SCIENTIFIC REPORTS
natureresearch

OPEN

Bayesian Optimization for Materials Design with Mixed Quantitative and Qualitative Variables

Yichi Zhang[1], Daniel W. Apley[2] & Wei Chen[1]*

Although Bayesian Optimization (BO) has been employed for accelerating materials design in computational materials engineering, existing works are restricted to problems with quantitative variables. However, real designs of materials systems involve both qualitative and quantitative design variables representing material compositions, microstructure morphology, and processing conditions. For mixed-variable problems, existing Bayesian Optimization (BO) approaches represent qualitative factors by dummy variables first and then fit a standard Gaussian process (GP) model with numerical variables as the surrogate model. This approach is restrictive theoretically and fails to capture complex correlations between qualitative levels. We present in this paper the integration of a novel latent-variable (LV) approach for mixed-variable GP modeling with the BO framework for materials design. LVGP is a fundamentally different approach that maps qualitative design variables to underlying numerical LV in GP, which has strong physical justification. It provides flexible parameterization and representation of qualitative factors and shows superior modeling accuracy compared to the existing methods. We demonstrate our approach through testing with numerical examples and materials design examples. The chosen materials design examples represent two different scenarios, one on concurrent materials selection and microstructure optimization for optimizing the light absorption of a quasi-random solar cell, and another on combinatorial search of material constitutes for optimal Hybrid Organic-Inorganic Perovskite (HOIP) design. It is found that in all test examples the mapped LVs provide intuitive visualization and substantial insight into the nature and effects of the qualitative factors. Though materials designs are used as examples, the method presented is generic and can be utilized for other mixed variable design optimization problems that involve expensive physics-based simulations.

With advances in computational engineering, materials design and discovery have been increasingly viewed as optimization problems with the goal of achieving desired material properties or device performance[1–3]. One challenge of designing new materials systems is the co-existence of qualitative and quantitative design variables associated with material compositions, microstructure morphology, and processing conditions. While microstructure morphology can be described using quantitative, variables such as those associated with correlation function[4], descriptors[3,5,6], and spectral density functions[1,2], many composition and processing conditions are discrete and qualitative by nature. For example, in polymer nanocomposite design, there are numerous choices of material constituents (e.g., the types of filler and matrix) and processing conditions (e.g., the type of surface treatment); each combination follows drastically different physical mechanisms with significant impact on the overall properties[3,7]. As illustrated in Fig. 1, the existence of both quantitative and qualitative material design variables results in multiple disjointed regions in the property/performance space. The combinatorial nature poses additional challenges in materials modeling and the search for optimal solution.

In contrast to the traditional trial-and-error based experiment approach to materials design, computational materials design methods have emerged as an efficient and effective alternative in the past decade, building upon the advancement of simulation techniques such as finite element analysis (FEA) and density functional theory (DFT) that can accurately model and predict material properties at different length scales[8,9]. Recent years have

[1]Mechanical Engineering, Northwestern University, Evanston, IL, US. [2]Industrial Engineering and Management Science, Northwestern University, Evanston, IL, US. *email: weichen@northwestern.edu

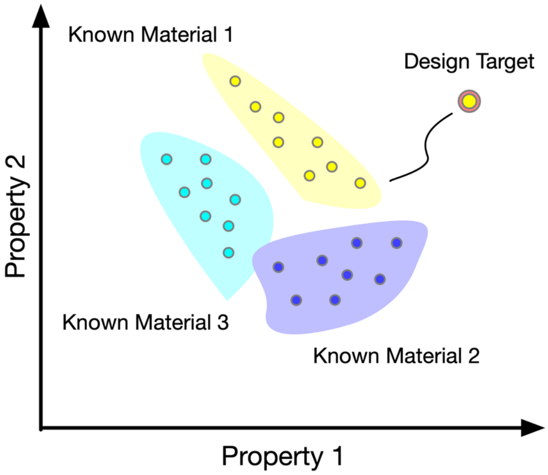

Figure 1. Target space exploration in materials design.

seen new developments in computational methods for designing materials with quantitative and qualitative variables simultaneously, e.g., structure optimization and materials selection of a multilayer beam structure supported by a spring using a genetic algorithm (GA)[10], and optimization of a thermal insulation system using mixed variable programming (MVP) and a pattern search algorithm[11,12]. However, directly employing expensive simulation models in mixed-variable optimization is still restrictive because optimization algorithms commonly require hundreds of evaluations of the objective (material properties) for problems with only quantitative variables and the computational demand is significantly higher with mixed-variables.

To address the issue of expensive simulation, a common strategy in simulation-based optimization is to build a metamodel, a.k.a., response surface model, based on data generated by simulations, and then directly use the metamodel for optimization. Design of experiments (DOE) methods, such as Latin-hypercube sampling (LHS)[13], are often used to improve the overall metamodel accuracy by generating samples that cover the design space as evenly as possible. However, this is not the most efficient approach if the design objective is well defined because it is apparent that more sample points should be selected close to the "optimal" locations rather than uniformly over the whole design space.

In contrast, Bayesian Optimization (BO) provides an adaptive paradigm to sample the design space more efficiently for identifying the global optimum. In particular, a prior response surface model of the objective is prescribed and then sequentially refined as data are observed via an acquisition function[14]. One essential advantage of using BO for materials design is the emergence of various materials databases, such as the polymer nanocomposites data resource NanoMine[15,16], and the open quantum materials database (OQMD) that stores high-throughput DFT data[17]. These databases provide valuable low-cost existing knowledge as a "starting point" for Bayesian inference to guide the rapid exploration of novel material designs. Recent years have seen a number of extensions of using BO in materials design, such as the optimization of the synthesis process of short polymer fibers[18], adaptive optimization of the elastic modulus of the MAX$_2$ phase[19], and the prediction of crystal structures[20]. Nevertheless, these materials design applications of BO are all limited to considering only quantitative design variables, such as the constriction angle, channel width and solvent speed[18], the *s*-, *p*-, and *d*-orbital radii of atoms[19], and structure descriptors[20].

In real applications, most materials design scenarios involve qualitative or categorical factors, such as compositions (selection of material types) and particle surface treatment conditions (e.g., octyldimethylmethoxysilane and aminopropyldimethylethoxysilane) in nanodielectrics[7]. The challenge of including qualitative design factors within BO lies in response surface modeling in the absence of any direct ordering and distance metrics between the factor levels like the inherent distance metric for quantitative design variables. Gaussian process (GP) models, a.k.a., kriging models have become the most popular method for modeling simulation response surfaces[21–23] and thus are widely employed in BO frameworks, because of its flexibility to capture complex nonlinear response surface as well as to quantify uncertainties in prediction. However, these standard GP models are only applicable for quantitative inputs (i.e., quantitative design variables). Most state-of-the-art BO implementations use 0/1 dummy variables to represent qualitative input levels and then fit standard GP models with quantitative inputs, which is essentially equivalent to fitting a multi-response GP model where a different response surface (over the quantitative inputs) is assumed for each combination of qualitative factor levels. This has been shown to be restrictive theoretically and incapable of capturing complex correlations between qualitative levels when the number of levels is large[24–26].

