be fixed.
- For all citations leave a space before the reference. i.e. Replace "word[citation]" (word\cite{ref}) with "word [citation]" (word~\cite{ref}).

Confidence: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper
Rating: 2: Ok but not good enough - rejection

Response to reviewer 4:

We thank the reviewer for their thoughtful suggestions and present the modifications to the manuscript here for clarity.

2. Abstract

- Line 6: "We specifically analyze the conditions under which lower-fidelity data can enhance performance compared to single-fidelity problem formulations." Consider rewriting given that the results you show are not general, but are specific to the synthetic problems you study.
- Line 8-9: Not clear which are the two key challenges as three problems are mentioned here.
- Line 9: Can add an adverb in the final sentence. Eg. "Finally, we discuss..."

We have modified the abstract to include these changes; please find the updated section here, for clarity:

"... Here, we investigate the application of MFBO to accelerate the identification of promising molecules or materials for target applications. We specifically analyse the conditions under which lower-fidelity data can enhance performance compared to single-

fidelity problem formulations through four case studies – addressing three key challenges: i) selecting the optimal acquisition function, ii) understanding the impact of the correlation between cost and data fidelity, and iii) how to assess the effectiveness of MFBO for chemical discovery."

Lines

3. Introduction

- Line 18: Consider replacing "Often this challenge" with "Ofter, the challenge of search in chemical space…" or something else that improves clarity.

We have modified this sentence to read: "Efficiently traversing chemical space can often be translated to an optimization problem, …"

- Line 26-27: Consider adding references.

We have included two additional references here:
R. Garnett, *Bayesian Optimization*, Cambridge University Press, Cambridge, 2023.
Digital Discovery, 2024, **3,** 1086-1100

- Line 30-31: Consider saying "including but not limited to" or adding "…, among others" at the end of the sentence. Examples of other BO applications in chemistry include https://doi.org/10.1038/s41586-021-03213-y, https://doi.org/10.1126/science.adk9227, etc.

We have modified the sentence to read:
"… and nanoporous materials, [5; 6] among others [7; 8]."
The two additional references are:
Nature, 2021, 590, 89–96
Science, 2024, 384, eadk9227

- Line 37-38: Please add references for each molecular encoding for consistency.

We thank the reviewer and have included additional references for consistency.
Commun. Mater., 2022, **3**, 1–18
WIREs Computational Molecular Science, 2022, **12**, 225, e1603
J. Cheminf., 2020, **12**, 1–22

- Line 32-54: Consider rewriting these paragraph for clarity. It is mention here that there are "several limitations" of BO. However, both the second and third points mentioned (encoding to machine readable format and ability to handle mixed domains) are not limitations. Also, the introduction of the kernel is abrupt and unclear.

We have modified the text to adjust the tone, as the second and third points are, indeed, not limitations, but considerations. We include the modifications in the main text here for clarity:

"While powerful, there are several important considerations when …"

We have moved the discussion on the kernel from this section of the manuscript. Indeed, this is not pertinent to the study that we performed; however, it is an important aspect of BO problem formulations. Therefore, we include it in the Conclusions and Outlook Section.

- Line 55: Consider replacing *"*As chemists, we often have access" to a sentence that includes non-chemist readers.

To be more inclusive, we have modified this to read: "Researchers often have access…"

- Line 61: Fix spacing in "(MFBO)"
- Line 64-65: This sentence is unclear.

We agree and have modified the sentence to read:

"Specifically, we examine the circumstances that lead to MFBO outperforming single-fidelity BO (SFBO)."

- Line 71: Unclear statement: "it is possible to optimize the objective function more economically."

We agree and have modified the statement to read:

"Lower-fidelity evaluations usually incur a lower cost, so provided there is some correlation between the low-fidelity and the high-fidelity target, it is possible to reach the optimum more economically (i.e. using fewer resources)."

- Line 76: Consider replacing "the algorithm." by "MFBO."

We have made the change as suggested.

4. Experimental methods
- Line 86: "We did not include Knowledge Gradient[22] acquisition function here, as preliminary 87 tests showed the computational cost of the approach was prohibitive." It is unclear what the Knowledge Gradient acquisition function is and what made it computationally expensive. Consider removing.

We have not removed the sentence from line 86 as we feel it is helpful for the reader familiar with MFBO.

- Line 145: "Relative Improvement" is used before it is defined. Not really clear from the text. Better to put a formula.

We have improved the clarity of this statement and included a formula, as suggested.

5. Results and discussion

- Consider relabeling problems from a number to their true description. The number label does not quantify difficulty and when referencing problems later in the text, this labeling is confusing (Eg. captions of Fig 3 and appendix Figures).

We thank the reviewer for this suggestion and have updated the figure captions accordingly.

6. Conclusions and Outlook

- Line 167: "These initial results were surprising" — How come it is more surprising a BO algorithm is better when using more data (despite lower quality data?). This should be explained in more detail.

We thank the reviewer for raising this point. It is surprising since the premise of MFBO is that including lower fidelity data with some correlation to the high-fidelity data would improve the performance of the search algorithm. We also could not find for the specific case of MFBO, any prior work discussing the limitation of the approach when the lower fidelity data prior to this work.
We have added a detailed introduction to single and multi-fidelity BO in the appendix. To clarify this point, and also

7. Appendix

- Figs 4-7: Captions of all these figures are unclear.
- Consider unifying histograms from Figures 4-7 into a 2x2 subplot.

Each of these points has been addressed; updated captions for Figures 4-7 are included here for clarity.

8. General comments

- Citation style is not consistent throughout the work. This should be fixed.
- For all citations leave a space before the reference. i.e. Replace "word[citation]" (word\cite{ref}) with "word [citation]" (word~\cite{ref}).

Each of these points have been addressed.