

Experiment Design Frameworks for Accelerated Discovery of Targeted Materials Across Scales

Raymundo Arroyave ^{1,*}, Anjana Talapatra ¹, Shahin Boluki ², Pejman Honarmandi ¹, Alexandros Solomou ³, Seyedeh Fatemeh Ghoreishi ⁴, Douglas Allaire ⁴ Xiaoning Qian ², Edward Dougherty ², and Dimitris C. Lagoudas ^{1,3}

¹ Department of Materials Science & Engineering, Texas A&M University, 3003 TAMU, College Station, Texas -77843

² Department of Electrical & Computer Engineering, Texas A&M University, College Station, Texas -77843

³ Department of Aerospace Engineering, Texas A&M University, College Station, TX 77843

⁴ Department of Mechanical Engineering, Texas A&M University, College Station, Texas -77843

Correspondence*: Corresponding Author
raaroyave@tamu.edu

ABSTRACT

Over the last decade, there has been a paradigm shift away from labor-intensive and time consuming materials discovery methods and materials exploration through informatics approaches is gaining traction at present. Current approaches are however typically centered around the idea of achieving this exploration through high-throughput experimentation/computation. Such approaches, however, do not account for the practicalities of resource constraints which eventually result in bottlenecks at various stage of the workflow. Regardless of how many bottlenecks are eliminated, the fact that ultimately a human must make decisions about what to do with the acquired information implies that HT frameworks face hard limits that will be extremely difficult to overcome. Recently, this problem has been addressed by framing materials discovery as an optimal experiment design problem. In this article, we discuss the need for optimal experiment design, the challenges in its implementation and finally discuss some successful examples of materials discovery via experiment design.

Keywords: Materials discovery, efficient experiment design, Bayesian optimization, information fusion, Materials Informatics, Machine learning

1 INTRODUCTION

Historically, the beginning of materials research centered around learning to use the elements and minerals discovered in nature. The chief challenge at the time was the separation of the pure metal from the mined ore which lead over time to the science of metallurgy - the foundation of current day materials research. Then, humans discovered that these pure metals could be combined to form alloys, followed by the

principles of heat treatments - advances that shaped history; since the ability to invent new and exploit known techniques to use metals and alloys to forge weapons for sustenance and defense was instrumental in the success, expansion and migration of early civilizations. Additionally, there is evidence that the oft quoted sequence of copper-tin bronze-iron which lend their names to the ‘ages’ of human progress, occurred in different parts of the world, sometimes even simultaneously Tylecote and Tylecote (1992). Thus, the desire to harvest materials from nature and use them to improve the quality of life is a uniquely human as well as universal trait.

With the acceleration of scientific advances over the last few centuries, man has moved on from developing applications based on available materials, to demanding materials to suit desired applications. Science and technology are continuously part of this contentious chicken and egg situation where it is folly to prioritize either - scientific knowledge for its own sake or the use of science as a tool to fuel new applications. Regardless, the majority of the scientific breakthroughs from the materials perspective have resulted from an Edisonian approach and guided primarily by experience, intuition and to some extent, serendipity. Further, bringing the possibilities suggested by such discoveries to fruition takes decades and considerable financial resources. Also, such approaches when successful, enable the investigation of a very small fraction of a given materials design space leaving vast possibilities unexplored. No alchemic recipes exist, however desirable, which given a target application and desired properties, enables one to design the optimized material for that application. However, to tread some distance on that alchemic road, recently, extensive work has centered on the accelerated and cost-effective discovery, manufacturing, and deployment of novel and better materials as promoted by the Materials Genome Initiative Holdren et al. (2011).

1.1 Challenges in Accelerated Materials Discovery Techniques

The chief hurdle when it comes to searching for new materials with requisite or better properties is the scarcity of physical knowledge about the class of materials that constitute the design space. Data regarding the structure and resultant properties may be available, but what is lacking is usually the fundamental physics that delineate the processing-structure-property-performance (PSPP) relationships in these materials. Additionally, the interplay of structural, chemical and microstructural degrees of freedom introduces enormous complexity, which exponentially increases the dimensionality of the problem at hand, limiting the application of traditional design strategies.

To bypass this challenge, the focus is on the use of data-to-knowledge approaches, the idea being to implicitly extract the material physics embedded in the data itself with the use of modern day tools - machine learning, design optimization, manufacturing scale-up and automation, multiscale modeling, and uncertainty quantification with verification and validation. Typical techniques include the utilization of High-Throughput (HT) computational Strasser et al. (2003); Curtarolo et al. (2013); Kirklin et al. (2013) and experimental frameworks Strasser et al. (2003); Potyrailo et al. (2011); HPE (????); Green et al. (2017), which are used to generate large databases of materials feature / response sets, which then must be analyzed Curtarolo et al. (2003) to identify the materials with the desired characteristics Solomou et al. (2018). HT methods, however, fail to account for constraints in experimental / computational) resources available, nor do they anticipate the the existence of bottle necks in the scientific workflow that unfortuantely render impossible the parallel execution of specific experimental / computational tasks.

Recently, the concept of optimal experimental design, within the overall framework of Bayesian Optimization (BO), has been put forward as a design strategy to circumvent the limitations of traditional (costly) exploration of the design space. This was pioneered by Balachandran et al. Balachandran et al. (2016) who put forward a framework that balanced the need to exploit available knowledge of the design

space with the need to explore it by using a metric (Expected Improvement (EI)) that selects the best next experiment with the end-goal of accelerating the iterative design process. BO-based approaches rely on the construction of a response surface of the design space and are typically limited to the use of a single model to carry out the queries. This is an important limitation, as often times, at the beginning of a materials discovery problem, there is not sufficient information to elucidate the feature set (i.e., model) that is most related to the specific performance metric to optimize.

Additionally, although these techniques have been successfully demonstrated in a few materials science problems Xue et al. (2016a,b); Dehghannasiri et al. (2017); Frazier and Wang (2016); Seko et al. (2015); Ueno et al. (2016); Seko et al. (2014); Ju et al. (2017); Gopakumar et al. (2018), the published work tends to focus on the optimization of a single objective Balachandran et al. (2016), which is far from the complicated multi-dimensional real world materials design requirements.

In this manuscript, we discuss the materials discovery challenge from the perspective of experiment design - i.e. goal-oriented materials discovery, wherein we efficiently exploit available computational tools, in combination with experiments, to accelerate the development of new materials and materials systems. In the following sections, we discuss the need for exploring the field of Materials discovery via the experiment design paradigm and then specifically discuss two approaches that address the prevalent limitations discussed above: i) a framework that is capable of adaptively selecting competing models connecting materials features to performance metrics through Bayesian Model Averaging (BMA), followed by optimal experimental design, ii) a framework for the fusion of information that exploits correlations among sources/models and between the sources and ‘ground truth’ in conjunction with a multi-information source optimization framework that identifies, given current knowledge, the next best information source to query and where in the input space to query it via a novel value-gradient policy and examples of applications of these approaches in the context of single-objective and multi-objective materials design optimization problems and information fusion in CALPHAD-based thermodynamic modeling.

2 EXPERIMENT DESIGN

The divergence of modern science from its roots in natural philosophy was heralded by the emphasis on experimentation in the sixteenth and seventeenth centuries as a means to establish causal relationships Hacking (1983). Experiments involve the manipulation of one or more independent variables followed by the systematic observation of the effects of the manipulation on one or more dependent variables. An experiment design then, is the formulation of a detailed experimental plan in advance of doing the experiment that when carried out will describe or explain the variation of information under conditions that are hypothesized to reflect the variation. An optimal experiment design maximizes either the amount of “information” that can be obtained for a given amount of experimental effort or the accuracy with which the results are obtained, depending on the purpose of the experiment. A schematic of the experiment design process is shown in Figure 1.