To overcome the aforementioned limitations, we propose a BO framework for data-driven materials design with mixed qualitative and quantitative variables (Fig. 2). The proposed framework is built upon a novel latent variable GP (LVGP) modeling approach that we recently developed for creating response surfaces with both qualitative and quantitative inputs[27]. The key idea of the LVGP approach is to map the qualitative factors into low-dimensional quantitative latent variable (LV) representations. Our study[27] showed that the LVGP modeling dramatically outperforms existing GP models with qualitative input factors in terms of predictive root mean

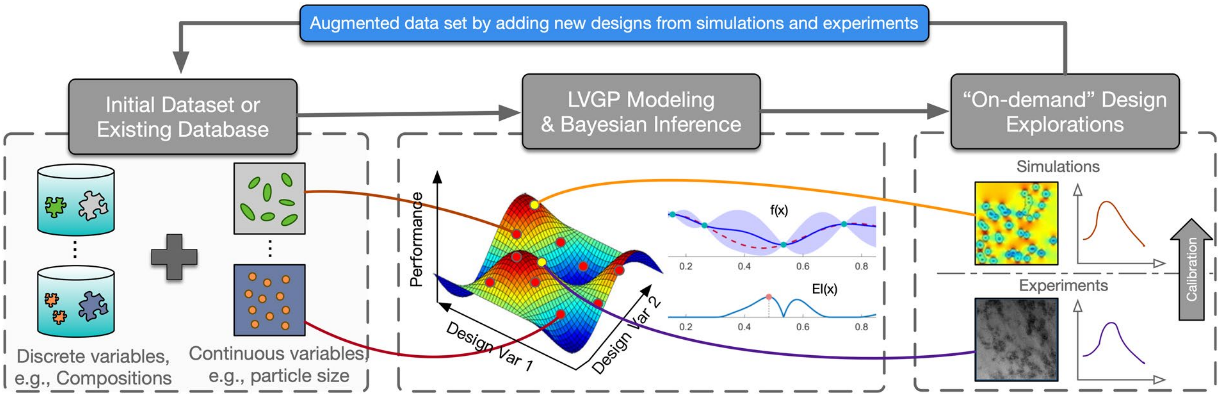
Augmented data set by adding new designs from simulations and experiments
Initial Dataset or Existing Database
LVGP Modeling & Bayesian Inference
"On-demand" Design Explorations
Discrete variables, e.g., Compositions
Continuous variables, e.g., particle size
Performance
Design Var 1
Design Var 2
f(x)
EI(x)
0.2
0.4
0.6
0.8
Simulations
Experiments
Calibration

Figure 2. Bayesian optimization framework for data-driven materials design.

squared error (RMSE). In addition to empirical evidence of far better RMSE predictive performance across a variety of examples, the LVGP approach has strong physical justification in that the effects of any qualitative factor on a quantitative response must *always* be due to some underlying quantitative physical input variables (otherwise, one cannot code the physics of the simulation model)[27]. Noting that the underlying physical variables may be extremely high-dimensional (which is why they are treated as a qualitative factor in the first place), the LVGP mapping serves as a low-dimensional LV surrogate for the high-dimensional physical variables that captures their collective effect on the response. Sufficient dimension reduction arguments[28,29] provide justification for this, since they imply that the collective effects of these high-dimensional underlying physical variables can usually be represented approximately as a function of manifold coordinates over some low-dimensional manifold in the high-dimensional space. We elaborate on this later. In addition to outstanding predictive performance, this LV mapping provides an inherent ordering and structure for the levels of the qualitative factor(s), such as the type of material constituents and processing types, which can provide substantial insights into their influence on the material properties/performance. In this manner the LVGP approach can model a large number of qualitative levels with a relatively small number of parameters, which improves the prediction while maintaining low computational costs. Moreover, in contrast to the existing methods for handling qualitative factors[24–26,30], LVGP is compatible with any standard GP correlation functions for quantitative inputs, including nonseparable correlation functions such as power exponential, Matèrn and lifted Brownian.

In this study, we integrate the LVGP approach with a BO framework (we call it LVGP-BO) for data-driven materials design that involves both qualitative and quantative matereials design values. We examine the LVGP-BO approach over a variety of mathematical and real materials design examples and demonstrate its superior optimization performance over other state-of-the-art methods.

Results

We present two mathematical examples and two materials design examples to demonstrate the efficacy of the proposed LVGP-BO approach for mixed-variable problems and the physical insights it brings. The proposed method is compared to the state-of-the-art BO implementation *bayesopt* in MATLAB[31], which represents qualitative factors as dummy variables first and then fits a standard GP model with quantitative inputs[26]. *Bayesopt* is chosen as the only benchmark for two reasons: (1) Most existing GP-based BO packages that support qualitative inputs (e.g., skopt, a popular python library used to optimize hyperparameters of machine learning models) all treat qualitative inputs as dummy variables and are fundamentally similar to the approach exploited by *bayesopt*. (2) The implementation of our approach relies on Matlab's optimization solver engine, so Matlab's built-in *bayesopt* implementation was chosen as the benchmark to reduce platform and solver engine differences. When comparing both methods, we start with the same set of initial datasets and stop at the maximal allowed number of iterations. The same procedure is repeated multiple times to assess the statistical stability of both methods. In both scenarios, expected improvement (EI)[32] is used as the acquisition function. We denote the LVGP-BO result as "LV-EI" and *bayesopt* as "MC-EI" because its GP model is equivalent to the multiplicative covariance (MC) model in literature[24,26].

Materials design examples.

High-performance light absorbing quasi-random solar cell. We first present a solar cell design problem, in which both the light scattering structure pattern and material selection are optimized simultaneously, in contrast to the original design problem that only considered tuning the scattering pattern represented by quantitative variables[2,33]. Fig. 3 shows the setup of this solar cell design problem: the middle layer with the quasi-random structure is the light-trapping layer of thickness t_1, patterned on the amorphous silicon (a-Si) based absorbing layer with a total thickness of t. The bottom silver layer prevents the light escaping on the back side. The light-trapping structure is optimized for maximizing light absorption at 650 nm by balancing the competing processes of light reflection and scattering. Besides designing the light trapping layer's pattern, we also consider the choices of materials. First, there are five types of a-Si's with different refractive indices, and on top of the light trapping layer, is a 70 nm thin layer of anti-reflection coating (ARC), which can be chosen from five different materials. Table 1 lists the refractive indices (n) of the five a-Si's and five ARC's and the extinction

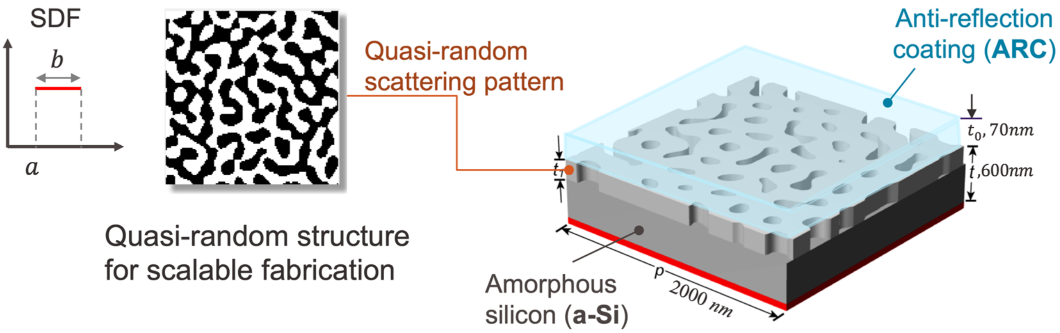
SDF
b
a
Quasi-random scattering pattern
Quasi-random structure for scalable fabrication
Anti-reflection coating (ARC)
t_0, 70nm
t,600nm
t_1
Amorphous silicon (a-Si)
p 2000 nm

Figure 3. Setup of the solar cell design problem: the middle layer is the quasi-random scattering pattern with thickness t_1, directly etched on top of the a-Si substrate of a total thickness of t = 600 nm. The top layer in light blue is an ARC coating to reduce reflection of incident light. We use the SDF based approach to generate the quasi-random scattering pattern.

	a-Si		ARC
Type	*n*	*k*	*n*
1	3.3569	0.0168	2.7225
2	3.1891	0.0160	3.0625
3	3.0212	0.0151	3.4225
4	2.8534	0.0143	3.6100
5	2.6855	0.0134	3.8025

Table 1. Refractive index for the candidate materials.

Quantitative Design Variable	Range
t_1	[50, 300] nm
ρ	[0.1, 0.9] nm
a	[0.0016, 0.0064] nm^{-1}
b	[0.0016, 0.0064] nm^{-1}
Qualitative Design Variable	**Levels**
a-Si Type	{1, 2, 3, 4, 5}
ARC Type	{1, 2, 3, 4, 5}

Table 2. Range of design variables in the solar cell design problem.

coefficient (k) of a-Si. Many existing materials systems involve such concurrent materials selection and structure optimization decisions shown in this example.

