MRS Communications

https://doi.org/10.1557/s43579-022-00250-0

MRS Communications

Early Career Materials Researcher Research Letter

MetaNOR: A meta-learnt nonlocal operator regression approach for metamaterial modeling

Lu Zhang, Huaiqian You, and Yue Yu, Department of Mathematics, Lehigh University, 27 Memorial Dr W, Bethlehem, PA 18015, USA

Address all correspondence to Yue Yu at **yuy214@lehigh.edu**

(Received 16 May 2022; accepted 23 August 2022)

Abstract

We propose MetaNOR, a meta-learnt approach for transfer-learning operators based on the nonlocal operator regression. The overall goal is to efficiently provide surrogate models for new and unknown material-learning tasks with different microstructures. The algorithm consists of two phases: (1) learning a common nonlocal kernel representation from existing tasks; (2) transferring the learned knowledge and rapidly learning surrogate operators for unseen tasks with a different material, where only a few test samples are required. We apply MetaNOR to model the wave propagation within 1D metamaterials, showing substantial improvements on the sampling efficiency for new materials.

Introduction

Metamaterial is a group of artificial heterogeneous materials exhibiting unusual yet desired frequency dispersive properties from its composite microstructure. These materials have been studied theoretically, numerically and experimentally,[1–3] to optimize its microstructure for dispersive properties and performances in different environments. However, their design and modeling is often computationally prohibitive, since they contain more than thousands of interfaces in its microstructures (example interfaces are shown in Fig. 1). For example, modeling dispersive properties via the stress wave propagation in metamaterials would require accurate bottom-up characterization of material interfaces and simulations from each individual layer (the microscopic scale), which are often of orders smaller than the problem domain length scale (the macroscopic scale). Therefore, even with sophisticated optimization techniques, designing such a material would still require running multiple full wave simulations for each candidate microstructure and consumes significant computation time.

To accelerate stress wave simulations, efficient surrogate models for metamaterials such as homogenized models are often employed.[4,5] They are posed as a single equation of the displacement field and can be readily used in simulations at the macroscopic scale. Among various choices of homogenized models, nonlocal surrogate models have received lots of attentions,[6–8] where integral operators are employed which embeds all time and length scales in their definitions. Comparing with continuum partial differential equation (PDE) counterparts, the nonlocal surrogate models provide a more natural affinity with the dispersive waves in a microstructure, and successfully reproduce many features of the decay and spreading of stress waves.[8] In nonlocal homogenization models the choice of kernel functions contains information about the small-scale response of the system. Hence, once equipped with the power of machine learning to identify the optimal form for the kernel function, effective nonlocal surrogate models can be obtained from high-fidelity (HF) simulations and/or experimental measurements, so as to reproduce the material response with the greatest fidelity. To this end, the nonlocal operator regression (NOR) approach,[6,9,10] is proposed to obtain large-scale nonlocal descriptions and capture small-scale material behavior that would remain hidden in classical approaches to homogenization. NOR and the general homogenized surrogate models have greatly accelerated the simulation of heterogeneous materials at macroscale, but the choice of homogenized models is often selected case by case, which makes rigorous model calibration for multiple microstructure challenging and time consuming. Moreover, in metamaterials and the general heterogeneous material modeling problems, the data acquisition is often very challenging and expensive, which makes it critical for any method to learn the material model with a limited number of measurements.

Motivated by further accelerating the design process for metamaterials, in this paper we leverage NOR and the general heterogeneous material homogenization procedure, to answer the following question: *Given knowledge on a number of materials with different microstructures, how can one efficiently learn the best surrogate model for a new and possibly unknown microstructure, with only a small set of training data (such as several pairs/measurements of displacement and loading fields from experiments)?*

To answer this question, we develop sample-efficient data-driven homogenized models for new metamaterial with unseen microstructure, which (1) allow for accurate simulations of wave propagation at much larger scales than the microstructure;

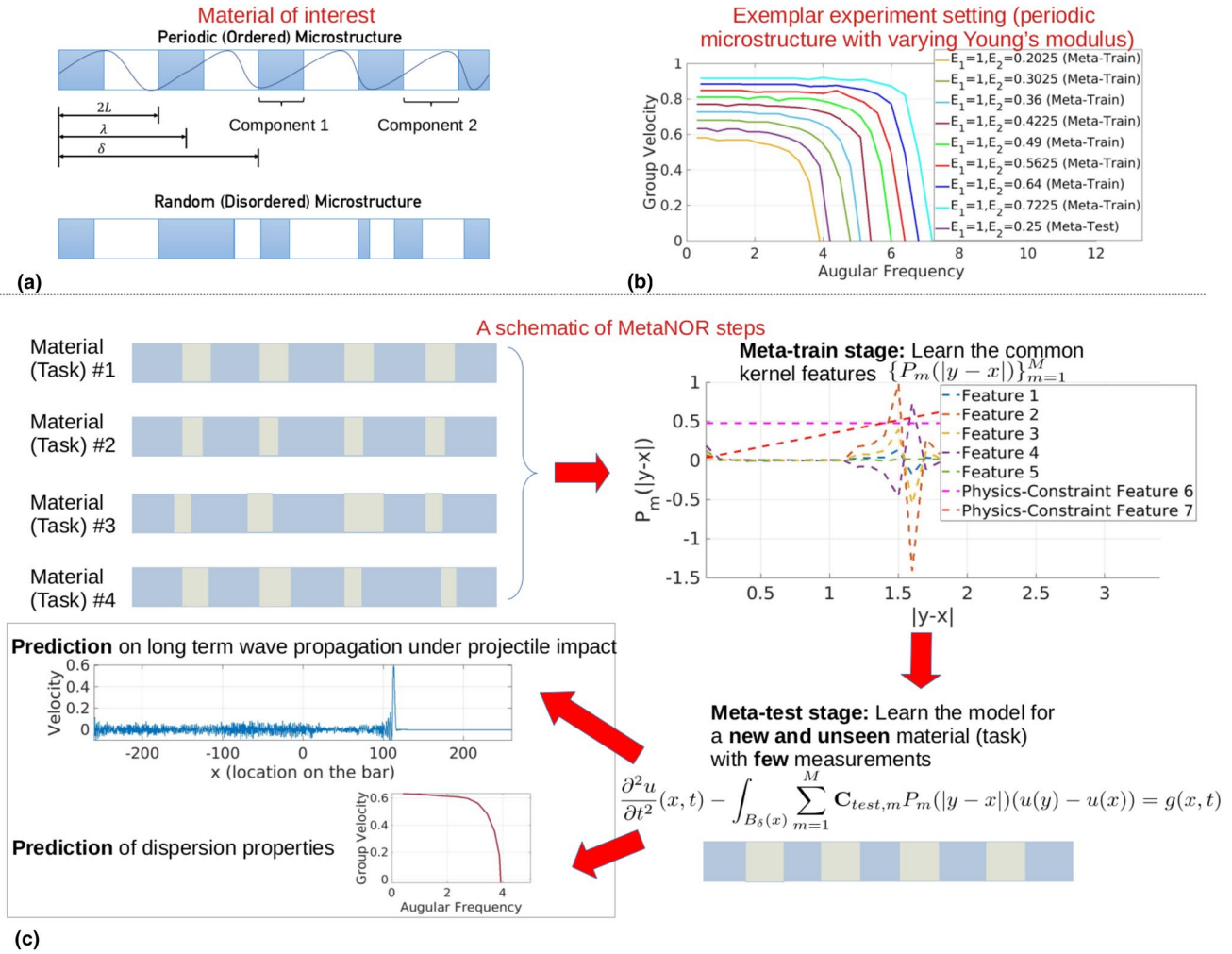


Figure 1. Problem of interests and a schematic of the proposed algorithm. (a) One-dimensional metamaterial composed by dissimilar components 1 and 2. Components 1 and 2 have the densities ρ_1 and ρ_2 and Young's modulus E_1 and E_2. The horizon δ, and the wave length λ, are reported for comparison. (b) A demonstration of our experiment setting. (c) A schematic of the main steps in MetaNOR.

(2) provide constitutive laws that can be readily applied in simulation problems posed on various environments with general geometries and complex time-dependent loadings; and (3) utilize the knowledge from previously studied materials to rapidly adapt to new microstructures. Specifically, inspired by a recent provable meta-learning approach of linear representations,[11] we propose meta nonlocal operator regression (MetaNOR), a sample-efficient learning algorithm for metamaterials where multiple tasks (microstructures) share a common set of low-dimensional features in their kernel space, for accurate and efficient adaptation to unseen tasks. The algorithm has three components. First, in the *meta-train* stage, we learn a common set of features in the nonlocal kernel space from multiple related tasks (i.e., related microstructures), by minimizing the corresponding empirical risk induced by nonlocal surrogate models. Second, in the *meta-test* stage, with estimated kernel feature representation shared across tasks, we transfer this knowledge to learn the model for a new and unseen microstructure, where only a few samples/measurements are required. Third, when partial physical knowledge is available, such as the effective wave speed for infinitely long wavelengths and/or the dispersion properties of the material for very long waves, we incorporate these physics-based constraints into the proposed meta-learning algorithm to further improve the learning accuracy and sample efficiency.

Main contributions: we summarize the main contributions of this paper as follows:

1. We design a novel meta-learning technique for learning nonlocal operators, by learning a common set of low-dimensional features on multiple known tasks in their kernel space and then transferring this knowledge to new and unseen tasks.
2. Our method is the first application of meta-learning approach on homogenized model for heterogeneous materials, which efficiently provides an associated model surro-

gates that are effective on applications with various loading and time scales.
3. We provide rigorous error analysis for the proposed algorithm, showing that the estimator converges as the data resolution refines and the number of sample increases. We verify its efficacy on a synthetic dataset.
4. We illustrate the proposed method on one-dimensional metamaterials, to confirm the applicability of our technique and the improved sample efficiency over baseline nonlocal operator regression estimators.

Nonlocal operator regression (NOR)

We first review related concepts of the general nonlocal operator regression (NOR) approach, and then demonstrate the equivalence of NOR and a linear regression model in the kernel space, which provides the foundation for us to formulate MetaNOR as a linear kernel feature learning problem with theoretical guarantees in the next section.

Throughout this paper, we use $\mathbf{A}^*$ to denote its optima for each vector or matrix or operator $\mathbf{A}$ of interests, $\mathbf{A}^+$ to denote its ground-truth, and $\mathbf{A}^\circ$ to denote an approximation solution to the optima. For any vector $\mathbf{v} = [v_1, \ldots, v_q] \in \mathbb{R}^q$, we use $||v||_{l^2} := \sqrt{\sum_{i=1}^q v_i^2}$ to denote its l^2 norm, and similarly $||v||_{l^1} := \sum_{i=1}^q |v_i|$ denotes the l^1 norm. For any matrix $\mathbf{A} = [A_{ij}] \in \mathbb{R}^{p\times q}$, we use $||\mathbf{A}||_F := \sqrt{\sum_{i=1}^p \sum_{j=1}^q A_{ij}^2}$ to denote its Frobenius norm, and use $||\mathbf{A}||$ to denote its spectral norm. $\gtrsim$ and $\lesssim$ denote greater than and less than, up to a universal constant, respectively. We use O, Ω and Θ as in the standard notations, and $\tilde{O}$ as an expression that hides polylogarithmic factor in problem parameters. $\mathbf{I}_p$ denotes the $p \times p$ identity matrix.

Nonlocal models and Kernel learning

Nonlocal Models[12] describe the state of a system, where any point depends on the state in a neighborhood of points. In heterogeneous materials, it is shown that nonlocality naturally appears in the homogenized model derived from micromechanical models,[13] which makes nonlocal operators good candidates for obtaining homogenized models for heterogeneous materials.[14]

Formally, on some bounded domain $\Omega \in \mathbb{R}^p$, we model the HF or ground-truth material response as a mapping between two function spaces, $\mathcal{L}^+ : \mathcal{U} \to \mathcal{F}$, where $\mathcal{U} = \{u(x), x \in \Omega\}$ (can be seen as the space of displacement fields) and $\mathcal{F} = \{f(x), x \in \Omega\}$ (can be seen as the space of loading fields) are Banach spaces. The goal of nonlocal operator learning is to then find a surrogate operator $\mathcal{L}_\theta : \mathcal{U} \to \mathcal{F}$ with parameter θ, such that $\mathcal{L}_\theta \approx \mathcal{L}^+$. In particular, we assume that the action of the ground-truth material model may be approximated by an integral operator of the form:

$$\mathcal{L}_\theta[u](x) = \int_\Omega \gamma_\theta(x,y)(u(y) - u(x))\mathrm{d}y, \tag{1}$$

where the kernel γ takes inputs spatial locations x and y. Inspired by the application of nonlocal diffusion operator in metamaterial problems,[6,7] we choose to take $\gamma_\theta(x,y) := \gamma_\theta(|y-x|)$ as a radial and sign-changing kernel function, which is compactly supported on the ball of radius δ centered at x, denoted as $B_\delta(x)$. Moreover, we represent γ_θ as a linear combination of basis polynomials:

$$\gamma_\theta(|y-x|) = \sum_{m=1}^M C_m P_m(|y-x|), \tag{2}$$

with properly chosen basis polynomials $\{P_m(|y-x|)\}$.

