Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Designing Shear-thinning Hydrogel Nanocomposites Using Mixed-variable Bayesian Optimization

Wei Chen

weichen@northwetsern.edu

Northwestern University

Gourav Kumbhojkar

Northwestern University https://orcid.org/0000-0003-4253-7247

George Varghese P J

Keren Zhao

Sinan Keten

Northwestern University https://orcid.org/0000-0003-2203-1425

Jingjie Hu

Article

Keywords:

Posted Date: June 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9886097/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Designing Shear-thinning Hydrogel Nanocomposites Using Mixed-variable Bayesian Optimization

Gourav Kumbhojkar[1], George Varghese P J[2], Keren Zhao[2], Sinan Keten[1,3], Jingjie Hu[2], Wei Chen[1*]

[1] Department of Mechanical Engineering, Northwestern University, Evanston, Illinois, USA.

[2] Department of Mechanical and Aerospace Engineering, North Carolina State University, Raleigh, NC, USA.

[3] Department of Civil and Environmental Engineering, Northwestern University, Evanston, Illinois, USA.

* Corresponding author (weichen@northwetsern.edu)

Abstract

Shear-thinning hydrogel composites are promising materials for transarterial embolization, a minimally invasive procedure used to occlude blood vessels, owing to their injectability, mechanical strength, and biocompatibility. However, designing these materials is challenging because their properties depend sensitively on both polymer type and composition, leading to a complex mixed-variable optimization problem. Here we show that a machine learning framework combining latent-variable Gaussian processes (LVGP) with Bayesian optimization (BO) can efficiently design hydrogel nanocomposites formulated using nanoclay and natural polymers such as chitosan, gelatin, and alginate, with targeted storage modulus. This approach enables simultaneous exploration of multiple polymer systems and reduces experimental effort by 67% compared to conventional Bayesian optimization methods that optimize multiple polymers one at a time. Furthermore, our model learns a latent representation that captures similarities between composition-property relationships of different polymers, providing insight into material behavior. These results demonstrate a scalable strategy for accelerating the design of multi-material systems and highlight the potential of machine learning to address complex formulation challenges in soft materials and biomedical engineering.

1. Introduction

Shear-thinning hydrogels are a class of materials that reversibly transition between solid-like and liquid-like states under applied stress [1]. This property enables them to flow under shear and rapidly recover their structure once the stress is removed, making them highly suitable for injectable biomedical applications. In particular, shear-thinning hydrogel composites (STHCs), composed of two or more interacting components, have emerged as promising candidates for transarterial embolization, a minimally invasive endovascular procedure used to occlude targeted blood vessels in conditions such as intracranial aneurysms,

arteriovenous malformations, and hypervascular tumors including hepatocellular carcinoma, renal cell carcinoma, and uterine fibroids [1–6].

Current embolic agents exhibit several limitations. Metallic coils are constrained by their fixed geometries and relatively large dimensions, limiting their effectiveness in small or irregular vasculatures [2,7,8]. Microspheres, available in sizes ranging from tens to hundreds of micrometers, can target smaller vessels but typically require sequential administration of multiple size ranges to achieve complete occlusion [2,9,10]. This size escalation increases procedural complexity and variability and may lead to incomplete embolization. Additionally, microspheres can fragment under injection pressure, increasing the risk of distal migration, non-target embolization, and recanalization [2]. Liquid embolics such as Lipiodol, n-butyl cyanoacrylate (nBCA), and Onyx also present challenges, including rapid washout, premature polymerization, solvent-related cytotoxicity, and the risk of catheter occlusion [2]. STHCs address many of these limitations by enabling size-independent vessel occlusion, precise catheter-based delivery, and solvent-free formulations [1,4,11]. Among these, nanoclay (Laponite)- polymer systems based on chitosan, gelatin, or alginate have gained significant attention [1,4,12]. Laponite exhibits an anisotropic charge distribution, with negatively charged faces and positively charged edges, forming a dynamic “house-of-cards” structure in aqueous environments [12]. This network can reversibly break and reform under shear, imparting intrinsic shear-thinning behavior. Electrostatic interactions between nanoclay particles and polymers further enable tunable mechanical properties and provide functional sites for incorporating contrast agents or therapeutics. Such functionalization enhances radiopacity for real-time imaging and broadens the therapeutic potential of these systems, positioning STHCs as a next-generation platform for embolization therapies [1].

The rheological behavior of STHCs is critical to their clinical performance. Under shear stress, these materials transition from an elastic, solid-like state to a viscous, liquid-like state, enabling smooth and controlled delivery through clinically relevant catheters. This shear-thinning behavior facilitates deep penetration into distal and complex vascular networks. Upon cessation of shear, STHCs rapidly recover their mechanical integrity, forming stable occlusions across a wide range of vessel diameters, including microvasculature. However, designing shear-thinning hydrogel nanocomposites with precise mechanical and rheological characteristics remains a challenge. The mechanical properties are sensitive to composition and the choice of polymer [1–4]. The complex interactions between charged NC particles and polymers lead to nonlinear trends in rheological properties. Therefore, identifying formulations with desired mechanical properties requires selecting the right composition as well as the polymer.

To ensure the stability of the occlusion, the storage (or elastic) modulus (G’) of the gel should be high enough to withstand the physiological blood pressure. It has been suggested that the initial G′ of the embolic agent needs to be higher than 800 Pa upon injection to resist a pressure up to 200 mmHg [5]. However, previous studies have established in vivo that G’ requirement depends on the hemodynamic pressure of the target vessel. At the lower end, ~200 Pa is sufficient for low-pressure peripheral veins and small vessels keeping in mind that gel also benefits from thrombus reinforcement [6]. Whereas for medium-sized arteries and capillary beds such as renal or splenic vessels, G' values in the range of ~800 Pa to ~14 kPa appear adequate [3]. For high blood pressure artery, such as hepatic artery, the final storage modulus of the gel should be more than 0.1 MPa for stable gel embolus formation [3,7]. Therefore, it is essential to select an appropriate STHC system tailored to the patient’s condition. Additionally, complex charge interactions in hydrogel nanocomposites lead to an unknown phase stability. The challenge is exacerbated by the unavailability of any quantitative metric to assess phase stability. However, systematic experimental

exploration of the design space is costly and time-consuming as each polymer is associated with a continuous composition space.

Machine learning-based optimization algorithms such as Bayesian optimization (BO) are particularly effective in finding the optimal design with only a few experimental observations [13,14]. Gaussian process (GP) [15] is a commonly used machine learning method to build surrogate models for material property prediction, but they are not well suited for handling categorical inputs. In hydrogel nanocomposite design, polymer type is a categorical variable with each polymer representing a distinct category. To handle such problems Zhang et al. [16] developed Latent variable Gaussian Process (LVGP) which maps categorical variables onto a continuous latent space. This allows LVGP to predict properties across different categories along with predicting correlations across categories and provide interpretable insights regarding similarities in observed properties of different categories [17–20]. On the other hand, in order to identify feasible designs, machine learning-based classification models are used where the unknown feasibility space of a material is modeled using Gaussian process classification [21,22].

