A Perspective on Bayesian Methods Applied to Materials Discovery and Design

Raymundo Arróyave[a,b,c], Danial Khatamsaz[b], Brent Vela[a], Richard Couperthwaite[a], Abhilash Molkeri[a], Prashant Singh[d], Duane D. Johnson[d,e], Xiaoning Qian[f], Ankit Srivastava[a], Douglas Allaire[b]

[a]Department of Materials Science and Engineering, Texas A&M University, College Station, TX 77843

[b]J. Mike Walker '66 Department of Mechanical Engineering, Texas A&M University, College Station, TX 77843

[c]Wm Michael Barnes '64 Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77843

[d]Ames Laboratory, U.S. Department of Energy, Iowa State University, Ames, IA 50011, USA

[e]Department of Materials Science & Engineering, Iowa State University, Ames, IA 50011, US

[f]Department of Electrical and Computing Engineering, Texas A&M University, College Station, TX 77843

Abstract

For more than two decades, there has been increasing interest in developing frameworks for the accelerated discovery and design of novel materials that could enable promising and transformative technologies. The Integrated Computational Materials Engineering (ICME) program called for integrating computational tools to establish linkages along process-structure-property-performance (PSPP) chains. The Materials Genome Initiative called for integrating experiments and computations within data science frameworks as a strategy to accelerate the materials development cycle. While these frameworks and paradigms have been quite influential, traditional ICME or data science-based approaches tend to have some limitations, mainly when querying the materials space is costly and very little information is available. Bayesian methods are more suitable in this context due to their efficiency gains. To this end, the materials discovery problem is framed as a Bayesian Optimization (BO). Different examples in which BO has been applied to solve materials discovery problems are presented. The methods/examples discussed include BO under model uncertainty, multi-information source BO, multi-objective and multi-constraint BO, and batch BO. Bayesian Materials Discovery is a promising area of research that is likely to become more influential as more attention is put on autonomous materials discovery platforms. Therefore, a discussion is provided on the potential development of such methods to increase the ability of existing platforms in materials discovery. The ultimate goal is to pave the way to autonomous materials discovery.

Keywords: Bayesian Optimization, Alloy Design, ICME, Materials Genome Initiative

Introduction

Until well into the 20th century, the materials discovery/development cycle was mostly carried out using highly heuristic approaches lacking well-defined frameworks. The lack of a systematic approach to materials design meant that very little could be learned/transferred from successful materials development programs. To address this issue, Olson[1] extended Cyril Stanley Smith's ideas on materials as hierarchical (multi-scale) systems[2] and proposed that materials systems were amenable to design using engineering design principles informed by scientific understanding. In that work, Olson recognized the importance of developing process-structure-property-performance (PSPP) relationships to establish causal connections between processing conditions and materials performance. Such links could then be used to understand how a given material could be modified to optimize its performance.

Olson's idea to use computational tools to establish such linkages was embraced as the underlying framework guiding Integrated Computational Materials Engineering (ICME)[3], as an approach to accelerate the insertion of new materials into technological applications. Provided such linkages could be transformed into quantitative relationships, PSPPs could, in turn, be inverted to optimize the materials to meet specific performance goals. Having established quantitative PSPP relationships, the actual

optimization over a given materials design space could be carried out with a number of algorithmic optimization schemes[4, 5]. Unfortunately, simulation-driven materials design requires solving many significant challenges, such as effectively linking different simulation tools [6] and addressing the considerable uncertainties in the models, model parameters, and the experiments used to validate them[7]. This problem is compounded by the high computational cost of many of the most precise and predictive tools at the disposal of ICME practitioners.

While the ICME's focus was on linking computational PSPP model chains, the Materials Genome Initiative (MGI)[8] put forward an aspirational program to accelerate the materials development cycle through the tight integration of computations and experiments through data analytics techniques. MGI has been embraced by almost every single community within the materials science discipline, and the field is undergoing an exponential expansion in the number of efforts taking place around the world involving the leveraging of data science, machine learning (ML), and artificial intelligence (AI) to accelerate our understanding of materials spaces. While many types of problems are being addressed using these tools, much work under the MGI paradigm requires a significant amount of data. Consequently, some of the most prominent applications of MGI principles to materials discovery have exploited extensive data, either obtained from experimental combinatorial synthesis of material libraries[9] or high-throughput (HTP) computations[10].

ICME and MGI-inspired approaches have resulted in numerous successful cases of materials discovery and design. However, such approaches tend to be limited, particularly considering that typical materials development efforts are always carried out under stringent resource constraints. ICME methods[11, 12], for example, (i) build and exploit process-structure-property-performance (PSPP) relationships[7] at a considerable computational expense; (ii) do not readily incorporate data from experiments within their framework; and (iii) are sequential, which means that they tend to be deployed by evaluating materials design choices one-at-a-time. Traditional HTP combinatorial computational[10] and experimental[13, 14] approaches, on the other hand, are (i) incapable of dealing with the high dimensionality and complexity of typical materials design spaces; and (ii) are 'one-shot' or 'open loop' schemes without a built-in iterative framework, and are unable to prescribe future actions given the current state of knowledge[15]. *Moreover, traditional ICME and HTP approaches are suboptimal in resource utilization*, as they do not guarantee that each proposed choice is optimal.

The limitations described above can potentially be overcome by using inherently resource-aware frameworks that provide a principled way to augment our knowledge of the state of a materials design space iteratively. Because of this, one of the best approaches to accelerate the design of a materials space is to frame this problem as an optimal experimental design task within a Bayesian Optimization (BO) setting[16-21]. As a result, the number of works that deploy BO in materials optimization tasks is growing rapidly[22-30]. Yet, most of the works published thus far have focused on translating the materials discovery problem into a single black-box optimization challenge. Here, we revisit some recent work in which traditional BO frameworks have been modified for them to be better adapted to solving problems in materials science, mainly when one can get access to multiple tools/models/experiments to query the material space and one (potentially) has access to experimental or computational platforms capable of operating in batch/parallel mode.

A Short Tutorial on Bayesian Optimization

Any Bayesian methods rely on the deployment of the Bayes theorem to update our prior knowledge (beliefs) about a system once new information has been acquired[31]. Bayesian Optimization (BO) is a sequential design strategy for increasing (optimizing) the system's performance or augmenting our knowledge about it. It relies on Bayesian updates of prior beliefs through acquiring new information. BO methods tend to be most effective when confronted with functions that: are costly to compute, are not analytical (are 'black boxes'), cannot be evaluated precisely due to intrinsic noise, and/or provide no facilities for estimating their gradients. We note that many resources explain in great detail what BO is, and the reader is kindly directed to some of these sources[32-34] if they want to understand these frameworks at a deeper level.

BO methods have many essential features that make them a highly effective optimization scheme. As a consequence of the sequential exploration of a design space, BO methods construct a statistical model of the ground truth being explored that is orders of magnitude cheaper to query than the ground truth itself. The primary ingredients of any BO algorithm are (1) a statistical (surrogate) model used to predict the outcomes of experiments yet to be carried out; (2) a policy that prescribes in a deterministic, principled, and unambiguous manner the best action to take to meet a given (set of) objective(s) given our current state of knowledge of the system[35]. Given an initial dataset, $\mathcal{D}$, a typical sequential BO scheme would then follow the following general algorithm (see also **Fig. 1**):

input: initial dataset $\mathcal{D}$
repeat
    $x \leftarrow POLICY(\mathcal{D})$
    $y \leftarrow OBSERVE(x)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(x, y)\}$
until termination condition reached
return $\mathcal{D}$

Algorithm 1: General Bayesian Optimization algorithm[32].

While, in principle, many types of statistical models can be used to simulate a given experimental space, the overwhelming majority of BO approaches use Gaussian Processes (GPs)[36-38]. A GP is a nonparametric statistical model for the objective function, *f*, defined as a stochastic process $p(f) = \mathcal{GP}(f; \mu, K)$ with a mean function, μ, and a covariance function (or kernel) K. The reason why $\mathcal{GP}s$ are used in BO is because of their mathematical properties (including smoothness, controllable modeled correlation among observed points, etc.). Besides their intrinsic capability to predict uncertainty in our knowledge of the objective function over the design space, the critical characteristic of $\mathcal{GP}s$ is the notion of a measure of the degree of correlation between observations: observations that are close tend to yield similar values of the objective function. The degree (and type) of correlation is encoded by the kernel, K. The latter can take a wide range of forms (as long as it has some mathematical properties) depending on the nature of the system being optimized.

The second ingredient of BO is the *policy* (or utility function) used to make optimal sequential decisions on the queries. The policy for decision-making is constructed from the posterior distribution of the surrogate model for the design space. Over the past decades, dozens of policies or *acquisition functions* have been developed and deployed in BO frameworks. Yet, all acquisition functions address the tension between exploitation (querying the design space in locations predicted to be optimal by the surrogate model) and exploration (querying the design space where there is uncertainty about the surrogate model). These acquisition functions are much cheaper to evaluate than the ground truth/objective function. When $\mathcal{GP}s$ are used as the surrogate function, many acquisition functions can be computed as analytical functions of the properties of the $\mathcal{GP}s$, such as the mean response and its variance[16, 32, 33]. Examples of such acquisition functions include Expected Improvement (EI), Probability of Improvement (PI), Upper Confidence Bound (UCB), etc. Unfortunately, the effectiveness of different policies is highly problem-dependent, and it is thus impossible to know *a priori* the optimal policy for a given problem.

