UNIVERSITY OF CALIFORNIA

Los Angeles

Deep Learning for the Design and Characterization of Nanophotonic Materials and Structures

A dissertation submitted in partial satisfaction of the

requirements for the degree of Doctor of Philosophy

in Materials Science and Engineering

by

Christopher Yeung

2022

ABSTRACT OF THE DISSERTATION

Deep Learning for the Design and Characterization of Nanophotonic Materials and Structures

by

Christopher Yeung

Doctor of Philosophy in Materials Science and Engineering

University of California, Los Angeles, 2022

Professor Aaswath P. Raman, Chair

A central challenge in contemporary materials and photonics research is understanding how intrinsic materials properties can be optimally combined with nano- or micro-scale structuring to deliver a target functionality. By leveraging subwavelength nanostructures and the intrinsic dispersion of constituent materials, tailored changes in the amplitude and phase of incident wavefronts can be precisely engineered, along with desired spectral characteristics. However, our ability to meet increasing demands in the performance of photonic structures faces roadblocks due to the complexity of the materials and structural design spaces that are currently accessible. Conventional optimization methods, which rely on numerical simulations that solve Maxwell's equations, have shown remarkable capabilities in designing nanophotonic structures and are now commonly used. However, they can be computationally costly and are often intractable for large-scale designs or high-dimensional design spaces. As a result, data-driven approaches based on machine learning (ML) have been extensively explored in order to tackle challenging photonics design problems. To this end, this work explores the application of various advance deep learning methods for the design and characterization of nanophotonic materials and structures.

The dissertation of Christopher Yeung is approved.

Yinmin Wang

Ali Mosleh

Mathieu Bauchy

Aaswath P. Raman, Committee Chair

University of California, Los Angeles

2022

Table of Contents

Acknowledgements

I would like to thank a number of people who have supported me over the years, without whom this work would not have been possible. First, I would like to thank Professor Aaswath Raman for the opportunity to work in his lab for the past four years. His insight and guidance were irreplaceable, and I will carry the lessons I have learned in his lab with me for the rest of my life. I would also like to thank the Professors on my committee: Morris Wang, Ali Mosley, and Mathieu Bauchy for their questions and suggestions which have greatly influenced the direction of this work. I would like to thank my lab mates and contemporaries: Jin Xu, John Brewer, Xin Huang, Jae Hwang, David Abraham, Dr. Jyotiroy Mandal, and Dr. Parthiban Santhanam, all of who provided wonderful feedback which improved the quality of this work and taught me valuable skills outside of my domain. May you all continue to lead long successful careers and find fulfillment and happiness in the process. To all the undergraduate assistants and visiting students who supported me through the years: David Ho, Brian King, Yusaku Kawagoe, Julia Liang, Kara Levy, Nancy Zhang, Ryan Tsai, and Benjamin Pham, I hope your time with me was equally rewarding and I hope I was able to teach you some valuable technical skills as well as general life lessons. To all the members of Professor Sam Emaminejad's lab, you also played a pivotal role in shaping my career interests and trajectory. To Dr. Mark Knight and Dr. Katherine Fountaine, thank you for the opportunity to work with Northrop Grumman during my doctoral studies and for lending your technical expertise.

Finally, I would like to thank my family for their love, support, and encouragement which helped shape who I am today. To my parents, sister, and loving fiancée, this journey would not have been possible without you.

Vita

2011-2012	Undergraduate Research Assistant University of California Irvine, Irvine, California
2009-2013	Bachelor of Science, Mechanical Engineering, Materials Science and Engineering University of California Irvine, Irvine, California
2012-2013	Design Engineer Intern Southwestern Performance Technology, Inc., Santa Ana, California
2013-2014	Quality Assurance Engineer MSC Software Corporation, Newport Beach, California
2014-2017	Masters of Science, Materials Science and Engineering University of California Los Angeles, Los Angeles, California
2014-2018	Product Manager MSC Software Corporation, Newport Beach, California
2016-2020	Founder CY Printing Studio, Los Angeles, California
2018-2019	Graduate Student Researcher – Emaminejad Lab University of California Los Angeles, Los Angeles, California
2019-2022	Research Scientist Northrop Grumman Corporation, Redondo Beach, California
2018-2022	Graduate Student Researcher – Raman Lab University of California Los Angeles, Los Angeles, California
2019-2022	Teaching Assistant University of California Los Angeles, Los Angeles, California
2019	UCLA Graduate Division University Fellowship Award University of California Los Angeles, Los Angeles, California
2022	Recognized Reviewer Award Materials Letters, Optics Communications, Elsevier
2022	Member of Technical Staff - Data Analytics Foundry Engineer Advanced Micro Devices, Inc. (AMD)

Select Presentations and Publications

Yeung, C.; Tsai, J.; King, B.; Kawagoe, Y.; Ho, D.; Knight, M. W.; Raman, A. P. Elucidating the Behavior of Nanophotonic Structures through Explainable Machine Learning Algorithms. *ACS Photonics* **2020**, *7*, 2309–2318.

Yeung, C.; Tsai, J. M.; King, B.; Pham, B.; Ho, D.; Liang, J.; Knight, M. W.; Raman, A. P. Multiplexed supercell metasurface design and optimization with tandem residual networks. *Nanophotonics* **2021**, *10*, 1133–1143.

Yeung, C.; Tsai, R.; Pham, B.; King, B.; Kawagoe, Y.; Ho, D.; Liang, J.; Knight, M. W.; Raman, A. P. Global Inverse Design across Multiple Photonic Structure Classes Using Generative Deep Learning. *Advanced Optical Materials* **2021**, *9*, 2100548.

> Resultant media coverage: UCLA Team Harnesses Machine Learning Imaging Technology for Materials Design. *UCLA Samueli School of Engineering* **2022**.
> https://www.mse.ucla.edu/ucla-team-harnesses-machine-learning-imaging-technology-for-materials-design/

Yeung, C.; Ho, D.; Pham, B.; Fountaine, K.; Zhang, Z.; Levy, K.; Raman, A. P. Enhancing Adjoint Optimization-Based Photonic Inverse Design with Explainable Machine Learning. *ACS Photonics* **2022**, *5*, 1577-1585.

Yeung, C.; Pham, B.; Tsai, R.; Fountaine, K.; Raman, A. P. DeepAdjoint: All-in-One Hybrid Global Photonics Inverse Design Framework Combining Machine Learning with Electromagnetic Optimization Algorithms. *In Review* **2022**.

Yeung, C. Explaining Adjoint Shape Optimization for Electromagnetic Design, *SPIE Photonics West*, San Francisco, California. January **2022**.

Yeung, C. Conditional Machine Learning-Based Inverse Design Across Multiple Classes of Nanophotonic Structures. *Conference on Lasers and Electro-Optics* (*CLEO)*, Los Angeles, California, May **2021**.

Yeung, C. Elucidating the Physics of Nanophotonic Structures Through Explainable Machine Learning Algorithms. *Frontiers in Optics / Laser Science (FiO+LS)*, Los Angeles, California, September **2020**.

Yeung, C. Inverse Design of Nanophotonic Structures with Interpretable Convolutional Neural Networks. *Photonics Online Meetup*, Los Angeles, California, January **2020**.

1. Introduction

Nanophotonics is the study of how light, or more generally electromagnetic radiation, interacts with nanoscale materials and structures to obtain systems or devices with optical properties that are not found in nature. Metasurfaces, for instance, hold the potential to become a vital component for many next-generation optical technologies due to their ability to manipulate the propagation of light within an ultracompact footprint [1]. More broadly, by leveraging subwavelength nanostructures and the intrinsic dispersion of constituent materials, tailored changes in the amplitude and phase of incident wavefronts can be precisely engineered, along with desired spectral characteristics. This new level of control has enabled and accelerated critical developments in fields such as flat optics [1-3], quantum communications [4], and holography [5,6]. However, our ability to meet increasing demands in the performance of metasurfaces, and photonic structures in general, faces roadblocks due to the complexity of the materials and structural design spaces that are currently accessible.

Conventional optimization methods, which rely on numerical simulations that solve Maxwell's equations, have shown remarkable capabilities in designing nanophotonic structures and are now commonly used [7]. However, they can be computationally costly and are often intractable for large-scale designs or high-dimensional design spaces [8,9]. As a result, data-driven approaches based on artificial intelligence (AI) and machine learning (ML) have been extensively explored in order to tackle challenging photonics design problems [10,11]. Current state-of-the-art machine learning methods involve training neural networks to learn the underlying relationships between photonic structures and corresponding optical phenomena. A trained neural network can, in principle, instantaneously generate designs with substantially lower computational costs than optimization-based methods. A wide range of neural network and machine learning architectures have been investigated for the design and characterization of materials [12-15]. In the photonics context, one-dimensional (1D) tandem networks were used to design core-shell nanoparticles [16], multilayer thin films [17], and supercell-class metasurfaces [18]. However, such network architectures are only applicable to simple photonic structures for which geometric and material properties can be described by a vector of discrete parameters [19]. In contrast, photonic devices with complex freeform geometries cannot be well-represented by discrete variables, but offer the potential to achieve new functionalities and greater device performance [20]. For these structures, image-based generative networks have successfully designed various types of metasurfaces, including ones with silver [21], gold [22], or silicon [23] meta-atoms and other topological features. Further studies have combined image-based ML with optimization algorithms to yield even greater model performance [24,25].

To contribute to the field of AI/ML for materials design and optimization, this dissertation presents a collection of works, where we explore and apply a wide range of deep learning methods for the design and characterization of nanophotonic materials and structures. This goal entails addressing several contemporary research challenges that remain in this particular field, including global inverse design across multiple photonic structure classes, handing the inverse design of multiplexed supercell cell structures with numerous subunit elements, and explaining the behavior of photonic structures through explainable artificial intelligence. Each project is divided into individual chapters, where specific examples or problems are introduced, then a corresponding method is proposed which addresses said problems. Key results of each method are presented, and potential future directions are discussed. Broader implications and fundamental contributions of this work lie in the exploration and evaluation of AI/ML algorithms for materials optimization, which are applicable to a number of materials-related domains beyond nanophotonics and optics.

2. Global Inverse Design Across Multiple Photonic Structure Classes Using Generative Deep Learning

2.1 Introduction

Understanding how nano- or micro-scale structures and material properties can be optimally configured to attain specific functionalities remains a fundamental challenge. Photonic metasurfaces, for instance, can be spectrally tuned through material choice and structural geometry to achieve unique optical responses. However, existing numerical design methods require prior identification of specific material-structure combinations, or device classes, as the starting point for optimization. As such, a unified solution that simultaneously optimizes across materials and geometries has yet to be realized. To overcome these challenges, we present a global deep learning-based inverse design framework, where a conditional deep convolutional generative adversarial network is trained on colored images encoded with a range of material and structural parameters, including refractive index, plasma frequency, and geometric design. We demonstrate that, in response to target absorption spectra, the network can identify an effective metasurface in terms of its class, materials properties, and overall shape. Furthermore, the model can arrive at multiple design variants with distinct materials and structures that present nearly identical absorption spectra. Our proposed framework is thus an important step towards global photonics and materials design strategies that can identify combinations of device categories, material properties, and geometric parameters which algorithmically deliver a sought functionality.

From the perspective of a researcher or practitioner in the field, enabling a desired set of optical characteristics today typically involves a prior understanding of the capabilities of different categories of devices or nanostructures. For instance, ultra-strong field confinement may lead one to start with a plasmonic architecture, while high transmission applications would lead one to ensure the use of materials that present low extinction coefficients in the wavelength range of operation. Designing photonic structures that meet application-specific objectives thus entails identifying the ideal intersection of material properties, structural composition, and fabrication process (or device class), as specific combinations are more likely to yield desired functional characteristics. It is only once a device or photonic structure category has been identified that numerical optimization methods typically enter the picture to optimize and refine performance characteristics.

Despite the significant progress in image-based photonics design, existing studies are limited to designing the two-dimensional structural topology (or geometry) for a single class of metasurface or nanophotonic structure. In addition, the material properties and out-of-plane parameters (*e.g.*, layer thicknesses) of the explored structures are typically held constant. The central limitation identified earlier remains: prior knowledge of which category of structures or devices may deliver a specific functionality is needed before initiating the optimization procedure (whether machine learning-based or otherwise). However, human intuition on the optimal nanostructure category — the initial conditions for a numerical optimization procedure — can often go awry when faced with competing design goals. Thus, a unified 'global' materials and photonics inverse design approach that can define both the materials and structure (beyond 2D) across multiple classes of photonic structures has yet to be demonstrated, but could fundamentally change how we approach the design and optimization of photonic structures and metamaterials. Moreover, such a capability could prove critical to the design of nonlinear and phase-changing platforms where optical response depends heavily on material composition and fabrication process [26].

In this study, we present an image-based deep learning framework for the inverse design of photonic structures across multiple materials and device categories. Our approach combines the advantages of material property and structural parameter prediction enabled by 1D tandem networks, with the freeform design capabilities of image-based deep learning. This is accomplished through a versatile image-encoding technique where material and structural parameters such as refractive indices, plasma frequencies, layer thicknesses, resonator geometries, and metasurface classes are embedded within the discrete 'RGB' channels of colored images. Although we show multiparametric encoding through different shades of color in a 3D array (as an initial demonstration), we note that this information can also be encoded via higher-dimensional matrices or data structures that extend beyond the 'RGB' color system. The encoded images are used to train a customized conditional deep convolutional generative adversarial network (cDCGAN), which we evaluate by inputting a variety of target absorption spectra. In response to the input spectra, the network generates corresponding metasurface designs that are validated through full-wave electromagnetic (EM) simulations. To determine network accuracy, performance, and generalizability, the simulated spectra are compared to the input targets. Through this process, we demonstrate that the network simultaneously optimizes the material properties and 2.5D structuring across multiple classes of metasurfaces, thus validating the feasibility of a global inverse design framework that accounts for all the parameters which govern the optical behavior of photonic structures. We note that 'global' in this context refers to the network's ability to perform a global search within the surveyed design space [8,26], which includes material properties and freeform topology, but the network does not guarantee that the final generated device is globally optimal.

2.2 Results and Discussion

We consider two classes of absorbing metasurfaces in developing and demonstrating our inverse design approach (Figure 1a). First, we consider metal-insulator-metal (MIM) structures, where a thin dielectric layer is sandwiched between two metal layers (one uniformly deposited and the other lithographically patterned). This class of metasurface exhibits a relatively broad Lorentzian-shaped absorption response supported by each individual resonator, which renders this type of structure highly-amenable to thermal emission and energy harvesting applications [27,28]. Next, we consider hybrid dielectric metasurfaces with a metal film substrate, which take advantage of a cavity effect to produce an asymmetric, narrow-band Fano resonance that is well-suited for optical sensing and detection [29].

As seen in Figure 1b, the first step of our encoding method involves capturing the planar geometries (*G*) and material properties of the metasurface resonator (*M*), followed by the thicknesses of the dielectric layer (*T*), for both MIM and hybrid dielectric metasurfaces. We then encode *G, M,* and *T* into the red, green, and blue channels of a colored image. Within our encoding scheme, the red-channel represents the plasma frequency ($M=\omega_P$) and shape of the metal resonator in an MIM structure. The green-channel represents the real refractive index ($M=n$) and shape of the dielectric resonator in a hybrid dielectric structure. The remaining pixels in the blue-channel are used to define the thickness of the dielectric layer (in nanometers) for both metasurface classes. Thus, a red-blue color scheme indicates MIM structures while green-blue indicates hybrid dielectric structures (red-green image combinations are undefined). With this strategy, in addition to representing resonator geometry, different colors on an image can be used to describe unique combinations of material and structural parameters, which in turn yield significantly more variation in achievable optical responses than single-material approaches.

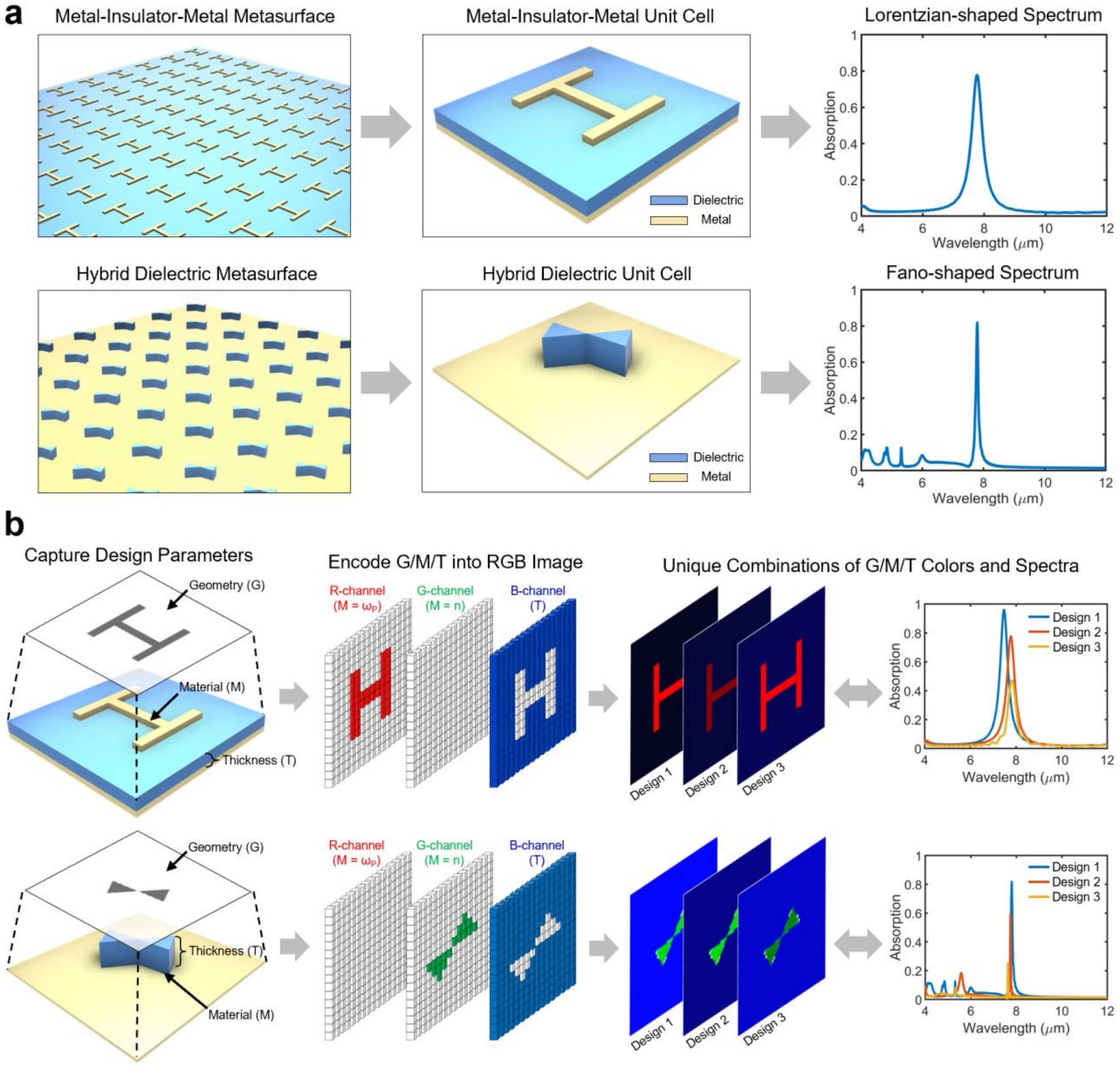


Figure 1. (a) MIM and hybrid dielectric metasurfaces with Lorentzian-shaped and Fano-shaped absorption responses, respectively. (b) Representing distinct classes of metasurfaces as color-encoded images. Metasurfaces are converted into images representing their planar geometries. Material properties, thickness values, and metasurface class are encoded into the images as various shades of color, allowing more degrees of freedom for metasurface design.

Though the described material properties can be denoted by individual values instead of entire image channels, the presented channel-encoding method offers several key advantages.

First, it combats the well-known noise-related artifacts found in image-based ML techniques such as generative adversarial networks (GANs) [8,21] by ensuring that the encoded properties are appropriately weighted towards the network's final predictions. A detailed analysis of models trained on several property-encoded neurons versus models trained on whole image channels is found in the Supporting Information. Additionally, in principle, our approach only requires small modifications to the input dimensions of an existing model (*e.g.*, changing from a 64×64 to 64×64×3 matrix), which allows us to leverage existing model optimization and training techniques without significantly increasing training costs. Furthermore, the presented method is capable of representing spatially-varying material properties along the entire physical structure, which enables the design of 3D or complex gradient-index and metal alloy-based structures that are, in principle, amenable to existing fabrication methods [49]. A demonstration of this design capability is shown in Figure S6.

Our training dataset consists of 20,000 metasurface unit cell designs, represented as image-vector pairs, derived from seven shape templates: cross, square, ellipse, bow-tie, H, V, and tripole-shaped. Detailed information regarding these designs is found in Figure S1 of the Supporting Information. MIM and hybrid dielectric structures are captured within 3.2×3.2 μm2 and 7.5×7.5 μm2 unit cells, respectively. Each design was converted into a 64×64×3 pixel 'RGB' image using the rules established above. A single pixel therefore corresponds to a minimum feature size of 50 nm (MIM) and 120 nm (hybrid dielectric), which is well-within feasible fabrication range [21,50]. Furthermore, we employed a Gaussian filtering post-processing procedure (described in the Supporting Information) to enhance device performance and fabricability. Finite-difference time-domain (FDTD) simulations were performed on the designs (Lumerical) to obtain an 800-point absorption spectrum vector (from 4-12 μm) for each structure. Low quality designs (defined in the

Supporting Information) were removed from the training set to maximize the model's utility and performance [30]. Figure S2 illustrates the peak absorptions and resonance wavelengths of the spectra represented in the final training dataset.

During the color-encoding step, the Drude model plasma frequencies of the metal resonators (ω_P=1.91 PHz for gold [31], ω_P=2.32 PHz for silver [32], and ω_P=3.57 PHz for aluminum [33]) were used to encode the red channel, and the real refractive indices of the dielectric resonators (n=2.41 for zinc selenide [34], n=3.42 for silicon [35], and n=4.01 for germanium [35]) were used to encode the green channel. The encoded material properties are based on optical constants from the same mid-infrared wavelength range as the simulations. A range of dielectric thickness values (100 nm to 950 nm) were used for the blue channel. To support the 'RGB' color scheme, all encoded values were normalized from 0 to 255.

Using the encoded images, we trained our image-based deep learning model using a GAN-based architecture. GANs have been recognized as the best performing type of generative network [19]; a class of neural networks that can directly find multiple solutions to a given problem. Other types of networks that fall in this category include variational autoencoders (VAEs) [52] and mixture density networks (MDNs) [53]. Recent developments in GAN technology have led to numerous GAN-variants, including but not limited to: the Self-Attention GAN (SAGAN) [36], Deep Regret Analytic GAN (DRAGAN) [37], StyleGAN [38], Wasserstein GAN (WGAN) [39], and the Least Squares GAN (LSGAN) [40]. Here, as an initial proof of concept, we tested our framework using a modified cDCGAN architecture, as shown in Figure 2a. cDCGANs have previously been used to generate domain-specific images in response to input conditions [41-43]. Implemented in the PyTorch framework, the cDCGAN consists of a generator and a discriminator. Initially, batches of absorption spectra (y) are fed into the generator, along with a latent vector (z), to generate ‘fake’ images (*G*) that are similar to the ‘real’ images (*x*) from the training set. The latent vector is sampled from a random uniform distribution and allows the generator to map a probability distribution to a design space, thereby enabling a one-to-many mapping [26]. Both *G* and *x* are then fed into the discriminator (*D*), which attempts to distinguish the generated images from the real. Thus, the generator is trained to produce convincing images that deceive the discriminator, while the discriminator is trained not to be deceived — a competition which leads to the joint and stepwise improvement of both networks via their loss functions. These loss functions are calculated using the binary cross-entropy criterion, and the complete model interaction is represented as:

$$min_G\ max_D\ l(G,D)\ =\ E_{x\rightarrow p_{data}(x)}\{log\ D(x,y)\}\ +\ E_{z\rightarrow p_z(z)}\{log(1-D(G(z,y))\}, \quad (1)$$

where *E* is the expected result, *pdata(x)* is the training data distribution, *pz(z)* is the latent vector distribution, *log(D(x,y))* + *log(*1−*D(G(z,y))]* is the discriminator loss (*LD*), and *log(D(G(z,y)))* is the generator loss (*LG*). During training, the objective is to maximize *LD* and *LG*. We note that our definition of the *LG* differs from the original GAN implementation, where *log(*1−*D(G(z,y))* is minimized instead, since this was shown to not provide sufficient gradients [53,54]. To improve the performance of the cDCGAN, we applied one-sided label smoothing and mini-batch discrimination [44,45]. Unlike previous cDCGAN implementations, our approach relies on adversarial training without explicitly guiding the generator towards known images [21], thereby achieving a greater degree of generalization that is unconstrained by pre-existing images. Over 40 different cDCGAN architectures were trained through extensive hyperparameter tuning, and the optimized architecture can be found on Figure S3. Several alternative parameter-encoding schemes were also trained and presented in Figure S4, where models trained on several neurons (to represent encoded properties) were compared to models trained using the entire ‘RGB’ channels. The validation losses of each method are reported in Table S1 and S2, and the color-encoding approach is shown to exhibit the best performance among the tested encoding schemes. After training the cDCGAN, we developed an image processing workflow to convert the generated images into full 3D metasurface designs (Figure 2b). In this workflow, the material property (ω_P or *n)* and thickness values (t) are calculated by taking the average pixel-values in their respective channels (based on structure classification), then reversing the normalization performed in the encoding step. Additional details regarding this process can be found in the Supporting Information.

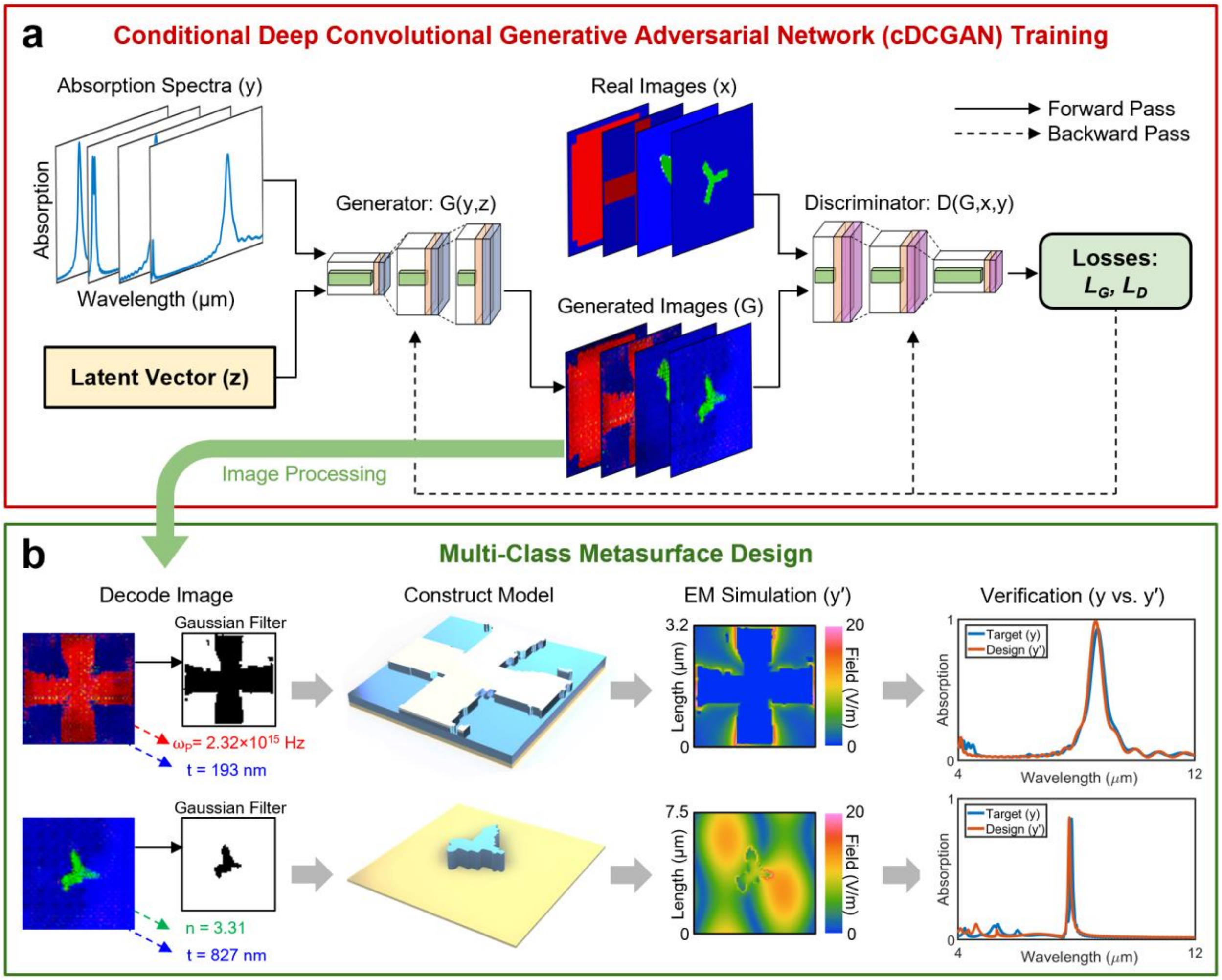


Figure 2. Schematic of the cDCGAN training and design process. (a) Both the generator and discriminator are neural networks that train in tandem to maximize the generator's accuracy. (b) After training, the generator can be used for multi-class metasurface inverse design. Images synthesized by the generator are decoded to construct 3D models of metasurfaces with unique material and structural parameters. The generated structures are then simulated to verify their adherence to the input target spectra.

