npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Deep learning approach based on dimensionality reduction for designing electromagnetic nanostructures

Yashar Kiarashinejad[1,2], Sajjad Abdollahramezani[1,2] and Ali Adibi[1]*

In this paper, we demonstrate a computationally efficient new approach based on deep learning (DL) techniques for analysis, design and optimization of electromagnetic (EM) nanostructures. We use the strong correlation among features of a generic EM problem to considerably reduce the dimensionality of the problem and thus, the computational complexity, without imposing considerable errors. By employing the dimensionality reduction concept using the more recently demonstrated autoencoder technique, we redefine the conventional many-to-one design problem in EM nanostructures into a one-to-one problem plus a much simpler many-to-one problem, which can be simply solved using an analytic formulation. This approach reduces the computational complexity in solving both the forward problem (i.e., analysis) and the inverse problem (i.e., design) by orders of magnitude compared to conventional approaches. In addition, it provides analytic formulations that, despite their complexity, can be used to obtain intuitive understanding of the physics and dynamics of EM wave interaction with nanostructures with minimal computation requirements. As a proof-of-concept, we applied such an efficacious method to design a new class of on-demand reconfigurable optical metasurfaces based on phase-change materials (PCMs). The experimental results of the fabricated devices are in good agreement with those predicted by the proposed approach. We envision that the integration of such a DL-based technique with full-wave commercial software packages offers a powerful toolkit to facilitate the analysis, design, and optimization of the EM nanostructures as well as explaining, understanding, and predicting the observed responses in such structures. It will thus enable to solve complex design problems that could not be solved with existing techniques.

npj Computational Materials (2020)6:12 ; https://doi.org/10.1038/s41524-020-0276-y

INTRODUCTION

The field of nanophotonics has been the subject of extensive expansion due to the unique capabilities of photonic nanostructures to control the propagation of electromagnetic (EM) waves. Owing to their constituent nanoscale features, which spectrally, spatially, and even temporally manipulate the optical state of the EM wave, nanophotonic devices extend all the functionalities realized by conventional optical devices in much smaller footprints. Combined with the advances in nanofabrication technologies, these nanostructures have been used to demonstrate devices with enormous potential for groundbreaking technologies addressing major challenges in state-of-the-art applications, such as optical communications,[1] signal processing,[2] biosensing,[3] energy harvesting,[4] and imaging,[5] to name a few. As an example, newly-emerged metasurfaces (MSs),[6–14] two-dimensional planar structures comprising of densely arranged periodic/aperiodic arrays of well-engineered dielectric or plasmonic inclusions, offer profound control of the EM wave dynamics including amplitude, phase, polarization, and frequency in the subwavelength regime.[15–18]

Despite extensive achievements in the fabrication and realization of photonic nanostructures, the efforts on the development of accurate and computationally efficient design and optimization approaches for these nanostructures are still at early stages.[19] With the fast progress in forming more complex nanostructures with several design parameters, the need for new design approaches that can keep pace with the computational requirements for analysis and understanding of all possible design options has become more imminent. In addition, realization of next-generation nanodevices with potentially new physics enabled through light-matter interaction at the nanoscale requires significant knowledge about the role of different design parameters in the functionality of a nanostructure.

Traditional design and optimization approaches for EM nanostructures rely on either using analytical (or semi-analytical) modeling[19–25] or brute-force analysis of the nanostructure through exhaustive search of the design parameter space.[26] The use of these approaches are limited to simple structures that could be either analytically modeled or completely studied by an exhaustive search technique with reasonable computation cost. To improve the computation efficiency of such design and optimization tools, evolutionary approaches (e.g., genetic algorithm[27,28] and particle swarm[29]) rely on starting from a random initial guess and converging to the final optimum. While reducing the computation cost compared to brute-force approaches, such techniques are not guaranteed to converge to the global optimum of a problem (even by allocation of extensive computational resources). They are also limited to a single design problem (i.e., the simulations must be completely repeated when a small change in the nanostructure happens) and are computationally expensive for large-scale problems due to the significant amount of iterations to find the optimum design for a given device functionality.

More recently, design and optimization approaches based on deep learning (DL) techniques have been proposed and implemented for the design of nanostructures.[30–37] Different reported approaches to date primarily rely on training a neural network (NN) (see Fig. 1a) using the response of a set of devices (found by numerical simulations) and using the trained NN to

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, 778 Atlantic Drive NW, Atlanta, GA 30332, USA. [2]These authors contributed equally: Yashar Kiarashinejad, Sajjad Abdollahramezani. *email: ali.adibi@ece.gatech.edu

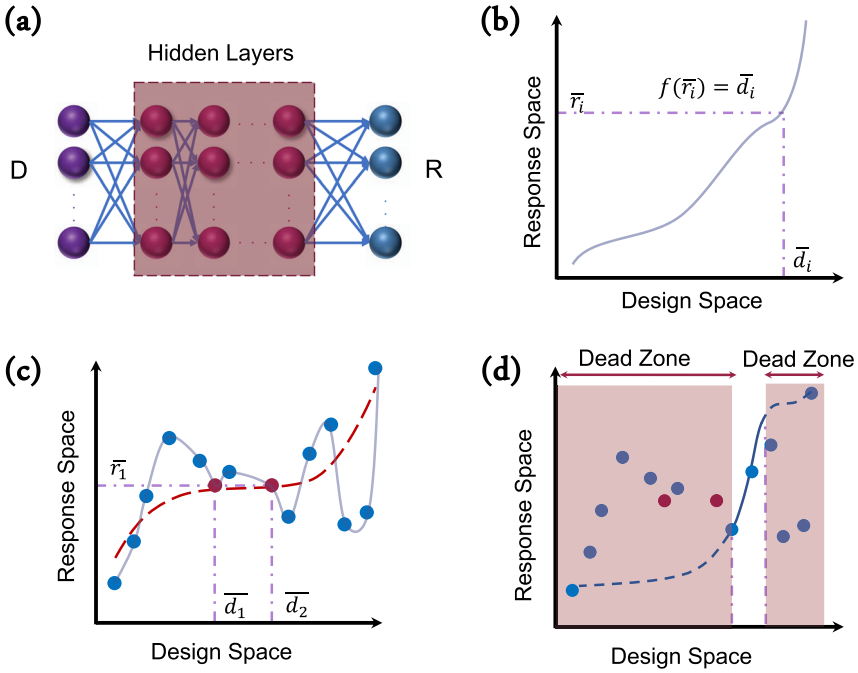

Fig. 1 Representation of a many to one problem. a A feed-forward NN for design and analysis of EM nanostructures; D and R represent design and response parameters, respectively. **b** Representation of a one-to-one design landscape (or manifold) as the simplest class of problems for solution with the NN in (**a**). **c** Representation of a general (non-one-to-one or many-to-one) design manifold. Red dots represent instances with same response features obtained with different sets of design parameters. The light-blue curve demonstrates the original design manifold while the dashed line shows the estimated one obtained with conventional methods for solving one-to-one problem (e.g., the NN in **a**). **d** Representation of the same design manifold as in (**c**) with a solution obtained by just training the NN in (**a**) for some intrinsically one-to-one region (outside the dead-zones); the non-optimal extrapolated manifold for the dead-zones is highlighted by red color.

solve the inverse design problem. Despite impressing progress in this area, the reported solutions mostly focus on solving simple problems with reasonably smooth optimization landscapes[31] that have a one-to-one mapping of design parameter space to the device response space (i.e., any given response can be obtained by only a single set of design parameters) as shown in Fig. 1b, where a vector of device response ($\overline{r}_i$) is achieved by a unique vector of design parameters ($\overline{d}_i$). Unfortunately, most nanostructures of interest do not have this property. Figure 1c shows the optimization landscape of a more general problem in which the one-to-one relation between design parameters and output response does not exist. This can result in convergence issues for the NN used for optimization (i.e., finding design parameters for a given output response). Efforts on converting the problem to a one-to-one mapping by removing some training data sets (see Fig. 1c)[38] do not essentially help in solving the problem as most of the design space is not covered by these training datasets. Such approaches at most result in a NN that smooths out the optimization dataset (see Fig. 1c) without converging to the global optimum. Other proposed approaches (e.g., the use of tandem networks[30]) rely on first training a NN that relates the design space to the response space (i.e., for the forward problem), then cascading it as a pre-trained NN with another NN that relates the response space to the design space (i.e., the inverse problem), and finally training the resulting network (from the response space to the design space) to avoid the non-one-to-one relation. However, such techniques do not solve the main problem; they at best smooth out the optimization landscape as shown in Fig. 1c. Another notable recent approach is based on using generative adversarial networks (GANs) to solve the inverse design problem.[32] This technique is built on training a network to solve the forward problem with zero error and use it to generate ground truth data in each iteration. Training such a forward-problem-solver network with zero error in a general design problem is a major challenge and may require excessive computational resources. In the reported design problem, each desired output needs extensive computation (200,000 iterations to reach the convergence region for each structure),[32] which may reduce the value of using GANs if a perfect forward problem solver exists with comparable computation complexity (similar computations can be used to solve the design problem by exhaustive search using the perfect forward-problem solver). Despite impressing results, the reported GAN-based approach will be limited to simple design problems with non-complex nanostructure. Also limiting the design space to a smooth (one-to-one) region is unable to address the nonuniqueness challenge. The success of such techniques highly depends on the complexity of the problem and the selection of the design parameters in the one-to-one region (outside the dead-zones in Fig. 1d) to converge to acceptable answers. As a result, these approaches can be used to design simple structures, which can also be designed using alternative approaches. Finding a reliable approach to fundamentally address this nonuniqueness issue (without limiting the optimization landscape to the one-to-one region (or extrapolating from it, see Fig. 1d) is still a major challenge in using DL based approaches for the design of EM nanostructures.

