
T-METASET: TAILORING PROPERTY BIAS OF LARGE-SCALE METAMATERIAL DATASETS THROUGH ACTIVE LEARNING

Doksoo Lee
 Dept. of Mechanical Engineering
 Northwestern University
 Evanston, IL 60208
 dslee@northwestern.edu

Yu-Chin Chan
 Siemens Corporate Technology
 Princeton, New Jersey 08540
 yu-chin.chan@siemens.com

Wei (Wayne) Chen
 Dept. of Mechanical Engineering
 Northwestern University
 Evanston, IL 60208
 wei.wayne.chen@northwestern.edu

Liwei Wang
 School of Mechanical Engineering
 Shanghai Jiao Tong University
 Shanghai, P.R. China, 200240
 iridescence@sjtu.edu.cn

Anton van Beek
 School of Mechanical and Materials Engineering
 University College Dublin
 Belfield, Dublin 4, Ireland, D04 V1W8
 anton.vanbeek@ucd.ie

Wei Chen*
 Dept. of Mechanical Engineering
 Northwestern University
 Evanston, IL 60208
 weichen@northwestern.edu

ABSTRACT

Inspired by the recent achievements of machine learning in diverse domains, data-driven metamaterials design has emerged as a compelling paradigm that can unlock the potential of multiscale architectures. The model-centric research trend, however, lacks principled frameworks dedicated to data acquisition, whose quality propagates into the downstream tasks. Often built by naive space-filling design in shape descriptor space, metamaterial datasets suffer from property distributions that are either highly imbalanced or at odds with design tasks of interest. To this end, we present t-METASET: an active-learning-based data acquisition framework aiming to guide both diverse and task-aware data generation. Distinctly, we seek a solution to a commonplace yet frequently overlooked scenario at early stages of data-driven design of metamaterials: when a massive ($\sim O(10^4)$) shape-only library has been prepared with no properties evaluated. The key idea is to harness a data-driven shape descriptor learned from generative models, fit a sparse regressor as a start-up agent, and leverage metrics related to diversity to drive data acquisition to areas that help designers fulfill design goals. We validate the proposed framework in three deployment cases, which encompass general use, task-specific use, and tailororable use. Two large-scale mechanical metamaterial datasets are used to demonstrate the efficacy. Applicable to general image-based design representations, t-METASET could boost future advancements in data-driven design.

Keywords Data acquisition · Data-driven design · Active learning · Variational autoencoder · Gaussian processes · Determinantal Point Processes · Metamaterials

1 Introduction

Metamaterials are artificially architected materials that support unusual properties from their structure rather than composition [1]. The recent advancements of computing power and manufacturing have fueled research on metama-

*Corresponding author: Wei Chen (weichen@northwestern.edu)

terials, including theoretical analysis, computational design, and experimental validation. Over the last two decades, outstanding properties and functionalities achievable by metamaterials have been reported from a variety of fields, such as optical [2], acoustic [3], thermal [4], and mechanical [5]. They have been widely deployed to applications in communications, aerospace, biomedical, and defense, to name a few [6]. From a design point of view, leveraging the rich designability in hierarchical systems is key to further disseminating metamaterials as a versatile material platform, which not only realizes superior functionalities but also facilitates customization and miniaturization. There has been growing demand for advanced design methods to harness the potential of metamaterials.

Data-driven metamaterials design (DDMD) offers a route to intelligently design metamaterials. In general, the approach builds on three main steps: data acquisition, model construction, and inference for design purposes. DDMD typically starts with a precomputed dataset that includes a large number of structure-property pairs [7, 8, 9, 10, 11]. Machine learning model construction follows to learn the underlying mapping from structure to property, and sometimes vice versa. Then the data-driven model is used for design optimization, such as at the “building block” or unit cell level, and optionally tiling in the macroscale as well when aperiodic designs are of interest [12, 13, 14, 15]. The key distinctions of DDMD against conventional approaches are that (i) DDMD accelerates multiscale design optimization via exploring the vast design space efficiently; (ii) it has little restrictions on analytical formulations of design interest; and (iii) some DDMD approaches enable on-demand design without iterations, which pays off the initial cost of data acquisition and model construction. Capitalizing on the advantages, DDMD has reported a plethora of achievements for diverse design problems in recent years [1, 8, 9, 10, 16, 17, 18].

Despite the recent surge of DDMD, rare attention has been given to data acquisition and data quality assessment – the very first step of DDMD. In data-driven design, *data is a design element*; a collection of data points forms a landscape to be learned by a model, which is an “abstraction” of the data, and to be explored by either model inference or modern optimization methods. Hence data quality ends up propagating into the subsequent stages. Yet the downstream impact of naive data acquisition is opaque to diagnose and thus challenging to prevent *a priori* [19]. Underestimating the risk, common practice in DDMD typically resorts to a large number of space-filling designs in the shape space spanned by the shape parameters. This *inevitably* hosts imbalance – distributional bias of data – in the property space [12, 20, 11, 21] formed by the property components. The downstream tasks involving a data-driven model – training, validation, and deployment to design – follow mostly without rigorous assessment on data quality in terms of diversity, design quality, and feasibility, among others. The practice overlooks not only *data imbalance itself* but also *the compounding ramifications* at the design stage, allowing both to impede solid deployment of DDMD.

To this end, Chan et al. presented METASET [11] as a subset selection framework that can identify small yet diverse subsets from a fully evaluated database. Key idea is to evaluate the properties of *all* the designs *a priori*, then downsample a balanced subset based on diversity metrics. Yet the approach lacks generality of data acquisition for DDMD in that: (i) design evaluation could be prohibitively expensive to build a massive ($\sim O(10^4)$) database with all the data evaluated; (ii) diversity alone does not offer data customization for specific design tasks.

To enhance the generality and efficiency of data acquisition for DDMD, we propose *task-aware* METASET (t-METASET) with special attention to starting with sparse observations. Herein, “task-aware” approaches rate individual data points based on the utility for a given *specific* design scenario, rather than on distributional metrics (e.g., diversity) for *general* use. The proposed framework handles data bias reduction (for generic use) and design quality (for particular use) simultaneously, by leveraging diversity and quality as the sampling criteria, respectively. We advocate that (i) building a good dataset should be an *iterative* procedure [22, 23]; (ii) diversity sampling [24] can efficiently suppress the property bias of multi-dimensional regression involved in most DDMD methods [11]; (iii) property bias control significantly improves fully aperiodic metamaterial designs, as shown by recent reports [12, 13, 14, 15]. Distinct from existing work, however, we primarily seek a solution to a commonplace – yet frequently overlooked – scenario that designers face during data preparation: a large-scale shape dataset has been generated and is about to be observed *without evaluated observations at the beginning*.

Our t-METASET incrementally “grows” a high-quality dataset that is not only diverse but also task-aware. Figure 1 illustrates a schematic of the t-METASET procedure. The central ideas are (i) to extract a compact shape descriptor from a shape-only dataset by unsupervised representation learning, (ii) to sequentially update a sparse regressor as a start-up “agent” under *sparse* observations, and (iii) to intelligently curate samples based on the prediction of the regressor, and batch sequential sampling [24] building on shape diversity, *estimated* property diversity, and user-defined quality. Starting from a massive library of building blocks, the active learning framework maneuvers the data acquisition so that it can tailor the data distribution based on both diversity (for generic use) and quality (for specific use) for given tasks.

In the context of DDMD, the intellectual contributions of t-METASET are three-fold:

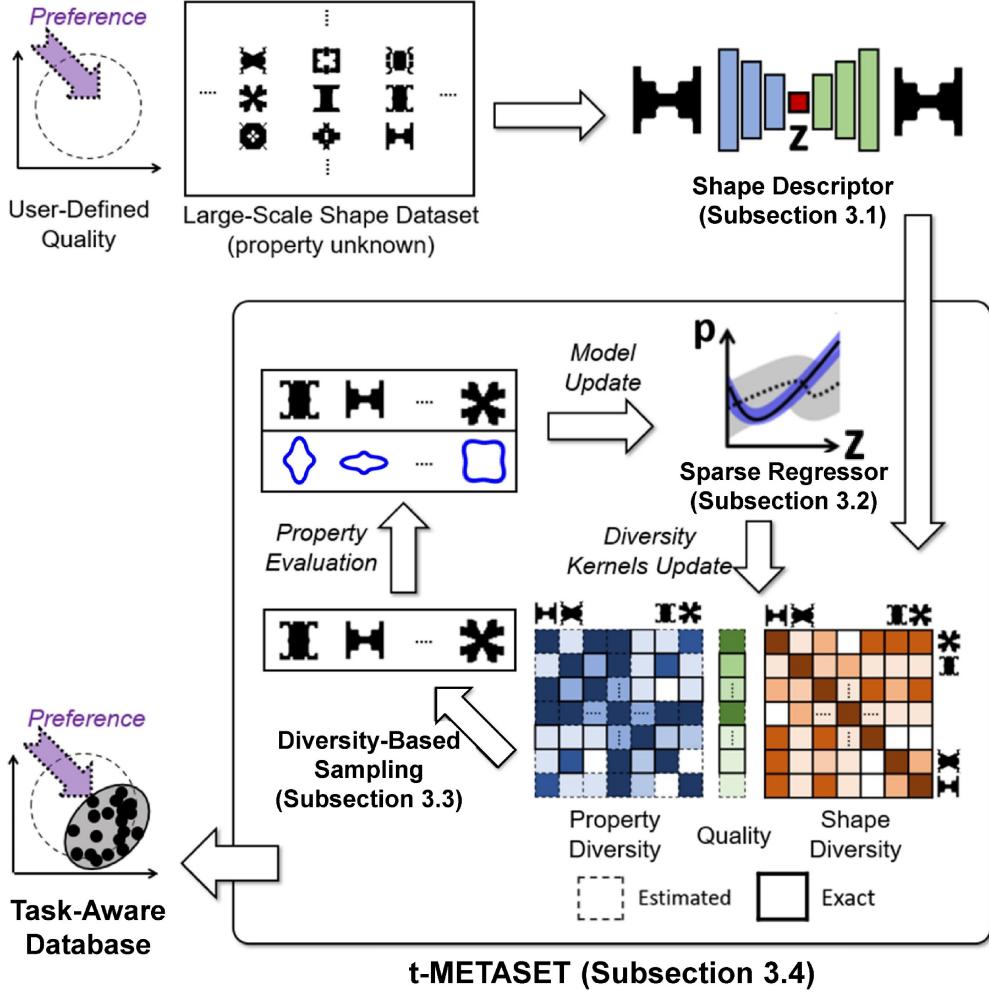


Figure 1: An overview of t-METASET. Given a shape-only dataset, a compact shape descriptor of microstructures is distilled by a generative model (e.g., variational autoencoder). A sparse regressor (e.g., a Gaussian process) learns the descriptor-property mapping in light of sparse observations. Harnessing the prediction over unseen shapes, diversity sampling (e.g., Determinantal Point Processes) drives active learning through diversity and, optionally, task-related quality metrics. Once evaluated, the observed batch refines the regressor. By repeating this routine, t-METASET suppresses undesirable distributional bias while boosting desirable one.

- Starting without evaluated designs, t-METASET offers a principled framework on how to build a diverse dataset during data acquisition with rigorous metrics and a small amount of heuristics;
- The framework provides a solution to *property bias* that both existing and newly created metamaterial datasets are prone to;
- The proposed t-METASET can produce *task-aware* datasets whose distributional characteristics can be tailored in response to user-defined design tasks, while securing shape and property diversity along the way.

We argue the advantages of t-METASET are: (i) scalability, (ii) modularity, (iii) customizability to general or specific tasks, (iv) freedom from restrictions on shape generation schemes, (v) no dependency on domain knowledge and, by extension, (vi) applicability over generic design datasets involving high-dimensional images. t-METASET is validated via two large-scale shape-only mechanical metamaterial datasets (containing 88,180 and 57,000 instances, respectively) that are built from different ideas, without preliminary downsampling. The validation involves three scenarios addressed by different sampling criteria: (i) only diversity aiming at general use (e.g., global metamodeling [25, 26]), (ii) quality-weighted diversity aiming at task-aware use, and (iii) shape-property joint diversity for tailororable use.

2 Property Bias: An Example of Lattice Mechanical Metamaterials

Property bias prevails in existing metamaterial datasets. To convey this point, we examine an example of a lattice-based 2-D mechanical metamaterial dataset. Lattice-based metamaterials have been intensely studied due to their outstanding performance-to-mass ratio, great heat dissipation, and negative Possion's ratio [1]. Wang et al. devised a lattice-based dataset [12], to be called \mathcal{D}_{lat} in this work. In the dataset, a unit cell (i.e., microstructure or building block) takes six bars aligned in different directions as its geometric primitives (see Figure 2(a)). All unit cells can be fully specified by four parameters associated with the thickness of each bar group. The shape generation scheme produces diverse geometric classes (i.e., baseline, family, motif, basis, and template), as displayed in Figure 2(b). Each class exhibits different topological features, which offer diverse modulus surfaces of homogenized elastic constants ($C_{11}, C_{12}, C_{13}, C_{22}, C_{23}, C_{33}$) (Figure 2(c)). Figure 2(d) shows the nearly uniform sampling in the parametric shape space $\Omega_w = [0, 1]^4$ used for data population. We removed repeated instances where the entire domain is either solid ($v_f = 1$) or void ($v_f = 0$); this explains why some regions in Figure 2(d) have no data points.

