We thank the reviewers for their careful review. In the manuscript, all edits/additions are highlighted in blue. In what follows we will try to address the comments to the best of our abilities:

Report of the First Referee – XR10345M/Talapatra

This paper presents a Bayesian framework for the optimization of materials properties that combines important ingredients: (i) an array of simple candidate predictive models for the properties of interest whose predictions are weighted (aka "model averaging") to yield the predictions (ii) a list of simple descriptor to aid in materials property prediction (iii) a formal treatment of model uncertainties and (iv) a selection rule to determine the most useful new data point to obtain. This general topic covered is clearly timely and helpful and the paper does bring in some useful formal concept from the statistics literature into the field of materials discovery. However, the paper, in its present form, has some issues.

We thank the reviewer for this appraisal of the present work.

A) Introduction sets up extremely high expectations and the reader is left disappointed when seeing what the practical implementation example actually delivers. The application example is fine - it is the initial claims that are overstated.

We have modified the introduction and modified the text commensurate with the specific contributions of this work.

A few examples:

1) it is stated that earlier models "are incapable of dealing with the high dimensionality (compositional, configurational, and microstructural degrees of freedom) and complexity (e.g., multi-physics)...", but the example given is hardly high-dimensional and does not include microstructure considerations or multi-physics aspects.

We have eliminated this statement.

2) It is also indicated that earlier approaches "preclude expansion of the Materials Design Space (i.e., identification of new materials features) as new information is acquired". It is not obvious, in the example presented, that the system actually learns new features to add to the list. The list of features is decided in advance. In any case, a system that is truly able to come up with new features on its own would probably pass a Turing test.

We have eliminated this statement and modified the paragraph as follows: *Traditional HT experimental [refs] and computational [ref] approaches, while powerful, have important limitations as they (i) employ hardcoded workflows and lack flexibility to iteratively learn and adapt based on the knowledge acquired to assure balanced exploration and exploitation of the MDS (ii) and tend to be suboptimal in resource allocation as these approaches generally rely on highly parallelized exploration of the MDS, even in regions that are of low value relative to the objective, or performance metric, that is sought after.*

3) The claim "Current paradigms are typically centered around the idea of performing this exploration through high-throughput experimentation/computation. Such approaches, however, do not account for (the always present) constraints in resources available." will certainly sound incorrect to most readers. In the example, each ab initio calculation essentially takes the same amount of time (and the precise time is not known in advance anyway), so it is not clear that the algorithm takes into account computation time. The algorithm does seek to reduce the total number of calculations needed to reach a given precision, but a lot of researchers are doing that.

This is a statement in the paragraph. We have added more references in the text to relevant recent work. This issue of optimal materials discovery, as pointed out by the reviewers, has started to become more widely explored in the literature. We state in the abstract:*Recently, this problem has been addressed by framing materials discovery as an optimal experiment design.* We are thus placing our contribution in a larger context. As to the issue of resource allocation, our framework does not account for computational cost of a single computation as we implicitly assume that the unit cost per structure is roughly of the same order of magnitude. We currently are working on a parallel effort in which we have multiple sources of information of clearly different computational costs and in that case we explicitly include computational cost when computing the utility of querying a source.

B) Please define the type of Gaussian Process Regression used. This is a very general term and does not point to a specific model unless one specifies the prior on the (conditional) means and variances.

The Gaussian Process Regression used was defined in Section II B of the manuscript.

C) The authors need to better define what the swarm plots (fig 6 and 9) represent. The explanation in the text would suggest that wide bars are desirable, but the graph shows that the width decreases with the number of steps. More fundamentally, what is the "x" axis? The swarm plots indicate the distribution of all the 1500 instances for each feature set or BMA variant. ($F_1, F_2..., F_6$, BMA_1, BMA_2. Bottom heavy, wide bars, with the width decreasing with the number of steps is desirable, since that would indicate that larger number of instances needed fewer number of steps to converge.

D) What was the Pearson correlation matrix used for? (BTW, perhaps it's not needed define what a correlation is...)

The secton on the Pearson correlation was included to give some context to the relationships between the features, but was not used in feature selection. the authors realize the section is superfluous to the manuscript and it has been moved to the supplemental document.