Suppose we are given observations of K pairs of functions $\{u_k(x), f_k(x)\}_{k=1}^K$ where $u_k(x)$ are samples in $\mathcal{U}$ and $\mathcal{L}^+[u_k] = f_k$, potentially with noise. Then, learning the nonlocal kernel γ_θ can be framed as an optimization problem. Here we consider the squared loss in the $L^2(\Omega)$ norm:

$$\mathcal{J}(\theta) = \mathbb{E}_u\left[||\mathcal{L}_\theta[u] - f||^2_{L^2(\Omega)}\right] \approx \frac{1}{K}\sum_{k=1}^K ||\mathcal{L}_\theta[u_k] - f_k||^2_{L^2(\Omega)}.$$

Therefore, learning the surrogate operator $\mathcal{L}_\theta$ is equivalent to finding optimal parameters $\theta = \{C_m\}$ by minimizing the objective $\mathcal{J}(\theta)$. Here, we stress that NOR aims to learn a surrogate operator for the ground-truth operator $\mathcal{L}^+$, rather than the displacement field solution $u(x)$ for a single instance loading field $f(x)$. Since our goal is to provide a homogenized surrogate model which can be readily applied in simulation problems with time-dependent loadings, the former setting is more appropriate, which directly approximates the solution operator and finds the solutions for different loadings at different time instants. Moreover, approximating the operator also possesses the notable advantages of resolution independence and convergence.

NOR for metamaterial homogenization

We now seek nonlocal homogenized models for the stress wave propagation problem in a one-dimensional metamaterial, i.e., $p = 1$, using NOR as described above. As illustrated in Fig. 1, each material is a heterogeneous bar formed by two dissimilar materials, with microstructure either made of periodic layers or randomly generated layers. The goal of NOR is to learn a surrogate model which is able to reproduce wave propagation on distances that are much larger than the size of the microstructure, without resolving the microscales.

We assume that there exists a (possibly unknown or nonfeasible) HF model that faithfully represents the wave propagation in detailed microstructures: for $(x,t) \in \Omega \times [0,T]$,

$$\frac{\partial^2 \hat{u}}{\partial t^2}(x,t) - \mathcal{L}^+[\hat{u}](x,t) = g(x,t). \tag{3}$$

Here $\mathcal{L}^+$ is the HF operator that considers the detailed microstructure, $\hat{u}(x,t)$ is the HF solution which can be provided either from fine-scale simulations or from experiment measurements in practice, and $g(x,t)$ represents a time-dependent force loading term. Analogously, we will refer to the homogenized effective nonlocal operator as $\mathcal{L}_\theta$, and assume that the surrogate model has the form

$$\frac{\partial^2 \tilde{u}}{\partial t^2}(x,t) - \mathcal{L}_\theta[\tilde{u}](x,t) = g(x,t), \tag{4}$$

for $(x,t) \in \Omega \times [0,T]$, augmented with Dirichlet-type boundary conditions from the HF solution on a layer of thickness δ that surrounds the domain, and the same initial conditions as in the HF model. Here $\tilde{u}(x,t)$ is the homogenized solution. As shown in,[15] the second-order-in-time nonlocal equation (4) is guaranteed to be well-posed as far as γ_θ is uniformly Lipschitz continuous. Therefore, when parameterizing the nonlocal kernel γ_θ as a linear combination of basis polynomials following (2), one can make sure that the learnt model can be readily applied in simulation problems.

To learn the optimal $\mathcal{L}_\theta$, suppose we have K observations of forcing terms $g_k(x,t)$ and the corresponding HF solution/experimental measurements $\hat{u}_k(x,t)$, $k = 1,\dots,K$, measured at time instance $t^n \in [0,T]$ and discretization points $x_i \in \Omega$. Without loss of generality, here we assume measurements are provided on uniformly spacing spatial and time instances, with fixed spatial grid size Δx and time step size Δt. Denoting the collection of discretization points as $\chi = \{x_i\}_{i=1}^L$, then the training dataset contains $N_{\text{train}} := LK\lfloor T/\Delta t \rfloor$ measurements in total, specifically,

$$\mathcal{D}_{\text{train}} = \{(\hat{u}_k(x_i,t^n), g_k(x_i,t^n))\}_{k,i,n=1}^{K,L,\lfloor T/\Delta t\rfloor}.$$

In NOR the squared loss then writes:

$$\begin{aligned}\mathcal{J}(\theta) \approx& \frac{1}{K}\sum_{k=1}^{K} \left|\left|\mathcal{L}_\theta[\hat{u}_k] - \mathcal{L}^+[\hat{u}_k]\right|\right|^2_{L^2(\Omega\times[0,T])} \\ =& \frac{1}{K}\sum_{k=1}^{K} \left|\left|\mathcal{L}_\theta[\hat{u}_k] - \frac{\partial^2 \hat{u}_k}{\partial t^2} + g_k(x,t)\right|\right|^2_{L^2(\Omega\times[0,T])}.\end{aligned} \tag{5}$$

To numerically evaluate the above loss, we discretize $\frac{\partial^2 \hat{u}_k}{\partial t^2}$ with the central difference scheme in time and Riemann sum approximation of the nonlocal operator in space:

$$\begin{aligned}&\frac{\partial^2 \hat{u}_k}{\partial t^2}(x,t) \approx \ddot{\hat{u}}_k(x,t) : \\ &\quad = \frac{1}{\Delta t^2}(\hat{u}_k(x,t+\Delta t) - 2\hat{u}_k(x,t) + \hat{u}_k(x,t-\Delta t)),\end{aligned} \tag{6}$$

$$\begin{aligned}\mathcal{L}_\theta[\hat{u}_k](x,t) =& \int_{B_\delta(x)} \gamma_\theta(|y-x|)(\hat{u}_k(y,t) \\ &- \hat{u}_k(x,t))\mathrm{d}y \approx \mathcal{L}_{\theta,\Delta x}[\hat{u}_k](x,t) \\ :=& \Delta x \sum_{x_j \in B_\delta(x)\cap\chi} \sum_{m=1}^{M} C_m P_m(|x_j - x|) \\ &(\hat{u}_k(x_j,t) - \hat{u}_k(x,t)) \\ =& \Delta x \sum_{\alpha=1}^{d}\sum_{m=1}^{M} C_m P_m(\alpha\Delta x)(\hat{u}_k(x+\alpha\Delta x,t) \\ &- 2\hat{u}_k(x,t) + \hat{u}_k(x-\alpha\Delta x,t)),\end{aligned} \tag{7}$$

for each $x = x_i \in \chi$, $t = t^n$ and $d := \lfloor \delta/\Delta x \rfloor$. Substituting the above schemes into (5), we then obtain

$$\mathcal{J}(\theta) \approx \frac{1}{K}\sum_{k=1}^{K}\sum_{x_i\in\chi}\sum_{n=1}^{\lfloor T/\Delta t\rfloor} \left(y_{k,i}^n - (\mathbf{s}_{k,i}^n)^T\mathbf{BC}\right)^2. \tag{8}$$

where the (reformulated) data pair $y_{k,i}^n \in \mathbb{R}$ and $\mathbf{s}_{k,i}^n \in \mathbb{R}^d$ are defined as

$$y_{k,i}^n := -\frac{1}{\Delta t^2}\left(2\hat{u}_k(x_i,t^n) - \hat{u}_k(x_i,t^{n+1}) - \hat{u}_k(x_i,t^{n-1})\right) - g_k(x_i,t^n), \tag{9}$$

$$\begin{aligned}[\mathbf{s}_{k,i}^n]_\alpha :=& \Delta x\left(\hat{u}_k(x_{i+\alpha},t^n) + \hat{u}_k(x_{i-\alpha},t^n) - 2\hat{u}_k(x_i,t^n)\right), \\ &\alpha = 1,\dots,d.\end{aligned} \tag{10}$$

The parameter vector $\mathbf{C} := [C_1,\dots,C_M] \in \mathbb{R}^M$, and the feature matrix $\mathbf{B} := (\mathbf{b}_1,\dots,\mathbf{b}_M) \in \mathbb{R}^{d\times M}$ is defined as:

$$\mathbf{B}_{\alpha m} = P_m(\alpha\Delta x),\ m = 1,\dots,M,\ \alpha = 1,\dots,d,$$

and the dimension of each feature equals to d. Therefore, the optimal parameters are obtained by solving a (constrained) optimization problem

$$\min_\theta \mathcal{J}(\theta) + \zeta\mathcal{R}(\theta), \quad \text{s.t. } \mathcal{L}_\theta \text{ satisfies physics-based constraints.} \tag{11}$$

Here $\mathcal{R}(\theta)$ is a regularization term which aims to prevent over-fitting in the inverse problem, and ζ is the regularization parameter. A commonly used regularization term is the Euclidean norm in the classical Tikhonov regularization, i.e., $\mathcal{R}(\theta) := ||\mathbf{C}||_{l^2}^2$. The physics constraints denote the additional conditions which enforce partial physical knowledge of the heterogeneous material, for which we will explain later on.

From (8), we can see that NOR is equivalent to a linear model with M-dimensional features in the kernel space. When taking P_m, $m = 1,\dots,d = M$ as the Lagrange basis polynomials, satisfying $P_m(\alpha\Delta x) = 1$ for $\alpha = m$ and $P_m(\alpha\Delta x) = 0$ for all $\alpha \neq m$, we note that $\mathbf{B}$ becomes an identity matrix and NOR will be equivalent to a linear regression problem. Therefore, algorithms and analysis on linear regression models can be immediately applied in NOR. In "Empirical experiments" section, this linear kernel regression setting will be employed as the baseline method on a new microstructure (task), which only uses data generated from that task.

Proposed meta-learning algorithm

We now consider the problem of meta-learning in NOR, such that multiple tasks share a common set of low-dimensional features in the kernel space. Given K_{train} observations $(\hat{u}_k(x,t), g_k(x,t))$ which belong to H unobserved underlying tasks, the meta-learning NOR model writes:

$$\frac{\partial^2 \hat{u}_k}{\partial t^2}(x,t) - \mathcal{L}_{\theta_{\eta(k)}}[\hat{u}_k](x,t) = g_k(x,t) + \epsilon_k(x,t). \tag{12}$$

Here, we assume that each task corresponds to a different microstructure and the underlying optimal surrogate kernel $\gamma_{\theta_{\eta(k)}}$, where $\eta(k) \in \{1, \ldots, H\}$ denotes the index of task associated with the k-th function pair. $\epsilon_k(x,t)$ is a smooth function related with the additive noise, describing the discrepancy between the ground-truth operator and the optimal surrogate operator, i.e., $\epsilon_k(x,t) := (\mathcal{L}^+ - \mathcal{L}_{\theta_{\eta(k)}})[\hat{u}_k](x,t)$. Our goal is to recover the underlining low-dimensional representation for the kernel space, and use this representation to recover a better estimate for a new and unseen task. Mathematically, we assume that there exists an unobserved kernel feature space $span(\{P_m(|y-x|)\}_{m=1}^M)$, such that $M \ll d$ and

$$\gamma_{\theta_\eta}(|y-x|) = \sum_{m=1}^{M} \mathbf{C}_{\eta,m} P_m(|y-x|), \tag{13}$$

where $\mathbf{C}_\eta$ is the parameter for the $\eta-$th task. For a new and unseen microstructure, we assume there also exists a surrogate model for it:

$$\frac{\partial^2 \hat{u}}{\partial t^2}(x,t) - \mathcal{L}_{\theta_{\text{test}}}[\hat{u}](x,t) = g(x,t) + \epsilon(x,t), \tag{14}$$

with its optimal surrogate kernel inside the unobserved kernel feature space $span(\{P_m(|y-x|)\}_{m=1}^M)$, i.e., there exists $\mathbf{C}_{\text{test}} \in \mathbb{R}^M$ such that

$$\gamma_{\theta_{\text{test}}}(|y-x|) = \sum_{m=1}^{M} \mathbf{C}_{\text{test},m} P_m(|y-x|). \tag{15}$$

Then, with only a few measurements given for a new and unseen microstructure, we aim to efficiently recover this optimal estimator $\gamma_{\theta_{\text{test}}}$.