In the GAN-metasurface design process, new materials were specified in the EM simulation software using the generated ω_P or n values. We note that new materials created in this manner may not be compatible with fabrication schemes which rely on conventional materials. However, the presented material definition scheme allows the model to freely predict a continuum of material properties that are otherwise lost or disregarded due to categorical approximations, which enables a wider range of material property-driven designs. For example, metamaterials using dielectrics embedded with custom nanoparticle formulations can yield materials with effective refractive indices that can be deterministically tuned [46-48,56]. Prior studies have also employed nanoscale metallic alloying to achieve tailored plasma frequencies.57 Highly granular material-level predictions, as we show are possible here, would therefore enable additional degrees of freedom for materials optimization, which may in turn yield novel optical responses.

We evaluated the performance of our trained cDCGAN and image processing method by inputting a set of absorption spectra (coupled with randomly sampled latent vectors) and analyzing the resulting designs. Since the GAN may produce a distribution of designs with potentially varying degrees of accuracy [8,51], ten different latent vectors were generated for each target spectrum, which were then used as inputs to the network. Each design is verified using numerical simulation, then the design (and corresponding latent vector) with the lowest mean-squared error to the target is reported as the final design. Figure S7 shows the distribution of designs (across different latent vectors) for several input targets, where we observe that each design variant has over 90% accuracy in comparison to the input target. Following this procedure, Figure 3 presents a series of tests performed with inputs that originate from the validation dataset (10% of the training dataset). Here, the blue lines represent randomly selected inputs (across both classes of structures), and the orange lines are the simulated spectra of the cDCGAN-generated designs. Images of the corresponding structures (direct outputs of the network) are shown to the right of each plot. Below each image are the associated material property (ω_P or n) and dielectric thickness values which are derived from the aforementioned decoding scheme. Figure S8 shows the equivalent results for inputs from the training dataset.

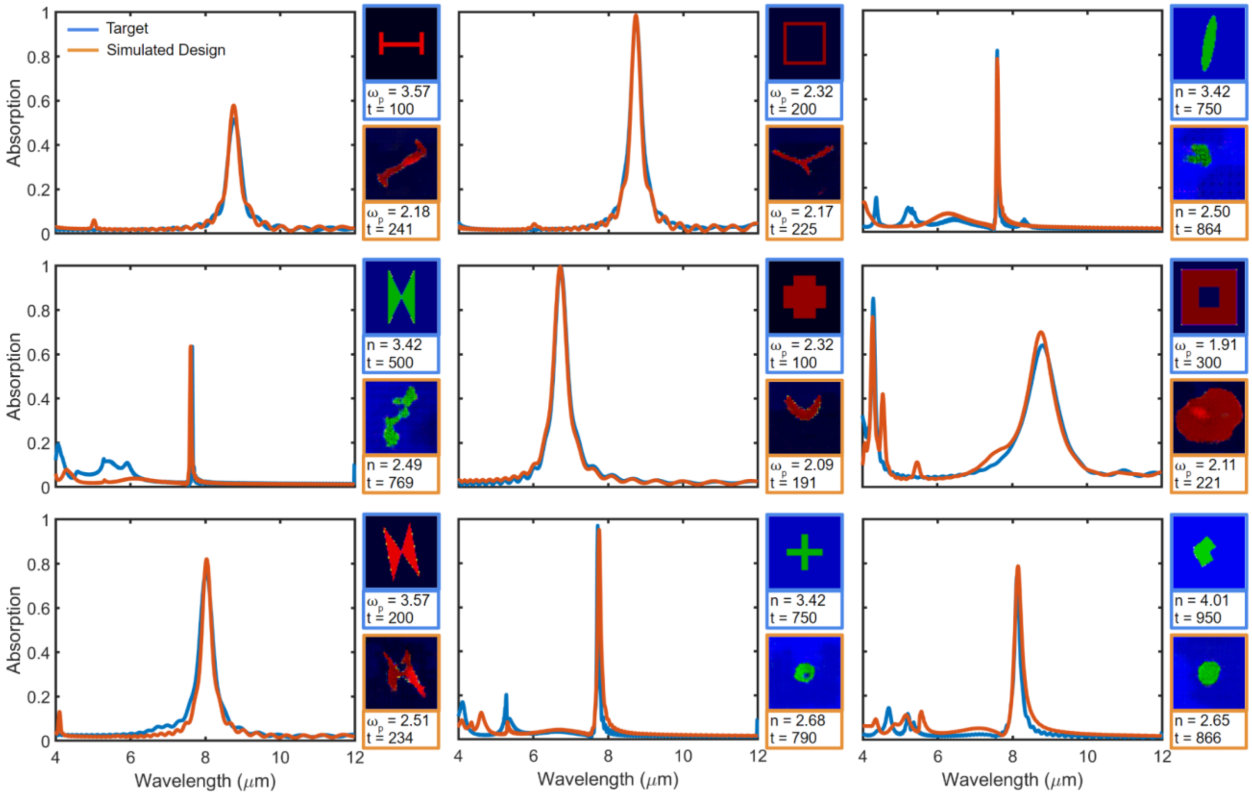


Figure 3. Randomly selected absorption spectra from the validation dataset (blue) which were designated as input targets for the cDCGAN. The simulated spectra of the cDCGAN-synthesized designs (orange) are plotted alongside the targets for comparison. Images representing the respective structures are shown to the right of each plot, with material and thickness information below each image. Units for plasma frequency (ω_P) values are in PHz and thicknesses (t) are in nanometers. The results here reveal that the network can identify the underlying relationships between structure, material, metasurface class, and optical response to provide new yet accurate solutions that extend beyond the known designs.

We observe that in each test case, the network predicted the class of structure that corresponds with its particular type of spectral response. Specifically, when Fano-shaped spectra of various hybrid dielectric structures were passed into the network, the network exclusively generated hybrid dielectric structures (or green-blue images). Similarly, Lorentzian-shaped inputs yielded only MIM structures (or red-blue images). The generated images suggest that the network was capable of: 1) learning the distinguishing features and optical responses between the two explored classes of metasurfaces, and 2) using this information to predict the appropriate class based on the nature of the input spectra. In addition, across a wide range of input spectra, we observe that the network synthesized designs that are noticeably different from the known structures (either in resonator shape or property/thickness). Despite this difference, the generated designs exhibit responses that strongly match the input targets. Thus, these results show that our network is not simply mimicking designs from the training dataset. To a degree, the cDCGAN is capable of learning the underlying relationships between structure, material, metasurface class, and optical response to provide new yet accurate design solutions that extend beyond the training data.

To assess our network's ability to solve arbitrarily-defined design problems, we tested the network using 'hand drawn' target spectra. These targets are derived from the Fano resonance and Lorentzian distribution functions and have no associated design or structure. We evaluated the cDCGAN's performance across a wide range of inputs by using each function to create 200 spectra with amplitudes ranging from 0.5-0.9, and resonance wavelengths ranging from 5-9 μm, for 400 total test spectra. Figure 4a and Figure 4b show several results of the Fano-shaped and Lorentzian-shaped targets, respectively, where a strong match between the targets and simulated designs can be observed. A statistical evaluation of the entire test dataset is reported in Figure 4c (for the Fano-shaped targets) and Figure 4d (for the Lorentzian-shaped targets). Here, the histograms illustrate the number of test spectra which reside in specific MSE value ranges. Dashed-red lines indicate the average mean-squared error (MSE) of the Fano-shaped and Lorentzian-shaped targets, which equal to approximately 8.5×10-3 and 2.9×10-3, respectively. Through these plots, we note that the accuracy of the Fano targets is lower than the accuracy of the Lorentzian targets. However, further analysis of the training dataset (Figure S2) and the individual test results (Figure S5) reveal that the low-accuracy regions of the Fano-shaped structures correspond to regions that are not well-represented by the training data, whereas the high accuracy of the Lorentzian-shaped spectra can be explained by the wide spectral range of the MIM structures. Therefore, the performance of the Fano-shaped designs can potentially be improved by expanding the training data and design space.

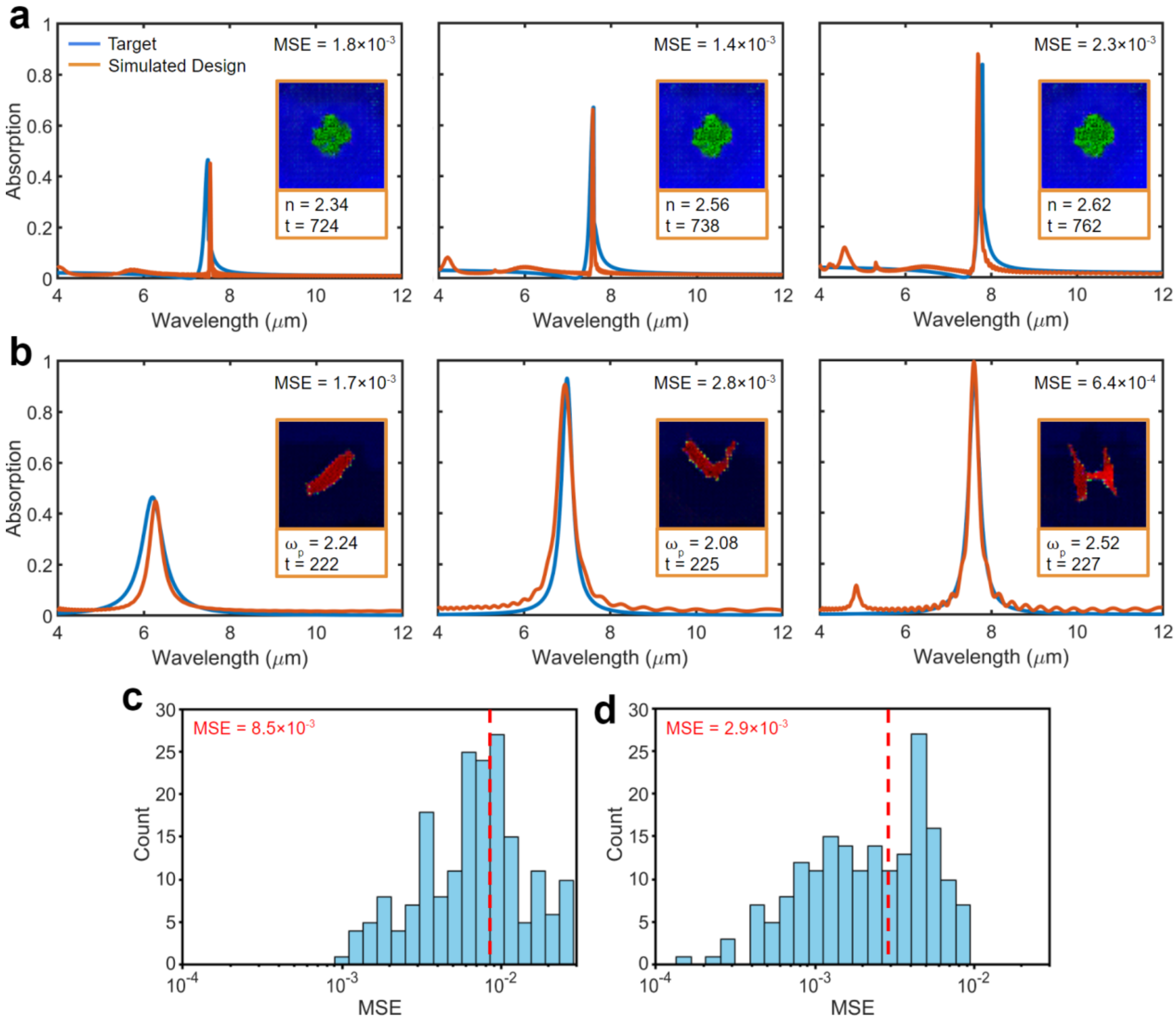


Figure 4. cDCGAN response to arbitrary 'hand drawn' targets for which there are no corresponding structures. The inset images show the synthesized images with material and structural information. (a) For Fano-shaped and (b) Lorentzian-shaped input targets, various hybrid dielectric and MIM structures with matching simulated responses are produced, respectively. Units for plasma frequency (ω_P) values are in PHz and thicknesses (t) are in nanometers. Statistical analyses across the entire test dataset (400 total spectra) for the (c) Fano-shaped and (d) Lorentzian-shaped targets.

In principle, the ‘one-to-many’ mapping capabilities of GANs allow the deep learning model to generate multiple answers to a given problem. In the context of photonics design, this ‘one-to-many’ feature could provide an assortment of design options from which the designer can select from. Accordingly, to harness the full potential of our property-embedded cDCGAN, we evaluate and report the network’s ability to generate multiple designs for a single target spectrum. To ensure consistency, this ‘diversity test’ was performed on several target spectra. As seen in Figure 5a and Figure 5b, we queried the cDCGAN with Fano-shaped and Lorentzian-shaped spectra, respectively. For each spectrum (shown in their individual plots), a second query was performed after resampling the latent vector and slightly perturbing the starting spectrum. While not perturbing the spectrum still produced unique results on the second run (as shown in Figure S7), adding small perturbations (less than 0.01 shifts in amplitude at various wavelengths) increased the overall uniqueness of the new designs. It can be observed that for each of the Fano-shaped and Lorentzian-shaped inputs, the network is able to generate two designs with distinct resonator geometry, material properties, and/or dielectric thicknesses. Importantly, though the designs have varying levels of differences, their absorption spectra remain approximately the same. The diversity of ‘one-to-many’ structures for a target spectrum is tied to the available shapes and materials that the network was able to learn from, and allows us to make use of the non-uniqueness problem that is traditionally a limiting factor in inverse design approaches in photonics. A training dataset with a larger variety of materials and geometries could certainly yield a wider panel of designs for a given target, thereby providing end-users a range of materials and geometric designs that can deliver the same spectral response.

While the presented inverse design framework was intended to generate arbitrary material predictions as a means to enable additional degrees of freedom for geometry and materials optimization, a key limitation of the presented approach thus far is that constituent materials with arbitrarily-defined properties are generally more difficult to fabricate or synthesize than conventional materials. Accordingly, to enhance the capabilities of the proposed framework in terms of their fabricability and accessibility, we demonstrate that the GAN can be used with a look-up table to substitute the predicted material properties with the closest properties derived from standard materials (shown in Figure 6). In particular, Figure 6a shows a series of tests where the input targets are Fano-shaped spectra. Here, the GAN predicted arbitrary geometries, thicknesses, and refractive index values of 2.48, 2.32, and 2.58 (from left to right). We observe that the simulated structures match well with the target responses (as previously demonstrated).

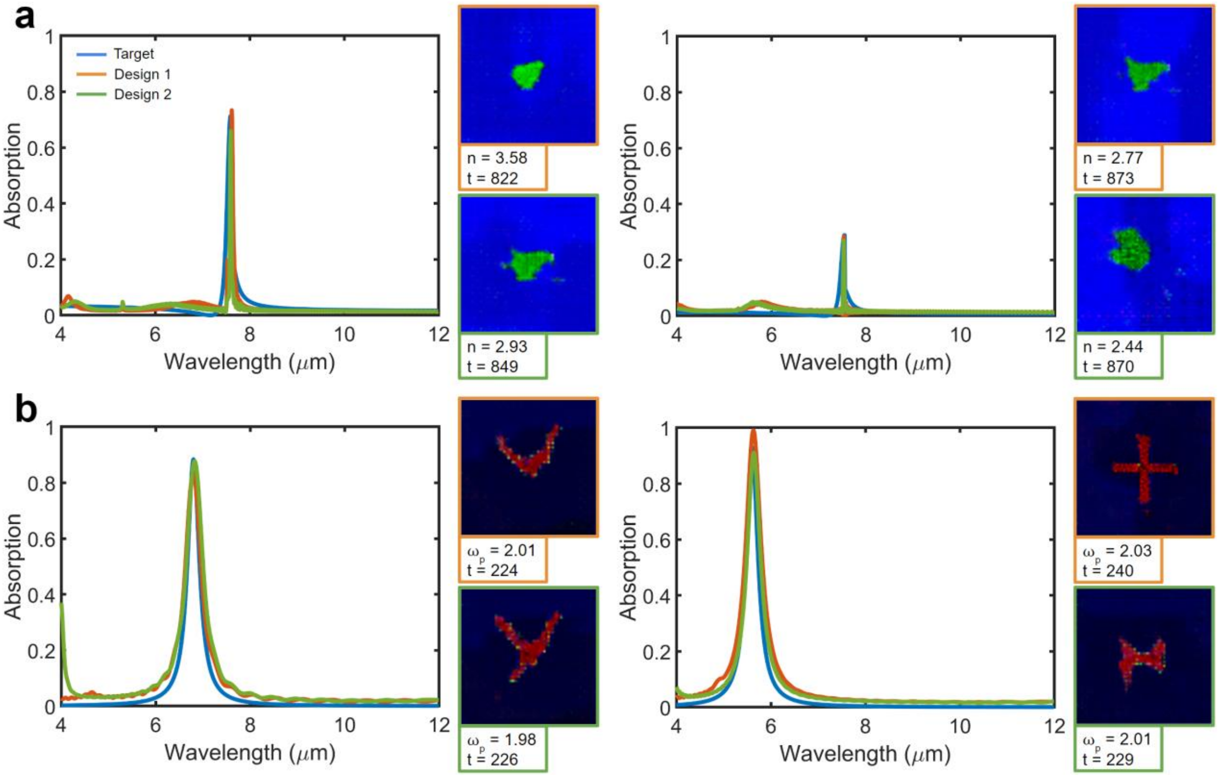


Figure 5. Demonstration of the 'one-to-many' mapping capabilities of the cDCGAN. Multiple structures with different materials and designs can be generated for a given (a) Fano-shaped or (b) Lorentzian-shaped target spectrum. Units for plasma frequency (ω_P) values are in PHz and thicknesses (t) are in nanometers.

Next, to implement the look-up table, we substitute the GAN-generated values of *n* with those of the closest materials found in a publicly-available database [61], including: CdSe (*n*=2.44), GaSe (*n*=2.38), and CdTe (*n*=2.68) [58-60]. In Figure 6b, we perform a similar set of tests with Lorentzian-shaped spectra, where the predicted materials are substituted with Au and Ag [31,32]. In both cases, after repeating the simulations, we observe that the material approximations maintain ~90% accuracy in comparison to the GAN's true predictions. Thus, we demonstrate an alternative approach at using our inverse design framework to achieve designs with greater accessibility (while maintaining reasonable accuracy). We also note that some materials identified through this approach are unique and do not exist in the training dataset (CdSe, GaSe, and CdTe). However, by virtue of the GAN-based approach outputting a new material parameter (refractive index or plasma frequency) as its prediction, we are able to identify other materials (beyond the training data) that can meet the requirements of a newly sought target. We believe this highlights a notable strength of our approach, because class-based machine learning-based methods are restricted to predicting material categories that are only available in the training dataset. As we demonstrate here, our approach enables a new degree of generalization and design flexibility by allowing practitioners to access more materials than those represented by the training data. While the particular examples we presented show that the GAN predicts values which fall within the range of real materials, we acknowledge that the GAN may also predict properties beyond the current scope of conventional materials. However, we expect the accuracy of such material approximations to improve as material libraries, and material accessibility in general, continue to develop and grow.

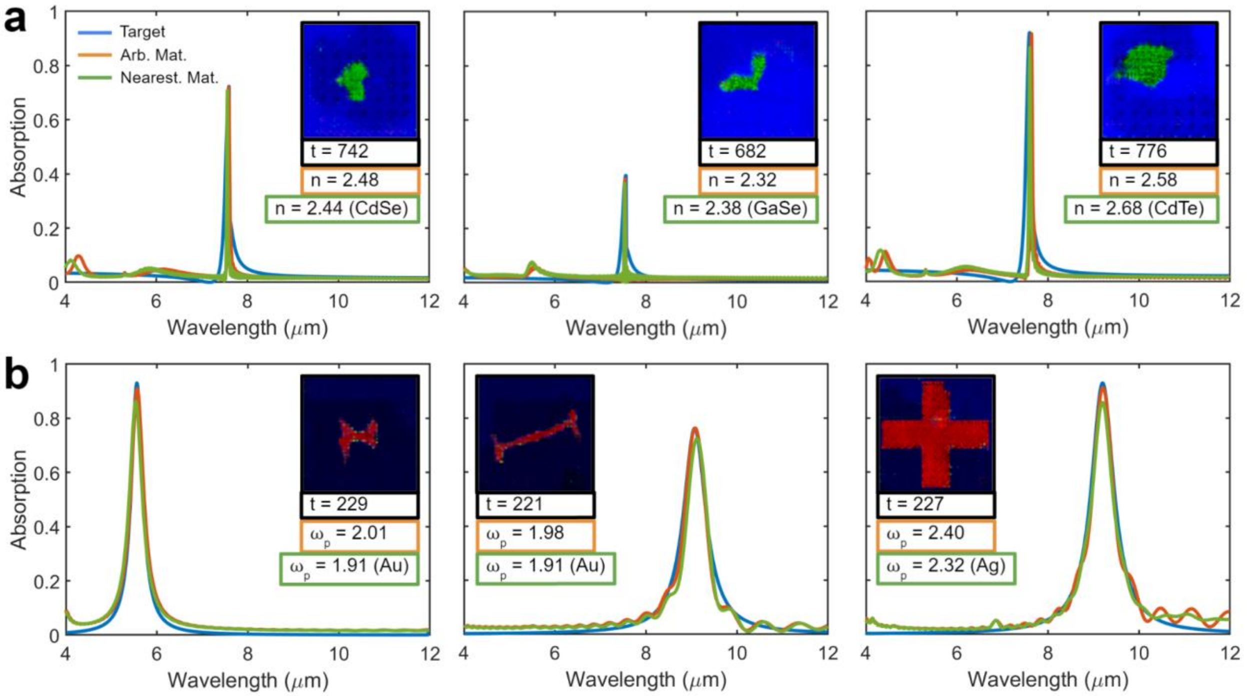


Figure 6. Applying similar materials to the cDCGAN predictions to increase fabricability. Comparison between (a) Fano-shaped and (b) Lorentzian-shaped input targets (blue). Units for plasma frequency (ω_P) values are in PHz and thicknesses (t) are in nanometers. Simulated results reveal that material approximations (green) maintain ~90% accuracy in comparison to the GAN-predicted materials (orange).

2.3 Conclusions

In summary, we present a deep learning-based photonics design framework that enables the simultaneous prediction of metasurface topology, material properties, and out-of-plane structural parameters across multiple classes of metasurfaces. Our framework is centered on a conditional deep convolutional generative adversarial network (cDCGAN) and a multiparametric-encoding strategy in which the colors of an image are encoded with various material and structural properties. By accounting for the global parameters which govern the optical behavior of metasurfaces (material, structure, and device class or fabrication process), our approach overcomes the key limitations of previously-demonstrated generative models, where only a few of the aforementioned design criteria were considered. Evaluation of our model's performance reveals that it is capable of generating not only accurate and distinct solutions from the training and validation datasets, but also multiple design alternatives and material recommendations for a single target by taking advantage of the ‘one-to-many’ mapping capabilities of GANs. To account for potential fabrication or material constraints, a property-based look-up mechanism can be paired with the model’s predictions to identify readily-available materials that serve as reasonably-accurate substitutes. The presented encoding scheme is easily adaptable to existing generative models that are integrated with optimization algorithms.

Though only two classes of metasurfaces were explored in this study (metal-insulator-metal and hybrid dielectric resonators), we believe that the results here validate the feasibility of a deep learning-based global photonics design solution aimed at describing all physical aspects of a structure. Alternative encoding schemes with greater complexity, such as higher-dimensional tensors, may therefore be employed to capture more categories of photonic designs as well as more information regarding a structure’s physical properties. To achieve a more generalized inverse design framework, future studies may directly incorporate other fundamental optical properties of materials (e.g., real and imaginary refractive indices, magnetic permeability, etc.) into the model. In this regard, a multi-pole Lorentz-Drude oscillator model with multiple parameters can also provide higher-accuracy fits over alternative wavelength ranges. More broadly, the presented methodology can be adapted to a wide range of materials design problems, including mechanical metamaterials and other synthesis-driven design challenges. Thus, our proposed framework offers a path towards a global machine learning platform that can allow practitioners and researchers to identify optimal combinations of materials, geometric parameters as well as device categories to meet complex and demanding performance goals in a range of physical systems.

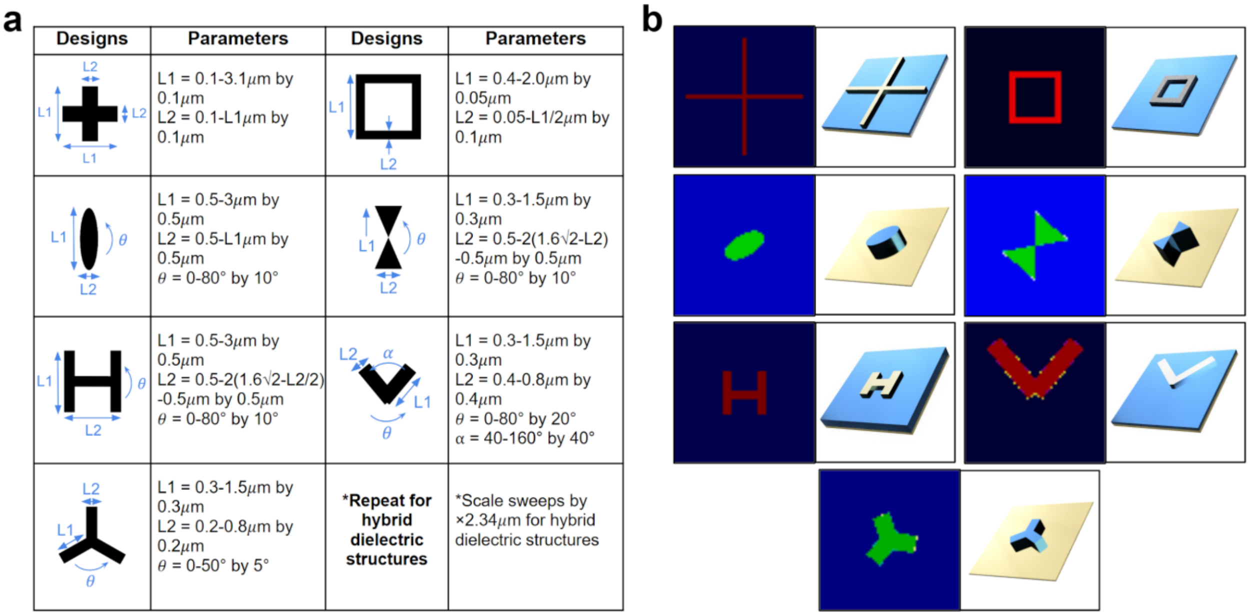
2.4 Supporting Information

Training Dataset

The training dataset for deep learning consists of precisely 18,770 metasurface unit cell designs (12,632 MIM and 6,138 hybrid dielectric structures). These designs were derived from seven starting shape templates: cross, square, ellipse, bow-tie, H, V, and tripole-shaped. As shown in Figure S1a, parameter sweeps were performed on each shape (for the MIM structures) to produce geometric variations. The tabulated parameter sweeps are captured within 3.2×3.2 μm^2 unit cells. For the hybrid dielectric structures, the same parameter sweeps were scaled by ×2.34 μm, and the unit cell dimensions for this group of structures were 7.5×7.5 μm^2. Both sets of structures are represented as 64×64×3 pixel images. Figure S1b shows several image pairs, which illustrate examples of finalized metasurface designs from each shape template. Images on the left are the color-encoded images used for deep learning, and images on the right represent the corresponding 3D physical models. As described in the main text, the colors on the image are used to indicate the metasurface class, resonator geometry, material choice, and dielectric thickness. Specifically, each structure possesses a 100 nm thickness gold substrate. For each MIM structure, the metal resonator is a 100 nm layer of gold, silver, or aluminum, while the dielectric material is Al_2O_3 with a thickness of 100 nm, 200 nm, or 300 nm. For each hybrid dielectric structure, the dielectric resonator is zinc selenide, silicon, or germanium, and its thickness is 500 nm, 750 nm, or 950 nm.