Rigorous coupled wave analysis (RCWA)[34,35] is employed to evaluate the light absorption coefficients of the reconstructed structures. RCWA is a Fourier-domain-based algorithm that can solve the scattering problems for both periodic and aperiodic structures. The length of the unit cell for RCWA calculation is set at 2000 nm. The light trapping pattern is represented using the SDF (spectral density function) method as we proposed in[33], which significantly reduces the dimensionality of quasi-random microstructures. In this case, we use a uniform SDF defined by its left end a and width b. These two parameters, along with the space filling ratio ρ, are the three design variables that describe the light trapping layer's pattern. The overall thickness t of a-Si layer is fixed as 600 nm, but the thickness of the light trapping layer t_1 is adjustable in this example. The feasible ranges of all design variables, including both quantitative and qualitative variables, are listed in Table 2.

In this design study, both the LV-EI and MC-EI methods start with 30 random initial samples where the quantitative variables $\{t_1, \rho, a, b\}$ are generated by the maximin LHD and the two materials type variables *a-Si Type* and *ARC type* are sampled uniformly. Both methods are terminated after the maximal allowed 100 iterations, and the same procedure is repeated for 20 replicates. Because there are certain uncertainties associated with the microstructure reconstruction process when generating the quasi-random scattering pattern using SDF, for each fixed pair of $\{\rho, a, b\}$, we generate three statistically equivalent microstructures and take the average of their light absorptions simulated in RCWA as the corresponding response. The acquisition function used in this case is EI with the "plug-in" $\mu_{min}(\boldsymbol{x})$, which is introduced in Method Section for the noisy response scenario. We use "EI" in our notation for simplicity. The left panel of Fig. 4 displays the optimization convergence history of both methods over 100 iterations. It is apparent that the LV-EI method converged significantly faster and also consistently achieved better solutions compared to MC-EI. The best solution found possessing a light absorption coefficient of 0.94, with the optimized quantitative design variables

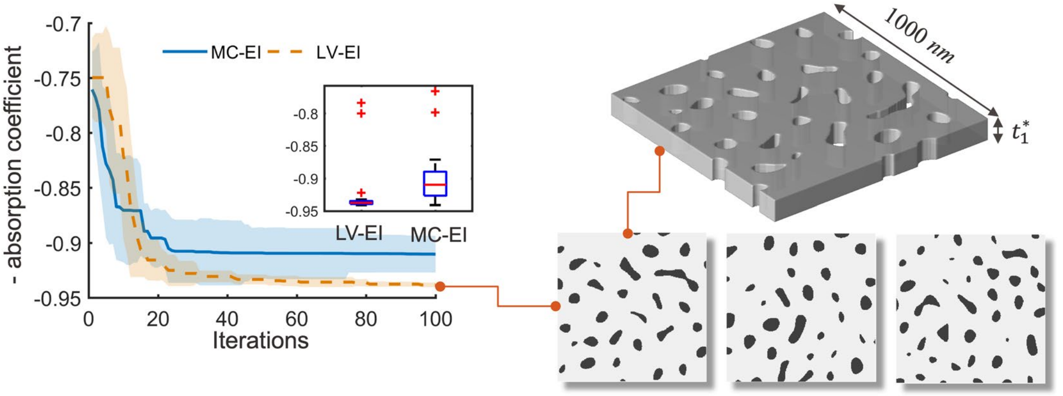

Figure 4. Optimization results of the light absorbing solar cell: the left panel displays the convergence plot of LV-EI and MC-EI method, with medians and median absolute deviations. The LV-EI outperforms MC-EI by consistently achieving better solutions with less uncertainty. The right panel shows three quasi-random patterns according to the results from LV-EI method. The top 3D structure is based on the first 2D pattern.

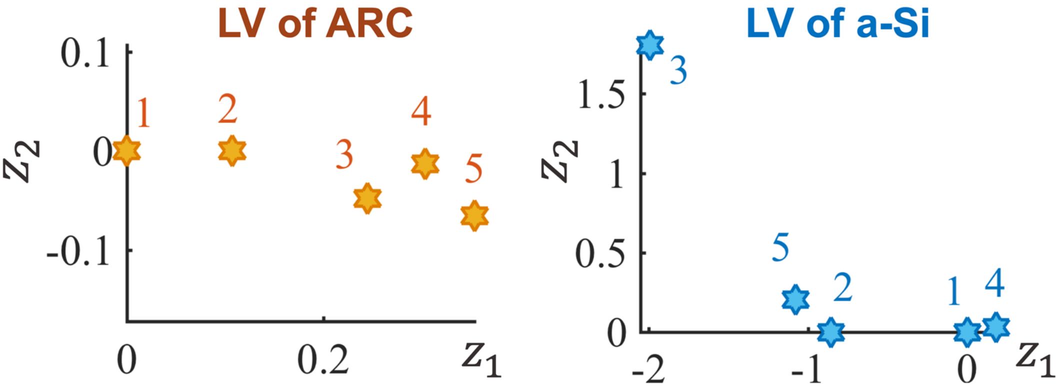

Figure 5. Estimated latent variable of the two qualitative factors represent different materials: the LV of ARC approximately lines up in a straight line following the order 1-2-3-4-5, consistent with their refractive indices; from the LV of a-Si, type 3 is far from the other four, indicating a distinct influence on response compared to the others.

$\{a^* = 0.0049\ nm^{-1},\quad b^* = 0.0035\ nm^{-1},\quad \rho^* = 0.875,\quad t_1^* = 100.95\ nm\}$ and the optimal choices of ARC and a-Si are type 5 and type 3 respectively. The right panel shows three random scattering structures using the optimal design variables.

The LVGP model represents the correlation between qualitative levels by the distances in the 2D latent variable space: larger distances means more difference and less correlation. In this test case, the estimated latent variables of ARC type and a-Si type are illustrated in Fig. 5, which shows that five different ARC materials are positioned approximately in a straight line with the sequence 1-2-3-4-5, consistent with the differences in their refractive indices listed in Table 2. This provides some insights and indicates that the refractive indices are the dominating characteristics of the ARC design factor compared to other simulation inputs, in terms of its effect on the properties. The relationships between the five a-Si's are not straightforward at first glance, because n and k have coupled effects of the overall performance: the real part n quantifies refraction and the imaginary part k represents the loss of flux intensity in the medium, so higher n and lower k are preferred for better absorption. As shown in the LV plot (Fig. 5), the 3rd type a-Si has very different effects on the response compared to the other four types. This can potentially be explained based on the physical parameters presented in Table 1 where a-Si type 3 has both moderate n and k, leading to the best combination of refraction and loss for achieving higher absorption than other four types of a-Si.

Combinatorial search of hybrid organic-inorganic perovskite (HOIP). We present here another materials design example adopted from literature with a combinatorial search of Hybrid Organic-Inorganic Perovskites (HOIPs)[36], where all design variables are qualitative. Such problems are quite popular in materials design with emphasis on design of materials constitutes. HOIPs are an exciting class of new materials that exhibit extremely promising photovoltaic (PV) properties. The goal is to search the perovskite compositional space for an ABX_3 combination with optimal intermolecular binding energy to a solvent molecule, S_0. The A-site cation has three candidates {MA, FA, Cs}, the B-site in HOIP is often occupied by metal cations and in this case, is fixed to be Pb. X denotes the