Now we look into the data distribution of \mathcal{D}_{lat} . The near-uniform sampling ensures good uniformity in the parametric shape space (Figure 2(d)). On the other hand, the corresponding property distributions in Figure 2(e) show considerable imbalance, which epitomizes that *data balance in parametric shape space does not ensure the same in property space*. We argue that such property imbalance is prevalent in many metamaterial datasets generated by space-filling design in parametric shape space [20, 11, 21, 27, 28]. We claim that: (i) metamaterial datasets collected based on naive sampling in parametric shape space are subject to substantial property bias [20, 11, 21, 27, 28], and more importantly, (ii) this is highly likely to hold true for datasets with generic design representations – beyond parametric ones – as well [11, 13, 10, 29, 30]. The general statement is, in part, grounded on the near-zero correlation between shape similarity and property similarity in large-scale metamaterial datasets ($\sim O(10^4)$), consisting of microstructures represented as pixel/voxel, observed by Chan et al [11]. Overlooking the significant property imbalance, many methods assume that the subsequent stages of DDMD can accurately learn and perform inference under such strong property imbalance, ignoring the compounding impact of data bias [31].

In addition, diversity alone does not ensure successful deployment of DDMD for design purposes. Imagine a case where a 50k-size dataset with perfect uniformity has been prepared, yet the region associated with a given design task (e.g., high performance-to-mass ratio; high stiffness anisotropy; manufacturability) happens to include a tiny portion of the dataset. This implies, provided a design task has been prescribed, that (i) designers would want to involve the utility of data points for the given task, on top of diversity, during both data acquisition and evaluation; (ii) it could be rather *desirable to promote artificial data bias* towards a certain direction/area associated with the task.

Property bias is inevitable without supervision. Properties – a function of a given shape – are unknown before evaluation. Obtaining their values is the major computational bottleneck [19], not only at the data preparation stage but also in the whole DDMD pipeline. An undesirable yet prevalent case is: one evaluates all the shape samples with time-consuming numerical analysis (e.g., finite element analysis (FEM); wave analysis) and trains a model on the data, only to end up with a property distribution that is severely biased outside where one had planned to deploy the data-driven model. To circumvent such unwanted scenarios, it is warranted to monitor property distributions at early stages and maneuver the sampling process in a supervised manner *during* data acquisition, not after. As a solution, we propose t-METASET, a task-aware data acquisition framework that tailors data distributions upon user-defined design tasks.

3 Proposed Method

In this section, we walk readers through the three components of the proposed t-METASET: shape descriptor (Section 3.1), sparse regressor (Section 3.2), and diversity-driven sampling (Section 3.3). Then the algorithm in its entirety is presented. (Section 3.4)

3.1 Shape Descriptor

To exploit topologically free variations of building block geometries, metamaterials design often involves a high-dimensional geometric space (e.g., 50×50 pixelated 2-D designs equates a 50^2 -D space). Exploring the vast design space is inefficient and not computationally affordable. Instead we wish to reparameterize instances in the ambient space using a compact yet expressive shape descriptor. The shape descriptor captures essential topological features of metamaterial building blocks and offers a low-dimensional design representation with an acceptable compromise of expressiveness.

In the literature of DDMD, shape descriptors of building blocks roughly fall into three categories: physical descriptors, spectral descriptors, and data-driven descriptors. First, physical descriptors represent a geometry based on geometric

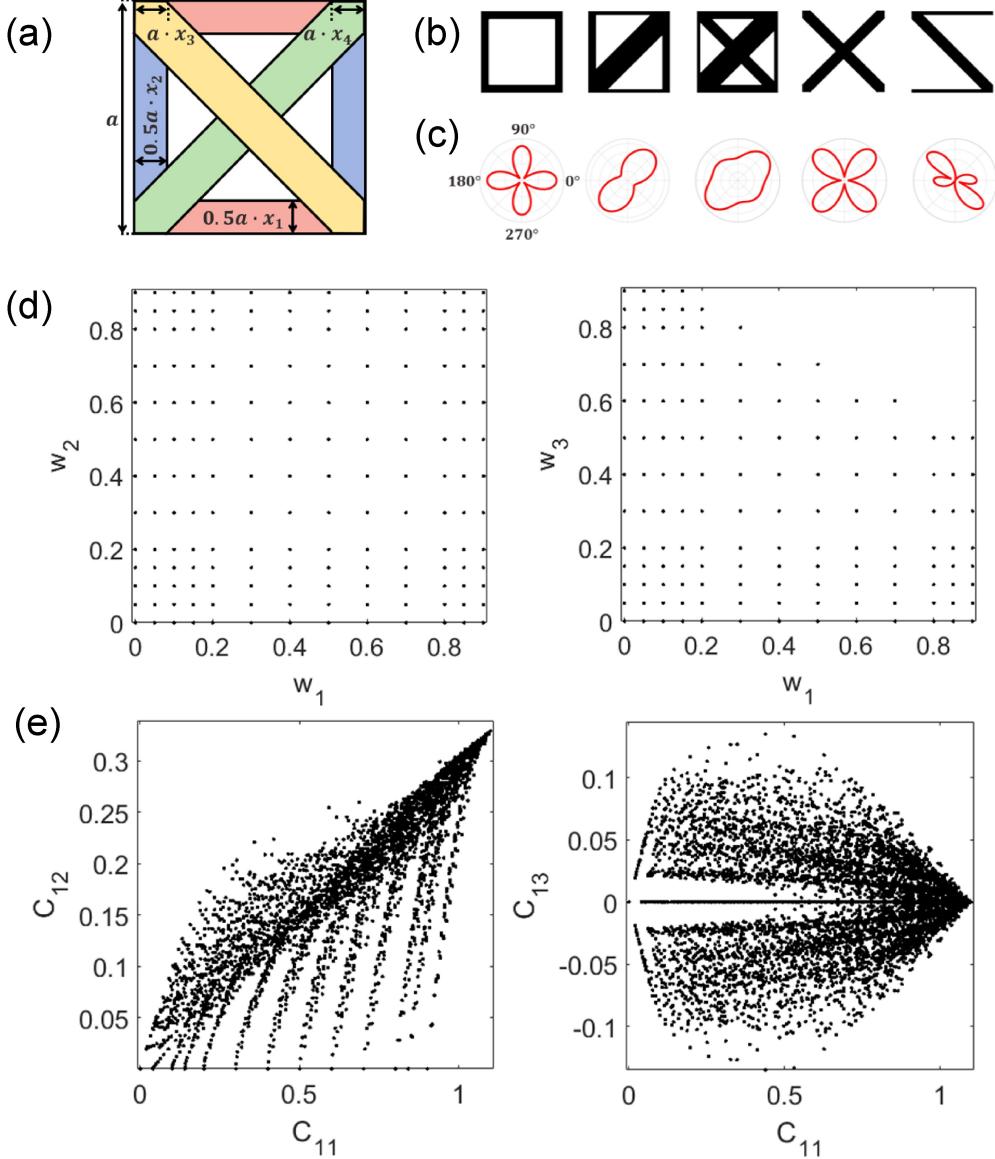


Figure 2: Illustration of \mathcal{D}_{lat} [12]. (a) Microstructure shape representation specific by thickness of each bargroup. (b) Five instances of generated microstructures. (c) The resulting surfaces of homogenized elastic modulus. (d) Data distribution in parametric shape spaces w_1-w_2 and w_1-w_3 . (e) Data distribution in projected property spaces $C_{11}-C_{12}$ and $C_{11}-C_{13}$.

features of interest, such as curvature, moment, angle, shape context [32]. Hence, the key advantage is high interpretability provided by the physical criteria. For example in DDMD, Chan et al. [11] employed the division point-based descriptor [33], which recursively identifies centroids of binary images at several granularity levels, and concatenates the coordinates as the descriptor. Second, spectral descriptors exploit finite-dimensional spectral decomposition of ambient shape space. Liu et al. [34] proposed a Fourier transform based descriptor as a topological encoding method for optical metasurfaces. The spectral descriptor enjoys representational parsimony, reconstruction capability (inverse Fourier transform), efficient symmetry handling, and a continuous latent space. Third, data-driven descriptors exploit data-driven feature engineering. Wang et al. [10] employed a variational autoencoder (VAE) [35] as a deep generative model for DDMD. It was demonstrated that the latent representation offers a compact shape similarity measure in light of given data, facilitates blending across microstructures, and encodes interpretable geometric patterns.

As a data-driven model involving unsupervised representation learning, VAE learns a compact latent representation that can be used as a shape descriptor [35]. We advocate the VAE descriptor as the shape descriptor of metamaterial unit

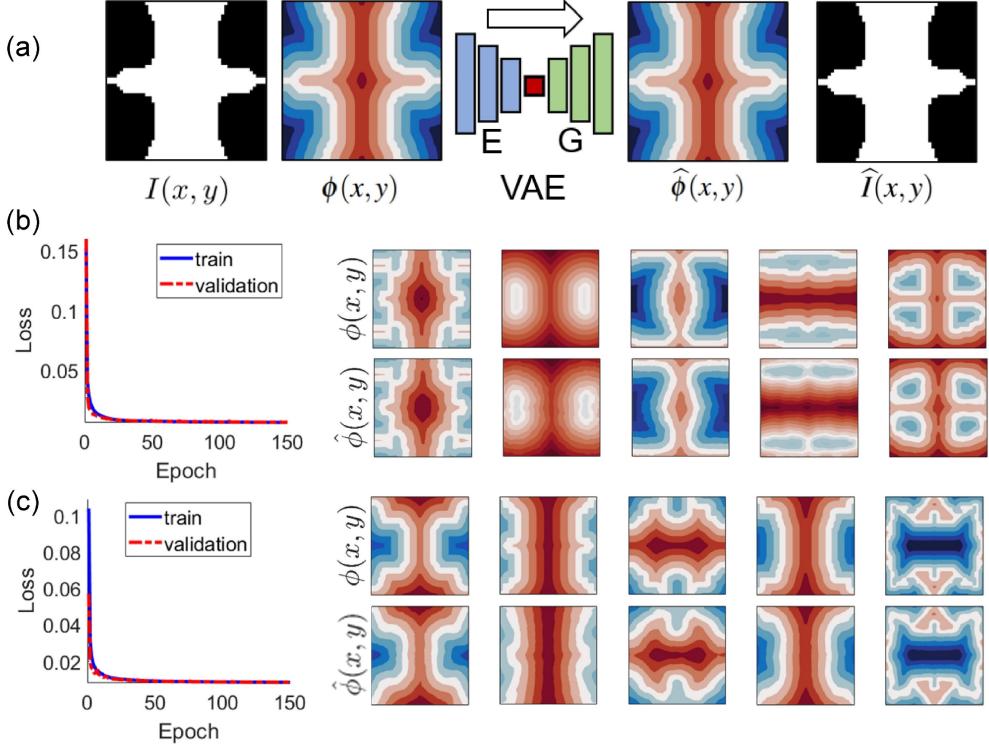


Figure 3: Shape VAE. (a) Schematic of the architecture. (b) The training result of \mathcal{D}_{mix} . (c) The training result of \mathcal{D}_{TO} .

cells based on two aspects. First, VAE enjoys the parsimony of a low-dimensional manifold, which is crucial to make a sparse regressor (Section 3.2) have compact yet expressive predictors, and to expedite the subsequent diversity-driven sampling (Section 3.3). Second, this work also takes advantage of the distributional regularization imposed on the encoder: the latent vectors are enforced to be roughly multivariate Gaussian. The regularization enforces built-in scaling across individual components of the latent representation, rendering diversity-based sampling robust to arbitrary scaling.

Figure 3(a) depicts the shape VAE used in our study. The VAE involves two key components, encoder E and decoder G . Assuming an input instance is given as a discretized image, the encoder involves a set of progressively contracting layers to capture underlying low-dimensional features, until it reaches the bottleneck layer, which provides the latent vector as $\mathbf{z} = E(\phi(x, y))$ where $\phi(x, y)$ is the signed distance field (SDF) of a binary microstructure image $I(x, y)$. The decoder, reversely, takes a latent variable from the information bottleneck and generates a reconstructed image as $\hat{\phi}(x, y) = G(\mathbf{z})$. In formulating the shape instances, we prefer the SDF representation to the binary one since (i) SDFs offer richer local information (distance and sign) that unsupervised representation learning can exploit [36], and (ii) the continuous surface-based representation tends to help generative models produce smoother synthesized instances [37].

Now we briefly introduce key formulations of VAE. A VAE assumes that given data have come from an underlying random process specified by a latent variable \mathbf{z} . Each instance ϕ and latent variable \mathbf{z} are viewed a realization of the conditional distribution $p_\theta(\phi|\mathbf{z})$ and prior distribution $p_\theta(\mathbf{z})$, respectively, where θ is the parameters that specify the distributions. The marginal likelihood of a given instance ϕ reads:

$$\log p_\theta(\phi) = KL[q_\psi(\mathbf{z}|\phi)||p_\theta(\mathbf{z}|\phi)] + \mathcal{L}(\theta, \psi; \phi), \quad (1)$$

where $KL[\cdot||\cdot]$ is the Kullback-Leibler divergence, a non-negative distance measure between two distributions; $q_\psi(\mathbf{z}|\phi)$ is the variational posterior that is specified by the parameter ψ and approximates the true posterior $p_\theta(\mathbf{z}|\phi)$ to bypass the intractability of the marginal distribution [35]; and $\mathcal{L}(\cdot)$ is the variational lower bound on the marginal likelihood. Usual practice for training is to rearrange the equation and to maximize the evidence lower bound:

$$\begin{aligned} \mathcal{L}(\theta, \psi; \phi) &= -KL[q_\psi(\mathbf{z}|\phi)||p_\theta(\mathbf{z}|\phi)] \\ &\quad + \mathbb{E}_{q_\psi(\mathbf{z}|\phi)}[\log(p_\theta(\phi|\mathbf{z}))]. \end{aligned} \quad (2)$$

The first right-hand side term $KL[\cdot||\cdot]$ involves the regularization loss that enforces the latent variable \mathbf{z} to be distributed as multivariate Gaussian, while the second term denotes the reconstruction loss. The approximated variational lower

bound allows stochastic gradient decent to be used for end-to-end training of the whole VAE. For efficient training, standard VAE assumes the prior distribution as $p_{\theta}(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the variational posterior as $q_{\psi}(\phi|\mathbf{z}') \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, respectively, where the reparameterization trick [35] involves a stochastic embedding \mathbf{z}' as $\mathbf{z}' = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \epsilon$ with a Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The training reduces to the following optimization problem:

$$\begin{aligned} \min_{\theta, \psi} [-\mathcal{L}(\theta, \psi; \phi)] &= \frac{1}{|\mathcal{S}|} \sum p_{\theta}(\phi|\mathbf{z}') \\ &\quad - \frac{1}{2} \sum [1 + \log(\boldsymbol{\sigma}^2) - \boldsymbol{\sigma}^2 - \boldsymbol{\mu}^2], \end{aligned} \quad (3)$$

where $|\mathcal{S}|$ is the number of shape data.