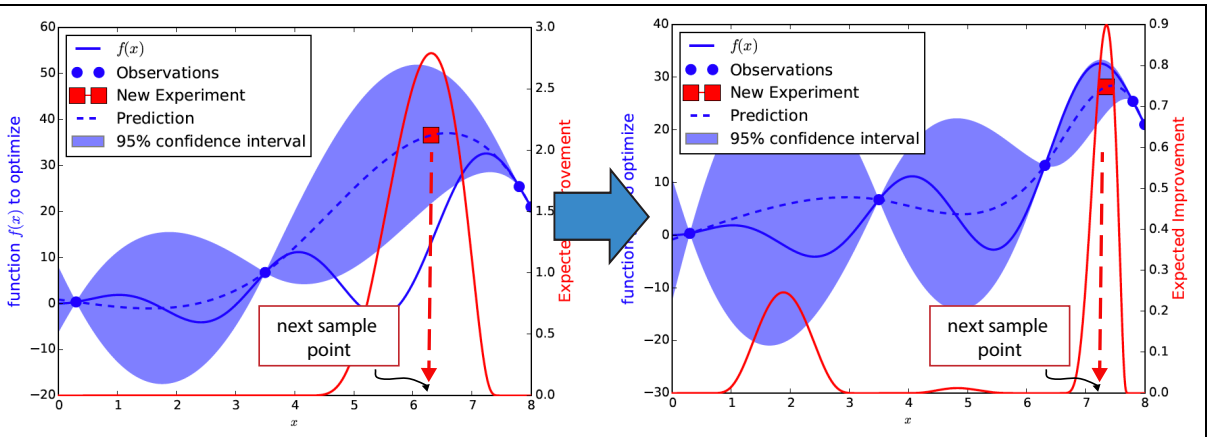

Figure 1. *Schematic illustration of Bayesian Optimization (BO): from a limited number of observations on a system (blue solid line), a stochastic model (dashed blue line and shaded area) is built. The following observation is determined by accounting for the tradeoff between the exploitation of the current knowledge and the exploration of the unknown regions of the design domain x. In this case, Expected Improvement (EI) is the metric used, and thus the policy falls within the Efficient Global Optimization (EGO) framework [35]. Reproduced from Talapatra et al. [39].*

Several approaches have employed BO in a materials science context, for example, in molecular design and discovery, smart and additive manufacturing design optimization of materials, such as piezoelectric materials, catalysts, and structural materials[29, 30, 39-41]. They have shown dramatic improvements in the efficiency with which experimental/computational resources are utilized to find materials with optimal properties/performance, even within exclusively experimentally-based frameworks [42]. Yet, most approaches to materials discovery that deploy BO-based frameworks are still limited. BO-based materials discovery has been based on translating existing methods to optimize general 'black box' functions. Yet, in materials science, one typically has multiple ways of querying a design space. Most approaches tend to deploy BO to solve single-objective optimization problems when realistic materials development efforts require optimality across many dimensions of the performance space while satisfying several constraints. Moreover, most BO-based materials discovery approaches developed to date do not take advantage of the quickly evolving high-throughput experimental and computational platforms. As will be seen below, there are many opportunities to tailor BO methods to materials discovery.

Bayesian Optimization under Feature Importance Uncertainty

In typical realistic materials discovery challenges, there is typically little to no prior knowledge of the behavior of the objective function to explore. For example, we may not know the 'shape' of the (potentially multi-objective) materials response space. Thus, it may not be possible to develop a realistic model for the correlation (or distance) between points in the design space, encoded in a $\mathcal{GP}$ as the covariance function, K. Furthermore, in many cases, there may not be sufficient information to gauge which features are most correlated with the optimization objective(s). This lack of knowledge stems from cost limitations on doing enough experiments due to the complexity of the objective(s) or the involvement of many features. For example, initially, we may not know which features of a given material are the most correlated with a particular property or performance metric that we want to optimize. In some early examples of BO applied to materials science, feature selection (or feature engineering) was proposed as a preliminary step before initiating a BO optimization loop. Unfortunately, in most practical applications, the feature and/or model selection step is highly uncertain due to the sparseness in the data available. This inability to narrow down the feature space compromises the ability of BO to optimally sample a given design space since the performance of BO methods tends to degrade as the dimensionality of the problem increases[43].

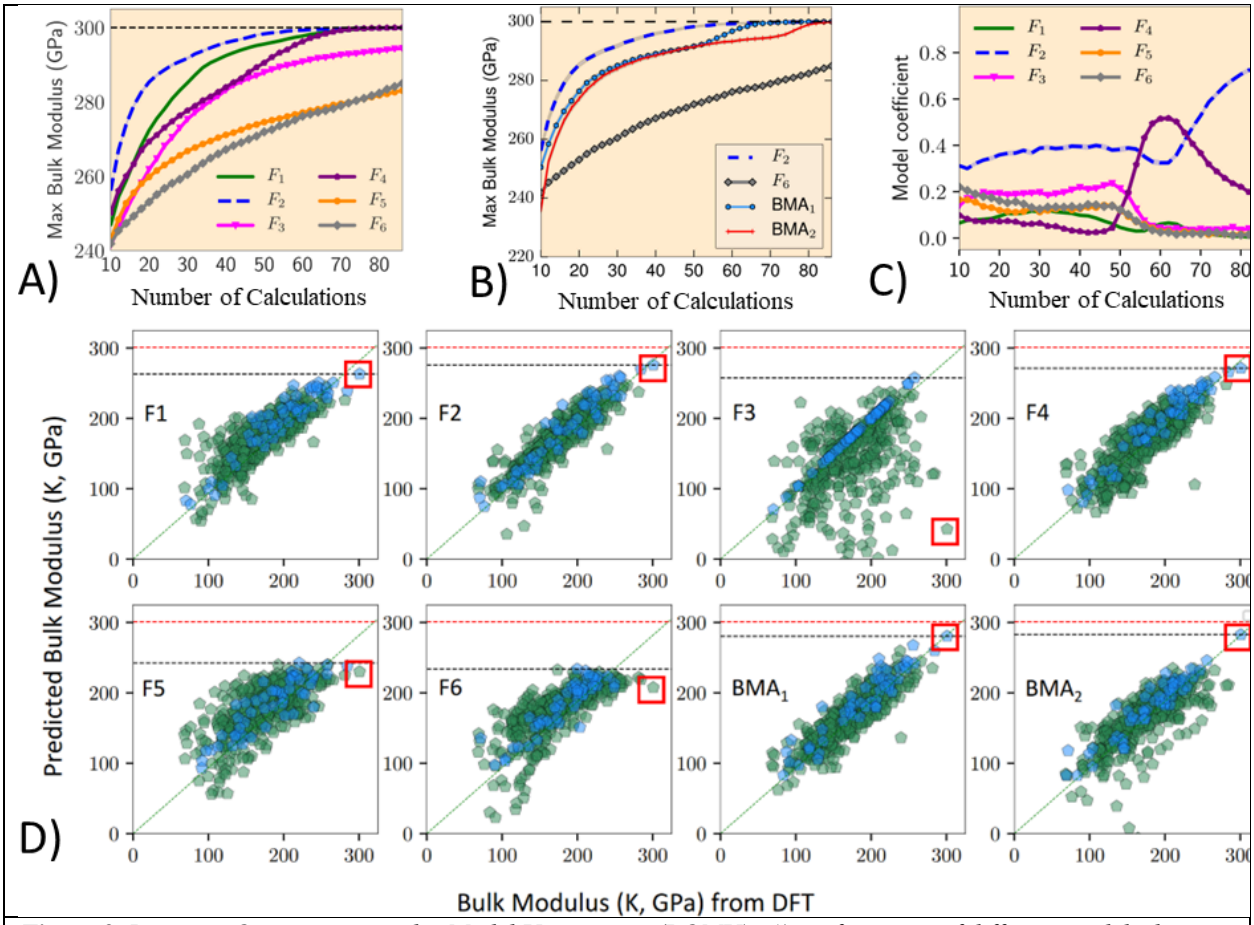

Figure 2. *Bayesian Optimization under Model Uncertainty (BOMU): A) performance of different models during a sequential BO task. By partitioning features into six different subsets, six different models are created to estimate the same quantity of interest (QoI). B) comparison between the best and the worst models and two effective models estimated through Bayesian Model Averaging (BMA). C) Evolution of the model probability during the sequential experimental campaign. As more data are added to the system, model probability, and importance change. D) Benchmarking results of using each model to learn and predict the bulk modulus of a set of samples. The red square shows the largest bulk modulus found in DFT calculations of queried samples in the optimization process, and the dashed red line shows the largest value ever seen. [39].*

In recent work[39, 44], it has been shown that it is possible to carry out feature selection while carrying out an optimal sequential experimental design. In that setting, the feature space was partitioned into feature subspaces that, together with specific covariance functions, K, constitute different models describing the objective space. To account for this model uncertainty, it is possible to weigh all the possible models by their probability of being the true model. When the weighing is based on the Bayesian model evidence, such model assembly is known as Bayesian Model Averaging (BMA)[45]. By incorporating BMA within BO, Bayesian Optimization under Model Uncertainty (BOMU) can simultaneously carry out feature selection and BO. In BOMU (see **Fig. 2**), the feature space available for exploration is partitioned into subsets corresponding to different hypotheses as to what features most correlate with the objective. Each feature set-kernel/covariance function combination is treated as an individual model, and each member of the model set $\{\mathcal{M}_1 \cdots \mathcal{M}_N\}$ would vary in their ability to predict the outcome of the observations that have yet to be made over the design space.