Another challenge in using DL techniques to design complex EM nanostructures is the large size of the response and design spaces resulting in the need to train a large NN. As an example, to study the spatial and spectral response of a MS with reasonable accuracy, the response space must constitute the sampled EM intensity in a two-dimensional space and in frequency with spatial and spectral resolutions smaller than the smallest spatial and spectral features of the output response, respectively. This typically results in thousands of data points in the response space and quickly rises as the structures with sharper spatial and spectral features are designed. Combined with ever-increasing number of design parameters in the nanostructures of recent interest, this

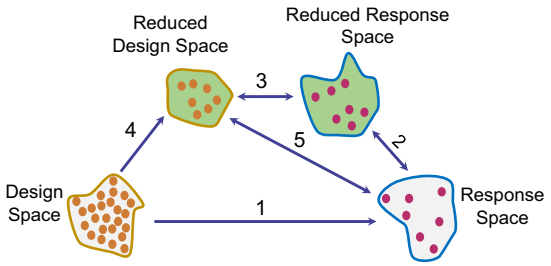

Fig. 2 Applying the DR technique to the response and design spaces. In an optimal implementation, paths 1 and 4 are many-to-one while paths 2, 3, and 5 are one-to-one. The directions of arrows indicated paths that are easily achievable due to their one-to-one nature.

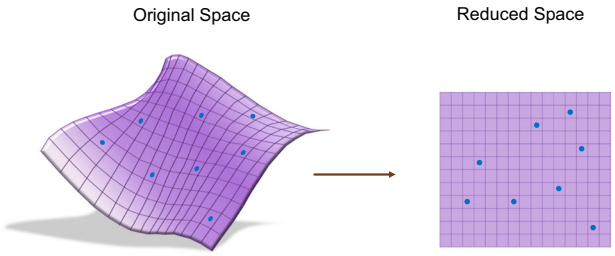

Fig. 3 An example for the one-to one DR. Each dot represents a point in the original space that corresponds to a point (shown by a dot) in the lower-dimensional reduced space.

results in a very large NN, which is difficult to be trained, even for the problems with one-to-one optimization landscapes.

In this paper, we demonstrate a new approach for designing complex EM nanostructures by addressing both the network-size issue and the nonuniqueness issue. Our approach is based on reducing the dimensionality of both the design space and the response space through training multi-layer NNs, called auto-encoders.[39] Once the dimensionality of the problem is reduced, the problem converts into a one-to-one problem in the reduced spaces, which can be solved with considerably less computational complexity. In addition, by reducing the envisioned design parameters to few number of more complex design parameters (e.g., a nonlinear function of the weighted sum of the original design parameters), we can obtain valuable intuitive understanding of the roles of different design parameters in the response of the nanostructure. Such an efficacious approach paves the way for understanding, design, and optimization of complex EM nanostructures with far less computation complexity than the alternative approaches. In addition, a trade-off between the accepted error and the complexity (and time) of the simulations can be used to solve different problems with desired degrees of computation complexity or to obtain quick (approximate) information about the role of design parameters in the overall device performance. Dimensionality reduction (DR) is a powerful technique in machine learning that has been used to effectively solve problems in a wide range of applications including robotics,[40] optical tomography,[41] face recognition,[42] handwritten digit classification,[43] remote sensing,[44] medical science,[45] genetics,[46] and electronics.[47]

To show the applicability of our approach, we demonstrate its use for designing a new class of reconfigurable MS based on PCMs to form wideband amplitude modulation of near-infrared (near-IR) light.

RESULTS

Dimensionality reduction of the design and response spaces in designing electromagnetic nanostructures

Figure 2 shows the schematic of the design approach based on DR of the design and response spaces assuming that the optimization landscape is nonunique (or many-to-one), i.e., more than one set of design parameters can result in the same response. The original forward problem is shown by path 1 in Fig. 2, where each point in the design space (that includes a vector of dimension D corresponding to a set of design parameters) correspond to a point in the response space (which includes a vector of dimension R) through a many-to-one relationship. A NN cannot be trained to inverse this relation as explained above. This is the main complication in the design and optimization problem. In our approach, we first use the DR technique to reduce the dimensionality of the response space as much as possible (i.e., reducing the size of the response vector $\bar{r}_i$ in Fig. 1b) while keeping the same number of points in the response space (see path 2 in Fig. 2). This concept is schematically shown in Fig. 3 in which a three-dimensional manifold in the response space is reduced to a two-dimensional manifold, which includes the same number of points in the response space, but each point is represented by a smaller size vector. Each feature in the reduced response space is related to the features of the original response space through a well-defined nonlinear function. This is a one-to-one process.

In the next step, we reduce the dimensionality of the design space as much as possible (see path 4 in Fig. 2). In this process, the redundant nature of the design space is removed resulting in a one-to-one relation between the reduced design space and the reduced response space (see path 3 in Fig. 2). After training the relevant DR mechanisms in Fig. 2, the relation between the original response space and the reduced design space (paths in Fig. 2) will be one-to-one and thus, it can be simply inverted. Thus, our design problem will relate the desired response to the reduced design parameters (see path 5 in Fig. 2). The reduced design parameters are related to the original design parameters through a one-to-many relation that is analytically available through the training process (i.e., in the form of a formula with a series of nested Tanh(.) functions that model different nodes of the trained NN for the encoder part of the pseudo-encoder). Thus, we can find several design options by converting the resulting optimum reduced design parameters to several sets of the original design parameters. At this stage, design constraints (e.g., fabrication imperfections, structure robustness, characterization limitations, etc.) can be taken into account to choose the final design parameters.

The heart of our approach is the effective implementation of the DR technique to maximally reduce the dimensionality of both design and response spaces, especially the former. Several DR techniques have been developed in machine learning to facilitate classification, data visualization, reduction of the computation cost, etc. Among different options, principal component analysis (PCA),[48] kernel principal component analysis (KPCA),[49] Laplacian eigen map,[50] locally linear embedding,[51] and autoencoder[39] are the most effective techniques. Considering the features of these techniques, we believe that the autoencoder is the most suitable approach for solving inverse problems in general and designing EM nanostructures in particular.

The general schematic of an autoencoder is shown in Fig. 4. Autoencoder is a multilayer NN that can encode the high-dimensional data into low-dimensional data (using the encoder part in Fig. 4) and use another NN (see the decoder part in Fig. 4) to decode and recover the high-dimensional data. In other words, the autoencoder in Fig. 4 is a feed-forward NN where the input layer and the output layer have the same structure and are connected to each other with one or more hidden layers. The number of neurons in the layer with minimum number of neurons represents the dimension of the reduced space. This layer is known as the bottleneck of the autoencoder. This way, an

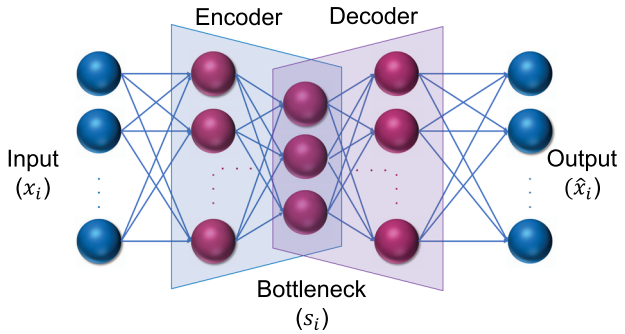

Fig. 4 Schematic architecture of an autoencoder for the DR technique. The left half (i.e., encoder) reduces the dimensionality (the bottleneck layer corresponds to the reduced space) while the right half (i.e., decoder) brings the data back to the original space. The complete autoencoder is trained to minimize the MSE.

autoencoder concentrates the data from a high-dimensional manifold in a given space around a low-dimensional manifold or a small set of such manifolds. The goal of an autoencoder is to map an original set of input data $\{x_1, x_2, \cdots, x_n\}$ to a lower dimensional set of output data $\{s_1, s_2, \cdots, s_n\}$ (at the bottleneck) in which x_i and s_i are vectors with size $k \times 1$ and $m \times 1$, respectively ($m < k$), and s_i contains the essential information of x_i.

To find the mapping from high-dimensional to low-dimensional data, the autoencoder in Fig. 4 should be trained with a sufficiently large training dataset. The training part of the autoencoder can be considered as an optimization problem where the algorithm minimizes a cost function. The cost function is a measurement of discrepancy between the output of the autoencoder and the input data. The mean-squared error (MSE) is used as the cost function of the autoencoder, and the error is minimized using the back-propagation method.[52] Assuming the output of the autoencoder structure in Fig. 6 for the input x_i is represented by $\hat{x}_i$, the reconstruction MSE of the trained autoencoder is defined as:

$$\mathrm{MSE} = \frac{1}{\acute{n}} \sum_{i=1}^{\acute{n}} \| x_i - \hat{x}_i \|_2^2 \tag{1}$$

Where $\acute{n}$ represents the number of validation (or test) instances (not used for training) that are used to validate the trained autoencoder (but not used for training) . The number of layers and the topology of the NN is also found using an ad-hoc method (by trial and error). The training dataset for the design of EM nanostructures is obtained by using numerical simulation of the structure using a random set of input design parameters.