Taking into consideration the large number of measurements often needed in materials research, design problems are formulated as a multi-dimensional optimization problem, which typically require training data in order to be solved. Prior knowledge regarding parameters and features affecting the desired properties of materials is of great importance. However, often prior knowledge is inadequate and the presence of large uncertainties is detrimental to the experimental design. Hence, additional measurements or experiments are necessary in order to improve the predictability of the model with respect to the design objective. Naturally, it is then essential to direct experimental efforts such that the targeted material may be found by

minimizing the number of experiments. This may be achieved via an experimental design strategy that is able to distinguish between different experiments based upon the information they can provide. Thus, the experiment design strategy results in the choosing of the next best experiment, which is determined by optimizing an *acquisition* function.

2.1 Experiment Design under Model Uncertainty

In most materials design tasks, there are always multiple information sources at the disposal of the material scientist. For example, the relationships between the crystal structure and properties/performance can in principle be developed through experiments as well as (computational) models at different levels of fidelity and resolution (atomistic scale, molecular scale, continuum scale). Traditional holistic design approaches such as Integrated Computational Materials Engineering (ICME), on the other hand, often proceed on the limited and unrealistic assumption that there is only one source available to query the design space. For single information sources and sequential querying, there are two traditional techniques for choosing what to query next in this context Lynch (2007). In most materials design tasks, there are always multiple information sources at the disposal of the material scientist. For example, the relationships between the crystal structure and properties/performance can in principle be developed through experiments as well as (computational) models at different levels of fidelity and resolution (atomistic scale, molecular scale, continuum scale). Traditional holistic design approaches such as Integrated Computational Materials Engineering (ICME), on the other hand, often proceed on the limited and unrealistic assumption that there is only one source available to explore the design space.

In the realm of single information sources and sequential querying, two approaches are most commonly used to choose the next query/experiment Scott et al. (2011). These are (i) efficient global optimization (EGO) Jones et al. (1998) and its extensions, such as sequential Kriging optimization (SKO) Huang et al. (2006) and value-based global optimization (VGO) Moore et al. (2014), and (ii) the knowledge gradient (KG) Frazier et al. (2008); Gupta and Miescke (1994, 1996). EGO uses a Gaussian process Rasmussen (2004) representation of available data, but does not account for noise Schonlau et al. (1996, 1998). SKO also uses Gaussian processes, but includes a variable weighting factor to favour decisions with higher uncertainty Scott et al. (2011). KG differs in that while it can also account for noise, it selects the next experiment on the basis of the expected value of the best material after the experiment. In the case of multiple uncertain sources of information (e.g., different models for the same problem), it is imperative to integrate all the sources to produce more reliable results Dasey and Braun (2007). In practice, there are several approaches for fusing information from multiple models. Bayesian Model Averaging (BMA) and fusion under known correlation Geisser (1965); Morris (1977); Winkler (1981); Ghoreishi and Allaire (2018) are two such model fusion techniques that enable robust design. These two approaches shall be discussed in detail in the following sections.

2.1.1 Bayesian Model Averaging (BMA)

The goal of any materials discovery strategy is to identify an action that results in a desired property, which is usually optimizing an objective function of the action over the Materials Design Space (MDS). In materials discovery, each action is equivalent to an input or design parameter setup. If complete knowledge of the objective function exists, then the materials discovery challenge is met. In reality however, this objective function is a blackbox, of which little if anything is known and the cost of querying such a function (through expensive experiments/simulations) at arbitrary query points in the MDS is very high. In these cases a parametric or nonparametric surrogate model can be used to approximate the true objective function. Bayesian Optimization (BO) Shahriari et al. (2016) corresponds to these cases, where the prior

model is sequentially updated after each experiment. Irrespective of whether prior knowledge about the form of the objective function exists and/or many observations of the objective values at different parts of the input space are available, there is an inherent feature selection step, where different potential feature sets might exist. Moreover, there might be a set of possible parametric families as candidates for the surrogate model. Even when employing nonparametric surrogate models, several choices for the functional form might be available. These translate into different possible surrogate models for the objective function. The common approach is to select a feature set and a single family of models and fix this selection throughout the experiment design loop; however, this is not a reliable approach due to the small initial sample size that is ubiquitous in materials science.

This problem was addressed by framing experiment design as Bayesian Optimization under Model Uncertainty (BOMU), and incorporating Bayesian Model Averaging (BMA) within Bayesian Optimization. The *acquisition* function used is the Expected Improvement (EI) which seeks to balance the need to exploit available knowledge of the design space with the need to explore it. Suppose that f' is the minimal value of the objective function f observed so far. Expected improvement evaluates f at the point that, in expectation, improves upon f' the most. This corresponds to the following utility function:

$$I = 0, \max(f' - f(x)) \quad (1)$$

If \hat{y} and s are the predicted value and its standard error at x respectively, then the expected improvement is given by:

$$E[I(x)] = (f' - \hat{y})\Phi\frac{(f' - \hat{y})}{s} + s\phi\frac{(f' - \hat{y})}{s} \quad (2)$$

where: $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions Jones et al. (1998).

The Bayesian Optimization under Model Uncertainty approach may then be described in words as follows:

- There is a collection of potential models (e.g. models based on different features sets)
- The models are averaged, based on the (posterior) model probabilities based on initial data set to form a BMA.
- Using the expected acquisition function under the BMA, an experiment is chosen that maximizes the expected acquisition.
- The experiment is run, each model is updated and the (posterior) model probabilities are updated.
- The expected acquisition under the updated BMA is computed and an experiment is chosen.
- This iteration is done until some stopping criteria (e.g. while objective not satisfied and budget not exhausted), and the best observation so far is selected as the final suggestion.

Incorporating BMA within Bayesian Optimization produces a system capable of autonomously and adaptively learning not only the most promising regions in the materials space but also the models that most efficiently guide such exploration. The framework is also capable of defining optimal experimental sequences in cases where multiple objectives must be met—we note that recent works have begun to address the issue of multi-objective Bayesian Optimization [38] in the context of materials discovery. Our approach, however, is different in that the multi-objective optimization is carried out simultaneously with feature selection. The overall framework for autonomous materials discovery is shown in Figure 2.

2.1.2 Error correlation-based model fusion (CMF) approach

As mentioned earlier, model-based ICME-style frameworks tend to focus on integrating tools at multiple scales under the assumption that there is a single model/tool which is significant at a specific scale of the problem. This ignores the use of multiple models that may be more/less effective in different regions of the performance space. Data-centric approaches, on the other hand, tend to focus (with some exceptions) on the brute-force exploration of the MDS, not taking into account the considerable cost associated with such exploration. In Ghoreishi et al. (2018), the authors presented a framework that addresses the two outstanding issues listed above in the context of the optimal microstructural design of advanced high strength steels. Specifically, they carried out the fusion of multiple information sources that connect microstructural descriptors to mechanical performance metrics. This fusion is done in a way that accounts for and exploits the correlations between each individual information source—reduced order model constructed under different simplifying assumptions regarding the partitioning of (total) strain, stress or deformation work among the phases constituting the microstructure—and between each information source and the ground truth—represented in this case by a full-field microstructure-based finite element model. This finite element model is computational, and is considered as a higher fidelity model as part of a multifidelity framework, the intention being to create a framework for predicting ground truth. Specifically, the purpose of the work was not to match the highest fidelity model, but to predict material properties when created at ground truth. There is usually no common resource tradeoff in this scenario, in contrast to traditional computational multifidelity frameworks that trade computational expense and accuracy.

In the framework, the impact a new query to an information source has on the fused model is of value. The search is performed over the input domain and the information source options concurrently to determine which next query will lead to the most improvement in the objective function. In addition, the exploitation of correlations between the discrepancies of the information sources in the fusion process is novel and enables the identification of ground truth optimal points that are not shared by any individual information sources in the analysis.