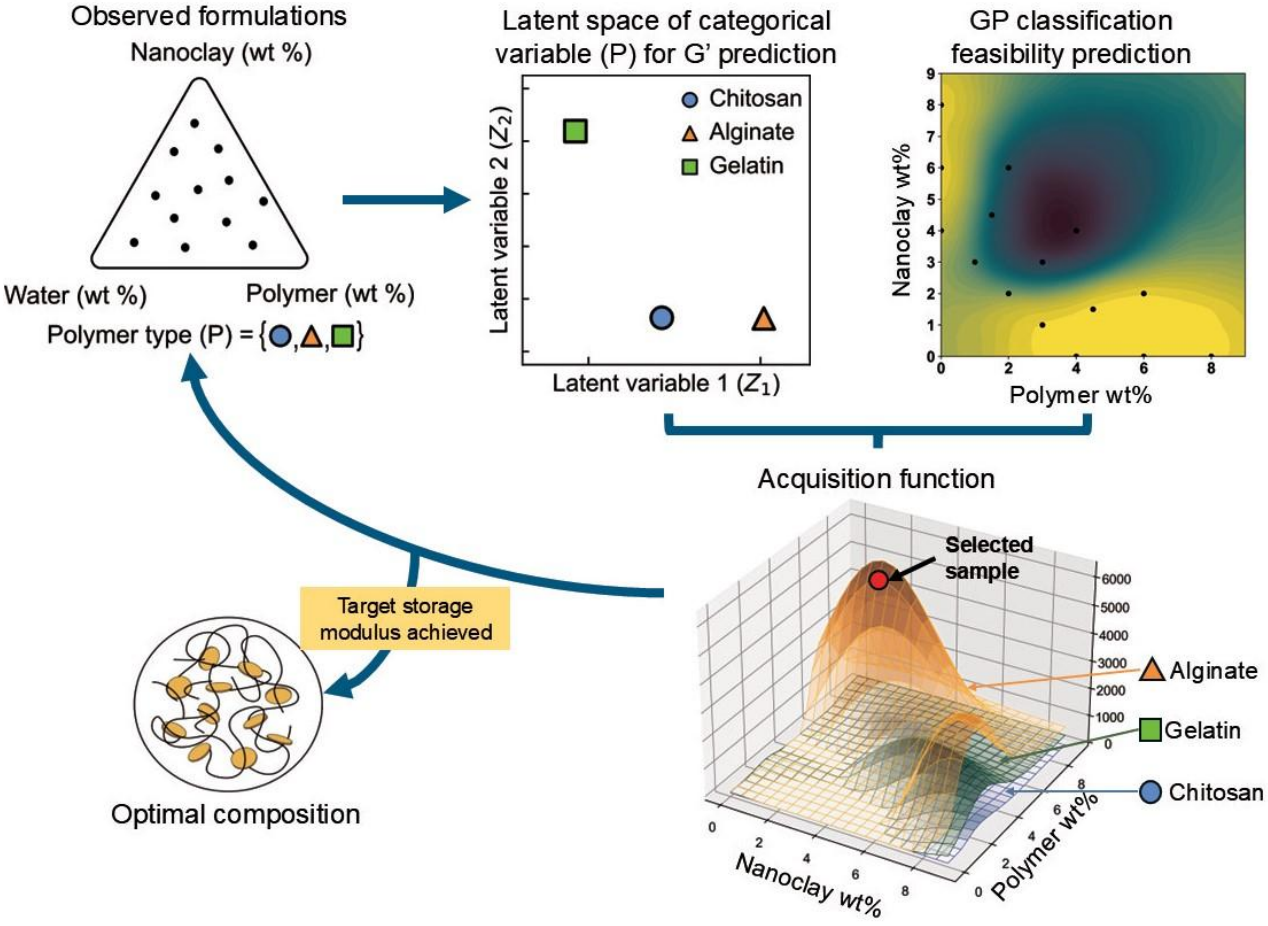

Figure 1. Machine learning-guided Bayesian optimization. Starting from an initial dataset of formulations with three different polymers (chitosan, alginate, and gelatin), a Latent Variable Gaussian Process (LVGP) model maps each material system on a two-dimensional latent space and predicts storage modulus (G') of all materials using a single model. A Gaussian Process classifier predicts feasibility which is quantified by probability of gel formation. An acquisition function (expected improvement) combines LVGP predictions and classification prediction to guide iterative sampling across all materials, toward the target modulus.

As shown in Figure 1, we employ a latent-variable Gaussian process (LVGP) to jointly model categorical (polymer type) and continuous (composition) design variables, enabling rapid prediction of storage modulus across hydrogel nanocomposites. Coupled with GP classification for phase stability, this surrogate is integrated within a Bayesian optimization (BO) framework to iteratively identify optimal and feasible formulations targeting ~30 kPa. Compared to conventional GP-based BO that requires separate models for each polymer, the LVGP-BO approach achieves the target using only 3 new experiments on top of a 38-sample initial dataset versus 9 new experiments used by GP-BO with same initial dataset, corresponding to a 67% reduction. By learning shared structure across polymers, the model captures common trends in composition–property relationships, while the learned latent space provides interpretable insights into similarities between material systems. Beyond identifying optimal formulations, this framework highlights the potential of machine learning–guided experimental design for efficiently exploring mixed-variable material spaces, with direct relevance to the design of shear-thinning hydrogel nanocomposites for biomedical applications such as endovascular embolization.