In the approach shown in Fig. 2, we first reduce the dimensionality of the response space by training an autoencoder (see Fig. 5a). In the next step, we form a pseudo-encoder that relates the original design space to the reduced response space as shown in Fig. 5b. The reason for naming the structure in Fig. 5b a pseudo-encoder is the fact that its input and output are from different spaces (in contrast to a conventional autoencoder in Fig. 4). By training the pseudo-encoder in Fig. 5b to reach the minimum size of the bottleneck layer, we reach the reduced design space. Each parameter in this space is related to the original design parameters through a nonlinear function defined by the NN structure of the pseudo-encoder from the original design space to the reduced design space (or the bottleneck) in Fig. 5b. The training approach is similar to that explained for a general autoencoder in Fig. 4 or Fig. 5a. The pseudo-encoder in Fig. 5b corresponds to the paths 3 and 4 in Fig. 2, i.e., these two paths are trained together.

Once the DR of the two spaces are complete, we form a NN by cascading the pseudo-encoder in Fig. 5b with the pre-trained decoder part of Fig. 5a to form a completely trained NN for solving the forward problem as shown in Fig. 5c. The resulting NN in Fig. 5c relates the original design parameters to the original response space using a unique set of analytic equations defined by different layers of the NN. While this analytic relation is complicated for a large-size network, it provides extremely valuable information about the roles of different design parameters in the response of the nanostructure with minimal computation complexity (technically by calculating the complex analytic formulas in a conventional environment like MATLAB). However, the goal of this paper is the design of EM nanostructures for which the inverse problem has to be solved. For this purpose, we will use a two-step approach. In the first step, we find the inverse of the part of the NN in Fig. 5c that relates the reduced design space to the output space. The resulting inverse network is shown in Fig. 5d. This is easily achievable as the relation between the reduced design space and the original response space is one-to-one (see path 5 in Fig. 2). The NN in Fig. 5d allows us to obtain the optimal reduced design parameters for any given desired response. This is the last part in our approach where the DL approaches can be used. The final step is to relate the reduced design parameters to the original design parameters (i.e., the inverse of path 4 in Fig. 2). This is a nonunique relation, i.e., it can provide several sets of design parameters from a given reduced set of design parameters. Fortunately, the encoder part of the pseudo-encoder in Fig. 5b relates the reduced design parameters to the original design parameters analytically (through the formulation of the underlying NN at different layers). We can use these equations to move layer-by-layer from the reduced design parameters to the original design parameters. In this backward process, we can reduce the number of possible solutions by imposing constraints such as fabrication limitations. This approach can provide many possible solutions for a design problem, which is expected due to the nonuniqueness of the problem. Note also that within this design problem, we can use the obtained knowledge about the role of the design parameters (using the forward solver in Fig. 5c) and the relation between the reduced design parameters and the original design parameters (using the encoder part of the pseudo-encode in Fig. 5b) to reduce the complexity in solving the design problem. In this paper, we use the analytic relation between the original and reduced design spaces to completely search the original design space to find the point(s) that correspond to the desired point in the reduced design space.

In addition to solving the nonuniqueness issue, the approach in Fig. 5 considerably reduces the computation cost by reducing the dimensionality of the two spaces. It is clear that the training of the pseudo-encoder that relates the design space to the reduced response space (see Fig. 5b) requires much less computation compared to training of a NN that relates the design space to the original (non-reduced) response space. Furthermore, the calculation of the inverse NN in Fig. 5d does not impose significant computation cost due to its one-to-one nature.

Application to the design of hybrid reconfigurable plasmonic-PCM metasurfaces

To show the applicability of the design approach, we consider a generic design problem for the implementation of a reconfigurable multifunctional MS enabling high performance optical modulation as shown in Fig. 6. The metasurface (MS) in Fig. 6 is composed of a periodic array of gold (Au) nanoribbons fabricated on top of a thin layer of germanium antimony telluride ($Ge_2Sb_2Te_2$ or in short GST), which is a non-volatile PCM whose index of refraction can be significantly modified (e.g., from 4.5 to 7 in the near-infrared region)[53] when it undergoes transition from the amorphous to the crystalline state in the near infrared regime or vice versa. In addition, using GST in intermediate states between amorphous and crystalline results in a wide range of tunability for

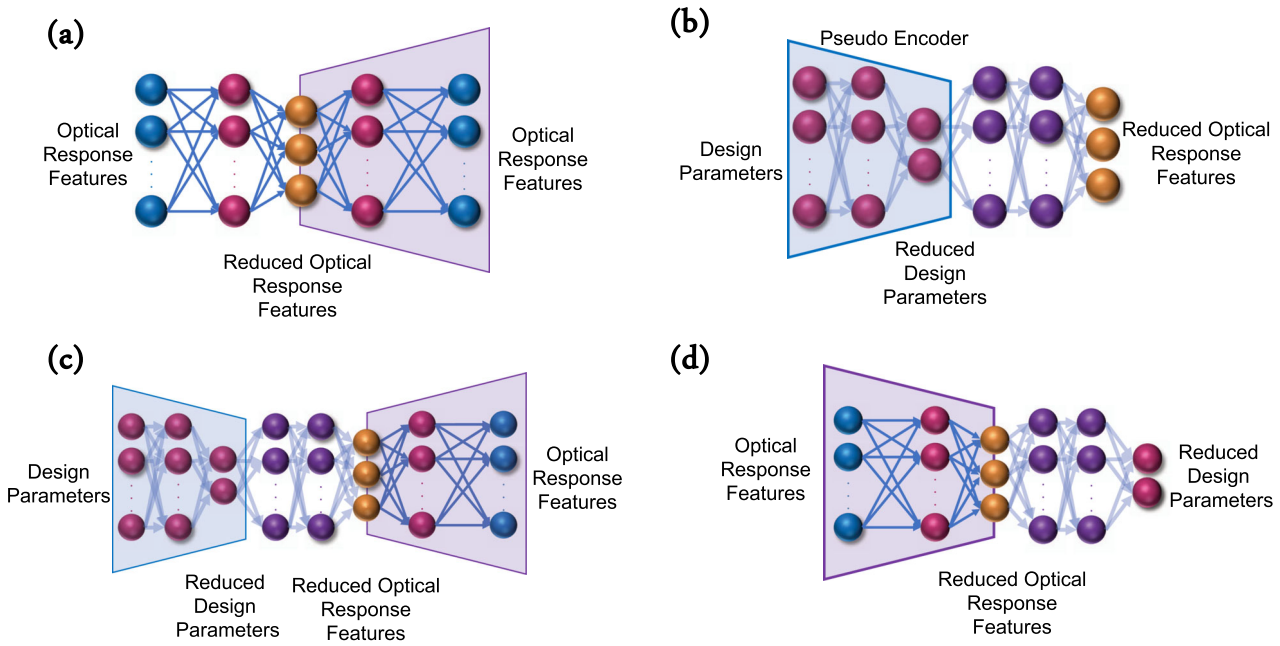

Fig. 5 Schematic of the NN architectures. a Using an autoencoder to reduce the dimensionality of the response space (i.e., extract reduced optical response features from the original response features). **b** The pseudo-encoder architecture, which relates the original design space to the reduced response space while reducing the dimensionality of the design space (i.e., extracting the reduced design parameters). **c** A complete model for the forward problem formed by cascading the pseudo-encoder architecture in (**b**) with the decoder part of the autoencoder in (**a**). **d** The semi-inverse-problem model, which relates the original response space to the reduced design space as a one-to-one problem.

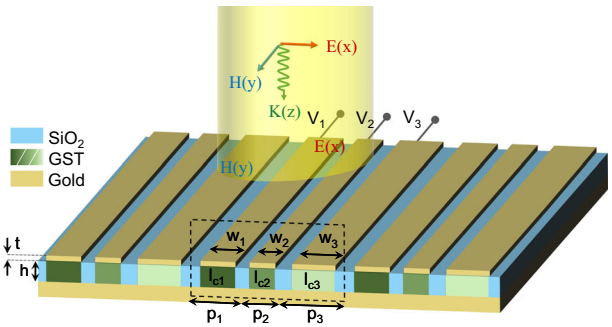

Fig. 6 3D representation of the MS under study. A MS with reconfigurable reflectivity formed by a periodic array of Au nanoribbons (thickness: t) on top of a thin layer (height: h) of GST on top of a SiO_2 substrate. The unit cell of the structure is composed of three Au nanoribbons with different widths (w_1, w_2, and w_3) and pitches (p_1, p_2, and p_3, respectively). Other design parameters are the crystallization state of GST under the three nanoribbons (shown by l_{c1}, l_{c1}, and l_{c3}, respectively) and the height of the GST layer (h). The phase of GST under each nanoribbon can be changed by applying a voltage (V_1, V_2, and V_3, respectively). The incident light normally illuminates the MS, and the spatial and spectral profiles of the reflection from the structure is calculated as its response.

its index of refraction. The GST layer deposited on an optically thick film of Au is patterned as shown in Fig. 6. By laterally applying electric signals to the Au nanoribbons, the state of the GST underneath that nanoribbon is controlled through resistive heating.[54] In addition, by controlling the electric stimulus intermediate states (between amorphous and crystalline) can be obtained for GST.[55] We limit the number of GST transition states to 11 (i.e., amorphous, crystalline, and 9 intermediate states).[56] The supercell (limited to three different building blocks to prevent excitation of high diffraction orders) of the MS in Fig. 6 is composed of three Au nanoribbons with different widths (w_1, w_2, and w_3) and 3 crystallization levels (l_{c1}, l_{c2}, l_{c3}, corresponding to three indices of refraction, see "Methods" for more details) of GST underneath with the same height (h). The pitches of the 3 building blocks of the supercell are represented by p_1, p_2, and p_3, respectively. As a result, the MS in this work has 10 design parameters (i.e., dimensionality of the design space is equal to 10) with different units (3 unit-less indices of refraction and 7 lengths with units of nanometers).