To take action (or query/observation), the policy or utility of observing every yet-to-be-queried point is evaluated as the weighed experiment utility, using the probability of a given model being the correct model as the weights. After an observation is made, the updated dataset is used to evaluate the probability

of the models, and a new cycle begins. As a consequence of this model selection step, models would change their probability or weight based on how good they are at predicting the augmented dataset. An interesting aspect of BOMU is that it has been observed that the model probabilities (**Fig. 2**) tend to evolve in a non-monotonic manner, with the importance of some models decreasing and increasing in importance as a function of the stage of an experimental campaign. BOMU is also capable of immediately detecting uninformative feature sets, reducing their probability to close to zero early in the experimental campaign.

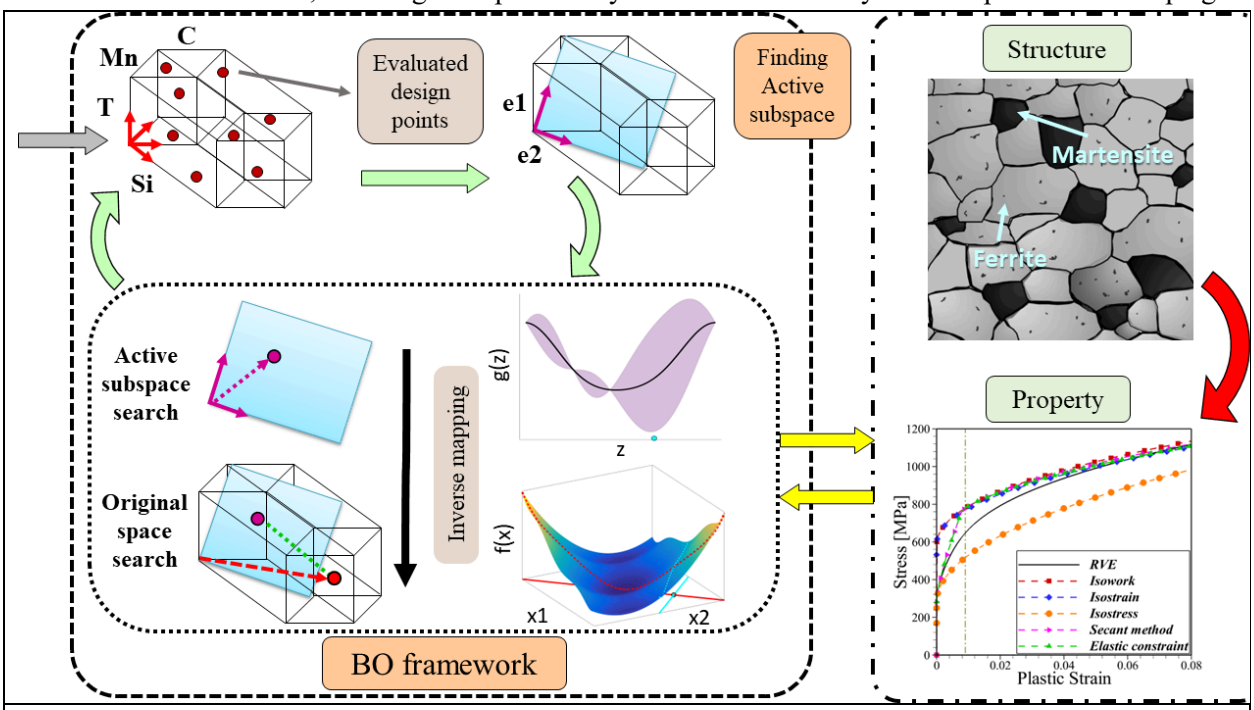

***Figure 3**. Implementation of adaptive active subspace method within a Bayesian Optimization (BO) framework. The basic idea is to find the active subspace, i.e., the directions in the material design space that give the most significant variation in the target property, by using the available data at every stage of the optimization task. Next, the design space is mapped to the active subspace, and the first step of the BO framework is applied to find the 'next best point' to evaluate within the active subspace. The 'next best point' is then mapped back to the original design space by implementing a second BO step. Finally, the design space is evaluated at this best point. This new data is added to the framework for the next iteration. In the example shown inside the BO framework window, the objective function f(**x**) is represented as g(z) in the active subspace. Thus, instead of variables x_1 and x_2, the objective function is estimated on a 1-dimensional active subspace space of z. This figure summarizes the implementation of the active subspace method for designing dual-phase steels in the presence of multiple information sources estimating the same property at different levels of fidelity[46].*

While BOMU carries out feature selection by evaluating the posterior probabilities of competing models, there are other approaches to BO with automated model selection. Having a feature space, the framework described by Malkomes and collaborators[47-49] will be capable of dynamically selecting optimal models (with different covariance functions/kernels) through the use of Bayesian optimization over the model space by using the concept of 'kernel of kernels', which can be used to measure the distance between the covariance functions (kernels) of Gaussian Process regression models. The framework generates arbitrary candidate kernel functions by utilizing the concept of 'kernel grammars'[50]. We note that BOMU can potentially explore the feature and model space and the policy used. In practice, different feature set-model-policy tuples would be evaluated, and their weight would be modified depending on their effectiveness.

BOMU is a practical approach when there is uncertainty in the feature space. However, a significant limitation is that the partitioning of the feature space is done in an ad-hoc manner. Thus there is always a risk of not capturing the truly important features/degrees of freedom. An alternative approach is the *active*

subspace method[51], which is a technique to identify the directions in a given design space that have the most significant change relative to a given objective, effectively reducing the dimension of the problem. Once a subspace has been identified, an acquisition function/policy is used to evaluate the best point within this lower-dimensional space. Finally, an observation is prescribed in the actual higher-dimensional design space by inverse mapping from the lower to the higher dimensional spaces.

Compared to other approaches for automated feature selection within BO, an advantage of the active subspace method is the ability to dynamically change the effective dimensionality of the problem as the experimental campaign progresses. An extensive evaluation of the active subspace method found that the effective dimension of a given materials design problem did not exceed two for most sequential experimental campaigns [46]. Crucially, the effective dimensions of the problem changed continuously through the design sequence, pointing to the fact that attempting to carry out feature selection before the BO stage of a materials design task may not be a robust strategy. The dynamic nature of the feature importance found in this work is in line with what was found in BOMU, where the weight of some models changed dramatically as the BO sequence progressed. These observations imply that in materials design/optimization problems, only a few features/degrees of freedom are active at any given time.

Multi-Information Source Bayesian Optimization

While mapping a materials discovery task to a 'black box' optimization can significantly accelerate the materials discovery process, such a simplistic approach does not take advantage of a crucial aspect of materials research. For example, 'black box' BO approaches do not account for the fact that it is always possible to use several methods to query a particular materials space. Such methods include, for example, different models (with different cost and fidelity) connecting different linkages along PSPP relationships or experiments measuring different attributes of the material to be investigated. As such, 'black box' BO is not tailored to materials design tasks. On the other hand, traditional ICME-based approaches rely on quantitative PSPP chains, whose linkages can be evaluated using computational models. A significant challenge for these approaches is that they tend to consider a single model per chain—i.e., they ignore the possibility of combining the predictive power of different models simultaneously. A further limitation is that most typical ICME approaches are not capable of combining experiments and simulations since they use experimental information only as a way to verify or validate the models used. Integrating multiple information sources into a materials discovery task could provide significant advantages.

Other fields in science and engineering have already developed sophisticated approaches to combine multiple information sources within optimization schemes. In these approaches, multiple information sources may be used to approximate the behavior of an expensive-to-query 'ground truth .'In such multi-fidelity approaches[52], it is possible to combine multiple information sources to have better information about the 'ground truth', at a reduced cost. **Fig. 4** shows how one can transfer knowledge among information sources to enhance learning through Bayesian information fusion. Practically, such knowledge transfer is constructed by learning and exploiting the statistical correlations among the different information sources. The underlying assumption of this type of information fusion is the existence of correlation among different information sources and between a given information source and the ground truth. In materials science, we can assume that such an assumption is warranted, given that all information sources for a given materials design task predict different aspects of the material's behavior and are thus causally correlated. The common goal in any application using multi-information source approaches is reducing associated costs. By collecting information from different sources, which are usually lower in fidelity but cheaper to evaluate, it is possible to obtain information about the ground truth (highest fidelity model) via the correlation between the models and the ground truth. Consequently, it allows for a cheaper learning process and prevents conducting expensive experiments, in contrast to single model approaches which have no choice but to perform costly experiments to learn an objective function.