Full-wave simulations were performed on each structure (under p-polarization at normal incidence) to produce a corresponding 800-point absorption spectrum across the mid-infrared wavelength range. Low quality designs which exhibited ‘flat’ (maximum absorption is less than 0.2) or ‘noisy’ (mean-squared error, or MSE, between the spectra and its average is greater than

0.05) spectral responses were removed from the training set to maximize the model's utility and performance. Figure S2 illustrates the peak absorptions and resonance wavelengths of all the spectra training dataset, organized by full width at half maximum (FWHM). The distribution of absorption spectra reveals that the MIM structures (FWHM >= 0.4 μm) cover a wide range of peak amplitudes and resonance wavelengths from 4-10 μm, while a majority of the hybrid dielectric structures (FWHM <= 0.2 μm) exhibit responses from 7.5-8.5 μm. The range of responses in the training dataset may be extended in future studies to enhance the network's predictive capabilities. On a distributed high-performance computing cluster with four dedicated compute nodes per simulation, where a node has a minimum of four 64-bit Intel Xeon or AMD Opteron CPU cores and 8 GB memory, each FDTD simulation took approximately 5 minutes to complete. Therefore, our training dataset equates to approximately 65 days of simulation time.



Designs	Parameters	Designs	Parameters

	L1 = 0.1-3.1μm by 0.1μm
	L2 = 0.1-L1μm by 0.1μm
			L1 = 0.4-2.0μm by 0.05μm
			L2 = 0.05-L1/2μm by 0.1μm

	L1 = 0.5-3μm by 0.5μm
	L2 = 0.5-L1μm by 0.5μm
	θ = 0-80° by 10°
			L1 = 0.3-1.5μm by 0.3μm
			L2 = 0.5-2(1.6√2-L2) -0.5μm by 0.5μm
			θ = 0-80° by 10°

	L1 = 0.5-3μm by 0.5μm
	L2 = 0.5-2(1.6√2-L2/2) -0.5μm by 0.5μm
	θ = 0-80° by 10°
			L1 = 0.3-1.5μm by 0.3μm
			L2 = 0.4-0.8μm by 0.4μm
			θ = 0-80° by 20°
			α = 40-160° by 40°

	L1 = 0.3-1.5μm by 0.3μm
	L2 = 0.2-0.8μm by 0.2μm
	θ = 0-50° by 5°
		***Repeat for hybrid dielectric structures**	*Scale sweeps by ×2.34μm for hybrid dielectric structures

Figure S1. (a) 2D images of template shapes used to derive unit cell designs for both classes of metasurfaces. The range of variation allowed for each parameter is listed to the right of the associated shape. (b) Color-encoded 2D images (left) representing metasurface resonators. Images with red colors represent MIM structures while images with green colors represent hybrid dielectric structures. Shades of blue represent the dielectric thickness for both classes of structures. 3D models of each resonator are shown to the right of their 2D representation.

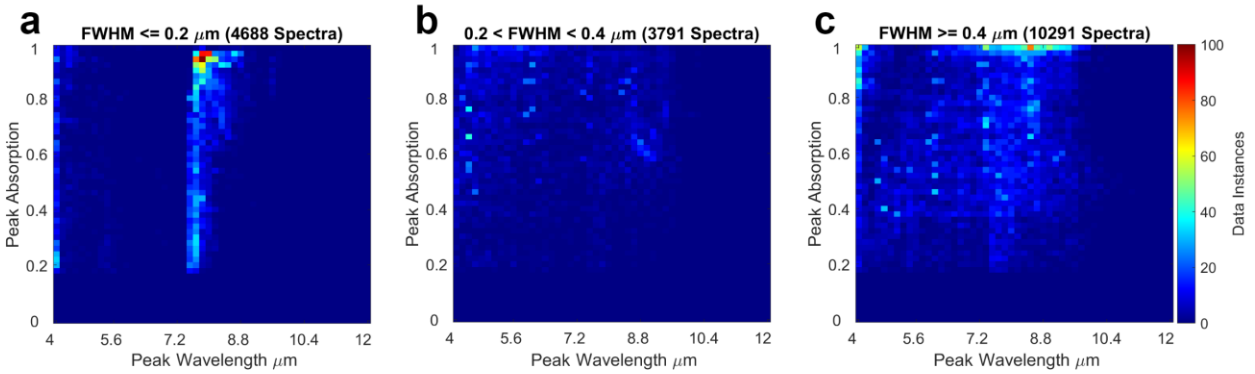


Figure S2. Visualization of peak absorption and wavelength values in the training set as distribution heatmaps. Regions of higher red-intensity indicate increased data instances within this range. The individual spectra within the training set are sorted into three subsets based on the FWHM of their absorption peak: (a) FWHM less than or equal to 0.2 µm (b) between 0.2 µm and 0.4 µm, and (c) greater than 0.4 µm.

Network Architecture Design and Optimization

Implemented in the PyTorch framework, the cDCGAN consists of two networks: a generator and a discriminator (illustrated in Figure S3). The optimized generator contains five transposed convolutional layers (with 1200, 1024, 512, 256, and 128 input channels or feature maps), while the discriminator has five convolutional layers (with 6, 64, 128, 256, and 512 input channels or feature maps). Each transposed convolutional layer in the generator is followed by a batch normalization and ReLU (rectified linear unit) activation layer, instead of the final layer, where a Tanh (hyperbolic tangent) activation is used. Similarly, in the discriminator, each convolutional layer is followed by a batch normalization and Leaky ReLU layer, and the final layer possesses a Sigmoid activation. At the generator input layer, the 800-point absorption spectra are concatenated with 400-point latent vectors to yield 1200-point input vectors. For the discriminator input, the absorption spectra are passed through a fully-connected layer and reshaped into a 64×64×3 matrix. These matrices were then concatenated with the real and generated images to form 64×64×6 inputs for the discriminator. Model training was performed on an NVIDIA Titan RTX GPU and took approximately 30 minutes to complete.

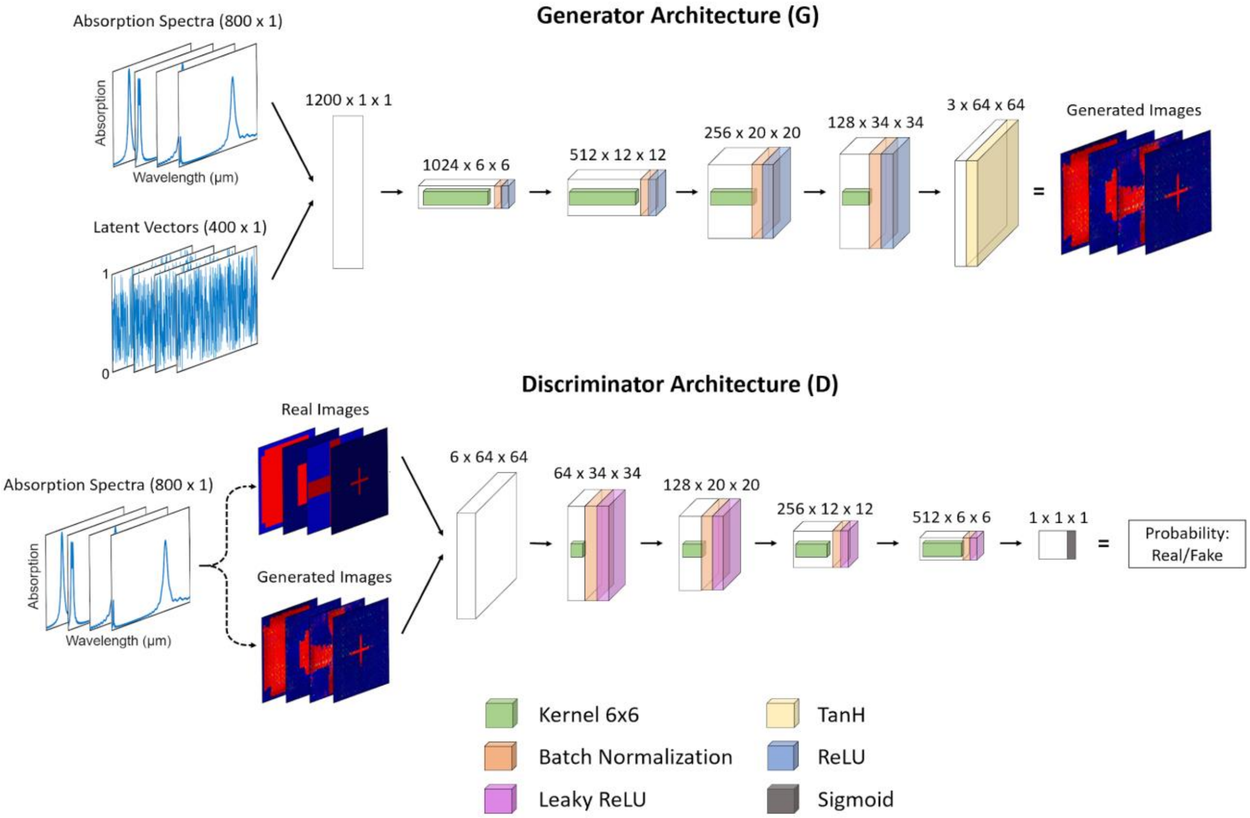


Figure S3. Schematic of the generator and discriminator architectures implemented in our cDCGAN model. Input and output types are shown for each layer along with the layer types and dimensions.

Evaluation of Multiparametric Encoding Methods

We tested the efficacy and performance of three different material and structural parameter encoding methods. Figure S4 shows the training progression of each encoding method at various epochs. As seen in Figure S4a, the first encoding method uses several neurons to represent the parameters by embedding them into a single row and column of pixels within the topological image (for each parameter). Specifically, in this encoding scheme, an initial 64×64 pixel image is converted to a 66×66 image, where the new rows and columns represent the material property and dielectric thickness of the metasurface design. The second method (Figure S4b), similar to the first, encodes the material and structure parameters as two rows and columns per parameter (yielding a 68×68 image). The third and final method that was investigated (Figure S4b) involves encoding the parameters into discrete 'RGB' channels of colored images, as described in the main text.

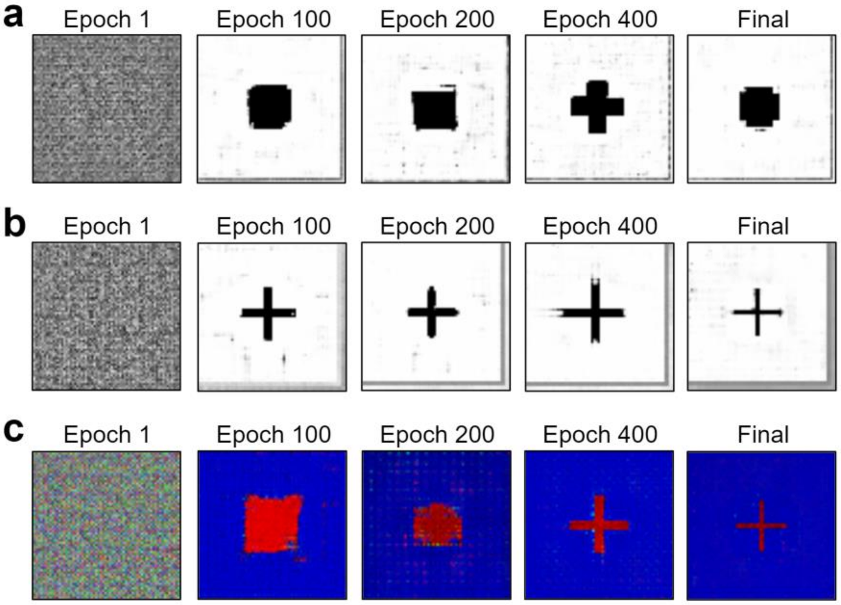


Figure S4. Examples of the generative model's training progress for three implementations of multiparametric encoding over hundreds of epochs. The tested implementations embedded parameter information as (a) a single-row and column vector concatenated with the 2D image, (b) double-row and column vectors concatenated with the image, and (c) normalized values within the 'RGB' channels of colored images.

Each encoding method was applied to the training dataset, and the corresponding datasets were separately used to train the cDCGAN. Hyperparameter tuning was performed via grid search, and the results for each dataset are presented in Table S1. Here, various feature maps, kernels, batch sizes, epochs, and miscellaneous pre- and post-processing steps were tested. Listed feature map sizes represent the lowest denomination of maps used in the intermediate layers (between the first and last layers) of the generator and discriminator. Reported losses are derived from the validation dataset.

Table S1. Hyperparameter optimization for various multiparametric encoding methods. Highlighted cells indicate the lowest validation loss for the corresponding method.

Model	**Feature Maps (G/D)**	**Kernel Size**	**Batch Size**	**Misc.**	**Epochs**	**Validation Loss (MSE)**
Single Row/Column Encoding (MIM Only)						
1	32/16	4	128		750	0.0264
2	64/32	4	128		750	0.0215
3	64/32	4	128		1000	0.025
4	66/66	4	16	Gaussian Filter (σ=0.75)	500	0.0269
5	66/66	4	16	Gaussian Filter (σ=0.75)	750	0.0332
6	66/66	4	128	Gaussian Filter (σ=0.75)	750	0.0127
7	66/66	4	128		750	0.0135
8	66/66	6	16	Gaussian Filter (σ=0.75)	500	0.0351
9	66/66	6	16	Gaussian Filter (σ=0.75)	750	0.0405
Double Row/Column Encoding (MIM Only)						
10	68/68	4	128		750	0.0413
11	68/68	5	16		250	0.0379
12	68/68	5	16		500	0.0292
13	68/68	5	32		500	0.0291
14	68/68	5	64		500	0.0157
15	68/68	6	16		500	0.0128
16	68/68	6	16	Gaussian Filter (σ=1)	500	0.0168
17	68/68	6	16	Gaussian Filter (σ=0.75)	500	0.0112
18	68/68	6	32		500	0.0177
19	68/68	6	32	Gaussian Filter (1)	500	0.0183
20	68/68	6	32	Normal Distribution (z)	500	0.0281
21	68/68	6	32	Noise 350 pts (z)	500	0.0328
22	68/68	6	32	Noise 450 pts (z)	500	0.0294
23	68/68	6	68		500	0.0247
24	68/68	6	128		500	0.0211
Color-Encoding (MIM Only)						
25	64/32	6	16	Boundary Thresh. (0.2) + GF	750	0.0044
26	64/32	6	16	Boundary Thresh. (0.1) + GF	750	0.0153
27	64/32	6	32	Boundary Thresh. (0.2) + GF	500	0.013
28	64/32	6	32	Boundary Thresh. (0.1) + GF	500	0.0142
29	64/32	6	32	Boundary Thresh. (0.2) + GF	750	0.0111
30	64/32	6	32	Boundary Thresh. (0.1) + GF	750	0.0116
31	64/32	6	64	Boundary Thresh. (0.2) + GF	500	0.0179
32	64/32	6	64	Boundary Thresh. (0.2) + GF	750	0.0285

We note that the highest-performing hyperparameters for a specific encoding method was not typically the optimal model for other encoding methods. As a result, each encoding method was optimized independently of prior models. To expedite our training efforts, only the MIM structures were used for the first round of optimization. Across all the explored encoding methods, we observe that the color-encoding approach yielded the lowest validation loss (0.0044) and highest performance. In addition to hyperparameter tuning, a Gaussian filter (GF) with binary thresholding offered substantial performance gains (where σ is the standard deviation of the Gaussian kernel) and significant reduction in noise-related artifacts, while modifying the latent vector (*z*) size and distribution (uniform to normal) resulted in no noticeable improvements.

When training the cDCGAN with the color-encoded images, the discriminator frequently overpowered the generator and resulted in mode collapse. Thus, we reduced the size of the discriminator's layers in comparison to the generator to balance the two networks. Furthermore, we developed an image processing workflow to convert the generated images into full 3D metasurface designs. Here, each generated image is decoded into three components: a resonator-only image, a material property value, and a dielectric thickness value. The resonator image specifies the existence (black pixels) or absence (white pixels) of planar features. These pixels are obtained by determining the boundaries between major color gradients on the GAN-generated color images (*e.g.*, red-blue or green-blue transition points), thus a boundary conversion threshold was applied in order to find the exact transition points. Here, we determined that the optimum threshold was a fifth of the maximum resonator color intensity (shown as 0.2 in Table S1 and S2) in the red or green color channel. If a pixel position possessed a red or green pixel value that exceeded the threshold, then the existence of a physical structure was indicated here. Prior to determining these feature boundaries, a binary classification is performed by calculating the dominant class-specific color, which is used to classify the structure (if red pixels are greater than green, then the structure is MIM, and vice versa for hybrid dielectric). The purpose of this procedure is to filter any stray red pixels that may be intermingled with green and vice versa, and to assign the appropriate boundary conditions and unit cell dimensions to the FDTD model. As described in the main text, material property and thickness values are then calculated by taking the average pixel-values in their respective channels (based on structure classification), then reversing the normalization performed in the encoding step.

Table S2. Final hyperparameter optimization with the entire training dataset. The highlighted cell indicates the model with the lowest validation loss.

Model	**Feature Maps (G/D)**	**Kernel Size**	**Batch Size**	**Misc.**	**Epochs**	**Validation Loss (MSE)**
Color-Encoding (MIM + DM Only)						
33	64/32	6	16	Boundary Thresh. (0.2) + GF	750	0.0128
34	64/32	6	16	Boundary Thresh. (0.2) + GF	1000	0.0094
35	64/32	6	32	Boundary Thresh. (0.2) + GF	500	0.0136
36	64/32	6	32	Boundary Thresh. (0.2) + GF	750	0.0086
37	64/32	6	32	Boundary Thresh. (0.2) + GF	1000	0.0135
38	64/32	6	32	Boundary Thresh. (0.2) + GF	700	0.0115
39	64/32	6	32	Boundary Thresh. (0.2) + GF	800	0.0125
40	128/64	6	16	Boundary Thresh. (0.2) + GF	500	0.0076
41	128/64	6	16	Boundary Thresh. (0.2) + GF	750	0.0106
42	128/64	6	16	Boundary Thresh. (0.2) + GF	1000	0.0125
43	128/64	6	32	Boundary Thresh. (0.2) + GF	1000	0.0144

After identifying that the color-encoding strategy resulted in the highest design accuracy (or lowest validation loss), a second round of optimization was conducted on the entire training dataset. As shown in Table S2, with the optimized post-processing procedure determined in the previous section, the highest-performance model was trained with a kernel size of 6, batch size of 16, 128 base generator feature maps, 64 base discriminator feature maps, and for 500 epochs.

Batch Testing

We evaluated the cDCGAN's performance across a wide range of new inputs by creating 200 Fano-shaped and Lorentzian-shaped spectra with amplitudes ranging from 0.5-0.9, and resonance wavelengths ranging from 5-9 μm, for 400 total test spectra. Figure S5 illustrates a comparison of 50 individual responses within this 'batch' test. Each tiled plot is presented with 4-12 μm wavelength and 0-1 absorption axes limits. Here, we observe that the Fano-shaped responses (Figure S5a) are most accurate between resonance wavelengths of 7.5-8.5 μm, while the Lorentzian-shaped responses (Figure S5b) maintain strong matches across the entire test data range. Regions of low accuracy correspond to the areas that are not well-represented by the training dataset (shown in Figure S2). Therefore, the performance of the cDCGAN may be improved by expanding the training data and design space.

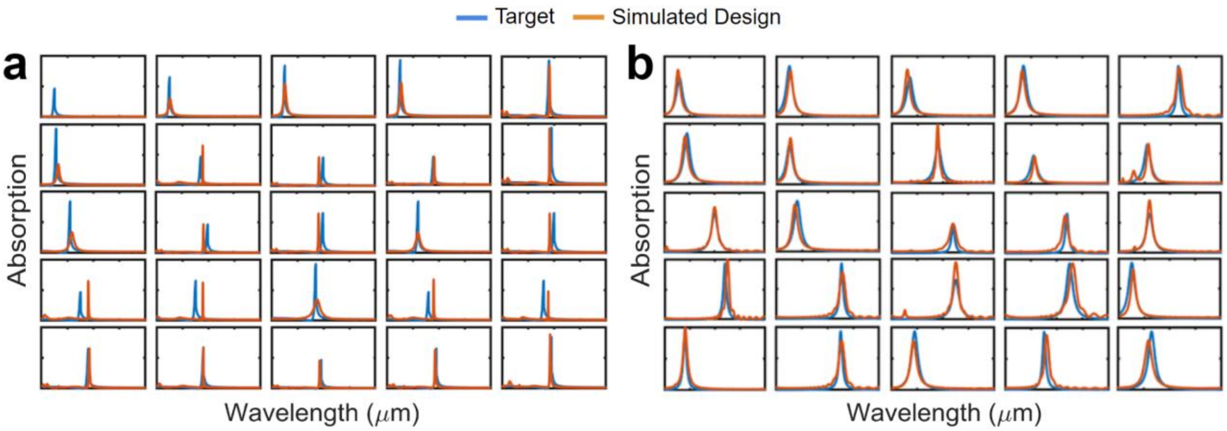


Figure S5. Test results of the cDCGAN using a diverse range of newly-constructed (a) Fano-shaped and (b) Lorentzian-shaped inputs. Blue lines represent input targets while the orange lines represent simulated designs produced by the cDCGAN.

Designing Complex Alloyed Structures

Though we limited this study to the application of uniform materials, we note that the devised color-encoding strategy is capable of representing spatially-varying material properties along the entire physical structure. This in turn sets the stage for future studies with much greater design complexity, such as 3D or complex metal alloy-based structures, with potentially greater control over the electromagnetic spectrum. A demonstration of this capability is shown in Figure S6, where the different shades of color on the cDCGAN-generated MIM structures are converted into different metals (shown in the inset images) based on their individual plasma frequencies, rather than the average over the channel. Notably, the simulated alloyed structures yield similar responses to the uniform material structures, beyond which there are no distinguishable advantages in the particular design space that was explored. Therefore, future studies utilizing a wider range of dissimilar materials (as well as the application of fabrication constraints tailored towards alloy-based design) may produce device properties that extend beyond the uniform material domain to gradient-index materials.

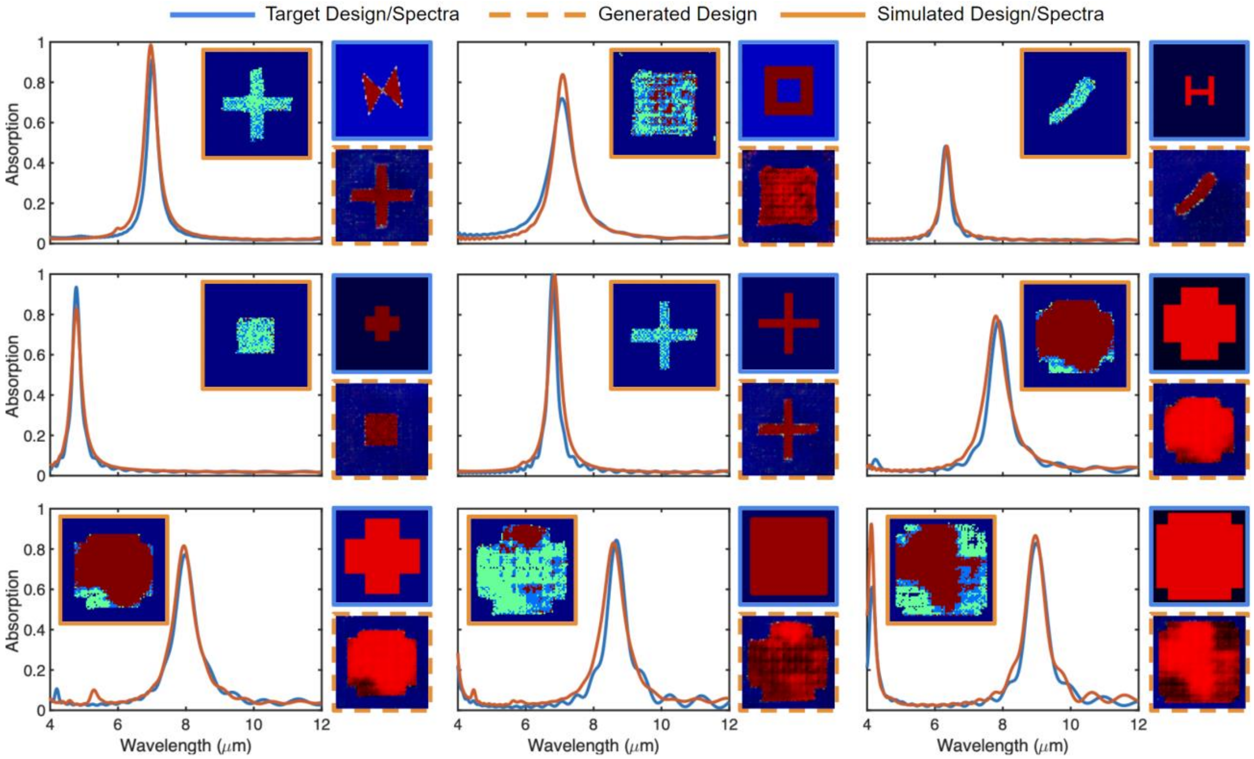


Figure S6. Demonstration of metal alloy-based structure design using the color-encoded cDCGAN. Blue lines represent the input spectra and reference designs. Dashed-orange lines represent the cDCGAN output, and solid-orange lines are alloyed structures created using the color gradients. We note that the simulated results match well with the input targets. The fabricability of these structures could potentially be improved with the addition of fabrication constraints such as minimum feature size.

Latent Vector Sampling and Model Validation

Since the GAN may produce a distribution of designs with potentially varying degrees of accuracy, ten different latent vectors were generated for each target spectrum, which were then used as inputs to the network. Each design is verified using numerical simulation, then the design (and corresponding latent vector) with the lowest mean-squared error to the target is used as the final design. Figure S7 shows the distribution of designs across different latent vectors for several input targets (a Lorentzian function centered at 7.2 μm and at 7.8 μm), where we observe that all the generated designs have over 90% accuracy in comparison to the input target. Following this procedure, Figure S8 presents a series of tests performed with inputs that originate from the training dataset. An equivalent analysis for the validation dataset can be found in the main text.

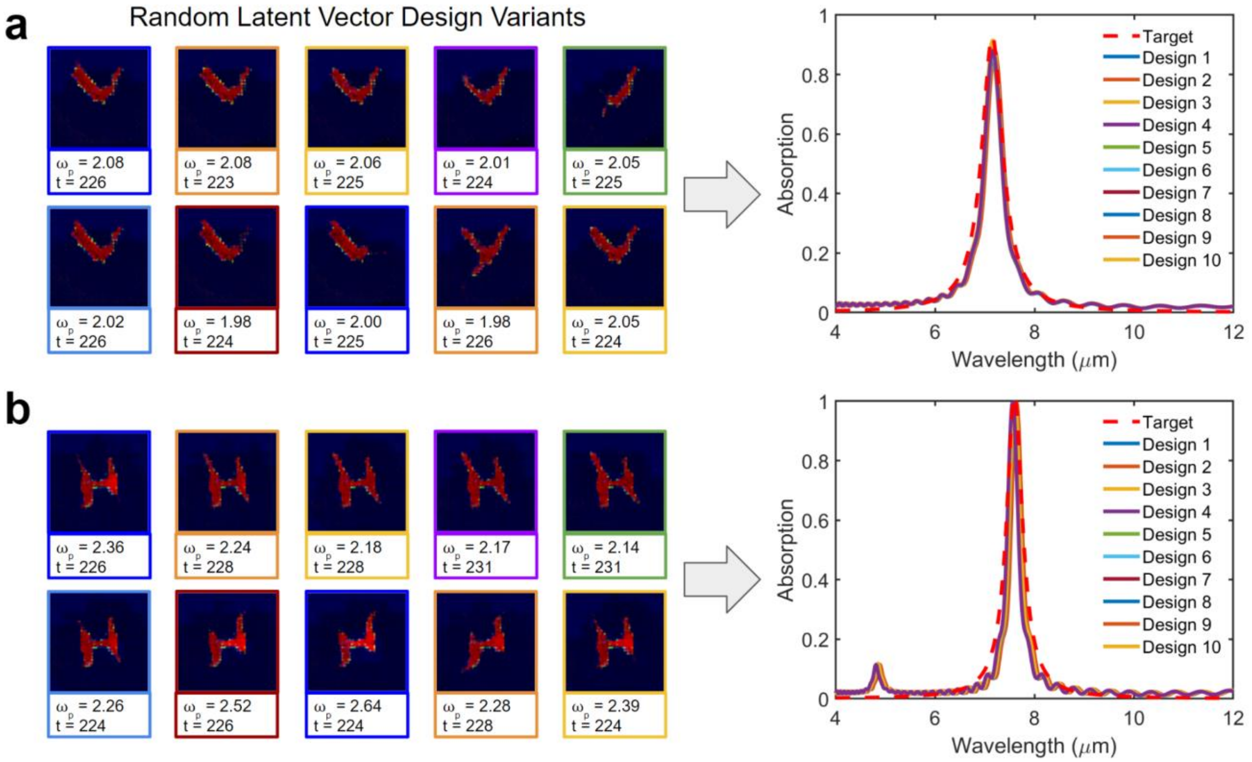


Figure S7. GAN-produced design variants achieved by pairing the target spectrum with 10 randomly-generated latent vectors. Input targets for Lorentzian functions centered at (a) 7.2 μm and (b) 7.8 μm are indicated by the dashed red lines. For each target, the design with the lowest mean-squared error is used as the final design.