As an interesting functionality, we are interested in amplitude modulation of the incident light at $\lambda = 1600$ nm with a considerable bandwidth around the central wavelength. The MS in Fig. 6 is illuminated with a plane wave of light with variable wavelengths in the desired range (from 1250 nm to 1850 nm). The polarization of the incident light is such that the electric field (i.e., E_x) is perpendicular to the grating direction of the MS. The response of the system is the MS reflectance (calculated as the far-field reflection intensity divided by the intensity of the incident field and integrated over a surface area equal to one supercell in the far-field). The resulting reflectance is sampled at 200 equally-spaced wavelengths in the 1250–1850 nm range. This results in a response space dimensionality of 200. To obtain the data for training and validation of the DR autoencoders in Fig. 5, we simulate the structure in Fig. 6 with 4000 randomly generated instances (3600 for training and 400 for validation) formed by randomly selecting the design parameters in the acceptable variation ranges shown in the caption of Fig. 6. The simulations were performed using the finite element method (FEM) in the COMSOL Multiphysics environment (see "Methods" for details).

In the next step, we use the training data to train a series of autoencoders with different numbers of hidden layers (ranging from 3 to 9) to study the DR of the response space from 200 to different values in the range of 1–20. For each autoencoder, MSE is calculated by finding the square of the norm of the difference between the reflectance vector obtained from the decoder part of the autoencoder and that obtained from the FEM simulations (also

called ground truth data) for each one of the validation data. Note that each vector has 200 elements corresponding to the reflectance at 200 selected wavelengths in the 1250–1850 nm range.

Figure 7a shows the calculated MSE as a function of the dimensionality of the reduced response space. Figure 7b shows the comparison of the actual reflectance spectrum for the original data and the reconstructed data using the autoencoder for different dimensionalities of the reduced response space. It is clear from both Fig. 7a, b that the dimension of the response space can be reduced from 200 to 10 with negligible MSE (less than 10^{-3}). This is a clear advantage of our optimization technique.

Among different autoencoder architectures tested for the response space, the one with five layers (with the number of neurons in consecutive layers being 200-50-10-50-200) is selected based on its low MSE and computation costs to form the pseudo-encoder architecture in Fig. 5b. We also choose four layers (10-20-15-x) for the encoder part of the pseudo-encoder in Fig. 5b. For each set of values for the dimensions of the reduced response space and the reduced design space, we train the resulting pseudo-encoder using the 3600 training data, and we calculate the MSE by comparing the output of the pseudo-encoder with the actual output using the 400 validation data. The results for four different dimensionalities of the reduced response space are shown in Fig. 8a. Figure 8b shows representative reflectance spectra for three different values of the dimensionality of the reduce design space. Using Fig. 8a, it is clear that the dimension of the design space can be reduced from 10 to 5 without imposing much error.

Using Figs. 7 and 8, we choose the dimensionality of the reduced response space and the reduced design space to be 10 and 5, respectively (10-20-15-5-20-30-20-10-50-200). This considerably reduces the computation time as the dimension of the resulting problem is defined in a 5×10 rather than 10×200. Using these values, the final NN architecture for the analysis (or the solution of the forward problem) of the MS in Fig. 6 is formed according to Fig. 5c. It is clear that the training of the pseudo-encoder that relates the design space to the reduced response space (see Fig. 5b) requires much less computation compared to training of a NN that relates the design space to the original (non-reduced) response space.

To form a platform for designing MSs with an arbitrary response, we first find the inverse of the network from the original response space to the reduced design space as shown in Fig. 5d. This is not computationally extensive due to the one-to-one nature of the problem. For this purpose, the pre-trained encoder part of the DR algorithm for the response space (left side of Fig. 5a) is combined with a NN that connects the reduced response space to the reduced design space. This added NN is trained using the same 3600 training data to form the inverse network that relates the desired response to the reduced design parameters. The resulting one-to-one trained platform (see Fig. 5d) results in finding the five reduced design parameters for the desired response. To find the 10 original design parameters, we solve the one-to-many problem through an analytical search approach using the encoder part of the pseudo-encoder for DR of the design space (first part of the platform in Fig. 5b). This encoder part relates the original design parameters analytically (through the NN formulation) to the reduced design parameters. Thus, the exhaustive search of the design space is not computationally extensive. We use MATLAB to perform this calculation (sweeping each parameter over 10 possible values) using the minimization of the MSE (defined by the integral of the square of the difference between the desired and the resulting light intensities over the operation bandwidth) as the optimization goal.

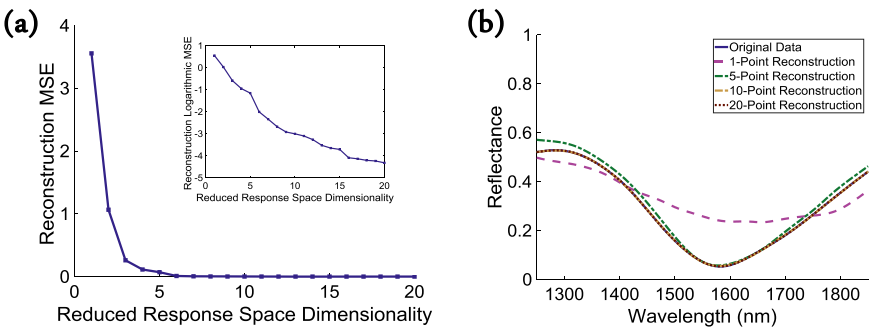

Fig. 7 Comparison of the MSE for different dimensionality of the reduced response space. a MSE of the DR mechanism for the response space of the structure in Fig. 6 as a function of the dimensionality of the reduced response space. The inset shows the same data in the logarithmic scale. **b** The reconstructed response of the nanostructure in Fig. 6 after DR of the response space as a function of the dimensionality of the reduced response space.

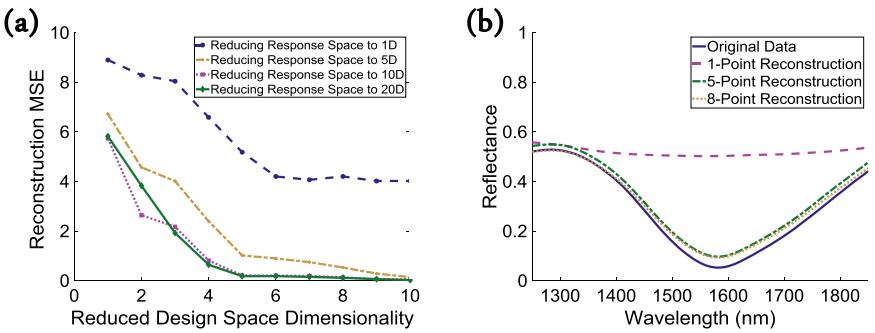

Fig. 8 Comparison of the MSE for different dimensionality of the reduced design space. a Performance of the DR technique (i.e., the architecture in Fig. 5c) for the analysis of the response of the nanostructure in Fig. 6 (i.e., solution of the forward problem) in terms of MSE of reflectance for different dimensionalities of the reduced response and design spaces. **b** A typical spectral reflectivity response of the structure in Fig. 6 for different dimensions of the reduced design space with a fixed dimension (=10) of the reduced response space along with the original reflectivity response.

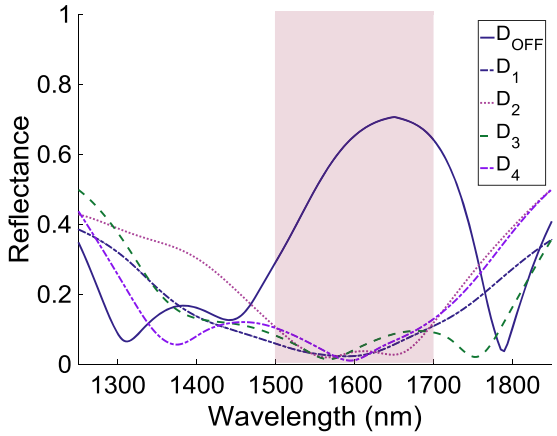

Fig. 9 Achieved spectral responses for the full absorption in the 1500–1700 nm wavelength range. Responses of the optimal (D_1) and three other reasonably good designed structures (D_2, D_3, and D_4) with the goal of achieving maximum absorption in the 1500–1700 nm wavelength region (shown by the shaded rectangle). The MSE and the values of the design parameters are shown in Table 1. The blue line shows the reflectance spectrum in the OFF-state (i.e., amorphous).