Comparing with NOR and the other classical machine learning methods, our method aims at five desirable properties: (1) The learning algorithm is sample-efficient on the new task, which implies that the optimal estimator $\gamma_{\theta_{\text{test}}}$ can be learnt even with very scarce measurements. (2) The estimator is resolution independent, in the sense that the learnt model can be applied to different resolution problems. (3) Beyond resolution invariance, we further aim for a robust consistent estimator, that is, the estimator converges as data resolution refines. (4) The method learns the nonlocal surrogate model directly from data, i.e., no preliminary knowledge on the governing law is required. (5) The learnt model is generalizable, meaning that it is applicable to problem settings that are substantially different from the ones used for training in terms of loading and domain/time scales. Hence, once the nonlocal surrogate model is learnt, one can further employ it in further prediction tasks with a longer simulation time, a larger computational domain, and on a different grid.

Before demonstrating our main algorithm, we first establish the connection of our meta-learning model with the linear representation model illustrated in.[11] Following a similar derivation in "NOR for metamaterial homogenization" section, we reformulate the training data following (9) and (10), then denote the collection of all (reformulated) data points as

$$\mathcal{D}_{\text{train}} = \{(y_{k,i}^n, \mathbf{s}_{k,i}^n)\}_{k,i,n=1}^{K_{\text{train}}, L_{\text{train}}, \lfloor T_{\text{train}}/\Delta t \rfloor} = \{(y_j, \mathbf{s}_j)\}_{j=1}^{N_{\text{train}}}.$$

With a slight abuse of notations, in the meta-learning algorithm we consider a uniform task sampling model which does not differentiate the datapoints from different sample k, time step n and spatial grid i, then use $\eta(j) \in \{1, \ldots, H\}$ to denote the index of the task associated with the datapoint j. Then, discovering the kernel basis $\{P_m(|y-x|)\}_{m=1}^M$ is equivalent to recovering a linear feature matrix $\mathbf{B} \in \mathbb{R}^{d \times M}$ with orthonormal columns, such that $P_m(\alpha \Delta x) = \mathbf{B}_{\alpha m}$ and

$$y_j = \mathbf{s}_j^T \mathbf{B} \mathbf{C}_{\eta(j)} + \epsilon_j, \tag{16}$$

where $\mathbf{C}_{\eta(j)}$ is the parameter for the $\eta(j)-$th task, and ϵ_j is additive noise. For the new task, with a collection of (reformulated) datapoints

$$\mathcal{D}_{\text{test}} = \{(y_j^{\text{test}}, \mathbf{s}_j^{\text{test}})\}_{j=1}^{N_{\text{test}}},$$

recovering the optimal kernel $\gamma_{\theta_{\text{test}}}$ is equivalent to recovering an optimal estimate $\mathbf{C}_{\text{test}}$, such that:

$$y_j^{\text{test}} = (\mathbf{s}_j^{\text{test}})^T \mathbf{B} \mathbf{C}_{\text{test}} + \epsilon_j^{\text{test}}. \tag{17}$$

Our meta-learning model has two stages: firstly, the linear feature matrix $\mathbf{B}$ is recovered from $\mathcal{D}_{\text{train}}$, the data from the first H known tasks, then the learnt feature representation will be employed to discover an estimate of the task parameter $\mathbf{C}_{\text{test}}^\circ$ from a (scarce) test dataset corresponding to this new task. In particular, we employ the provable meta-learning algorithm recently proposed in.[11] In the following we briefly describe the main steps, with further theoretical results elaborated in "Prediction error bounds" section and discussions in "Physics-based constraints" section.

Step 1 *Data preprocessing for learning tasks* We first normalize each pair of solution $\hat{u}_k(x,t)$ and forcing term $g_k(x,t)$ with respect to the L^2 norm of $\hat{u}_k(x,t)$, then generate the reformulated data pairs following (9) and (10). To further ensure that the dataset satisfies the sub-Gaussian requirement (see Assumption 3.1), we normalize the training data pairs such that $\mathbb{E}[\mathbf{s}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{s}\mathbf{s}^T] = \mathbf{I}_d$.

Step 2 *Meta-train to learn Kernel features* As the first stage of meta-learning, we solve for $\mathbf{B} \in \mathbb{R}^{d \times M}$ and try to recover $\mathbf{W} := (\mathbf{C}_1, \ldots, \mathbf{C}_H)^T \mathbf{B}^T \in \mathbb{R}^{H \times d}$ with $rank(\mathbf{W}) = r < d$. In particular, we consider the Burer–Monteiro factorization of $\mathbf{W} = \mathbf{U}\mathbf{V}^T$ with $\mathbf{U} \in \mathbb{R}^{H \times M}$, $\mathbf{V} \in \mathbb{R}^{d \times M}$, and solve the following optimization problem

$$\min_{\mathbf{U},\mathbf{V}} \frac{H}{N_{\text{train}}} \sum_{j=1}^{N_{\text{train}}} \left(y_j - \mathbf{e}_{\eta(j)} \cdot \left[\mathbf{s}_j^T \mathbf{V}\mathbf{U}^T\right]\right)^2 + \frac{1}{4}||\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V}||_F^2, \tag{18}$$

where $N_{\text{train}} = K_{\text{train}} L_{\text{train}} \lfloor T_{\text{train}}/\Delta t \rfloor$ is the total number of training datapoints and $\mathbf{e}_j$ is the $j-$th standard basis vector in $\mathbb{R}^M$. The estimated feature matrix $\mathbf{B}^\circ$ can then be extracted, as an orthonormal basis from the column space of $\mathbf{V}^\circ$. As shown in,[16] all localminima of this optimization problem, $\mathbf{V}^\circ$, would be in the neighborhood of the optimal, that means, the approximated basis $\mathbf{B}^\circ$ would provide a good estimate to the optimal low-dimensional feature space.

Step 3 *Meta-test to transfer features to new tasks* As the second stage of meta-learning, we substitute the learnt feature matrix $\mathbf{B}^\circ$ from the first stage, and estimate the new task parameter $\mathbf{C}_{\text{test}}$ as follows:

$$\mathbf{C}_{\text{test}}^\circ = \underset{\mathbf{C}_{\text{test}}}{\text{argmin}} \sum_{j=1}^{N_{\text{test}}} \left(y_j^{\text{test}} - (\mathbf{s}_j^{\text{test}})^T \mathbf{B}^\circ \mathbf{C}_{\text{test}}\right)^2. \tag{19}$$

Since this is an ordinary least-square objective, an analytical solution can be obtained:

$$\mathbf{C}_{\text{test}}^\circ = \left(\sum_{j=1}^{N_{\text{test}}} (\mathbf{B}^\circ)^T \mathbf{s}_j^{\text{test}} (\mathbf{s}_j^{\text{test}})^T \mathbf{B}^\circ\right)^\dagger (\mathbf{B}^\circ)^T \sum_{j=1}^{S_{\text{test}}} \mathbf{s}_j^{\text{test}} y_j^{\text{test}} \tag{20}$$

where † indicates the Moore–Penrose pseudo-inverse.

Step 4 *Postprocessing to obtain a continuous model* To construct the continuous model which can be employed in further prediction tasks with various resolutions, we employ the B-spline basis functions consisting of piece-wise polynomials with degree 2. In particular, we construct the basis polynomials as $P_m^\circ(|y-x|) := \sum_{\alpha=1}^d \mathbf{B}_{\alpha m}^\circ N_{\alpha,2}(|y-x|)$, $m = 1, \ldots, M$, where $N_{\alpha,2}$ are constructed with evenly spaced knots on interval $[0, (d+1)\Delta x]$. Substituting the chosen polynomials $\{P_m(|y-x|)\}$ into (2), we obtain the learnt nonlocal surrogate model for the new microstructure, which is defined by a continuous nonlocal kernel

$$\gamma_{\theta_{\text{test}}}^\circ(|y-x|) := \sum_{m=1}^M (\mathbf{C}_{\text{test},m}^\circ) P_m^\circ(|y-x|). \tag{21}$$

Note that this kernel is indeed a twice differentiable function for $|y-x| \le \delta$, the resultant nonlocal surrogate model is therefore well-posed and defined in a continuous way.

Prediction error bounds

We now provide error bounds for MetaNOR based on the results for linear regression provided in.[11] Throughout this section, we use $||\mathbf{v}||_{l^2(\Omega)} := \sqrt{\Delta x \sum_{x_i \in \chi} v_i^2}$ to denote the domain-associated l^2 norm for a vector with values on χ. This norm can be seemed as a discretized approximation for the $L^2(\Omega)$ norm. Consider a solution pair $(\mathbf{C}^\circ, \mathbf{B}^\circ)$ which corresponds to a local minimizer of (18) and (20), we use $\mathcal{L}_{\theta_{\text{test}}}^\circ$ to denote the corresponding nonlocal operator generated from the learnt nonlocal kernel $\gamma_{\theta_{\text{test}}}^\circ$, and $\mathcal{L}_{\theta_{\text{test}},h}^\circ$ to denote its approximation by Riemann sum following (7).

We now provide the error estimates for the kernel estimator, $||\gamma_{\text{test}} - \gamma_{\theta_{\text{test}}}^\circ||_{l^2([0,\delta])}$, and for the prediction error. For the later, we consider a given time-dependent loading $g(x, t)$ with $x \in \Omega_{\text{pred}}$ and $t \in [0, T_{\text{pred}}]$, then use the learnt nonlocal model to predict the material response of our test microstructure, i.e., to provide an approximated displacement solution. Here, we stress that the prediction domain Ω_{pred} and time interval $[0, T_{\text{pred}}]$ may be different from the training datasets. We then discretize Ω_{pred} and $[0, T_{\text{pred}}]$ with grid sizes Δx and Δt, respectively, and denote the spatial grid set as χ_{pred}. For simplicity of analysis, here we take the same discretization sizes as those in the training dataset. However, since a continuous model is learnt, in practice one may employ different resolution or even discretization methods, as will be numerically demonstrated in the empirical experiment of "Verification on synthetic datasets" section . We consider δ as a physical parameter, i.e., as a fixed value, and hence $d = \Theta(\Delta x^{-1})$. Denoting $\hat{u}(\cdot, t^n)$ as the ground-truth HF solution subject to loading $g(x, t)$ and $\bar{u}^n(x_i)$ as the numerical solution satisfying

$$\ddot{\bar{u}}^n(x_i) - \mathcal{L}_{\theta_{\text{test}},h}^\circ[\bar{u}^n](x_i) = g(x_i, t^n) \tag{22}$$

for $x_i \in \chi_{\text{pred}}$ and $n = 1, \ldots, \lfloor T_{\text{pred}}/\Delta t \rfloor$, we aim to provide the error bound in the discretized energy norm for displacement prediction:

$$||\bar{u}^n - \hat{u}(\cdot, t^n)||_{E(\Omega)} := \sum_{i=1}^{L_{\text{pred}}} \left(\frac{e_i^n - e_i^{n-1}}{\Delta t}\right)^2 + 2\Delta x \sum_{\alpha=1}^d \sum_{i=1-\alpha}^{L_{\text{pred}}} (\mathbf{B}^\circ \mathbf{C}_{\text{test}}^\circ)_\alpha (e_{i+\alpha}^n - e_i^n)^2,$$
$$\text{where } e_i^n := \hat{u}(x_i, t^n) - \bar{u}^n(x_i).$$

We first detail three required assumptions for the analysis. In the following derivations, we always assume that the statements below are true, and therefore will not list them in the statement of theorems again.

Assumption 3.1 For both the training and test datasets, the vectors $\mathbf{s}_j$ are i.i.d. designed with zero mean, covariance $\mathbb{E}[\mathbf{s}\mathbf{s}^T] = \mathbf{I}_d$, and are $\mathbf{I}$-sub-Gaussian. The additive noise variables $\epsilon_j = y_j - \mathbf{s}_j^T \mathbf{B}\mathbf{C}_{\eta(j)}$ are also i.i.d. sub-Gaussian with variance parameter 1. Moreover, ϵ_j are independent of $\mathbf{s}_j$.

For the H training tasks, we define the population task diversity matrix and condition numbers as $\mathbf{T} = (\mathbf{C}_1, \ldots, \mathbf{C}_H)^T \in \mathbb{R}^{H \times M}$, $\nu := \sigma_M(\mathbf{T}^T\mathbf{T}/H)$ and $\kappa := \frac{1}{\nu}\sigma_1(\mathbf{T}^T\mathbf{T}/H)$.