2. Results and discussion

2.1 Experimental characterization and data collection

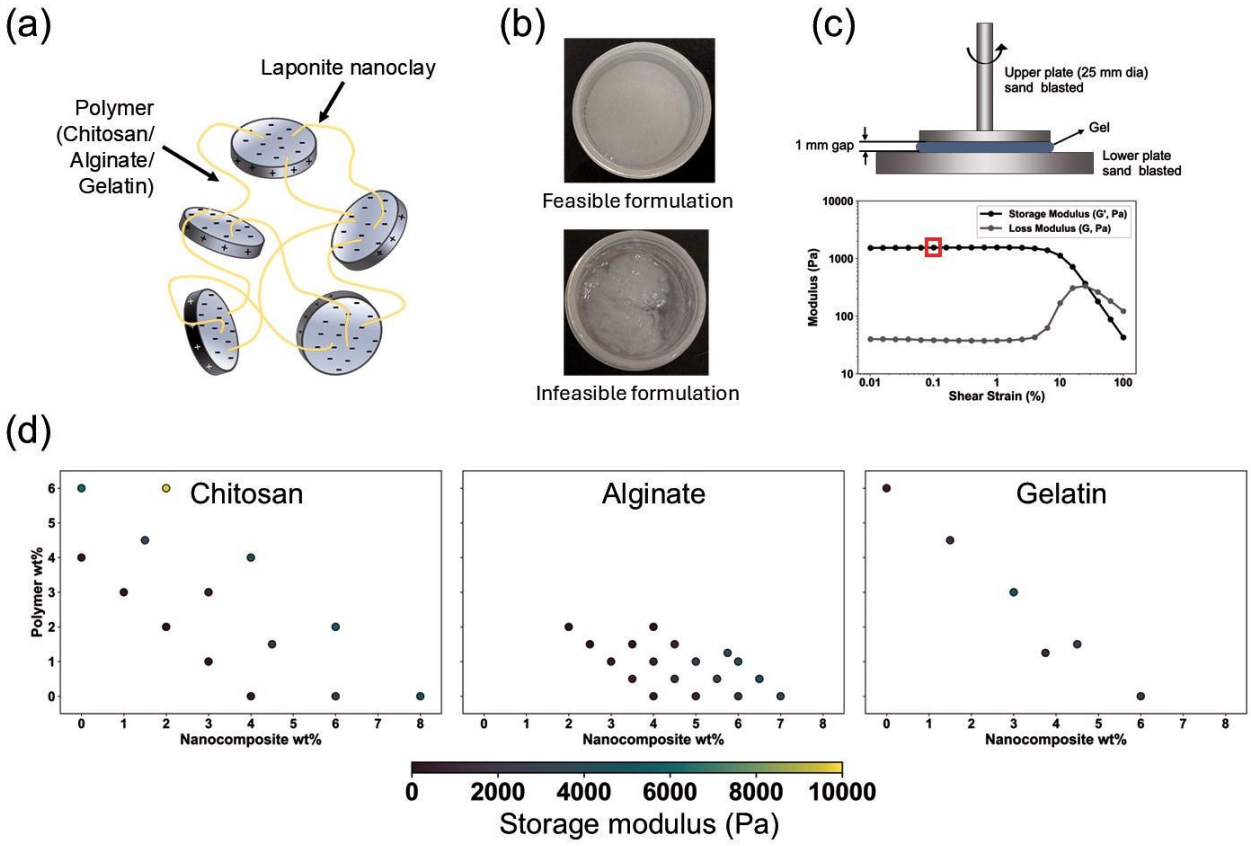

Figure 2. Material preparation and initial data. (a) Schematic of network structure formed by polymer and nanoclay. The anisotropic charge distribution of NC particles is represented with negatively charged surface and positively charged rims. (b) Infeasible formulations are identified based on homogeneity of the prepared gel. (c) Rheology of feasible gels is characterized by

measuring storage and loss modulus. Storage modulus at 0.1% shear strain (highlighted by red box in bottom figure) is treated as objective function of designing new formulations. (d) Initial dataset prepared with varying weight fractions of polymer and nanoclay. Three different polymers are used – chitosan, alginate, and gelatin. The number of samples and range of composition of gels in each polymer is different.

Figure *2* shows the data preparation method and the collected initial data points. An initial dataset comprising 38 gel formulations is constructed based on prior studies [1,2,23,24]. Gels are formulated by mixing nanoclay and polymer solutions which form network structure due to electrostatic interactions. Some formulations of gelatin-based gel exhibited inhomogeneous or phase-separated gels, as shown in Figure 2 (b). Amplitude sweep tests are performed on all gels to measure their storage (elastic) modulus (G') and loss (viscous) modulus (G"). The storage modulus G'quantifies the elastic energy stored in the material, whereas the loss modulus G" represents the viscous component, corresponding to the energy dissipated during deformation. The storage modulus value at 0.1% shear strain, which falls under the linear viscoelastic (LVE) region, is treated as the quantity of interest. We aim to find formulations with high storage modulus at this shear strain. LVE region is the low shear strain region where the G' of the gels are almost constant. The formulations in the initial dataset span three polymers, chitosan, alginate, and gelatin, with 14 chitosan-based gels, 18 alginate-based gels, and 6 gelatin-based gels. The weight fractions of polymer and nanoclay are varied from 0 to 9%, while ensuring that all formulations remain homogeneous and do not undergo phase separation. The initial dataset is used to train machine learning surrogate models.

The G′ of STHCs is strongly influenced by the choice of polymer system. Chitosan-based formulations consistently exhibit overall higher G′ values compared to alginate- and gelatin-based counterparts. For instance, a formulation containing 4.5 wt% nanoclay (NC) and 1.5 wt% polymer showed a significantly higher G′ for NC-chitosan gels (1614.40 ± 90.41 Pa) than for NC-alginate gels (801.32 ± 30.49 Pa). Similarly, at 1.5 wt% NC and 4.5 wt% polymer, NC-chitosan gels (3370.07 ± 375.77 Pa) exhibit higher G' compared to NC-gelatin gels (1297.05 ± 60.13 Pa). These differences arise from the nature and extent of electrostatic interactions between NC and the respective polymers.

NC consists of disc-shaped synthetic silicate nanoparticles (20-30 nm in diameter and ~1 nm in thickness) with anisotropic charge distribution with negatively charged faces and positively charged rims (**Figure *2*** (a)). In aqueous media, NC forms a dynamic "house-of-cards" structure through face–rim interactions, which can reversibly break and reform under shear, imparting shear-thinning behavior. Polymers such as chitosan and alginate interact electrostatically with NC, forming physically crosslinked networks. Chitosan, a positively charged polymer due to its protonated amine groups ($-NH_3^+$), primarily interacts with the negatively charged faces of NC, whereas alginate which carries negatively charged carboxylate groups ($-COO^-$) interacts with the positively charged rims of NC. The substantially greater available surface area associated with the faces relative to the rim likely accounts for the higher storage moduli observed in chitosan-based STHCs compared to their alginate-based counterparts. However, alginate and chitosan gels with low polymer concentration show similar values of storage moduli, indicating that formulations with reduced polymer content may provide similar mechanical stability independent of polymer type. This indicates that both polymer type and concentration can be tuned to achieve desired mechanical properties.

Gelatin-based systems introduce additional complexity due to their amphoteric nature. Type A gelatin is positively charged under neutral to acidic conditions (isoelectric point ~7-9) and becomes negatively charged at higher pH. As a result, its interaction with NC depends on environmental conditions, leading to

variable network structures and, in some cases, phase separation at specific NC-to-gelatin ratios. Further investigation is required to fully elucidate these behaviors.

For effective embolization, the G′ of STHCs must be sufficient to withstand physiological blood pressures, preventing fragmentation and recanalization. Reported values indicate that gels with G′ ~200 Pa can occlude low-pressure peripheral veins, whereas G′ on the order of 0.1 MPa is required for high-pressure vessels such as hepatic arteries [2,3,10]. Importantly, the mechanical properties of NC-polymer STHCs can be precisely tuned by adjusting total solid content and NC-to-polymer ratios. This tunability enables the design of application-specific embolic materials, offering greater flexibility and versatility compared to conventional embolic agents.

2.2 Predictive modeling

The LVGP model is trained on the complete initial dataset combining all three materials. Model performance is evaluated using leave-one-out cross-validation (LOOCV), and generalizability is quantified using the R^2metric. The LVGP model demonstrates strong predictive performance for alginate and chitosan, with R^2values of 0.90 and 0.95, respectively (see **Error! Reference source not found.**). In contrast, the limited number of initial samples for gelatin adversely affects the model's predictive accuracy. For comparison, standard Gaussian process (GP) models are also trained on the initial dataset. Unlike LVGP, a separate GP model is fitted independently for each material. Further details of model

validation studies are provided in the supplementary information (see **Error! Reference source not found.**).