Table 1. Values of the design parameters for four totally different set of design parameters (d_1, d_1^*, d_2, and d_2^*) which each two results in almost identical responses (see Fig. 13).

Design	h	l_{c1}	l_{c2}	l_{c3}	p_1	p_2	p_3	w_1	w_2	w_3
d_1	50	0.7	0.7	0.8	600	700	800	480	140	160
d_1^*	150	0.7	0.3	0.9	1000	600	400	500	180	120
d_2	50	0.7	0.3	0.5	700	600	500	490	180	500
d_2^*	75	1	0.7	0.8	400	1000	700	320	500	560

Δd_n(10-D)/Δd_n(5-D) for (d_1, d_1^*) and (d_2, d_2^*) are (0.83/0.024) and (1.72/0.07), respectively. h, w_i, and p_i {$i = 1, 2, 3$} are in nm.

Figure 9 shows the results for the design of prefect light absorber for operation in the 1500–1700 nm wavelength range using the MS in Fig. 6. The desired response is zero reflectivity over the entire operation bandwidth. The overall MSE for the response of the optimal structure in Fig. 6 is 0.0147 MSE. The reflectance for two other (non-optimal) designs with considerably different design parameters are also shown in Fig. 9. The set of design parameters along with the MSE for these three structures are listed in Table 1.

To ensure the selected structure is indeed the optimal design, we compared the MSE for its response with those of all training instances. This comparison is graphically shown in Fig. 1 below, and it confirms that the selected structure has the lowest MSE. The advantage of the optimal design is clear both qualitatively (by comparing the frequency response of the optimal structure with those of the best three responses (with minimal MSE) in the training data set as shown in Fig. 10a) and quantitatively (by comparing the MSE of the optimal response with those of all 3600 training dataset as shown in Fig. 10b).

Understanding the physics of light-matter interaction

A main advantage of our approach is the possibility of investigating the underlying physics of the device operation and obtaining intuitive information about the roles of different design parameters on its response. To show this capability, we use our approach with a pseudo-encoder (10-4-10-50-20-10) to model the MS in Fig. 6. Figure 11a shows the resulting pseudo-encoder with the dimension of the reduced design space being 4 with green and red arrows representing positive and negative weights, respectively. Note that the DR of the design space is performed with only one encoder layer. Figure 11b shows the values of the weights for the mono-layer encoder. Each weight is multiplied by its corresponding design parameter to form the inputs to the node of the bottleneck layer. The larger the weight, the stronger the contribution of the corresponding design parameter will be. This strength is also shown in Fig. 11a by the thickness of the arrows that connect the nodes of the two layers. As shown in Fig. 11b, the height of the structure (h) plays an important role in changing the response compared to other design parameters as *h* connects to all 4 nodes in the bottleneck layer with reasonably strong weights. Moreover, the crystallization levels l_{c1}, l_{c2}, and l_{c3} can only change one of the reduced response features as they mainly connect to only one node (the purple node) in the bottleneck layer. As a result, as long as the total input to that purple node is fixed, the response will stay the same regardless of how the values of l_{c1}, l_{c2}, and l_{c3} change. This conclusion is reached by assuming a small error in training the pseudo-encoder and neglecting the small weights (or arrows in Fig. 11a) that connect l_{c1}, l_{c2}, and l_{c3} to the nodes of the bottleneck layer. To test this conclusion, we vary l_{c1}, l_{c2}, and l_{c3} while keeping their weighted sum (according to the trained pseudo-encoder) and all other 7 design parameters for the MS in Fig. 6 constant, and we calculate the response of the MS using brute-force COMSOL simulations (no pseudo-encoder intervention). The results for two different weighted sums of l_{c1}, l_{c2}, and l_{c3} are shown in Fig. 11c. Figure 11c clearly confirms our observation from the trained pseudo-encoder that l_{c1}, l_{c2}, and l_{c3} effectively act as one design parameter (through their weighted sum). Figure 11d shows the results of COMSOL simulations when the structure height (*h*) is changed while keeping all other 9 design parameters fixed. The large range of variation of the response in Fig. 11d clearly shows the importance of h as a design parameter. It is interesting to see from Fig. 11d that different responses for different values of h have low correlation while the responses for different values of the weighted sum of l_{c1}, l_{c2}, and l_{c3} (i.e., blues curves and red curves in Fig. 11c) show a similar trend with different locations of peaks and valleys. This suggests that the parameter *h* can be used to obtain different classes of responses while the weighted sum of l_{c1}, l_{c2}, and l_{c3} can be used to finely tune a given class of response. The details of the design parameters for each case are shown in Table 2.

The important observations about the role of different design parameters were obtained from our deep-learning approach without taking any information about the physics of the structure into account. Nevertheless, these observations agree with the physical intuition about the structure in Fig. 6. Each unit cell in this structure is composed of three plasmonic building blocks formed between the Au layer underneath and each Au nanoribbon on the top GST layer (see Fig. 6). Since the supermode of each building block is formed by coupling of the surface plasmon polaritons at the two Au layers, its properties strongly depend on the height of the GST layer (h), which directly controls the coupling strength.[57] Thus, strong dependence of the MS response on *h* is expected. Figure 12 shows the electric field patterns for the unit cell structure in Fig. 6 for two different values of h, confirming the strong dependence of the spatial mode profile on h.

To consider the effect of variation of the crystallization fractions (l_{c1}, l_{c2}, and l_{c3}), we note that the reflection response of the overall MS is essentially the sum of three responses defined by the three plasmonic resonators in each unit cell. By combining three wideband resonances with different resonance wavelengths, a wideband reflection response is obtained. Figure 12 shows the variation of the reflectance of the MS with frequency for a given set of l_{c1}, l_{c2}, and l_{c3} values (0.6, 0.7, and 0.8, respectively). The inset

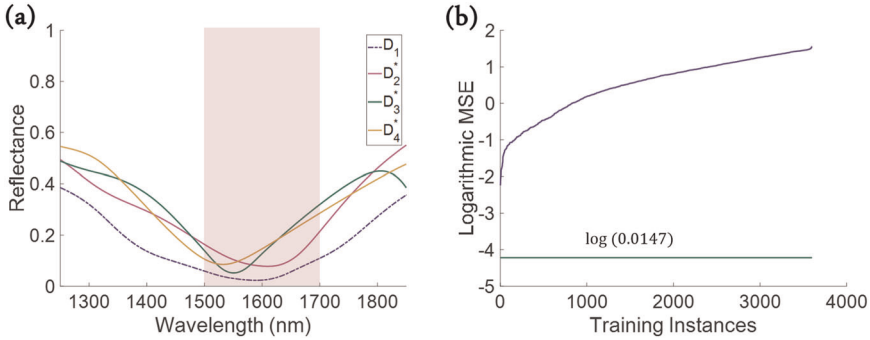

Fig. 10 Representation of the optimal/sub-optimal responses and the MSEs for all training datapoints. a Responses of the optimal (represented by dashed line) found by our approach and three best designed structures in the training dataset. **b** Logarithmic MSE for all training data (best one to worst one sorted from 1 to 3600) and for the optimal structure found by our technique.

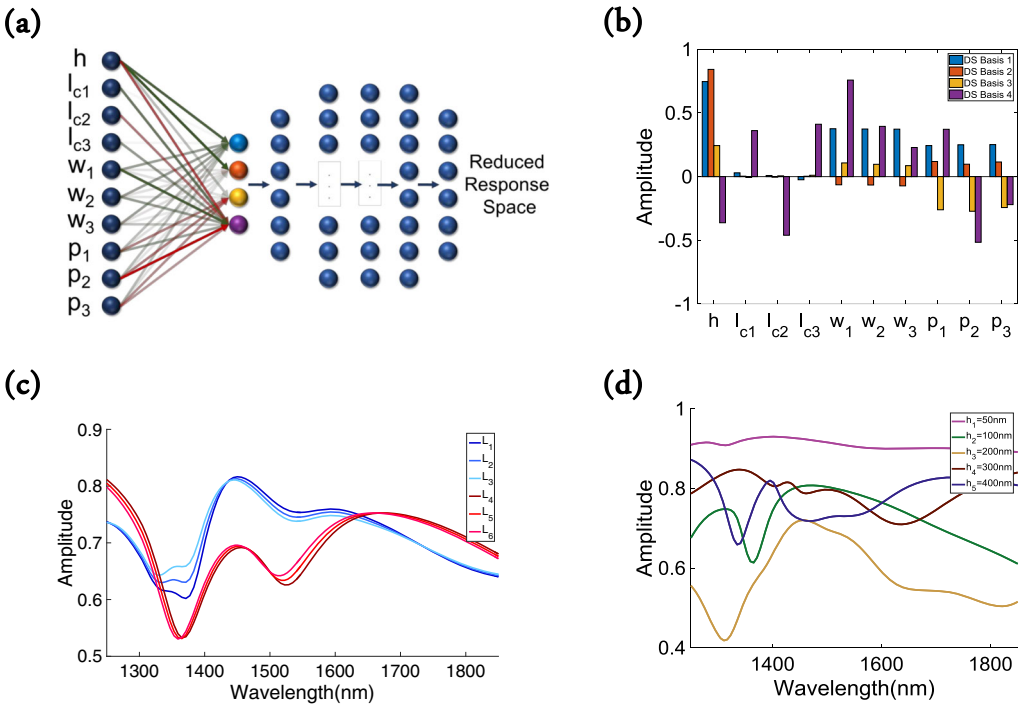

Fig. 11 Leveraging the developed algorithm as toolkit for knowledge discovery. a The pseudo-encoder architecture trained for the problem in Fig. 6, which relates the original design space to the reduced response space while reducing the dimensionality of the design space Weights in a color map for the one-layer DR of design space. Color corresponds to the node. h has strengths to all 4 nodes, but l_{c1}, l_{c2}, and l_{c3} primarily connect to the purple. **b** Strength of the weights which are connected to different design parameters. **c** Red curves correspond to three different sets of l_{c1}, l_{c2}, and l_{c3} where their weighted sum for the purple node is the same. All other parameters are fixed. Blue: similar three curves but for a different weighted sum. **d** Variation of the response using COMSOL (no NN) where *h* varies and all parameters are fixed.

shows the field profiles of the three plasmonic resonators within each unit cell at different wavelength regimes.