E Minor issues:

1) In the bibliography, make sure to enclose chemical formulas in { } in your BibTeX file, otherwise the capital letters get lost.

We have fixed this issue.

2) Figure 1 would be easier to understand if the vertical scale were the same in both panels. Currently, it looks like the true model f(x) changes. Also, for clarity, include the red curve in the legend.

The figure has been corrected.

3) In the present case, the Pareto front should perhaps be defined as "Specifically, the Pareto front here is the 1-dimensional design curve over which any improvement in one material property (i.e bulk modulus K) is only achieved through a corresponding sacrifice of another property (here, shear modulus G)."

We have modified the sentence as suggested.

Report of the First Referee – XR10345M/Talapatra

In this paper, Talapatra et al discuss an experimental design approach based on Bayesian Optimization under Model Uncertainty and Bayesian Model Averaging to sequentially guide density functional theory (DFT) calculations for exploring the materials design space (MDS). They first perform high-throughput DFT calculations (a total of 403 calculations) on a materials class, referred to as MAX phases (M_2AX and M_3AX_2), and calculate the bulk and shear modulus from the elastic constants. They also represent each MAX compound using fifteen features. From this list, they consider six feature sets and each feature set is made of four features. For the single objective optimization, two cases were considered: maximize bulk modulus and (ii) minimize shear modulus. They start from different initial training data set sizes (N=2,5,10,15, and 20) and estimate the average number of calculations required to find the best compound that satisfy their optimization objective. They also set a budget of 80 calculations and select two calculations at a time. They also develop an approach for multiobjective optimization (maximize bulk modulus and minimize shear modulus) and demonstrate a method to find the Pareto front. The authors have also sufficiently reviewed the related research in the literature.

We thank the reviewer for his/her comments.

The novelty of this work is in the application of Bayesian Model Averaging (BMA) to estimate the mean and uncertainty for sequential experimental design. In the BMA, a committee of weighted predictive models is used as opposed to a single model and each model uses a subset of feature as defined by the authors. The relative models weights also change adaptively at the end of each iteration, which is very interesting. I do not consider their multiobjective optimization to be completely new, because of the related work in the literature (listed as one of my comments below).

We have rewritten the novelty statement. At the time of writing we were not aware of a MO-based framework in the context of optimal experimental design but have added a reference to Gopakumar et al. The other reference was not considered in the manuscript as there was no notion of Bayesian-based optimal experimental design.

This is an interesting work and the authors have performed rigorous simulations to justify the results. I have several technical questions for the authors, which I hope will improve the presentation of the results in the paper:

– On Page 5, the authors state that " ... to the best of the knowledge of these authors no prior work on optimal materials discovery considers the multi-objective." Here are two recent papers on multi-objective optimization: (i) Mannodi-Kanakkithodi et al Computational Materials Science 125 pp. 92-99 (2016) and (ii) Gopakumar et al Scientific Reports 8 article number 3738 (2018). How does the current work deviate from these two papers?

There are several papers on Multi-Objective optimization of materials. We think that the most relevant one is Gopakumar et al and we include a reference to this work. The key difference is the addition of the feature selection step.

– The idea of building a committee of models using different feature sets is not new. Random forests are one of the well-known examples in the machine learning literature. Can the authors compare the performance of their approach with the strategy akin to random forest (i.e., randomly choose subsets of features and build Gaussian process models for each randomly feature set)? This is important because without a comparison it is difficult to assess the efficacy of the current approach.

Here, we have considered a fixed set of feature sets in order to show the power of our approach in terms of both experimental design performance and also identifying the best model among the considered models. Note that selecting a single feature set based on the limited initial data that might be misleading is very common in the materials experimental design papers. The novelty of our approach is that we also consider uncertainty over the model space for experimental design. The original Random Forest algorithm is a model-free approach, with no notion of model probability. Therefore, it does not naturally lead to experimental design without carefully combining it with new probabilistic interpretation of the random forest results. In addition, as random forest is essentially a bagging approach, the prediction performance may not be desirable when we only have a very limited number of training samples in MDS applications.
We can indeed combine the idea of Random Forests, i.e. randomly choosing subsets of features and build Gaussian process models for each feature set as suggested, with our approach, where instead of simple averaging we still follow our approach by weighting each model based on its posterior probability given the observed data. This is especially useful when the space of considered models is too large. Moreover, a better approach is to even optimize the search in the model space instead of randomly picking them. In fact, one of the advantages of having model probabilities in a Bayesian approach is the ability to leverage it for a stochastic search in model space. We are currently working on developing this method on a parallel effort.