Figure 3(b) and (c) report the VAE training results of each dataset, 2-D multiclass blending dataset (\mathcal{D}_{mix}) [13] and 2-D topology optimization dataset (\mathcal{D}_{TO}) [38, 39], respectively. A concise description of the datasets can be found in Section 4.1. The VAE architecture was set based on that of Wang et al. [10]. The dimension of the latent space is set as 10, with the trade-off between dimensionality and reconstruction error taken into account. The Adam optimizer [40] was used to train the VAE with the following setting: learning rate 10^{-4} , batch size 128, epochs 150, and dropout probability 0.4. Each shape dataset is split into training set and validation set with the ratio of 80% and 20%, respectively. In Figure 3(b) and (c), each training history shows stable convergence behavior for both training and validation. From the plots of SDF instances on the right side, we qualitatively confirm good agreement between the input instances (top) and their reconstruction (bottom), for both training results.

3.2 Sparse Regressor

In t-METASET, a sparse regressor enables active learning and task-aware distributional control under epistemic uncertainty (i.e., lack of data). In Section 3.2.1 we elaborate on why a Gaussian process (GP) is a good choice as the sparse regressor and introduce key formulations of multi-output GPs. Section 3.2.2 details roughness parameters of a GP and how they are harnessed for sampling mode transition in t-METASET.

3.2.1 Gaussian Processes

We implement a GP regressor as the “agent” of data acquisition in this work. The mission is to learn the underlying structure-property mapping from sparse data, and to pass predictions over unseen shapes as $\hat{\mathbf{p}} = \mathcal{GP}(\mathbf{z})$ to batch sequential sampling. In this study, the GP takes the VAE latent shape descriptor as its input, which offers substantial dimension reduction ($50^2\text{-D} \rightarrow 10\text{-D}$ in this work). We advocate a GP as the sparse agent due to three key advantages: (i) model parsimony congruent with sparse observations at early stages; (ii) decent modeling capacity of nonlinear structure-property regression (i.e., $\mathbf{z} \rightarrow \mathbf{p}$); (iii) roughness parameters as an indicator of model convergence, to be used for sampling mode transition (detailed in Section 3.2.2).

Building on the advantages of the GP, our novel idea on task-aware property bias control is to (i) construct an *estimated* property similarity kernel $\widehat{L_p}$ (Section 3.3.1) from the GP prediction $\hat{\mathbf{p}} = \mathcal{GP}(\mathbf{z})$, as the counterpart of the shape kernel $\widehat{L_z}$, and (ii) employ conditional Determinantal Point Processes (DPP) [24] – a probabilistic approach to diversity modeling – on the estimated property kernel $\widehat{L_p}$ to recursively sample a batch based on the expected property diversity. The property kernel $\widehat{L_p}$ estimates property similarity, *prior to design evaluation*, not only between train-train pairs, but also train-unseen and unseen-unseen ones. In this way, the sampler of t-METASET recommends a batch \mathcal{B} hinging on both estimated property diversity and shape diversity. It is important to note that, at an incipient phase, we do not rely on $\widehat{L_p}$, as the predictive performance of a multivariate multiresponse GP ($\mathbb{R}^{D_z} \rightarrow \mathbb{R}^{D_p}$, where D_p is the property dimensionality) trained on tiny data is not reliable. We determine the turning point – when to start to respect the GP prediction – based on the convergence history of a set of the GP hyperparameters: roughness parameters (i.e., scale parameters).

As a background to the roughness parameters, we introduce key formulations of GPs. A GP is a collection of random variables, any of whose finite subset is distributed as multivariate Gaussian [41]. Given a set of observations, a GP with D_p responses is fully specified by its mean and covariance functions as

$$f \sim \mathcal{GP}(\boldsymbol{\mu}(\mathbf{z}), \text{cov}(\mathbf{z}, \mathbf{z}')), \quad (4)$$

where $\boldsymbol{\mu}(\cdot)$ is the mean function; $\text{cov}(\cdot, \cdot)$ is the covariance function; f is a function viewed as a realization from the underlying distribution. For the multivariate input \mathbf{z} and the multiresponse outputs \mathbf{p} in our study, the covariance function reads: $\text{cov}(\mathbf{z}, \mathbf{z}') = \Sigma \otimes r(\mathbf{z}, \mathbf{z}')$, where Σ is the $D_p \times D_p$ dimensional multiresponse prior variance, \otimes is

the Kronecker product, and $r(\cdot, \cdot)$ is the correlation function. In this work we use the squared exponential correlation function given as

$$r(\mathbf{z}, \mathbf{z}') = \exp(-(z - z')\Theta(z - z')) \quad (5)$$

where $\Theta = \text{diag}(10^\omega)$ and $\omega = [\omega_1, \dots, \omega_{D_z}]^T$ is the vector of roughness parameters [42]. Given a dataset $\mathcal{D} = \{((\mathbf{z}_1, \dots, \mathbf{z}_{D_z}), (\mathbf{p}_1, \dots, \mathbf{p}_{D_p}))\}_{i=1}^n$, a point estimate of the hyperparameters can be found through maximizing the Gaussian likelihood function:

$$[\hat{\beta}, \hat{\Sigma}, \hat{\omega}] = \arg \min_{[\beta, \Sigma, \omega]} \left[\frac{n}{2} \log(\det(\Sigma)) + \frac{1}{2} \log(\det(\mathbf{R})) + \frac{1}{2\sigma^2} (\mathbf{p} - \mathbf{1}\beta)^T \mathbf{R}^{-1} (\mathbf{p} - \mathbf{1}\beta) \right] \quad (6)$$

where $\mathbf{1}$ is an $n \times D_p$ dimensional vector of ones, $\beta = [\beta_1, \dots, \beta_{D_p}]^T$ is $1 \times D_p$ dimensional vector of weights, $\log(\cdot)$ is the natural logarithm, \mathbf{R} is the $n \times n$ correlation matrix with (i, j) -th element R_{ij} given as $r(\mathbf{z}_i, \mathbf{z}_j)$ for $i, j = 1, \dots, n$, and $\det(\cdot)$ is the matrix determinant operation. In the above formulation of the likelihood we have assumed a constant prior mean function as $\mathbf{1}\beta$. More complex basis functions can be used to represent the prior mean (e.g., linear, or quadratic); however, this is not advised as this information is typically not known *a priori*, and is likely to compromise model accuracy when chosen incorrectly.

After approximation of the hyperparameters, the posterior predictive distribution for an unobserved input \mathbf{z}_{new} can be obtained by conditioning the prior distribution on the observed data \mathcal{D} [43]. Specifically, the mean and the covariance of the posterior predictive distribution is given as

$$\begin{aligned} \mu(\mathbf{z}_{new}) &= \hat{\beta} + \mathbf{r}^T(\mathbf{z}_{new}) \mathbf{R}^{-1} (\mathbf{p} - \mathbf{1}\hat{\beta}), \\ cov(\mathbf{z}_{new}) &= \hat{\Sigma}[r(\mathbf{z}_{new}, \mathbf{z}_{new}) - \mathbf{r}^T(\mathbf{z}_{new}) \mathbf{R}^{-1} \mathbf{r}(\mathbf{z}_{new}) + \mathbf{W}^T (\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1}) \mathbf{W}], \end{aligned} \quad (7)$$

where $\mathbf{r}(\mathbf{z}_{new})$ is an $n \times 1$ dimensional vector whose i -th element is given as $r(\mathbf{z}_{new}, \mathbf{z}_i)$ for $i = 1, \dots, n$, $\mathbf{W} = \mathbf{1}' - \mathbf{1}^T \mathbf{R}^{-1} \mathbf{r}(\mathbf{z}_{new})$, and $\mathbf{1}'$ is a $D_p \times 1$ dimensional vector of ones.

3.2.2 Roughness parameters

Informally speaking, roughness parameters $\omega = [\omega_1, \dots, \omega_{D_z}]^T$ dictate fluctuation levels of responses w.r.t. each predictor (each component of \mathbf{z} in our study), in light of given data. Bostanabad et al. [42] used the fluctuations of roughness parameters with Eq. 5 and their estimated variance to qualitatively determine if sufficient samples were collected during GP training. Building on that, we monitor the roughness parameters ω and take the convergence of roughness parameters as a proxy for model convergence. The roughness residual serves as the transition criterion across sampling modes. We define the convergence criterion involving the roughness residual metric Δ as follows:

$$\Delta^{(t+1)} = \sqrt{\frac{1}{D_z} \|\omega^{(t+1)} - \omega^{(t)}\|^2} \leq \tau \quad (t = 1, \dots, t_{max}) \quad (8)$$

where τ is a threshold associated with the sampling mode transition. At an early stage the roughness residual exhibits a “transient” behavior. As a stream of data comes in, the residual converges to zero, implying a mild convergence of the GP. In this work, we set two different values of threshold namely, τ_1 and τ_2 where $\tau_1 > \tau_2$. We assume each convergence criterion is met if the residuals of five consecutive iterations are below the threshold. τ_1 is to identify a mild convergence, indicated by the larger tolerance. Once met, t-METASET initiates Stage II, where estimated property diversity serves as the main sampling criterion. Meanwhile, the smaller threshold τ_2 is used to decide when to stop the GP update: as the size of training data accumulates, the variations of roughness parameters get unnoticeable [42], whereas the computational cost of fitting the GP rapidly increases as $\sim O(|\mathcal{D}^{(t)}|^3)$ due to the inversion of covariance matrix \mathbf{R} . We prioritize speed, at the modest cost of prediction accuracy. Detailed implementation with the other pillars can be found in Section 3.4. When reporting the results of t-METASET, we will include the history of the residuals, in addition to that of diversity metrics.

3.3 Diversity-Based Sampling

In this section, we elaborate on diversity-based batch sequential sampling. It maneuvers the data acquisition, leveraging both the compact shape descriptor distilled by the VAE (Section 3.1) and iteratively refined prediction offered by the GP agent (Section 3.2), from beginning to end of t-METASET. Recalling the mission of t-METASET – task-aware generation of balanced datasets – we advocate DPP-based diversity sampling primarily based on three key advantages: (i) DPPs offer a variety of practical extensions (e.g., cardinality constraint, conditioning) that facilitate the active learning of t-METASET; (ii) The probabilistic modeling from DPP captures the trade-off between diversity and quality; (iii) Importantly, DPPs are flexible in terms of handling distributional characteristics in that most object-driven sampling

approaches [44] support either exploration (diversity of input) or exploitation (quality of output), while DPPs do all the combinations of diversity (input/output) and quality (shape/property/joint) without restrictions.

t-METASET builds on a few extensions of DPPs. Section 3.3.1 provides fundamental concepts related to DPP. Section 3.3.2 introduces conditional DPPs that are key for DPP-based active learning, and brings up the scalability issue of massive similarity kernels. As a workaround, a large-scale kernel approximation scheme is introduced in Section 3.3.3. Section 3.3.4 addresses how to accommodate design quality into DPP, which enables “task-aware” dataset construction.

3.3.1 Similarity and DPP

In general, an instance of interest could be represented as a vector \mathbf{x} . A similarity metric between items i and j can then be quantified as a monotonically decreasing function of the distance in the virtual item space as

$$s_{ij} = T(h(\mathbf{x}_i, \mathbf{x}_j)), \quad (9)$$

where s_{ij} is the pairwise similarity between items i and j , $h(\cdot, \cdot)$ is a distance function, T is a monotonically decreasing transformation (i.e., the larger a distance, the smaller the similarity is). One way to represent all the pairwise similarities of a given set is to construct the $n \times n$ similarity matrix L as $L_{ij} = s_{ij}$, where $n = |L|$ is the set cardinality (i.e., dataset size). The matrix is often called a *similarity kernel* in that it converts a pair of items into a distance measure (or a similarity measure, equivalently). While any combinations of similarity and transformation are supported by the formalism above, usual practice favors transformations that result in positive semi-definite (PSD) kernels for operational convenience, such as matrix decomposition. Following this, we employ Euclidean distance $h(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\|\mathbf{x}_i - \mathbf{x}_j\|^2}$ and the square exponential transformation. The resulting similarity kernel reads:

$$L_{ij} = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_L^2}\right), \quad (10)$$

where σ_L is a length-scale parameter (i.e., bandwidth) that tunes the correlation between items.