Due to pronounced light-matter interaction of the supermode with the GST layer at the higher wavelength (i.e., 1650–1850 nm), we expect that most of the resistive loss occurs in the building block with high crystallization level (i.e., l_{c3}) accommodating more free charge carriers. This effect is clarified in the inset of Fig. 12 at higher wavelengths (red border) showing that a good portion of absorption takes place in the rightmost building block (i.e., l_{c3}). Figure 12 also shows that the absorption loss in the middle wavelength window (i.e., 1450–1650 nm, shown by green) occurs mostly in the building block with lower crystallization level (i.e., the leftmost building blocks with (i.e., l_{c1} and l_{c2}). Finally, Fig. 12 shows similar contributions from the three building blocks at lower wavelengths (e.g., 1250–1450 nm). This is due to the fact that by increasing the level of crystallization in this regime, the optical constant of GST varies significantly leading to decrease in the light-matter interaction. This explains the collective role of l_{c1}, l_{c2}, and l_{c3} observed through training the pseudo-encoder. Note that obtaining this observation from the basic device properties was not as trivial as that of the role of *h*.

While some of the conclusions about the role of design parameters in Fig. 6 obtained by training the pseudo-encoder could also be obtained by the underlying mode properties of the DL-based (e.g., by analyzing the modes of the plasmonic resonators), the ability of our approach in providing useful information about the physics of wave-matter interaction in non-trivial structures (e.g., nonlinear and dispersive metamaterials) will be extremely valuable. Indeed, by using this approach to find and understand new phenomena in such non-trivial structures,

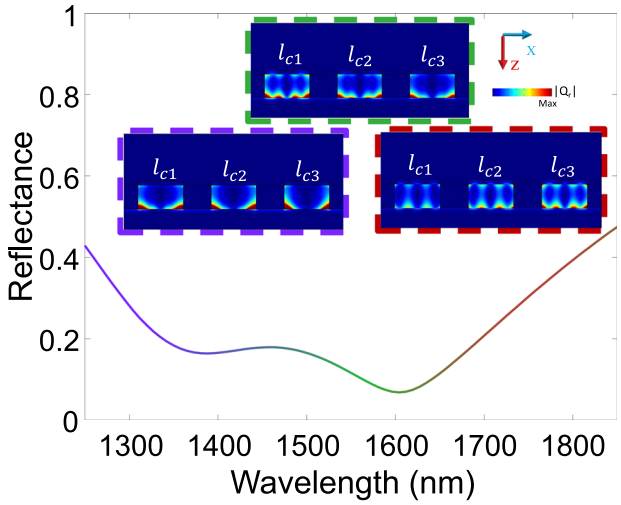

Fig. 12 Simulated absorption spectra for the proposed broadband near-perfect absorber. Inset shows the resistive loss corresponding to the lower wavelengths (i.e., 1250–1450 nm), intermediate wavelengths (i.e., 1450–1650 nm), and higher wavelengths (i.e., 1650–1850 nm) of the curve. l_{c1}, l_{c2}, and l_{c3} represent 0.6, 0.7, and 0.8 crystallization level of the GST layer, respectively.

new ideas for forming new classes of devices can be generated. This is a major advantage of our approach over all existing design approaches, especially those that rely on multiple brute-force simulations of the structure for different design parameters.

Note that the MSE (about 10%) in relating the design and response spaces using the mono-layer pseudo-encoder in Fig. 12a is larger than that of a more complex pseudo-encoder (e.g., that in Fig. 5). Nevertheless, the intuitive understanding of the roles of the design parameters achieved with the simple pseudo-encoder is completely valid. While the mono-layer structure provides simple and helpful information about the roles of the design parameters, more sophisticated relations (and physics) can be learned by using a pseudo-encoder with more layers and studying the NN weights in different layers. It is also evident that the algorithms selected for understanding the physics of the wave-matter interaction are in general different from those used for design and optimization of the structure to achieve a desired response. In the latter the minimization of the MSE in the input-output relation is critical while in the former, it is of secondary importance.

DISCUSSION

Figure 9 and Table 1 clearly show the ability of our approach in designing MSs with considerably reduced computation complexity. Figure 9 obviously verifies the good modulation depth (between ON- and OFF-state) of the optimized structure. They also show the importance of the understanding of the many-to-one nature of the design problem. Considering the many options for the original design parameters that correspond to a single set of reduced design parameters, we can easily enforce the fabrication restrictions and other design preferences in the last part of the design approach and find a set of design parameters that results in a close-to-optimal response. It is important to note that the availability of the analytic relation between the original and reduced design spaces makes the brute-force optimization (e.g., using analytic search) computationally feasible even for a large number of design parameters. Nevertheless, more sophisticated constrained optimization techniques can be used to solve the last (i.e., many-to-one) part of the design problem with explicit inclusion of the fabrication and other design-related constraints. Such techniques are currently under investigation and will be the subject of future publications.

A unique feature of our DR-based approach is the computation simplicity while appreciating the many-to-one nature of the problem. By not considering the latter explicitly, several other existing NN-based techniques are technically limited to only smooth-enough problems, or they require apriori assumptions to limit the search for the optimal solution in to a given region in the design space, where the relation to the response space is one-to-one (as discussed in "Introduction"). Nevertheless, by reducing the dimensionality of the problem, our approach requires less computation than any other alternative. For example, in the design problem studied here, we reduced a 10×200 dimensional problem to a 5×10 one.

Compared to brute-force optimization approaches (e.g., exhaustive search), our technique requires far less computation. For example, by assuming only 10 possible values for the 7 analog variables (i.e., h, w_1, w_2, and w_3, p_1, p_2, and p_3) and 11 values for the discrete ones (i.e., l_{c1}, l_{c2}, l_{c3}) in the design problem in Fig. 6, the exhaustive search algorithm requires the complete EM simulation of the structure for more than 10^{10} times, which is essentially intractable. However, our optimization requires only 4000 EM simulations along with the training process that requires far less computations. Indeed, the entire training of the forward and inverse parts of the platform in Fig. 5 for the design problem in Fig. 5 (results are shown in Fig. 9) took less than 3 hours using a simple personal computer with a 3.4 GHz core i7-6700 CPU and 8 GB of random access memory (RAM).

It is important to note that the key computation advantage of our proposed technique is for the complete optimization process. Here, the roles of DR and the training of the autoencoder and the pseudo-encoder are to: (1) convert a large non-one-to-one problem into a combination of a large one-to-one problem and a small non-one-to-one problem, and (2) enable a reliable approach for finding the global optimum without requiring intractable brute-force approaches, and (3) provide intuitive information about the roles of design parameters to form a smarter training set to further reduce computation. As a result, we are able to demonstrate an optimization technique, which requires far less computation than the existing techniques (e.g., brute-force approaches, evolutionary approaches, and simple neural-network-based approaches).

The simulation of the structure for achieving training data is indeed the most computationally intense part of the solution. The main advantage of the DR technique is to avoid using the conventional brute-force (e.g., exhaustive search) or evolutionary techniques. In addition, compared to techniques based on training a conventional NN to solve the problem, the DR technique is superior by: (1) addressing the many-to-one issue and making a more reliable path for design and optimization, (2) providing intuitive information about the dynamics of the problem, and (3) providing intuitive information about the relative importance of different design parameters, which can be used to form a smarter grid for generation of the training data (i.e., using less simulation of the electromagnetic problem), and (4) requiring less computation for training (under the same size of the training data) by breaking the training process into two steps of training the autoencoder and training the pseudo-encoder, which have considerably fewer number of nodes.

While the role of DR techniques in reducing the computation complexity and the severity of the nonuniqueness challenge is clear, there is a possibility that the relation between the reduced design space and the response space in Fig. 5d remains mildly many-to-one (although much more manageable than that between the original design space and the response space). The performance of this technique becomes closer to one-to-one once the optimal dimension for the reduced design space is selected. While this is currently performed using trial and error, more rigorous approaches for improving this property of the DR techniques should be considered in future. Nevertheless, by treating the last stage of

solution (i.e., from the reduced design space to the original design space) as a many-to-one problem, the risk of missing viable solutions is highly reduced compared to existing techniques. A rigorous mathematical study of the conditions for the dimensionality of the reduced spaces from machine-learning point of view can provide more solid guidelines in selecting the dimensionality of the reduced spaces. However, such a rigorous mathematical study is outside the scope of this paper.