A fundamental hypothesis of this approach is that any model can provide potentially useful information to a given task. This technique thus takes into account all potential information any given model may provide and fuses unique information from the available models. The fusion goal then is to identify dependencies, via estimated correlations, among the model discrepancies. With these estimated correlations, the models are fused following standard practice for the fusion of normally distributed data. To estimate the correlations between the model deviations when they are unknown, the reification process Allaire and Willcox (2012); Thomison and Allaire (2017) is used, which refers to the process of treating each model, in turn, as ground truth. The underlying assumption here is that the data generated by the reified model represents the true quantity of interest. These data are used to estimate the correlation between the errors of the different models and the process is then repeated for each model. The detailed process of estimating the correlation between the errors of two models can be found in Refs. Allaire and Willcox (2012); Thomison and Allaire (2017). A flowchart of the approach is shown in figure 3. The next step then is to determine which information source should be queried and where to query it, concurrently, so as to produce the most value with the tacit resource constraint in mind. For this decision, a utility, which is referred to as the value-gradient utility is used, which accounts for both the immediate improvement in one step and expected improvement in two steps. The goal here is to produce rapid improvement, with the knowledge that every resource expenditure could be the last, but at the same time, to be optimally positioned for the next resource expenditure. In this sense, there is equal weight accorded to next step value with next step (knowledge) gradient information, hence the term value-gradient. As mentioned in the previous section, in

the BMA approach, the Expected Improvement (EI) metric is used to choose the next query point, while in this approach, the value gradient is used.

The knowledge gradient, which is a measure of expected improvement, is defined as:

$$\nu^{KG}(x) = E[V^{N+1}H^{N+1}(x)) - V^NH^N|H^N] \quad (3)$$

The KG policy for sequentially choosing the next query is then given as:

$$x^{KG} = \operatorname{argmax}_{x \in \chi} \nu^{KG}(x) \quad (4)$$

and the value-gradient utility is given by:

$$U = \mu_{fused}^* + \max_{x \in \chi} \nu^{KG}(x) \quad (5)$$

where μ_{fused}^* is the maximum value of the mean function of the current fused model and $\max_{x \in \chi} \nu^{KG}(x)$ is the maximum expected improvement that can be obtained with another query as measured by the knowledge gradient over the fused model.

2.2 Application of Experiment Design Framework: Examples

2.2.1 Multi-Objective Bayesian Optimization

A Multi-objective Optimal Experimental Design (OED) framework (see Figure 4) based on the Bayesian optimization techniques was reported by the authors in Solomou et al. (2018). The specific objective of this Bayesian Optimal Experimental Design (BOED) framework in this study was to provide an Optimal Experiment Selection (OES) policy to guide an efficient search of precipitation strengthened NiTi Shape memory alloys (SMAs) with selected target properties by efficiently solving a multi-objective optimization problem. The EHVI Emmerich et al. (2011) acquisition function was used to perform multi-objective optimization. EHVI balances the trade-off between exploration and exploitation for multi-objective BOED problems, similar to EI for single-objective problems. EHVI is a scalar quantity that allows a rational agent to select, sequentially, the next best experiment to carry out, given current knowledge, regardless of the number of objectives, or dimensionality, of the materials. The optimal solutions in an optimization problem are typically referred as Pareto optimal solutions or Pareto front or Pareto front points. The Pareto optimal solutions in a selected multi-objective space, correspond to the points of the objective space that are not dominated by any other points in the same space.

For the NiTi SMA, the considered input variables were the initial homogeneous Ni concentration of the material before precipitation (c) and the volume fraction (v_f) of the precipitates while the objective functions were functions of the material properties of the corresponding homogenized SMA. The framework was used to discover precipitated SMAs with (objective 1) an austenitic finish temperature $A_f = 30^\circ C$, (objective 2) a specific thermal hysteresis that is defined as the difference of austenitic finish temperature and martensitic start temperature, $A_f - M_s = 40^\circ C$. The problem was solved for two case studies, where the selected continuous MDS is discretized with a coarse and a dense mesh respectively. The refined MDS has $n_T = 21021$ combinations of the considered variables c and v_f . The utility of the queried materials by the BOED framework within a predefined experimental budget is compared with the utility of the corresponding queried materials following a Pure Random Experiment Selection (PRES) policy

and a Pure Exploitation Experiment Selection (PEES) policy within a predefined experimental budget. An experimental budget of is assumed of $n_B = 20$ material queries and for the case of the OES and PEES policies the experimental budget is allocated to $n_I = 1$ randomly queried material and to $n_E = 19$.

The results are indicated in Figure 5a. It is seen that the OES policy, even under the limited experimental budget, queries materials that belong to the region of the objective space which approaches the true Pareto front. This is clear by comparing the Pareto front calculated based on the results of the OES (blue dash line) with the true Pareto front found during the case study 1 (red dot line). The results also show that the materials queried by the PRES policy are randomly dispersed in the objective space as expected while the materials queried by the PEES policy are clustered in a specific region of the objective space which consists of materials with similar volume fraction values which is anticipated courtesy the true exploitative nature of the policy. On further analysis is is seen that the OES on average queries materials with better utility in comparison to the other two policies, while the PRES policy exhibits the worst performance.

Same trends of performance are maintained through the equivalent comparisons conducted for various experimental budgets as shown in Figure 5b which also indicates similar curves for the coarse mesh. It is apparent that if the OES policy is employed to query a material in a discrete MDS with defined variables bounds, its relative performance in comparison to the PRES policy becomes more definitive as the discretization of the MDS is further refined, as the gap between PRES and OES policies for the case of the dense discretized MDS (red lines) is much bigger than that in the case of the coarse discretized MDS (blue lines) optimally queried materials. The results of the BOED framework thus demonstrate that the method could efficiently approach the true Pareto front of the objective space of the approached materials discovery problems successfully. Such treatment was also carried out for a three objective problem with the additional objective of maximizing the maximum saturation strain (H_{sat}) that the material can exhibit and similar conclusions were drawn.

2.2.2 Bayesian Model Averaging: Search for MAX phase with maximum bulk modulus

The BMA framework was demonstrated by efficiently exploring the MAX ternary carbide/nitride space through Density Functional Theory (DFT) calculations by the authors in Talapatra et al. (2018). The problem was formulated with the goal of identifying the material/materials with the maximum bulk modulus K with a resource constraint of permitting experiments totally querying 20% of the MDS. The case of the maximum bulk modulus K is designed as a single-objective optimization problem while the second problem is designed as a multi-objective problem. Features describing the relation between the material and objective functions were obtained from literature and domain knowledge. Feature correlations were used to finalize six different feature sets which are denoted as F_1, F_2, F_3, F_4, F_5 and F_6 . Complete details may be found in Talapatra et al. (2018). Some representative results are shown here. Calculations were carried out for different number of initial data instances $N = 2, 5, 10, 15, 20$. The performance trends for all problems across different values of N are consistent. The technique is found to not significantly depend on quantity of initial data. Here, representative results for $N = 5$ are shown.

Figure 5c indicates the maximum bulk modulus found in the experiment design iterations based on each model (feature set) averaged over all initial data set instances for $N = 5$. The dotted line in the figure indicates the maximum bulk modulus = 300 GPa that can be found in the MDS. F_2 is found to be the best performing feature set on average, converging fastest to the maximum bulk modulus. F_6 and F_5 on the other hand, are uniformly the worst performing feature sets on average, converging the slowest. It is evident that using a regular optimization approach will work so long there is, apriori, a good feature set. Unfortunately, due to small sample size and large number of potential predictive models, the feature selection step may

not result in the true best predictive model for efficient Bayesian Optimization. Small sample sizes pose a great challenge in model selection due to inherent risk of imprecision and, and no feature selection method performs well in all scenarios when sample sizes are small. Thus, by selecting a single model as the predictive model based on small observed sample data, one ignores the model uncertainty.