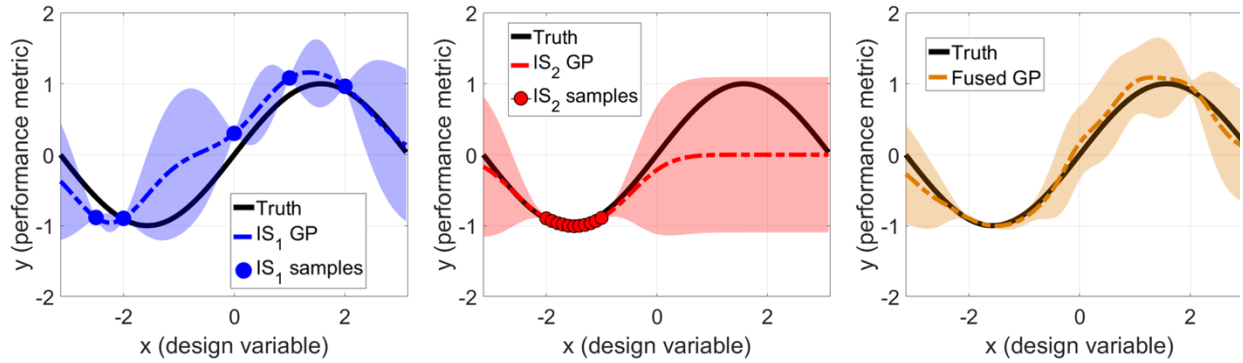

Figure 4: Left: Fitting a GP to information source IS 1. Center: Fitting a GP to information source IS 2. Right: Fused GP from IS 1 and IS 2 GPs. Learning the relationship between IS 1 and IS 2 could improve the fused GP even more in the region where we have no samples of IS 2.

In many materials design applications, collecting experimental data is considered the most accurate way of estimating a material's property of interest. However, this comes at the expense of substantial experimental costs (money, time, computational resources, etc.), which makes it almost impossible to complete sufficient numbers of experiments to make any conclusions about the optimal design, particularly in high-dimensional feature spaces. In such cases, using less accurate approaches (exploiting lower fidelity computational models, for example) to estimate a quantity of interest (QoI) and correlating it to the ground truth (highest fidelity model or experimental data) can be beneficial, cost-wise. It is often the case that a set of QoIs can be estimated through different kinds of experiments varying in cost and accuracy (fidelity), yielding different results. It is also usually the case that computational models are exploited to estimate a set of QoIs to avoid the challenges of collecting expensive experimental data. Computational models representing the same system of materials are also different in cost and fidelity since different assumptions and mathematical simplifications are made to create such models. Assuming that any information source contains some useful information about the QoIs, collecting this information from several lower fidelity models helps to reduce the overall design costs. Another point is that the fidelity of information sources may vary through the input space; thus, exploiting multiple information sources is beneficial because more useful sources can be deployed depending on the region of interest.

Recently, such a multi-information source framework was used to predict the crystallization tendency of polymers. In work by Venkatram and collaborators[53], it was argued that predicting a new polymer crystallinity is not trivial due to experimental challenges such as having the necessary infrastructure and sample preparation. On the other hand, group contribution methods are less accurate but much cheaper to obtain. Thus, they fused data from group contribution methods and fewer experimental data available from the literature, obtaining a more accurate model without performing any expensive experiments. Another recent application was the work by Pilania *et al.* on the bandgap prediction of solids using multi-fidelity approaches[54]. In that work, a larger number of samples from a cheap model (PBE) and fewer samples from a more expensive model (HSE06) are used to obtain a fused model and make predictions. The work by Tran and collaborators[18] present an application in optimizing the bulk modulus in a ternary composition space using multi-information source BO. A similar work employed multi-information source BO in geometry optimization of CO via employing different DFT functionals as information sources[55].

A general multi-information source BO framework has been proposed and deployed to optimize normalized strain hardening rate in designing dual-phase steels[56, 57]. In this framework, one first constructs intermediate surrogate models (GPs for convenience due to their mathematical properties) for each available information source. For example, these information sources could be different models connecting materials design inputs and outputs to varying levels of fidelity and physical rigor. Then, using a so-called reification[58] procedure, a fused model is constructed from all information sources by learning the statistical correlation among all the sources, including the 'ground truth'. The fused model, described as a GP, $\mathcal{GP}(f;\mu,K)$ is then used to evaluate the value of the acquisition function on each candidate point

in the experimental space yet to be queried. Having established the best point to query, the multi-information source Bayesian optimization (MISBO) framework then selects the next information source to query, given an information source fidelity, uncertainty, and cost.

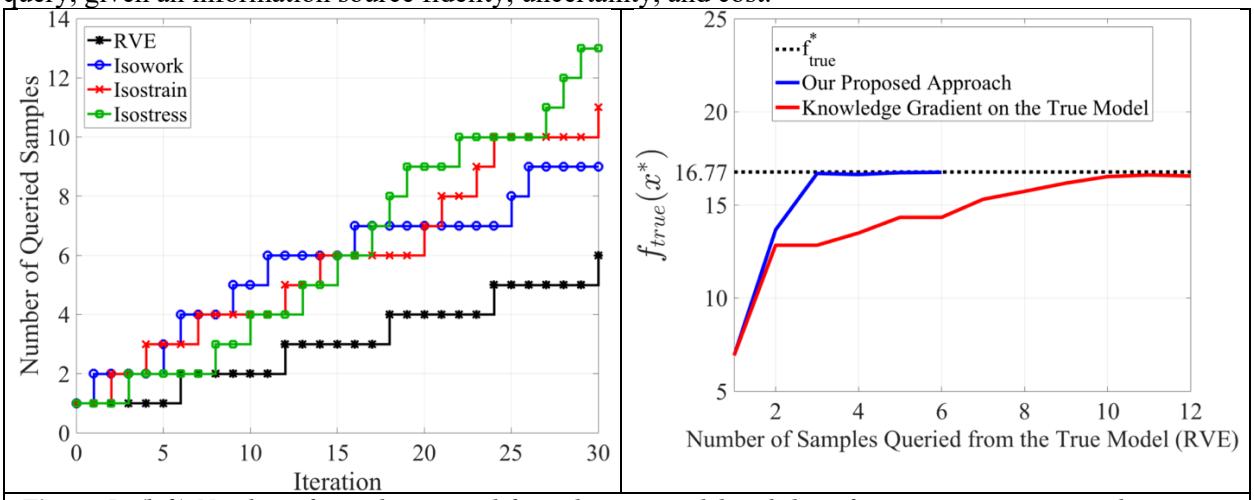

Figure 5: *(left) Number of samples queried from the true model and the information sources in each iteration. (right) The optimal solution was obtained by the proposed approach and by applying the knowledge gradient on a GP of only the true data for different numbers of samples queried from the true model[56].*

MISBO is much more efficient at solving materials optimization problems[56] than even state-of-the-art BO methods that only query the ground truth—see **Fig. 5**. Since every information source is mapped to a GP, it has been shown that MISBO can be used in problems in which explicit PSPP connections can be exploited by mapping different linkages along the PSPP chain to GPs and then connecting the inputs and outputs of the different GPs along a model chain using simple statistical correlations. As such, MISBO can be mapped directly to ICME frameworks[57], the advantage being a much more efficient utilization of the available information. An additional benefit of MISBO is that it is possible to account explicitly for the cost of a given information source, which is crucial when there are stringent budget constraints[57]. Other groups [59] have presented similar approaches. Pilania *et al.* [54], for example, have shown how it is possible to combine low-fidelity approximate predictions for materials properties to supplement high-fidelity, expensive computational simulations, arriving at accurate predictive models at a relatively low cost. MISBO differs from other multi-fidelity approaches in that it does not assume the existence of a hierarchy among the different models as they are all *simultaneously* fused to predict the behavior of the 'ground truth'.

Bayesian Optimization under Multiple Objectives and Constraints

By far, most approaches to BO have focused on deploying acquisition functions tailored to balance the exploration and exploitation of single-objective design spaces. The reason for this is that in single-objective optimization problems, the objective of the decision-making process is clear: to optimize the value of a target system response. However, in materials science, it is often the case that a materials design task has multiple objectives to optimize. For example, one would like materials to be strong and ductile or materials with higher energy storage capacity and very short charging times. In many cases, these objectives compete against each other, resulting in performance tradeoffs: strong materials are not necessarily the most ductile, for example.

A multi-objective optimization problem can thus be solved in two different ways:

One strategy is to construct a single utility function from different objectives[60]. Such a utility function is built based on the preferences of the decision-maker(s) and thus has some degree of subjectivity. Once this single objective function has been constructed, the optimization is carried out in the usual way, using BO or non-BO methods. Building a single utility function implies that (subjective) preferences do not

change as more information is acquired about the system. If the utility function changes because of changes in the preferences for different objectives, the optimization would have to be carried out from scratch.