To show the effect of the DR on the non-uniqueness challenge, we select four totally different sets of design parameters (d_1, d_1^*, d_2, and d_2^*) with each two (i.e., d_1 and d_1^* on one hand and d_2 and d_2^* on the other hand) result in almost identical responses (see Fig. 13). We define a normalized distance metric (Δd_n) to evaluate the ability of the pseudo-encoder in solving the non-uniqueness challenge.

$$\Delta d_n(i,j) = \binom{2}{N} \frac{||d_i - d_j||_2^2}{\sum_{k,l \atop k \neq l} ||d_l - d_k||_2^2}, \tag{2}$$

where N and Δd_n (i, j) represent the number of training samples and the normalized distance between the two design sets i and j, respectively. In the ideal case, we expect Δd_n to become zero in the reduced design space for original designs with the same response (e.g., d_1 and d_1^*). Our calculations show that Δd_n drops by a factor of 30 (0.83–0.024 for the distance between d_1 and d_1^*, and from 1.72 to 0.07 for the distance between d_2 and d_2^* after reducing the dimensionality of the design space using the pseudo-encoder. We tried this study with several cases and found similar reduction in Δd_n for structures with similar responses. Note that the little difference between the actual responses of the two designs (e.g., for d_1 and d_1^* in Fig. 13) contributes partially to the non-zero Δd_n. This clearly shows that the DR performed by the pseudo-encoder is capable of reducing the severity of the non-uniqueness problem considerably.

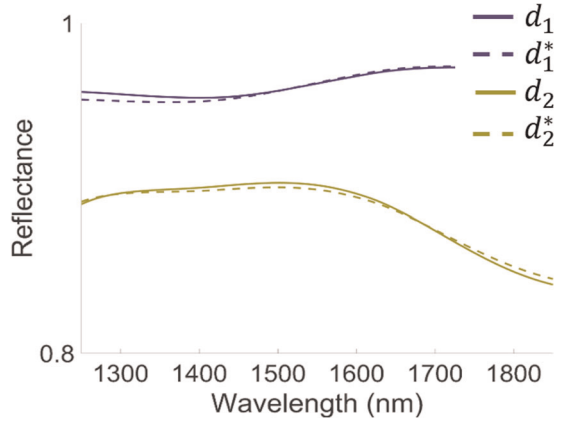

Fig. 13 Comparison of the mutual distances in original and reduced spaces. Four totally different set of design parameters (d_1, d_1^*, d_2, and d_2^*) which each two results in almost identical responses. The Δd_n and the values of the design parameters are shown in Table 2.

Table 2. The design parameters and the resulting MSE for the optimal design and three good designs for the structure in Fig. 6 to achieve maximum absorption in the 1500–1700 nm wavelength region. h, w_i, and p_i {$i = 1, 2, 3$} are in nm.

Design	h	l_{c1}	l_{c2}	l_{c3}	p_1	p_2	p_3	w_1	w_2	w_3	MSE
D_1	190	0.5	0.6	0.7	650	650	550	350	500	200	0.0147
D_2	190	0	0.2	0.8	650	650	350	450	250	250	0.0149
D_3	190	0.5	0.1	0.7	650	450	450	200	350	300	0.0152
D_4	190	0.3	0.6	0.8	650	550	550	250	300	450	0.0172

Note that the level of computation (for training, finding the inverse network, and moving from the reduced design space to the original design space) for this approach depends on the problem complexity. It is expected that structures with sharp spectral features require more training instances to converge. Also, in extreme cases of responses with radically varying spectral (or spatial) features and very well selected design parameters (with not much redundancy in the design and response spaces), the DR technique may not reduce the size of the overall network considerably. Nevertheless, even for such extreme cases, this technique can solve the non-uniqueness issue and result in the global optimum although at less computation advantage compared to majority of the mainstream design cases. To show the ability of our technique to design structures with sharp features, we have applied it to several such structures, and the results are shown in the Supplementary Information. In addition, as an experimental proof-of-concept, a simpler version of the investigated reconfigurable MS here was fabricated and tested. The measured reflection responses are in agreement with those results obtained using the DR technique (see Supplementary Information).

While the selected structure in Fig. 6 has enough complexity through 10 design parameters to show the capability of the DR technique, our technique can be used for studying far more complex structures with a reasonable number of training data. To push the limits of applicability of this technique with reasonable computation, development of more intelligent sampling techniques for reducing the number of simulations for obtaining training data will be helpful.

In summary, we demonstrated here a new DL-based approach for the design of EM nanostructures with a wide range of design possibilities. We showed that by reducing the dimensionality of the response and design spaces using an autoencoder and a psuedo-encoder, we can convert the initial many-to-one problem into a one-to-one (or in the worst case, close to one-to-one) problem plus a simple one-to-many problem that can be solved using brute-force analytical formulas. The resulting approach considerably reduces the computational complexity of both the forward problem and the inverse (or design) problem. In addition, it allows for the inclusion of the design restrictions (e.g., fabrication limitations) without adding computational complexities. It also provides valuable information about the roles of design parameters in the response of the EM structure, which can potentially enable new phenomena and devices. Finally, this technique can be extended to solve many different optimization problems in a wide range of disciplines as long as enough data for training the incorporated NNs are provided.

METHODS

The full-wave EM simulations were carried out using the finite element method (FEM) enabled by linking the commercial software package COMSOL Multiphysics 5.3 (wave optics module) with MATLAB to expedite the design, optimization, and analysis processes. Floquet periodic and perfectly-matched layer boundary conditions were used along transverse x-axis and the z-axis in Fig. 6, respectively. The structure was assumed infinite along the y-direction. A linearly polarized planewave of light, excited the MS in the wavelength range of 1250–1850 nm. The refractive index (n) and absorption coefficient (k) data for amorphous and crystalline GST, Au, and SiO_2 were obtained from the literature.[58,59] The computation domain was meshed using triangular elements with a maximum size of λ_0/(10 n) in SiO_2 and GST, and of λ_0/50 in Au with λ_0 being the free-space wavelength. The effective dielectric constant associated with the intermediate states of GST were approximated via the effective medium theory. Among different options, Lorentz-Lorenz formula more accurately describes effective permittivity $\epsilon_{eff}(\lambda_0)$ as:[60]

$$\frac{\epsilon_{eff}(\lambda_0)-1}{\epsilon_{eff}(\lambda_0)+2} = l_c \times \frac{\epsilon_c(\lambda_0)-1}{\epsilon_c(\lambda_0)+2} + (l_c - 1) \times \frac{\epsilon_a(\lambda_0)-1}{\epsilon_a(\lambda_0)+2}, \tag{3}$$

where $\epsilon_c(\lambda_0)$ and $\epsilon_a(\lambda_0)$ are the permitivitties of the crystalline and amorphous GST, respectively, and l_c, ranging from 0 (amorphous) to 1 (crystalline), is the crystallization fraction of GST.

DATA AVAILABILITY

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

CODE AVAILABILITY

An implementation of the algorithm described in the paper is available at https://github.com/PRGatech/DimensionalityReduction.

Received: 2 August 2019; Accepted: 9 January 2020;
Published online: 04 February 2020