Assumption 3.2 The H underlying task parameters $\mathbf{C}_j^*$ satisfies $||\mathbf{C}_j||_{l^2} = \Theta(1)$, i.e., they are asymptotically bounded below and above by constants. Moreover, the population task diversity matrix is well-conditioned, i.e., $\kappa \le O(1)$, which indicates that $\nu \ge \Omega(1/M)$.

Moreover, we make the following additional assumptions associated with the stability and consistency of the numerical scheme:

Assumption 3.3 The high-fidelity solution for our prediction task $\hat{u} \in C^2(\Omega \times [0, T_{\text{pred}}])$ and Δt is sufficiently small such that it satisfies $\Delta t \leq \min[(8\Delta x||\mathbf{B}^\circ\mathbf{C}^\circ||_{l^1})^{-1}, (2T_{\text{pred}})^{-1}]$. Moreover, the modeling error, $\epsilon(x,t)$, is bounded by a constant E for all $x \in \Omega_{\text{pred}}$ and $t \in [0, T_{\text{pred}}]$.

We now proceed to provide error bounds to our linear kernel representation learning setting.

Theorem 3.4 *Suppose we are given N_{train} total training datapoints from H diverse and normalized tasks, and N_{test} numbers of test datapoint on a new task with unknown microstructure. If the number of meta-train samples N_{train} satisfies $N_{\text{train}} \gtrsim \text{polylog}(N_{\text{train}}, d, H)(\kappa M)^4 \max\{H, d\}$, the number of meta-test samples N_{test} satisfies $N_{\text{test}} \gtrsim M \log(N_{\text{test}})$, and the optimal test microstructure satisfies $||\mathbf{C}^*_{\text{test}}||_{l^2} \leq O(1)$, then any local minimizer of (18) and the learnt kernel converges to the underlying optimal kernel with the following error bound*:

$$||\gamma_{\text{test}} - \gamma^\circ_{\theta_{\text{test}}}||^2_{l^2([0,\delta])} \leq \tilde{O}\left(\Delta x \frac{\max\{H,d\}M^2}{N_{\text{train}}} + \Delta x \frac{M}{N_{\text{test}}}\right),$$

and the corresponding approximated solution $\bar{u}$ has the following excess prediction error bound for $n = 1, \ldots, \lfloor T_{\text{pred}}/\Delta t \rfloor$:

$$||\bar{u}^n - \hat{u}(\cdot, t^n)||^2_{E(\Omega)} \leq \tilde{O}\Big(E^2 + \Delta t^4 + \Delta x^2 + \Delta x\Big[\frac{\max(H,d)M^2}{N_{\text{train}}} + \frac{M}{N_{\text{test}}}\Big]\Big),$$

with probability at least $1 - O((\text{poly}(d))^{-1} + N_{\text{test}}^{-100})$.

The proof is obtained by applying Theorems 2 and 4 in.[11] A more detailed proof is provided in Appendix.

Remark 1 This theorem indicates that when a sufficiently large training dataset is provided, for a new microstructure with very scarce measurements ($N_{\text{test}} \ll N_{\text{train}}/(\max(H,d)M)$), we have an approximated kernel error bound as $\tilde{O}\left(\left(\frac{M\Delta x}{N_{\text{test}}}\right)^{1/2}\right)$ and the energy error bound for the prediction task as $\tilde{O}\left(E + (\Delta t)^2 + \left(\frac{M\Delta x}{N_{\text{test}}}\right)^{1/2}\right)$. Hence, when the nonlocal model serves as a good surrogate for the material response, i.e., E is negligible, the estimator from MetaNOR provides a converging kernel and solution for further prediction tasks.

Physics-based constraints

As illustrated [6], when some physical knowledge is available, these knowledge can be incorporated into the optimization problem as physics-based constraints (11). In particular, when the effective wave speed for infinitely long wavelengths, c_0, is available, the corresponding constraint writes:

$$\int_0^\delta \xi^2 \gamma_\theta(|\xi|)\mathrm{d}\xi = \bar{\rho}c_0^2, \tag{23}$$

where $\bar{\rho}$ is the effective material density. Discretizing (23) by Riemann sum, we obtain the first constraint of $\{C_m\}$:

$$\bar{\rho}c_0^2 = \sum_{m=1}^{M} C_m \sum_{\alpha=1}^{d} \alpha^2 \Delta x^3 P_m(\alpha\Delta x) = \sum_{m=1}^{r} C_m A_{1m} \tag{24}$$

where $A_{1m} := \sum_{\alpha=1}^{d} \alpha^2 \Delta x^3 P_m(\alpha\Delta x)$. Furthermore, when the curvature of the dispersion curve in the low-frequency limit, R, is also available, the corresponding constraint writes:

$$\int_0^\delta \xi^4 \gamma_\theta(|\xi|)\mathrm{d}\xi = -4\bar{\rho}c_0^3 R. \tag{25}$$

Discretizing (25) yields the second constraint of $\{C_m\}$:

$$-4\bar{\rho}c_0^3 R = \sum_{m=1}^{M} C_m \sum_{\alpha=1}^{d} \alpha^4 \Delta x^5 P_m(\alpha\Delta x) = \sum_{m=1}^{M} C_m A_{2m} \tag{26}$$

where $A_{2m} := \sum_{\alpha=1}^{d} \alpha^4 \Delta x^5 P_m(\alpha\Delta x)$. Therefore, these two physics-based constraints are imposed as linear constraints for $\{C_m\}$. In empirical tests, we will refer to the experiments with these constraints applied as the "constraint" cases. In this work, we consider the heterogeneous bar composed by alternating layers of two dissimilar materials, with (averaged) layer size $L_1 = (1-\phi)L$, $L_2 = (1+\phi)L$ for components 1 and 2, respectively. Then the effective material density, Young's modulus, the wave speed, and the dispersion curvature are given by $\bar{\rho} = ((1-\phi)\rho_1 + (1+\phi)\rho_2)/2$, $\bar{E} = 2/((1-\phi)E_1^{-1} + (1+\phi)E_2^{-1})$, $c_0 = \sqrt{\bar{E}/\bar{\rho}}$, and $R = 0$. To apply (24) and (26), we reformulate the constraint optimization problem such that an unconstraint optimization problem of the form (8) is obtained. Detailed derivation is provided in the Appendix.

Empirical experiments

We evaluate MetaNOR on both synthetic and real-world datasets. On each dataset, we compare our MetaNOR approach with baseline NOR in (11). For NOR, we use linear regression with the L-curve method to select the proper regularization parameter ζ. In synthetic datasets, we generate the data from a known nonlocal diffusion equation and study the convergence of estimators to the true kernel. We also apply our method to a real-world dataset for stress wave propagation in 1D metamaterials.

In meta-training, we solve the optimization problem (18) using SciPy's L-BFGS-B optimization module. The maximum iteration step is set to 10000. In meta-testing, the solutions are solved with NumPy's linalg module.

Verification on synthetic datasets

We consider a synthetic dataset generated from a nonlocal diffusion equation

$$\mathcal{L}_{\gamma_\eta^+}[u_k](x) := \int_{B_\delta(x)} \gamma_\eta^+(|y-x|)(u_k(y)-u_k(x))\mathrm{d}y = g_k(x).$$

Here, η denotes the index of task, with $\eta \in \{1,\dots,8\}$. Each task is associated with a sine-type kernel:

$$\gamma_\eta^+(|y-x|) := \exp(-\eta(|y-x|))\sin(6|y-x|)\mathbf{1}_{[0,10]}(|y-x|),$$

with the estimated support of kernel as $\delta = 11$. To generate the training and test function pairs $(u_k(x), g_k(x))$, for each task the kernel acts on the same set of function $\{u_k\}_{k=1,2}$ with $u_1(x) = \sin(x)\mathbf{1}_{[-\pi,\pi]}(x)$ and $u_2(x) = \cos(x)\mathbf{1}_{[-\pi,\pi]}(x)$, and the loading function $\mathcal{L}_{\gamma_\eta^+}[u_k] = g_k$ is computed by the adaptive Gauss–Kronrod quadrature method, both on the computational domain $\Omega = [-40, 40]$. To create discrete datasets with different resolutions, we consider $\Delta x \in 0.0125 \times \{1, 2, 4, 8\}$. In meta-training, we use all samples from 7 "known tasks" $\eta \in \{1, 2, 3, 4, 6, 7, 8\}$. Then, the goal is to learn a good estimator for the "unknown" new task with $\eta = 5$.

Effect of low-dimensional feature selection: In this experiment we aim to verify the low-dimensional structure of the kernel space and select a proper value of M, with all test measurements employed, i.e., $N_{\text{test}} = 2 \times 80/\Delta x$. In Fig. 2(b) we demonstrate the learnt kernel for $M \in \{1, 2, 4, 7\}$ and $\Delta x \in \{0.0125, 0.1\}$, together with the averaged loss on all test samples (denoted as "loss") and the $l^2([0,\delta])$ errors for the kernel (denoted as "kernel error"). It is observed that the learnt kernel is visually consistent with the true kernel when $M \geq 4$. Hence in the following investigations we fix $M = 4$ for all cases.**Figure 2.** Problem settings and convergence study results for the MetaNOR verification on synthetic datasets.

Sampling efficiency on the new task: We now demonstrate the performance of the estimator in the small test measurement regime. We randomly select $N_{\text{test}} \in \{10, 20, 40, 80, 160, 320\}$ measurements from all available data on the test task, and study the convergence of the learnt kernel as N_{test} increases. In this experiment we fix $M = 4$ and $\Delta x = 0.0125$. To generate a fair comparison, the means and standard errors are calculated from 10 independent simulations. The errors of learnt kernels and the averaged loss on all test samples from MetaNOR and NOR are reported in Fig. 2(c). Averaged convergence rates are calculated on the relatively small data regime, i.e., for $N_{\text{test}} \leq 160$, as: rate $= \log_{16}[\text{error}(N_{\text{test}} = 10)/\text{error}(N_{\text{test}} = 160)]$. One can see that the kernel error from MetaNOR decreases almost linearly with the increase of N_{test} – a half order faster than the bound suggested in Theorem 3.4. This fact indicates a possible improvement of the analysis in the future work. On the other hand, NOR exhibits a much larger error and test loss in the same small test data regime, highlighting the advantage of our MetaNOR in sample efficiency.

Resolution independence and convergence We now study the performance of the estimator in terms of its convergence as the data mesh refines. Two types of experiments are designed, both with limited measurements ($N_{\text{test}} = 320$). First, we keep the same resolutions (Δx) in all tasks, to study the convergence of estimators to the true kernel as Δx decreases. Additionally, to verify the consistency of estimators across different resolutions, we further investigate their performances when the training tasks and test tasks have different resolutions. In Fig. 2(d) the kernel errors and test losses are reported, as functions of Δx in the test case (denoted as Δx_{test}). For the first study, a 0.88 order convergence is observed, which is consistent with the error bound from Theorem 3.4. For the second study, one can see that no matter if we extract the features from a relatively coarse grid ($\Delta x_{\text{train}} = 0.1$) or a fine grid ($\Delta x_{\text{train}} = 0.0125$), the resultant estimator on the test task pertains a similar accuracy or even achieves convergence as the test grid size Δx_{test} refines, when learning from fine measurements. These results highlight the advantage of our method on learning the kernel and the corresponding continuous nonlocal operator instead of learning the solution: the resultant model is not tied to the input's resolution.

Application to wave propagation in metamaterials

We now apply MetaNOR to model the propagation of stress waves in one-dimensional metamaterials. Two experiments are considered:

1. [*Varying Disorder Parameter, see Fig.* 3(a)] We aim to transfer the knowledge between different disordered microstructures, where the size of each layer is defined by a random variable. For component 1, the layer size $L_1 \sim \mathcal{U}[(1-D)(1-\phi)L, (1+D)(1-\phi)L]$, and for component 2 the layer size $L_2 \sim \mathcal{U}[(1-D)(1+\phi)L, (1+D)(1+\phi)L]$. Here $D \in [0.05, 0.5]$ is the disorder parameter for each task, and the Young's modulus $E_1 = 1$ and $E_2 = 0.25$ are fixed. For this experiment we train with 9 microstructures, then test the meta-learned parameter on a new microstructure with $D = 0.3$.
2. [*Varying Young's Modulus, see Fig.* 3(d)] We aim to transfer the knowledge between varying components. Periodic layers are considered ($D = 0$) with fixed Young's modulus $E_1 = 1$ in component 1 and varying Young's modulus ($E_2 \in [0.2025, 0.7225]$) in component 2. For this experiment we train with 8 microstructures, then test the meta-learned parameter on an new microstructure with $E_2 = 0.25$.