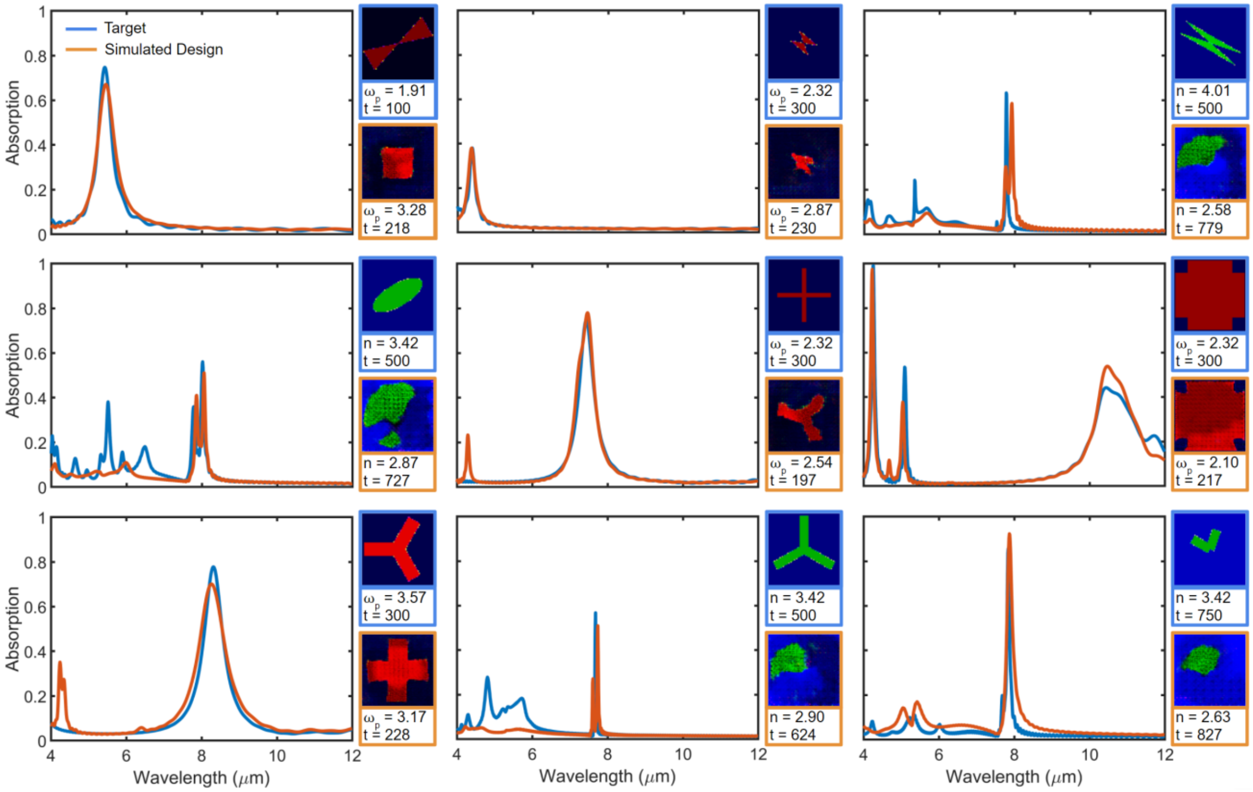


Figure S8. Randomly selected absorption spectra from the training dataset (in blue) which were designated as input targets for the cDCGAN. The simulated spectra of the cDCGAN-synthesized designs (in orange) are plotted alongside the targets for comparison. Images representing the respective structures are shown to the right of each plot, with material and thickness information below each image. Units for plasma frequency (ω_P) values are in PHz and thicknesses (t) are in nanometers. The results here reveal that the network is not copying the training dataset, but to a degree, it is identifying the underlying relationships between structure, material, metasurface class, and optical response to provide new yet accurate solutions that extend beyond the training dataset.

2.5 References

[1] D. Neshev, I. Aharonovich, *Light Sci. Appl.* **2018,** *7*, 58.

[2] M. Khorasaninejad, W.T. Chen, R.C. Devlin, J. Oh, A.Y. Zhu, F. Capasso, *Sci*. **2016**, *352*, 1190.

[3] D. Lin, P. Fan, E. Hasman, M. L. Brongersma, *Sci.* **2014**, *345*, 298.

[4] D. Neshev, I. Aharonovich, *Nano Lett.* **2017**, *14,* 2634.

[5] K. Park, Z. Deutsch, J.J. Li, D. Oron, S. Weiss, *ACS Nano.* **2012**, *6*, 10013.

[6] W. Wan, W. Qiao, D. Pu, R. Li, C. Wang, Y. Hu, H. Duan, L.J. Guo, L. Chen, *iSci.* **2020**, *23*, 100773.

[7] S. Molesky, Z. Lin, A.Y. Piggott, W. Jin, J, Vuckovic, A.W. Rodriguez, *Nature Photon.* **2018**, *12*, 659.

[8] J. Jiang, J.A. Fan, *Nano Lett.* **2019**, *19*, 5366.

[9] K.H. Matlack, M. Serra-Garcia, A. Palermo, S.D. Huber, C. Daraio, *Nat. Mater*. **2018**, *17*, 323.

[10] C. Yeung, J. Tsai, B. King, Y. Kawagoe, D. Ho, M.W. Knight, A.P. Raman, *ACS Photon*. **2020**, *7*, 23.

[11] S. So, T. Badloe, J. Noh, J. Bravo-Abad, J. Rho, *Nanophoton*. **2020**, *9*, 1041.

[12] B. Sanchez-Lengeling, A. Apuru-Guzik, *Sci.* **2018**, *361*, 360.

[13] K. Tu, H. Huang, S. Lee, W. Lee, Z. Sun, A. Alexander-Katz, C.A. Ross, *Adv. Mat.* **2020**, *32*.

[14] B. Han, Y. Lin, Y. Yang, N. Mao, W. Li, H. Wang, K. Yasuda, X. Wang, V. Fatemi, L. Zhou, J.I. Wang, Q. Ma, Y. Cao, D. Rodan-Legrain, Y. Bie, E. Navarro-Moratalla, D. Klein, D. MacNeill, S. Wu, H. Kitadai, X. Ling, P. Jarillo-Herrero, J. Kong, J. Yin, T. Palacios, *Adv. Mat*. **2020**, 2000953.

[15] G. H. Gu, C. Choi, Y. Lee, A.B. Situmorang, J. Noh, Y. Kim, Y. Jung, *Adv. Mat*. **2020**, 1908965.

[16] S. So, J. Mun, J. Rho, *ACS Appl. Mat. & Interfaces*. **2019**, *11*, 24264.

[17] D. Liu, Y. Tan, E. Khoram, Z. Yu, *ACS Photon*. **2018**, *5*, 1365.

[18] C. Yeung, J. Tsai, B. King, D. Ho, J. Liang, M.W. Knight, A.P. Raman, *arXiv preprint*. **2020**, 2008.00587.

[19] J. Jiang, M. Chen, J.A. Fan, *Nature Rev. Mat.* **2020**, *17*, 1.

[20] D. Sell, J. Yang, S. Doshay, R. Yang, J.A. Fan, *Nano Lett*. **2017**, *17*, 3752

[21] S. So, J. Rho, *Nanophoton*. **2019**, *8*, 1255.

[22] Z. Liu, D. Zhu, S.P. Rodrigues, K. Lee, W. Cai, *Nano Lett*. **2018**, *18*, 6570.

[23] F. Wen, J. Jiang, J.A. Fan, *ACS Photon*. **2020**, *7*, 2098.

[24] J. Jiang, D. Sell, S. Hoyer, J. Hickey, J. Yang, J.A. Fan, *ACS Nano*. **2019**, *13*, 8872.

[25] Z. A. Kudyshev, A.V. Kildishev, V.M. Shalaev, A. Boltasseva, *Appl. Phys. Rev*. **2020**, *7*, 021407.

[26] W. Ma, Z. Liu, Z.A. Kudyshev, A. Boltasseva, W. Cai, Y. Liu, *Nat. Photon*. **2020**, 1.

[27] X. Liu, T. Tyler, T. Starr, A.F. Starr, N.M. Jokerst, W.J. Padilla, *Phys. Rev. Lett.* **2011**, *107*, 045901.

[28] P. Neutens, P.V. Dorne, I.D. Vlaminck, L. Lagae, G. Borghs, *Nat. Photon.* **2009**, *3*, 283.

[29] S. Chen, Z. Chen, J. Liu, J. Cheng, Y. Zhou, L. Xiao, K. Chen, *Nanomat*. **2019**, *9*, 1350.

[30] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K.A. Persson, G. Ceder, A. Jain, *Nat*. **2019**, *571*, 95.

[31] M. Kreiter, S. Mittler, W. Knoll, J. R. Sambles, *Phys. Rev. B*. **2002**, *65*, 125415.

[32] M. G. Blaber, M.D. Arnold, M.J. Ford, *The Journal of Phys. Chem*. **2009**, *113*, 3041.

[33] M. A. Ordal, R.J. Bell, R.W. Alexander, L.L. Long, M.R. Querry, *Applied Optics*. **1985**, *24*, 4493.

[34] M. R. Querry, *Optical Constants of Minerals and Other Materials from the Millimeter to the Ultraviolet*, **1987**.

[35] E. D. Palik, *Handbook of Optical Constants of Solids*. **1998,** *3*.

[36] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, *International Conference on Machine Learning*. **2019**, 7354.

[37] N. Kodali, J. Abernethy, J. Hays, Z. Kira, *arXiv preprint*. **2019**, 1705.07215.

[38] T. Karras, S. Laine, T. Aila, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. **2019**, 4401.

[39] M. Arjovsky, S. Chintala, and L. Bottou *Proceedings of Machine Learning Research*. **2017**, *70*, 214.

[40] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, S. P. Smolley, *Proceedings of the IEEE International Conference on Computer Vision*. **2017**, 2794.

[41] M. Mirza, S. Osindero, *arXiv preprint*. **2014**, 1411.1784.

[42] A. Radford, L. Metz, *arXiv preprint*. **2015**, 1511.06434.

[43] W. Tang, G. Li, X. Bao, T. Li, *arXiv preprint*. **2018**, 1810.08534.

[44] R. Müller, S. Kornblith, G. Hinton, *Adv. in Neural Information Processing Systems*. **2019**, 4694.

[45] T. Salimans, I. Goodfellows, W. Zaremba, V. Cheung, A. Radford, X. Chen, *arXiv preprint*. **2016**, 1606.03498.

[46] K. Koshelev, Y. Kivshar, *ACS Photonics*. **2020**.

[47] S. Makarov, S. Kudryashov, I. Mukhin, A. Mozharov, V. Milichko, A. Krasnok, P Belov, *Nano Lett.* **2015**, *15*, 6187.

[48] Q. Liu, S.C. Qillin, D.J. Masiello, P.A. Crozier, *Phys. Rev. B.* **2019**, *99*, 165102.

[49] Q. Wu, B. Dang, C. Lu, G. Xu, G. Yang, J. Wang, X. Chuai, N. Lu, D. Geng, H. Wang, L. Li, *Nano Lett.* **2020**, *20*, 8015.

[50] Z. Liu, D. Zhu, K. Lee, A. Kim, L. Raju, W. Cai, *Adv. Mater*. **2019**, 32, 1904790.

[51] J. Jiang, J. Fan, *Nanophoton*. **2020**, *10*, 361.

[52] D. P. Kingma, M. Welling, *arXiv preprint*. **2013**, 1312.6114.

[53] R. Unni, K. Yao, Y. Zheng, *ACS Photonics*. **2020.**

[54] N. Inkawhich, *DCGAN Tutorial - PyTorch Tutorials 1.9.0+cu102 Documentation*. **2020.**

[55] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, *NeurIPS Proceedings.* **2014.**

[56] P. Kim, C. Li, R. E. Riman, J. Watkins, *ACS Appl. Mater. Interfaces* **2018**, 10, 10, 9038–9047.

[57] Y. Hashimoto, G. Seniutinas, A. Balčytis, S. Juodkazis, Y. Nishijima, *Sci. Rep.* **2016**, 6, 25010.

[58] M. P. Lisitsa, L. F. Gudymenko, V. N. Malinko, S. F. Terekhova, *Phys. Status Solidi* **1969**, *31*, 389.

[59] K. Kato, F. Tanno, N. Umemura, *Appl. Opt.* **2013**, *52*, 2325.

[60] A. G. DeBell, E. L. Dereniak, J. Harvey, J. Nissley, J. Palmer, A. Selvarajan, W. L. Wolfe, *Appl. Opt.* **1979**, *18*, 3114.

[61] M. N. Polyanskiy, Refractive index database, https://refractiveindex.info, accessed: June **2021**.

3. Multiplexed Supercell Metasurface Design and Optimization with Tandem Residual Networks

3.1 Introduction

Nanophotonic materials, including metasurfaces and metamaterials, have greatly expanded our ability to tailor light-matter interaction and deliver new functionalities for information processing and sensing applications [1-4]. As demand for advanced capabilities and high-performance nanophotonic devices grow, multimodal implementations with interconnected ensembles of optical sub-components, including supercells, have shown great promise in delivering tailored responses with respect to many optical characteristics [5-8]. For example, complex spatial arrangements within photonic crystal circuits have yielded high-efficiency spatial mode conversion [9]. Similarly, by employing metasurfaces that contain periodic arrays of meta-atoms with different geometric parameters, a range of useful behaviors including out-of-plane beam deflection and mirroring can be demonstrated [10]. Although the incorporation of numerous distinct subunit elements within a photonic structure is desirable, it is accompanied by an exponential increase in design costs as a result of the increased dimensionality of the associated design space [11].

A particular category of periodic metasurface structures that has shown promise in supercell configurations is the metal-insulator-metal (MIM) metasurface absorber. Periodic MIM absorbers yield strong resonances that are narrowband in nature, where the wavelength of the resonance peak can be shifted by changing the shape of the resonator [34-37]. By adopting simple supercell configurations, which contain more than one resonator geometry, multi-resonant and broadband absorption behavior has previously been realized [38-40]. The design and optimization of more complex supercells with hybridized behavior, however, remains an open challenge, but holds the potential of yielding a broader range of spectral responses than previously achieved.

Conventional design processes for periodic and complex supercell metasurfaces rely on electromagnetic (EM) simulations that are iteratively optimized by tuning key design parameters until the desired optical properties are obtained. Techniques that have been employed include evolutionary algorithms [12], topology optimization [13-15], and adjoint-based methods [16,17]. In the context of supercells and complex/non-periodic arrangements, methods such as Schur complement domain decomposition and overlapping-domain approximation have yielded compelling results [18,19]. As the unit cell of a metasurface increases in size and complexity, however, computation times from iterative optimization can rapidly escalate from hours to potentially days or weeks. Additionally, optimizations must be repeated and reconfigured for every new target, thus requiring a substantial amount of computational resources and, oftentimes, prior intuition on the capability of a particular class of nanophotonic structures. These computational costs are further compounded by the fact that only the final optimized results are preserved; any prior data generated in an optimization cycle is not typically reused in the future [41]. As a result, iterative design methods also become increasingly inefficient over time [20].

In response to the need for more efficient design strategies, data-driven approaches based on machine learning, such as deep neural networks (DNNs), have found applications in nanophotonic design [21]. DNNs are now well established in many fields, including: natural language processing, drug discovery, materials design, and medical diagnosis [22,24]. In the photonics context, DNNs have shown promise in designing a diverse range of high-performance structures by directly predicting key geometric parameters (*e.g.*, resonator widths, lengths, radii, etc.). By leveraging a one-time investment of EM simulation training data, DNNs can generate designs at orders-of-magnitude faster speeds than traditional optimization algorithms. An accurate DNN can also be paired with numerical optimization methods to save simulation time, where the DNN identifies solutions near the global minimum and the optimization refines the performance further [30]. With training datasets ranging from several hundred [45] to several thousand [41] instances, previously explored machine learning and DNN-based photonics design include the forward and inverse modeling of multi-shell nanoparticles, multi-layer thin-films, and various classes of metasurfaces [25-29]. A forward-modeling DNN takes structural parameters as inputs and predicts optical properties such as the absorption spectra. In contrast, an inverse-modeling DNN accepts target optical properties as inputs and generates matching structural parameters. Further advancements in DNNs have led to the development of the tandem network, which is designed to overcome the nonuniqueness scattering problem [27,41]. However, prior tandem networks have relied on traditional fully connected or dense networks, while tandem implementations with recent architectural advancements such as residual networks or ResNets (which address the well-known vanishing gradient problem [46]) remain unexplored.

While promising, prior studies of DNNs for nanophotonic design have primarily focused on individual scatterers or periodic structures with single-unit cell elements and relatively narrowband operation [31-33]. Recent studies involving the design of structures with multiple optical elements have assumed that they are separately constructed and then assembled into a multi-element structure [47,48]. This approach makes the limiting assumption that the coupling between adjacent elements is sufficiently weak, and further does not affect their cross-section. Moreover, in studies where coupling cannot be neglected, separately-trained models were required in order to design metasurfaces with specific numbers of elements [48], which limits scalability. Other work with unit cells consisting of multiple neighboring elements do not solve the inverse problem, but instead develop a fast and accurate proxy or surrogate model for forward design [49]. Therefore, an ML-based strategy for complex supercells that: 1) directly solves the inverse design problem, 2) generates structures with a wide range of unique elements, and 3) considers strong coupling interactions or mode hybridization between individual elements is lacking today, but could allow for the demonstration of complex nanophotonic architectures with a broader range of spectral responses.

In this article, we investigate the inverse design of large multiplexed supercell metasurfaces with over 100 subunit elements that can achieve a diverse set of broadband spectral responses. Specifically, we focus on engineering arbitrary bandwidth absorbers operating in the mid- and long-wave infrared regime (4-12 µm) by designing supercell MIM metamaterial absorbers through a deep learning approach. To navigate the large design space that comes with the increased dimensionality of supercells and to address the vanishing gradient problem associated with deep network architectures, we employed a tandem residual network (shown conceptually in Figure 1a). We demonstrate that with a training dataset of several thousand simulations, in a high-dimensional design space with over three trillion possible design combinations, the network can successfully design narrowband, multi-resonance, and broadband-absorption supercell metasurfaces with high degrees of accuracy. Furthermore, we show that the network itself can be harnessed to approximate the structure-property relationships of the explored class of metasurfaces.

3.2 Results and Discussion

Nanophotonic supercell structures such as MIM metasurfaces are capable of producing unique optical responses that extend beyond the sum of their parts. Specifically, in addition to the superposition of individual responses, distinct responses may also arise from the interaction or hybridization between neighboring elements [50]. Several examples of such interactions are presented in Figure S1, where the absorption spectra and EM field profiles of various supercell designs are shown. In this figure, we show that specific arrangements of identical subunit resonators can yield absorption peaks with different amplitudes and wavelengths, or new peaks entirely, in comparison to the response of the individual elements. These examples reveal that the relative positions of individual elements are critical as the characteristics of the peaks can depend strongly on which resonators are adjacent to each other. Therefore, supercell-class metasurfaces can potentially access new domains of functionalities by leveraging multiscale optical phenomena. To this end, a comprehensive inverse design scheme for supercell structures must consider the geometries and physical arrangements of individually-integrated structures as well as their collective EM interactions.

Figure 1b presents the detailed implementation of our supercell inverse design strategy. First, we defined a supercell layout of MIM resonators (labeled “1” in Figure 1b). The layout contains an assortment of 100 nm-thick gold cross-shaped resonators with a 100 nm gold backing and 200 nm Al2O3 spacer. This class of metasurfaces was derived from existing literature on selective thermal emitters and exhibits narrowband resonances in the mid-infrared (MIR) range [42]. A single supercell design is represented by an array of cross-shaped resonator lengths (ranging from 1.4-3 µm in 0.2 µm steps), each with fixed widths (500 nm). The resonator arrays (labeled “2” in Figure 1b) embody a quadrant of the supercell and resembles a hexagonal close-packed (HCP) lattice with a twin boundary, where the individual resonators are mirrored along the diagonal plane. The quadrant is then mirrored along the x- and y-axes to create a four-fold symmetric supercell. The HCP configuration is designed to maximize the area density (and therefore the resonance efficiency) of the supercell, while the four-fold symmetry ensures the structure is s- and p-polarization independent under normal incidence. We limited our supercell size to 25 unique resonators per quadrant (12.8 × 12.8 μm2 before four-fold symmetry) to maximize the resonance modes within the 4-12 μm window while simultaneously attempting to minimize simulation time. Thus, a unique supercell design is represented by $D_A=[l_1, l_2, \ldots, l_{25}]$, with l_{25} being the length of the *25*-th resonator (where $D_A, \ldots, D_n$ are vectors with distinct *l*-values). These vectors were then used as the supercell design parameters for deep learning.

We converted the supercell design parameters into three-dimensional MIM structure models (labeled "3" in Figure 1b) and performed full-wave EM simulations (Lumerical FDTD) on these models over the spectral range of 4-12 μm at normal incidence (labeled "4" in Figure 1b), obtaining an 800-point "ground truth" absorption spectrum for each structure (labeled "5" in Figure 1b). Using this approach, we simulated the absorption spectrum (*A*) for pseudo-randomly generated design parameters (*D*) to create training data pairs (*D*, *A*) for the neural network. As discussed in the next section, the deep learning model "learns" by comparing the ground truth spectra (*A*) to the network-predicted spectra (A', labeled "6" in Figure 1b).

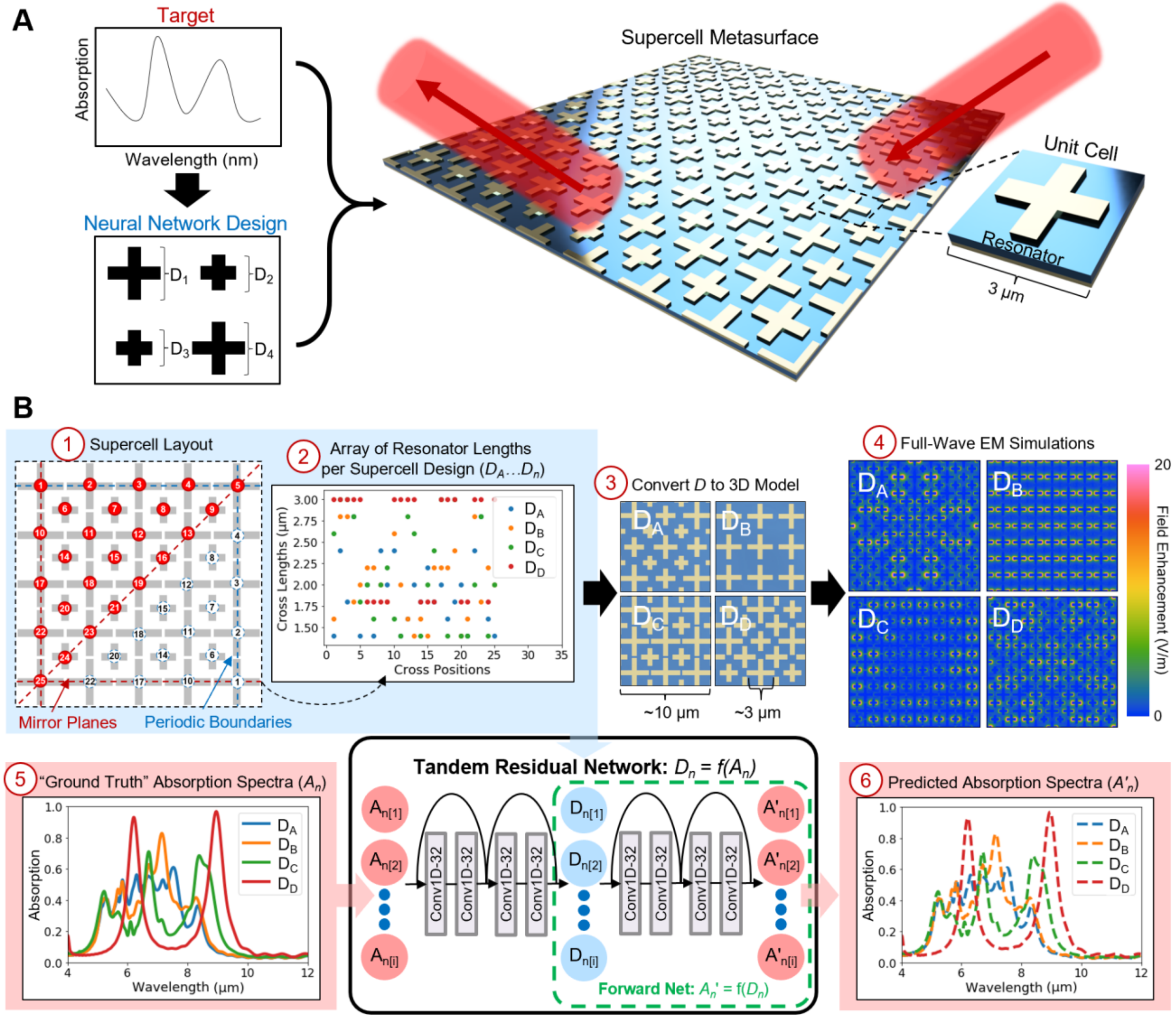


Figure 1. Inverse design of supercell metasurfaces with a range of underlying symmetries using a tandem residual network approach. (a) A target absorption spectrum is defined, and the matching design parameters for a multiplexed array of MIM resonators are generated. (b) Data preparation schematic for deep learning. Supercell design parameters (1) representing resonator lengths and positions (2) are converted into 3D models (3). Full-wave electromagnetic simulations are performed on the models (4). Design parameters along with corresponding "ground truth" (5) and predicted (6) absorption spectra are used to train the tandem residual network.

The performance of a tandem network hinges on the accuracy of the forward-modeling network as well as the breadth and size of the training dataset. Thus, we sought to optimize the architecture of the forward-modeling network and to ensure that the size of our training dataset maximizes the network's implementation efficiency and predictive capabilities. Unlike previous implementations of the tandem architecture, our approach utilizes one-dimensional convolutional neural networks (1-D CNNs) instead of dense networks. 1-D CNNs have been used in various scientific domains [51,52], with recent works showing that they are capable of outperforming dense networks in terms of regression fidelity and generalization capabilities [53]. This is enabled by the convolutional layers of the CNN, which are optimized to extract highly discriminative features using a large set of 1-D filter kernels [51]. Furthermore, our particular CNN consists of residual building blocks, which leverage identity shortcuts or skipped connections to address the vanishing gradient problem and achieve better performance than "plain" networks of the same depth [54-56]. The corresponding ResNet was trained in the forward-modeling configuration to predict an absorption spectrum (A'), given a set of design parameters (D) as inputs. We evaluated and compared the performances of the dense network, CNN, and ResNet in Figure S2 and S3, where it can be observed that the ResNet achieved the lowest validation loss out of the three model types. Our optimized forward-modeling ResNet architecture consists of a 25-neuron input layer (matching the vector size of the supercell design parameters D, with values normalized from 0 to 1), two residual blocks, followed by an 800-neuron dense layer. Each residual block contains two 1-D convolutional layers with 32 filters, kernel size of 3, and zero-padding. In addition, the Adam optimizer, batch size of 10, and ReLU activation functions yielded the lowest validation loss.

Using the same hyperparameters as the forward network, except with an inverted sequence of input and output layers (with 800 and 25 filters, respectively), we designed an inverse-modeling network for the prediction of design parameters (D') given an input A. However, plain inverse modeling networks are known to encounter the nonuniqueness problem [27,41], where the multiple mappings between an EM response and its available structural parameters may confound the network's learning process. To illustrate this problem in the context of our training data, Figure

S4 shows several examples where two substantially different supercell design layouts map to nearly-identical dual-band and triple-band responses. Due to the considerable degrees of freedom in a supercell design, the nonuniqueness problem in a supercell architecture is exacerbated relative to single-element and periodic structures, and is crucial to address. Thus, to account for this issue, we implemented the tandem architecture by coupling the inverse-modeling network with a pretrained forward-modeling network.

First, we trained a standard tandem dense network by minimizing the loss function between the input absorption spectrum (*A*) and the spectrum predicted by the forward-modeling network (A'), where A' is generated by the same D' predicted by the inverse-modeling network from above. As in Ref. [27], we define the tandem network's loss as the mean squared error between A and A': $MSE = \frac{1}{n}\Sigma(A'_i - A_i)^2$. This loss calculation is distinct from the plain inverse modeling network, which is programmed to minimize the loss between the designs from the training dataset (D) and the predicted designs (D'). Since D' may offer a completely different solution than D that correctly maps to the target response (due to the issue of nonuniqueness), a plain inverse-modeling network can struggle to minimize loss or converge, whereas in the tandem network, the loss function converges so long as the target and predicted spectra (*A* and A') are similar [27,41]. In other words, since the task at hand requires solving a one-to-many problem, the tandem network finds the optimum response for an input target rather than mixing the corresponding outputs (which can lead to a suboptimal solution).