To circumvent this problem the Bayesian Model Averaging (BMA) method was used. Regression models based on aforementioned six feature subsets, were adopted in the BMA experiment design. The BMA coefficients were evaluated in two ways: first-order (BMA_1) and second-order (BMA_2) Laplace approximation. Figure 5e shows the comparison of the average performance of both the first-order and second-order BMA over all initial data set instances with the best performing model (F_2) and worst performing model (F_6). It can be seen that both the first-order and second-order BMA performance in identifying the maximum bulk modulus is consistently close to the best model (F_2). BMA_1 performs as well as if not better than F_2 . Figure 5f shows the corresponding swarm plots indicating the number of calculations required to discover the maximum bulk modulus in the MDS for $N = 5$ using BMA_1 and BMA_2 . It can be seen that for a very high percentage of cases the maximum bulk modulus can be found within the designated budget. In Figures 5g and 5h, the average model coefficients (posterior model probabilities) of the models based on different feature sets over all instances of initial data set are shown with the increasing number of calculations for BMA_1 and BMA_2 respectively. Thus we see that, while prior knowledge about the fundamental features linking the material to the desired material property is certainly essential to build the Materials Design Space (MDS), the BMA approach may be used to auto-select the best features/feature sets in the MDS, thereby eliminating the requirement of knowing the best feature set a priori. Also, this framework is not significantly dependent on the size of the initial data, which enables its use in materials discovery problems where initial data is scant.

2.2.3 Multi-Source Information Fusion: Application to Dual-Phase materials

2.2.4 Bayesian Model Averaging and Information Fusion: CALPHAD-based thermodynamic modeling

Calculation of phase diagrams (CALPHAD) is one of the fundamental tools in alloy design and an important component of ICME. Uncertainty quantification of phase diagrams is the first step required to provide confidence for decision making in property- or performance-based design. In work that was the first of its kind Honarmandi et al. (2019), the authors independently generated four CALPHAD models describing Gibbs free energies for the $Hf - Si$ system. The MCMC Metropolis Hastings toolbox in Matlab was then utilized for probabilistic calibration of the parameters in the applied CALPHAD models. These results are shown for each model in Figure 6 where it is seen that there is a very good agreement between the results obtained from model 2 and the data with a very small uncertainty band and consequently a small Model Structure Uncertainty (MSU) Choi et al. (2008). Models 3 and 4 on the other hand show large uncertainties for the phase diagrams which are mostly attributed to MPU. In the context of BMA, the weight of the applied models was calculated to be 0.1352, 0.5938, 0.1331, and 0.1379, respectively, indicating that Model 2 thus has three times the weight of the other models, which otherwise have similar Bayesian importance, which is consistent with the phase diagram results in Figure 6. The phase diagram obtained using BMA is shown in 7. The posterior modes of the probability distributions in the BMA model exactly correspond to the posterior modes of the probability distributions in model 2. Thus, the best model results can be considered as the optimum results for the average model, but with broader uncertainties, contributed by the inferior models. In BMA, each model has some probability of being true and the fused estimate is a weighted average of the models. This method is extremely useful in the case of model-building process based on a weighted average over the models' responses, and/or less risk (more confidence) in design

based on broader uncertainty bands provided by a weighted average over the uncertainties of the models' responses. Error Correlation-based Model fusion was then used to fuse the models together in two ways: i) Fuse models 1, 3 , 4 to examine whether the resulting fused model maybe closer to the data and reduce the uncertainties and ii) Fuse models 1,2,3 and 4 together. Figure 8a shows that the approach can provide a phase diagram in much better agreement with data and less uncertainties compared to phase diagrams obtained from each one of the applied models individually. This result implies that random CALPHAD models can be fused together to find a reasonable estimation for phase diagram instead of trial-and-error to find the best predicting model. It is also apparent that better predictions can be achieved as shown in 8b if model 2 (the best model) is also involved in the model fusion. The information fusion technique allowed the acquisition of more precise estimations and lower uncertainties compared to results obtained from each individual model. In summary, the average model obtained from BMA shows larger 95% confidence intervals compared to any one of the individual models, which can provide more confidence for robust design but is likely too conservative. On the other hand, the error correlation-based technique can provide closer results to data with less uncertainties than the individual models used for the fusion. The uncertainty reductions through this fusion approach are also verified through the comparison of the average entropies (as a measure of uncertainty) obtained for the individual and fused models. Therefore, random CALPHAD models can be fused together to find reasonable predictions for phase diagrams with no need to go through the cumbersome task of identifying the best CALPHAD models.

3 CONCLUSIONS AND RECOMMENDATIONS

In this work, we have reviewed some of the most important challenges and opportunities related to the concept of optimal experiment as an integral component for the development of Materials Discovery frameworks. As our understanding of the vagaries implicit in different design problems progresses, tailoring experiment design strategies around the specific material classes under study while further developing the experiment design frameworks will become increasingly feasible and successful. As techniques improve, we will be able to access and explore increasingly complex materials design spaces, opening the door to precision tailoring of materials to desired applications. Challenges in the form of the availability and generation of sufficient and relevant data of high quality need to be continuously addressed. The optimal way to accomplish this would be the implementation of universal standards, centralized databases and the development of an open access data-sharing system in conjunction with academia, industry, government research institutions and journal publishing agencies which is already underway.

4 CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

5 AUTHOR CONTRIBUTIONS

All authors have contributed equally to this work.

6 FUNDING

The authors acknowledge the support of NSF through the projects NSF-CMMI-1534534, NSF-CMMI-1663130, and DGE- 1545403. RA and ED also acknowledge the support of NSF through Grant No. NSF-DGE-1545403. XQ acknowledges the support of NSF through the project CAREER: Knowledge-driven Analytics, Model Uncertainty, and Experiment Design, NSF-CCF-1553281 and NSF-CISE-1835690 (with RA). AT and RA also acknowledge support by the Air Force Office of Scientific Research under AFOSR-FA9550-78816-1- 0180 (Program Manager: Dr. Ali Sayir). RA and DA also acknowledge the support of ARL through [grant No. W911NF-132-0018].

7 ACKNOWLEDGMENTS

Calculations were carried out in the Texas A&M Supercomputing Facility.

REFERENCES

- (????)
- Allaire, D. and Willcox, K. (2012). Fusing information from multifidelity computer models of physical systems. In *Information Fusion (FUSION), 2012 15th International Conference on* (IEEE), 2458–2465
- Balachandran, P. V., Xue, D., Theiler, J., Hogden, J., and Lookman, T. (2016). Adaptive strategies for materials design using uncertainties. *Scientific reports* 6, 19660
- Choi, H.-J., McDowell, D. L., Allen, J. K., and Mistree, F. (2008). An inductive design exploration method for hierarchical systems design under uncertainty. *Engineering Optimization* 40, 287–307
- Curtarolo, S., Hart, G. L., Nardelli, M. B., Mingo, N., Sanvito, S., and Levy, O. (2013). The high-throughput highway to computational materials design. *Nature materials* 12, 191
- Curtarolo, S., Morgan, D., Persson, K., Rodgers, J., and Ceder, G. (2003). Predicting crystal structures with data mining of quantum calculations. *Physical review letters* 91, 135503
- Dasey, T. J. and Braun, J. J. (2007). Information fusion and response guidance. *Lincoln Laboratory Journal* 17
- Dehghannasiri, R., Xue, D., Balachandran, P. V., Yousefi, M. R., Dalton, L. A., Lookman, T., et al. (2017). Optimal experimental design for materials discovery. *Computational Materials Science* 129, 311–322
- Emmerich, M. T., Deutz, A. H., and Klinkenberg, J. W. (2011). Hypervolume-based expected improvement: Monotonicity properties and exact computation. In *Evolutionary Computation (CEC), 2011 IEEE Congress on* (IEEE), 2147–2154
- Frazier, P. I., Powell, W. B., and Dayanik, S. (2008). A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization* 47, 2410–2439
- Frazier, P. I. and Wang, J. (2016). Bayesian optimization for materials design. In *Information Science for Materials Discovery and Design* (Springer). 45–75
- Geisser, S. (1965). A bayes approach for combining correlated estimates. *Journal of the American Statistical Association* 60, 602–607
- Ghoreishi, S. F. and Allaire, D. L. (2018). A fusion-based multi-information source optimization approach using knowledge gradient policies. In *2018 AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*. 1159
- Ghoreishi, S. F., Molkeri, A., Srivastava, A., Arroyave, R., and Allaire, D. (2018). Multi-information source fusion and optimization to realize icme: Application to dual-phase materials. *Journal of Mechanical Design* 140, 111409