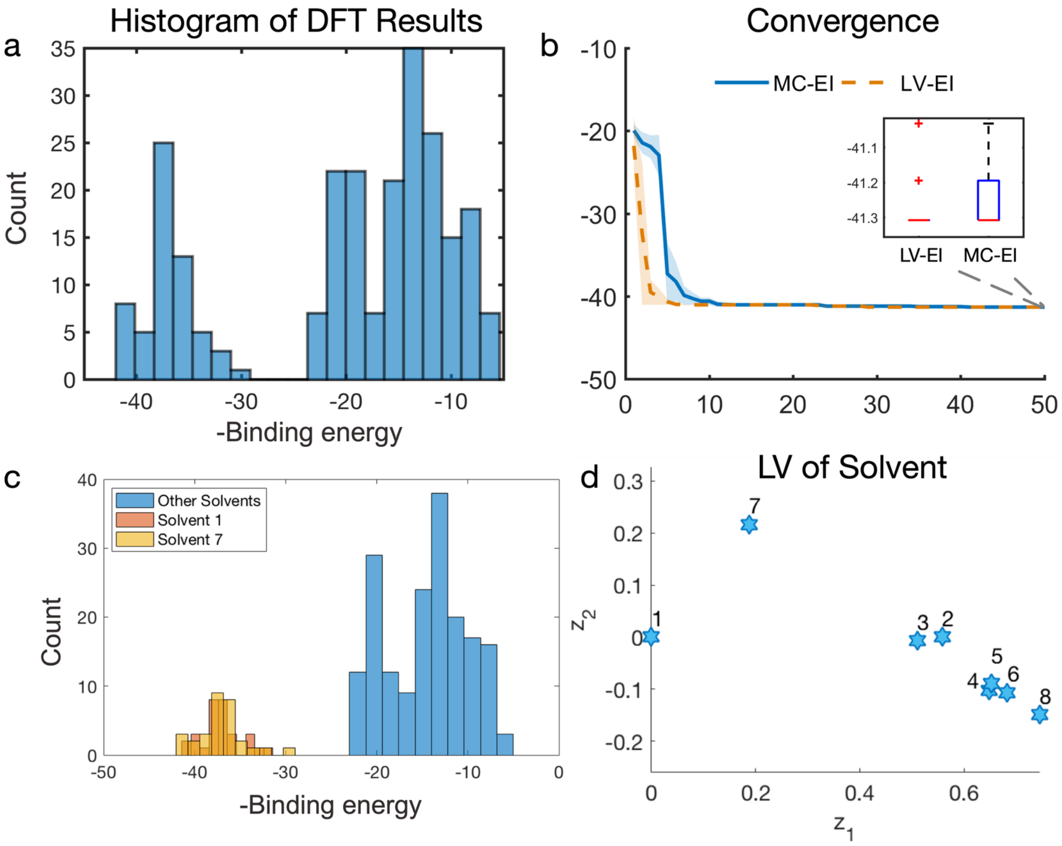

Figure 6. (**a**) histogram of the simulated binding energy from DFT, (**b**) optimization convergence plot of LV-EI and MC-EI methods with 10 initial random samples, median and median absolute deviation at each iteration are plotted, (**c**) negative binding energy distribution by solvent types, (**d**) estimated latent variable for solvent types.

halide, which has three options {Cl, Br, I} and in this design scenario, mixed halides are allowed, which means the three X's in ABX_3 can be different. There are eight different solvents S_0's to explore, and the binding energies between ABX_3 and S_0 are results from DFT calculations.

We took the problem only with qualitative variables for two reasons: (1) GP models in general struggle to handle qualitative variables and pure combinatorial search would be a challenging task for GP based Bayesian optimization and we wanted to test our approach's performance under this setting. (2) The use of only qualitative variables is not necessarily the best formulation to solve this specific design problem. It is possible that other formulations (e.g., using domain knowledge to create quantitative features as surrogates for the qualitative variables that one suspects may have large influence on the response) would have been better for this specific application, but we want to emphasize the generality of our approach and how it can be used without such domain knowledge.

There are five qualitative design variables: the first denotes the choice of the A-site cation (three levels), three other variables indicate the selection of each of the three halides in the configuration (three levels each), and the fifth variable represents the type of the solvent (eight levels). Out of all the possible combinations, 240 are stable, whose binding energies are precomputed by DFT. The code and dataset are made available by the authors of[36] at https://github.com/clancyLab/NCM2018, where the DFT calculations are done through the Physical Analytics Pipeline (PAL) developed by the Clancy Research Group from Cornell University[36]. Fig. 6(a) shows the distributions of the binding energies of the 240 samples, which indicates that majority of the samples have negative binding energies larger than −30, and the best solution has binding energy around −41.3. This ground truth was determined via exhaustive simulation, which the mixed-variable BO algorithms seeks to avoid. Their efficiency can be assessed based on whether the method can identify the true optimal solution from among the 240 combinations using as few evaluations as possible.

When comparing the LV-EI and MC-EI methods, we repeat our tests for 30 replicates, and in each replicate, both methods start with the same set of 10 initial random samples chosen from the samples with binding energy larger than −30 (to make the optimization more challenging), and the optimization was terminated after another 50 iterations. With respect to the capability of identifying the correct optimal solution, from Fig. 6b, we note both methods work well in this combinatorial search problem: the objective function drops quickly within less than ten iterations, and in most of the replicates both methods found the global optimum. Nevertheless, the proposed LV-EI method outperforms MC-EI with smaller variance and quicker convergence. Moreover, the LV-EI method is more robust. Specifically, the boxplot in the inset shows that LV-EI found the exact best solution in 28 of the 30 replicates, whereas MC-EI found the exact best solution in only 22 of 30 replicates.

To better understand the correctness of the fitted LVGP model, we visualize the estimated LVs for solvent type in Fig. 6d: solvent types 1 and 7 are located far from the other six types of solvents, which indicates they might have effects on the binding energy more distinct than the others. To validate this finding, we analyze the distribution for binding energies of the 240 samples in the dataset by singling out solvent types 1 and 7 in Fig. 6c. All the samples with large binding energy (>30) are formed with solvent types 1 and 7 and using other six types of solvents resulted in much smaller binding energies, which is consistent with our interpretation of the estimated LVs in Fig. 6d. Our LVGP model successfully found that THTO is a superior solvent for dissolving ABX3. Using conventional approaches without latent space estimation, it would be difficult to draw such design insights from the original raw dataset. Even without any knowledge or description about the eight (8) different types of solvents, we can tell how much solvent 1 and 7 are different from the other six solvents simply from their locations in the latent space (Fig. 6d) as they are both far from the other six solvents. In conclusion, the proposed mixed variable LVGP-based Bayesian optimization approach effectively searches a combinatorial design space, efficiently identifies the global optimal solution, and provides additional design insights through the LV representation of qualitative factors.

Mathematical examples.

Two mathematical examples are used to illustrate the capability of LVGP-BO in finding the global optimization solutions for mixed variable problems.

Branin function. The first mathematical function considered is the Branin-Hoo function[37], which originally has two quantitative variables x_1 and x_2, while in this example we convert x_2 to be qualitative for testing purpose:

$$f(x_1, x_2) = \left(x_2 - \frac{5.1}{4\pi^2}{x_1}^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10,$$

where $x_1 \in [-5, 10]$ and x_2 is qualitative, with four levels corresponding to the values $\{0, \ 5, \ 10, \ 15\}$. As plotted in Fig. 7a this function has 6 local minimum and a global minimum $f(-2.6, 10) \approx 2.79118$. We can also observe that levels 1 and 2 of x_2 are closely correlated and levels 3 and 4 are closely correlated, while levels 1 and 3 are quite different from each other. Both methods start with 10 random initial points and continue sampling for another 30 iterations. Fig. 7c shows the convergence history of the LV-EI and MC-EI method, from which we see that the LV-EI converges much faster than MC-EI and after 30 iterations LV-EI achieved a more accurate solution as illustrated in the inset of Fig. 7.

Goldstein-price function. The second mathematical example is the Goldstein-price function[37], which also has two input variables x_1 and x_2, where x_2 is made qualitative:

$$f(x_1, x_2) = [1 + (x_1 + x_2 + 1)^2(19 - 14x_1 + 3{x_1}^2 - 14x_2 + 6x_1x_2 + 3{x_2}^2)],$$

where $x_1 \in [-2, 2]$ and x_2 is qualitative, with five levels corresponding to the values $\{-2, \ 1, \ 0, \ 1, \ 2\}$. The global minimum is $f^*(0, \ -1) = 3$. In this example, both methods start with 20 random initial points, and continue sampling for another 30 iterations. Fig. 7d shows the convergence history of the LV-EI and MC-EI method, from which, we see that LV-EI converges much faster than MC-EI consistently out of the 30 replicates. The LV-EI method is also more robust as it has a smaller variance than MC-EI. According to the inset boxplot, we also note that after 30 iterations, most of the 30 replicates of LV-EI are very close to the real global minimum 3 while MC-EI has a much larger gap to the real solution. The two numerical tests shown here illustrate our proposed LVGP based BO approach as an effective method for global optimization of mixed-variable problems.