To validate that the tandem network reaches said optimum response, in Figure S5, we show the tandem neural network architecture's ability to resolve the nonuniqueness issue by testing the accuracy of designs for which an explicit nonunique relationship exists, which were found in Figure S4. As seen in Figure S5A, dual-band and triple-band spectra were passed into the inverse modeling network, and a poor match between the input spectra and the simulated design parameters can be observed. However, when the same spectra were passed into the tandem network (Figure S5B), the accuracy between the input spectra and the simulated parameters is substantially improved. Thus, we find that the tandem dense network effectively addresses the nonuniqueness issue, and yields superior accuracy over a plain inverse modeling network for the multiplexed supercell metasurfaces evaluated here. Furthermore, in Figure S6, we verify that the quantity of our training data was capable of maximizing the network's ability to learn supercell designs.

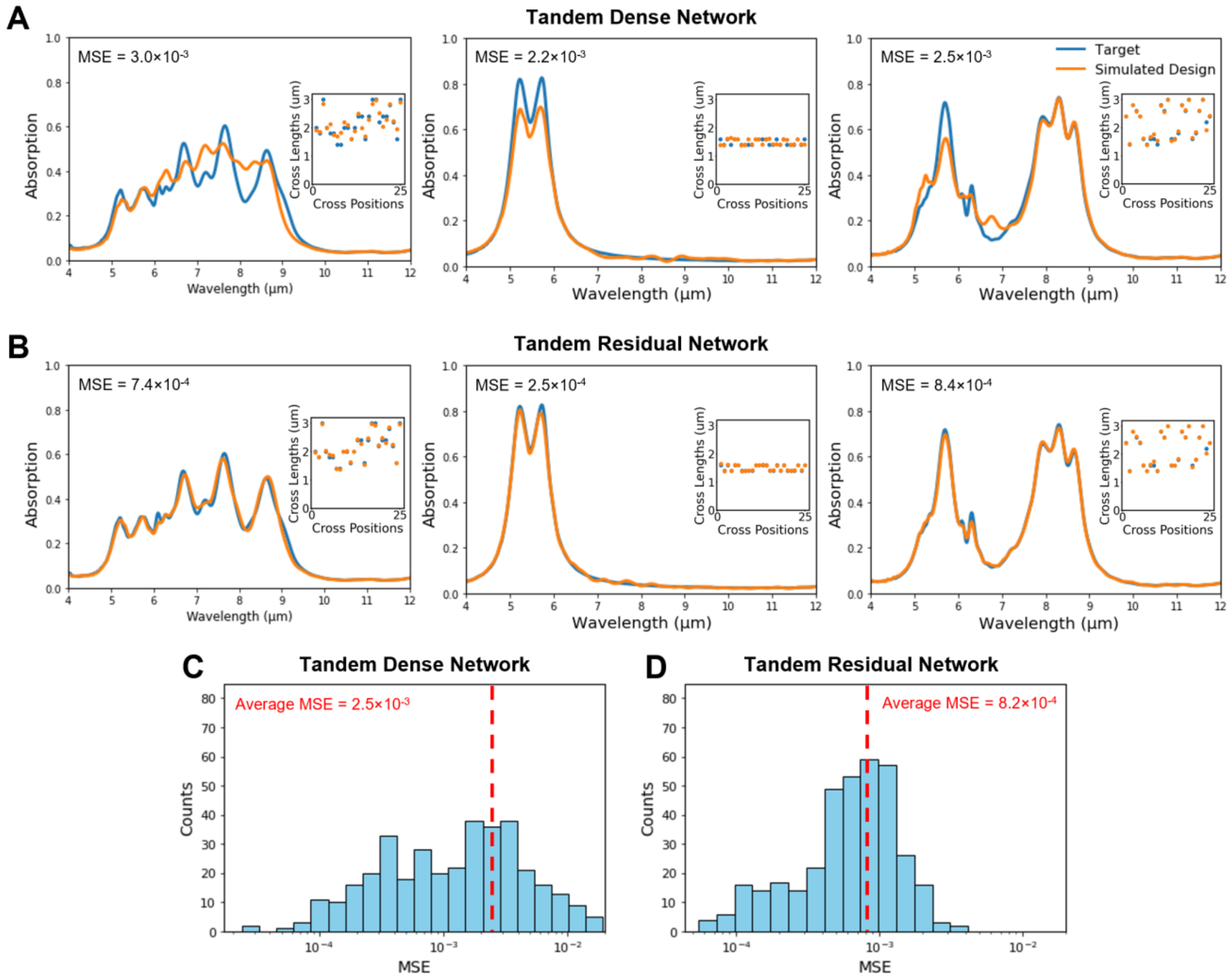


Figure 2. Performance evaluation based on mean-squared error (MSE) of target spectra and simulated designs for (a) the tandem dense and (b) tandem residual networks. The tandem residual network achieves a lower MSE for three distinct targets. Inset images show the design parameters

for the corresponding spectra. Statistical analyses across the entire validation dataset (over 300 spectra) for the tandem dense (c) and residual (d) networks indicate that the latter achieves a lower average MSE.

Using the customized loss function approach described above, we implemented a tandem residual network and compared its performance to the tandem dense network. Figure 2 presents several example test results from both networks; with new spectral targets from the validation dataset. These test results are obtained by simulating the predicted D', then comparing the target and simulated spectra (shown as blue and orange lines, respectively). It can be observed that across various spectral response patterns, the tandem residual network (Figure 2b) produces supercell designs with greater accuracy than the tandem dense network (Figure 2a). A larger statistical evaluation using the entire validation dataset (360 input spectra) reveals that the average validation MSE is approximately 2.5 × 10-3 for the tandem dense network (Figure 2c) and 8.2 × 10-4 for the tandem residual network (Figure 2d). Moreover, we observe that the tandem residual network exhibits a lower overall distribution of errors. As a result, we demonstrate that a tandem architecture composed of 1-D convolutional layers and residual building blocks is well-suited for supercell design and can outperform a tandem network (of the same network depth) that is based on fully connected layers.

We utilized the tandem residual network to generate new supercell metasurface designs with a broad range of spectral properties. Figure 3 presents a series of test cases comparing the target network inputs to the simulated results of the corresponding output designs. The inset images show the spatial geometries of each supercell designed by the network. For example, as shown in Figure 3a, after specifying a narrowband target with a full width half maximum (FWHM) of 0.5 µm, the network generated a periodic layout that matches the target spectra with over 90% accuracy as well as results from prior literature [42]. Similarly, in Figure 3b and 3c, dual-narrowband and triple-narrowband designs were created (with sharp resonances at two and three discrete wavelengths) that closely match their respective targets. In these multi-resonance structures, the supercells include additional cross dimensions that are associated with distinct resonances.

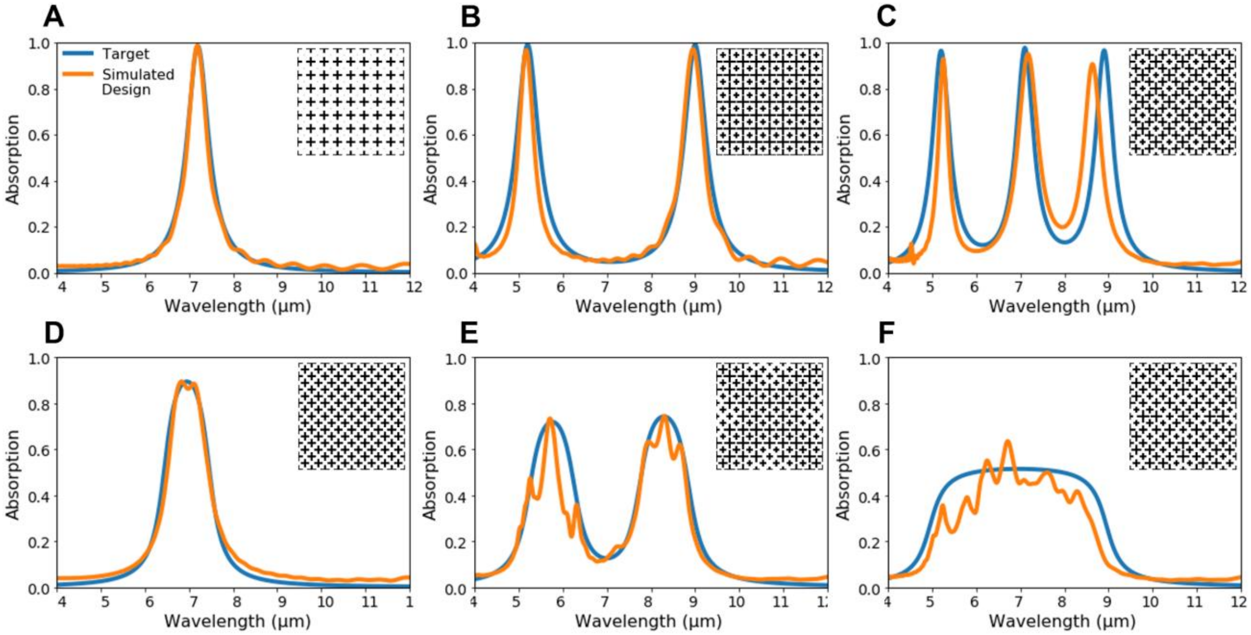


Figure 3. Inverse design of new supercell metasurfaces with the tandem neural network. The structures exhibit (a) narrowband, (b) dual-narrowband, (c) triple-narrowband, (d) broadband, (e) dual-broadband, and (f) graybody behaviors. Blue lines indicate the target spectra used as inputs to the network, and orange lines represent the simulated results of the output design parameters. Inset images show the physical layouts of the network-generated supercells.

The ability to construct an array of resonator geometries suggests that different resonant modes can be superimposed to achieve responses of arbitrary bandwidth [43]. Accordingly, we tasked the neural network with designing metasurfaces with various broadband characteristics (FWHM > 1 μm). In Figure 3d, a broadband structure with a FWHM of 1.5 μm is shown, and in Figure 3e, we increased the complexity of the target to design a structure with dual-broadband absorption peaks. Lastly, in Figure 3f, we demonstrate the design of a broadband graybody structure that encompasses the entire MIR range of resonance wavelengths captured by the training dataset (5-9 µm).

In the design of the aforementioned multi-resonance and broadband structures, the network not only defined the resonator dimensions required to achieve resonances at the target wavelengths, but also determined their appropriate placements within the lattice in order to reach the target absorption amplitudes. For example, as shown in Figure 4a, the network-designed triple-narrowband structure possesses three primary cross lengths that are responsible for resonances at 5.2, 7.2, and 8.6 µm. The high absorption amplitudes are attributed to the periodic and short-range ordered arrangements (repeating patterns spanning 1-2 subunit cell distances) of the resonators, which result in the strong dipole resonances seen in the electric field enhancement plots. When short-range order is converted to long-range order (patterns spanning beyond 2 subunit cell distances), additional response types are enabled. In particular, we observe in Figure 4b that the network can alter the relative positions of the same subunit resonators to produce a new response with lower peak amplitudes and a peak shift. As seen in the EM fields, this response is achieved by the new interactions that emerge from the modified arrangement of resonators, as well as modifications of the cross section of a given resonator by its neighbors. By introducing a larger assortment of cross geometries with more complex interactions, the net absorption spectra can also produce a graybody response (Figure 4c). Thus, by systematically predicting the subunit resonator dimensions as well as their spatial positions, the trained network can modulate absorption peak phase and amplitude by designing multiplexed metasurfaces with a range of underlying symmetries.

It is a challenging design task to combine multiple distinct resonant modes in a single metastructure while leveraging, or alternatively minimizing, hybridization between modes and maintaining high absorption per unit area [43,44]. Thus, multiplexed resonator structures impose an inherent tradeoff between broadband response and maximum absorption. Here, we seek to investigate the structure-property relationships of the explored metasurface class by leveraging the near-instantaneous calculation speed of the neural network. Previous studies have used pre-trained ML models for design space exploration and pattern discovery [32,57-59]. For instance, it has been shown that for a constrained domain, the fast inference speed of ML models can produce reasonably accurate estimates of an optical system's physical responses so that unnecessary exploration of the solution space can be avoided [60]. Similarly, to enable the exploration of our supercell design space, we use the pre-trained forward-modeling network as a validation mechanism for the design parameters predicted by the tandem network. As one example, we specified design targets using Lorentzian functions of increasing bandwidth (FWHM of 0.2-4 µm centered at 7 µm), illustrated in Figure 5a. The tandem network outputs were then fed into the forward-modeling network, and the resulting design predictions were compared to the initial targets. Full-wave simulation results of the tandem network-designed structures are also presented in Figure 5a, indicating that the network-predicted results match well with the ground truth. In this approach, the forward network effectively serves as a high-speed surrogate EM solver, replacing the FDTD software that was used to generate the training data.

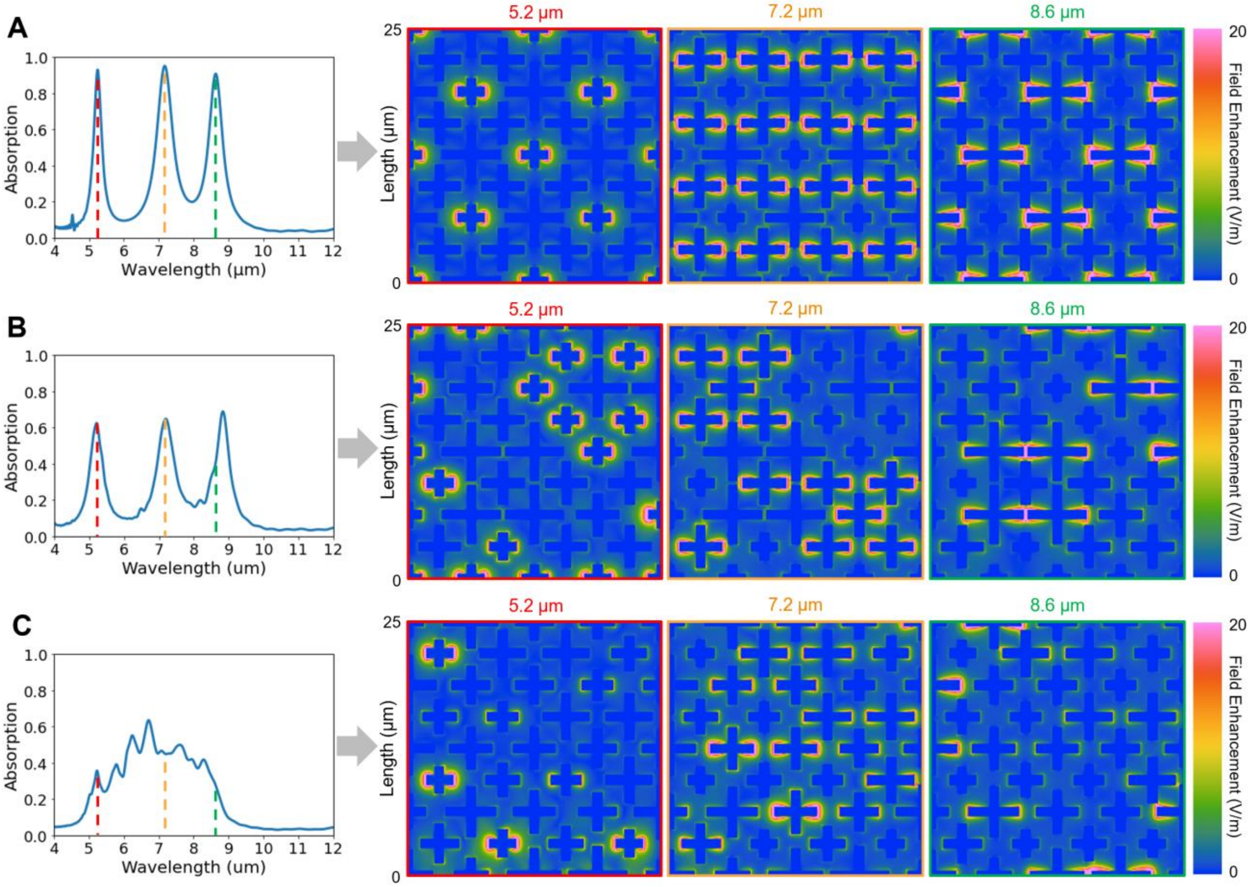


Figure 4. Relationships between supercell absorption properties and their subunit resonator spatial distributions. Simulated absorption spectra (of the tandem network-designed structures) and corresponding electric field profiles are shown for triple-narrowband structures with (a) high and (b) low absorption peaks and a (c) graybody structure. These plots reveal the dependence of absorption response on resonator geometry and position relative to other elements.

The design predictions reveal that when an unobtainable target was specified, the network designs a structure with the closest possible solution in the context of the supercell design space which it was trained on. As a result, we observe that as the target bandwidth increases, the discrepancy between the target response and the closest design (measured by MSE) increases as well (Figure 5b). This, in turn, allows us to numerically infer a relationship between the broadband response and the maximum obtainable absorption for this class of resonant metasurfaces, as estimated by the machine learning algorithm. By fitting these observations, we can further derive an estimate for maximum absorption at various bandwidths ($R^2 = 0.98$)

$$A_{max} = 0.0004f^2 - 0.0302f + 1.0154, \tag{1}$$

where A_{max} is the model's estimate of maximum absorption and f is the FWHM in THz. While we show that the structure-property relationships of a design space can be easily represented using a forward-modeling network, we note that the relation in Eqn. 1 (or any relation captured in a similar manner) is not universally applicable, but subject to the same parametric restrictions that were imposed on the training data (*i.e.*, cross widths fixed at 500 nm and cross lengths within the range of 1.4-3 µm). Metasurfaces with dimensions beyond the range restricted by the training data may exhibit a relationship that is different from Eqn. 1. However, the training dataset may simply be updated to incorporate a wider range of geometries to account for such parametric restrictions.

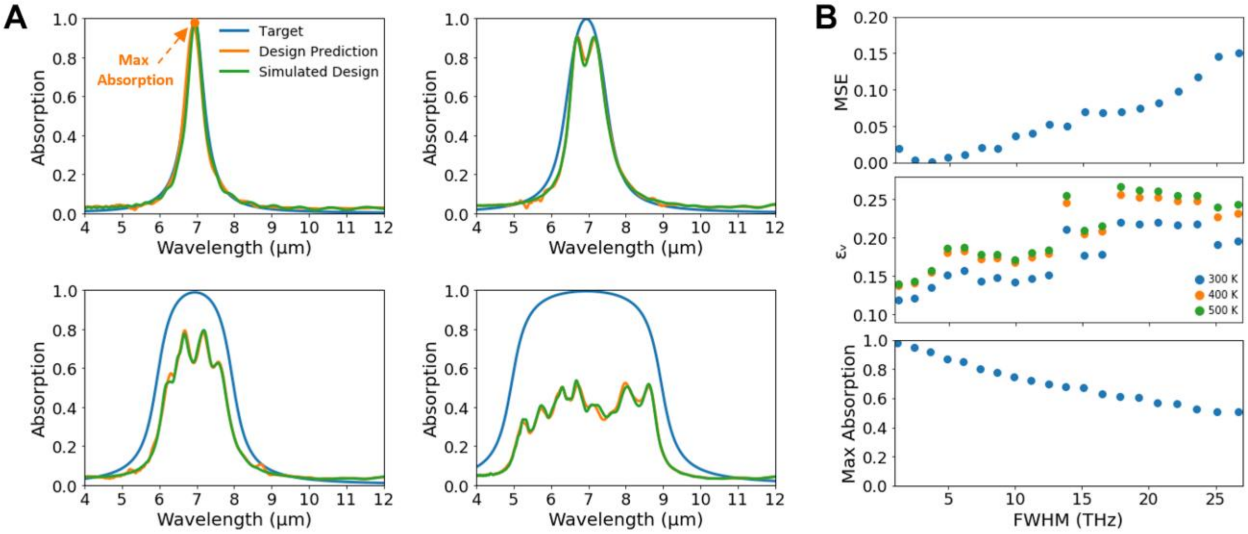


Figure 5. Probing metasurface design relations using the tandem network. (a) Tandem network input targets (blue lines) for various Lorentzian functions (FWHM of 0.6, 1, 2, and 3 µm centered at 7 µm) and the corresponding forward network-predicted results (orange lines). The network is

unable to identify designs that exceed the model's estimate of bandwidth / maximum absorption of the explored class of supercell metasurfaces. Full-wave simulation results are shown (green lines) for comparison, indicating that the network-predicted results match well with the ground truth. (b) Network-determined design trends and metrics, including the MSE between target and design responses, thermal emittance of the metasurface, and max absorption as functions of FWHM (THz).

As an additional example of discovering application-specific design insights through the neural network, we can calculate the average normal-incidence emissivity of the optimized supercell metasurfaces within defined target bandwidths:

$$\underline{\varepsilon} = \frac{\int_{v_1}^{v_2} I_{BB}(T,v)\cdot\varepsilon(v)dv}{\int_{v_1}^{v_2} I_{BB}(T,v)dv}. \tag{2}$$

Here, $I_{BB}(T,v) = \frac{2hv^3}{c^2}\frac{1}{e^{\frac{hv}{k_BT}}-1}$ is the spectral radiance of a blackbody at temperature *T*, where *h* is Planck's constant, *kB* is the Boltzmann constant, *c* is the speed of light, and *v* is frequency. The lower and upper bounds of the integral (*v*1 and *v*2) are derived from the evaluated spectral range (4-12 μm). ε*(v)* is the metasurface's spectral emittance, which is equal to *A(v)* by Kirchoff's law. In this case, by querying the neural network in a cyclic manner to solve for emittance (at various temperatures) as a function of the target bandwidth, we can find the relationship between the two parameters (Figure 5b) in a remarkably short time frame (less than one minute). Overall, we observe that as the sought bandwidth (FWHM) increases, the MSE between the target (with maximum absorption across the entire bandwidth) and design response increases and the maximum absorption point of the achievable design decreases. Furthermore, the integrated normal incidence emittance increases as the additional bandwidth compensates for the decreases in the peak absorption/emittance value. However, as can be seen in Figure 5b, the precise relationship is complex and depends both on the bandwidth being specified and the temperature of the metasurface because the blackbody spectral radiance changes with temperature. Thus, by training a neural network that is tasked with the inverse design of complex supercell metasurfaces, we demonstrate that the same framework can be strategically leveraged to rapidly identify design trends and dependencies associated with application-specific properties (within the parameter ranges represented by the trained class of metasurfaces).

3.3 Conclusions

In this article, we demonstrated a machine learning approach to the inverse-design of multiplexed supercell metasurfaces with over 100 subunit elements. The added degrees of freedom offered by a supercell architecture, relative to periodic single-element structures, yields new tailored capabilities including multi-resonant and broadband responses. By forming a cascaded architecture with an inverse-modeling and forward-modeling network, we show that a tandem network effectively overcomes the nonuniqueness problem present in supercell architectures, and can successfully learn a high-dimensional design space of over three trillion possible designs using only 3,600 data instances. Moreover, we present a network architecture based on 1-D convolutional layers and residual building blocks that is capable of generating designs with greater accuracy than a conventional tandem network based on fully connected layers. Through the superposition and coupling of multiple resonant modes in a compact region, the tandem residual network can efficiently design supercell structures with a range of symmetries that yield narrowband, broadband, and multi-resonant responses. The network not only predicts the geometric parameters for an array of resonators (*e.g.*, resonator widths, lengths, radii, etc.), but also their optimum spatial arrangement towards satisfying a specified target. Therefore, the presented approach enables additional degrees of complexity in metasurface design by directly generating structures with a wide range of unique elements while accounting for coupling between these elements. Though we sought to maximize implementation efficiency by minimizing the required training data, we expect that the performance of our tandem network can be improved with more training data and a larger network architecture. Furthermore, we demonstrate that the network itself can be utilized to approximate the structure-property relationships of the investigated class of metasurfaces (within the parameter ranges represented by the training data). By using the forward-modeling network as a full-wave EM simulator, high-speed parameter sweeps can be performed to capture property-specific design trends such as maximum absorption and thermal emittance as a function of bandwidth. Importantly, our results show that DNN-based approaches can efficiently design and characterize large-scale supercell metasurfaces with numerous discrete resonators. We believe our results can expedite the development of supercell-class nanophotonic structures and materials, which may in turn yield new tailored capabilities not achievable through conventional periodic nanostructures.

3.4 Supporting Information

Forward-Modeling Network Optimization

To train the tandem network for inverse design, we first optimized the architecture of the forward-modeling network through extensive hyperparameter tuning. Figure S1 shows the tuning results, where we compare the validation loss of the starting controlled architecture to the losses of networks trained after changing a single dependent variable. Implemented through the TensorFlow framework, the controlled architecture consists of 2 hidden layers, each with 100 neurons, sigmoid activation functions, a batch size of 10 data instances, and the Adam optimizer.

The learning rate is 0.001 with an exponential decay of 10^{-5}. The tested dependent variables include: number of hidden layers, number of neurons within each layer, activation function, batch size, and optimizer.

Figure S1a shows a comparison of different batch sizes (10, 100, and 1000), where we observe noticeable increases in loss as the batch size was increased. In Figure S1b, we tested three commonly used optimization algorithms: Adam, stochastic gradient descent (SGD), and RMSprop. SGD yielded higher losses than the other algorithms while Adam and RMSprop resulted in similar losses. However, RMSprop plateaued much sooner than Adam, indicating that the network was able to improve further with Adam. In Figure S1c, we compared the following activation functions: Sigmoid, TanH (hyperbolic tangent), ReLU (rectified linear unit), Leaky ReLU, and Parametric ReLU. Here, we observe that the Sigmoid and TanH functions resulted in the lowest losses, while the loss progression was more stable and ultimately lower with Sigmoid. We then tested various numbers of neurons and layers (Figure S1d and S1e), and found that increasing from 100 to 200 neurons garnered small improvements, while increasing the number of layers alone yielded insignificant changes. Figure S1f shows the full integration of the individually-optimized hyperparameters and the use of more elaborate combinations of neurons and layers, which resulted in considerable overall performance improvements. From these tests, we found that the 50-100-200-400 neuron architecture has the best performance. Adding more neurons to the optimized architecture did not improve the performance any further and unnecessarily increased training time.

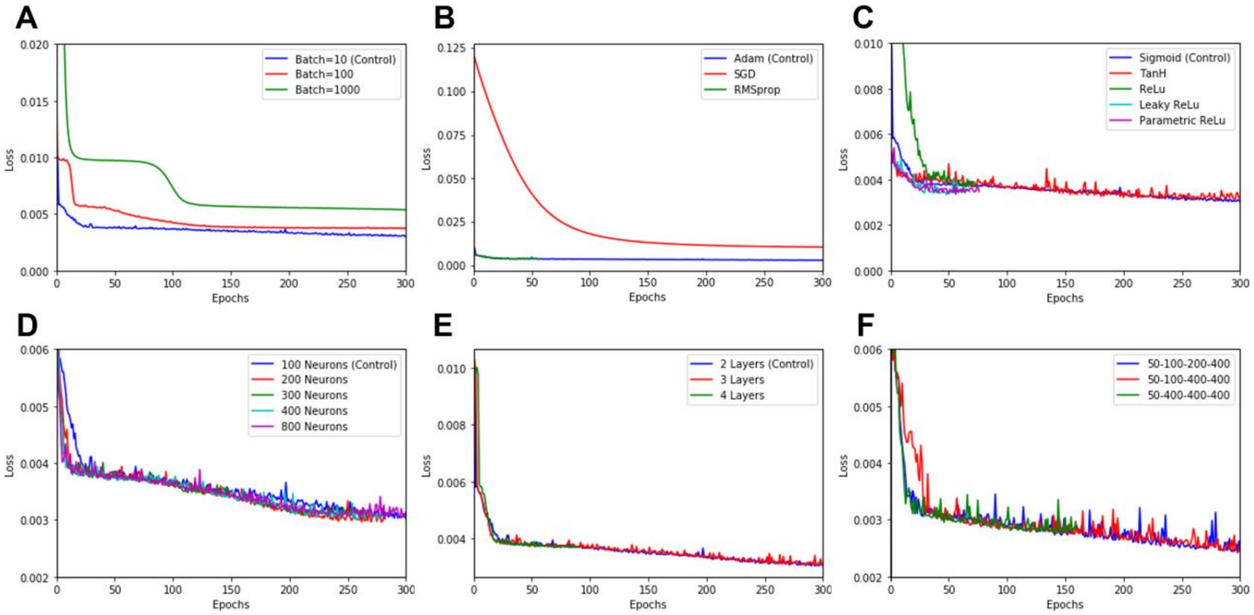


Figure S1. Forward network hyperparameter tuning. Validation loss of a controlled architecture in comparison to various dependent variables, including: (a) batch size, (b) optimizer, (c) activation function, (d) number of neurons per layer, and (e) number of hidden layers. (f) Final optimized architecture comparison.