An alternative approach is to not establish any subjective preference for the different individual objectives and instead try to discover the Pareto front in the multi-objective space[61]. The Pareto front is the set of non-dominated solutions being optimal along with one of the directions in the objective space. Once a Pareto front is discovered, a preference model (or utility function) could be constructed from the Pareto set—any utility function constructed with dominated solutions (i.e., solutions enclosed by the Pareto front) would always be suboptimal. While discovering the Pareto front allows more flexibility in the decision-making process, the discovery of the Pareto set is challenging, particularly in high-dimensional spaces.

In the context of BO, the problem is paradoxical: on the one hand, one is interested in discovering the Pareto front. On the other hand, however, sequential BO policies always propose a single action to take at any given time. Therefore, one must construct a scalar function that measures the utility of carrying out an experiment at a point in the design space, regardless of the number of objectives. There are not as many acquisition functions in multi-objective BO as in single-objective BO. One of the most widely used metrics in multi-objective BO is the Expected Hyper-Volume Improvement (EHVI) [62]. This metric measures the expected change in the Pareto front (towards the direction of non-domination) when a query is made at a given point. Under this policy, the optimizer can identify the Pareto front under resource constraints. This approach has been used before in multi-objective materials optimization—see **Fig. 6**[63]. The efficiency of multi-objective BO methods tend to outperform other optimization approaches, particularly when the number of objectives increases.

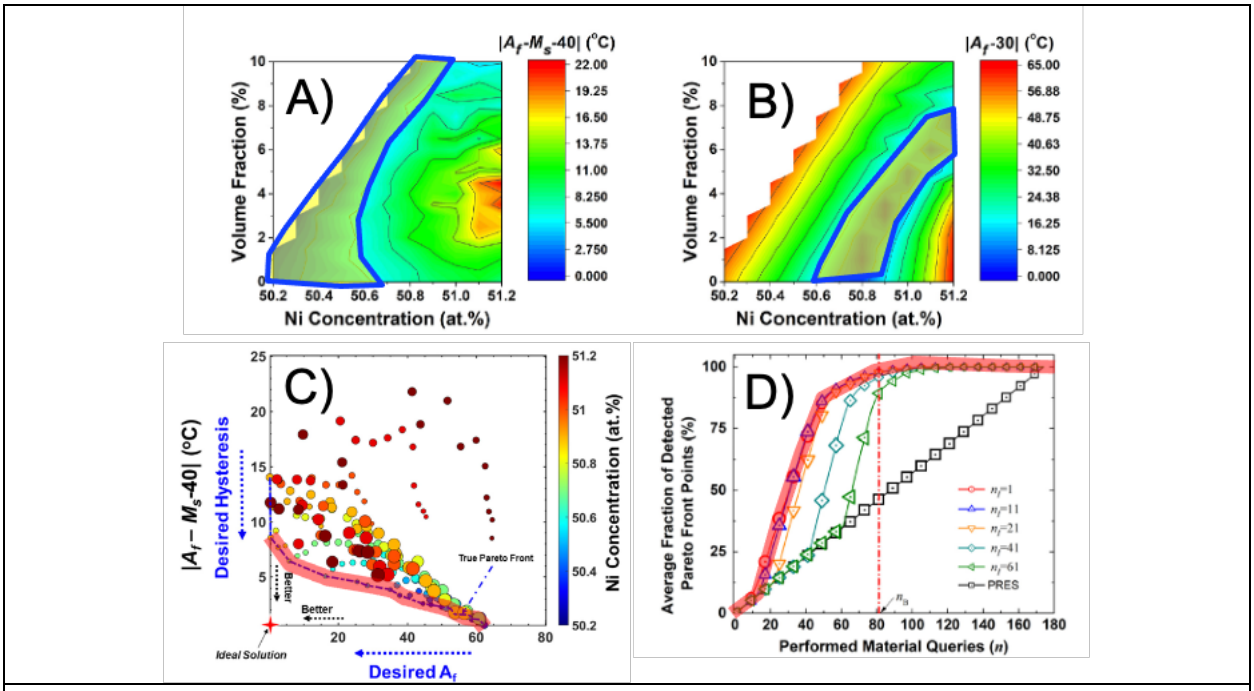

***Figure 6**: (A and B): Multi-objective space in the optimization of chemical composition and precipitate volume fraction to optimize the performance of a precipitation-strengthened shape memory alloy (SMA). C) Pareto front identified using an optimal Bayesian policy. D) Optimal BO approaches outperform other policies and can discover a larger fraction of the Pareto front with fewer queries to the system[63].*

While the ability to solve multi-objective problems in a Bayesian optimal setting brings materials discovery to a more realistic setting, one must consider many problems in materials science in which the performance requirements are so stringent that the discovery process is constraint-dominated[64]. When there is insufficient information about the materials design space, one needs to allocate resources to identify

the feasible space (i.e., the region in the design space with solutions that satisfy all the performance constraints) and find the Pareto set. In some cases, the constraints can be evaluated in terms of easy and fast-to-evaluate analytical models constructed as a function of the design variables. However, in many cases, assessing the constraints constitutes a costly endeavor. For example, evaluating whether an alloy is resistant to oxidation may involve a complex experimental protocol lasting 24-100 hours.

The identification of the feasible space can be mapped to a Bayesian Classification (BC) task, in which the boundary separating the feasible and infeasible regions in the materials design space is found in the least number of queries possible. Ideally, such a BC task would be set up using multiple information sources, as in MISBO[56, 57]. In recent work, it has been shown how it is possible to combine multi-information source BO with multi-information source BC to solve classification/optimization materials design problems—see **Fig. 7**. While the results of this approach are promising, there are significant opportunities for further development. For example, it is unclear how to optimally divide resources between optimization and classification in a principled, algorithmic manner. One approach could be first identifying the feasible space, then allocating all the remaining resources to identify the Pareto set. However, this approach may be too risky when resources are limited, as the total (experimental/computational) budget could be spent well before the exploration of the Pareto set begins. Other approaches may switch between classification and optimization as the campaign progresses, resulting in suboptimal performance. Perhaps building criteria for uncertainty measures in the feasible/unfeasible boundary and the Pareto set could be constructed to optimize the process.

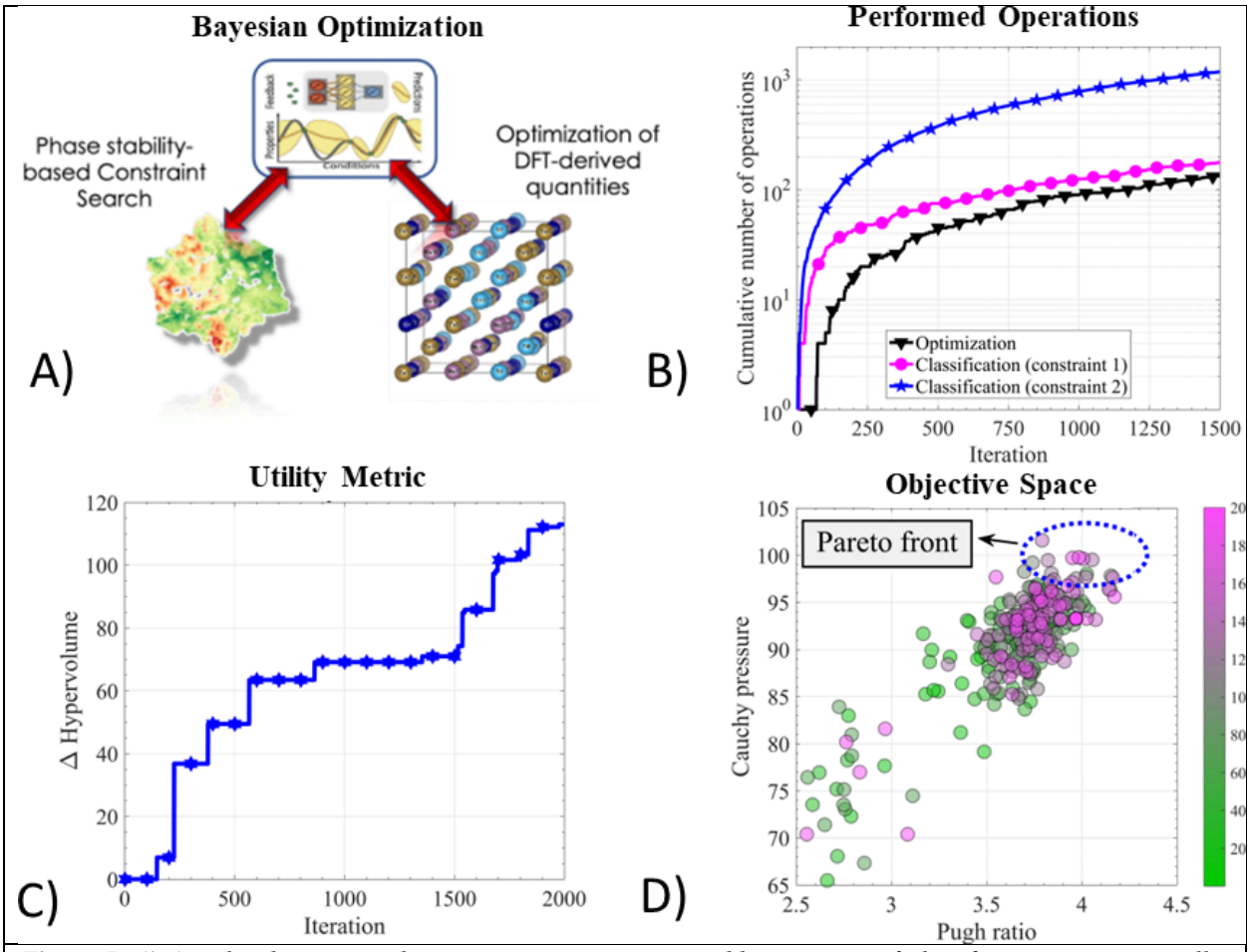

***Figure 7**: A) A multi-objective, multi-constraint optimization problem consists of identifying regions in an alloy space that satisfy some minimal performance requirements (i.e., meet performance constraints) while simultaneously optimizing a set of intrinsic properties. B) operations performed during the design process are*