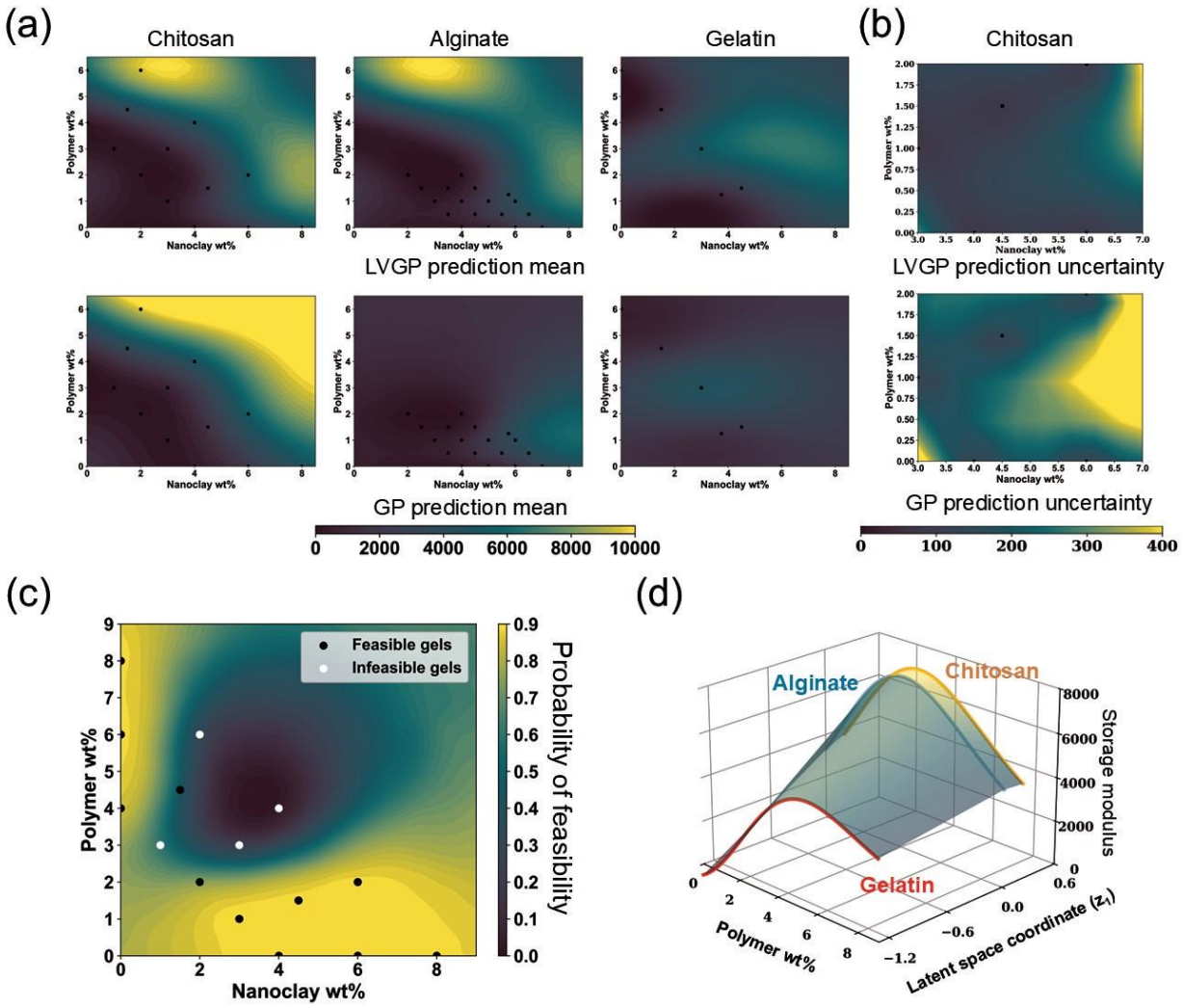

Figure 3. Prediction model trained on initial dataset. (a) Predicted trends in storage moduli differ between LVGP (top row) and GP (bottom row) surrogate model, especially in Chitosan and alginate predictions. (b) Model prediction uncertainty for chitosan hydrogels in LVGP and GP models. LVGP shows lower prediction uncertainty than GP. (c) Feasibility prediction contours for gelatin-based gels using GP classification model. (d) LVGP prediction at constant nanoclay weight fraction (5 wt%) to show learnt distributions for three materials over latent space.

Figure 3(a) compares the mean predictions of the storage modulus obtained using the GP and LVGP models. The LVGP model, which combines datasets from all three materials into a single model, exhibits similar prediction contours for chitosan and alginate. This arises because LVGP learns two types of correlations: (i) correlations between polymer concentration and Nanoclay concentration within each material, and (ii) correlations across different materials to capture polymer-specific effects (e.g., chitosan, alginate, and gelatin). In the initial dataset, chitosan and alginate show similar storage moduli in the low polymer concentration regime. As a result, the LVGP model learns a strong correlation between these two materials, leading to similar prediction contours. The strength of correlation is based on the proximity between two polymers in LVGP's latent space (see Methods). Figure 3(d) illustrates this by showing the similarity in predictive distributions and latent space coordinates for alginate and chitosan at a fixed

nanoclay weight fraction. Additionally, the learnt correlations across materials lead to lower prediction uncertainty in one material if observations are taken in another material (Figure 3(b)). In contrast, this behavior is not observed in the GP predictions, since separate GP models are trained for each material and therefore cannot capture correlations across materials.

To predict feasibility of the gel formulations Dirichlet-based GP classification model [28] is used. Here we only use the classification model to predict the feasibility of gelatin-based gels. Figure 3(C) shows the predicted probability of feasibility for gelatin-based gels. Gels with POF higher than 0.7 are considered feasible. The performance of the classification model is evaluated using Leave-one-out cross-validation where the model shows F1 score of 0.7. Further details of the model validation are provided in the supplementary information (**Error! Reference source not found.**).

For Chitosan-based gels, feasibility is ensured by constraining the total solid content (polymer wt% + nanoclay wt%) to be less than 9%. In contrast, for alginate-based gels the feasible region is bounded by a linear constraint of the form $y = 4 - \frac{2}{15}x$, implying that the allowable nanoclay content decreases proportionally with increasing polymer content. Due to the simplicity of the feasibility constraint for chitosan and alginate-based gels, a predictive model is not necessary.

2.3 LVGP captures the similarity between Alginate and Chitosan

The latent space mapping of categorical variable (polymer type) allows LVGP to learn cross-material correlations. We compare learned kernel parameters of LVGP and regular GP to further understand the latent mapping and draw interpretable insights about similarities and differences in property trends of different polymers.