Training Dataset Analysis

The primary advantage of the tandem network architecture is its ability to address the nonuniqueness scattering problem (as described in the main text), where drastically different design parameters can meet a similar target response (*i.e.*, a one-to-many problem). To illustrate this problem in the context of our training dataset, Figure S2 shows examples where two substantially different supercell design configurations map to nearly-identical dual-band (Figure S2a) and triple-band (Figure S2b) responses. In Figure S3a, the nonunique dual-band and triple-band spectra were passed into the inverse modeling network, and a poor match between the input spectra and the simulated spectra of the optimized design is observed. However, when the same spectra were passed into the tandem network (Figure S3b), the accuracy between the input spectra and the simulated parameters is substantially improved. Thus, we validate that the tandem network is better than the inverse modeling network at resolving the nonuniqueness issue.

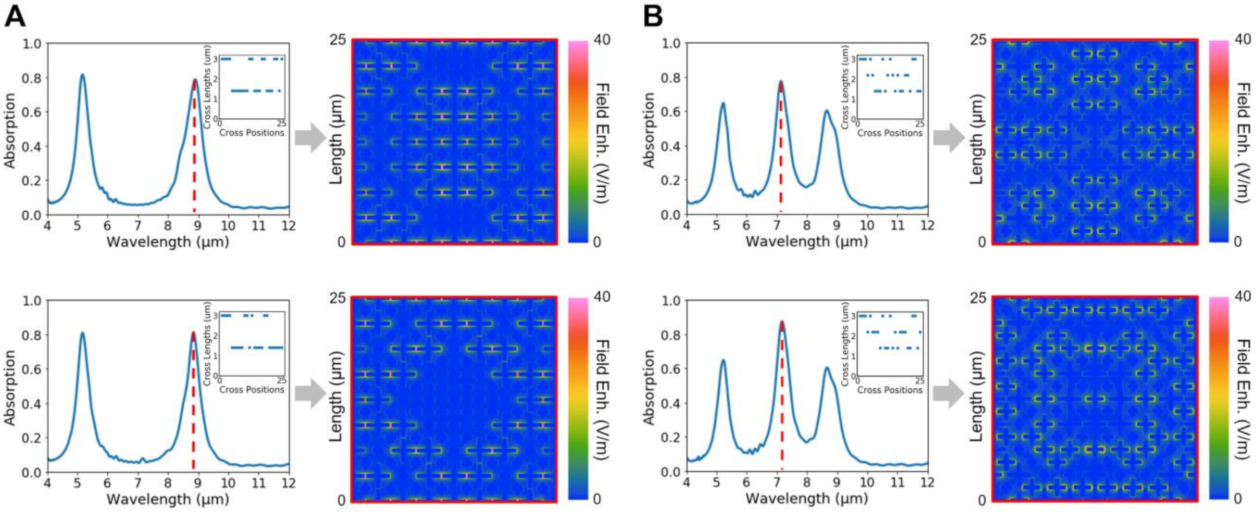


Figure S2. Examples of nonuniqueness in the training dataset. Two very different supercell design layouts have nearly-identical (a) dual-band and (b) triple-band absorption responses. Inset images show the design parameters corresponding to the shown absorption spectra.

To expedite our deep learning efforts, we sought to minimize the total simulation time and therefore the amount of training data. However, we also had to ensure that the dataset size was large enough to maximize the network's ability to learn supercell designs. In that regard, as shown in Figure S4, we trained the optimized tandem network architecture with increasing increments of training data and evaluated the corresponding validation losses. First, we trained the network with a training set size of 300 and recorded the network's final validation loss using two different validation datasets. The first validation dataset was a fixed group of 300 data instances. This dataset was intended to monitor the network's growth as it trained toward a predetermined set of goals. The second validation dataset came from randomly splitting 10% of the data instances within the available training set. The validation loss derived from the second dataset informs how well the network is able to generalize from the amount of data it learned from. We repeated this validation process after increasing the training set size in 300-increment steps, and found that both validation losses began to converge to 2.5 × 10^{-3} at approximately 3,000 data instances. The reported losses represent the averaged results of 5 training cycles for each training set size. Error bars were omitted due to negligible differences in the range of losses. Thus, we demonstrate that the final model resulted in the optimal performance while minimizing the amount of data required for deep learning.

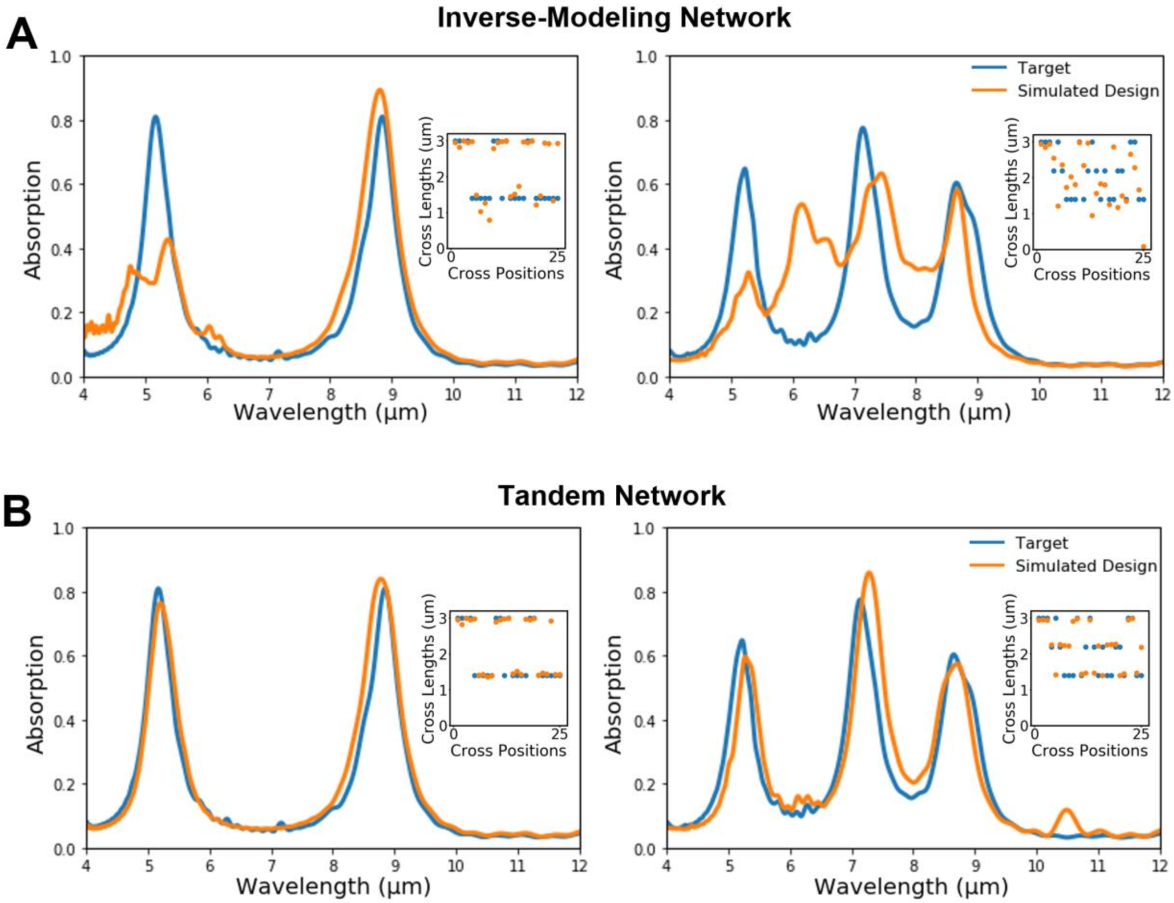


Figure S3. Comparisons between target absorption spectra (from the previous Figure) and the simulated spectra for the optimized designs from the (a) inverse-modeling network and the (b) tandem network. The tandem network has greater accuracy in predicting design parameters than the inverse network due to stronger convergence. Inset images show design parameters for the corresponding spectra.

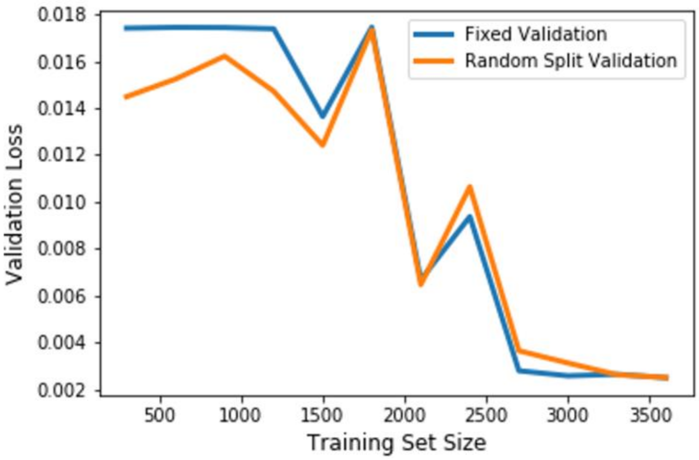


Figure S4. Training set size analysis. Validation loss vs. training set size with fixed and randomly split validation datasets.

3.5 References

[1] Olthaus J, Schrinner P, Reiter D. Optimal Photonic Crystal Cavities for Coupling Nanoemitters to Photonic Integrated Circuits. Adv Quantum Technol 2020, 3:1900084.

[2] Yoshimi H, Yamaguchi T, Ota Y, Arakawa Y, Iwamoto S. Slow light waveguides in topological valley photonic crystals. Opt Lett 2020, 45:2648-2651.

[3] Bin Tarik F, Famili A, Lao Y, Ryckman J. D. Robust optical physical unclonable function using disordered photonic integrated circuits. Nanophotonics 2020, 20200049.

[4] Mittapalli V, Khan H. Excitation Schemes of Plasmonic Angular Ring Resonator-Based Band-Pass Filters Using a MIM Waveguide. Photonics 2019, 6(2):41.

[5] Ding F, Wang Z, He S, Shalaev VM, Kildishev AV. Broadband high-efficiency half-wave plate: a supercell-based plasmonic metasurface approach. ACS Nano 2015, 9(4):4111-9.

[6] Aoni RA, Rahmani M, Xu L, et al. High-efficiency visible light manipulation using dielectric Metasurfaces. Sci Rep 2019, 9(1):1-9.

[7] Wu PC, Tsai WY, Chen WT, et al. Versatile polarization generation with an aluminum plasmonic metasurface. Nano Lett 2017, 17(1):445-52.

[8] Ma Q, Chen L, Jing HB, et al. Controllable and programmable nonreciprocity based on detachable digital coding metasurface. Adv Opt Mater 2019, 7(24):1901285.

[9] Liu V, Miller DA, Fan S. Ultra-compact photonic crystal waveguide spatial mode converter and its connection to the optical diode effect. Opt Express 2012, 20(27):28388-97.

[10] Guo X, Ding Y, Chen X, Duan Y, Ni X. Molding Free-Space Light with Guided-Wave-Driven Metasurfaces. 2020, arXiv preprint, arXiv:2001.03001.

[11] Hegde RS. Deep learning: a new tool for photonic nanostructure design. Nanoscale Adv 2020, 2: 1007–1023.

[12] Gondarenko A, Lipson M. Low Modal Volume Dipole-like Dielectric Slab Resonator. Opt Express 2008, 16:17689-17694.

[13] Kao CY, Osher S, Yablonovitch E. Maximizing Band Gaps in Two-Dimensional Photonic Crystals by Using Level Set Methods. Appl Phy. B: Lasers Opt 2 2005, 81:235-244.

[14] Piggott AY, Lu J, Lagoudakis KG, et al. Inverse Design and Demonstration of a Compact and Broadband on-Chip Wavelength Demultiplexer. Nat Photonics 2015, 9(6):374-377.

[15] Shen B, Wang P, Polson R, Menon R. An Integrated Nanophotonics Polarization Beamsplitter with 2.4x2.4 μm2 Footprint. Nat Photonics 2015, 9:378-382.

[16] Oskooi A, Mutapcic A, Noda S, et al. Robust Optimization of Adiabatic Tapers for Coupling to Slow-Light Photonic-Crystal Waveguides. Opt Express 2012, 20:21558-21575.

[17] Seliger P, Mahvash M, Wang C, Levi A. Optimization of Aperiodic Dielectric Structures. J Appl Phys 2006, 100:034310.

[18] Verweij S, Liu V, Fan S. Accelerating simulation of ensembles of locally differing optical structures via a Schur complement domain decomposition. Opt Lett 2014,39(22):6458-61.

[19] Lin Z, Johnson SG. Overlapping domains for topology optimization of large-area metasurfaces. Opt Express 2019, 27(22):32445-53.

[20] Elesin Y, Lazarov BS, Jensen JS, Sigmund O. Time domain topology optimization of 3D nanophotonic devices. Photonic Nanostruct 2014, 12(1):23-33.

[21] Yeung C, Tsai JM, King B, Kawagoe Y, Ho D, Knight M, Raman AP. Elucidating the Behavior of Nanophotonic Structures Through Explainable Machine Learning Algorithms. ACS Photonics 2020.

[22] Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H. State-of-the-art in artificial neural network applications: A survey. Heliyon 2018, 4(11):e00938.

[23] Muhammad W, Hart GR, Nartowt B, Farrell JJ, Johung K, Liang Y, Deng J. Pancreatic cancer prediction through an artificial neural network. Frontiers in Artificial Intelligence 2019, 2:2.

[24] Conduit B, Jones N, Stone H, Conduit G. Design of a nickel-base superalloy using a neural network. Mat and Des 2017, 131:358-365.

[25] So S, Rho J. Designing nanophotonic structures using conditional deep convolutional generative adversarial networks. Nanophotonics 2019, 8:1255–1261.

[26] Liu Z, Zhu D, Rodrigues SP, Lee KT, Cai W. Generative Model for the Inverse Design of Metasurfaces. Nano Lett 2018, 18:6570–6576.

[27] Liu D, Tan Y, Khoram E, Yu Z. Training deep neural networks for the inverse design of nanophotonic structures. ACS Photonics 2018, 5(4):1365-9.

[28] Peurifoy J, Shen Y, Jing L,et al. Nanophotonic particle simulation and inverse design using artificial neural networks. Sci Adv 2018, 4(6):eaar4206.

[29] An S, Fowler C, Zheng B, et al. A deep learning approach for objective-driven all-dielectric metasurface design. ACS Photonics 2019, 6(12):3196-207.

[30] Hegde R. Photonics Inverse Design: Pairing Deep Neural Networks With Evolutionary Algorithms. IEEE J. Sel. Top. Quantum Electron 2019, 26(1):2933796.

[31] Ma W, Cheng F, Liu Y. Deep-learning-enabled on-demand design of chiral metamaterials. ACS Nano 2018, 12(6):6326-34.

[32] Inampudi S, Mosallaei H. Neural network based design of metagratings. Appl Phys Lett 2018, 112(24):241102.

[33] Harper ES, Coyle EJ, Vernon JP, Mills MS. Inverse design of broadband highly reflective metasurfaces using neural networks. Phys Rev B 2020, 101(19):195104.

[34] Ogawa S, Kimata M. Metal-insulator-metal-based plasmonic metamaterial absorbers at visible and infrared wavelengths: a review. Materials 2018, 11(3):458.

[35] Vorobyev AY, Topkov AN, Gurin OV, Svich VA, Guo C. Enhanced absorption of metals over ultrabroad electromagnetic spectrum. Appl Phys Lett 2009, 95(12):121106.

[36] Ye YQ, Jin Y, He S. Omnidirectional, polarization-insensitive and broadband thin absorber in the terahertz regime. JOSA B 2010, 27(3):498-504.

[37] Chen HH, Su YC, Huang WL, Kuo CY, Tian WC, Chen MJ, Lee SC. A plasmonic infrared photodetector with narrow bandwidth absorption. Appl Phys Lett 2014, 105(2):023109.

[38] Ma Y, Chen Q, Grant J, Saha SC, Khalid A, Cumming DR. A terahertz polarization insensitive dual band metamaterial absorber. Opt Lett 2011, 36(6):945-7.

[39] Shen X, Cui TJ, Zhao J, Ma HF, Jiang WX, Li H. Polarization-independent wide-angle triple-band metamaterial absorber. Opt Express 2011, 19(10):9401-7.

[40] Luo H, Cheng YZ, Gong RZ. Numerical study of metamaterial absorber and extending absorbance bandwidth based on multi-square patches. Eur Phys J B 2011, 81(4):387-92.

[41] Gao L, Li X, Liu D, Wang L, Yu Z. A bidirectional deep neural network for accurate silicon color design. Adv Mater 2019, 31(51):1905467.

[42] Liu X, Tyler T, Starr T, Starr AF, Jokerst NM, Padilla WJ. Taming the blackbody with infrared metamaterials as selective thermal emitters. Phys Rev Lett 2011, 107(4):045901.

[43] Fan RH, Xiong B, Peng RW, Wang M. Constructing metastructures with broadband electromagnetic functionality. Adv Mater 2019, 1904646.

[44] Ma W, Wen Y, Yu X. Broadband metamaterial absorber at mid-infrared using multiplexed cross resonators. Opt Express 2013, 21(25):30724-30.

[45] Jiang J, Sell D, Hoyer S, Hickey J, Yang J, Fan JA. Free-Form Diffractive Metagrating Design Based on Generative Adversarial Networks. ACS Nano 2019, 13(8): 8872-78.

[46] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proc IEEE Conference on Computer Vision and Pattern Recognition 2016, 770-8.

[47] Liu Z, Zhu D, Lee K, Kim A, Raju L, Cai W. Compounding Meta-Atoms into Metamolecules with Hybrid Artificial Intelligence Techniques. Adv Mater 2020, 32:1904790.

[48] Naseri P, Hum S. A Generative Machine Learning-Based Approach for Inverse Design of Multilayer Metasurfaces. 2020, arXiv preprint, arXiv:2008.02074.

[49] Zhelyeznyakov M, Brunton S, Majumdar A. Deep learning to accelerate Maxwell's equations for inverse design of dielectric metasurfaces. 2020, arXiv preprint, arXiv:2008.10632.

[50] Prodan E, Radloff C, Halas N J Nordlander P. A Hybridization Model for the Plasmon Response of Complex Nanostructures. Science 2003, 302: 1089171.

[51] Ince T, Kiranyaz S, Eren L, Askar M, Gabbouj M. Real-Time Motor Fault Detection by 1-D Convolutional Neural Networks. IEEE Trans Ind Electron 2016, 63:7067-75.

[52] Xiao B, Xu Y, Bi X, Zhang J, Ma, X. Heart sounds classification using a novel 1-D convolutional neural network with extremely low parameter consumption. Neurocomputing 2020, 392:153-9.

[53] Chao Q, Tao J, Wei X, Wang Y, Meng L, Liu C. Cavitation intensity recognition for high-speed axial piston pumps using 1-D convolutional neural networks with multi-channel inputs of vibration signals. Alexandria Eng J 2020.

[54] Tahersima M, Kojima K, Koike-Akino, T, et al. Deep Neural Network Inverse Design of Integrated Photonic Power Splitters. Sci Rep 2019, 9:1368.

[55] Sajedian I, Kim J, Rho J. Finding the optical properties of plasmonic structures by image processing using a combination of convolutional neural networks and recurrent neural networks. Microsystems Nanoeng 2019, 5:27.

[56] Jiang J, Fan J A. Multiobjective and categorical global optimization of photonic structures based on ResNet generative neural networks. Nanophotonics 2020.

[57] An S, Zheng B, Shalaginov M, et al. A freeform dielectric metasurface modeling approach based on deep neural networks. 2019, arXiv preprint, arxiV:2001.00121.

[58] Melati D, Grinberg Y, Dezfouli M, et al. Mapping the global design space of nanophotonic components using machine learning pattern recognition. Nat Commun 2019, 10:4775.

[59] Nadell C C, Huang B, Malof, J M, Padilla W J. Deep learning for accelerated all-dielectric metasurface design. Opt Express 2019, 27(20):27523-35.

[60] Ma W, Liu Z, Kudyshev Z, Boltasseva A, Cai W, Liu Y. Deep Learning for the design of photonic structures. Nature Photonics 2020, 1-14.

4. Elucidating the Behavior of Nanophotonic Structures Through Explainable Machine Learning Algorithms

4.1 Introduction

Nanophotonic structures and devices have enabled a broad range of transformative technologies including photonic integrated circuits for optical communication [1-3], and metasurfaces that compactly control the propagation of electromagnetic waves [4-6]. The conventional approach to designing nanophotonic structures is via numerical simulations based on fundamental physical laws (*e.g.*, Maxwell's Equations). This design technique, which we here refer to as 'forward design' [7], is well established, but depends on computationally expensive trial-and-error processes to obtain target functionalities. To address the limitations of forward design, 'inverse design' methods have been developed to generate nanophotonic structures that meet predefined targets8. Methodologies such as topology [9-11,31] and adjoint-based optimization [12,13,32] have shown promising results in designing complex structures that deliberately interact with electromagnetic fields, often at sub-wavelength scales, to enable a desired response. While inverse design algorithms can yield high-performance designs that go beyond human intuition, the algorithms can miss globally optimal designs [7,8], produce unstable results [14], are often constrained by long runtimes. Additionally, inverse design methods typically operate as 'black boxes' and cannot explain the underlying relationship between a designed physical structure and its electromagnetic response.

In recent years, machine learning (ML) techniques have emerged as alternate strategies for both forward and inverse design of photonic structures. Tandem neural networks have been used to design multilayer thin films based on target transmission spectra [15], and for spatially complex geometries, generative adversarial networks (GANs) have produced images of structure designs, given an input of desired spectral properties [16-18]. Convolutional neural networks (CNNs) have also been used to map physical geometries to spectral and spatial properties in two- and three-dimensional settings, respectively [33,34]. However, the internal decision models built by these machine learning algorithms are not well-understood; their contents, similar to conventional inverse design approaches, are widely regarded as ‘black boxes’ [19,20]. This challenge emerges from the fact that supervised ML algorithms, including neural networks, learn by optimizing up to millions of internal variables (weights and biases) to fit the training data [21]. Consequently, it is exceptionally challenging to explain *why* a machine learning algorithm makes one prediction over another. Thus, the lack of explainability is a key limitation for both conventional and ML-based inverse design strategies [35].

In response to the long-standing ‘black box’ problem of ML, explainable artificial intelligence (XAI) and ML approaches have become a topic of active inquiry. This rapidly developing field aims to analyze and understand ML models in general, with domain-specific demonstrations of scientific insights that might in turn emerge [36,37]. XAI approaches include sensitivity analysis, Taylor decomposition, deconvolution, guided backpropagation, and layer-wise relevance propagation [38-41]. For image-based classification and regression these methods typically create a salience map (or heatmap) to highlight small portions of the computation that are most relevant to the context at hand, thus explaining the features that contribute most to a model’s predictions. While explainability approaches have recently shown promise in better understanding machine learning outcomes in chemical, biological, and physical models [42-45], their use in photonics remains largely unexplored.

Motivated by these developments, in this article, we uncover what a class of machine learning algorithms (CNNs) has learned regarding the underlying physical principles which govern specific classes of nanophotonic structures. CNNs are deep neural networks widely used for image analysis and classification [22]. As shown in Figure 1a, we first created two-dimensional images representing the geometries of metal-dielectric-metal metamaterial resonators. This class of metamaterials was selected due to their ease of fabrication, compact structure, and ability to achieve high absorbance across a broad wavelength range, making them amenable to a wide range of spectral applications while enabling the rapid generation of training data for deep learning. Next, we demonstrate that the CNN can accurately perform forward design by learning the relationships between the metamaterial-structure images and their absorption spectra over mid-infrared wavelengths (Figure 1b). We then use the Deep SHapley Additive exPlanations (SHAP) framework [24] to ‘open the black box’ and explain the CNN’s predictions (Figure 1c). Deep SHAP combines DeepLIFT [46], a previously employed method for decomposing output predictions via backpropagation, with Shapley values, a metric that determines feature relevance, to generate pixel-by-pixel explanation heatmaps [24]. The explanations obtained with Deep SHAP show that the CNN has learned important physical behaviors of the class of metamaterials studied, including the relationships between structural elements and optical responses for both simple and freeform resonator geometries. Our approach uncovers specific geometric contributions to ML predictions of nanophotonic device properties, and thus allow us to both better understand the behavior of complex nanophotonic devices and identify pathways to improved designs (Figure 1d).

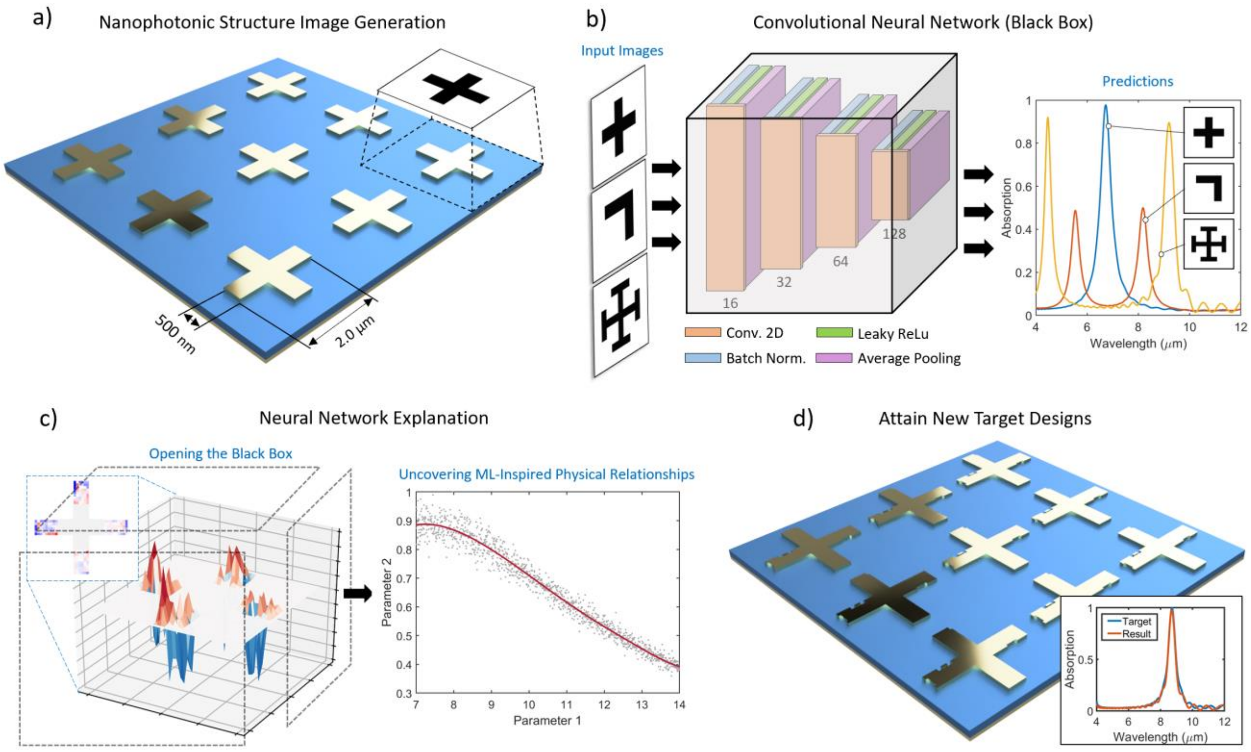


Figure 1. (a) Converting 3D metal-dielectric-metal metamaterials into 2D representations for image-based machine learning. (b) Training a convolutional neural network (CNN) to predict the electromagnetic response of input images. (c) Elucidating the underlying physics learned by a CNN by explaining the relationships between structural features and predicted parameters. (d) Leveraging the explained relationships to construct new designs with new target properties.

4.2 Results and Discussion

Forward Design Convolutional Neural Network (CNN) Development and Evaluation

We first developed and trained a CNN for the forward design of nanophotonic structures, such that when it is given an image of a nanophotonic structure as input, it outputs the associated absorption spectrum over a particular wavelength range. To constrain the problem further, we focused on a specific class of nanophotonic structure: metal-insulator-metal (MIM) metamaterials designed to operate at mid-infrared wavelengths [23]. We performed three-dimensional finite-difference time domain (FDTD) simulations of 10,000 unique structures in Lumerical, and generated 10,000 two-dimensional images of the resonator layers and their corresponding absorption spectra. The simulated structures, previously demonstrated in literature to possess selective thermal emissivity over a large bandwidth [23], consist of a 100 nm gold bottom layer, a 200 nm Al2O3 dielectric middle layer, and a 100 nm gold resonator top layer (within a 3.2 μm × 3.2μm unit cell). The dimensions of the top layer were progressively adjusted for design variation. The designs consist of cross-shapes, box-shapes (hollow and solid), cross-shapes with perpendicular resonators along the arm tips, and the inverted versions of each shape (as shown in Figure S1 of the Supporting Information). The models were then converted into two-dimensional images, and each image was associated with an 800-point vector of absorption values (ranging from 0 to 1) across fixed wavelengths (4 μm to 12 μm). Periodic boundary conditions were applied along the x- and y-planes. Each image was resized to 40 × 40 pixels and converted to grey-scale for ease of training.