DPPs provide an elegant probabilistic modeling that favors a subset comprised of diverse instances [24]. They have been employed for a variety of applications that take advantage of set diversity, such as recommender systems [45], summarization [46], object retrieval [47]. The defining property of DPPs is:

$$p(X = \mathcal{A}) \propto \det(L_{\mathcal{A}}), \quad (11)$$

where $L_{\mathcal{A}}$ is a subset of a ground set $L = L_{\mathcal{V}}$ indexed by \mathcal{A} , and $p(X = \mathcal{A})$ is the probability to sample \mathcal{A} . The property has an intuitive geometric interpretation: $\det(\mathcal{A})$ is associated with the hypervolume spanned by the constituent instances. If the catalog \mathcal{A} includes any pair of items that is almost linearly dependent on each other, the corresponding volume would be nearly zero, making \mathcal{A} unlikely to be selected. De-emphasizing such cases, the DPP-based sampling serves as a subset recommender that favors a subset of diverse items. In this study, we set the batch size k to be constant at $k = 10$ using k -DPP [48] as follows:

$$p(X = \mathcal{B}) = \frac{\det(L_{\mathcal{B}})}{\sum_{|\mathcal{B}'|=k} \det(L_{\mathcal{B}'}),} \quad (12)$$

where $L_{\mathcal{B}}$ denotes a submatrix indexed by the items that constitute a batch, or subset, $\mathcal{B} \in \mathcal{V}$.

3.3.2 Conditional DPP

Our data acquisition grows a dataset using active learning. At each iteration, the similarity kernels should be recursively updated so that sampling a new batch leverages the latest information of all evaluated observations. This enables the sampler to (i) avoid drawing duplicate samples that have been observed, and to (ii) promote samples that are diverse not only within a given batch, but *across* a sequence of batches [49]. In DPP, such a kernel update is supported via conditioning a DPP on the instances observed so far. DPPs are closed under conditioning operations; i.e., a conditional DPP is also a DPP [50, 49, 51]. This implies that DPP-based sampling can be iteratively applied to similarity kernels to achieve across-batch diversity, as well as within-batch diversity [49]. Let \mathcal{B} and \mathcal{V} be the batch and the ground set at the i -th iteration, respectively. Given the DPP kernel $L^{(i)}$ at that iteration, a recursive formula for the conditional kernel $L^{(i+1)}$ reads:

$$L^{(i+1)} = \left(\left(L^{(i)} + I_{\bar{\mathcal{B}}} \right)^{-1}_{\bar{\mathcal{B}}} \right)^{-1} - I, \quad (13)$$

where $\bar{\mathcal{B}} = \mathcal{V} \setminus \mathcal{B}$. Due to the cascaded matrix inversions involving cubic time complexity, the equation does not scale well to the large-scale kernels with instances $\sim O(10^4)$. Furthermore, t-METASET demands at least a few hundreds of conditioning. Even just storing a $88,180^2$ -size similarity kernel for \mathcal{D}_{TO} with double precision takes up about 62 gigabytes. In brief, Eq. 13 is intractable for large-scale similarity kernels of our interest.

3.3.3 Large-Scale Kernel Approximation

To circumvent the scalability issue, we leverage large-scale kernel approximation [52]. Recalling that we have employed the Gaussian similarity kernel (Section 3.3.1), we harness the shift-invariance (i.e., $L(\mathbf{x}, \mathbf{y}) = L(\mathbf{x} - \mathbf{y})$) by implementing random Fourier feature (RFF) [52] as an approximation method. It builds on the Bochner theorem [53], which states that the Fourier transform \mathcal{F} of a properly scaled shift-invariant (i.e., stationary) kernel L is a probability measure $p(f)$ as follows:

$$L(\mathbf{x} - \mathbf{y}) = \int_{\Omega} p(f) \exp(j f' (\mathbf{x} - \mathbf{y})) df, \quad (14)$$

where j is the imaginary unit $\sqrt{-1}$, $p(f) = \mathcal{F}[L(\mathbf{x} - \mathbf{y})]$ is the probability distribution, D_V ($\ll n$) is the feature dimension, and $\mathbf{x}, \mathbf{y} \in \Omega$. By setting $\zeta_f(\mathbf{x}) = \exp(j f' \mathbf{x})$, we recognize that $L(\mathbf{x}, \mathbf{y}) = \mathbb{E}_f[\zeta_f(\mathbf{x}) \zeta_f(\mathbf{y})^*]$, implying that $\zeta_f(\mathbf{x}) \zeta_f(\mathbf{y})^*$ is an unbiased estimate of the kernel to be approximated. The estimate variance is lowered by concatenating D_V ($\ll n$) realizations of $\zeta_f(\mathbf{x})$. For a real-valued Gaussian kernel L , the probability distribution $p(f)$ is also Gaussian, and $\zeta_f(\mathbf{x})$ reduces to cosine. Under all the considerations so far, the $D_V \times n$ RFF becomes:

$$V(\mathbf{x}) = \sqrt{\frac{2}{D_V}} [\cos(f'_1 \mathbf{x} + b_1), \dots, \cos(f'_{D_V} \mathbf{x} + b_{D_V})]^T, \quad (15)$$

where $\{f_1, \dots, f_{D_V}\} \stackrel{iid}{\sim} \mathcal{N}(0, 1)$ and $b_1, \dots, b_{D_V} \stackrel{iid}{\sim} \mathcal{U}[0, 2\pi]$. Given an RFF V , the updated feature V' conditioned on a batch \mathcal{B} has the following closed-form expression [51]:

$$V' = V_{\bar{\mathcal{B}}} Z^{\mathcal{B}} (I - V_{\mathcal{B}}^T (V_{\mathcal{B}} V_{\mathcal{B}}^T)^{-1} V_{\mathcal{B}}), \quad (16)$$

where the true kernel can be estimated via $L \approx V'(V')^T$. Now the matrix inversions become amenable as the time complexity decreases to $O(|\mathcal{B}|^3)$ with $|\mathcal{B}| = k \ll n$.

3.3.4 Quality-Weighted Diversity for Task-Aware Sampling

Lastly, we take into account user-defined quality, in addition to diversity, to construct datasets that are not only balanced but also *task-aware*. This study is dedicated to *pointwise* design quality, where a pointwise $n \times 1$ quality vector $q(\mathbf{z}, \hat{\mathbf{p}})$ associated with a design task serves as an additional weight to a feature V' . The resulting feature $D_V \times n$ matrix V'' reads

$$V'' = \overbrace{[q(\mathbf{z}, \hat{\mathbf{p}}^{(t)}) \dots q(\mathbf{z}, \hat{\mathbf{p}}^{(t)})]}^{D_V}^T \circ V', \quad (17)$$

where \circ denotes the Hadamard product (i.e., elementwise multiplication).

The quality-weighted DPP sampling could seem similar to Bayesian optimization (BO) [44] in that: (i) quality contributes to exploitation given design attributes of interest, whereas diversity supports exploration, and (ii) both use sequential sampling, taking a GP as the surrogate. We highlight their differences as: (i) t-METASET does not take the uncertainty provided by the GP regressor – at least under the current setup – as a sampling criterion; (ii) diversity is the main driver of the sequential DPP sampling, whereas in BO, exploration (diversity) is ultimately a means for exploitation (quality); (iii) t-METASET is primarily driven by *pairwise* DPP kernels, taking a pointwise quality as an option, whereas BO is driven by a *pointwise* acquisition function; (iv) t-METASET handles quality that accommodates distributional attributes of shape, property, and even the combination of them, while for BO no acquisition functions have been proposed that explicitly consider property distribution; (v) t-METASET has more flexibility in terms of tailoring distributional characteristics, while standard BO ends up biasing both shape and property distributions to reach the global optimum of a black-box cost function. Quantitative comparisons between t-METASET and BO would be an interesting topic but is currently beyond the scope of this work, as t-METASET can only downsample out of $|\mathcal{S}|$ *finite* points in the VAE latent space, whereas standard BO takes *infinitely many* continuous inputs into account. The validation would be viable under the following extensions: (i) the decoder of the VAE joins the t-METASET algorithm to generate new shapes $\hat{\phi}(x, y) = G(\mathbf{z})$, not existing in the given shape dataset \mathcal{S} , and (ii) continuous DPP [54] can be employed to recommend diverse samples from a *continuous* landscape, learned from the discrete data points provided by users. This is our future work.

3.4 The t-METASET Algorithm

In this section, we detail how to seamlessly integrate the three main components introduced: (i) the latent shape descriptor from the shape VAE, (ii) a sparse regressor as the start-up agent, and (iii) the batch sequential DPP-based sampling that suppresses undesirable bias while enforcing an intentional one. Visual illustration of t-METASET is

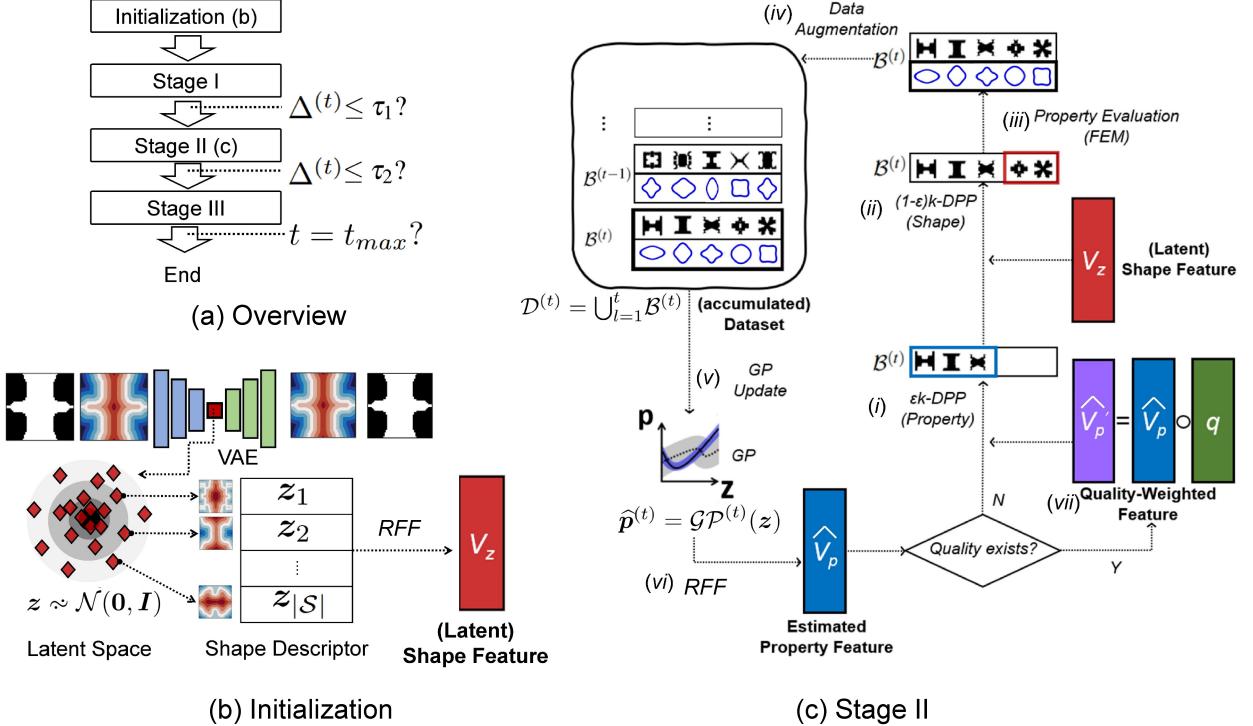


Figure 4: A visual overview of t-METASET. (a) Sampling mode transition associated with the roughness residual $r^{(t)}$. (b) Initialization. The VAE is trained on the given shape-only dataset \mathcal{S} . The latent variables are roughly distributed as multivariate Gaussian. The latent representation is taken as the shape descriptor, whose concatenation forms the $|\mathcal{S}| \times D_z$ matrix, where $|\mathcal{S}|$ is the shape set cardinality ($\sim O(10^4)$). RFF follows to extract a $|\mathcal{S}| \times D_V$ -sized feature of shape feature to be used for the DPP sampling based on shape diversity. (c) A simplified flowchart of Stage II. Details are stated in the main body. Stage I shares the structure as Stage II but is driven only by shape feature (no step (iii)). Stage III is equivalent to Stage II except for the GP update (step (v)). The proposed data acquisition ends when user-defined termination criteria are met (e.g., maximum iteration).

presented in Figure 4. Figure 4(a) shows a flow of t-METASET, whose transition is determined by the roughness residual of the GP agent. Given a shape-only, Figure 4(b) depicts the initialization of t-METASET supported by VAE shape descriptor 3.1 and large-scale kernel approximations 3.3.3. The key sampling procedure of t-METASET is illustrated in Figure 4(c).

3.4.1 Initialization

Figure 4(b) illustrates the initialization of t-METASET, which involves VAE training, latent shape descriptor, and RFF extraction from the descriptor. The framework takes the following input arguments: the shape-only dataset \mathcal{S} comprised of SDF instances $\phi(x, y)$, batch cardinality k , the ratio of property samples in each batch ϵ , and optionally a pointwise quality function $q(z, \hat{p})$ that reflects a design task if declared in advance. A shape VAE is trained on \mathcal{S} with the dimension of latent space D_z , which is 10-D herein (Figure 4(b)).

Then we draw the $D_V \times n$ -sized RFF V_z (15) of the $n \times n$ shape similarity kernel, L_z . This feature is to be recursively updated by conditioning on a series of collected batches. For initialization of conditional DPPs over the shape feature, we follow the procedure of Affandi et al. [49].

3.4.2 Stage I

During Stage I, the GP model’s roughness parameter ω shows large fluctuation due to lack of data. The sampling only relies on shape diversity, because the property prediction of the GP given unseen latent variables is not reliable yet. This stage also can be viewed as initial exploration driven by the pairwise shape dissimilarity – as an analog to initial passive space-filling design – where $|\mathcal{D}| \sim O(10^4)$ discrete data points are given as a pool for sampling.