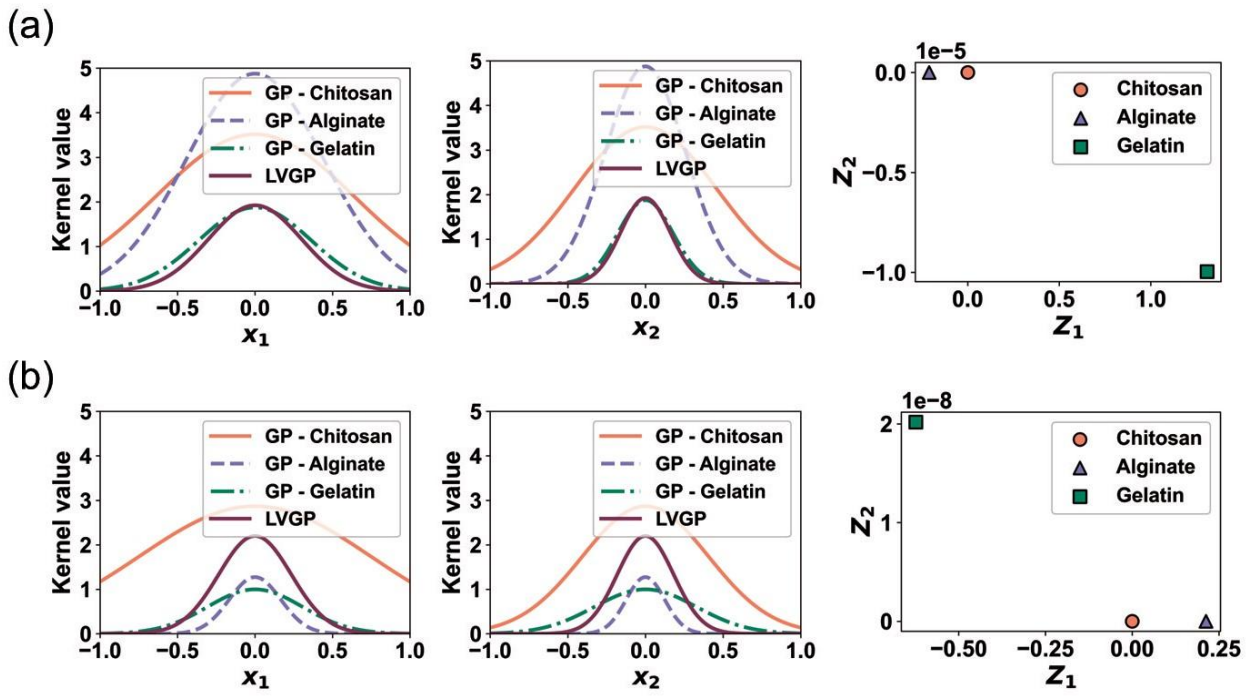

Figure 4. Learnt kernels and latent space coordinates. Kernels learnt by GP and LVGP (a) before and (b) after BO sampling over both the inputs x_1 (polymer wt%) and x_2 (nanoclay wt%). Using LVGP, only one kernel is used for all three materials. The LVGP kernel has smaller length scales and has smaller peaks than GP kernels. Third column shows latent space learnt by LVGP

where each polymer material is assigned a point in two-dimensional continuous latent space based on relative similarities between observed values.

Figure 4 shows the Radial Basis Function (RBF) kernels learnt by each model. The GP kernels are fit using individual datasets of each material and therefore have different length scales and output scales for chitosan, alginate, and gelatin. The length scale in a GP kernel determines the sharpness of spatial correlations. Smaller length scales indicate sharp and highly nonlinear trends in property values. In hydrogel nanocomposite application, spatial correlations correspond to similarity in storage moduli of gel formulations that are similar each other. Before collecting new samples using Bayesian optimization, GP kernels show relatively longer length scales and larger output scales (peak value of kernel) for chitosan compared to GP kernels for alginate and gelatin. This indicates that the GP model predicts a relatively smooth response surface over chitosan formulations. However, LVGP kernel shows sharper length scales which indicate prediction of sharp nonlinear trends. This is also observed after collecting the BO samples. This behavior is attributed to the fact that LVGP learns kernel parameters using data set of all three material instead of a single material and therefore it tries to capture trends which are observed across different materials, thereby learning a smaller length scale which can model sharp nonlinearities.

Cross-material correlations are learned using the latent space coordinates in LVGP. Latent space coordinates are learnt during the model fitting process along with kernel parameters. These coordinates are optimized such that the trend similarities between data points across different materials can be modeled using an exponential kernel in latent space (See Methods). We observe proximity of chitosan and alginate in latent space (Figure 4), indicating similarity in their material property trends, which is in line with observations in initial dataset which show similarity in the storage moduli trends of these two materials in the low polymer concentration regime (Figure 2). These similarities can be attributed to the complex interplay between charge on the polymer and surface area of the NC particles.

2.4 Adaptive Sampling using Bayesian optimization

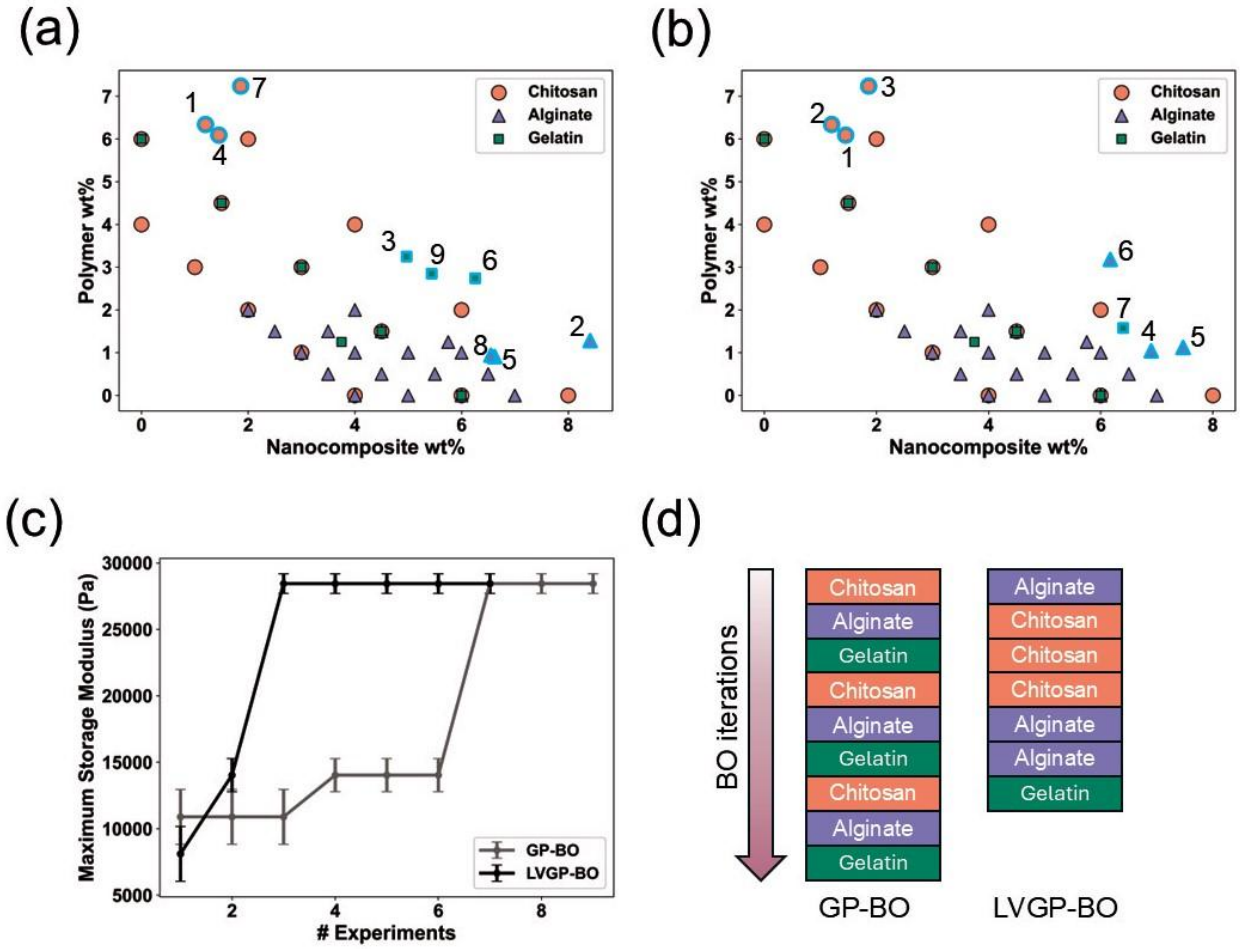

Figure 5. Sequential sampling using Bayesian optimization. (a-b) Gel formulations before and after collecting new samples using GP-BO and LVGP-BO, respectively. (c) Comparison of sampling performance of GP-based BO and LVGP-based BO.(d) Sequence of polymer selected by sampling processes using GP-BO and LVGP-BO algorithm, respectively.

Bayesian optimization is used to identify optimal gel formulation that leads to storage modulus in the range of 30000 Pa. The initial dataset consists of formulations with maximum storage modulus of around 8000 Pa. The iterative algorithm of Bayesian optimization includes fitting the predictive model, calculating acquisition function values and then selecting formulation with highest acquisition value. Here, two algorithms are used- one with Gaussian Process models as the predictive model (GP-BO) and other with LVGP as the predictive model (LVGP-BO).