categorized as either classification or optimization depending on querying constraint models or objective functions. C) As more iterations are completed in the optimization loop, better non-dominated designs are found. Hypervolume is a metric to measure the quality of discovered non-dominated designs, and its improvements are directly related to discovering a better Pareto front. D) Queried samples and non-dominated designs were queried during the process of optimizing the Pugh Ratio and Cauchy pressure as an indication of the ductility of an alloy. The colormap is used to illustrate the order of queries [65].

Bayesian Optimization over High-throughput Parallel Computational/Experimental Settings

One of the significant drawbacks of ICME and modern BO-based materials optimization approaches is their sequential nature and their inability to make use of high-throughput workflows pervasive in HPC environments [66, 67], as well as modern synthesis [68, 69] and characterization [70-72] facilities. In addition, traditional ICME approaches suffer from their non-optimality already, so in the context of materials design, providing BO with the ability to make *multiple optimal decisions* at once is perhaps the only viable approach toward parallelizing a design framework [73].

The approach for Batch BO by Joy and collaborators [73] is among the most promising thus far. Their method is based on the premise that when only a few data points are available, it is impossible to have much confidence in the covariance structure of the GP models describing the correlations between data. A pragmatic solution is to *make no assumption* regarding the shape (e.g., smoothness) of the response surface but instead assume that any possible covariance structure, within bounds, is possible. The hyperparameters describing the model of the observations are sampled as extensively as possible, and then multiple recommendations for the next point to explore are elicited. To generate the batch of recommendations, the possible recommendations found are clustered using a k-medoids [74] clustering approach. The number of clusters matches the size of the batch of recommendations required. This approach provides excellent flexibility and is relatively easy to integrate with any method that already uses Gaussian Process functions to model the objective function. **Fig. 8,** shows a demonstration of Batch BO and an example in which Batch BO schemes have been integrated with a multi-information source fusion framework developed previously to accelerate the design of dual-phase steels [56, 75]. The preliminary results show that using batch queries; it is possible to dramatically improve the rate of convergence towards a global optimum and reduce the uncertainty of (and increase confidence in) the optimization framework.

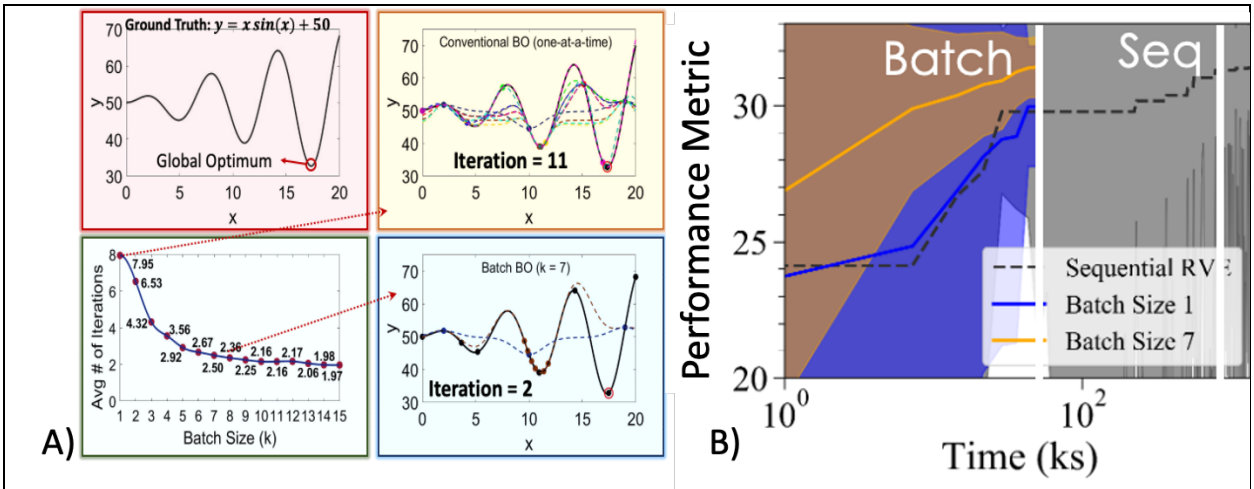

Figure 8*: A) Demonstration of batch Bayesian BO. The sequential approach is much less efficient than batch querying. B) BBO results in orders of magnitude faster materials optimization than standard sequential BO[76].*

Conclusions and Outlook

When resources are limited, it is clear that Bayesian methods offer the best approach to identifying optimal **and** feasible regions of a materials design space. While off-the-shelf BO approaches can, in principle, be applied to materials discovery and design, very few materials problems can be mapped to the optimization of 'black box' functions. In this short overview, it has been shown that there are modifications to BO that are better suited to materials science problems. Combining multiple information sources with high throughput (HTP) experimental and/or computational platforms could be deployed to solve materials discovery problems with multiple objectives and constraints. While promising, there is much more opportunity for further development.

The performance of BO depends on the adopted surrogate models approximating underlying objective functions, for which Gaussian Processes (GPs) are typically used. However, in many materials discovery applications, the inherent smooth assumptions of GPs may not always hold. Therefore, more adaptive and flexible Bayesian surrogate models in BO may need to be developed to overcome the weaknesses of widely used GP-based methods when faced with relatively high dimensional design space or non-smooth patterns of objective functions. Recently, it has been shown that Bayesian Multivariate Adaptive Regression Splines (BMARS)[77] and Bayesian Additive Regression Trees (BART)[78] are flexible stochastic surrogate ML models that can provide better performance when applying BO to complex materials spaces[79].

In the case of multi-objective BO, it is not uncommon for the different objectives to exhibit some degree of correlation. In the case of materials optimization, this makes sense because other properties or performance metrics of a given material are ultimately connected since they most likely share the same underlying process-structure (PS) relationships. Unfortunately, in traditional multi-objective BO, each objective is modeled individually, and such correlations are not considered when making predictions using GPs. Recently, Deep Gaussian Processes (DGP) have been proposed as ML surrogate models that can be used to build multivariate surrogates to catch this correlation and improve the prediction ability of GPs[80-82]. A DGP uses multiple layers of GPs to map from a set of features to latent spaces layer by layer to reach a target objective space. This can potentially be useful in representing different objectives in materials discovery as it is probable that a set of properties in a given material system to be correlated to some degree.

Another challenge in the deployment of Bayesian optimization approaches is the ability to handle qualitative data. In real materials discovery tasks, quantitative and qualitative design variables are required to fully define the materials systems. Existing BO approaches mostly work with quantitative data and cannot involve categorical data or catch the correlation between qualitative and quantitative design variables. Recently, some works have proposed employing Latent variable Gaussian Process (LVGP) models to map categorical variables into continuous-variable latent spaces[83]. Developing mixed-variable BO frameworks has been a subject of recent research[84-89] and more platforms capable of handling both types of design variables are expected to be developed in the coming years.

A natural extension of the mathematical frameworks presented here is the deployment of automated, closed-loop materials discovery platforms[90]. While somewhat limited, the first examples of automated scientific discovery platforms consist of integrated robotic platforms that execute actions that are, in turn, selected using either BO or Reinforcement Learning (RL) approaches. The field of autonomous experimentation is evolving in a highly accelerated manner, and such platforms have already been deployed to solve problems involving a wide range of materials classes and applications[91-93].

Further developments in autonomous materials discovery will likely arise from advances in the decision support tools—BO being one of the dominant paradigms due to its flexibility and applicability in the sparse data regime—used to guide the platform towards a particular objective. Currently, for example, human intuition is seldom incorporated into these platforms. Yet, expert human intuition may be highly valuable in guiding the discovery process. A significant challenge, in this case, would be how to algorithmically introduce human expert opinion within a formal closed-loop BO framework. In other contexts, human-in-the-loop BO (HITL-BO) has been implemented by eliciting human experts to express their preference among BO-generated choices[94]. Similar approaches could potentially be used in the context of BO-

assisted materials discovery. However, there is no work (to the best of the authors' knowledge) that has exploited such ideas as of this writing.