3.4.3 Stage II

Figure 4(c) provides an overview of Stage II – the core sampling stage of t-METASET. As more data come in, the roughness residual $\Delta^{(t)}$ (Eq. 8) approaches zero and becomes stable. Provided that the roughness residual falls under the first threshold τ_1 for five consecutive iterations, the t-METASET framework assumes that the GP prediction is ready to be appreciated. t-METASET proceeds to the next sampling phase Stage II, where t-METASET harnesses the *estimated* property diversity, in addition to shape diversity, as the main criterion. The key is to introduce the RFF of the *estimated* property \widehat{V}_p , building on the GP prediction $\widehat{p} = \mathcal{GP}(z)$.

Now we detail each step described in Figure 4(c). (i) Given a ratio of property samples ϵ in a given batch, the DPP sampler draws $\epsilon k \in \mathbb{N}$ instances from the property RFF V_p based on property diversity, weighted by task-related quality when a task is specified. (ii) The rest of the batch is filled by $(1 - \epsilon)k$ samples from the shape RFF, to complement possible lack of exploration in the shape descriptor space Ω_z . Herein, the shape RFF must be updated with respect to batch \mathcal{B}_ϵ first, to reflect the latest information. Once sampled, the shape feature is updated again with respect to the rest of the shapes in $\mathcal{B}_{1-\epsilon}$ just selected, for the next iteration. (iii) The microstructures of the batch are observed by design evaluation – FEM with energy-based homogenization [55, 56] in this study – to obtain the true properties (e.g., $p = \{C_{11}, C_{12}, C_{22}\}$). (iv) The true properties replace the GP prediction in the given batch $\mathcal{B}^{(t)}$. (v) Then the evaluated batch updates the GP to refine the property prediction as $\widehat{p}^{(t)} = \mathcal{GP}^{(t)}(z)$ for the next iteration. (vi) The refined prediction demands the update of a new property RFF, as well as the conditioning of it on the entire dataset $\mathcal{D}^{(t)} = \bigcup_{t=1}^{t_{max}} \mathcal{B}^{(t)}$ collected so far. (vii) If a quality function $q(z, \widehat{p})$ over design attributes has been specified, it can be incorporated into the latest property RFF by invoking Eq. 17 to prompt a task-aware dataset.

3.4.4 Stage III

Stage III shares all the settings of Stage II except for the GP update. The main computational overhead of Stage II comes from GP fitting as it involves matrix inversion with the time complexity $\sim O(|\mathcal{D}^{(t)}|^3)$. To bypass the overhead, we stop updating the GP if the roughness residual falls under τ_2 for five consecutive iterations. During Stage III, our algorithm can quickly identify diverse instances from a large-scale dataset ($\sim O(10^4)$), without the scalability issue. The main product of t-METASET is a high-quality dataset $\mathcal{D}^{t_{max}} = \bigcup_{t=1}^{t_{max}} \mathcal{B}^{(t)}$, which is not only diverse but task-aware.

4 Results

In this section, the results of t-METASET are presented. As benchmarks, the two large-scale mechanical metamaterial libraries [13, 39] are used for validation. Data description on the two datasets is provided in Section 4.1. We propose an interpretable diversity metric in Section 4.2 for fair evaluation of t-METASET. To accommodate various end-uses in DDMD, we validate t-METASET under three hypothetical deployment scenarios: (i) *diversity only* for generic use (balanced datasets; Section 4.3), (ii) *quality-weighted diversity* for particular use (task-aware datasets; Section 4.4), and (iii) *joint diversity* for tailorable use (tunable datasets; Section 4.5). Basic settings include: batch cardinality as $k = 10$; property sample ratio during Stage II as $\epsilon = 0.8$; the RFF size as $D_V = 3,000$; maximum iteration as $t_{max} = 500$; first and second threshold of roughness parameters as $\tau_1 = 0.02$ and $\tau_2 = 0.01$, respectively; iteration tolerance of roughness convergence as $i_{tol} = 5$. Lastly, we focus on producing datasets with sizes of either 3,000 or 5,000 (i.e., $t_{max} = 300$ or 500, respectively).

4.1 Datasets

We introduce two mechanical metamaterial datasets, in addition to \mathcal{D}_{lat} , to be used for validating t-METASET: (i) 2-D multiclass blending dataset (\mathcal{D}_{mix}) [13], and (ii) 2-D topology optimization dataset (\mathcal{D}_{TO}) [39]. Table 1 compares key characteristics of the datasets. Figure 5 illustrates each dataset and shape generation heuristic. Note that the purpose of involving the two datasets is to corroborate the versatility of our t-METASET framework, which can accommodate a wide range of datasets born from different methods for different end-uses in a unified way. What we aim to provide is quality assessment of subsets *within* one of the datasets, *not across* them. In addition, while all the datasets in the original references provide the homogenized properties, we assume in all the upcoming numerical experiments that only the shapes are given, *without any property evaluated a priori*. \mathcal{D}_{TO} is publicly available for download at <https://ideal.mech.northwestern.edu/research/software/>.

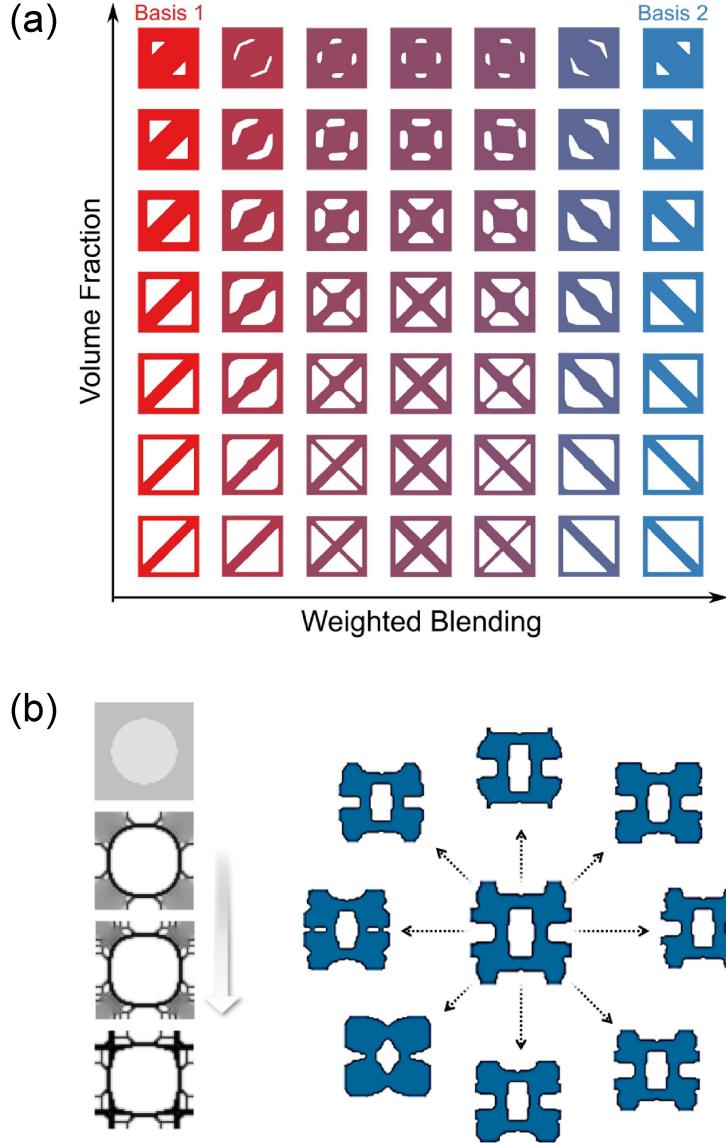


Figure 5: Illustration on shape generation schemes of each dataset. (a) \mathcal{D}_{mix} [13]: example of blending the SDFs of basis shapes and varying their volume fractions to produce new unit cells. (b) \mathcal{D}_{TO} [39]: (left) an example of design evolution by inverse topology optimization with respect to a target property; (right) the stochastic shape perturbation applied to a given microstructure.

4.2 Diversity Metric: Distance Gain

We devise an interpretable diversity metric for assessing the capability of t-METASET against benchmark sampling. In the literature of DDMD, Chan et al. [11] compared the determinant of jointly diverse subsets' similarity kernels against those of *iid* replicates, following the usual practice of reporting set diversity in the DPP literature [24] as the metric to quantify the efficiency of the proposed downsampling. We point out possible issues of using either similarity or determinant for diversity evaluation: (i) similarity values s_{ij} depend on data preprocessing; (ii) a decreasing transformation from distance to similarity $s_{ij} = T(h(\mathbf{x}_i, \mathbf{x}_j))$ for constructing DPP kernels also involves arbitrary scaling, depending on the type of associated transformation T and their tuning parameters (e.g., the bandwidth σ_L of Gaussian kernels in Eq. 10); (iii) the raw values of both similarity and determinant enable the “better or worse” type comparison yet lack intuitive interpretation on “*how much better or worse*”.

Table 1: Dataset Description

	\mathcal{D}_{lat} [12]	\mathcal{D}_{mix} [13]	\mathcal{D}_{TO} [39]
Cardinality	9,882	57,000	88,180
Shape primitive	Bar	SDF of basis unit cell	N/A (used TO)
Shape population	Parametric sweep	Continuous sampling of basis weights & Blending	Stochastic shape perturbation & Iterative sampling
Topological freedom	Predefined	Quasi-free	Free
Property	$\{C_{11}, C_{12}, C_{22}, C_{13}, C_{23}, C_{33}\}$	$\{C_{11}, C_{12}, C_{22}\}$	$\{C_{11}, C_{12}, C_{22}\}$
FEM discretization	100×100	50×50	50×50
FEM solver		Energy-based homogenization [55, 56]	

To this end, we propose a distance-based metric that is more interpretable and less arbitrary. Given a dataset \mathcal{D} , we compute the mean Euclidean distance \bar{d} of pairwise distances of attributes (shape/property) as

$$\bar{d}(\mathcal{D}) = \frac{1}{|\mathcal{D}|^2} \sum_{j=1}^{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} h(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{|\mathcal{D}|^2} \sum_{j=1}^{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \sqrt{\|\mathbf{x}_i - \mathbf{x}_j\|^2}. \quad (18)$$

Intuitively, the larger $\bar{d}(\mathcal{D})$ is, the more diverse \mathcal{D} is. Since this mean metric still depends on data preprocessing, as similarity does, the key idea herein is to normalize $\bar{d}(\mathcal{D})$ with that of an *iid* counterpart $\bar{d}(\mathcal{D}_{iid})$ with the same cardinality $|\mathcal{D}_{iid}| = |\mathcal{D}|$ so that data preprocessing does not affect it. To account for the stochasticity of *iid* realizations, we generate $n_{rep} = 30$ replicates, take the mean of each mean distance, and compute the relative gain h_G as:

$$h_G = \frac{\bar{d}(\mathcal{D})}{\frac{1}{n_{rep}} \sum_{t=1}^{n_{rep}} \bar{d}((\mathcal{D}_{iid})_t)}, \quad (19)$$

where $(\mathcal{D}_{iid})_t$ denotes the t -th *iid* replicate with $|(\mathcal{D}_{iid})_t| = |\mathcal{D}|$. We call the metric *distance gain*, as it *relatively* gauges how much more diverse a given set is compared to a set of *iid* samples. For example, the gain of 1.5 given a property set \mathcal{P} implies that the Euclidean distances between property pairs are 1.5 times larger on average than those of \mathcal{P}_{iid} in the property space. The proposed metric offers an intuitive interpretation based on distance, avoids the dependency on both data scaling and distance-to-similarity transformation, and thus offers a means for consistent diversity evaluation of a given dataset. In addition, the metric generalizes to sequential sampling with $h_G^{(t)}$ at the t -th iteration as well, allowing quantitative assessment across datasets at different iterations (i.e., different sizes). Hence, we report all the upcoming results based on the distance gain proposed.

4.3 Scenario I: Diversity Only

Figure 6 shows the t-METASET results applied to D_{TO} only based on diversity. From Figure 6(a), we observe the evolution of the distance gain as a relative proxy for set diversity at each iteration. At Stage I, the proposed sampling solely relies on shape diversity. The shape gain exceeds unity at the early stage, meaning the exploration by t-METASET shows better shape diversity than that of the *iid* replicates. Meanwhile, the property diversity of t-METASET is even less than the *iid* counterpart; this is another evidence that shape diversity barely contributes to property diversity [11]. During this transient stage, t-METASET keeps monitoring the residual of roughness parameters. Figure 6(b) shows the history up to a few hundred observations; the residuals with little data stay unstable, indicating large fluctuations of the hyperparameters. The mild convergence defined by τ_1 occurs at the 19-th iteration with $10 \times 19 = 190$ observations. This is approximately twice larger than the rule-of-thumb for the initial space-filling design: $D_z \times 10 = 100$ [57]. Rigorous comparison between our pairwise initial exploration and space-filling design (e.g., Latin hypercube sampling [58]) is future work.