For GP-BO, the new candidate formulation is selected based on maximum value of acquisition function which is calculated separately for each material using weight fraction of polymer and nanoclay as the inputs. Therefore, one GP-BO iteration includes evaluating three gels, one for each polymer, leading to three experiments per iteration. On the other hand, in LVGP-BO, there is only one acquisition function which considers the polymer as well as the weight percentage of polymer and nanoclay as the input and provide the acquisition value. Therefore, one LVGP-BO iteration includes evaluation of only one gel

leading to one experiment in each iteration. Performance of the two algorithms is compared in terms of number of experiments required to reach the desired storage modulus of 30,000 Pa.

Figure 5 shows the distribution of formulations before and after collecting new samples. As shown in **Figure *5***a, sampling with GP-BO leads to total 9 experiments within 3 iterations. Since each GP only models one material, a single GP-BO iteration samples one formulation from each material, leading to 3 experiments per iteration. Furthermore, due to lack of knowledge about cross-material correlations, GP-BO was observed to exploit already explored formulations as seen by the proximity of selected alginate-based formulations.

With LVGP-BO, total 7 new experiments are carried out (**Figure *5***b). Although, the target storage modulus of 30000 Pa is reached within only 3 iterations (i.e. 3 experiments) as shown in **Figure *5***c, we continue to run LVGP-BO for another 4 iterations such that at least one new formulation in each polymer system is evaluated. The primary reason for speed-up observed in finding the optimal formulation with LVGP-BO is the learnt correlation in latent space which allows the model to quickly identify high performing chitosan-based formulations. **Figure *5***d shows the sequence of formulations sampled by both methods.

The key difference between GP and LVGP-based sampling comes from the ability of LVGP to build a single model for all three materials. This allows one sample after evaluating formulations from three materials. The latent space of LVGP which allows for learning cross-material correlations evolves as more samples are collected during BO iterations. The underlying distribution of storage moduli dictates the distance between different materials in the latent space. These distances control the strength of correlations across materials and allow LVGP to learn similarities between material responses. The learnt correlation across materials is helpful here as it allows LVGP to share trends from alginate to chitosan and avoid sampling low polymer concentration regions of chitosan and reach optimal formulation faster.

Conclusion

In this work, we demonstrated Bayesian optimization-based design approach for shear-thinning hydrogel nanocomposites using latent variable Gaussian process (LVGP)-based Bayesian optimization under mixed-variable design space and unknown feasibility constraints. By jointly modeling categorical (polymer type) and continuous (composition) variables within a single surrogate model, LVGP learns correlation across materials and reduces experimental burden. Compared to conventional GP-based Bayesian optimization, the LVGP-BO achieved the target storage modulus with fewer experiments, illustrating the benefits of cross-material correlation learning.

Beyond hydrogel nanocomposite design, the proposed approach is broadly applicable to materials systems involving discrete material choices combined with continuous processing or compositional variables. Examples include alloy design with discrete phase selections, polymer blends with categorical monomer families, composite systems with different filler chemistries, and biomaterials with multiple scaffold types. In many such systems, unknown feasibility constraints such as phase instability, manufacturability limits, or biological compatibility further complicate experimental exploration. The integration of GP-based probabilistic classification for feasibility with LVGP-based regression provides a general strategy for design under uncertainty.

More broadly, this work highlights how interpretable latent space coordinates provided by LVGP can reveal underlying similarities across material classes, offering not only predictive acceleration but also mechanistic insights. By embedding materials into a learned continuous latent space, the framework provides a data-driven representation of material similarity that may guide future multi-material optimization and transfer learning strategies in materials science.

Together, our results demonstrate that mixed-variable Bayesian optimization frameworks can serve as powerful tools for accelerating materials discovery in experimentally constrained settings, particularly when categorical material choices and feasibility uncertainties coexist.

Methods

Experimental synthesis

Materials

Laponite-XLG nanoclay (NC) was purchased from BYK USA Inc., Texas, USA (Lot No: 0002303785). Chitosan nanofibers (Sugino Machine Ltd, Japan), medium viscosity alginic acid sodium salt derived from brown algae (Sigma-Aldrich, St. Louis, MO, USA), gelatin type A from porcine skin (Sigma, St. Louis, MO, USA), molecular grade water (Intermountain Life Sciences, Utah, USA) were used for engineering the gels.

Gel preparation

NC-polymer gels were engineered by mixing a pre-determined weight ratios of NC, polymer (either gelatin, alginate or gelatin) and molecular grade water using a specialized high-speed mixer (DAC 330-100 SE, FlackTek Speed Mixer, USA) at 3000 rpm for 5 minutes and repeated for three times. The list of engineered gels is given in SI.

Rheology

All rheological studies were conducted at 25 °C using a rheometer (MCR 302e, Anton Paar, Austria). A sandblasted 25 mm upper plate and a sandblasted lower plate were used, with a 1 mm gap in between for all tests. A solvent trap was used to maintain humidity and prevent the gels from drying. Amplitude sweep test were carried out with from 0.01% to 100% shear strain to assess the linear viscoelastic (LVE) region of gel, as well as its storage modulus (G') and loss modulus (G"). The angular frequency was kept at 10 rad/s. The test was conducted three times for each gel and the average G' at 0.1% shear strain (which falls under the LVE region) was calculated.

Predictive modelling and design

Gaussian Process (GP) regression

Gaussian process (GP) regression is a non-parametric Bayesian modeling approach that defines a prior distribution over functions, characterized by a mean function and a covariance (kernel) function [15].

Assuming the output values (y) are observed Gaussian aleatoric noise ϵ zero mean and σ_n^2 variance, we can write the input-output relationship as follows.

$$y = f(x) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_n^2)$$

Given observed input–output pairs, the GP posterior is obtained by conditioning Gaussian prior on data -

$$f(x) \sim GP(m(x), k(x, x'))$$

Here $m(x)$ denotes the mean function and $k(x, x')$ denotes the kernel function. In this work, we use a zero-mean function $m(x) = 0$ and the Radial Basis Function (RBF) kernel $k(x, x') = \exp\left(-\frac{||x - x'||^2}{2l^2}\right)$ where $||.||$ denotes Euclidean norm. The kernel function encodes assumptions about smoothness and similarity.

The GP posterior provides a predictive posterior distribution of the output values at a new unobserved point x_{new} –

$$y_{new} | x_{new}, y_{old}, x_{old} \sim N(\hat{y}_{new}, cov(y_{new}))$$

As written above, the posterior predictive distribution is a normal distribution with mean $\hat{y}_{new}$ and posterior covariance $cov(y_{new})$ which are given as follows –

$$\hat{y}_{new} = m(x_{new}) + K(x_{new}, x_{old})(K(x_{old}, x_{old}) + \sigma_n^2 I) y_{old}$$

$$cov(y_{new}) = K(x_{new}, x_{new}) - K(x_{new}, x_{old})(K(x_{old}, x_{old}) + \sigma_n^2 I) K(x_{old}, x_{new})$$

For n observed points and n^*unobsreved points, $K(x_{new}, x_{old})$ denotes a $n^* \times n$ matrix of covariances between all pairs of observed and unobserved points. The kernel function $k(x, x')$ is used for computing the covariances. Other matrices $K(x_{new}, x_{new}), K(x_{old}, x_{old})$, and $K(x_{old}, x_{new})$ are built in the same way.