Discussion

In this paper, we integrate a novel latent variable approach to GP modeling into Bayesian Optimization to support a variety of materials design applications with mixed qualitative and quantitative design variables. The LVGP approach not only provides superior model predictive accuracy compared to existing GP models, it also offers meaningful visualization of the correlation between qualitative levels (Figs. 5 and 6d). The proposed Bayesian Optimization approach is especially useful for materials design as it can fully utilize the existing materials databases and sequentially explores the unknown design space via Bayesian inference and "on-demand" simulations/experiments. Our proposed method overcomes the challenge of using BO for problems with qualitative input factors and has achieved superior performance compared to the state-of-the-art MATLAB BO implementation the using dummy variable representation for qualitative factors, which is demonstrated through both mathematical and real materials design examples (Figs. 4, 6b and 7c,d). For materials design, we successfully utilized the proposed BO framework to improve the light absorption of the quasi-random solar cell by designing the microstructure and selecting materials constituents simultaneously. Moreover, we showed that our method is efficient and effective for optimizing the material constituents of a hybrid organic-inorganic perovskite, a more challenging combinatorial search problem.

While this paper is focused on design of new materials and materials systems, the method presented is generic and can be used for other challenging engineering optimization problems where qualitative and quantitative design variables co-exist. The current BO framework will be further extended for multi-objective problems and the inclusion of physical constraints by refining the sampling strategy in association with the acquisition function.

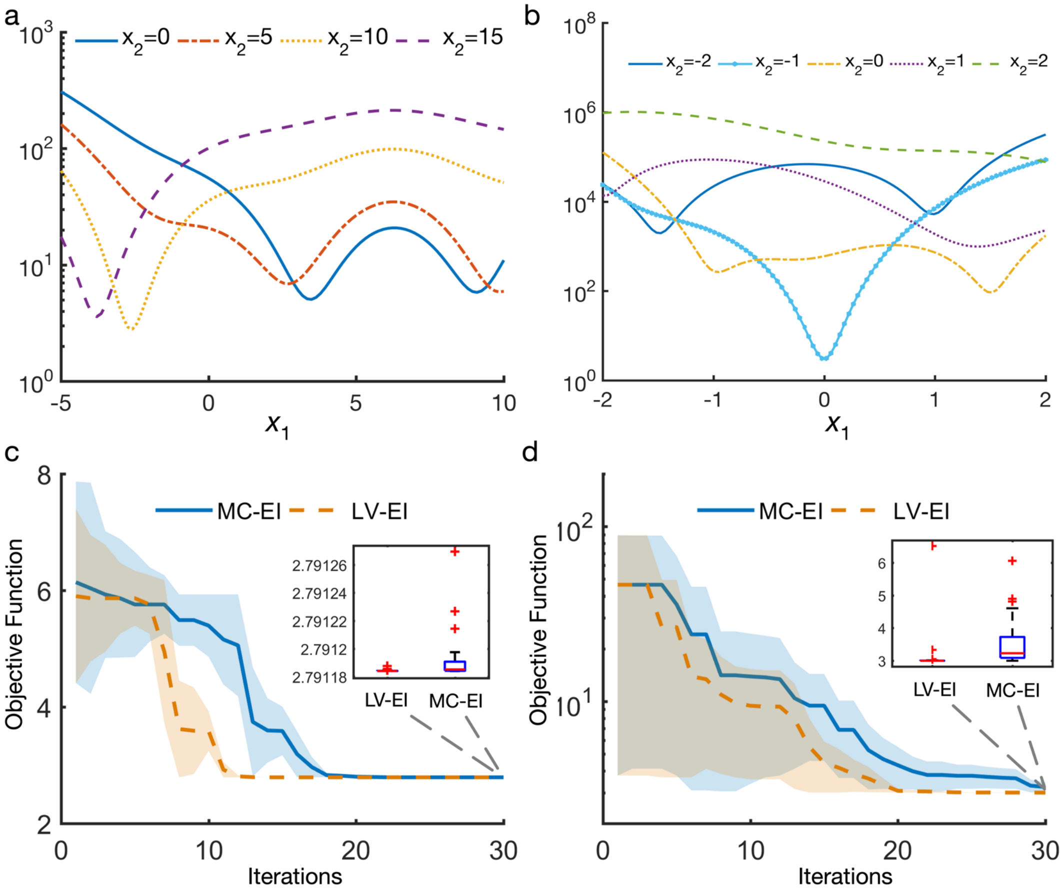

Figure 7. (**a**) Plot of the Branin function, (**b**) Plot of the Gold-Stein Price function, and problems having multiple local minima, and y-axes are in log-scale to visualize the local minimum better; (**c**) Optimization history of 30 replicates for the Branin example, starting with 10 initial sample points; (**d**) Optimization history of 30 replicates for the Gold-Stein Price example, with 20 initial sample points, y-axis in log scale. The middle lines represent the median values of the 30 replicates at each iteration, and the shaded bounds represent median +/− median absolute deviations (MAD). Insets in (**c**) and (**d**) display the boxplots of the minimum obtained after 30 iterations;.

Methods

Latent variable gaussian process (LVGP) for both qualitative and quantitative factors.

We first briefly review the technical details of the standard GP model for quantitative variables, and then describe our novel LVGP model to handle qualitative factors. To make the discussion more concrete, let $y(\cdot)$ denote the true physical response surface model with inputs $\boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{t})$ where $\boldsymbol{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$ represents p quantitative variables and qualitative factors $\boldsymbol{t} = \left(t_1, \ldots, t_q\right) \in \{1, 2, \ldots, m_1\} \times \{1, 2, \ldots, m_2\} \times \ldots \times \left\{1, 2, \ldots, m_q\right\}$, where the jth qualitative factor t_j has m_j levels that are coded (without loss of generality) as $\left\{1, 2, \ldots, m_j\right\}$. A GP model with only quantitative input variables is commonly assumed to be of the form:

$$y(\boldsymbol{x}) = \mu + Z(\boldsymbol{x}), \tag{1}$$

where μ is a constant prior mean term, $Z(\cdot)$ is a zero-mean Gaussian process with stationary covariance function $K(\cdot,\cdot) = \sigma^2 R(\cdot,\cdot)$, σ^2 is the prior variance, and $R(\cdot,\cdot) = R(\cdot,\cdot|\phi)$ denotes the correlation function with parameters ϕ. A Gaussian correlation function is commonly used:

$$R(\boldsymbol{x}, \ \boldsymbol{x}') = \exp\left\{-\sum_{i=1}^{p} \phi_i (x_i - x'_t)^2\right\}, \tag{2}$$

which represents the correlation between $Z(\boldsymbol{x})$ and $Z(\boldsymbol{x}')$ for any two input locations $\boldsymbol{x} = (x_1, \ \ldots, \ x_p)$ and $\boldsymbol{x}' = (x_1', \ \ldots, \ x_p')$, where $\boldsymbol{\phi} = \left(\phi_1, \ \ldots, \ \phi_p\right)^T$ is the vector of correlation parameters to be estimated via MLE, along with μ and σ^2. The correlation between $y(\boldsymbol{x})$ and $y(\boldsymbol{x}')$ depend on the spatial distance between $\boldsymbol{x}$ and $\boldsymbol{x}'$ and the correlation parameters. Other choices of correlation functions include power exponential, Matèrn[38] and lifted Brownian[39]. These types of correlation functions cannot be directly applied to problems with qualitative factors