The HF dataset we rely on is generated by a classical wave solver, where all material interfaces are treated explicitly, and therefore small time and step discretization sizes are required. This solver and its results will be referred to as Direct Numerical Solution (DNS). In all training data, we set $L = 0.2$, $\rho_1 = \rho_2 = 1$, $\phi = 0$, and the domain $\Omega = [-50, 50]$. Following the settings in,[6] two types of data, including 20 simulations from the oscillating source dataset and 20 simulations from the plane wave dataset, are generated for meta-training and meta-test in each task. Parameters for the training and the optimization algorithm are set

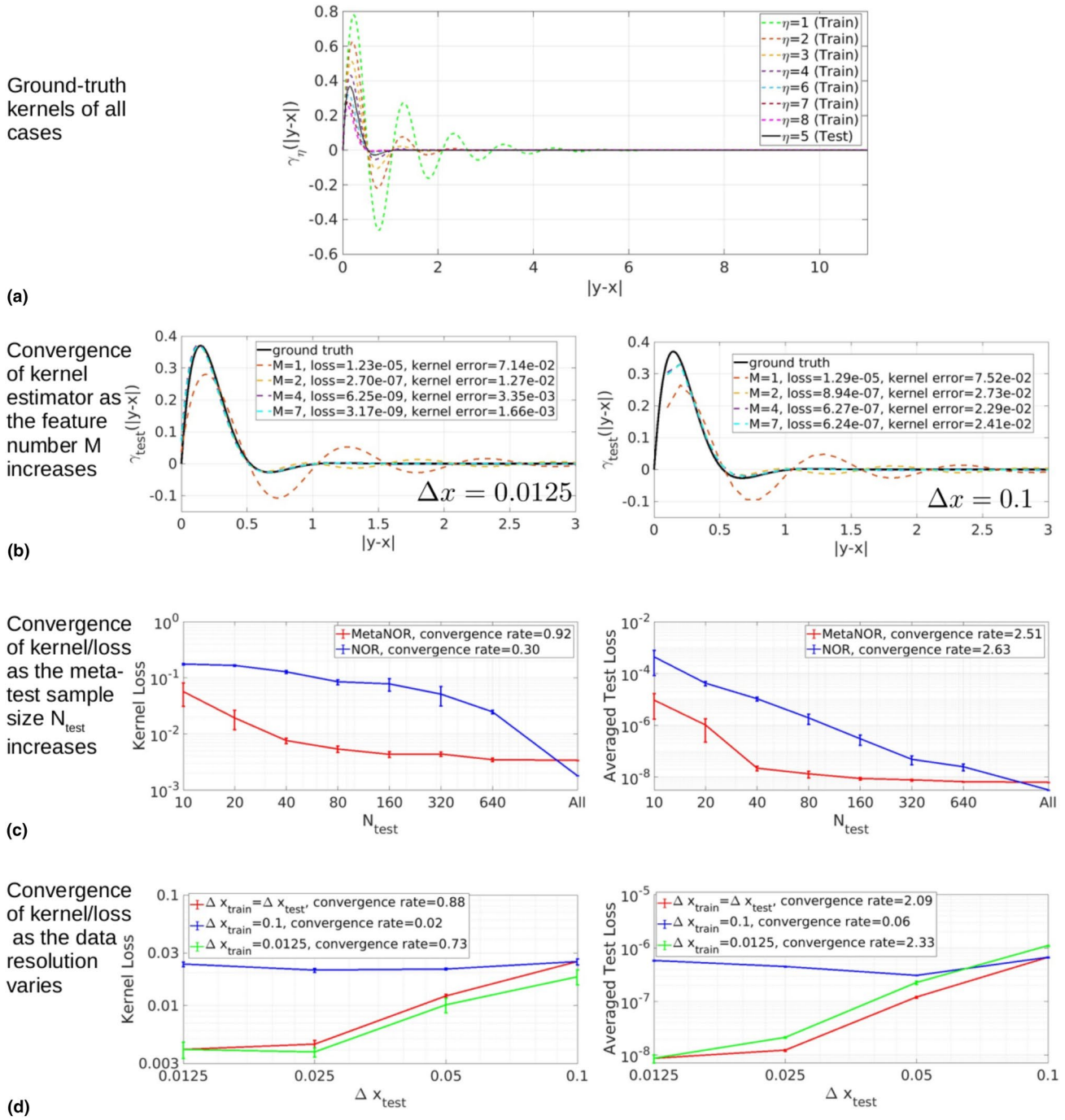


Figure 2. Problem settings and convergence study results for the MetaNOR verification on synthetic datasets.

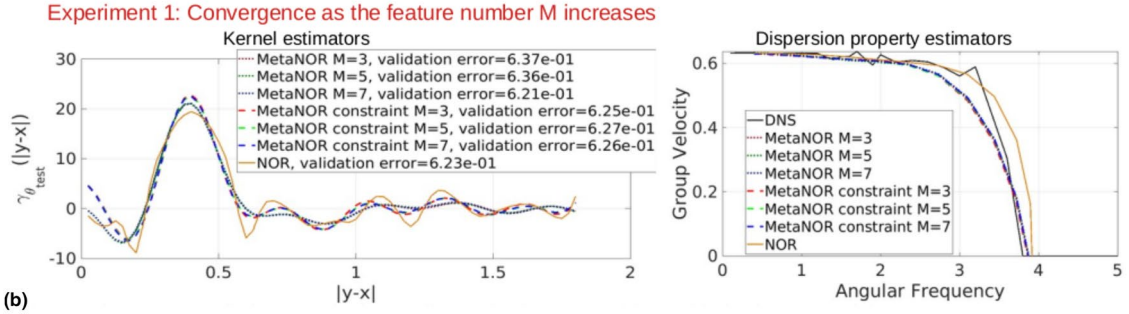
to $\Delta x \in \{0.025, 0.05, 0.1\}$, $\Delta t = 0.02$, $T = 2$ and $\delta = 1.8$. Additionally, we create two validation datasets, denoted as the wave packet dataset and the projectile impact dataset respectively, both very different from the meta-training and meta-test datasets. They consider a much longer bar ($\Omega_{wp} = [-133.3, 133.3]$ for wave packet and $\Omega_{impact} = [-267, 267]$ for impact), under a different loading condition from the training dataset, and with a much longer simulation time ($T_{wp} = 100$ and $T_{impact} = 600$). Full details are provided in Appendix.

Comparison metrics Notice that in this case, the data is not faithful to the nonlocal model, but generated from an HF model with microscale details. Therefore, there is no ground-truth kernel and we demonstrate the performance of estimators by studying their capability of reproducing the dispersion relation and the wave motion on the two validation datasets, and compare them with the results computed with DNS. The dispersion curve provides the group velocity profile as a function of frequency for each microstructure, which directly depicts the dispersion properties in

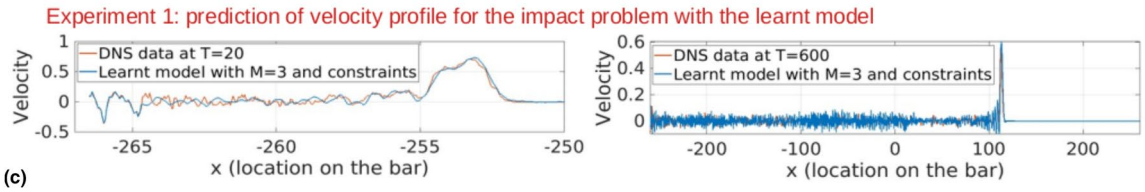
Experiment 1 settings: varying disorder parameter D

Task	Train 1	Train 2	Train 3	Train 4	Train 5	Train 6	Train 7	Train 8	Train 9	Test
D	0.05	0.10	0.15	0.20	0.25	0.35	0.40	0.45	0.50	0.30

(a)



(b)

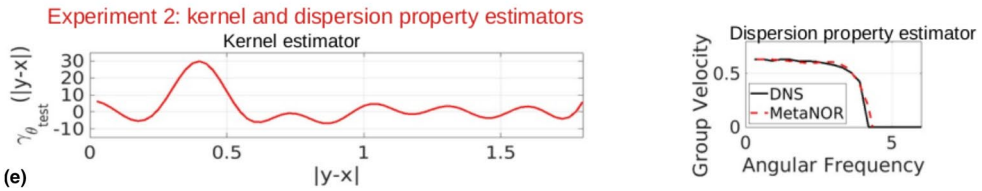


(c)

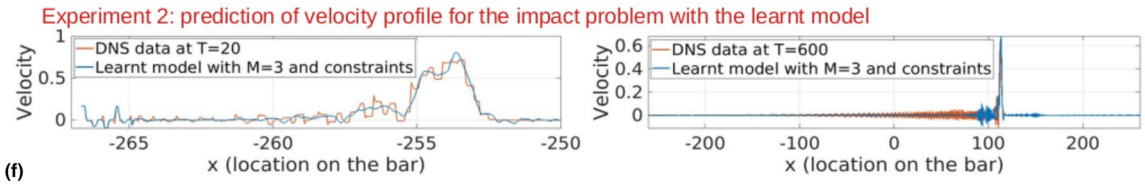
Experiment 2 settings: varying Young's modulus

Task	Train 1	Train 2	Train 3	Train 4	Train 5	Train 6	Train 7	Train 8	Test
E_2	0.2025	0.3025	0.36	0.4225	0.49	0.5625	0.64	0.7225	0.25

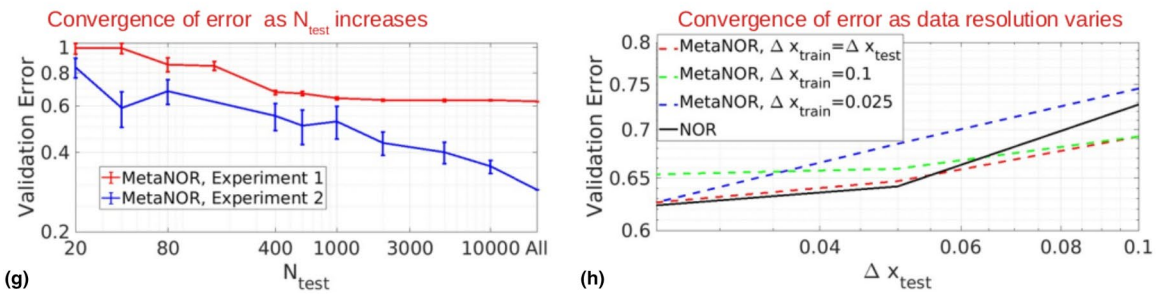
(d)



(e)



(f)



(g) (h)

Figure 3. Problem settings and numerical results for the MetaNOR application to wave propagation modeling problem in 1D metamaterials.

this microstructure. We further report the relative prediction error in the discrete energy norm, $\left|\left|\bar{u}^n - \hat{u}(\cdot, t^n)\right|\right|_{E(\Omega)} / \left|\left|\hat{u}(\cdot, t^n)\right|\right|_{E(\Omega)}$, on the wave packet dataset. Last, we use the learnt kernel to perform long-term prediction tasks on the projectile impact dataset, to validate the model stability and generalizability.

Model validation To investigate the low-dimensional structure of kernel space, in Fig. 3(b) we report the estimated kernels in experiment 1, their corresponding group velocities, and validation errors for $\Delta x = 0.025$ and $M \in \{3, 5, 7\}$. We can observe that, while all MetaNOR models have successfully reproduces the DNS dispersion relation, the "constraint" cases have achieved a better prediction accuracy comparing with the ones without physical constraints. Hence, in the following studies we mainly focus on "constraint" cases. We employ the constraint model with $M = 3, \Delta x = 0.05$, to predict the short-term ($T = 20$) and long-term ($T = 600$) velocity profiles subject to projectile impact, and report the results in Fig. 3(c). The results are consistent with DNS simulations, verifying that our optimal kernel can accurately predict the short- and long-time wave propagation. We then perform similar tests in experiment 2, with the predicted kernel, dispersion relation, and wave propagation prediction results provided in Fig. 3(e) and (f). All these results indicate that there exists a common set of low-dimensional features for all microstructures, and MetaNOR provides a good surrogate model based on these features in the low-dimensional kernel space.