– The authors select two compounds at a time for update. How is the selection done?

The selection for the compound(s) to query is based on the optimal policy used: EI or EHVI. Thus the candidates with the maximum and second maximum EI/EHVI are selected for update. This has been made clear in the manuscript.

– The notation for Features in Table 1 and Figure 4 does not match. In Figure 4, I see rad and vol. But, I do not see it in Table 1.

rad and *vol* are quantities derived from the lattice parameters and did not offer any additional information. Consequently although they were initially considered, they were not used to build the feature sets and do not feature in Table 1.

– The authors visualize the correlation matrix in Figure 4, but they do not use this Figure for any feature selection. This figure is not adding anything to the manuscript. I would recommend the authors to put this Figure in the Supplemental document (after they fix my previous comment).

The figure has been moved to the Supplemental section

– In the results, the authors state that Feature F_2 wins consistently (in both single-objective problem). Why? What is so unique about m, Z, I_{dist} and e/a that is leading to this result? Where is the physical

insight?

In this work, the problem selected is one which is very well understood by material scientists. The initial 15 features were selected based off domain knowledge and are all known to affect to the bulk modulus. We found it especially interesting, that the lattice parameters (a,c), which were the most relatively difficult features to collate and were calculated via DFT, did not feature in the best feature set F_2. As noted in the manuscript,'*The C, m parameters are related to the bonding character. These are composition-weighted values of the empirical constants reported by Makino* et al.*[ref], who proposed that the bulk modulus* **K** *of elemental substances can be determined by the relation $K = Cr_{ps}^{-m}$; where r_{ps} is the effective pseudopotential radius. The $\frac{e}{a}$ ratio, which is the average number of itinerant electrons per atom, plays a significant role in the bonding of a solid and is closely related to the valence electron concentration C_v [ref]... The atomic number Z, which denotes the number of electrons is the foremost factor that determines the chemical bonding behavior of a material and defines its chemical properties. The weighted interatomic distance I_{dist} was calculated from the elemental values, which were sourced from the CRC Handbook of Chemistry and Physics [ref].*'. Thus, the best feature set F_2 is a combination of an empirical constant (m) based on the bonding character, the atomic number Z, and interatomic distance I_dist, and the $\frac{e}{a}$ ratio which are electronic structure properties.

– *What is a first-order BMA (BMA_1) and second-order BMA (BMA_2) Laplace approximation? What is the difference between them? How are they calculated? Why did the authors stop at second-order and not go further?*

First-order and second-order BMA refers to the level of approximation for the probability of the data given a model (details are provided in the last paragraph of Section II A). We employ this approximation(s) because the probability functions are not necessarily Gaussian-distributed and does not have a closed form. We stopped at a second-order approximation due to the considerable computational cost associated with higher-order expansions. Also, posterior probabilities are usually rather highly peaked, so stopping at second-order is considered reasonable. We found that either approximation yielded good results.