- Gopakumar, A. M., Balachandran, P. V., Xue, D., Gubernatis, J. E., and Lookman, T. (2018). Multi-objective optimization for materials discovery via adaptive design. *Scientific reports* 8, 3738
- Green, M. L., Choi, C., Hattrick-Simpers, J., Joshi, A., Takeuchi, I., Barron, S., et al. (2017). Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Applied Physics Reviews* 4, 011105
- Gupta, S. S. and Miescke, K. J. (1994). Bayesian look ahead one stage sampling allocations for selecting the largest normal mean. *Statistical Papers* 35, 169–177
- Gupta, S. S. and Miescke, K. J. (1996). Bayesian look ahead one-stage sampling allocations for selection of the best population. *Journal of statistical planning and inference* 54, 229
- Hacking, I. (1983). *Representing and intervening*, vol. 279 (Cambridge University Press Cambridge)
- Holdren, J. P. et al. (2011). Materials genome initiative for global competitiveness. *National Science and technology council OSTP. Washington, USA*
- Honarmandi, P., Duong, T. C., Ghoreishi, S. F., Allaire, D., and Arroyave, R. (2019). Bayesian uncertainty quantification and information fusion in calphad-based thermodynamic modeling. *Acta Materialia* 164, 636–647
- Huang, D., Allen, T. T., Notz, W. I., and Miller, R. A. (2006). Sequential kriging optimization using multiple-fidelity evaluations. *Structural and Multidisciplinary Optimization* 32, 369–382
- Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global optimization* 13, 455–492
- Ju, S., Shiga, T., Feng, L., Hou, Z., Tsuda, K., and Shiomi, J. (2017). Designing nanostructures for phonon transport via bayesian optimization. *Physical Review X* 7, 021024
- Kirklin, S., Meredig, B., and Wolverton, C. (2013). High-throughput computational screening of new li-ion battery anode materials. *Advanced Energy Materials* 3, 252–262
- Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists* (Springer Science & Business Media)
- Moore, R. A., Romero, D. A., and Paredis, C. J. (2014). Value-based global optimization. *Journal of Mechanical Design* 136, 041003
- Morris, P. A. (1977). Combining expert judgments: A bayesian approach. *Management Science* 23, 679–693
- Potyrailo, R., Rajan, K., Stoewe, K., Takeuchi, I., Chisholm, B., and Lam, H. (2011). Combinatorial and high-throughput screening of materials libraries: review of state of the art. *ACS combinatorial science* 13, 579–633
- Rasmussen, C. E. (2004). Gaussian processes in machine learning. In *Advanced lectures on machine learning* (Springer). 63–71
- Schonlau, M., Welch, W. J., and Jones, D. (1996). Global optimization with nonparametric function fitting. *Proceedings of the ASA, section on physical and engineering sciences* , 183–186
- Schonlau, M., Welch, W. J., and Jones, D. R. (1998). Global versus local search in constrained optimization of computer models. *Lecture Notes-Monograph Series* , 11–25
- Scott, W., Frazier, P., and Powell, W. (2011). The correlated knowledge gradient for simulation optimization of continuous parameters using gaussian process regression. *SIAM Journal on Optimization* 21, 996–1026
- Seko, A., Maekawa, T., Tsuda, K., and Tanaka, I. (2014). Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single-and binary-component solids. *Physical Review B* 89, 054303

- Seko, A., Togo, A., Hayashi, H., Tsuda, K., Chaput, L., and Tanaka, I. (2015). Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization. *Physical review letters* 115, 205901
- Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. (2016). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE* 104, 148–175
- Solomou, A., Zhao, G., Boluki, S., Joy, J. K., Qian, X., Karaman, I., et al. (2018). Multi-objective bayesian materials discovery: Application on the discovery of precipitation strengthened niti shape memory alloys through micromechanical modeling. *Materials & Design* 160, 810–827
- Strasser, P., Fan, Q., Devenney, M., Weinberg, W. H., Liu, P., and Nørskov, J. K. (2003). High throughput experimental and theoretical predictive screening of materials- a comparative study of search strategies for new fuel cell anode catalysts. *The Journal of Physical Chemistry B* 107, 11013–11021
- Talapatra, A., Boluki, S., Duong, T., Qian, X., Dougherty, E., and Arróyave, R. (2018). Autonomous efficient experiment design for materials discovery with bayesian model averaging. *Physical Review Materials* 2, 113803
- Thomison, W. D. and Allaire, D. L. (2017). A model reification approach to fusing information from multifidelity information sources. In *19th AIAA Non-Deterministic Approaches Conference*. 1949
- Tylecote, R. F. and Tylecote, R. (1992). *A history of metallurgy* (Institute of materials London)
- Ueno, T., Rhone, T. D., Hou, Z., Mizoguchi, T., and Tsuda, K. (2016). Combo: An efficient bayesian optimization library for materials science. *Materials discovery* 4, 18–21
- Winkler, R. L. (1981). Combining probability distributions from dependent information sources. *Management Science* 27, 479–488
- Xue, D., Balachandran, P. V., Hogden, J., Theiler, J., Xue, D., and Lookman, T. (2016a). Accelerated search for materials with targeted properties by adaptive design. *Nature communications* 7, 11241
- Xue, D., Balachandran, P. V., Yuan, R., Hu, T., Qian, X., Dougherty, E. R., et al. (2016b). Accelerated search for batio3-based piezoelectrics with vertical morphotropic phase boundary using bayesian learning. *Proceedings of the National Academy of Sciences* 113, 13301–13306