Another critical issue is that human-driven materials discovery tends to be a very dynamic and flexible process. Yet, the BO policies, actions, and experiments to query the materials design space are hard-coded. A further limitation is that traditional BO approaches tend to operate under the assumption that the problem itself (i.e., the definition of objectives, constraints, degrees of freedom) is fixed. However, it is conceivable that a non-trivial materials discovery loop requires significantly more flexibility when defining the problem space itself. For example, performance metrics that were initially considered to be objectives may be more suitable to be considered as constraints, constraints may need to be softened or hardened, or new degrees of freedom may need to be incorporated into the design space.

While there are many more limitations to current BO-based materials discovery platforms, solutions that provide enhanced flexibility and functionality will likely be gradually deployed as the community pursues further autonomy. It is thus likely that future platforms will be able to gather information, develop new knowledge, and evolve into fully autonomous systems. In the near future, further advancements to these autonomous materials discovery platforms will involve innovations in the BO space, robotics, and data integration. While progress has been slow, advancement is accelerating dramatically, and autonomous experimentation in materials science is likely to become the dominant paradigm for exploring complex, multi-objective and multi-dimensional materials design spaces in the decades to come.

Acknowledgments

The authors would like to acknowledge the groups of Profs. Ed Dougherty, Dimitris Lagoudas, Ibrahim Karaman, and Seyede Fetemeh Ghoreishi. The authors also acknowledge Drs. Sahin Boluki, Anjana Talapatra, and Jaylen James. The authors acknowledge the financial support from NSF through Grants No. NSF-CMMI-1663130, NSF-CISE-1835690, NSF-CDSE-2001333 and NSF-DMREF-2119103. BV acknowledges the support of NSF through Grant No. NSF-DGE-1545403. The authors also wish to acknowledge the support from the U.S. Department of Energy (DOE) ARPA-E ULTIMATE Program through Project DE-AR0001427. Calculations used to demonstrate many of the BO methods described in this work were carried out at the High-Performance Research Computing (HPRC) facility at Texas A&M University.

Conflict of Interest: The corresponding author states that there is no conflict of interest

Data Availability Statement:

Data shown in this paper can be made available upon reasonable request.

References

1. Olson, G.B., Science, 1997. **277**(5330): p. 1237-1242.
2. Smith, C.S., *A search for structure. Selected essays on science, art and history*. 1981.
3. M., P.T., J.E. Allison, D.G. Backman, M.C. Boyce, M. Gersh, E.A. Holm, R. LeSar, M. Long, A. Powell IV, J.J. Schirra, D. Demania Whitis, and C. Woodward, *Integrated computational materials engineering: a transformational discipline for improved competitiveness and national security*. 2008, National Academies Press: Washington, D. C.
4. David E., G., *Genetic Algorithms in search, optimization and machine learning*. 1989: Addison-Wesley.
5. Fonseca, C.M. and P.J. Fleming. *Genetic algorithms for multiobjective optimization: Formulation, discussion and generalization*. in *Fifth International Conference on Genetic Algorithms*. 1993. San Mateo, CA: Morgan Kaufmann.
6. Voorhees, P. and G. Spanos, TMS, Warrendale, PA, 2015. **14**.

7. Panchal, J.H., S.R. Kalidindi, and D.L. McDowell, Computer-Aided Design, 2013. **45**(1): p. 4-25.
8. National Science and Technology Council, *Materials genome initiative for global competitiveness*. June 2011, Executive Office of the President: Washington, DC. p. 1-18.
9. Engstrom, J.R. and W.H. Weinberg, AIChE journal, 2000. **46**(1): p. 2-5.
10. Curtarolo, S., G.L. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nature materials, 2013. **12**(3): p. 191.
11. Allison, J., D. Backman, and L. Christodoulou, Jom, 2006. **58**(11): p. 25-27.
12. Allison, J., Jom, 2011. **63**(4): p. 15.
13. Koinuma, H. and I. Takeuchi, Nature materials, 2004. **3**(7): p. 429.
14. Xiong, Z., Y. He, J.R. Hattrick-Simpers, and J. Hu, ACS combinatorial science, 2017. **19**(3): p. 137-144.
15. Talapatra, A., S. Boluki, P. Honarmandi, A. Solomou, G. Zhao, S.F. Ghoreishi, A. Molkeri, D. Allaire, A. Srivastava, and X. Qian, Frontiers in Materials, 2019. **6**: p. 82.
16. Brochu, E., V.M. Cora, and N. De Freitas, arXiv preprint arXiv:1012.2599, 2010.
17. Astudillo, R. and P. Frazier, *Bayesian Optimization of Composite Functions*, in *Proceedings of the 36th International Conference on Machine Learning*, C. Kamalika and S. Ruslan, Editors. 2019, PMLR: Proceedings of Machine Learning Research. p. 354--363.
18. Tran, A., J. Tranchida, T. Wildey, and A.P. Thompson, The Journal of Chemical Physics, 2020. **153**(7): p. 074705.
19. Liang, Q., A.E. Gongora, Z. Ren, A. Tiihonen, Z. Liu, S. Sun, J.R. Deneault, D. Bash, F. Mekki-Berrada, S.A. Khan, K. Hippalgaonkar, B. Maruyama, K.A. Brown, J. Fisher Iii, and T. Buonassisi, npj Computational Materials, 2021. **7**(1): p. 188.
20. Frazier, P.I., *Bayesian Optimization*, in *Recent Advances in Optimization and Modeling of Contemporary Problems*. p. 255-278.
21. Arróyave, R. and D.L. McDowell, Annual Review of Materials Research, 2019. **49**(1): p. 103-126.
22. Sharpe, C., C.C. Seepersad, S. Watts, and D. Tortorelli. *Design of Mechanical Metamaterials via Constrained Bayesian Optimization*. in *ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. 2018.
23. Wahab, H., V. Jain, A.S. Tyrrell, M.A. Seas, L. Kotthoff, and P.A. Johnson, Carbon, 2020. **167**: p. 609-619.
24. Xie, Y., C. Zhang, H. Deng, B. Zheng, J.-W. Su, K. Shutt, and J. Lin, ACS Applied Materials & Interfaces, 2021. **13**(45): p. 53485-53491.
25. Pedersen, J.K., C.M. Clausen, O.A. Krysiak, B. Xiao, T.A.A. Batchelor, T. Löffler, V.A. Mints, L. Banko, M. Arenz, A. Savan, W. Schuhmann, A. Ludwig, and J. Rossmeisl, Angewandte Chemie International Edition, 2021. **60**(45): p. 24144-24152.
26. Greenhill, S., S. Rana, S. Gupta, P. Vellanki, and S. Venkatesh, IEEE Access, 2020. **8**: p. 13937-13948.
27. Agarwal, G., H.A. Doan, L.A. Robertson, L. Zhang, and R.S. Assary, Chemistry of Materials, 2021. **33**(20): p. 8133-8144.
28. Peng, J., J.K. Damewood, J. Karaguesian, R. Gómez-Bombarelli, and Y. Shao-Horn, Joule, 2021. **5**(12): p. 3069-3071.
29. Xue, D., P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, Nature communications, 2016. **7**: p. 11241.
30. Xue, D., P.V. Balachandran, R. Yuan, T. Hu, X. Qian, E.R. Dougherty, and T. Lookman, Proceedings of the National Academy of Sciences, 2016. **113**(47): p. 13301-13306.
31. Good, I.J., I. Hacking, R. Jeffrey, and H. Törnebohm, 1966.
32. Garnett, R., *Bayesian Optimization*. 2022: Cambridge University Press.
33. Frazier, P.I., arXiv preprint arXiv:1807.02811, 2018.
34. Shahriari, B., K. Swersky, Z. Wang, R.P. Adams, and N. De Freitas, Proceedings of the IEEE, 2015. **104**(1): p. 148-175.