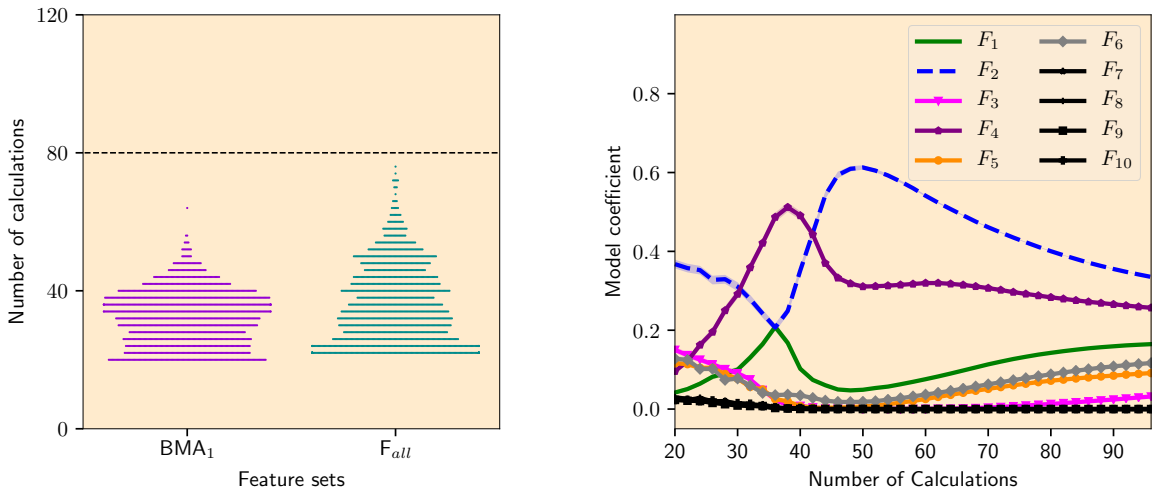
– *In almost all cases, F_2 performs better than BMA_1 or BMA_2 (eg., Figure 6c and Figure 9a). Why should one then even consider BMA as an approach if it not leading to improved results ("the major innovation")? I do not see the point.*

A posteriori, it is obvious that F_2 is the best feature set and it is thus not surprising that it performs better than BMA_1 or BMA_2 as in the universe of feature sets compared, F_2 contains the most useful information needed to navigate the multi-dimensional space being explored. However, this information is not known when we start carrying out the optimal exploration of the materials design space. Please note, however, that even F_2 is sub-optimal relative to some of its subsets. In Figures 12a and 12b we show how adding features without useful information can make the process of optimal discovery less efficient. The value of BMA-based adaptive feature selection is that it mitigates the effect of non-informative features by focusing on small feature sub-sets of the entire feature space and then using the Bayesian evidence to adaptively project the optimal exploration along the most informative directions.

– *On Page 10, the authors motivate the need for feature subsets and the challenges due to the high-dimensional space. In general, I agree with their assessment. But, I find that the argument is lacking specificity. Can the authors define high-dimensionality in the context of this work? Typically, high-dimensionality refers to p (# of dimensions) >> n (# of samples). In this work, $p > n$, only for $n < 15$. It will be interesting to check the performance of their approach by including all features at least for one example, where N=20. In my opinion, this is not high-dimensional.*

The authors agree with the reviewer's definition: high-dimensionality refers to p (# of dimensions) >> n (# of samples).

To showcase the utility of our BMA approach, we simulate a high-dimensional case by adding 16 non-informative random features, which we compose into subsets F_7, F_8, F_9, and F_{10}. We carry out two types of calculation using the larger set of 29 (13+16) features. First, we use the BMA_1 approach to find material with maximum K using F_1,...F_{10}; and we use the regular EGO-GP model to find the material with maximum K using all 29 features. The results for the same are plotted in Figure 1. Firstly, we see in Figure 1a, that in this case (an actual high dimensional case with a number of non-informative random features), the BMA approach outperforms using all features together. Additionally, tracking the model probabilities as in Figure



(a) A subfigure

(b) A subfigure

Figure 1: Representative results for single objective optimization – minimization of shear modulus for N=20 for the case of 29 features: a) average model probabilities for maximizing bulk modulus using BMA_1 and F_all b) swarm plots indicating the distribution of the number of calculations required for convergence using BMA_1 and F_{all}.

1b, shows us that the BMA approach effectively picks up the F_2 set as the best feature set, rejects the random feature sets F_7, ...F_{10} (average model probabilities are negligible) and performed better than using F_2 standalone (in Figure 5d in the manuscript).

– I do not see any where in the paper (or the Supplemental document) a Figure showing the performance of Gaussian Process Model vs the actual DFT result. What is the relative model quality when trained using F_1, F_2, ..., F_6, BMA_1 and BMA_2? There is no discussion on the relationship between relative model quality and the relative speed of optimization. This important discussion is lacking in the paper and without this result, I am not able to make any recommendation about this work.