Once the first convergence criterion on the GP roughness ω is met, t-METASET starts to respect the GP prediction and, by extension, the RFF of the estimated property DPP kernel as well. During Stage II, shape diversity decreases to less than unity. This implies that pursuing property diversity compromises shape diversity. After about 300 iterations, each gain seems to stabilize with minute fluctuations, and reach a plateau of about 1.3 for property and 0.95 for shape, respectively. Beyond the maximum iteration set as 500, we forecast that the mean of property Euclidean distances – the numerator of property gain – will eventually decrease because: (i) we have finite $|D_{TO}| = 88,180$ shapes to sample from; (ii) the property gamut $\partial\Omega_p^{(t)}$ at the t -th iteration incrementally grows yet ultimately converges to the

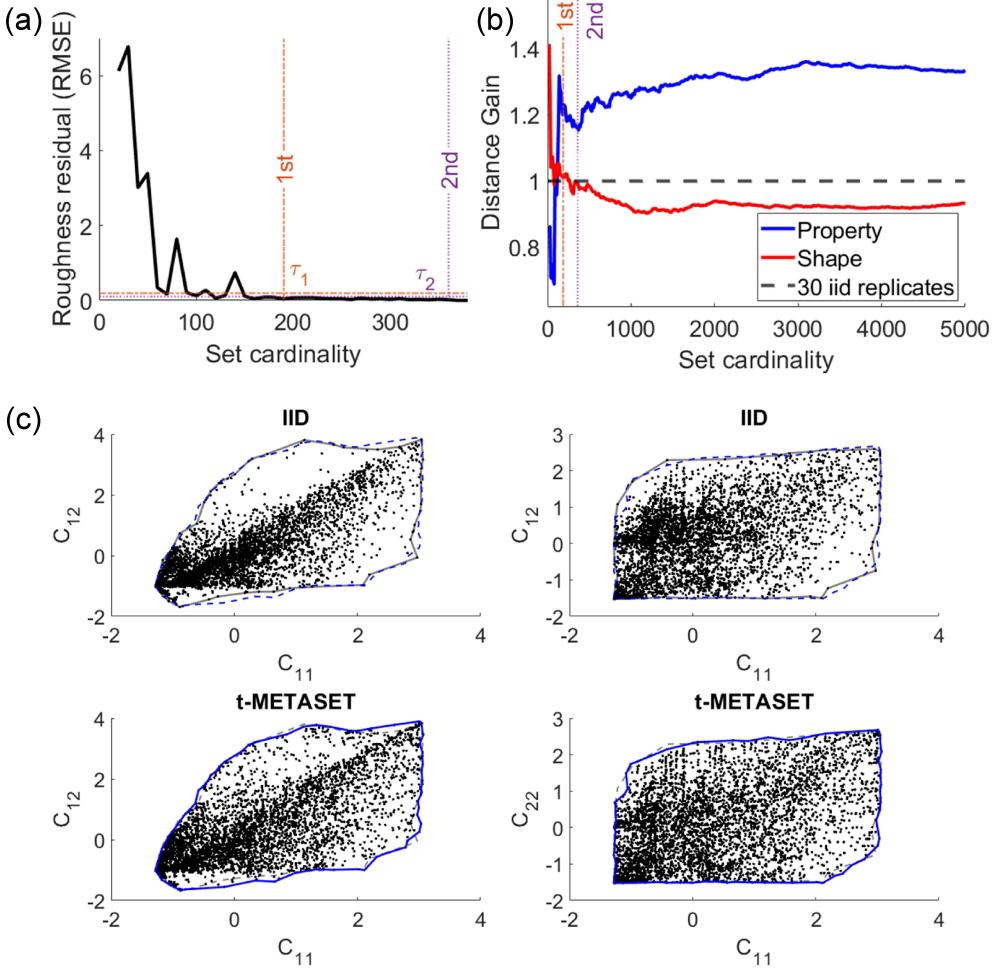


Figure 6: Scenario I result for \mathcal{D}_{TO} [39]. (a) History of roughness residual. The two vertical lines indicate the first and second roughness convergence, respectively. (b) History of distance gains. The horizontal dotted line denotes the distance gain of 30 iid replicates, which is unity by the definition in Section 4.2. The two vertical lines indicate the first (τ_1) and second (τ_2) convergence, respectively. (c) Property distribution in projected property space.

finite gamut as $\partial\Omega_p^{(t)} \rightarrow \partial\Omega_p^*$, where $\partial\Omega_p^*$ denotes the property gamut of fully observed \mathcal{D}_{TO} , which obviously exists yet is unknown in our scenarios; (iii) adding more data points within the confined boundary $\partial\Omega_p^*$ would decrease pairwise distances on average. The convergence behavior of the numerator of the property gain may possibly give a hint to answering the fundamental research question in data-driven design: “*How much data do we need?*”. In addition, adjusting the batch composition – the ratio of property versus shape – would lead to different results. The parameter study on ϵ is addressed and discussed in Section 4.5.

Figure 6(c) shows a qualitative view of the resulting property distributions. Figure 6(c) shows the data distribution in the projected property space, whose property components have been standardized. In the C_{11} - C_{12} space, the iid realization shows significant bias on the southeast region near $[-1 \leq C_{11} \leq 1] \times [-1.5 \leq C_{12} \leq 1]$, whereas only tiny samples are located on the upper region. Other 3,000-size iid realizations also result in property bias; local details are different, but the overall trend of distributional bias is more or less the same. On the other hand, the property distribution of t-METASET shows significantly reduced bias in the property spaces, in terms of projected pairwise distances and the property gamut $\partial\Omega_p$ as well.

4.4 Scenario II: Quality-Weighted Diversity (Task-Aware Generation)

Regarding task-aware acquisition of datasets, the scope of this work is dedicated to *pointwise* quality, where the task-related “value” of each observation is modeled based on a score function. It can be a function of properties (e.g.,

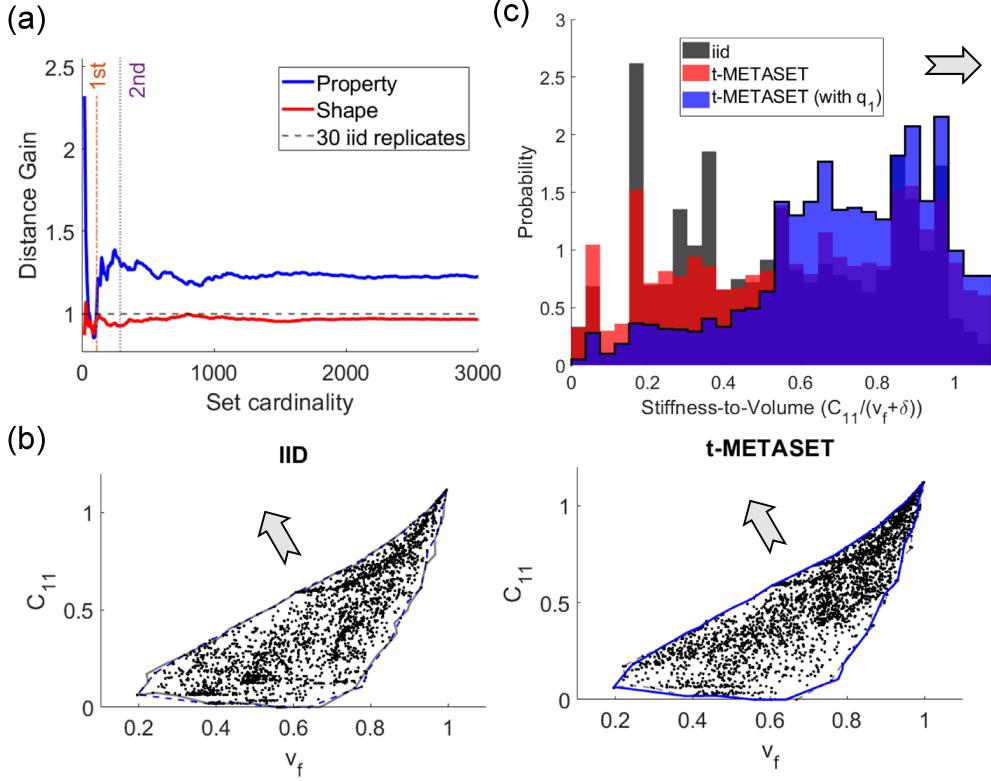


Figure 7: Task II-1 (stiffness-to-mass ratio) result for \mathcal{D}_{mix} [13]. Each arrow indicates the distributional bias preferred by the given task. (a) History of distance gains. (b) Data distribution: (left) iid and (right) t-METASET with the quality of interest (q_1). (c) Histogram of task-related quality (stiffness-to-volume): iid (gray); t-METASET (red); t-METASET with quality q_1 (blue).

stiffness anisotropy), shape (e.g., boundary smoothness), or both (stiffness-to-mass ratio). With proper formulation and scaling, the quality function can be included in t-METASET as a secondary sampling criterion. We present two examples, each of which involves either (i) only property (Section 4.4.1) or (ii) both shape and property (Section 4.4.2). All the results in this subsection assume the maximum cardinality is fixed as $|\bigcup_{t=1}^{t_{max}} \mathcal{B}^{(t)}| = 3,000$.

4.4.1 Task II-1: Promoting High Stiffness-to-Mass Ratio

Outstanding stiffness-to-mass ratio is one of the key advantages of mechanical metamaterial systems compared to conventional structures [1]. If lightweight design is of interest, users could attempt to prioritize observations with high stiffness-to-mass ratios. We take C_{11} as an example with an associated score $q(\cdot)$ formulated as

$$q_1(z, \hat{p}) = \frac{\widehat{C}_{11}}{v_f + \delta} \quad (20)$$

where v_f is the volume fraction of a given binary shape $I(x, y)$ implicitly associated with z , and δ is a small positive number to avoid singularity. Here, we use raw (not standardized) values of C_{11} to ensure that all the values are nonnegative. Note that the property \hat{p} takes both (i) ground-truth properties from the finite element analysis and (ii) predicted properties from the regressor \mathcal{GP} . To accommodate various datasets at different scales without manual scaling, we standardize q_1 into q'_1 . Then it is passed to the following Sigmoid transformation:

$$a_1(\cdot) = 1 - \frac{1}{(1 + \exp(-20(\cdot)))}, \quad (21)$$

where $a_1(\cdot)$ is the decreasing Sigmoid activation function. To accommodate the design attributes associated with the quality function $a_1(q')$, the RFF V_p of the property diversity kernel \widehat{L}_P has the pointwise quality on board according to Eq. 17.

Figure 7 presents the result for \mathcal{D}_{mix} . As indicated by the arrow, the quality function aims to bias the distribution in the $C_{11}-v_f$ space towards the northwest direction. In Figure 7(b), the resulting distribution of t-METASET shows an even stronger bias to the upper region than that of the *iid* replicates, whereas the data points near the bottom right gamut are more sparse. Figure 7(c) provides even more intuitive evidence: t-METASET without the quality function does not show distributional difference with the *iid* case. In contrast, the quality-based t-METASET leads to the strongly biased distribution – virtually opposite to the *iid* one – congruent with the enforced quality over high stiffness-to-volume ratio. Both plots corroborate that t-METASET can accommodate the preference of high stiffness-to-volume ratio, *even when starting with no property at all*. Along the way, t-METASET addresses property diversity as well, as indicated by the distance gain of property that exceeds unity (Figure 7(a)).

4.4.2 Task II-2: Promoting High Stiffness Anisotropy

Property anisotropy of unit cells is another key quality that mechanical metamaterials could leverage to achieve strong directional performances at system levels. With \mathcal{S}_{TO} , we attempt to deliberately bias the property distribution towards strong elastic anisotropy between C_{11} and C_{22} . We devise the anisotropy index as an associated quality function:

$$q_2(\mathbf{z}, \hat{\mathbf{p}}) = \frac{|\arctan(\widehat{C_{22}}/\widehat{C_{11}}) - \pi/4|}{\pi/4}, \quad (22)$$

where $\widehat{C_{11}}$ and $\widehat{C_{22}}$ denote the raw non-negative elastic constants predicted by the GP model, without any normalization; $\arctan(\widehat{C_{22}}/\widehat{C_{11}}) \in [0, \pi/2]$ is the polar angle in the C_{11} - C_{22} space; if isotropic (i.e., $C_{11} = C_{22}$), the index is 0, whereas either $C_{22}/C_{11} \rightarrow 0^+$ or $C_{22}/C_{11} \rightarrow \infty$, the index goes to 1. By the definition, the quality function ranges within $[0, 1]$. Without further scaling, we directly pass it to a monotonically increasing Sigmoid activation:

$$a_2(\cdot) = \frac{1}{1 + \exp(-20((\cdot) - 0.5))}. \quad (23)$$

Similar to the first example above, $a_2(q_2)$ is incorporated into the RFF of property through Eq. 17.

Figure 8 illustrates the result for \mathcal{D}_{TO} under the anisotropy preference. The two arrows indicate the bias direction of interest: Samples with isotropic elasticity on the line $C_{22} = C_{11}$, denoted as the green dotted line, are least preferred. From the scatter plot of Figure 8(a), the distribution of t-METASET exhibits clear bias towards the preferred direction compared to the *iid* case, while samples near the isotropic line are sparse except near the origin. The trend is even more apparent in the histograms of Figure 8(b): both the results from *iid* and vanilla t-METASET share a similar distribution in terms of polar angle. In contrast, task-aware t-METASET exhibits a bimodal distribution that is highly skewed to either 0 or $\pi/2$.

In Figure 8(a), we recognize an interesting point that reveals the power of t-METASET: unlike the other cases introduced, the shape gain also exceeds unity at the plateau stage, at mild cost of the property gain. Note that we did *not* enforce the framework to assign more resources on shape diversity. The quality function $q_2(\cdot, \cdot)$ has been defined over only the two properties C_{11} and C_{22} , *not* shape. Furthermore, during Stage II, t-METASET can take only two samples from shape diversity in each batch due to the setting $\epsilon = 0.8$, commonly shared by the other cases introduced. This indicates that: the decent exploration in the shape space – the shape gain comparable to the property gain during Stage II – is *what t-METASET autonomously decided via active learning to fulfill the mission specified by the given task*. The result demonstrates the ability of t-METASET to, given a large-scale dataset and on-demand design quality, decide how to properly tradeoff distributional biases in shape/property space, thereby efficiently addressing the design goals without human supervision.