35. Jones, D.R., M. Schonlau, and W.J. Welch, Journal of Global optimization, 1998. **13**(4): p. 455-492.
36. Williams, C. and C. Rasmussen, Advances in neural information processing systems, 1995. **8**.
37. Planas, R., N. Oune, and R. Bostanabad, Journal of Mechanical Design, 2021. **143**(11).
38. Eweis-Labolle, J.T., N. Oune, and R. Bostanabad, Journal of Mechanical Design, 2022. **144**(9).
39. Talapatra, A., S. Boluki, T. Duong, X. Qian, E. Dougherty, and R. Arróyave, Physical Review Materials, 2018. **2**(11): p. 113803.
40. Yuan, R., Z. Liu, P.V. Balachandran, D. Xue, Y. Zhou, X. Ding, J. Sun, D. Xue, and T. Lookman, Advanced materials, 2018. **30**(7): p. 1702884.
41. Tran, K. and Z.W. Ulissi, Nature Catalysis, 2018. **1**(9): p. 696-703.
42. Xue, D., D. Xue, R. Yuan, Y. Zhou, P.V. Balachandran, X. Ding, J. Sun, and T. Lookman, Acta Materialia, 2017. **125**: p. 532-541.
43. Raponi, E., H. Wang, M. Bujny, S. Boria, and C. Doerr. *High dimensional Bayesian optimization assisted by principal component analysis*. in *International Conference on Parallel Problem Solving from Nature*. 2020. Springer.
44. Talapatra, A., S. Boluki, T. Duong, X. Qian, E. Dougherty, and R. Arróyave, arXiv preprint arXiv:1803.05460, 2018.
45. Wasserman, L., Journal of mathematical psychology, 2000. **44**(1): p. 92-107.
46. Khatamsaz, D., A. Molkeri, R. Couperthwaite, J. James, R. Arróyave, A. Srivastava, and D. Allaire, Materials & Design, 2021. **209**: p. 110001.
47. Malkomes, G. and R. Garnett. *Automating Bayesian optimization with Bayesian optimization*. in *Advances in Neural Information Processing Systems*. 2018.
48. Malkomes, G., C. Schaff, and R. Garnett, Advances in Neural Information Processing Systems, 2016. **29**.
49. Schlessinger, L., G. Malkomes, and R. Garnett. *Automated model search using bayesian optimization and genetic programming*. in *Workshop on Meta-Learning at Advances in Neural Information Processing Systems*. 2019.
50. Duvenaud, D., J. Lloyd, R. Grosse, J. Tenenbaum, and G. Zoubin. *Structure discovery in nonparametric regression through compositional kernel search*. in *International Conference on Machine Learning*. 2013. PMLR.
51. Ghoreishi, S.F., S. Friedman, and D.L. Allaire, Journal of Mechanical Design, 2019. **141**(7).
52. Fernández-Godino, M.G., C. Park, N.-H. Kim, and R.T. Haftka, arXiv preprint arXiv:1609.07196, 2016.
53. Venkatram, S., R. Batra, L. Chen, C. Kim, M. Shelton, and R. Ramprasad, The Journal of Physical Chemistry B, 2020. **124**(28): p. 6046-6054.
54. Pilania, G., J.E. Gubernatis, and T. Lookman, Computational Materials Science, 2017. **129**: p. 156-163.
55. Herbol, H.C., M. Poloczek, and P. Clancy, Materials Horizons, 2020. **7**(8): p. 2113-2123.
56. Ghoreishi, S.F., A. Molkeri, A. Srivastava, R. Arroyave, and D. Allaire, Journal of Mechanical Design, 2018. **140**(11): p. 111409.
57. Khatamsaz, D., A. Molkeri, R. Couperthwaite, J. James, R. Arróyave, D. Allaire, and A. Srivastava, Acta Materialia, 2021. **206**: p. 116619.
58. Allaire, D. and K. Willcox. *Fusing information from multifidelity computer models of physical systems*. in *2012 15th international conference on information fusion*. 2012. IEEE.
59. Chen, S., Z. Jiang, S. Yang, and W. Chen, AIAA Journal, 2016. **55**(1): p. 241-254.
60. Wang, H., M. Olhofer, and Y. Jin, Complex & Intelligent Systems, 2017. **3**(4): p. 233-245.
61. Ngatchou, P., A. Zarei, and A. El-Sharkawi. *Pareto multi objective optimization*. in *Proceedings of the 13th International Conference on, Intelligent Systems Application to Power Systems*. 2005. IEEE.

62. Emmerich, M.T., A.H. Deutz, and J.W. Klinkenberg. *Hypervolume-based expected improvement: Monotonicity properties and exact computation*. in *Evolutionary Computation (CEC), 2011 IEEE Congress on*. 2011. IEEE.
63. Solomou, A., G. Zhao, S. Boluki, J.K. Joy, X. Qian, I. Karaman, R. Arróyave, and D.C. Lagoudas, Materials & Design, 2018. **160**: p. 810-827.
64. Arróyave, R., S. Gibbons, E. Galvan, and R. Malak, JOM, 2016. **68**(5): p. 1385-1395.
65. Khatamsaz, D., B. Vela, P. Singh, D.D. Johnson, D. Allaire, and R. Arroyave, Acta Materialia (submitted), 2022.
66. Jain, A., G. Hautier, C.J. Moore, S.P. Ong, C.C. Fischer, T. Mueller, K.A. Persson, and G. Ceder, Computational Materials Science, 2011. **50**(8): p. 2295-2310.
67. Curtarolo, S., W. Setyawan, G.L. Hart, M. Jahnatek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, and O. Levy, Computational Materials Science, 2012. **58**: p. 218-226.
68. Ortiz, B.R., J.M. Adamczyk, K. Gordiz, T. Braden, and E.S. Toberer, Molecular Systems Design & Engineering, 2019. **4**(2): p. 407-420.
69. Moorehead, M., K. Bertsch, M. Niezgoda, C. Parkin, M. Elbakhshwan, K. Sridharan, C. Zhang, D. Thoma, and A. Couet, Materials & Design, 2020. **187**: p. 108358.
70. Potyrailo, R.A. and I. Takeuchi, Measurement Science and Technology, 2004. **16**(1): p. 1.
71. Ludwig, A., npj Computational Materials, 2019. **5**(1): p. 70.
72. Oellers, T., V.G. Arigela, C. Kirchlechner, G. Dehm, and A. Ludwig, ACS Combinatorial Science, 2020. **22**(3): p. 142-149.
73. Joy, T.T., S. Rana, S. Gupta, and S. Venkatesh, Knowledge-Based Systems, 2020. **187**: p. 104818.
74. Kaufman, L. and P.J. Rousseeuw, Y. Dodge, Ed, 1987: p. 405-416.
75. Ghoreishi, S.F., A. Molkeri, R. Arróyave, D. Allaire, and A. Srivastava, Acta Materialia, 2019. **180**: p. 260-271.
76. Couperthwaite, R., A. Molkeri, D. Khatamsaz, A. Srivastava, D. Allaire, and R. Arròyave, JOM, 2020. **72**(12): p. 4431-4443.
77. Friedman, J.H., The annals of statistics, 1991. **19**(1): p. 1-67.
78. Chipman, H.A., E.I. George, and R.E. McCulloch, The Annals of Applied Statistics, 2010. **4**(1): p. 266-298.
79. Lei, B., T.Q. Kirk, A. Bhattacharya, D. Pati, X. Qian, R. Arroyave, and B.K. Mallick, npj Computational Materials, 2021. **7**(1): p. 1-12.
80. Hebbal, A., M. Balesdent, L. Brevault, N. Melab, and E.-G. Talbi, Optimization and Engineering, 2022.
81. Sauer, A., R.B. Gramacy, and D. Higdon, Technometrics, 2021: p. 1-15.
82. Damianou, A. and N.D. Lawrence, *Deep Gaussian Processes*, in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, M.C. Carlos and R. Pradeep, Editors. 2013, PMLR: Proceedings of Machine Learning Research. p. 207--215.
83. Zhang, Y., S. Tao, W. Chen, and D.W. Apley, Technometrics, 2020. **62**(3): p. 291-302.
84. Zhang, H., W. Chen, A. Iyer, and D.W. Apley, arXiv preprint arXiv:2207.04994, 2022.
85. Zhang, Y., D.W. Apley, and W. Chen, Scientific reports, 2020. **10**(1): p. 1-13.
86. Xu, H., C.-H. Chuang, and R.-J. Yang. *Mixed-Variable Metamodeling Methods for Designing Multi-Material Structures*. in *ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. 2016.
87. Pelamatti, J., L. Brevault, M. Balesdent, E.-G. Talbi, and Y. Guerin, Journal of Global Optimization, 2019. **73**(3): p. 583-613.
88. Manson, J.A., T.W. Chamberlain, and R.A. Bourne, Journal of Global Optimization, 2021. **80**(4): p. 865-886.
89. Iyer, A., Y. Zhang, A. Prasad, S. Tao, Y. Wang, L. Schadler, L.C. Brinson, and W. Chen. *Data-Centric Mixed-Variable Bayesian Optimization for Materials Design*. in *ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. 2019.

90. Häse, F., L.M. Roch, and A. Aspuru-Guzik, Trends in Chemistry, 2019. **1**(3): p. 282-291.
91. Stach, E., B. DeCost, A.G. Kusne, J. Hattrick-Simpers, K.A. Brown, K.G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, and I. Foster, Matter, 2021. **4**(9): p. 2702-2726.
92. Flores-Leonar, M.M., L.M. Mejía-Mendoza, A. Aguilar-Granda, B. Sanchez-Lengeling, H. Tribukait, C. Amador-Bedolla, and A. Aspuru-Guzik, Current Opinion in Green and Sustainable Chemistry, 2020. **25**: p. 100370.
93. Tabor, D.P., L.M. Roch, S.K. Saikin, C. Kreisbeck, D. Sheberla, J.H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, and H. Tribukait, Nature Reviews Materials, 2018. **3**(5): p. 5-20.
94. Zhou, Y., Y. Koyama, M. Goto, and T. Igarashi, arXiv preprint arXiv:2010.03190, 2020.