A section comparing the performance of Gaussian Process Model vs the actual DFT result has been added to the supplemental material and is reproduced below:

Figure 2 shows the results for one example of maximization of bulk modulus for all the feature sets and the BMA approach. The example started with 10 measured values, and after forty measurements, there are now 50 materials with known bulk modulus (K). These are then used to train the regressor; and the predicted K and actual K (from DFT) are plotted for all of the materials (in green), including those in blue whose true K values are known (in blue). It can be seen that the model error is high for F_3, F_5 and F_6 and low for F_1, F_2, F_4, BMA_1 and BMA_2. This is akin to the results included in the manuscript. In figure 3, we visualize the bulk modulus estimated from GP model in all cases for the composition with the largest K value from DFT. Error bars indicate the variance of the GP model. It is seen that GP models based off both BMA_1 and BMA_2 get very close to the ground truth similar to F_2. Thus, inspite of poor performing models based off F_3, F_5 and F_6, the GP models based off the BMA approach perform as well as the best standalone model F_2

Also as noted by Balachandran et. al, 'Although we would prefer regressor models with lower error, we remark that high model error is likely to be a common situation in data-driven approach to materials design and discovery with small data sets and a vast unexplored search space.' We would like to note that in the context of Bayesian Optimization, the actual accuracy of the model used to predict the data is not the relevant metric to be used when evaluating models or next acquisition points, especially when only a limited number of training samples are available. The problem that BO is trying to solve is to find the arguments of the black-box optimization function that maximize the utility function of carrying out the observations

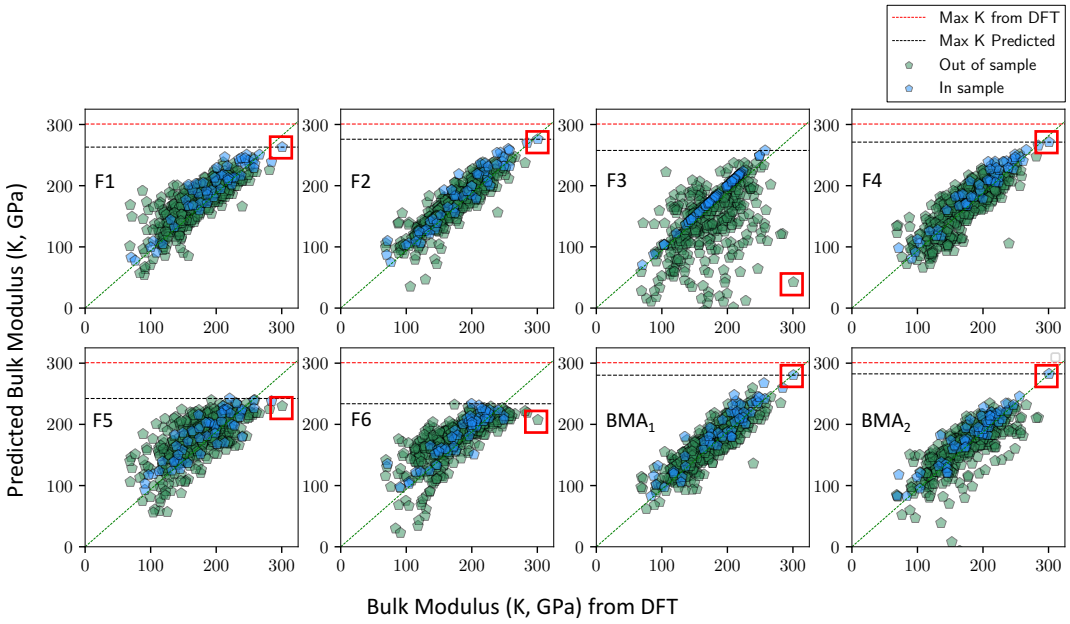


Figure 2: *In a single trial, after ten initial(N=10) and forty subsequent measurements, the GP model is fit to the n = 50 training points (in blue), and applied both to in-sample and out-of-sample data. Here, the red dashed line corresponds to the largest DFT value seen so far, and the black dashed line is the largest estimated value. Composition with the largest K value from DFT is highlighted (red square) to show the model error.*

in a yet-to-be-explored location of the design space. In 1-D optimization, a typical utility function is the so-called Expected Improvement, EI, which is constructed from the mean response of the GPR as well as the uncertainty associated with such response surface.