8 FIGURES

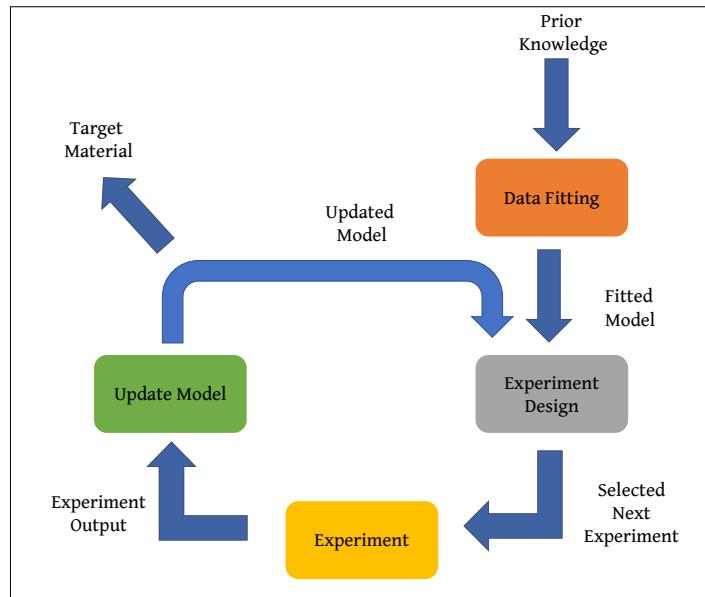


Figure 1. A schematic of the recursive experiment design process

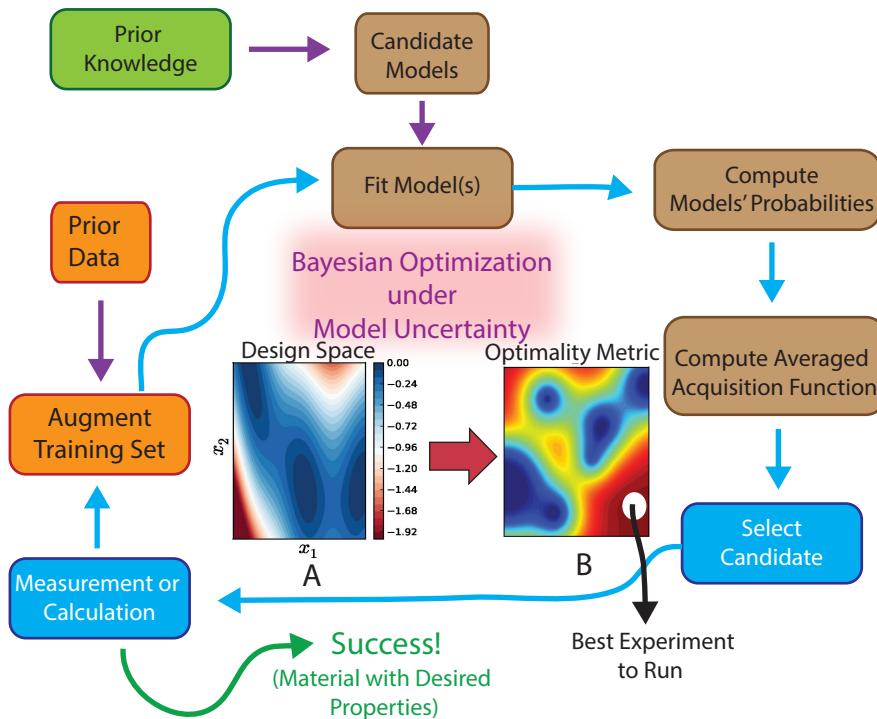


Figure 2. Schematic of the proposed framework for an autonomous, efficient materials discovery system as a realization of Bayesian Optimization under Model Uncertainty (BOMU). Initial data and a set of candidate models are used to construct a stochastic representation of an experiment/simulation. Each model is evaluated in a Bayesian sense and its probability is determined. Using the model probabilities, an effective acquisition function is computed, which is then used to select the next point in the materials design space that needs to be queried. The process is continued iteratively until target is reached or budget is exhausted. Used with permission from Talapatra et al. (2018).

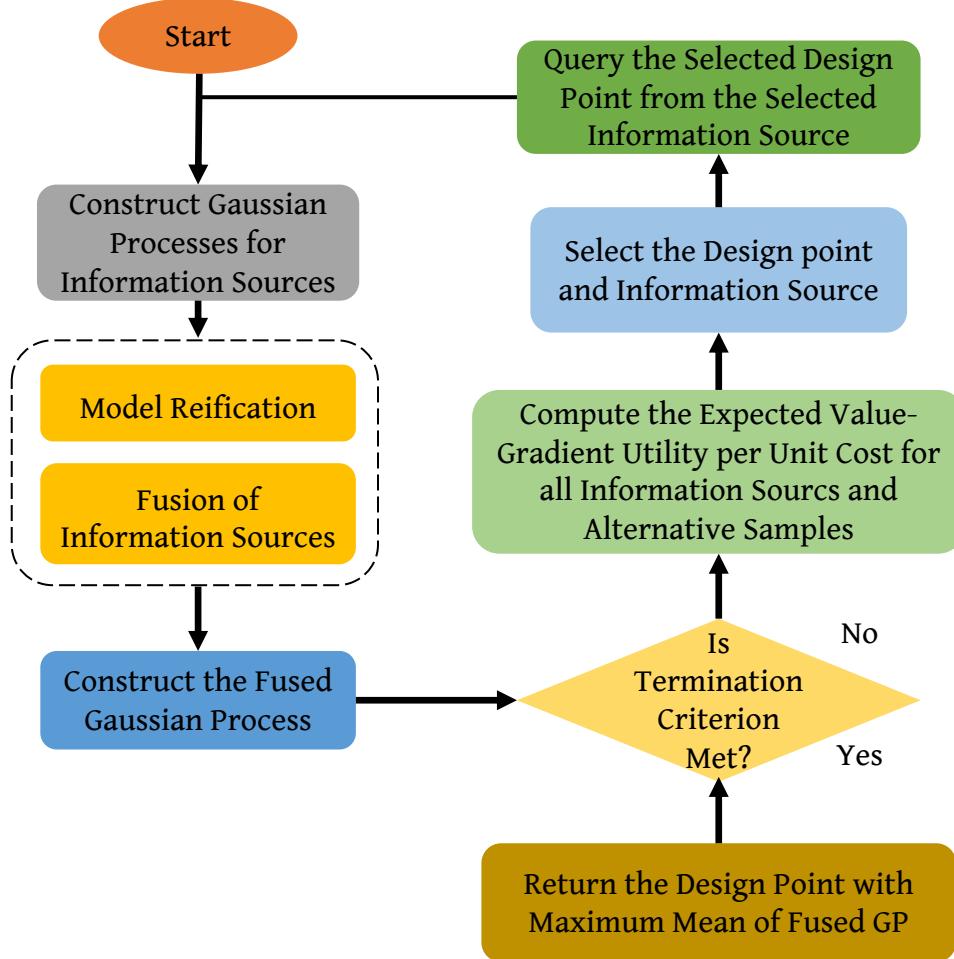


Figure 3. Flowchart of the information fusion approach. Adapted from Ghoreishi et al. (2018)

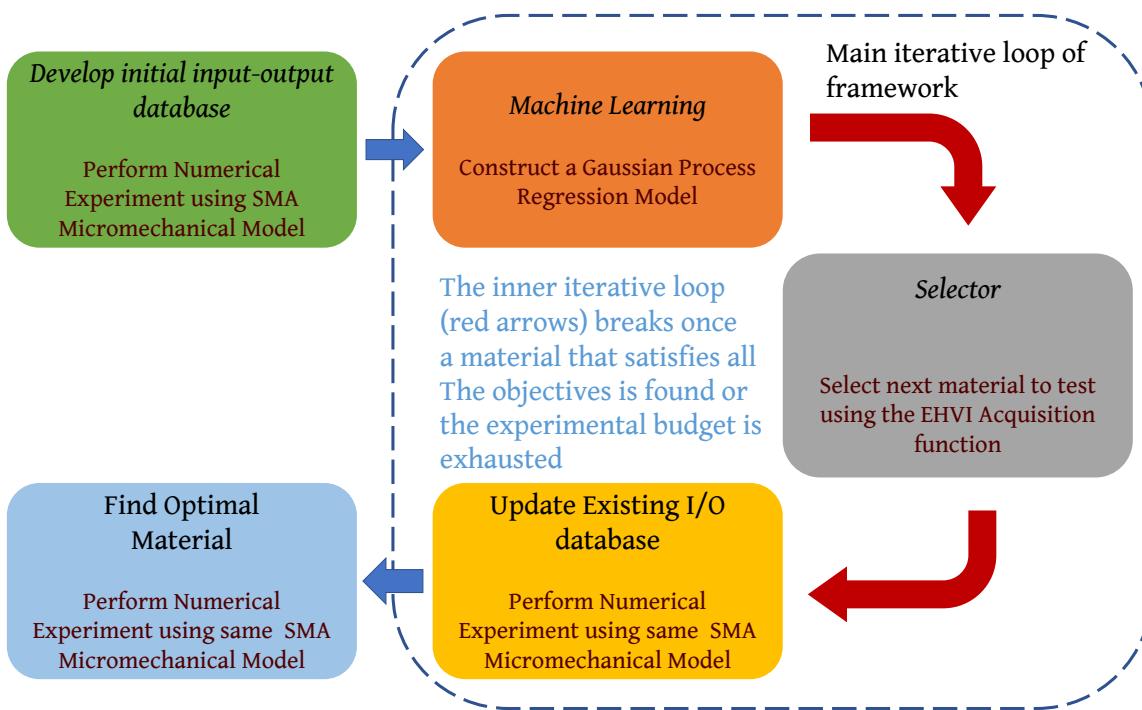


Figure 4. Autonomous closed-loop, multi-objective Bayesian Optimal Experimental Design framework. Adapted from Solomou et al. (2018)