Latent Variable Gaussian Process (LVGP)

The RBF kernel function of GPs requires a distance metric in input space. Such a distance metric is not available when inputs are categorical in nature. To model categorical variables using GP, Latent Variable Gaussian Process (LVGP) model was proposed [16]. LVGP maps discrete-valued categorical inputs onto a continuous valued latent space where distance metric between categorical inputs can be defined.

Let each design x consist of q continuous variables x_{con} and m categorical variables $x_{cat} = [t_1, t_2, \dots, t_m]$, where $t_j \in \{1, 2, \dots, l_j\}$. LVGP maps each level of t_j to a continuous valued latent vector $z_j(t_j) \in \mathbb{R}^2$. Therefore, the full design space is $x_{map} = [x_{con}, z] \in \mathbb{R}^{q+2m}$.

LVGP uses following kernel to compute correlation between two designs –

$$k(x, x') = \exp\left(-\frac{||x_{con} - x'_{con}||^2}{2 \times l^2} - \sum_{j=1}^{m} ||z_j(t_j) - z_j(t'_j)||^2 \right)$$

Now, LVGP can provide predictive posterior distribution for unobserved design x_{new} in the same way as Gaussian Process regression. **Figure *6*** shows the workflow involved in converting categorical variables to latent space and obtaining the predictive distributions for multiple categories with LVGP model.

We note that the LVGP kernel can be seen as a product of two RBF kernels. The first part is RBF kernel for continuous design variables (x_{con}), whereas the second part the RBF over categorical design variables which are mapped to continuous latent variables (z_j).

$$k(x,x') = \exp\left(-\frac{\left||x_{con}-x'_{con}||\right|^2}{2\times l^2} - \sum_{j=1}^{m}\left||z_j(t_j)-z_j(t'_j)\right|\right|^2\Bigg)$$
$$= \exp\left(-\frac{\left||x_{con}-x'_{con}||\right|^2}{2\times l^2}\right)\times \exp\left(-\sum_{j=1}^{m}\left||z_j(t_j)-z_j(t'_j)\right|\right|^2\Bigg)$$
$$= k_1(x_{con},x'_{con})\times k_2\big(z(t),z(t')\big)$$

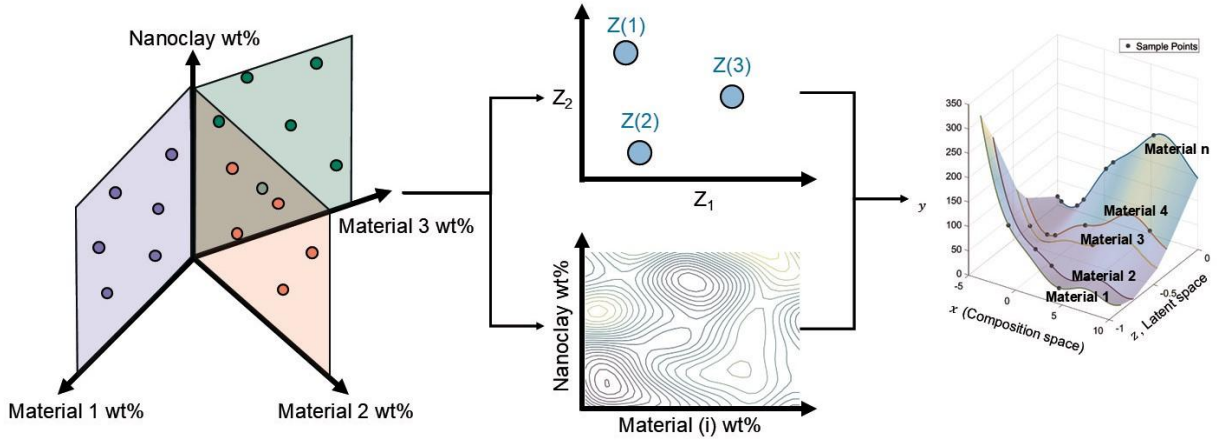

Figure 6. Latent Variable Gaussian Process (LVGP). LVGP takes data points from different categories as inputs and fits a Radial Basis function over the quantitative and qualitative variables. It learns the hyperparameters of the RBF kernel over quantitative variables as well as the latent space coordinates of categorical variables. The prediction is made using a product kernel of the two RBFs.

Gaussian process classification

Gaussian process classification (GPC) assumes a latent function $f_c(x)$, drawn from a GP prior, that maps the inputs to real-valued logits, which are converted to class probabilities via a link function like the sigmoid or SoftMax [15]. Calculating the posterior distribution of the logits is a computationally expensive step due to non-Gaussian nature of its likelihood. Dirichlet-based Gaussian process classification (GPD) proposed by Milios et al. [25] reduces the computational cost of posterior calculation. This method interprets classification labels as realizations of a Dirichlet-distribution, $\pi \sim \mathrm{Dir}(\alpha)$, where π is the class probability vector and α are hyperparameters. For a label y belonging to class k, the corresponding concentration parameters are set as: $\alpha_k = 1+\alpha_\varepsilon$, $\alpha_i = \alpha_\varepsilon$ for all $i \neq k$, with α_ε being a small positive constant ensuring valid Dirichlet parameters. These Dirichlet-distributed labels are then transformed into a latent Gaussian space via a log-normal approximation, allowing Gaussian process regression to efficiently approximate the transformed labels: $y_i \sim \mathrm{N}(f_i(x), \sigma^2_i)$, with $\sigma^2_i = \log(1 + 1/\alpha_i)$. We use Dirichlet-based GP classification to model feasibility and compute the probability of feasibility as the posterior probability of the feasible class.

Bayesian optimization

Bayesian optimization is a sequential global optimization algorithm which is quite frequently used for optimization of functions whose full form is not known but only a few observations are available and function evaluations are expensive [13,14]. Bayesian optimization includes two components – a surrogate model to predict the objective function and an acquisition function which estimates the value of observing a new design by balancing exploration of design space to build better surrogate model or exploitation of region where surrogate model provides accurate predictions.

Gaussian process is often employed as the surrogate model due to its efficient data handling and uncertainty quantification capabilities [15]. The acquisition function is defined using the predictive posterior of GP.

Expected Improvement (EI) is a well-known acquisition function which assigns acquisition value to a design x based on the expected value of improvement in objective function if the function were to be evaluated at x. The improvement is defined as $I(x) = \hat{y}(x) - y^*$ where $\hat{y}_{new}$ is the predictive mean and y^* is the best value of objective function among available observations. The acquisition function is defined as –

$$EI(x) = \mathbb{E}[I(x)] = \int_{-\infty}^{\infty} I(z)\phi(z)dz$$

Where $\phi(z)$ is the probability density of a standard normal distribution. Under Gaussian prior, expected improvement function has analytical form –

$$EI(x) = (\hat{y}_{new} - y^*)\phi\left(\frac{(\hat{y}_{new} - y^*)}{cov(y_{new})}\right) + cov(y_{new}) \times \phi\left(\frac{(\hat{y}_{new} - y^*)}{cov(y_{new})}\right)$$

Where $cov(y_{new})$ is the predictive posterior variance at design x.

The above acquisition function assigns high value to a design when its predictive posterior mean is higher than the best objective value so far, prioritizing exploitation. It also assigns high value to designs where the predictive posterior variance is large prioritizing exploration. To promote selection of feasible designs, this acquisition function is further modified by multiplying it with probability of feasibility associated with a new design –

$$EI_c(x) = EI(x) \times pof(x)$$

Where $EI_c(x)$ stands for constrained Expected Improvement [26] where probability of feasibility is the output of a GP classification model.