The EI policy balances exploration and exploitation and as such is not really concerned with how good the model is at predicting the mean response surface. Rather, EI provides an indication of where next to explore given the current model response and the uncertainty in that response. While the acquisition function uses information from the GPR, the ultimate goal is not to have a perfect predictor but rather to optimally find optimal regions in the space to optimize. An acquisition policy based solely on the mean response surface of a given model only exploits current knowledge and precludes any exploration of unknown regions in the feature space

In our framework we always take the DFT calculations as ground truth. Some models tend to be better than other models in the Bayesian sense when confronted with the DFT data, but here we are not concerned with how good such predictions are, as the underlying assumption is that once an optimal region is identified, it will be definitely and conclusively observed through a query to a DFT-based 'oracle'.

We would also like to note that the measured model weights constitute indirect information about model quality: the feature sets are weighed based on their Bayesian importance given the data available and the model constructed from them, including not only its mean response surface but also the predicted uncertainty in unexplored regions. A better model, with higher weight would be one that is able to predict more or less accurately the observations, but with less uncertainty in regions yet to be explored. Feature sets with higher weights correspond to models that are better predictors of optimal performance, based on the acquisition function used. Indirectly, one could assume that the same models are better at predicting the actual data as well.

– *This work comes across as a methods paper and there are no new materials or physical insights. Therefore, I am not convinced that this work is suited for publication in the Physical Review Materials journal.*

Given these major concerns, I do not recommend the current version of the manuscript for publication

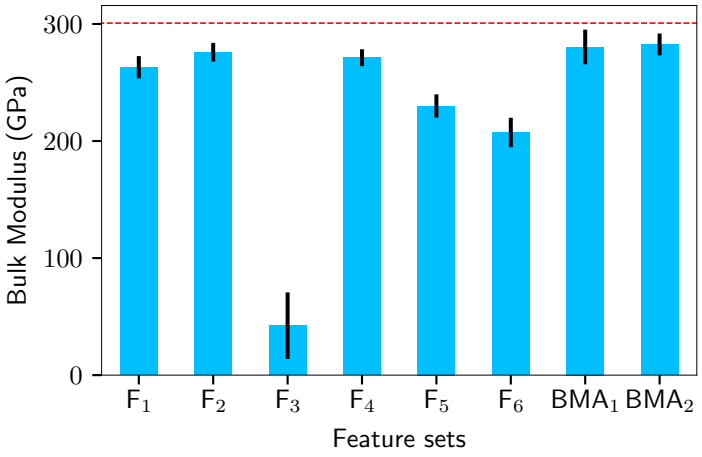


Figure 3: *Bar plots showing estimated bulk modulus for the composition with the largest bulk modulus from DFT for all cases in a single trial, after ten initial(N=10) and forty subsequent measurements, the GP model is fit to the $n = 50$ training points, and applied both to in-sample and out-of-sample data. Error bars indicate the variance of the GP model. Red dashed line corresponds to target maximum bulk modulus from DFT.*

in the Physical Review Materials journal.

We understand the concerns of the reviewer. However, our paper was deemed as within scope by the editors of PRM before it was sent for review. We have also obtained clarification from the journal that papers on new methods for materials are within the scope of PRMaterials (there is a dedicated section, Section M3-A "Development of new methods for materials). Additionally, the *physical insight*, in our opinion, was included a priori, in the initial 15 features considered. While the paper is essentially about the presentation of (what we think is) a novel framework, we have added a section that discusses in some detail, a posteriori, the underlying reasons for why some feature sets are more effective than others at guiding the exploration of the model materials design space.

Finally, the authors thank the referees for their valuable suggestions. We believe they have added necessary detail and clarity to this work.