We emphasize that the two results came from the same algorithmic settings of t-METASET shared with the other cases, except for the quality functions. Hence, the two case studies, investigated with respect to different datasets and different quality functions, demonstrate that t-METASET has fulfilled the mission: growing task-aware yet balanced datasets by active learning.

4.5 Scenario III: Joint Diversity

The proposed t-METASET can tune shape-property joint diversity when building datasets. Chan et al. [11] demonstrated that, given a *fully observed* dataset, the DPP-based sampling method can identify representative subsets with adjustable joint diversity [11]. It is grounded on the fact that any linear combination of PSD shape and property kernels can create a joint diversity kernel $L_J = (1 - \epsilon)L_s + \epsilon L_p$ that is also PSD, where L_s is a shape similarity kernel involving a shape descriptor s . Yet the linear combination approach does not apply to our proposed t-METASET, driven by the RFF V , because the linear combination of the feature V does not guarantee the resulting joint kernel to be PSD.

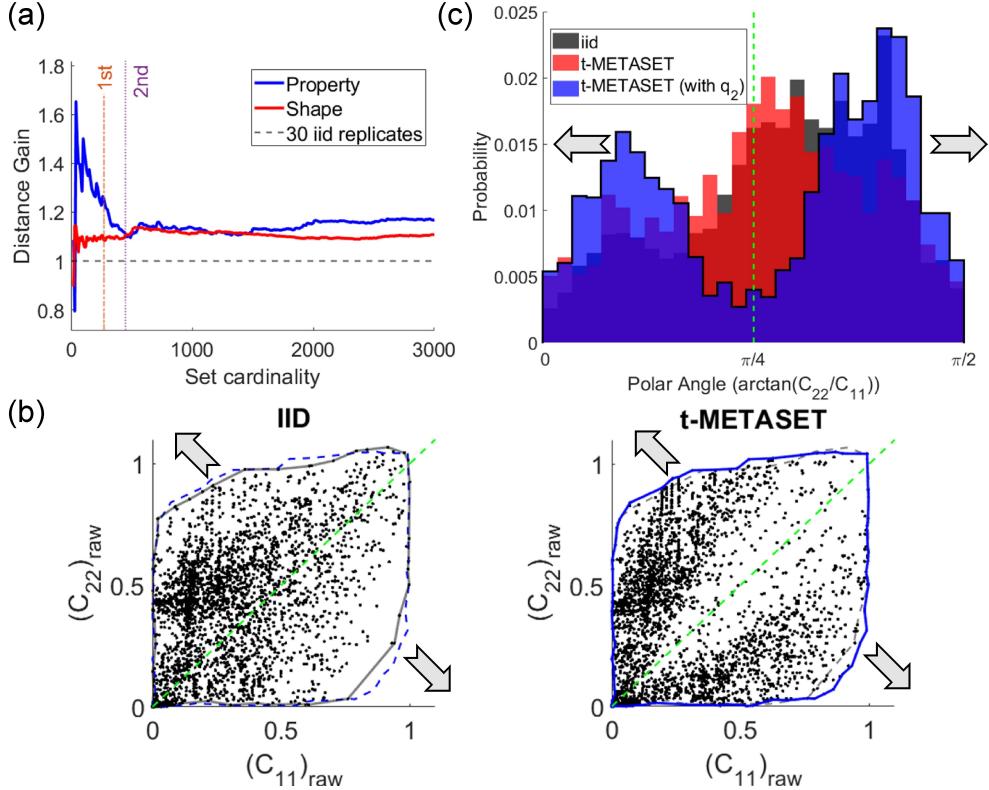


Figure 8: Task II-2 (stiffness anisotropy) result for \mathcal{D}_{TO} [39]. Each arrow indicates the distributional bias preferred by the given task. The green dotted line denotes the subregion where $C_{22} = C_{11}$, least preferred by the given task. (a) History of distance gains. (b) Data distribution: (left) *iid* and (right) t-METASET with the quality of interest (q_2). (c) Histogram of task-related quality (polar angle): *iid* (gray); t-METASET (red); t-METASET with quality q_2 (blue).

Instead, our framework tunes joint diversity by adjusting the shape/property sampling ratio ϵ of a batch. Figure 9 shows the parameter study over the batch composition ϵ with respect to D_{mix} and D_{TO} with $|\bigcup_{t=1}^{t_{max}} \mathcal{B}^{(t)}| = 5,000$. Both results manifest (i) better average diversity in terms of Euclidean distances than that of the *iid* replicates, and (ii) the tradeoff between shape diversity and property diversity. Additionally, the results support the previous finding that the correlation between shape diversity and property diversity is near-zero [11]. The substantial distinction of t-METASET lies in: We *sequentially* achieve the jointly diverse datasets, *beginning from scratch in terms of property*. In addition, t-METASET allows users to dynamically adjust ϵ as well, based on either real-time monitoring over diversity gains or user-defined criteria. This capacity could possibly help designers steer the sequential data acquisition at will, especially if growing a large-scale dataset ($\sim \mathcal{O}(10^4)$) is of interest, since applying a single sampling criterion over the whole generation procedure might not necessarily result in the best dataset for given design tasks.

4.6 Algorithm Efficiency

The t-METASET procedure is, in essence, a sequential decision making problem that selects from a given large pool of instances ($\sim \mathcal{O}(10^4)$). Its scalability comes primarily from the RFF based kernel approximation, and secondarily from the compact 10-D shape descriptor distilled from the VAE training. To give readers a glimpse of the scalability of the proposed data acquisition, Fig. 10 shows the history of wall time (i.e., elapsed real time) per iteration for \mathcal{D}_{TO} , where approximately 88k instances are included. The test was run using a desktop with Intel(R) Xeon(R) W-2295 CPU @ 3.00GHz, 18 cores/36 threads, RAM 256 Gb.

For each iteration, we look into the trend of the wall time based on three key steps: (i) GP updating, (ii) DPP sampling, and (iii) RFF updating. In the early stages, the incurred time for the GP update escalates rapidly over the dataset size, dominated by the inversion of covariance matrices. Once the second condition of roughness convergence is met ($(\Delta^{(t)}) \leq \tau_2$; denoted as “2nd” in Fig. 10), the improvements of GP updates over new batches become marginal. The sequential updates are then replaced by the preposterior analysis [59], whose computational cost gradually increases

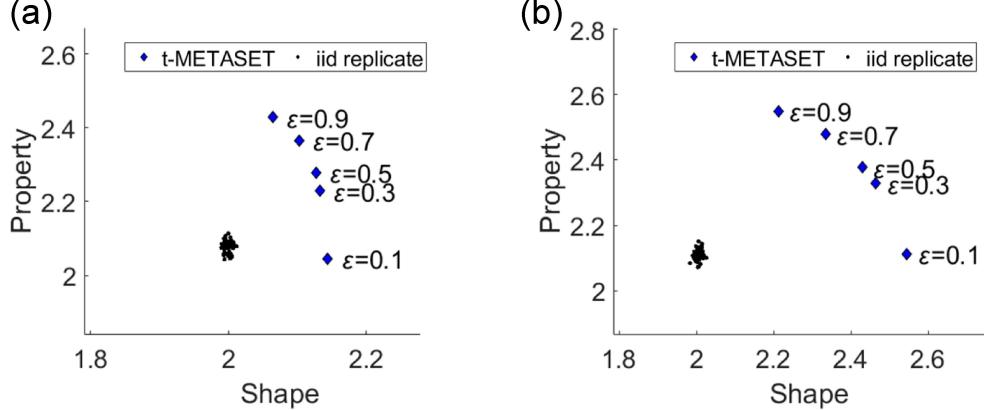


Figure 9: Scenario III (joint diversity) results for \mathcal{D}_{mix} [13] and \mathcal{D}_{TO} [39]. ϵ denotes the ratio of property samples in each batch. (a) Mean Euclidean distances for \mathcal{D}_{mix} . (b) Mean Euclidean distances for \mathcal{D}_{TO} .

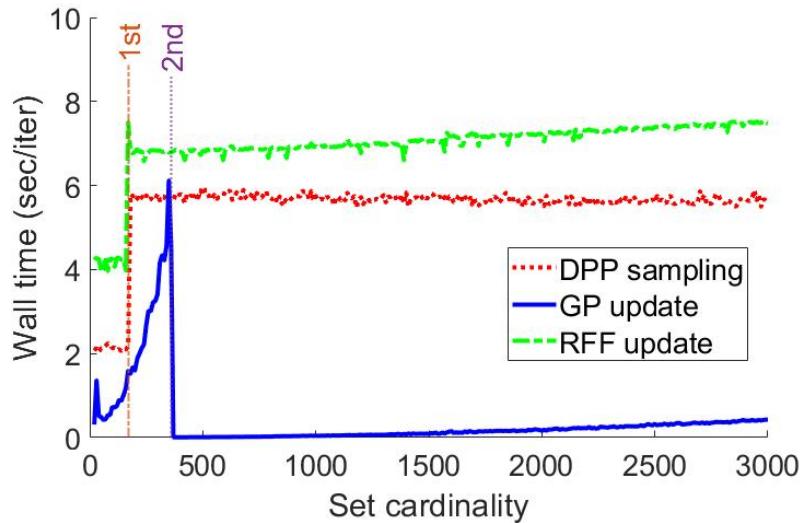


Figure 10: Elapsed real time measurement of a replicate of \mathcal{D}_{TO} as a function of dataset size. “1st” and “2nd” denote the sampling transition points corresponding to the thresholds τ_1 and τ_2 , respectively.

as the cardinality grows. Meanwhile, the conditional k -DPP for sequential diversity sampling takes up a moderate portion of time at each iteration. At the start-up phase, the DPP sampling is performed only once per iteration for the sampling based on shape diversity. As of Stage II, DPP sampling runs twice; once based on property diversity (which is optionally weighted by quality), followed by the other based on shape diversity. The incurred wall time shows little dependence on cardinality $|\mathcal{D}^{(t)}|$, as its time complexity primarily depends on the number of replicates in RFF (i.e., D_V). Lastly, the main computational overhead of the t-METASET procedure involves updating the RFF. In Stage I, only the shape RFF is updated. Once the first convergence of roughness parameters is met ($\Delta^{(t)} \leq \tau_1$; denoted as “1st” in Fig. 10), the property RFF is calculated and updated. Thereafter, every iteration involves (i) updating the shape RFF, (ii) constructing a property RFF that mirrors the latest GP, and (iii) conditioning the property RFF on the sample collected up to the current iteration.

5 Conclusion

We presented the task-aware METASET (t-METASET) framework dedicated to metamaterials data acquisition congruent with user-defined design tasks. Distinctly, t-METASET specializes in a data-driven scenario that designers often encounter in early stages of DDMD: a massive shape library has been prepared with no properties observed for a new

design case. The central idea of t-METASET for building a task-aware dataset, in general, is to (i) leverage a compact yet expressive shape descriptor (e.g., VAE latent representation) for shape dimension reduction, (ii) sequentially update a sparse regressor (e.g., GP) for nonlinear regression with sparse observations, and (iii) sequentially sample in the shape descriptor space based on estimated property diversity and estimated quality (e.g., DPP) for distribution control over shape and property. t-METASET contributes to the design field by: (i) proposing a data acquisition method *at early data-driven stages under large epistemic uncertainty*, (ii) *sequentially combating property bias*, and (iii) accommodating *task-aware design quality* as well. Starting without evaluated properties, all the results tested on two large-scale metamaterial datasets (\mathcal{D}_{mix} and \mathcal{D}_{TO}) were automatically achieved by t-METASET in three different scenarios without human supervision. We argue t-METASET can handle a variety of image-based datasets for design in general, by virtue of scalability, modularity, task-aware data customizability, and independence from both shape generation heuristics and domain knowledge.

Whereas the present scope of t-METASET is dedicated to metamaterials, the framework is applicable to other material systems where the structures, such as microstructure morphology, can be quantified. Three example scenarios in which t-METASET has potential are provided here:

- A low-dimensional representation is prescribed by a designer. This applies not only to metamaterials with an explicit parameterization (e.g., the lattice-type building block specified by four parameters [12] in Section 2), but to other systems as well (e.g., quasi random organic photovoltaic cells represented with a 2-D spectral density function [60, 61])
- A mixed categorial and quantitative representation is given. A key modification in t-METASET would be to replace the vanilla GP with a latent variable Gaussian Process (LVGP) [62]. An example is the multi-class lattice metamaterial dataset in Ref. [21]. Therein, any instance of a material is specified as $\mathbf{z} = (c_i, \rho)(i = 1, 2, \dots)$, where a qualitative variable c_i is the class index of lattice-type building blocks, and a quantitative variable ρ is the volume fraction.
- No representation is given (the scenario of primary interest in this work). Unsupervised representation learning can be harnessed, as has been employed in this work, to prepare a compact yet expressive descriptor in light of a dataset.

Thus we argue our framework can address other systems, such as those represented by user-defined descriptors or by pixels/voxels, *beyond* metamaterials. A possible issue, in particular when dealing with a system with 3-D volume elements (e.g., polymer nanocomposite), is that the dimensionality of a shape descriptor could be too large for a vanilla GP to handle, even after dimension reduction. Two workarounds for this case are: (i) employing extended GPs dedicated to high-dimensional data [63, 64] or (ii) using other surrogates with more modeling capability (e.g., a moderately sized neural network).