After generating the training data, we trained multiple CNN architectures, with 10% of the training dataset used for validation, to determine the optimum hyperparameters. Table S1 presents each of the trained models along with their validation root-mean-square error (RMSE) and training time. Figure 2a shows the predicted output spectra of the CNN when six new and unknown images were used as inputs, as well as the FDTD simulation results corresponding to each image. The simulations were performed by converting the images into the top layer of the MIM structure. We observe through comparison of the simulations and the CNN predictions that the network exhibits a high degree of accuracy in predicting the absorption spectra of a broad range of resonator geometries not present in the training set. The wavelength and amplitude of the predicted resonance peaks are aligned with the simulated peaks (with over 95% mean absolute accuracy). Away from the peaks, we note some minor variation relative to the FDTD simulated results. On average, each prediction was generated in 0.270 ± 0.043 seconds (n = 10), while each simulation took approximately 30 minutes (yielding a 6,500x improvement). The results here demonstrate that the CNN successfully performed the devised forward-design task with high accuracy.

The high accuracy of the CNN's predictions raises an intriguing question: has the CNN, to some extent, learned the physical relationships between the class of nanophotonic structures we explored and their absorption spectra? Normally, the information required to answer this question is embedded within the neural network's many thousands of internal weights and parameters, which is represented by a hierarchy of filters (or neurons). CNNs extract information from images by applying these filters to an input image [28]. The filters are optimized such that the error is minimized when comparing the CNN's output to the target output. Figure 2b shows several examples of the two-dimensional filters that the CNN is composed of (among over 100,000 available filters). The dark squares indicate small weight values and the light squares represent large weight values. As shown in Figure 2c, we can apply these filters to an input image, and capture the CNN's behavior in a series of feature maps. These feature maps can provide insights into the CNN's response to specific areas of the image at any point in the model. For example, in Layer 3, the CNN places more weight on the edges of the cross, while placing less weight on the inside, indicating that generic features such as lines and edges are captured in the initial layers. However, as the input progresses deeper into the model, the feature maps provide progressively less interpretable information. The connection between mapped features and the network's final output remain hard to discern, especially at the deeper layers. Thus, although we can identify the shapes and features extracted by the network's initial layers, it is challenging to synthesize this information into an understandable explanation of the model's decision. Attempting to 'open the black box' in this sense provides limited utility with regards to model interpretation and verification, since analyzing individual filters within a network does not guarantee a coherent explanation for an entire model or even a specific prediction. Furthermore, this form of internal analysis is model specific. Different architectures may yield various feature-map responses in the corresponding layers, leading to inconsistencies in explanations and interpretations.

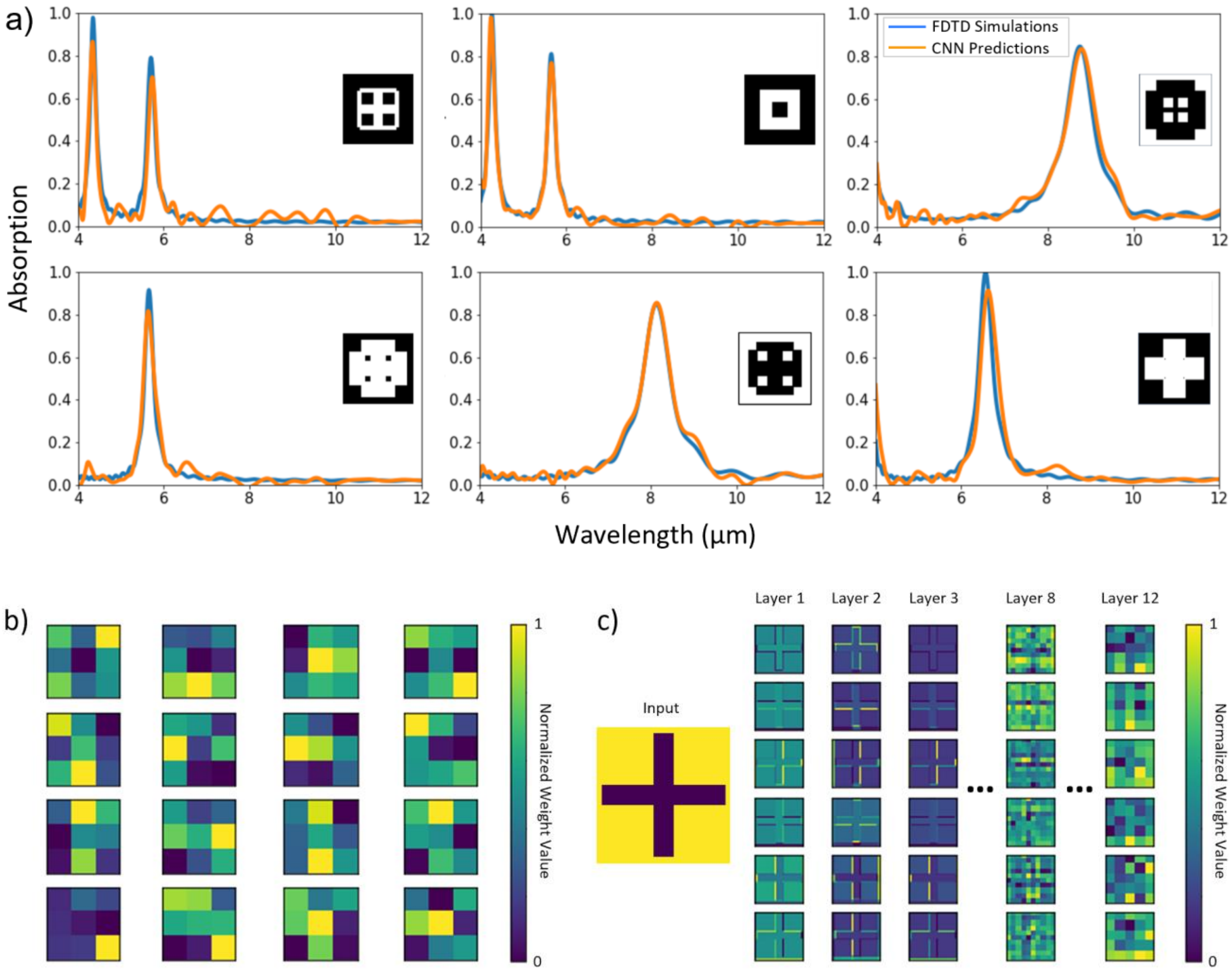


Figure 2. (a) CNN-predicted absorption spectra *vs.* FDTD simulations of six new nanophotonic structures (shown in the inset images), revealing the high accuracy of the CNN in performing forward design-based multiphysics structural analysis. (b) Examples of the filters and weights within the CNN. Dark squares represent small weights and light squares represent large weights. (c) Features maps showing what happens inside the model in response to an input image. Generic features such as lines and edges are captured in the initial layers, but the maps are less interpretable as we progress deeper into the model.

Explaining the CNN's Predictions

To explain the CNN's behavior and draw useful conclusions from the network's internal model, we instead use the recently developed Deep SHAP method (hereon referred to as SHAP), which attempts to explain model decisions by calculating feature contributions. The methods unified by SHAP are model agnostic and grounded in game theory, leading to more consistent and robust explanations. Instead of compelling the user to analyze thousands of feature maps, SHAP produces a single integrated relevancy-based heatmap that explains a prediction, with results that are output specific and aligned with human intuition, while addressing the previously reported saturation problem and thresholding artifact [46]. With the SHAP values, we can thus explain the contribution of a given geometric feature (represented by its pixels) of a nanophotonic structure to the structure's electromagnetic response at each wavelength. SHAP values are calculated through the following equation:

$$\Phi_i \; (f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \backslash i) \right], \tag{1}$$

where, Φ i is the SHAP value, x' are simplified inputs that mapped binary values into the original input space (x), M is the number of simplified input features, z' is a subset of non-zero indices in x', $f_x(z')$ is a model trained with the feature present, and $f_x(z' \backslash i)$ is a model trained with the feature withheld24. The SHAP algorithm captures the effect of withholding a feature, then iterates the computation across all possible subsets $(z' \subseteq x')$. In general, by removing specific features and calculating the change in the output, the algorithm can determine the contribution of these features (positive or negative) towards a specific prediction (*i.e.*, if the change of the output is large, the feature has a large contribution, and vice versa).

We performed SHAP explanations on the CNN model trained on 10,000 images of nanophotonic structures previously described. The explanation is represented as a heatmap, where red pixels represent positive contributions of a base image towards the model's prediction, and blue pixels represent negative contributions. As shown in Figure 3a, SHAP explanation heatmaps were captured at 6.0, 6.4, 6.8, 7.2, 7.6, 8.0, 8.4, and 8.8 μm with single-reference backgrounds (described in the Supporting Information), while the base image possessed a Lorentzian absorption peak at 5.2 μm and arm lengths of 1.4 μm. These explanations reveal the features, or lack thereof, that the CNN deems critical towards achieving an absorption resonance at the designated wavelengths. Specifically, as the resonance wavelength increases, the explanations show regions of blue pixels which gradually migrate from the center of the image to the edges, indicating that starting from the base image, the antenna arm lengths must become longer in order to achieve resonance at larger wavelengths. Conversely, Figure 3b shows that for a base image with longer initial arm lengths (2.9 μm), the arms must become shorter in order to achieve resonance at smaller wavelengths. This behavior is evident from the regions of blue pixels converging towards the center of the image as the resonance wavelength decreases. Both cross-arm tests indicate that the CNN has effectively inferred the relationship between antenna arm length and resonance wavelength.

At the same time, we observe varying degrees of red and blue pixels throughout the explanation heatmaps. For example, on the 8.8 μm explanation with the 5.2 μm base image, there are higher-intensity blue pixels on the top and left arms of the cross, indicating that the CNN weighs each arm differently in determining the resonance wavelength, when in reality, all of the arms are equally important to achieving resonance at the designated wavelength. In addition, the magnitude of the blue pixels are greater towards the edges of the structure, while the remaining areas have red pixels scattered throughout. Both results can be attributed to the filters developed by the CNN during training. CNNs tend to develop edge detection filters, since non-edge patterns (*e.g.*, a patch of black pixels) do not typically provide sufficient information to differentiate discrete objects22. Therefore, our CNN was tasked with creating the minimum set of filters that captures the most important features and distinctions (*i.e.*, the cross-arm edges) required to correlate the images to their respective absorption spectra. Naturally, this determines the range of the CNN's feature recognition capabilities and the extent of which it can generalize (or accurately predict new and unknown images), which may be limited to an unknown degree. However, we can alleviate this uncertainty by using the SHAP explanations to identify the sections of the structure that strongly contributed to resonance as well as the sections that contributed only weakly. With this information, we can infer what kind of relationships the CNN learned (or failed to learn), thereby allowing us to determine potential failure modes of the trained model. Thus, in addition to uncovering the physical relationships learned by the CNN, the presented CNN-explanation approach is also effective at determining the limitations and risks associated with a ML model trained on nanophotonic simulations by providing insight into the model's behavior.

To validate the SHAP explanations in their identification of features that contribute to resonance at specific wavelengths, we used the SHAP value heatmaps from Figure 3 to modify the base image, such that the resonance occurs at alternate wavelengths (Figure S2). We compared the SHAP explanations, and the design validations derived from them, against a standard antenna-based analytical relationship between the MIM resonator arm lengths and resonance wavelengths:

$$\lambda = \left(2n_{eff}\right)L + C. \tag{2}$$

Here, λ is the resonance wavelength, n_{eff} is the effective index of the transverse electric (TE) mode, L is the length of the resonator, and C is a correction phase term [25-27]. The comparison between the SHAP-generated and the FDTD simulated structures demonstrates that the information extracted by the CNN aligns strongly with the physical relationship established in Eqn. 2.

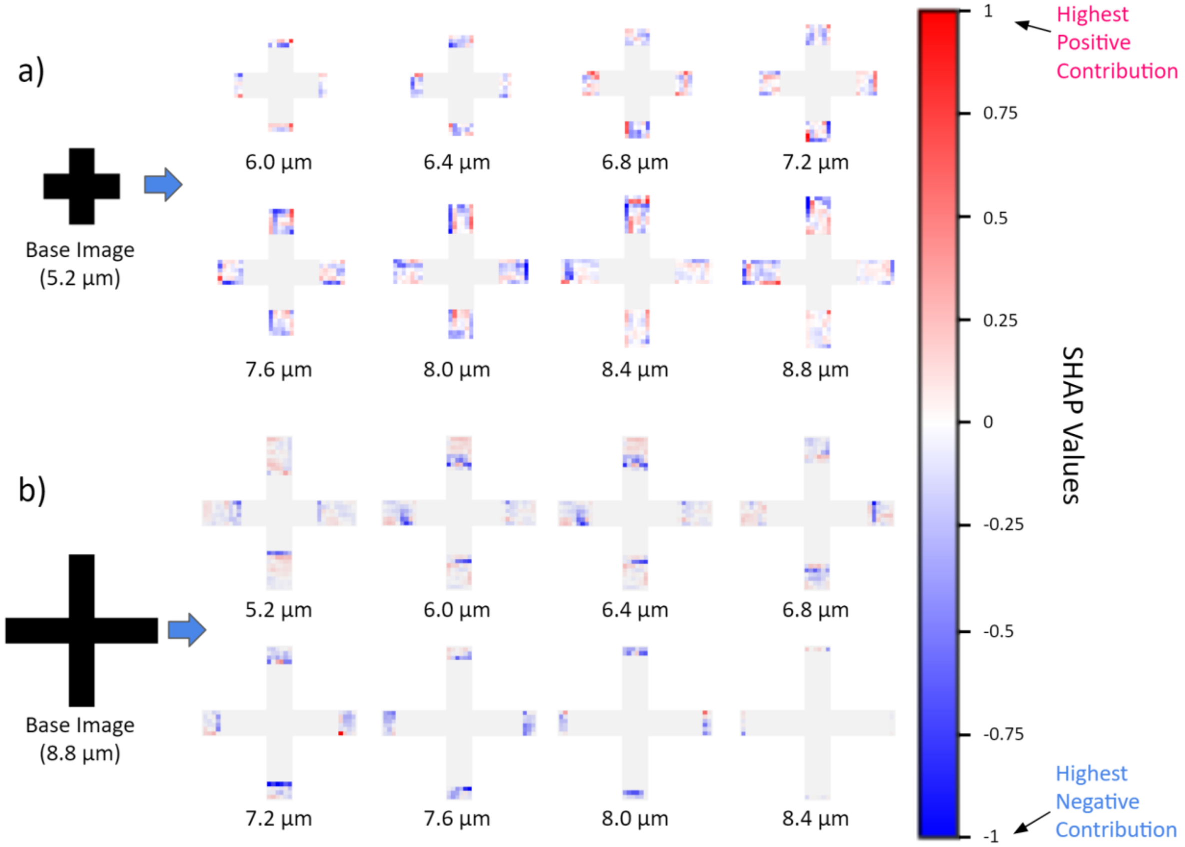


Figure 3. SHAP explanations for a (a) 'short-arm' cross (1.4 µm lengths) at increasing resonance wavelengths and a (b) 'long-arm' cross (2.9 µm lengths) at decreasing resonance wavelengths, revealing the CNN learned that the cross-arms must increase to achieve resonance at longer wavelengths and vice versa.

As an additional demonstration of the presented explanation method on complex shape-property relationships, we sought to explain the structural elements that distinguish a structure's response between single and double resonance behaviors within a given bandwidth. Figure 4 presents a series of test cases, where explanations of a dual absorption peak structure (L-shaped) and a single-peak structure (I-shaped) were captured at the peak wavelengths of each structure (marked in Figure 4a). The SHAP explanation heatmaps at the designated wavelengths are shown in Figure 4b, where the I-shaped image was used as the background for the L-shaped image and vice versa. The complete distribution of SHAP values from each heatmap are plotted and quantified in Figure 4c and 4d for the I-shaped resonator and L-shaped resonator, respectively. The inset bar graphs present the average SHAP values across each explanation. From these plots, we observe that at the peak/target wavelengths of the background image, the explanation of the base image at those wavelengths (indicated by the red-dashed boxes in Figure 4b) yield higher-magnitude and more negative SHAP values (blue pixels) than the explanations at non-peak wavelengths. Thus, the results here reveal that the CNN uses the inclusion of the horizontal-bar on the I-shaped structure to determine the presence of two absorption peaks at 5.5 µm and 8.2 µm, while the removal of the bar on the L-shaped structure renders a single peak at 6.3 µm. Using the same design validation strategy from the previous section, we confirmed that the SHAP explanations correctly identified the structural areas that contribute to single and dual resonance (Figure S4).

Furthermore, the explanations provide more granular details on which areas of each nanophotonic structure contributes to each resonance peak. For example, for the dual-peak L-shaped structure, the explanation at each peak (5.5 µm and 8.2 µm) illustrates different red-pixel dominant regions (features contributing to resonance at these wavelengths). We note that the heatmap's spatial distribution bears resemblance to the spatial nature of the resonances on either peak. In particular, the electric field concentrations in these structures vary at different resonance wavelengths (Figure S3). Similar to the electric field of the L-shaped structure at 8.2 µm, SHAP informs us that roughly the entire horizontal-arm length evenly contributes to resonance, while at 5.5 µm, the center of the arm contributes more to the resonance, which aligns with the nature of the E-field distribution for this resonance (Figure S3).

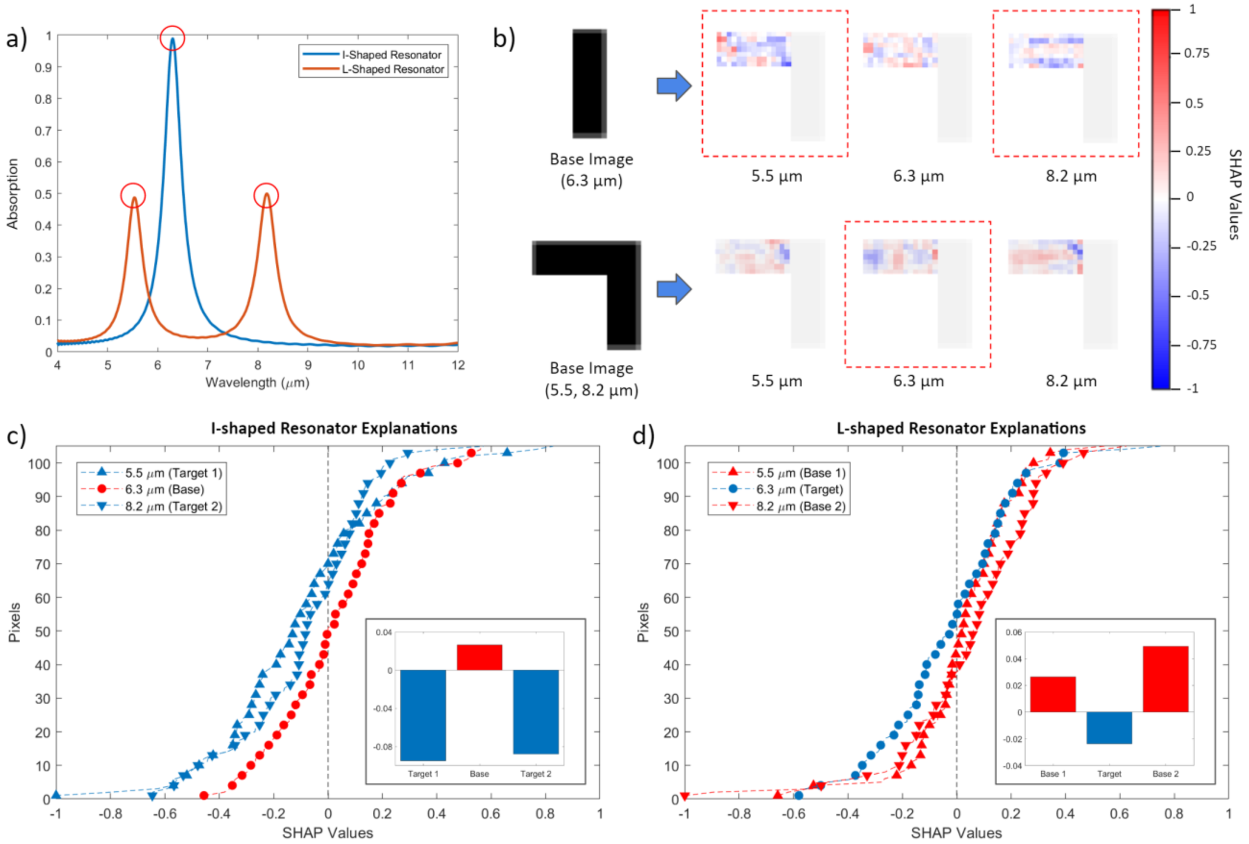


Figure 4. (a) Absorption spectra of a single-peak I-shaped resonator and a dual-peak L-shaped resonator. Red circles indicate the resonance wavelengths. (b) SHAP explanations of the resonators at the previously identified resonance wavelengths. Red dashed boxes indicate the explanations for obtaining new target resonance wavelengths of the opposing shape. Distribution of SHAP values across the explanation pixel-maps for the (c) I-shaped resonator and the (d) L-shaped resonator. Inset bar graphs represent the average SHAP values of each explanation, where the negative SHAP values (blue pixels) are dominant on all target explanations.

Explaining the Response of Complex Freeform Structures

We next investigate, using SHAP's explanations, the physical insight into the optical response of complex freeform resonator geometries that do not lend themselves to intuitive or previously understood physical models. Figure 5 shows a series of complex freeform metal-insulator-metal metamaterials we examine, as well as the multiple absorption peaks each structure supports. We use SHAP to generate explanations for wavelengths associated with each structure's peaks. In Figure 5, the blue pixels indicate the negative contributions of empty space towards the absorption peak. Therefore, the blue pixels indicate regions of the structure that positively contribute towards the absorption value at a particular wavelengths, while the red pixels indicate the regions that negatively contribute. As seen in Figure 5, the SHAP heatmaps highlight the spatial regions of each freeform shape most responsible for the noted absorption peaks. For example, in Figure 5a, the bottom and left edges of the structure appear to contribute the most to the absorption peak at 7.8 μm, while the top portion contributes to the peak at 8.9 μm. Similarly, in Figure 5c, we observe that the concave part of the structure on the upper-left strongly contributes to resonance at 7 μm, while the same region is not responsible for, and in fact suppresses, the resonance at 4.7 μm. In the case of Figure 5c then, this suggests that both spatial regions are responsible for the combined two peaks observed. Collectively, the negatively and positively contributing regions explain the overall absorption response that is observed at different wavelengths, and establishes a physical picture of the behavior of each freeform shape.

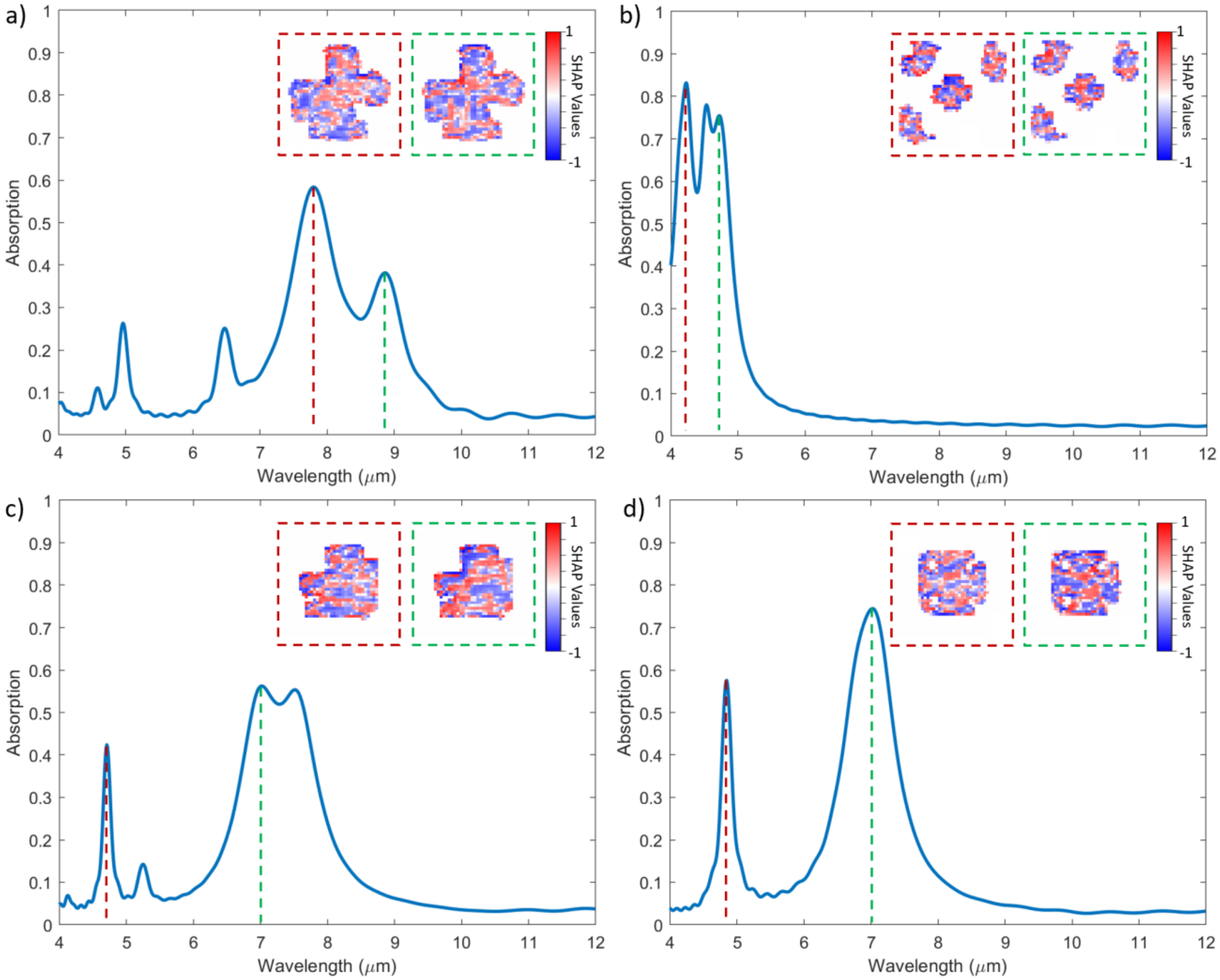


Figure 5. (a-d) SHAP explanations for freeform structures at various points of interest, revealing the range of structural elements which contribute to various absorption peaks.

Validating the Freeform Structure Explanations

To validate SHAP's explanations of which regions of freeform shapes contribute positively or negatively to the absorption value at a particular wavelength, we modify the original freeform structures based on these contributions, and simulate the modified structures in Lumerical. If SHAP's explanations are accurate, we expect the modified structures to enhance or suppress absorption at a specified wavelength based on the SHAP value for a given spatial region. Focusing the analysis on one resonance wavelength per design, Figure 6a illustrates the explanation of the freeform structure for the absorption peak at 7.8 μm, while 6b, 6c, and 6d show the explanations for their corresponding structures at 4.2 μm, at 7 μm, and 7.1 μm, respectively. After validating the SHAP values, we found that the absorption amplitude of the original structure can be tuned by selectively adding or removing the structural elements informed by SHAP. For example, in Figure 6a, the absorption spectra peaks at 0.6. After generating a structure using primarily blue pixels (shown below the SHAP heatmap and to the right with a blue border), the absorption rose to 0.9. Conversely, using primarily red pixels (which negatively contribute to the absorption value at that wavelength) the generated structure (shown below the SHAP heatmap and to the left with a red border) yields an absorption value near 0.1. We emphasize that this validation step is not necessarily a design strategy for photonic structures. Instead, it serves to confirm that the SHAP values at different spatial regions of the freeform structure do in fact correspond to its ability to enhance or suppress absorption at a particular wavelength.

We note that in deciding which pixels to transform, we used only the largest 95% of the absolute SHAP values to account for noise. We also observe that this absorption enhancement and suppression strategy is consistent across multiple designs (Figure 6b-d), although the degree of absorption intensity change varies by design, and the resonance wavelengths of the new structures deviate slightly from the original structure. However, from the SHAP heatmaps, we are able to describe the spatial components of an arbitrary structure to discover the regions responsible for absorption at a specific wavelength, and use this information to tune the properties of the metasurface. Thus, the presented XAI method provides an explanation of the behavior of complex structures, where the relationships between structure and property are not readily apparent, and opens the door to new strategies for nanophotonics design.

While promising as a first demonstration, we note some limitations in our current results which are linked to an important current limitation in the Deep SHAP method: its inability to account for feature dependence [29,30]. This in turn could have inhibited the identification of key structural features required for a resonance. Despite the minor discrepancies between the target and resulting resonance wavelengths in our validation studies, the general patterns identified by the explanations still offer significant insights into the features which contribute to resonance; a critical element which was not accessible in previous ML studies pertaining to photonic structures.