Sample efficiency We now consider both experiment settings with $M = 3$ and $\Delta x = 0.025$, and randomly pick $N_{\text{test}} \in [10, 10^4]$ numbers of datapoints on the new and unseen microstructure. For each N_{test}, we repeat the experiment for 10 times, to plot the mean and standard error of results. Note that under this scarce sample setting the estimated model from NOR gets unstable and fails the prediction task. Hence we only report the MetaNOR results. From Fig. 3(g) we can see that as the number of test sample increases, for both experiments the validation error decreases. Notice that because of the unavoidable modeling error due to the discrepancy between nonlocal surrogates and the HF model, as shown in Theorem 3.4, one should not expect the prediction error to converge to zero. This result again verifies the robustness of MetaNOR in the small data regime.

Resolution independence and convergence Lastly, we consider experiment setting 1 with $M = 7$, to study the performance of the estimator in terms of its convergence as the data mesh refines. In Fig. 3(h) the validation errors are reported for different combinations of Δx_{train} and Δx_{test}. The learnt estimator again demonstrates an improved accuracy as the data resolution refines.

Conclusion

We proposed a meta-learning approach for the nonlocal operator regression, by taking advantages of a common set of low-dimensional features in a multi-task setting for accurate and efficient adaption to new unseen tasks. Specifically, we reformulate the nonlocal operator regression as a linear kernel regression problem and propose MetaNOR as a linear kernel feature learning algorithm with provable guarantees. We apply such a method to metamaterial problems and show the superior transfer capability, showing meta-learning is a promising direction for heterogeneous material discovery. Future work could extend our method to obtain sharper estimates and apply to more general material types.

Acknowledgments

The authors would also like to thank Dr. Stewart Silling for sharing the DNS codes and for the helpful discussions.

Author contributions

LZ: methodology, validation, formal analysis, investigation, data curation, visualization; HY: methodology, investigation, data curation; YY: conceptualization, resources, funding acquisition, writing, supervision, project administration.

Funding

Authors are supported by the National Science Foundation under award DMS 1753031, and the AFOSR grant FA9550-22-1-0197. Portions of this research were conducted on Lehigh University's Research Computing infrastructure partially supported by NSF Award 2019035.

Data availability

The data that support the findings of this study are available from the corresponding author, Yue Yu, upon reasonable request.

Declarations

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix A: Related works

Material discovery

Using machine learning techniques for material discovery is gaining more attention in scientific communities.[17–20] They has been applied to materials such as thermoelectric material,[21] metallic glasses,[22] high-entropy ceramics,[23] and so on. Learning models for metamaterials has also gained popularity with recent approaches such as.[9] For a comprehensive review on the application of machine learning techniques to property prediction and materials development for energy-related fields, we refer interested readers to.[24].

Meta-learning

Meta-learning seeks to design algorithms that can utilize previous experience to rapidly learn new skills or adapt to new environments. There is a vast literature on papers proposing meta-learning[25] methods, and they have been applied to patient survival analysis,[26] few short image classification,[27] and natural language processing,[28] just to name a few. Recently, provably generalizable algorithms with sharp guarantees in the linear setting are first provided.[11]

Transfer and meta-learning for material modeling

Despite its popularity, few work has studied material discovery under meta or even transfer setting.[29] proposes a transfer-learning technique to exploit correlation among different material properties to augment the features with predicted material properties to improve the regression performance.[30] uses an ensemble of model and a meta-model to help discovering candidate water splitting photocatalysts. To the best of our knowledge, our work is the first application of transfer or meta-learning to heterogeneous material homogenization and discovery.

.

Appendix B: Detailed proof for the error bounds

In this section we review two main lemmas from,[11] which provide a theoretical prediction error bound for the meta-learning of linear representation model as illustrated in (16) and (17). Then we employ these two lemmas and detailed the proof of Theorem 3.4, which provides the error bound for the meta-learning of kernel representations and the resultant prediction tasks.

Lemma B.1 (11, Theorem 2) *Assume that we are in a uniform task sampling model. If the number of* meta-train samples N_{train} *satisfies* $N_{\text{train}} \gtrsim \text{polylog}(N_{\text{train}}, d, H)(\kappa M)^4 \max\{H, d\}$ *and given any local minimum of the optimization objective (18), the column space of* $\mathbf{V}^*$*, spanned by the orthonormal feature matrix* $\mathbf{B}^\circ$ *satisfies*

$$\sin\theta(\mathbf{B}^\circ, \mathbf{B}) \le O\left(\sqrt{\frac{\max\{H,d\}M\log N_{\text{train}}}{\nu N_{\text{train}}}}\right), \quad \text{(B.1)}$$

with probability at least $1 - 1/\text{poly}(d)$.

Note that Assumption 3.2 guarantees that $\nu \ge \Omega(1/M)$ and the above theorem yields

$$\sin\theta(\mathbf{B}^\circ, \mathbf{B}) \le \tilde{O}\left(\sqrt{\frac{\max\{H,d\}M^2}{N_{\text{train}}}}\right), \quad (2)$$

with probability at least $1 - 1/\text{poly}(d)$.

Lemma B.2 (11, Theorem 4) *Suppose the parameter associated with the new task satisfies* $||\mathbf{C}_{\text{test}}||_{l^2} \le O(1)$*, then if an estimate* $\mathbf{B}^\circ$ *of the true feature matrix* $\mathbf{B}$ *satisfies* $\sin\theta(\mathbf{B}^\circ, \mathbf{B}) \le \varpi$ *and* $N_{\text{test}} \gtrsim M\log N_{\text{test}}$*, then the output parameter* $\mathbf{C}^\circ_{\text{test}}$ *from (20) satisfies*

$$\left|\left|\mathbf{B}^\circ\mathbf{C}^\circ_{\text{test}} - \mathbf{B}\mathbf{C}_{\text{test}}\right|\right|^2_{l^2} \le \tilde{O}\left(\varpi^2 + \frac{M}{N_{\text{test}}}\right), \quad (3)$$

with probability at least $1 - O(N_{\text{test}}^{-100})$.

Combining Lemma B.2 with Lemma B.1, we obtain the following result for applying (18) and (20) as a linear feature meta-learning algorithm:

$$\left|\left|\mathbf{B}^\circ\mathbf{C}^\circ_{\text{test}} - \mathbf{B}\mathbf{C}_{\text{test}}\right|\right|^2_{l^2} \le \tilde{O}\left(\frac{\max\{H,d\}M^2}{N_{\text{train}}} + \frac{M}{N_{\text{test}}}\right), \quad (4)$$

with probability at least $1 - O((\text{poly}(d))^{-1} + N_{\text{test}}^{-100})$.

We now proceed to provide the proof for Theorem 3.4. In the following, we use C to denote a generic constant which is independent of Δx, Δt, M, H, N_{train} and N_{test}, but might depend on δ.

Proof With (4), we immediately obtain the $l^2([0,\delta])$ error estimate for the learnt kernel $\gamma^\circ_{\text{test}}$ as

$$\begin{aligned}\left|\left|\gamma_{\text{test}} - \gamma^\circ_{\theta_{\text{test}}}\right|\right|^2_{l^2([0,\delta])} &= \Delta x\sum_{\alpha=1}^{d}\left(\sum_{m=1}^{M}(\mathbf{C}^\circ_{\text{test}} - \mathbf{C}_{\text{test}})_m P_m(|\alpha\Delta x|)\right)^2 \\ &= \Delta x\left|\left|\mathbf{B}\mathbf{C}_{\text{test}} - \mathbf{B}^\circ\mathbf{C}^\circ_{\text{test}}\right|\right|^2_{l^2} \\ &= \tilde{O}\left(\Delta x\frac{\max\{H,d\}M^2}{N_{\text{train}}} + \Delta x\frac{M}{N_{\text{test}}}\right),\end{aligned}$$

with probability at least $1 - O((\text{poly}(d))^{-1} + N_{\text{test}}^{-100})$.

For the error bound in the discretized energy norm, we notice that the ground-truth solution $\hat{u}$ satisfies:

$$\begin{aligned}\ddot{\hat{u}}(x_i, t^n) &= \mathcal{L}_{\theta_{\text{test}},h}[\hat{u}](x_i, t^n) + g(x_i, t^n) \\ &\quad + \epsilon(x_i, t^n) + \left[\ddot{\hat{u}}(x_i, t^n) - \frac{\partial^2\hat{u}}{\partial t^2}(x_i, t^n)\right] \\ &\quad + \left[\mathcal{L}_{\theta_{\text{test}}}[\hat{u}](x_i, t^n) - \mathcal{L}_{\theta_{\text{test}},h}[\hat{u}](x_i, t^n)\right]\end{aligned}$$

for all $x \in \chi_{\text{pred}}$, $n = 1, \ldots, \lfloor T_{\text{pred}}/\Delta t\rfloor$. Subtracting this equation with (22) and denoting $e_i^n := \hat{u}(x_i, t^n) - \bar{u}^n(x_i)$, we then obtain

$$\begin{aligned}\frac{e_i^{n+1} - 2e_i^n + e_i^{n-1}}{\Delta t^2} &= \Delta x\sum_{\alpha=1}^{d}(\mathbf{B}^\circ\mathbf{C}^\circ_{\text{test}})_\alpha(e^n_{i+\alpha} \\ &\quad + e^n_{i-\alpha} - 2e_i^n) + (\epsilon_{\text{all}})_i^n,\end{aligned} \quad (5)$$

where

$$\begin{aligned}(\epsilon_{\text{all}})_i^n :&= \epsilon(x_i, t^n) + \Delta x\sum_{\alpha=1}^{d}(\mathbf{B}\mathbf{C}_{\text{test}} - \mathbf{B}^\circ\mathbf{C}^\circ_{\text{test}})_\alpha(\hat{u}(x_{i+\alpha}, t^n) \\ &\quad + \hat{u}(x_{i-\alpha}, t^n) - 2\hat{u}(x_i, t^n)) \\ &\quad + \left[\ddot{\hat{u}}(x_i, t^n) - \frac{\partial^2\hat{u}}{\partial t^2}(x_i, t^n)\right] + \left[\mathcal{L}_{\theta_{\text{test}}}[\hat{u}](x_i, t^n)\right. \\ &\quad \left.-\mathcal{L}_{\theta_{\text{test}},h}[\hat{u}](x_i, t^n)\right].\end{aligned}$$

With Assumption 3.3, we have the truncation error for the Riemann sum part as $\left|\mathcal{L}_{\theta_{\text{test}}}\hat{u}(x,t) - \mathcal{L}_{\theta_{\text{test}},h}\hat{u}(x,t)\right| \le C\Delta x$ for a constant C independent of Δx and Δt but might depends on δ.

Similarly, we have the truncation error for the central difference scheme as $\left|\ddot{\hat{u}}(x_i,t^n)-\frac{\partial^2\hat{u}}{\partial t^2}(x_i,t^n)\right|\le C(\Delta t)^2$ with the constant C independent of Δx, Δt, and δ. Moreover, (4) yields

$$\begin{aligned}
&\left|\Delta x\sum_{\alpha=1}^{d}(\mathbf{B}\mathbf{C}_{\text{test}}-\mathbf{B}^{\circ}\mathbf{C}^{\circ}_{\text{test}})_{\alpha}(\hat{u}(x_{i+\alpha},t^n)+\hat{u}(x_{i-\alpha},t^n)-2\hat{u}(x_i,t^n))\right|\\
&\le \Delta x\left|\sum_{\alpha=1}^{d}(\mathbf{B}\mathbf{C}_{\text{test}}-\mathbf{B}^{\circ}\mathbf{C}^{\circ}_{\text{test}})_{\alpha}(\alpha\Delta x)^2\max_{(x,t)\in\Omega_{\text{pred}}\times[0,T_{\text{pred}}]}\left|\frac{\partial^2\hat{u}}{\partial x^2}\right|\right|\\
&\le \Delta x\delta^2\sum_{\alpha=1}^{d}\left|(\mathbf{B}\mathbf{C}_{\text{test}}-\mathbf{B}^{\circ}\mathbf{C}^{\circ}_{\text{test}})_{\alpha}\right|\max_{(x,t)\in\Omega_{\text{pred}}\times[0,T_{\text{pred}}]}\left|\frac{\partial^2\hat{u}}{\partial x^2}\right|\\
&\le \Delta x\delta^2\sqrt{d}\left|\left|\mathbf{B}\mathbf{C}_{\text{test}}-\mathbf{B}^{\circ}\mathbf{C}^{\circ}_{\text{test}}\right|\right|_{l^2}\max_{(x,t)\in\Omega_{\text{pred}}\times[0,T_{\text{pred}}]}\left|\frac{\partial^2\hat{u}}{\partial x^2}\right|\\
&\le \tilde{O}\left(\sqrt{\frac{\max\{H,d\}M^2}{N_{\text{train}}}+\frac{M}{N_{\text{test}}}}\right)\sqrt{\Delta x\delta^5}\max_{(x,t)\in\Omega_{\text{pred}}\times[0,T_{\text{pred}}]}\left|\frac{\partial^2\hat{u}}{\partial x^2}\right|,
\end{aligned}$$

with probability at least $1-O((\text{poly}(d))^{-1}+N_{\text{test}}^{-100})$. Hence we have the bound for ϵ_{all}:

$$\left|(\epsilon_{\text{all}})_i^n\right|\le E+\tilde{O}\left(\Delta x+(\Delta t)^2+\sqrt{\left(\frac{\max\{H,d\}M^2}{N_{\text{train}}}+\frac{M}{N_{\text{test}}}\right)\Delta x}\right).$$