The imperative future work is inference-level validation of dataset quality, which aims to shed light on the downstream impact of data quality at the deployment stage of data-driven models. Among a plethora of such models, we are particularly interested in conditional generative models [65, 66] due to their on-the-fly inverse design capability, which is expected to be highly sensitive to data quality [67, 68]. The validation would further demonstrate the efficacy of t-METASET at the downstream stages of DDMD, in addition to the intuitive metric level we have shown. Moreover, we point out two interesting topics to be explored: (i) the proposed diversity gain as a termination indicator of data generation, which could offer insight into “*how much data?*” (detailed in Section 4.3), and (ii) quantitative comparison between the quality-weighted diversity sampling (Section 3.3.4) presented in this work and BO [44, 69].

Through producing and sharing open-source datasets, t-METASET ultimately aims to (i) provide a methodological guideline on how to generate a dataset that can meet individual needs, (ii) publicly offer datasets as a reference to a variety of benchmark design problems in different domains, and (iii) help designers diagnose their dataset quality on their own. This lays a solid foundation for the future advancement of data-driven design.

Acknowledgment

We acknowledge funding support from the National Science Foundation (NSF) through the CSSI program (Award # OAC 1835782).

References

- [1] Xianglong Yu, Ji Zhou, Haiyi Liang, Zhengyi Jiang, and Lingling Wu. Mechanical metamaterials associated with stiffness, rigidity and compressibility: A brief review. *Progress in Materials Science*, 94:114–173, 2018.

- [2] Costas M Soukoulis and Martin Wegener. Past achievements and future challenges in the development of three-dimensional photonic metamaterials. *Nature photonics*, 5(9):523–530, 2011.
- [3] Steven A Cummer, Johan Christensen, and Andrea Alù. Controlling sound with acoustic metamaterials. *Nature Reviews Materials*, 1(3):1–13, 2016.
- [4] Robert Schittny, Muamer Kadic, Sébastien Guenneau, and Martin Wegener. Experiments on transformation thermodynamics: molding the flow of heat. *Physical review letters*, 110(19):195901, 2013.
- [5] Muamer Kadic, Tiemo Bückmann, Robert Schittny, and Martin Wegener. Metamaterials beyond electromagnetism. *Reports on Progress in physics*, 76(12):126501, 2013.
- [6] Ruopeng Liu, Chunlin Ji, Zhiya Zhao, and Tian Zhou. Metamaterials: reshape and rethink. *Engineering*, 1(2):179–184, 2015.
- [7] Bo Zhu, Mélina Skouras, Desai Chen, and Wojciech Matusik. Two-scale topology optimization with microstructures. *ACM Transactions on Graphics (TOG)*, 36(4):1, 2017.
- [8] Zhaocheng Liu, Dayu Zhu, Sean P Rodrigues, Kyu-Tae Lee, and Wenshan Cai. Generative model for the inverse design of metasurfaces. *Nano letters*, 18(10):6570–6576, 2018.
- [9] Wei Ma, Feng Cheng, Yihao Xu, Qinlong Wen, and Yongmin Liu. Probabilistic representation and inverse design of metamaterials based on a deep generative model with semi-supervised learning strategy. *Advanced Materials*, 31(35):1901111, 2019.
- [10] Liwei Wang, Yu-Chin Chan, Faez Ahmed, Zhao Liu, Ping Zhu, and Wei Chen. Deep generative modeling for mechanistic-based learning and design of metamaterial systems. *Computer Methods in Applied Mechanics and Engineering*, 372:113377, 2020.
- [11] Yu-Chin Chan, Faez Ahmed, Liwei Wang, and Wei Chen. Metaset: Exploring shape and property spaces for data-driven metamaterials design. *Journal of Mechanical Design*, 143(3):031707, 2021.
- [12] Liwei Wang, Anton van Beek, Daicong Da, Yu-Chin Chan, Ping Zhu, and Wei Chen. Data-driven multiscale design of cellular composites with multiclass microstructures for natural frequency maximization. *Composite Structures*, 280:114949, 2022.
- [13] Yu-Chin Chan, Daicong Da, Liwei Wang, and Wei Chen. Remixing functionally graded structures: data-driven topology optimization with multiclass shape blending. *Structural and Multidisciplinary Optimization*, 65(5), April 2022.
- [14] Daicong Da, Yu-Chin Chan, Liwei Wang, and Wei Chen. Data-driven and topological design of structural metamaterials for fracture resistance. *Extreme Mechanics Letters*, 50:101528, 2022.
- [15] Liwei Wang, Jagannadh Boddapati, Ke Liu, Ping Zhu, Chiara Daraio, and Wei Chen. Mechanical cloak via data-driven aperiodic metamaterial design, 2021.
- [16] Evan W Wang, David Sell, Thaibao Phan, and Jonathan A Fan. Robust design of topology-optimized metasurfaces. *Optical Materials Express*, 9(2):469–482, 2019.
- [17] Sunae So, Jungho Mun, and Junsuk Rho. Simultaneous inverse design of materials and structures via deep learning: demonstration of dipole resonance engineering using core–shell nanoparticles. *ACS applied materials & interfaces*, 11(27):24264–24268, 2019.
- [18] Caglar Gurbuz, Felix Kronawetter, Christoph Dietz, Martin Eser, Jonas Schmid, and Steffen Marburg. Generative adversarial networks for the design of acoustic metamaterials. *The Journal of the Acoustical Society of America*, 149(2):1162–1174, 2021.
- [19] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. “everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.
- [20] Jun Wang, Wei Chen, Daicong Da, Mark Fuge, Rahul Rai, et al. Ih-gan: A conditional generative model for implicit surface-based inverse design of cellular structures. *arXiv preprint arXiv:2103.02588*, 2021.
- [21] Liwei Wang, Siyu Tao, Ping Zhu, and Wei Chen. Data-driven topology optimization with multiclass microstructures using latent variable gaussian process. *Journal of Mechanical Design*, 143(3):031708, 2021.
- [22] Andrew Ng. A chat with andrew on mlops: From model-centric to data-centric ai, 2021.
- [23] Eliza Strickland. Andrew ng, ai minimalist: The machine-learning pioneer says small is the new big. *IEEE Spectrum*, 59(4):22–50, 2022.
- [24] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

- [25] Ruichen Jin, Wei Chen, and Agus Sudjianto. On sequential sampling for global metamodeling in engineering design. In *International design engineering technical conferences and computers and information in engineering conference*, volume 36223, pages 539–548, 2002.
- [26] Haitao Liu, Yew-Soon Ong, and Jianfei Cai. A survey of adaptive sampling for global metamodeling in support of simulation-based complex engineering design. *Structural and Multidisciplinary Optimization*, 57(1):393–416, 2018.
- [27] Jiaqi Jiang and Jonathan A Fan. Global optimization of dielectric metasurfaces using a physics-driven neural network. *Nano letters*, 19(8):5366–5372, 2019.
- [28] Sensong An, Clayton Fowler, Bowen Zheng, Mikhail Y Shalaginov, Hong Tang, Hang Li, Li Zhou, Jun Ding, Anuradha Murthy Agarwal, Clara Rivero-Baleine, et al. A deep learning approach for objective-driven all-dielectric metasurface design. *ACS Photonics*, 6(12):3196–3207, 2019.
- [29] Sensong An, Bowen Zheng, Hong Tang, Mikhail Y Shalaginov, Li Zhou, Hang Li, Myungkoo Kang, Kathleen A Richardson, Tian Gu, Juejun Hu, et al. Multifunctional metasurface design with a generative adversarial network. *Advanced Optical Materials*, 9(5):2001433, 2021.
- [30] Eric B Whiting, Sawyer D Campbell, Lei Kang, and Douglas H Werner. Meta-atom library generation via an efficient multi-objective shape optimization method. *Optics Express*, 28(16):24229–24242, 2020.
- [31] Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):1–50, 2016.
- [32] Ismail Khalid Kazmi, Lihua You, and Jian Jun Zhang. A survey of 2d and 3d shape descriptors. In *2013 10th International Conference Computer Graphics, Imaging and Visualization*, pages 1–10. IEEE, 2013.
- [33] Georgios Vamvakas, Basilis Gatos, and Stavros J Perantonis. Handwritten character recognition through two-stage foreground sub-sampling. *Pattern Recognition*, 43(8):2807–2816, 2010.
- [34] Zhaocheng Liu, Zhaoming Zhu, and Wenshan Cai. Topological encoding method for data-driven photonics inverse design. *Optics express*, 28(4):4825–4835, 2020.
- [35] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [36] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5868–5877, 2017.
- [37] Wentai Zhang, Zhangsihao Yang, Haoliang Jiang, Suyash Nigam, Soji Yamakawa, Tomotake Furuhata, Kenji Shimada, and Levent Burak Kara. 3d shape synthesis for conceptual design and optimization using variational autoencoders. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 59186, page V02AT03A017. American Society of Mechanical Engineers, 2019.
- [38] Martin Philip Bendsoe and Ole Sigmund. *Topology optimization: theory, methods, and applications*. Springer Science & Business Media, 2003.
- [39] Liwei Wang, Yu-Chin Chan, Zhao Liu, Ping Zhu, and Wei Chen. Data-driven metamaterial design with laplace-beltrami spectrum as “shape-dna”. *Structural and multidisciplinary optimization*, 61(6), 2020.
- [40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [41] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.
- [42] Ramin Bostanabad, Yu-Chin Chan, Liwei Wang, Ping Zhu, and Wei Chen. Globally approximate gaussian processes for big data with application to data-driven metamaterials design. *Journal of Mechanical Design*, 141(11), 2019.
- [43] Anton Van Beek, Siyu Tao, Matthew Plumlee, Daniel W Apley, and Wei Chen. Integration of normative decision-making and batch sampling for global metamodeling. *Journal of Mechanical Design*, 142(3):031114, 2020.
- [44] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.
- [45] Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. Bayesian low-rank determinantal point processes. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 349–356, 2016.
- [46] Wei-Lun Chao, Boqing Gong, Kristen Grauman, and Fei Sha. Large-margin determinantal point processes. In *UAI*, 2015.

- [47] Raja Hafiz Affandi, Emily Fox, Ryan Adams, and Ben Taskar. Learning the parameters of determinantal point process kernels. In *International Conference on Machine Learning*, pages 1224–1232. PMLR, 2014.
- [48] Alex Kulesza and Ben Taskar. k-dpps: Fixed-size determinantal point processes. In *ICML*, 2011.
- [49] Raja Hafiz Affandi, Alex Kulesza, and Emily B Fox. Markov determinantal point processes. *arXiv preprint arXiv:1210.4850*, 2012.
- [50] Alexei Borodin and Eric M Rains. Eynard–mehta theorem, schur process, and their pfaffian analogs. *Journal of statistical physics*, 121(3):291–317, 2005.
- [51] Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. Low-rank factorization of determinantal point processes. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [52] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.
- [53] Walter Rudin. *Fourier analysis on groups*. Courier Dover Publications, 2017.
- [54] Raja Hafiz Affandi, Emily B Fox, and Ben Taskar. Approximate inference in continuous determinantal point processes. *arXiv preprint arXiv:1311.2971*, 2013.
- [55] Liang Xia and Piotr Breitkopf. Design of materials using topology optimization and energy-based homogenization approach in matlab. *Structural and multidisciplinary optimization*, 52(6):1229–1241, 2015.
- [56] Erik Andreassen and Casper Schousboe Andreasen. How to determine composite material properties using numerical homogenization. *Computational Materials Science*, 83:488–495, 2014.
- [57] Jason L Loepky, Jerome Sacks, and William J Welch. Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, 51(4):366–376, 2009.
- [58] Wei-Liem Loh. On latin hypercube sampling. *The annals of statistics*, 24(5):2058–2080, 1996.
- [59] Anton van Beek, Umar Farooq Ghumman, Joydeep Munshi, Siyu Tao, TeYu Chien, Ganesh Balasubramanian, Matthew Plumlee, Daniel Apley, and Wei Chen. Scalable adaptive batch sampling in simulation-based design with heteroscedastic noise. *Journal of Mechanical Design*, 143(3), 2021.
- [60] Umar Farooq Ghumman, Akshay Iyer, Rabindra Dulal, Joydeep Munshi, Aaron Wang, TeYu Chien, Ganesh Balasubramanian, and Wei Chen. A spectral density function approach for active layer design of organic photovoltaic cells. *Journal of Mechanical Design*, 140(11), 2018.
- [61] Akshay Iyer, Rabindra Dulal, Yichi Zhang, Umar Farooq Ghumman, TeYu Chien, Ganesh Balasubramanian, and Wei Chen. Designing anisotropic microstructures with spectral density function. *Computational Materials Science*, 179:109559, 2020.
- [62] Yichi Zhang, Siyu Tao, Wei Chen, and Daniel W Apley. A latent variable approach to gaussian process modeling with qualitative and quantitative factors. *Technometrics*, 62(3):291–302, 2020.
- [63] Santu Rana, Cheng Li, Sunil Gupta, Vu Nguyen, and Svetha Venkatesh. High dimensional bayesian optimization with elastic gaussian process. In *International conference on machine learning*, pages 2883–2891. PMLR, 2017.
- [64] Rohit Tripathy, Ilias Bilionis, and Marcial Gonzalez. Gaussian processes with built-in dimensionality reduction: Applications to high-dimensional uncertainty propagation. *Journal of Computational Physics*, 321:191–223, 2016.
- [65] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [66] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.
- [67] Yufeng Zheng, Yunkai Zhang, and Zeyu Zheng. Continuous conditional generative adversarial networks (cgan) with generator regularization. *arXiv preprint arXiv:2103.14884*, 2021.
- [68] Amin Heyrani Nobari, Wei Chen, and Faez Ahmed. Pcdgan: A continuous conditional diverse generative adversarial network for inverse design. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 606–616, 2021.
- [69] Siyu Tao, Anton Van Beek, Daniel W Apley, and Wei Chen. Multi-model bayesian optimization for simulation-based design. *Journal of Mechanical Design*, 143(11), 2021.