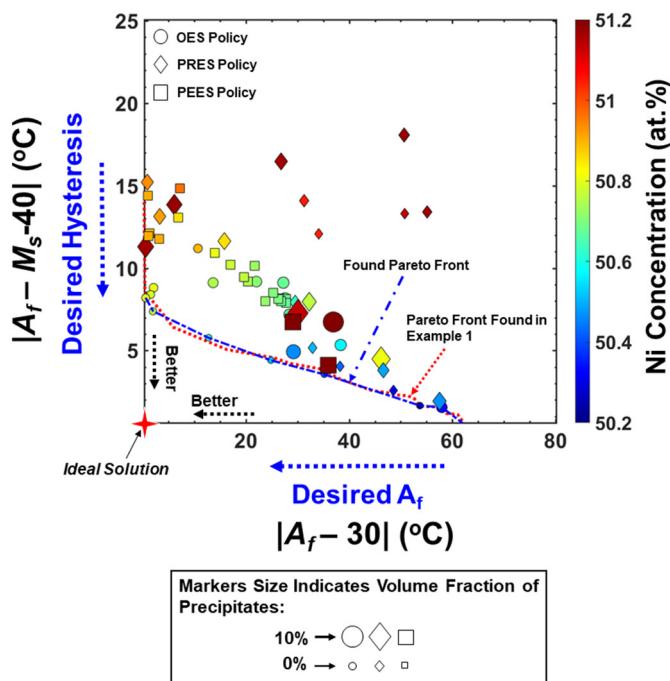


Figure 5a. Calculated objective space and Pareto front using the OES, PRES and PEES policies under the $n_B = 20$ experimental budget. Reproduced with permission from Solomou et al. (2018)

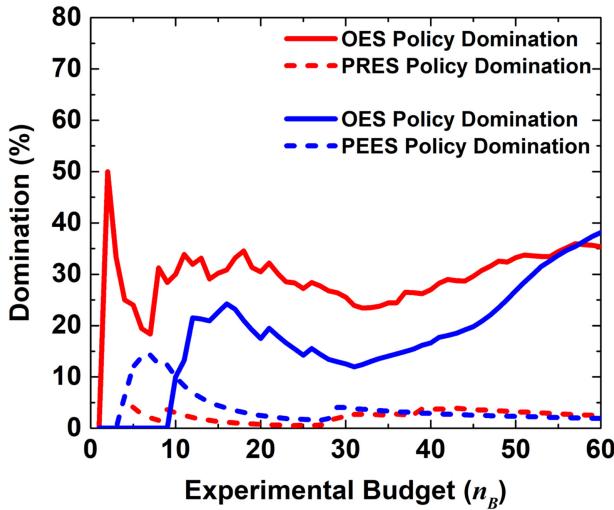


Figure 5b. Comparison of the utility of the queried materials by the OES, PRES and PEES policies as function of the experimental budget for the 2-objectives materials discovery problem. Reproduced with permission from Solomou et al. (2018)

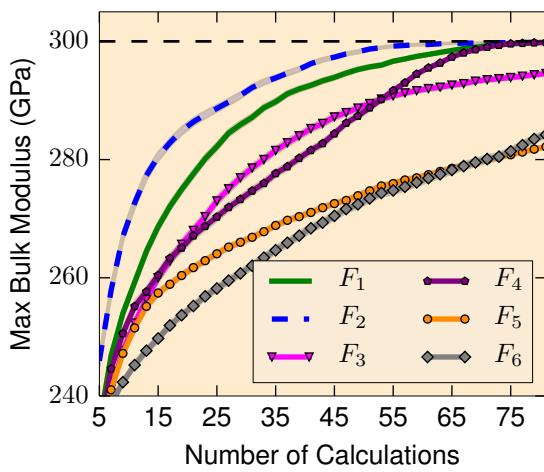


Figure 5c. Representative results for single objective optimization – maximization of bulk modulus for N=5: a) Average maximum bulk modulus discovered using all described feature sets,

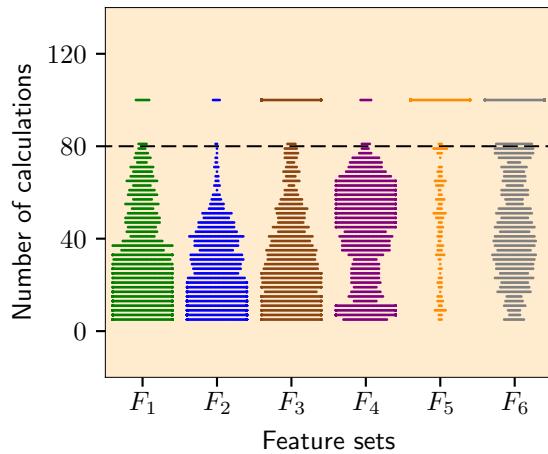


Figure 5d. Representative results for single objective optimization – maximization of bulk modulus for N=5: swarm plots indicating the distribution of the number of calculations required for convergence using all described feature sets

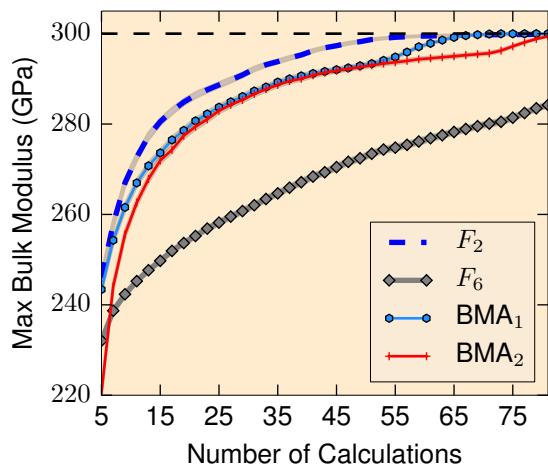


Figure 5e. Representative results for single objective optimization – maximization of bulk modulus for N=5: average maximum bulk modulus discovered using the best feature set F_2 , worst feature set F_6 , BMA₁ and BMA₂

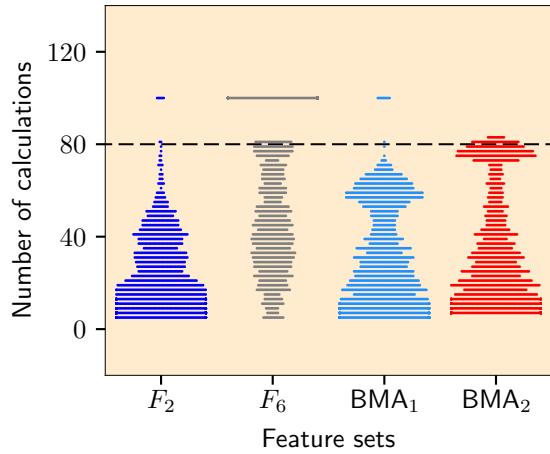


Figure 5f. Representative results for single objective optimization – maximization of bulk modulus for N=5: swarm plots indicating the distribution of the number of calculations required for convergence using the best feature set F_2 , worst feature set F_6 , BMA₁ and BMA₂