A gradient-based optimizer is used to select the design with highest acquisition value under the modified acquisition function for next function observation. The new observation is then added to the training data of GP model for retraining, and the BO iterations are performed again till a pre-defined stopping criterion is reached. In this work, the stopping criterion is based on target value of the objective function viz. storage modulus.

Data availability

The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files. Other relevant data are available from the corresponding author upon reasonable request.

Code availability

The codes that were used to train the machine learning model and perform Bayesian optimization in this study are available from the corresponding author upon reasonable request.

Author contributions

W.C. and J.H. conceived the project. G.V. and K.Z. designed and performed the experiments. G.K. developed the machine learning and sampling code and performed data analysis. G.K. and G.V. wrote the paper. S.K., W.C., and J.H. provided feedback. All authors reviewed the paper.

Competing interests

The authors declare no competing interests.

Acknowledgement

The G.K., S.K., and W.C. acknowledge NSF MRSEC award DMR-2308691 for funding support. G.V., K.Z., and J.H. acknowledge NIH R03, ORAU grant and NCSU for funding support.

References

[1] G. V. P.j., K. Zhao, P. Chen, and J. Hu, Chitosan-nanoclay embolic material for catheter-directed arterial embolization, J. Biomed. Mater. Res. A **112**, 914 (2024).

[2] K. Zhao, G. Varghese P J, P. Chen, and J. Hu, Developing a transcatheter injectable nanoclay-alginate gel for minimally invasive procedures, J. Mech. Behav. Biomed. Mater. **152**, 106448 (2024).

[3] J. Hu, H. Albadawi, B. W. Chong, A. R. Deipolyi, R. A. Sheth, A. Khademhosseini, and R. Oklu, Advances in Biomaterials and Technologies for Vascular Embolization, Adv. Mater. **31**, 1901071 (2019).

[4] J. Hu, H. Albadawi, Z. Zhang, M. A. Salomao, S. Gunduz, S. Rehman, L. D'Amone, J. L. Mayer, F. Omenetto, and R. Oklu, Silk Embolic Material for Catheter-Directed Endovascular Drug Delivery, Adv. Mater. **34**, 2106865 (2022).

[5] A. Fatimi, F. Zehtabi, and S. Lerouge, Optimization and characterization of injectable chitosan-iodixanol-based hydrogels for the embolization of blood vessels, J. Biomed. Mater. Res. B Appl. Biomater. **104**, 1551 (2016).

[6] H. Albadawi, I. Altun, J. Hu, Z. Zhang, A. Panda, H. Kim, A. Khademhosseini, and R. Oklu, Nanocomposite Hydrogel with Tantalum Microparticles for Rapid Endovascular Hemostasis, Adv. Sci. **8**, 2003327 (2021).
[7] A. Poursaid, M. M. Jensen, I. Nourbakhsh, M. Weisenberger, J. W. Hellgeth, S. Sampath, J. Cappello, and H. Ghandehari, Silk-Elastinlike Protein Polymer Liquid Chemoembolic for Localized Release of Doxorubicin and Sorafenib, Mol. Pharm. **13**, 2736 (2016).
[8] J. Hu, I. Altun, Z. Zhang, H. Albadawi, M. A. Salomao, J. L. Mayer, L. P. M. P. Hemachandra, S. Rehman, and R. Oklu, Bioactive-Tissue-Derived Nanocomposite Hydrogel for Permanent Arterial Embolization and Enhanced Vascular Healing, Adv. Mater. **32**, 2002611 (2020).
[9] D. Yang, Recent Advances in Hydrogels, Chem. Mater. **34**, 1987 (2022).
[10] J. M. Dodda, K. Deshmukh, D. Bezuidenhout, and Y.-C. Yeh, *Hydrogels: Definition, History, Classifications, Formation, Constitutive Characteristics, and Applications*, in *Multicomponent Hydrogels*, edited by J. M. Dodda, K. Deshmukh, and D. Bezuidenhout, 1st ed. (The Royal Society of Chemistry, 2023), pp. 1–25.
[11] S. Guleria, L. Chopra, and Manikanika, Temperature responsive hydrogels for biomedical applications, Mater. Today Proc. **92**, 356 (2023).
[12] S. M. Mantooth, A. M. Hancock, P. M. Thompson, G. Varghese P J, D. M. Meritet, M. R. Vrabel, J. Hu, and D. A. Zaharoff, Characterization of an Injectable Chitosan Hydrogel for the Tunable, Localized Delivery of Immunotherapeutics, ACS Biomater. Sci. Eng. **10**, 905 (2024).
[13] P. I. Frazier and J. Wang, *Bayesian Optimization for Materials Design*, in *Information Science for Materials Discovery and Design*, edited by T. Lookman, F. J. Alexander, and K. Rajan (Springer International Publishing, Cham, 2016), pp. 45–75.
[14] Y. Zhang, D. W. Apley, and W. Chen, Bayesian Optimization for Materials Design with Mixed Quantitative and Qualitative Variables, Sci. Rep. **10**, 4924 (2020).
[15] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, 2005).
[16] Y. Zhang, S. Tao, W. Chen, and D. W. Apley, A Latent Variable Approach to Gaussian Process Modeling with Qualitative and Quantitative Factors, Technometrics **62**, 291 (2020).
[17] Y.-P. Chen, L. Wang, Y. Comlek, and W. Chen, A Latent Variable Approach for Non-Hierarchical Multi-Fidelity Adaptive Sampling, Comput. Methods Appl. Mech. Eng. **421**, 116773 (2024).
[18] Y. Comlek, S. K. Ravi, P. Pandita, S. Ghosh, L. Wang, and W. Chen, Heterogeneous Multi-Source Data Fusion Through Input Mapping and Latent Variable Gaussian Process, J. Mech. Des. **147**, (2025).
[19] S. K. Ravi et al., Interpretable multi-source data fusion through Latent Variable Gaussian Process, Eng. Appl. Artif. Intell. **145**, 110033 (2025).
[20] L. Wang, S. Yerramilli, A. Iyer, D. Apley, P. Zhu, and W. Chen, Scalable Gaussian Processes for Data-Driven Design Using Big Data With Categorical Factors, J. Mech. Des. **144**, (2021).
[21] R. Hickman, M. Aldeghi, and A. Aspuru-Guzik, Anubis: Bayesian Optimization with Unknown Feasibility Constraints for Scientific Experimentation, preprint, Chemistry, 2023.
[22] G. Kumbhojkar, Z. Wang, S. Keten, and W. Chen, Multi-objective Bayesian Optimization For Design Under Unknown Feasibility Constraints, J. Mech. Des. 1 (2026).
[23] K. Zhao, P. Chen, Z. Wang, G. Varghese P.J., J. Liu, and J. Hu, A multi-modal embolic gel system for long-term fluorescence imaging and photothermal therapy, Biomater. Adv. **174**, 214298 (2025).
[24] P. J. G. Varghese, P. Chen, K. Zhao, M. Saha, and J. Hu, An Investigation on the Effect of Contrast Agents in the Chitosan-Nanoclay Shear Thinning Hydrogel for Trans-Catheter Arterial Embolization, JOM **76**, 5643 (2024).

[25] D. Milios, R. Camoriano, P. Michiardi, L. Rosasco, and M. Filippone, *Dirichlet-Based Gaussian Processes for Large-Scale Calibrated Classification*, in *Advances in Neural Information Processing Systems*, Vol. 31 (Curran Associates, Inc., 2018).
[26] R. B. Gramacy and H. K. H. Lee, *Optimization Under Unknown Constraints*, arXiv:1004.4027.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI.docx
- SIfigure1.png