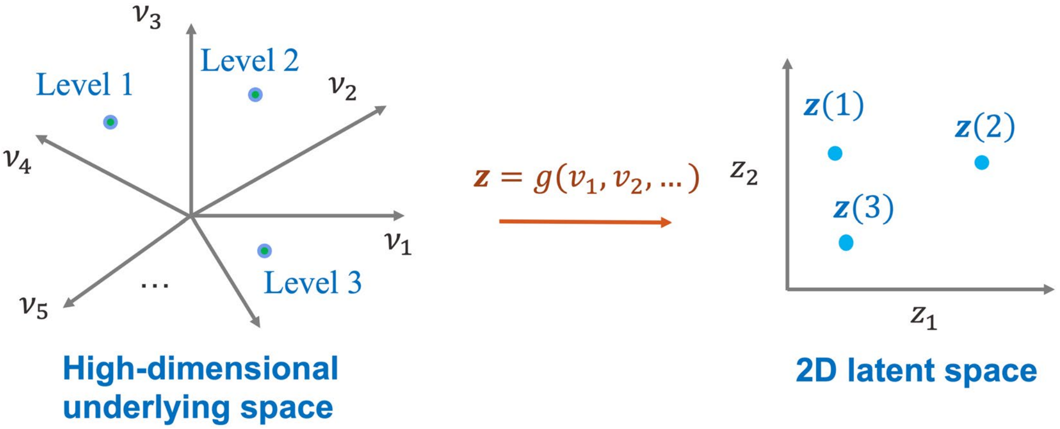

Figure 8. For a single factor $\boldsymbol{t}$ with three levels, depiction of the mapping from the true high-dimensional underlying quantitative variables to the 2D latent variables.

because the distances between levels of qualitative factors are not defined, and the levels have no natural ordering since $\boldsymbol{t}$ are assumed to be nominal categorical factors.

The LVGP approach[27] provides a natural and convenient way to handle qualitative input variables by mapping the levels of each qualitative factor t to a 2-dimensional (2D) continuous latent space. This has strong physical justification, which may explain the outstanding predictive performance of the LVGP approach[27]: For any real physical system with a qualitative input factor t, there are always underlying quantitative physical variables $\{v_1, v_2, \ldots\} = \{v_1(t), v_2(t), \ldots\}$ (perhaps very high-dimensional, poorly understood, and difficult to treat individually as quantitative input variables) that account for the differences between the responses at different levels of the qualitative factors. As shown in Fig. 8, the three qualitative levels are associated with points in a high-dimensional space of $\{v_1, v_2, \ldots\}$, and the distances between the three points indicates the differences between the three levels. Our LVGP model uses a low-dimensional (2D) representation $\boldsymbol{z}(t) = g(v_1(t), v_2(t), \ldots)$ to approximate the actual distances between the three levels. This implicitly makes the rather mild assumption that the collective effect of $\{v_1(t), v_2(t), \ldots\}$ on y, as t varies across its levels, depends predominantly on some low-dimensional combinations $\boldsymbol{z}(t) = g(v_1(t), v_2(t), \ldots)$ of the underlying high-dimensional variables (reminiscent of sufficient dimension reduction arguments[28,29]). In many applications, a two-dimensional representation $\boldsymbol{z}(t)$ suffices to approximate the high dimensional data.

To make these arguments more concrete, first consider the following beam bending example[27], in which the qualitative factor t corresponds to six different beam cross-sectional shapes. The underlying numerical variables $\{v_1(t),\ v_2(t),\ v_3(t),\ \ldots\}$ for cross-section type t would be the complete cross-sectional geometric positions of all the elements in the finite element mesh of the beam. The physics of this beam bending problem is well studied that we know the beam deflection depends on $\{v_1(t),\ v_2(t),\ v_3(t),\ \ldots\}$ via the moment of inertia $I = I(t) = I(v_1(t),\ v_2(t),\ v_3(t),\ \ldots)$. Therefore, the impact of the underlying high-dimensional numerical variables $\{v_1(t),\ v_2(t),\ v_3(t),\ \ldots\}$ on the response can be reduced to a single numerical variable $I(t)$. Reference[27] demonstrated that the $\boldsymbol{z}(t)$ estimated by the LVGP approach was a direct representation of $I(t)$. Second, consider the case where no definitive prior knowledge or a simulator with transparent mechanisms are available and the actual underlying quantitative space are high dimensional, an example of which is modeling of shear modulus (the response, y simulated by DFT) of material compounds belonging to the family of M2AX phases[27]. The M, A and X atoms have ten, two and twelve candidate choices respectively. Our study showed that the LVGP model using three qualitative factors that represent atom types with a 2D latent representation significantly outperformed the GP model with seven quantitative atom descriptors that were believed to have a large impact on the compound shear modulus. In this case, even though the actual underlying numerical spaces ares very high dimensional (formed by various properties of each atom), the proposed LVGP model still managed to achieve superior modeling results with a low-dimensional latent representation.

In the relatively rare situations that the impact of each qualitative variable cannot be represented in a two-dimensional space, the latent variable component $\boldsymbol{z}^j(t_j) = (z_1^j(t_j),\ z_2^j(t_j))$ in (3) can be easily extended to three and higher dimensions $\boldsymbol{z}^j(t_j) = (z_1^j(t_j),\ \ldots,\ z_{dim\text{-}_z}^j(t_j))$ to increase the model's flexibility. In our published R package "LVGP"[40], the dimension of latent variable space dim_$_z$ is treated as a tunable model hyperparameter, so that users have the flexibility to choose the appropriate dimension of latent space if needed.

When there are multiple qualitative factors, let $\boldsymbol{z}^j(t_j) = (z_1^j(t_j),\ z_2^j(t_j))$ denotes the 2D mapped LV for the qualitative factor t_j. For a Gaussian correlation function, the LVGP approach then assumes the correlation function

$$cor\Big(y\Big(\boldsymbol{x}, \boldsymbol{t} = \Big(t_1, \ldots, t_q\Big)\Big), y\Big(\boldsymbol{x}', \boldsymbol{t}' = t'_1, \ldots, t'_q\Big)\Big) = \exp\left\{-\sum_{i=1}^{p}\phi_i(x_i - x_i')^2 - \sum_{j=1}^{q}\Big\|\boldsymbol{z}^j\Big(t_j\Big) - \boldsymbol{z}^j\Big(t_j'\Big)\Big\|_2^2\right\}, \tag{3}$$

where $||\cdot||$ denotes the L_2 norm. Note that the 2D mapped LVs $\boldsymbol{z}^j(t_j) = (z_1^j(t_j),\ z_2^j(t_j))$ are unknown and will be estimated along with μ, σ^2 and ϕ through maximum likelihood estimation (MLE). Under model (3), the log-likelihood function is:

$$l(\mu, \sigma^2, \phi, \mathbf{Z}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\ln|\mathbf{R}(\phi, \mathbf{Z})| - \frac{1}{2\sigma^2}(\mathbf{y} - \mu\mathbf{1})^T\mathbf{R}(\phi, \mathbf{Z})^{-1}(\boldsymbol{y} - \mu\mathbf{1}), \tag{4}$$

where n is the sample size, $\mathbf{1}$ is an n-by-1 vector of ones, y is the n-by-1 vector of observed response values, $\mathbf{Z} = (\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^q)$, $\boldsymbol{Z}^j = (\boldsymbol{z}^j(1), \ldots, \boldsymbol{z}^j(m_j))$ represents the values of the LV corresponding to the m_j levels of the qualitative variable t_j, and $\mathbf{R} = \mathbf{R}(\phi, \mathbf{Z})$ is the n-by-n correlation matrix whose elements are obtained by plugging pairs of the n sample values of $(\boldsymbol{x},\ t)$ into (3). The MLE of μ and σ^2 in (4) can be represented in terms of the correlation matrix:

$$\hat{\mu} = (1^T\mathbf{R}^{-1}\mathbf{1})^{-1}\mathbf{1}^T\mathbf{R}^{-1}\mathbf{y}$$

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \hat{\mu}\mathbf{1})^T\mathbf{R}^{-1}(\mathbf{y} - \hat{\mu}\mathbf{1}). \tag{5}$$

Substituting the above into (4) and neglecting constants, the log-likelihood function becomes:

$$l(\phi, \mathbf{Z}) \sim -n\ln(\hat{\sigma}^2) - \ln|\mathbf{R}(\phi, \mathbf{Z})|, \tag{6}$$

which is maximized over the correlation matrix $\mathbf{R}$ that depends on the correlation parameters ϕ and the values of the mapped latent variables in $\mathbf{Z}$.