To show the $l^2(\Omega)$ error for e_i^n, we first derive a bound for its error in the (discretized) energy norm. Multiplying (5) with $\frac{e_i^{n+1}-e_i^n}{\Delta t}$ and summing over $\chi_{\text{pred}}=\{x_i\}_{i=1}^{L_{\text{pred}}}$ yields:

$$\begin{aligned}
&\sum_{i=1}^{L_{\text{pred}}}\frac{(e_i^{n+1}-2e_i^n+e_i^{n-1})(e_i^{n+1}-e_i^n)}{\Delta t^3}\\
&=\frac{\Delta x}{\Delta t}\sum_{i=1}^{L_{\text{pred}}}\sum_{\alpha=1}^{d}(\mathbf{B}^{\circ}\mathbf{C}^{\circ}_{\text{test}})_{\alpha}(e_{i+\alpha}^n+e_{i-\alpha}^n-2e_i^n)(e_i^{n+1}-e_i^n)\\
&\quad+\frac{1}{\Delta t}\sum_{i=1}^{L_{\text{pred}}}(\epsilon_{\text{all}})_i^n(e_i^{n+1}-e_i^n).
\end{aligned}$$

With the formulation $a(a-b)=\frac{1}{2}(a^2-b^2+(a-b)^2)$, we can rewrite the left hand side as

$$\begin{aligned}
&\sum_{i=1}^{L_{\text{pred}}}\frac{(e_i^{n+1}-2e_i^n+e_i^{n-1})(e_i^{n+1}-e_i^n)}{\Delta t^3}\\
&\ge\frac{1}{2\Delta t}\sum_{i=1}^{L_{\text{pred}}}\left[\left(\frac{e_i^{n+1}-e_i^n}{\Delta t}\right)^2-\left(\frac{e_i^n-e_i^{n-1}}{\Delta t}\right)^2\right.\\
&\quad\left.+\left(\frac{e_i^{n+1}-2e_i^n+e_i^{n-1}}{\Delta t}\right)^2\right]\\
&\ge\frac{1}{2\Delta t}\sum_{i=1}^{L_{\text{pred}}}\left[\left(\frac{e_i^{n+1}-e_i^n}{\Delta t}\right)^2-\left(\frac{e_i^n-e_i^{n-1}}{\Delta t}\right)^2\right].
\end{aligned}$$

For the first term on the right hand side, with the formulations

$$\begin{aligned}
\sum_{i=1-\alpha}^{L}a_i(b_{i+\alpha}-b_i)&=\sum_{i=1}^{\alpha}a_{L+i}b_{L+i}\\
&\quad-\sum_{i=1}^{\alpha}a_{i-\alpha}b_{i-\alpha}-\sum_{i=1-\alpha}^{L}b_{i+\alpha}(a_{i+\alpha}-a_i),
\end{aligned}$$

$a(b-a)=\frac{1}{2}(b^2-a^2-(a-b)^2)$, Assumption 3.3, and the exact Dirichlet-type boundary condition, i.e., $e_i^n=0$ for $i<1$ and $i>L_{\text{pred}}$, we have

$$\begin{aligned}
&\frac{\Delta x}{\Delta t}\sum_{i=1}^{L_{\text{pred}}}\sum_{\alpha=1}^{d}(\mathbf{B}^{\circ}\mathbf{C}^{\circ}_{\text{test}})_{\alpha}(e_{i+\alpha}^n+e_{i-\alpha}^n-2e_i^n)(e_i^{n+1}-e_i^n)\\
&=-\frac{\Delta x}{\Delta t}\sum_{\alpha=1}^{d}\sum_{i=1-\alpha}^{L_{\text{pred}}}(\mathbf{B}^{\circ}\mathbf{C}^{\circ}_{\text{test}})_{\alpha}(e_{i+\alpha}^n-e_i^n)\\
&\quad(e_{i+\alpha}^{n+1}-e_{i+\alpha}^n-e_i^{n+1}+e_i^n)\\
&=-\frac{\Delta x}{2\Delta t}\sum_{\alpha=1}^{d}\sum_{i=1-\alpha}^{L_{\text{pred}}}(\mathbf{B}^{\circ}\mathbf{C}^{\circ}_{\text{test}})_{\alpha}\\
&\quad\left[(e_{i+\alpha}^{n+1}-e_i^{n+1})^2-(e_{i+\alpha}^n-e_i^n)^2\right.\\
&\quad\left.-(e_{i+\alpha}^{n+1}-e_{i+\alpha}^n-e_i^{n+1}+e_i^n)^2\right]\\
&\le-\frac{\Delta x}{2\Delta t}\sum_{\alpha=1}^{d}\sum_{i=1-\alpha}^{L_{\text{pred}}}(\mathbf{B}^{\circ}\mathbf{C}^{\circ}_{\text{test}})_{\alpha}\\
&\quad\left[(e_{i+\alpha}^{n+1}-e_i^{n+1})^2-(e_{i+\alpha}^n-e_i^n)^2\right]\\
&\quad+\frac{\Delta x}{\Delta t}\sum_{\alpha=1}^{d}\sum_{i=1-\alpha}^{L_{\text{pred}}}(\mathbf{B}^{\circ}\mathbf{C}^{\circ}_{\text{test}})_{\alpha}\\
&\quad\left[(e_{i+\alpha}^{n+1}-e_{i+\alpha}^n)^2+(e_i^{n+1}-e_i^n)^2\right]\\
&\le-\frac{\Delta x}{2\Delta t}\sum_{\alpha=1}^{d}\sum_{i=1-\alpha}^{L_{\text{pred}}}(\mathbf{B}^{\circ}\mathbf{C}^{\circ}_{\text{test}})_{\alpha}\\
&\quad\left[(e_{i+\alpha}^{n+1}-e_i^{n+1})^2-(e_{i+\alpha}^n-e_i^n)^2\right]\\
&\quad+2\frac{\Delta x}{\Delta t}\sum_{i=1}^{L_{\text{pred}}}\left|\left|\mathbf{B}^{\circ}\mathbf{C}^{\circ}_{\text{test}}\right|\right|_{l^1}(e_i^{n+1}-e_i^n)^2\\
&\le-\frac{\Delta x}{2\Delta t}\sum_{\alpha=1}^{d}\sum_{i=1-\alpha}^{L_{\text{pred}}}(\mathbf{B}^{\circ}\mathbf{C}^{\circ}_{\text{test}})_{\alpha}\\
&\quad\left[(e_{i+\alpha}^{n+1}-e_i^{n+1})^2-(e_{i+\alpha}^n-e_i^n)^2\right]\\
&\quad+\frac{1}{4}\sum_{i=1}^{L_{\text{pred}}}\left(\frac{e_i^{n+1}-e_i^n}{\Delta t}\right)^2.
\end{aligned}$$

For the second term on the right hand side we have

$$\frac{1}{\Delta t}\sum_{i=1}^{L_{\text{pred}}}(\epsilon_{\text{all}})_i^n(e_i^{n+1}-e_i^n) \le \sum_{i=1}^{L_{\text{pred}}}((\epsilon_{\text{all}})_i^n)^2+\frac{1}{4}\sum_{i=1}^{L_{\text{pred}}}\left(\frac{e_i^{n+1}-e_i^n}{\Delta t}\right)^2.$$

Putting the above three inequalities together, we obtain

$$\begin{aligned}&\sum_{i=1}^{L_{\text{pred}}}\left[(1-\Delta t)\left(\frac{e_i^{n+1}-e_i^n}{\Delta t}\right)^2-\left(\frac{e_i^n-e_i^{n-1}}{\Delta t}\right)^2\right]\\&+2\Delta x\sum_{\alpha=1}^{d}\sum_{i=1-\alpha}^{L_{\text{pred}}}(\mathbf{B}^\circ\mathbf{C}^\circ_{\text{test}})_\alpha\left[(e_{i+\alpha}^{n+1}-e_i^{n+1})^2-(e_{i+\alpha}^n-e_i^n)^2\right]\\&\le 2\Delta t\sum_{i=1}^{L_{\text{pred}}}((\epsilon_{\text{all}})_i^n)^2.\end{aligned}$$

With the discrete Gronwall lemma and the bound of Δt in Assumption 3.3, for $n=1,\dots,\lfloor T_{\text{pred}}/\Delta t\rfloor$ we have

$$\begin{aligned}&\sum_{i=1}^{L_{\text{pred}}}\left(\frac{e_i^n-e_i^{n-1}}{\Delta t}\right)^2+2\Delta x\sum_{\alpha=1}^{d}\sum_{i=1-\alpha}^{L_{\text{pred}}}(\mathbf{B}^\circ\mathbf{C}^\circ_{\text{test}})_\alpha(e_{i+\alpha}^n-e_i^n)^2\\&\le 2L_{\text{pred}}((1-\Delta t)^{-n}-1)\max_{i,n}\\&\quad|(\epsilon_{\text{all}})_i^n|^2\le 4\exp(T_{\text{pred}})L_{\text{pred}}\max_{i,n}|(\epsilon_{\text{all}})_i^n|^2\\&\le L_{\text{pred}}\tilde{O}\left(E^2+(\Delta x)^2+(\Delta t)^4+\left(\frac{\max\{H,d\}M^2}{N_{\text{train}}}+\frac{M}{N_{\text{test}}}\right)\Delta x\right),\end{aligned}$$

with probability at least $1-O((\text{poly}(d))^{-1}+N_{\text{test}}^{-100})$, which provides the error bound in the discrete energy norm. □

Appendix C: Reduction of two physics constraints

In this section we further expend the discussion on physics-based constraints in "Prediction error bounds" section. The overall strategy is to fix the last two polynomial features,

$$P_{M-1}(\xi)=\beta_1:=\left(\sum_{\alpha=1}^{d}\alpha^2\Delta x^3\right)^{-1}$$

and

$$P_M(\xi)=\beta_2\xi:=\left(\sum_{\alpha=1}^{d}\alpha^3\Delta x^4\right)^{-1}\xi$$

into the set of basis polynomials. We note that these two polynomials satisfy

$$\sum_{\alpha=1}^{d}\alpha^2\Delta x^3P_{M-1}(\alpha\Delta x)=1,\ \sum_{\alpha=1}^{d}\alpha^2\Delta x^3P_M(\alpha\Delta x)=1,$$

and

$$\sum_{\alpha=1}^{d}\alpha^4\Delta x^5P_{M-1}(\alpha\Delta x)=\frac{\sum_{\alpha=1}^{d}\alpha^4\Delta x^2}{\sum_{\alpha=1}^{d}\alpha^2},\ \sum_{\alpha=1}^{d}\alpha^4\Delta x^5P_M(\alpha\Delta x)=\frac{\sum_{\alpha=1}^{d}\alpha^5\Delta x^2}{\sum_{\alpha=1}^{d}\alpha^3}.$$

Then (24) writes

$$\bar\rho c_0^2=\sum_{m=1}^{M-2}C_mA_{1m}+C_{M-1}+C_M,$$

and (26) writes

$$-4\bar\rho c_0^3R=\sum_{m=1}^{M-2}C_mA_{2m}+\frac{\sum_{\alpha=1}^{d}\alpha^4\Delta x^2}{\sum_{\alpha=1}^{d}\alpha^2}C_{M-1}+\frac{\sum_{\alpha=1}^{d}\alpha^5\Delta x^2}{\sum_{\alpha=1}^{d}\alpha^3}C_M.$$