REFERENCES

1. Melikyan, A. et al. High-speed plasmonic phase modulators. *Nat. Photonics* **8**, 229 (2014).
2. Zhu, T. et al. Plasmonic computing of spatial differentiation. *Nat. Commun.* **8**, 15391 (2017).
3. Rodrigo, D. et al. Mid-infrared plasmonic biosensing with graphene. *Science* **349**, 165–168 (2015).
4. Liu, X. et al. Taming the blackbody with infrared metamaterials as selective thermal emitters. *Phys. Rev. Lett.* **107**, 045901 (2011).
5. Huang, L. et al. Three-dimensional optical holography using a plasmonic metasurface. *Nat. Commun.* **4**, 2808 (2013).
6. Yu, N. et al. Light propagation with phase discontinuities: generalized laws of reflection and refraction. *Science* **334**, 1210713 (2011).
7. Kildishev, A. V., Boltasseva, A. & Shalaev, V. M. Planar photonics with metasurfaces. *Science* **339**, 1232009 (2013).
8. Arbabi, A., Horie, Y., Bagheri, M. & Faraon, A. Dielectric metasurfaces for complete control of phase and polarization with subwavelength spatial resolution and high transmission. *Nat. Nanotechnol.* **10**, 937 (2015).
9. Khorasaninejad, M. et al. Metalenses at visible wavelengths: Diffraction-limited focusing and subwavelength resolution imaging. *Science* **352**, 1190–1194 (2016).
10. Jahani, S. & Jacob, Z. All-dielectric metamaterials. *Nat. Nanotechnol.* **11**, 23 (2016).
11. Lin, D., Fan, P., Hasman, E. & Brongersma, M. L. Dielectric gradient metasurface optical elements. *Science* **345**, 298–302 (2014).
12. Taghinejad, M. et al. Ultrafast control of phase and polarization of light expedited by hot-electron transfer. *Nano Lett.* **18**, 5544–5551 (2018).
13. AbdollahRamezani, S., Arik, K., Khavasi, A. & Kavehvash, Z. Analog computing using graphene-based metalines. *Opt. Lett.* **40**, 5239–5242 (2015).
14. Abdollahramezani, S. et al. Reconfigurable multifunctional metasurfaces employing hybrid phase-change plasmonic architecture. Preprint at https://arxiv.org/abs/1809.08907 (2018).
15. Sun, S. et al. High-efficiency broadband anomalous reflection by gradient metasurfaces. *Nano Lett.* **12**, 6223–6229 (2012).
16. Arbabi, A., Arbabi, E., Horie, Y., Kamali, S. M. & Faraon, A. Planar metasurface retroreflector. *Nat. Photonics* **11**, 415 (2017).
17. Decker, M. et al. High-efficiency dielectric huygens' surfaces. *Adv. Optical Mater.* **3**, 813–820 (2015).
18. Chen, W. T. et al. A broadband achromatic metalens for focusing and imaging in the visible. *Nat. Nanotechnol.* **13**, 220 (2018).
19. Molesky, S. et al. Inverse design in nanophotonics. *Nat. Photonics* **12**, 659 (2018).
20. Piggott, A. Y., Petykiewicz, J., Su, L. & Vučković, J. Fabrication-constrained nanophotonic inverse design. *Sci. Rep.* **7**, 1786 (2017).
21. Lu, J. & Vučković, J. Nanophotonic computational design. *Opt. Express* **21**, 13351–13367 (2013).
22. Su, L., Piggott, A. Y., Sapra, N. V., Petykiewicz, J. & Vuckovic, J. Inverse design and demonstration of a compact on-chip narrowband three-channel wavelength demultiplexer. *ACS Photonics* **5**, 301–305 (2017).
23. Frellsen, L. F., Ding, Y., Sigmund, O. & Frandsen, L. H. Topology optimized mode multiplexing in silicon-on-insulator photonic wire waveguides. *Opt. Express* **24**, 16866–16873 (2016).
24. Piggott, A. Y. et al. Inverse design and implementation of a wavelength demultiplexing grating coupler. *Sci. Rep.* **4**, 7210 (2014).
25. Englund, D., Fushman, I. & Vuckovic, J. General recipe for designing photonic crystal cavities. *Opt. Express* **13**, 5961–5975 (2005).
26. Seidel, S. Y. & Rappaport, T. S. Site-specific propagation prediction for wireless in-building personal communication system design. *IEEE Trans. Vehicular Technol.* **43**, 879–891 (1994).
27. Gondarenko, A. & Lipson, M. Low modal volume dipole-like dielectric slab resonator. *Opt. Express* **16**, 17689–17694 (2008).
28. HÅkansson, A. & Sánchez-Dehesa, J. Inverse designed photonic crystal demultiplex waveguide coupler. *Opt. Express* **13**, 5440–5449 (2005).
29. Ma, Y. et al. Ultralow loss single layer submicron silicon waveguide crossing for soi optical interconnect. *Opt. Express* **21**, 29374–29382 (2013).
30. Liu, D., Tan, Y., Khoram, E. & Yu, Z. Training deep neural networks for the inverse design of nanophotonic structures. *ACS Photonics* **5**, 1365–1369 (2018).
31. Peurifoy, J. et al. Nanophotonic particle simulation and inverse design using artificial neural networks. *Sci. Adv.* **4**, eaar4206 (2018).
32. Liu, Z., Zhu, D., Rodrigues, S., Lee, K.-T. & Cai, W. A generative model for the inverse design of metasurfaces. *Nano Lett.* **18**, 6570–6576 (2018).
33. Tahersima, M. H. et al. Deep neural network inverse design of integrated nanophotonic devices. *Sci. Rep.* **9**, 1–9 (2019). Preprint at https://www.nature.com/articles/s41598-018-37952-2.
34. Zhang, T. et al. Spectrum prediction and inverse design for plasmonic waveguide system based on artificial neural networks. *Photon. Research.* **7**, 368–380 (2019). Preprint at https://www.osapublishing.org/prj/abstract.cfm?uri=prj-7-3-368.
35. Ma, W., Cheng, F. & Liu, Y. Deep-learning enabled on-demand design of chiral metamaterials. *ACS Nano* **12**, 6326–6334 (2018).
36. Qu, Y., Jing, L., Shen, Y., Qiu, M. & Soljacic, M. Migrating knowledge between physical scenarios based on artificial neural networks. *Nano Lett.*, **6**, 1168–1174 (2019). Preprint at https://pubs.acs.org/doi/10.1021/acsphotonics.8b01526.
37. Inampudi, S. & Mosallaei, H. Neural network based design of metagratings. *Appl. Phys. Lett.* **112**, 241102 (2018).
38. Kabir, H., Wang, Y., Yu, M. & Zhang, Q.-J. Neural network inverse modeling and applications to microwave filter design. *IEEE Trans. Microw. Theory Tech.* **56**, 867–879 (2008).
39. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
40. Ciocarlie, M., Goldfeder, C. & Allen, P. K. Dimensionality reduction for hand-independent dexterous robotic grasping. *IEEE/RS International Conference on Intelligent Robots and Systems*, 3270 (2007). https://ieeexplore.ieee.org/abstract/document/4399227.
41. Bhowmik, T., Liu, H., Ye, Z. & Oraintara, S. Dimensionality reduction based optimization algorithm for sparse 3-d image reconstruction in diffuse optical tomography. *Sci. Rep.* **6**, 22242 (2016).
42. He, X., Yan, S., Hu, Y., Niyogi, P. & Zhang, H.-J. Face recognition using laplacianfaces. *IEEE Trans. pattern Anal. Mach. Intell.* **27**, 328–340 (2005).
43. Hinton, G. E., Dayan, P. & Revow, M. Modeling the manifolds of images of handwritten digits. *IEEE Trans. Neural Netw.* **8**, 65–74 (1997).
44. Efremenko, D., Doicu, A., Loyola, D. & Trautmann, T. Optical property dimensionality reduction techniques for accelerated radiative transfer performance: Application to remote sensing total ozone retrievals. *J. Quant. Spectrosc. Radiat. Transf.* **133**, 128–135 (2014).
45. Breger, A. et al. Supervised learning and dimension reduction techniques for quantification of retinal fluid in optical coherence tomography images. *Eye* **31**, 1212 (2017).
46. Kim, P. M. & Tidor, B. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.* **13**, 1706–1718 (2003).
47. Choi, S., Shin, J. H., Lee, J., Sheridan, P. & Lu, W. D. Experimental demonstration of feature extraction and dimensionality reduction using memristor networks. *Nano Lett.* **17**, 3113–3118 (2017).
48. Jolliffe, I. T. *Springer series in statistics* 29 (Springer-Verlag, New York, 2002).
49. Schölkopf, B., Smola, A. & Müller, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* **10**, 1299–1319 (1998).
50. Belkin, M. & Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**, 1373–1396 (2003).
51. Roweis, S. T. & Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326 (2000).
52. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning internal representations by error propagation (No. ICS–8506). California Univ San Diego La Jolla Inst for Cognitive Science (1985).
53. Wuttig, M., Bhaskaran, H. & Taubner, T. Phase-change materials for non-volatile photonic applications. *Nat. Photonics* **11**, 465 (2017).
54. Tuma, T., Pantazi, A., LeGallo, M., Sebastian, A. & Eleftheriou, E. Stochastic phase-change neurons. *Nat. Nanotechnol.* **11**, 693 (2016).
55. Ríos, C. et al. Integrated all-photonic non-volatile multi-level memory. *Nat. Photonics* **9**, 725 (2015).
56. Feldmann, J. et al. Calculating with light using a chip-scale all-optical abacus. *Nat. Commun.* **8**, 1256 (2017).
57. Maier, S. A. *Plasmonics: fundamentals and applications* (Springer Science & Business Media, 2007).
58. Shportko, K. et al. Resonant bonding in crystalline phase-change materials. *Nat. Mater.* **7**, 653 (2008).

59. Huang, Y.-W. et al. Gate-tunable conducting oxide metasurfaces. *Nano Lett.* **16**, 5319–5325 (2016).
60. Chu, C. H. et al. Active dielectric metasurface based on phase-change medium. *Laser Photonics Rev.* **10**, 986–994 (2016).

ACKNOWLEDGEMENTS

The authors thank Ali A. Eftekhar, M. Zandehshahvar, and O. Hemmatyar for helpful discussion. This work was funded by Defense Advanced Research Projects Agency (DARPA) (D19AC00001, Dr. M. Fiddy), and in part by the Office of Naval Research (ONR) (N00014-18-1-2055, Dr. B. Bennett).

AUTHOR CONTRIBUTIONS

Y.K. and S.A. contributed equally to this work. The initial idea was developed by Y.K. and S.A., and its implementation was discussed by all authors. Y.K. performed the training optimization of the autoencoder and the pseudo-encoder. S.A. developed the simulation results for training and validation, fabricated the sample, and conducted optical characterization. S.A. proposed the initial idea for the electromagnetic nanostructure for light absorption, which was discussed in further details by all authors. All authors participated in the data analysis, writing, and reading the paper. A.A. managed the project.

COMPETING INTERESTS

The authors declare no competing interest.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-0276-y.

Correspondence and requests for materials should be addressed to A.A.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.