After the MLEs of ϕ and Z are obtained, the LVGP response predictions at any new point x^* is:

$$\hat{y}(\boldsymbol{x}^*) = \hat{\mu} + \boldsymbol{r}(\boldsymbol{x}^*)\mathbf{R}^{-1}(\mathbf{y} - \hat{\mu}\mathbf{1}), \tag{7}$$

where $\boldsymbol{r}(\boldsymbol{x}^*) = (R(\boldsymbol{x}^*, \boldsymbol{x}^{(1)}), \ldots, R(\boldsymbol{x}^*, \boldsymbol{x}^{(n)}))$ is a vector of the pairwise correlation between x^* and each data point $\boldsymbol{x}^{(j)},\ j = 1, \ldots, n$. Moreover, to quantify predictive uncertainty, the variance of the error for this prediction is:

$$\hat{s}^2(\boldsymbol{x}^*) = \hat{\sigma}^2(\boldsymbol{r}(\boldsymbol{x}^*) - \boldsymbol{r}(\boldsymbol{x}^*)\mathbf{R}^{-1}\boldsymbol{r}(\boldsymbol{x}^*)^T). \tag{8}$$

When the actual model is non-deterministic, we add an extra "nugget" parameter λ to each diagonal element of the correlation matrix R to account for the noise of the response and it is estimated along with ϕ and Z.

LVGP-BO framework for materials design. Our proposed Bayesian Optimization framework for data-driven materials design consists four major steps (as shown in Fig. 1): (1) Step 1 involves creating a materials dataset (either physical or/and computer data) based on the information gathered from literature, lab experiments and simulations, (2) Step 2 fits the LVGP model using the available dataset and provides uncertainty quantification of model prediction based on the nature of data, (3) Step 3 makes inference about where to sample the next point (either physical or computer experiments) based on an acquisition function that balances sampling where the response appears to be optimized (exploitation) vs. where the predictive uncertainty is high (exploration), and (4) Step 4 evaluates the chosen design point(s) to augment the materials database and update the metamodel. As this procedure keeps going, more sample points will be sequentially added to update the metamodel prediction and identify the global optimum solution.

There are two key components of the BO framework, a metamodel that provides predictions with uncertainty quantification, and a criterion that determines where to sample next. The LVGP model introduced in the previous section provides robust approximations of the actual response surface model with mixed variable types, as well as the uncertainty quantification. The next critical step is to consider where to sample next based on inferences of the fitted model (Step 3 "on-demand" design exploration on Fig. 2), through a measure of the value of the information gained from sampling at a certain point, known as acquisition functions. Three commonly used acquisition functions in the literature are expected improvement (EI)[32], probability of improvement (PI), and lower confidence bound (LCB). For deterministic responses, expected improvement (EI) is the most widely used and works well over a variety of problems, while for noisy responses, EI with plug-in and knowledge gradient (KG) are proper choices[37,41]. Detailed benchmark studies of different acquisition functions are available in[37].

EI balances exploitation and exploration as elaborated in Fig. 9: a GP model is first fitted based on the four sample points (Fig. 9a), and the fitted mean prediction $\hat{y}(\boldsymbol{x})$ is the blue curve, while the actual response $y(\boldsymbol{x})$ is the red dashed curve. Instead of picking the minimizer of $\hat{y}(\boldsymbol{x})$ as the next sample point, EI also considers the model uncertainties. Mathematically, EI quantifies the possible improvement of a particular point by incorporating both $\hat{y}(\boldsymbol{x})$ and the associated uncertainty $\hat{s}^2(\boldsymbol{x})$:

$$EI(\boldsymbol{x}) = E[\max(0, \Delta(\boldsymbol{x}))] = \hat{s}(\boldsymbol{x})\phi\left(\frac{\Delta(\boldsymbol{x})}{\hat{s}(\boldsymbol{x})}\right) + \Delta(\boldsymbol{x})\Phi(\frac{\Delta(\boldsymbol{x})}{\hat{s}(\boldsymbol{x})}), \tag{9}$$

Where $\hat{y}(\boldsymbol{x})$ and $y_{\min}$ are the mean prediction of the fitted GP model and the minimal value observed so far, $\Delta(\boldsymbol{x}) = y_{min} - \hat{y}(\boldsymbol{x})$, ϕ and Φ are the probability density function (PDF) and cumulative density function (CDF) of standard normal distribution. In practice, a trade-off parameter ξ can be included in $\Delta(\boldsymbol{x}) = y_{min} - \hat{y}(\boldsymbol{x}) - \xi$ to balance exploration and exploitation and higher ξ values lead to more exploration. Equation (9) is a special case

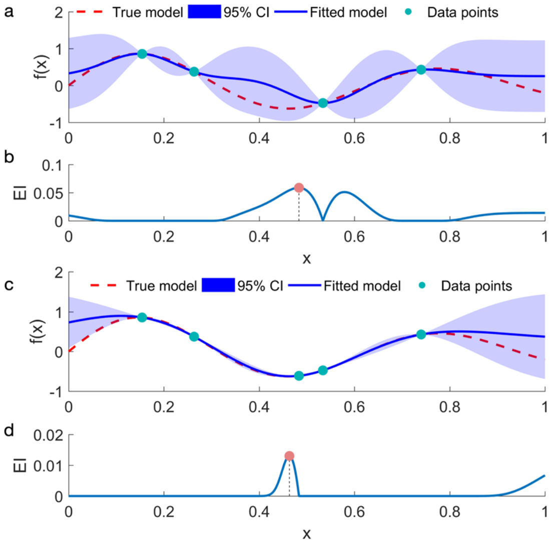

Figure 9. GP model evolution using EI as the acquisition function: (**a**) initial GP model based on four data points, (**b**) EI of the initial GP model, (**c**) updated GP model after sampling one additional data point based on EI, (**d**) EI of the updated GP model.

(0) Generate initial dataset $\boldsymbol{D}_0$ (1) **For** $n = 1, 2, \ldots,$ **do**
(2) Fit the latent variable GP model $\hat{y}_n(\boldsymbol{x}; \boldsymbol{D}_{n-1})$
(3) Select the next sampling point $\boldsymbol{x}_{n+1}$ by maximizing *EI*:
$\boldsymbol{x}_{n+1} = \underset{\boldsymbol{x}}{\textbf{argmax}} \boldsymbol{EI}(\boldsymbol{x}; \hat{y}_n)$
(4) Query simulations/experiments to obtain y_{n+1}
(5) Augment data $\boldsymbol{D}_{n+1} = \{\boldsymbol{D}_n, \ (\boldsymbol{x}_{n+1}, y_{n+1})\}$
(6) **End for**.

Table 3. Bayesian Optimization Algorithm.

where the trade-off parameter $\xi = 0$ and is used for the EI implementation across all examples in this paper. The maximizer of EI in Fig. 9b is chosen as the next sampling point and a new GP model is fitted with this additional data point (shown in Fig. 9c). The updated mode provides a more accurate approximation of the true model with much less uncertainty. The updated EI profile (Fig. 9d) also indicates that the region around the true response optimizer has a large expected improvement. This BO algorithm is summarized in Table 3.

Data availability

The source code of this work will be made available upon request to the corresponding author.