Denoting

$$\Lambda:=\begin{bmatrix}1 & 1\\ \frac{\sum_{\alpha=1}^{d}\alpha^4\Delta x^2}{\sum_{\alpha=1}^{d}\alpha^2} & \frac{\sum_{\alpha=1}^{d}\alpha^5\Delta x^2}{\sum_{\alpha=1}^{d}\alpha^3}\end{bmatrix},$$

and

$$\mathbf{H}:=\begin{bmatrix}\Delta x^2 & 4\Delta x^2 & \cdots & d^2\Delta x^2\\ \Delta x^4 & 16\Delta x^4 & \cdots & d^4\Delta x^4\end{bmatrix},$$

then

$$\begin{aligned}\begin{bmatrix}C_{M-1}\\ C_M\end{bmatrix}&=\Lambda^{-1}\begin{bmatrix}\bar\rho c_0^2-\sum_{m=1}^{M-2}C_mA_{1m}\\ -4\bar\rho c_0^3R-\sum_{m=1}^{M-2}C_mA_{2m}\end{bmatrix}\\&=\Lambda^{-1}\begin{bmatrix}\bar\rho c_0^2\\ -4\bar\rho c_0^3R\end{bmatrix}-\Delta x\Lambda^{-1}\mathbf{HBC}.\end{aligned}$$

Substituting this equation into the loss function in (8), for each x_i we obtain

$$\begin{aligned}&\left(y_{k,i}^n-(\mathbf{s}_{k,i}^n)^T\mathbf{BC}\right)^2\\&=\left(y_{k,i}^n-(\mathbf{s}_{k,i}^n)^T\left(\sum_{m=1}^{M-2}C_m\mathbf{b}_m+C_{M-1}\mathbf{b}_{M-1}+C_M\mathbf{b}_M\right)\right)^2\\&=\left(y_{k,i}^n-(\mathbf{s}_{k,i}^n)^T\sum_{m=1}^{M-2}C_m\mathbf{b}_m-(\mathbf{s}_{k,i}^n)^T[\mathbf{b}_{M-1},\mathbf{b}_M]\Lambda^{-1}\begin{bmatrix}\bar\rho c_0^2\\ -4\bar\rho c_0^3R\end{bmatrix}\right.\\&\quad\left.+\Delta x(\mathbf{s}_{k,i}^n)^T[\mathbf{b}_{M-1},\mathbf{b}_M]\Lambda^{-1}\mathbf{HBC}\right)^2\\&=\left(y_{k,i}^n-(\mathbf{s}_{k,i}^n)^T[\mathbf{b}_{M-1},\mathbf{b}_M]\Lambda^{-1}\begin{bmatrix}\bar\rho c_0^2\\ -4\bar\rho c_0^3R\end{bmatrix}\right.\\&\quad\left.-(\mathbf{s}_{k,i}^n)^T(\mathbf{I}-\Delta x[\mathbf{b}_{M-1},\mathbf{b}_M]\Lambda^{-1}\mathbf{H})\mathbf{BC}\right)^2\\&=\left(\tilde y_{k,i}^n-(\tilde{\mathbf{s}}_{k,i}^n)^T\mathbf{B}\tilde{\mathbf{C}}\right)^2,\end{aligned}$$

where

$$\tilde{y}^n_{k,i} := y^n_{k,i} - (\mathbf{s}^n_{k,i})^T [\mathbf{b}_{M-1}, \mathbf{b}_M] \Lambda^{-1} \begin{bmatrix} \bar{\rho} c_0^2 \\ -4\bar{\rho} c_0^3 R \end{bmatrix},$$

$\tilde{\mathbf{s}}^n_{k,i} := (\mathbf{I} - \Delta x [\mathbf{b}_{M-1}, \mathbf{b}_M] \Lambda^{-1} \mathbf{H})^T \mathbf{s}^n_{k,i}$, $\tilde{\mathbf{C}} := [C_1, \ldots, C_{M-2}]$, and $\mathbf{I}$ is an $d \times d$ identity matrix. Therefore, the analysis and algorithm can also be extended to the "constraints" cases.

Appendix D: detailed parameter and experiment settings

Meta-train and meta-test datasets

To demonstrate the performance of MetaNOR on both periodic and disordered materials, in empirical experiments we generate four types of data from the DNS solver for each microstructure. For each sample, the total training domain $\Omega = [-50, 50]$ and the training data is generated up to $T = 2$. The spatial and temporal discretization parameters in the DNS solver are set to $\Delta t = 0.01$, and $\max |\Delta x| = 0.01$. The other physical parameters are set as $L = 0.2$, $E_1 = 1$, $\rho_1 = \rho_2 = 1$, and $\phi = 0$. In experiment 1, we fix $E_2 = 0.25$ and set the disorder parameter $D \in [0.05, 0.50]$. In experiment 2, we set $E_2 \in [0.2025, 0.7225]$ and the disorder parameter $D = 0$. The training and testing data are obtained from the DNS data via linear interpolation with $\Delta t = 0.02$ and $\Delta x = 0.05$. The two types of data are chosen to follow a similar setting as in:[[6]]

1. *Oscillating source.* We let $\hat{u}(x, 0) = \frac{\partial \hat{u}}{\partial t}(x, 0) = 0$, $g(x, t) = \exp^{-(\frac{2x}{5kL})^2} \exp^{-(\frac{t-0.8}{0.8})^2} \cos^2(\frac{2\pi x}{kL})$, where $k = 1, 2, \ldots, 20$.
2. *Plane wave.* We set $g(x, t) = 0$, $\hat{u}(x, -200) = 0$, and $\frac{\partial \hat{u}}{\partial t}(-50, t) = \cos(\omega t)$. In experiment 1 (random microstructures), we set $\omega = 0.20, 0.40, \ldots, 4.0$. In experiment 2 (periodic microstructures), we set $\omega = 0.30, 0.60, \ldots, 6.0$.

In these two types of loading scenarios, the displacement $\hat{u}(x, t)$ is mostly zero when $x > 10$, which makes the corresponding datapoints carry very little information. To utilize the sample datapoints more efficiently, for the type 1 data, we only use datapoints from the $x \in [-10, 10]$ region, and for the type 2 data we only use datapoints from the $x \in [-38, -18]$ region.

Validation dataset: wave packet

We create a validation dataset, denoted as the wave packet dataset, which considers a much longer bar ($\Omega_{wp} = [-133.3, 133.3]$), and with a 50 times longer simulation time ($t \in [0, 100]$). The material is under a different loading condition from the training dataset, $g(x, t) = 0$ and $\frac{\partial \hat{u}}{\partial t}(-133.3, t) = \sin(\omega t) \exp\left(-(t/5 - 3)^2\right)$, for $\omega = 1$, 2, 3. To provide a metric for the estimator accuracy, we calculate the averaged displacement error in the discretized energy norm at the last time step. This error metric is referred to as the "validation error", which checks the stability and generalizability of the estimators.

Application D: Projectile impact simulations

To demonstrate the performance of learnt model in long-term simulation, we simulate the long-term propagation of waves in this material due to the impact of a projectile. In particular, in this problem a projectile hits the left end of the bar at time zero, which generates a velocity wave that travels into the microstructures.

To demonstrate the generalization capability of our approach on different domains, boundary conditions, and longer simulation time, we consider a drastically different setting in this simulation task. In particular, a much larger domain, $\Omega_{impact} = (-267, 267)$, and a much longer simulation time $t \in [0, 600]$ are considered. Notice that our training dataset are only generated up to $T = 2$, this long-term simulation task is particularly challenging not only because it has a different boundary condition setting from all training samples, but also due to the large aspect ratio between training time scale and simulation time scale. On the left end of the domain, we prescribe the velocity as $\frac{\partial \hat{u}}{\partial t}(-267, 0) = 1$, and zero velocity on elsewhere.

References

1. S. Yao, X. Zhou, G. Hu, Experimental study on negative effective mass in a 1d mass-spring system. N. J. Phys. **10**(4), 043020 (2008)
2. H. Huang, C. Sun, Wave attenuation mechanism in an acoustic metamaterial with negative effective mass density. N. J. Phys. **11**(1), 013003 (2009)
3. J.M. Manimala, H.H. Huang, C. Sun, R. Snyder, S. Bland, Dynamic load mitigation using negative effective mass structures. Eng. Struct. **80**, 458–468 (2014)
4. Q. Du, B. Engquist, X. Tian, Multiscale modeling, homogenization and nonlocal effects: Mathematical and computational issues (2019)
5. L. Barker, A model for stress wave propagation in composite materials. J. Compos. Mater. **5**(2), 140–162 (1971)
6. H. You, Y. Yu, S. Silling, M. D'Elia, Data-driven learning of nonlocal models: from high-fidelity simulations to constitutive laws. (MLPS, AAAI Spring Symposium, 2021)
7. S.A. Silling, Propagation of a stress pulse in a heterogeneous elastic bar. J. Peridyn. Nonlocal Model. **3**, 1–21 (2021)
8. K. Deshmukh, T. Breitzman, K. Dayal, K. Multiband homogenization of metamaterials in real-space: Higher-order nonlocal models and scattering at external surface. preprint (2021)
9. H. You, Y. Yu, N. Trask, M. Gulian, M. D'Elia, Data-driven learning of robust nonlocal physics from high-fidelity synthetic data. Comput. Methods Appl. Mech. Eng. **374**, 113553 (2021)
10. H. You, Y. Yu, S. Silling, M. D'Elia, A data-driven peridynamic continuum model for upscaling molecular dynamics. Comput. Methods Appl. Mech. Eng. **389**, 114400 (2022)
11. N. Tripuraneni, C. Jin, M. Jordan, Provable meta-learning of linear representations. In: International Conference on Machine Learning, pp. 10434–10443 (2021). PMLR
12. M. Beran, J. McCoy, Mean field variations in a statistical sample of heterogeneous linearly elastic solids. Int. J. Solids Struct. **6**(8), 1035–1054 (1970)
13. S.A. Silling, Origin and effect of nonlocality in a composite. J. Mech. Mater. Struct. **9**(2), 245–258 (2014)
14. Q. Du, B. Engquist, X. Tian, Multiscale modeling, homogenization and nonlocal effects: mathematical and computational issues. Contemp. Math. **754**, 18 (2020)

15. Q. Du, Y. Tao, X. Tian, A peridynamic model of fracture mechanics with bond-breaking. J. Elast. **25**, 1–22 (2017)
16. R. Ge, C. Jin, Y. Zheng, No spurious local minima in nonconvex low rank problems: a unified geometric analysis. In: International Conference on Machine Learning, pp. 1233– 1242 (2017). PMLR
17. Y. Liu, T. Zhao, W. Ju, S. Shi, Materials discovery and design using machine learning. J. Mater. **3**(3), 159–177 (2017)
18. S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. Nat. Commun. **9**(1), 1–8 (2018)
19. K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science. Nature **559**(7715), 547–555 (2018)
20. J. Cai, X. Chu, K. Xu, H. Li, J. Wei, Machine learning-driven new material discovery. Nanoscale Adv. **2**(8), 3115–3130 (2020)
21. Y. Iwasaki, I. Takeuchi, V. Stanev, A.G. Kusne, M. Ishida, A. Kirihara, K. Ihara, R. Sawada, K. Terashima, H. Someya et al., Machine-learning guided discovery of a new thermoelectric material. Sci. Rep. **9**(1), 1–7 (2019)
22. F. Ren, L. Ward, T. Williams, K.J. Laws, C. Wolverton, J. Hattrick-Simpers, A. Mehta, Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. Sci. Adv. **4**(4), 1566 (2018)
23. K. Kaufmann, D. Maryanovsky, W.M. Mellor, C. Zhu, A.S. Rosengarten, T.J. Harrington, C. Oses, C. Toher, S. Curtarolo, K.S. Vecchio, Discovery of high-entropy ceramics via machine learning. NPJ Comput. Mater. **6**(1), 1–9 (2020)
24. A. Chen, X. Zhang, Z. Zhou, Machine learning: accelerating materials development for energy storage and conversion. InfoMat **2**(3), 553–576 (2020)
25. C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning, pp. 1126– 1135 (2017). PMLR
26. Y.L. Qiu, H. Zheng, A. Devos, H. Selby, O. Gevaert, A meta-learning approach for genomic survival analysis. Nat. Commun. **11**(1), 1–11 (2020)
27. Y. Chen, C. Guan, Z. Wei,X. Wang, W. Zhu, Metadelta: A meta-learning system for few-shot image classification. arXiv preprint arXiv:2102.10744 (2021)
28. Yin, W.: Meta-learning for few-shot natural language processing: a survey. arXiv preprint arXiv:2007.09604 (2020)
29. B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, T.Y.J. Han, Reliable and explainable machine-learning methods for accelerated material discovery. NPJ Comput. Mater. **5**(1), 1–9 (2019)
30. H. Mai, T.C. Le, T. Hisatomi, D. Chen, K. Domen, D.A. Winkler, R.A. Caruso, Use of meta models for rapid discovery of narrow bandgap oxide photocatalysts. iScience **24**, 103068 (2021)