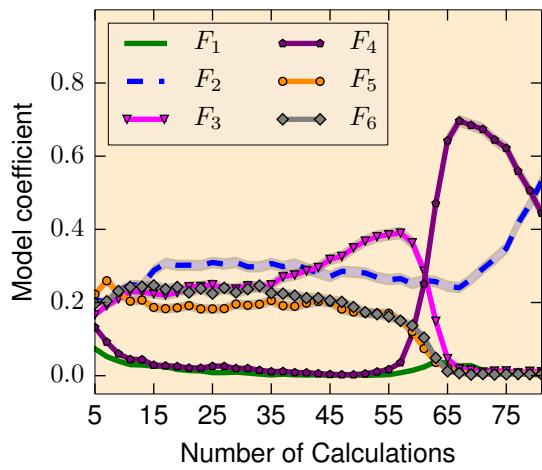


Figure 5g. Representative results for single objective optimization –maximization of bulk modulus for N=5: Average model probabilities for maximizing bulk modulus using BMA₁

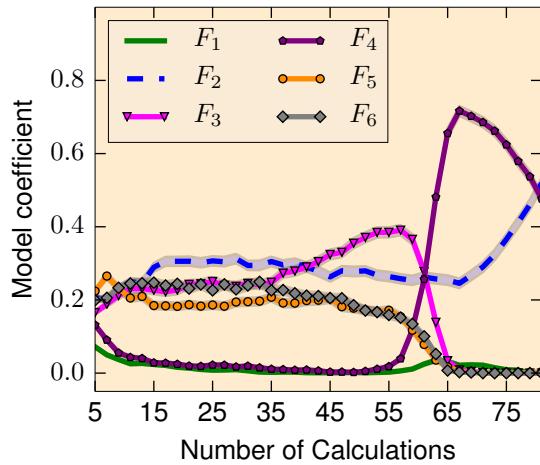


Figure 5h. Representative results for single objective optimization – maximization of bulk modulus for N=5: Average model probabilities for maximizing bulk modulus using BMA₂

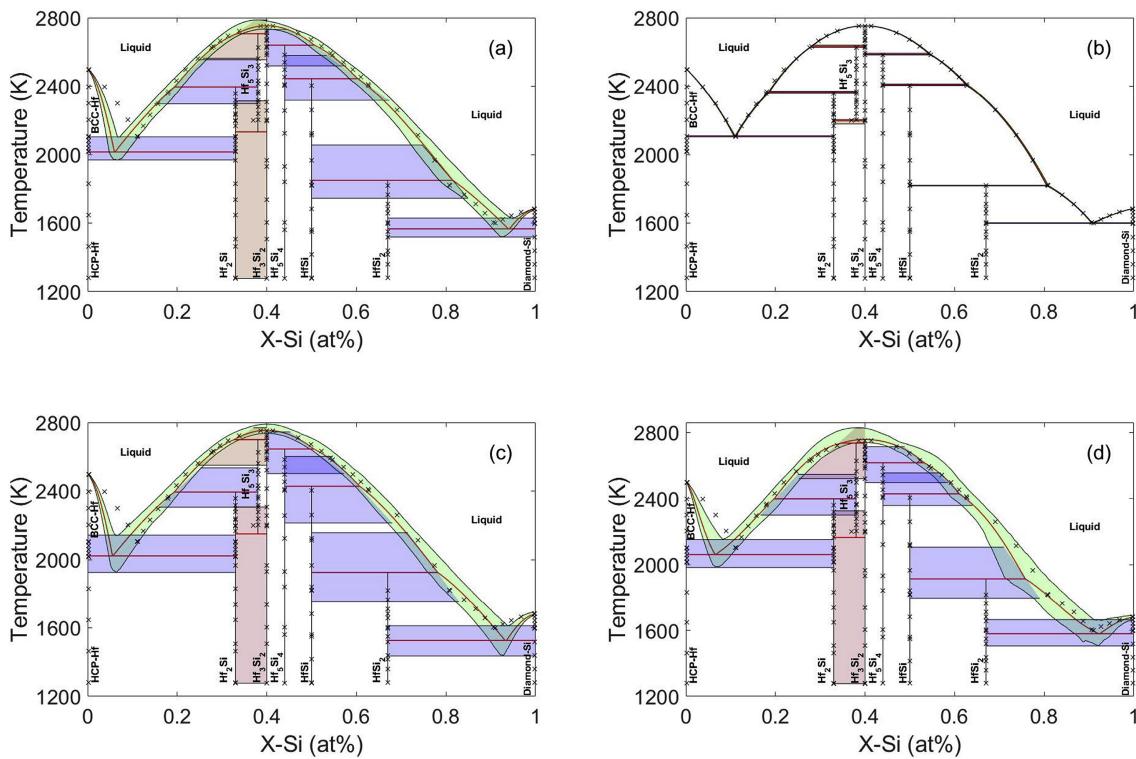


Figure 6. Optimum HfeSi phase diagrams and their 95% BCIs obtained from models 1-4 (a-d) after uncertainty propagation of the MCMC calibrated parameters in each case. Reproduced with permission from Honarmandi et al. (2019)

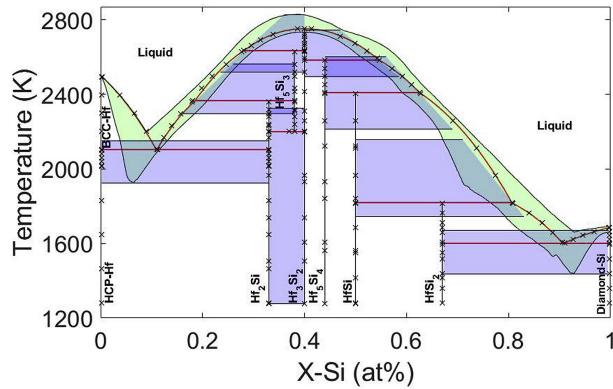


Figure 7. Posterior modes and 95% BCIs at different compositions/regions in Hf-Si phase diagram obtained after BMA. Reproduced with permission from Honarmandi et al. (2019)

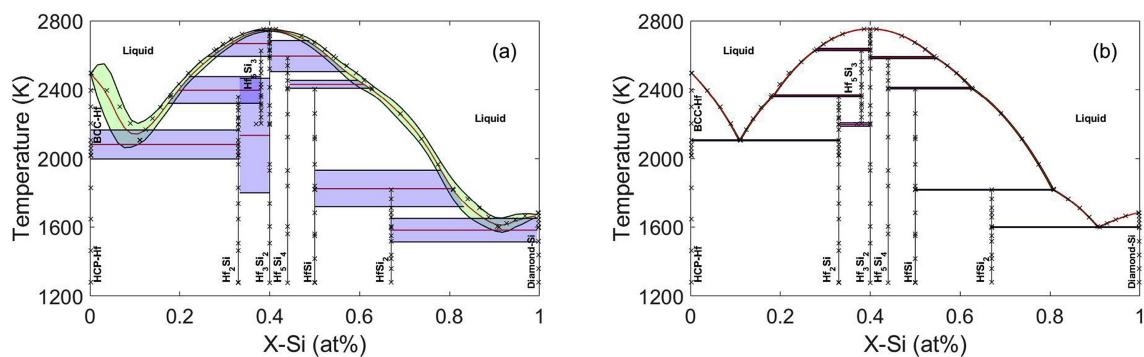


Figure 8. Error correlation-based model fusions of a) three models (1, 3, and 4) and b) all four models. Reproduced with permission from Honarmandi et al. (2019)