Received: 27 September 2019; Accepted: 5 February 2020;
Published online: 18 March 2020

References

1. Yu, S. *et al.* Design of Non-Deterministic Quasi-random Nanophotonic Structures Using Fourier Space Representations. *Sci. Rep.* **7**, 1–10 (2017).
2. Yu, S. *et al.* Characterization and Design of Functional Quasi-Random Nanostructured Materials Using Spectral Density Function. *J. Mech. Des.*, https://doi.org/10.1115/1.4036582 (2017).
3. Zhang, Y. *et al.* Microstructure reconstruction and structural equation modeling for computational design of nanodielectrics. *Integr. Mater. Manuf. Innov.* **4**, 14 (2015).
4. Torquato, S. S. & Tatistical, D. Escription of M Icrostructures. *Annu. Rev. Mater. Res.* **32**, 77–111 (2002).
5. Xu, H., Dikin, D. A., Burkhart, C. & Chen, W. Descriptor-based methodology for statistical characterization and 3D reconstruction of microstructural materials. *Comput. Mater. Sci.* **85**, 206–216 (2014).
6. Xu, H., Liu, R., Choudhary, A. & Chen, W. A Machine Learning-Based Design Representation Method for Designing Heterogeneous Microstructures. *J. Mech. Des.* **137**, 051403 (2015).
7. Hassinger, I. *et al.* Toward the development of a quantitative tool for predicting dispersion of nanocomposites under non-equilibrium processing conditions. *J. Mater. Sci.* **51**, 4238–4249 (2016).
8. Orio, M., Pantazis, D. A. & Neese, F. Density functional theory. *Photosynthesis Research*, https://doi.org/10.1007/s11120-009-9404-8 (2009).
9. Szabo, B. A. (Barna, A., Babuška, I. & Babuška, I.). *Finite element analysis*. (Wiley, 1991).

10. Tang, X., Bassir, D. H. & Zhang, W. Shape, sizing optimization and material selection based on mixed variables and genetic algorithm. *Optim. Eng.* **12**, 111–128 (2011).
11. Kokkolaras, M., Audet, C. & Dennis, J. E. Mixed Variable Optimization of the Number and Composition of Heat Intercepts in a Thermal Insulation System. *Optim. Eng.* **2**, 5–29 (2001).
12. Abramson, M. A. Mixed Variable Optimization of a Load-Bearing Thermal Insulation System Using a Filter Pattern Search Algorithm. *Optim. Eng.* **5**, 157–177 (2004).
13. Park, J. S. Optimal Latin-hypercube designs for computer experiments. *J. Stat. Plan. Inference* **39**, 95–111 (1994).
14. Melorose, J., Perroy, R. & Careas, S. Taking the Human Out of the Loop: A Review of Bayesian Optimization. **1** (2015).
15. Zhao, H. *et al.* Perspective: NanoMine: A material genome approach for polymer nanocomposites analysis and design. *APL Mater.*, https://doi.org/10.1063/1.4943679 (2016).
16. Zhao, H. *et al.* NanoMine schema: An extensible data representation for polymer nanocomposites. *APL Mater.*, https://doi.org/10.1063/1.5046839 (2018).
17. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *JOM*, https://doi.org/10.1007/s11837-013-0755-4 (2013).
18. Li, C. *et al.* Rapid Bayesian optimisation for synthesis of short polymer fiber materials. *Sci. Rep.*, https://doi.org/10.1038/s41598-017-05723-0 (2017).
19. Balachandran, P. V., Xue, D., Theiler, J., Hogden, J. & Lookman, T. Adaptive Strategies for Materials Design using Uncertainties. *Sci. Rep.* **6**, 19660 (2016).
20. Yamashita, T. *et al.* Crystal structure prediction accelerated by Bayesian optimization. *Phys. Rev. Mater.*, https://doi.org/10.1103/PhysRevMaterials.2.013803 (2018).
21. Fang, K., Li, R. Z. & Sudjianto, A. Design and modeling for computer experiments. *Computer Science and Data Analysis Series*, https://doi.org/10.1111/j.1467-985X.2006.00446_7.x (2006).
22. Sacks, J., Welch, W. J., Mitchell, T. J. & Wynn, H. P. Design and Analysis of Computer Experiments. *Stat. Sci.* **4**, 409–423 (1989).
23. Santner, T. J., Williams, B. J. & Notz, W. I. The Design and Analysis of Computer Experiments. **286**, https://doi.org/10.1007/978-1-4757-3799-8 (2003).
24. McMillan, N. J., Sacks, J., Welch, W. J. & Gao, F. Analysis of protein activity data by gaussian stochastic process models. *J. Biopharm. Stat.* **9**, 145–160 (1999).
25. Qian, P. Z. G., Wu, H. & Wu, C. F. J. Gaussian Process Models for Computer Experiments With Qualitative and Quantitative Factors. *Technometrics* **50**, 383–396 (2008).
26. Zhang, Y. & Notz, W. I. Computer experiments with qualitative and quantitative variables: A review and reexamination. *Qual. Eng.* **27**, 2–13 (2015).
27. Zhang, Y., Tao, S., Chen, W. & Apley, D. W. A Latent Variable Approach to Gaussian Process Modeling with Qualitative and Quantitative Factors. *Technometrics* 1–19, https://doi.org/10.1080/00401706.2019.1638834 (2019).
28. Cook, R. D. & Ni, L. Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Am. Stat. Assoc.* **86**, 316–327 (2005).
29. Li, K. C. Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.* **86**, 316–327 (1991).
30. Deng, X., Lin, C. D., Liu, K. W. & Rowe, R. K. Additive Gaussian Process for Computer Models With Qualitative and Quantitative Factors. *Technometrics* **59**, 283–292 (2017).
31. MATLAB and Statistical and Machine Learning Toolbox Release, The MathWorks, Inc., Natick, Massachusetts, United States (2016b).
32. Jones, D. R., Schonlau, M. & Welch, W. J. Efficient Global Optimization of Expensive Black-Box Functions. *J. Glob. Optim.* **13**, 455–492 (1998).
33. Yu, S. C. *et al.* Characterization and Design of Functional Quasi-Random Nanostructured Materials Using Spectral Density Function. *J. Mech. Des.* **139**(7), 071401. https://doi.org/10.1115/1.4036582 **139**, 135–145 (2016).
34. Li, L. New formulation of the Fourier modal method for crossed surface-relief gratings. *J. Opt. Soc. Am. A*, https://doi.org/10.1364/JOSAA.14.002758 (1997).
35. Moharam, M. G., Gaylord, T. K., Pommet, D. A. & Grann, E. B. Stable implementation of the rigorous coupled-wave analysis for surface-relief gratings: enhanced transmittance matrix approach. *J. Opt. Soc. Am. A*, https://doi.org/10.1364/JOSAA.12.001077 (1995).
36. Herbol, H. C., Hu, W., Frazier, P., Clancy, P. & Poloczek, M. Efficient search of compositional space for hybrid organic–inorganic perovskites via Bayesian optimization. *npj Comput. Mater.*, https://doi.org/10.1038/s41524-018-0106-7 (2018).
37. Picheny, V., Wagner, T. & Ginsbourger, D. A benchmark of kriging-based infill criteria for noisy optimization. *Struct. Multidiscip. Optim.* **48**, 607–626 (2013).
38. Rasmussen, C. E., Williams, C. K. I., Processes, G., Press, M. I. T. & Jordan, M. I. *Gaussian Processes for Machine Learning*, the MIT Press, 2006, ISBN 026218253X. (2006).
39. Plumlee, M. & Apley, D. W. Lifted Brownian Kriging Models. *Technometrics* **59**, 165–177 (2017).
40. Tao, S. & Zhang, Y. LVGP: Latent Variable Gaussian Process Modeling with Qualitative and Quantitative Input Variables, https://cran.r-project.org/web/packages/LVGP/LVGP.pdf (2018).
41. Vazquez, E., Villemonteix, J., Sidorkiewicz, M. & Walter, É. Global optimization based on noisy evaluations: An empirical study of two statistical approaches. *J. Phys. Conf. Ser.*, https://doi.org/10.1088/1742-6596/135/1/012100 (2008).

Acknowledgements

This work was supported in part by AFOSR (FA9550-18-1-0381) and National Science Foundation (NSF) Grants including CMMI 1662435, EEC 1530734, and DMREF 1729743.

Author contributions

The LVGP approach was developed jointly by Y. Zhang, D. Apley, and W. Chen. Y. Zhang and W. Chen formulated the idea of extending this approach to Bayesian Optimization. Code implementation was done by Y. Zhang, and the applications to numerical examples and material design problems were completed by Y. Zhang with continuous discussions and iterations with W. Chen and D. Apley. All authors reviewed the manuscript and contributed to the revision.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.C.

Reprints and permissions information is available at www.nature.com/reprints.