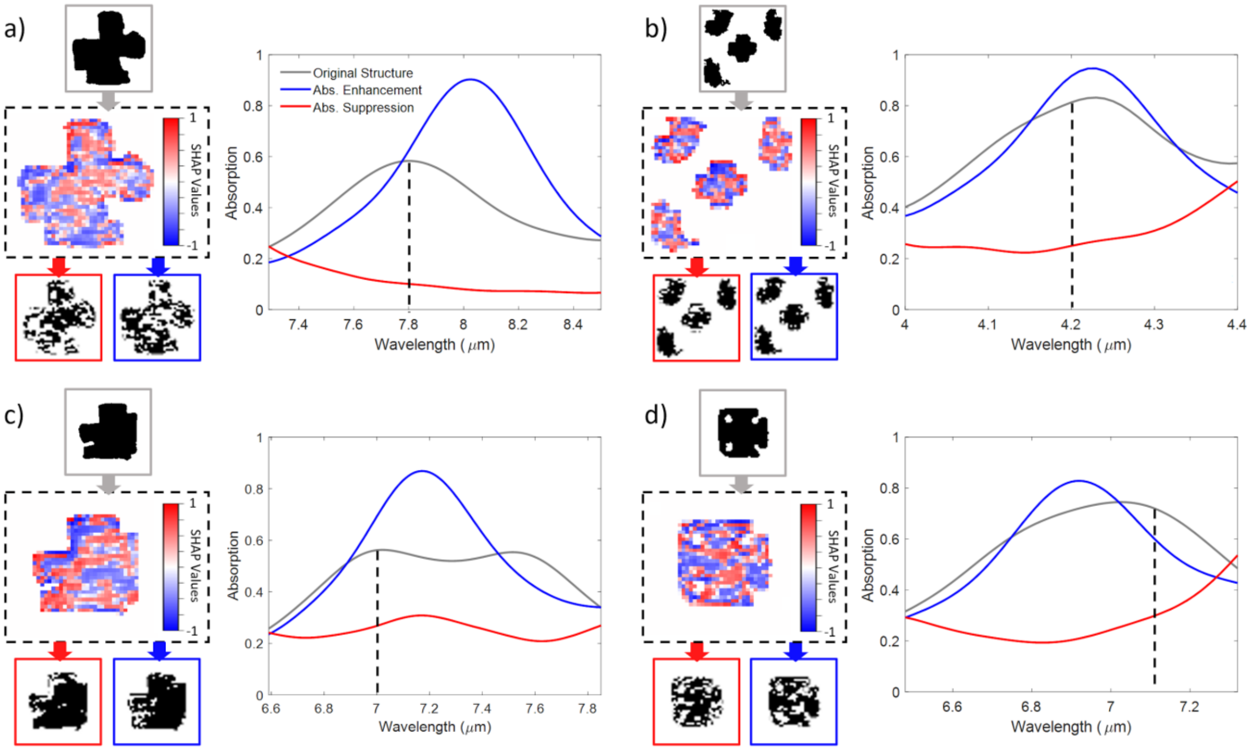


Figure 6. Validating the feature contributions highlighted by SHAP. Modifications were made to the freeform structures based on their SHAP-determined feature contributions at (a) 7.8 μm, (b) 4.2 μm, (c) 7 μm, and (d) 7.1 μm. For each structure, the SHAP explanations at the corresponding absorption peaks are shown. The SHAP-determined regions of positive (red pixels) and negative contributions (blue pixels) were used to generate structures which were then simulated using a full-field electromagnetic solver and resulted in absorption peak suppression or enhancement, respectively.

4.3 Conclusions

In summary, we show that convolutional neural networks can predict the optical properties of nanophotonic structures with remarkable precision, serving as an ultra-fast electromagnetic simulator for constrained domains of structures that also contain valuable information about the behavior of nanophotonic structures. Accordingly, we demonstrated an explanation algorithm (Deep Shapley Additive Explanations, or SHAP) that identifies the contributions of individual image features (on a pixel-by-pixel level) towards each of the network's predictions. The trained CNN predicted the spectra of new and unknown structures with over 95% accuracy, and orders of magnitude faster (~0.3 seconds) than conventional simulation (~30 minutes, yielding a 6,500x improvement). By examining the SHAP explanations, both qualitative and quantitative relationships between structure and spectra can be obtained (*i.e.*, resonator arm length vs resonance wavelength), and the explanations themselves can be used to enable unique design strategies through ML-inspired physics discovery. The explanations also revealed what the CNN did *not* learn, thus exposing potential limitations and risks associated with the trained model. Importantly, the presented explainable artificial intelligence method shows that the patterns and principles encoded within the ML model can be extracted to derive valuable insights into the nanophotonic structure behavior, even in complex freeform structures whose behavior is typically not easy to understand. While we chose to study a class of metamaterial resonators and their corresponding absorption spectra, we emphasize that the approach we have developed is applicable to any class of photonic structure or device for which a sufficiently large training dataset can be assembled by simulations, and any relevant optical device property, including focal depth, field of view and polarization sensitivity. Future studies could thus encompass using emerging explainability algorithms along with the explanation of additional device-property relationships. In the long term, combining explainability with machine learning may enable new discoveries in the physics of highly complex nanophotonic structures, and in turn yield new functionalities and capabilities not possible today.

4.4 Supporting Information

CNN Training Data and Network Architecture Optimization

To train our CNN, we performed three-dimensional finite-difference time domain (FDTD) simulations of 10,000 unique nanophotonic structures in Lumerical, and generated 10,000 two-dimensional images of the resonator layers and their corresponding absorption spectra. Figure S1 shows example 2D images of the designs used in the training dataset. The designs consist of cross-shapes, box-shapes (hollow and solid), cross-shapes with perpendicular resonators along the arm tips, and the inverted versions of each shape. To simulate the absorption spectra, these designs were simulated within 3.2 μm × 3.2 μm periodic arrays. Each structure contained a 100 nm gold bottom layer, a 200 nm Al_2O_3 dielectric middle layer, and a 100 nm gold top resonator layer with various dimensions. Mesh sizes of 20 nm × 20 nm × 20 nm were maintained across the entire simulation domain, and a plane-wave source at normal incidence was applied across a wavelength window of 4 μm to 12 μm.

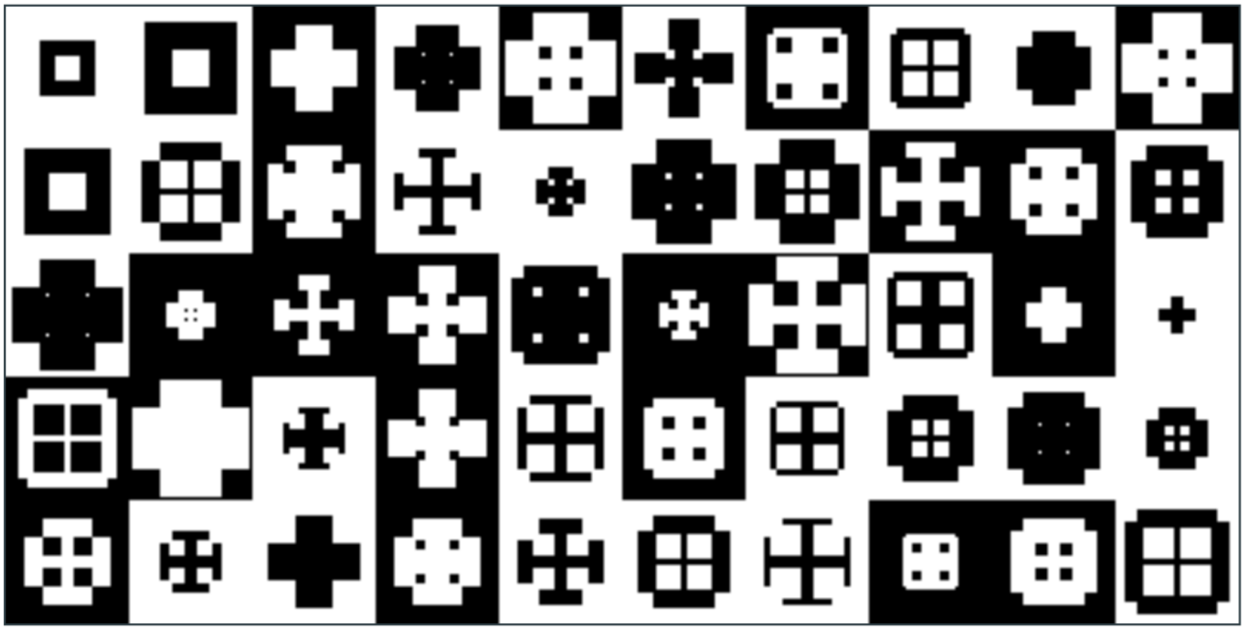


Figure S1. A subset of the training data images, consisting of: cross-shapes, box-shapes (hollow and solid), cross-shapes with perpendicular resonators along the arm tips, and the inverted versions of each shape.

The CNN was implemented using TensorFlow and Keras and trained on one Intel Core i5-8600T CPU for 300 epochs. Table S1 presents each of the trained models along with their validation root-mean-square error (RMSE) and training time. Model 1 served as the starting point, which consisted of three convolutional layer-stacks, each proceeding with a batch normalization layer, rectified linear unit (ReLU) activation layer, and average pooling layer (except the final stack). Each convolutional layer used 3 × 3 filters, numbering in 8, 16, and 32 in each subsequent layer. The pooling layer used 2 × 2 windows with a stride of 2. By testing incremental changes to the model (Model 2-8), we determined that a four-stack architecture with leaky ReLU layers trained with the adaptive moment estimation (Adam) algorithm yielded the lowest error (Model 9). In addition, the CNN was trained with a learning rate of 0.001, beta1 of 0.9, beta2 of 0.999, and test dataset of 10%.

Model 1			Model 2			Model 3		
Layers	Param.	Options	Layers	Param.	Options	Layers	Param.	Options
conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU	3x3,8 2x2, 2 3x3,16 2x2, 2 3x3,32	sgdm 256 minibatch 100 epochs	conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU	3x3,16 2x2, 2 3x3,32 2x2, 2 3x3,64	sgdm 256 minibatch 100 epochs	conv2d leakyReLU avgPool conv2d leakyReLU avgPool conv2d leakyReLU	3x3,16 2x2, 2 3x3,32 2x2, 2 3x3,64	sgdm 256 minibatch 100 epochs
RMSE	0.15313		RMSE	0.10648		RMSE	0.11762	
Time	63 min		Time	167 min		Time	218 min	
Model 4			**Model 5**			**Model 6**		
Layers	Param.	Options	Layers	Param.	Options	Layers	Param.	Options
conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU	3x3,8 2x2, 2 3x3,16 2x2, 2 3x3,32 2x2, 2 3x3,64 2x2, 2 3x3,128	sgdm 256 minibatch 100 epochs	conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU	3x3,8 2x2, 2 3x3,16 2x2, 2 3x3,32	adam 256 minibatch 100 epochs	conv2d ReLU maxPool conv2d ReLU maxPool conv2d ReLU	3x3,8 2x2, 2 3x3,16 2x2, 2 3x3,32	sgdm 256 minibatch 100 epochs
RMSE	0.13289		RMSE	0.11497		RMSE	0.16737	
Time	87 min		Time	77 min		Time	58 min	
Model 7			**Model 8**			**Model 9**		
Layers	Param.	Options	Layers	Param.	Options	Layers	Param.	Options
conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU	3x3,8 2x2, 2 3x3,16 2x2, 2 3x3,32	sgdm 256 minibatch 300 epochs	conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU	3x3,8 2x2, 2 3x3,16 2x2, 2 3x3,32	sgdm 128 minibatch 100 epochs	conv2d leakyReLU avgPool conv2d leakyReLU avgPool conv2d	3x3,16 2x2, 2 3x3,32 2x2, 2 3x3,64 2x2, 2	adam 128 minibatch 300 epochs

						leakyReLU avgPool conv2d leakyReLU	3x3,128	
RMSE	0.097562		RMSE	0.14086		RMSE	0.07709	
Time	229 min		Time	42 min		Time	340 min	

Table S1. Table of trained CNN architectures and the corresponding RMSE values.

SHAP Explanation Validation for Single-Resonance Structures

The DeepExplainer module from the Deep SHAP Python library was used to explain the predictions of the CNN. To generate SHAP values for deep learning models, the SHAP algorithm approximates the conditional expectations of SHAP values using a selection of background samples. The background dataset is used to determine the impact of a feature by replacing the feature with values from the background. In doing so, the algorithm can simulate 'missing' features and calculate the impact on the model output[1]. To minimize the noise that was generated by the SHAP explanations, we performed the SHAP explanations under single-reference background conditions. For the single-reference background, we used an image with a specific absorption peak as the background (*e.g.*, 8.0 μm), and captured the explanation at this peak wavelength. This process was repeated for all target wavelengths.

Following the generation of the explanations, the explanations were used for design validations towards target resonances. Validation was performed by converting the blue pixels in the heatmaps to black pixels on the base image. Figure S2b and S2c show the spectra of the validated structures and the original FDTD simulated structures, respectively. These validated designs were then compared with the corresponding FDTD simulated background images (as shown in Figure S2a) to ensure that the CNN learned the relationship between cross-arm length and resonance wavelength. In Figure S2d, the resonant wavelengths at peak absorption and the antenna arm lengths of both sets of structures are plotted (with linear fits of R^2=0.998). The FDTD-simulated structures yield an n_{eff} of 1.13 and C of 2.21, while the SHAP-validated structures display an n_{eff} of 1.15 and C of 2.10, yielding an n_{eff} error of 1.8% and a C error of 4.9%.

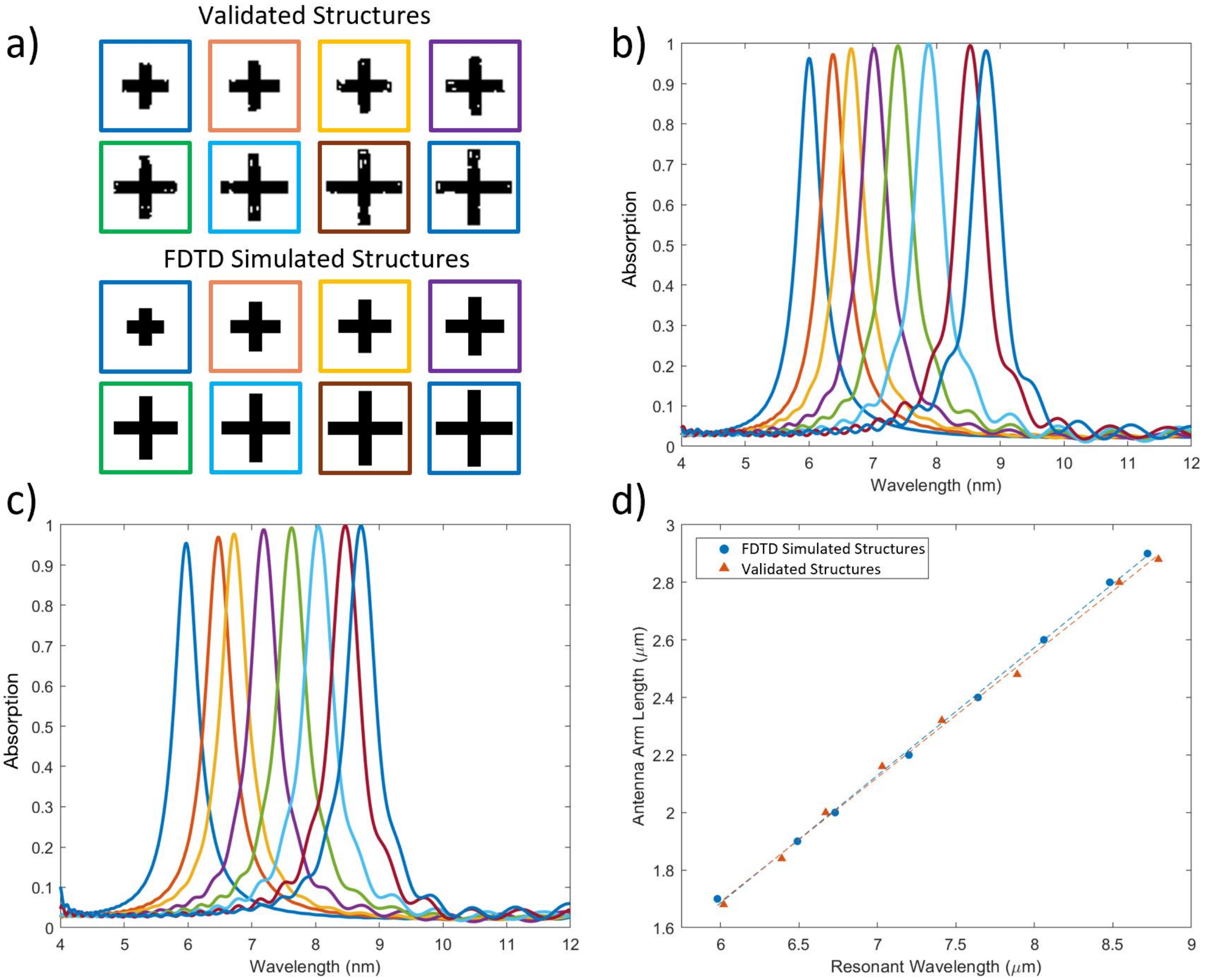


Figure S2. (a) Images of the SHAP-validated and FDTD simulated structures. The absorption spectra for the corresponding (b) validated and (c) simulated structures. Image-border colors correspond to the plot colors. (d) Comparison of the physical relationship between antenna arm length and resonant wavelength for the two sets of structures (linear fit of plots shown with R^2=0.998).

SHAP Explanation Validation for Multi-Resonance Structures

Figure S3 shows the electric field profiles of various MIM structure designs. We note that the SHAP explanation heatmap's spatial distribution bears resemblance to the spatial nature of the resonances on either peak. In particular, the electric field concentrations in these structures vary at different resonant wavelengths.

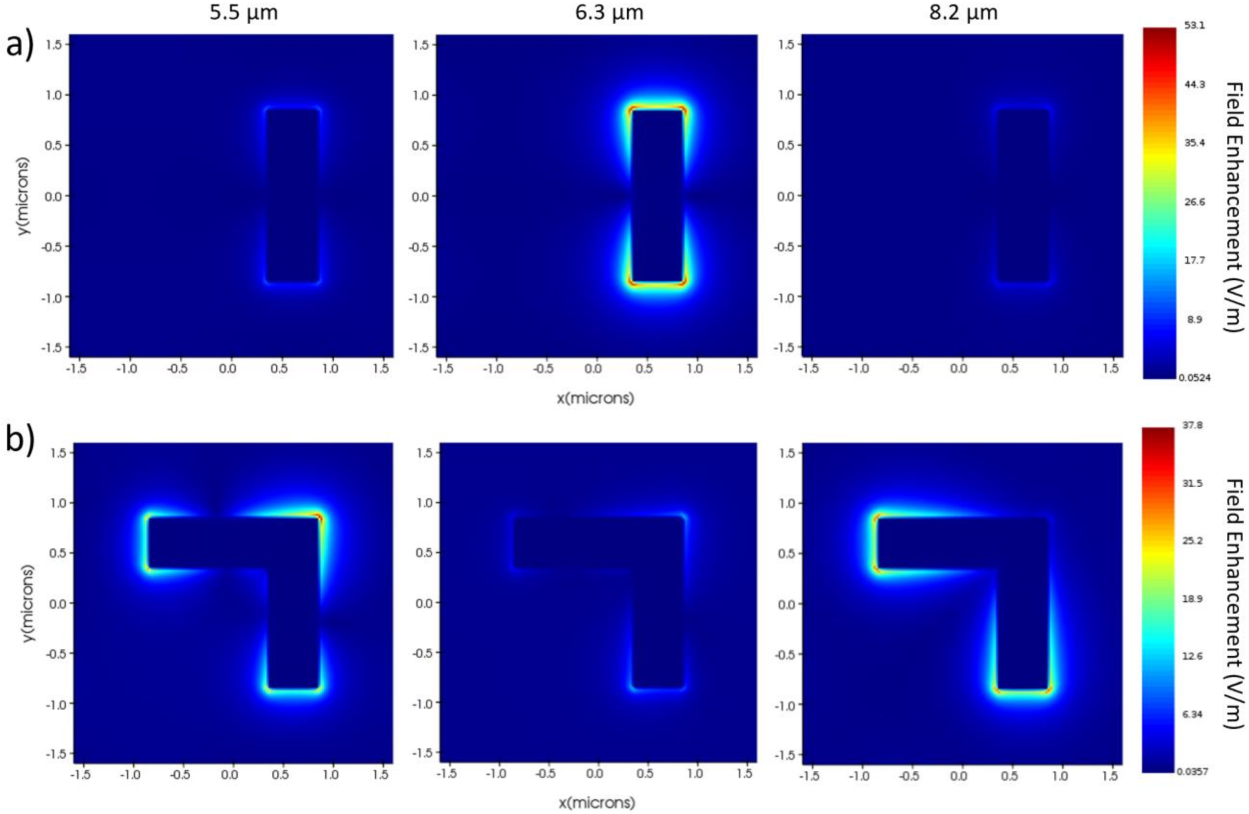


Figure S3. The electric field simulation profile of (a) an L-shaped resonator and (b) an I-shaped resonator at resonance wavelengths of 5.5 µm, 6.3 µm, and 8.2 µm.

We performed the same design validation method from the previous section to assess the accuracy of the SHAP explanations for complex, multi-resonance structures. In Figure S4a, the L-shaped structure was validated by utilizing the explanation generated at 6.3 µm, then converting all of the blue pixels to the opposite state on the original image. The resulting structure exhibited a single absorption peak of approximately 0.9 at 5.4 µm. Using the same approach, we attempted the reverse scenario of generating a dual-peak structure from a single-peak structure (Figure S4b). We leveraged the explanation from one of the dual-peak wavelengths (as either wavelength resulted in negligible differences) and applied it to the design validation process. The validated structure possessed an absorption peak of ~0.6 at 4.8 µm and ~0.48 at 6.9 µm. The design validation studies demonstrate that complex spectral targets can be met by converting the pixels identified by the SHAP heatmaps, and thus that the heatmaps themselves reveal useful information about the relationship between geometric features and their electromagnetic response. In the first case, by focusing the image conversion process on the explanations of a single target wavelength, we converted a dual-peak structure into a single-peak structure. In the second case, the single-peak structure was converted into a dual-peak structure by using the SHAP values of two target wavelengths.

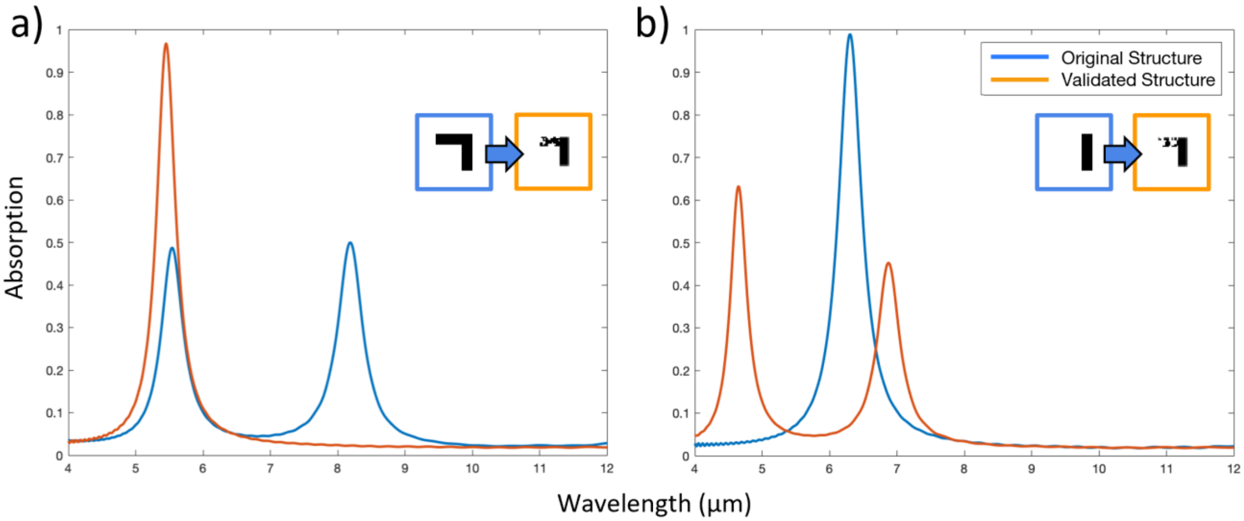


Figure S4. SHAP-validation for a (a) single-peak structure and a (b) dual-peak structure by utilizing the SHAP values at targeted resonance wavelengths for image conversion.

4.5 References

[1] Chen, R. et al. Nanophotonic integrated circuits from nanoresonators grown on silicon. Nat. Commun. 5, 4325 (2014).

[2] Pelton, M. Modified spontaneous emission in nanophotonic structures. Nature Photonics 9, 427-435 (2015).

[3] Odebo Länk, N., Verre, R., Johansson, P. & Käll, M. Large-Scale Silicon Nanophotonic Metasurfaces with Polarization Independent Near-Perfect Absorption. Nano Lett. 17, 3054-3060 (2017).

[4] Pralle, M. U. et al. Photonic crystal enhanced narrow-band infrared emitters. Appl. Phys. Lett. 81, 4685-4687 (2003).

[5] Lin, S. Y., Moreno, J. & Fleming, J. G. Three-dimensional photonic-crystal emitter for thermal photovoltaic power generation. Appl. Phys. Lett. 83, 380-382 (2003).

[6] Laroche, M., Carminati, R. & Greffet, J. J. Coherent thermal antenna using a photonic crystal slab. Phys. Rev. Lett. 96, 123903 (2006).

[7] Yao, K., et al. Intelligent nanophotonics: Merging photonics and artificial intelligence at the nanoscale. Nanophotonics 8, 339-366 (2019).

[8] Molesky, S. et al. Inverse design in nanophotonics. Nature Photonics 12, 659-670 (2018).

[9] Lin, Z., Groever, B., Capasso, F., Rodriguez, A. W. & Lončar, M. Topology-Optimized Multilayered Metaoptics. Phys. Rev. Appl. 9, 044030 (2018).

[10] Sell, D., Yang, J., Doshay, S. & Fan, J. A. Periodic Dielectric Metasurfaces with High-Efficiency, Multiwavelength Functionalities. Adv. Opt. Mater. 5, 1700645 (2017).

[11] Yang, J. & Fan, J. A. Topology-optimized metasurfaces: impact of initial geometric layout. Opt. Lett. 42, 3161-3164 (2017).

[12] Jensen, J. S. & Sigmund, O. Topology optimization for nano-photonics. Laser and Photonics Reviews 5, 308-321 (2011).

[13] Xiao, T. P. et al. Diffractive Spectral-Splitting Optical Element Designed by Adjoint-Based Electromagnetic Optimization and Fabricated by Femtosecond 3D Direct Laser Writing. ACS Photonics 3, 886-894 (2016).

[14] Burger, M., Osher, S. J. & Yablonovitch, E. Inverse Problem Techniques for the Design of Photonic Crystals. IEICE Transactions on Electronics 87, 258-263 (2004).

[15] 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).

[16] Jiang, J. et al. Free-form diffractive metagrating design based on generative adversarial networks. ACS Nano 13, 8873-8878 (2019).

[17] So, S. & Rho, J. Designing nanophotonic structures using conditional deep convolutional generative adversarial networks. Nanophotonics 8, 1255-1261 (2019).

[18] Liu, Z., Zhu, D., Rodrigues, S. P., Lee, K. & Cai, W. A Generative Model for Inverse Design of Metamaterials. Nano Lett. 18, 6570-6576 (2018).

[19] Olden, J. D. & Jackson, D. A. Illuminating the ‘black box’: A randomization approach for understanding variable contributions in artificial neural networks. Ecol. Modell. 154, 135-150 (2002).

[20] Qiu, F. & Jensen, J. R. Opening the black box of neural networks for remote sensing image classification. Int. J. Remote Sens. 25, 1749-1768 (2004).

[21] Han, S., Pool, J., Tran, J. & Dally, W. J. Learning both weights and connections for efficient neural networks. Advances in Neural Information Processing Systems (2015). arXiv:1506.02626v3.

[22] Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 84-90 (2017).

[23] Liu, X. et al. Taming the blackbody with infrared metamaterials as selective thermal emitters. Phys. Rev. Lett. 107, 045901 (2011).

[24] Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (2017). arXiv:1705.07874.

[25] Wada, K. & Kimerling, L. C. Photonics and electronics with germanium. Photonics and electronics with germanium (2015). doi:10.1002/9783527650200.

[26] Omeis, F. et al. Metal-insulator-metal antennas in the far-infrared range based on highly doped InAsSb. Appl. Phys. Lett. 111, 121108 (2017).

[27] Chen, K., Adato, R. & Altug, H. Dual-band perfect absorber for multispectral plasmon-enhanced infrared spectroscopy. ACS Nano 6, 7998-8006 (2012).

[28] Brachmann, A. & Redies, C. Using convolutional neural network filters to measure left-right mirror symmetry in images. Symmetry (Basel). 8, 1-10 (2016).

[29] Molnar, C. 5.10 SHAP (SHapley Additive exPlanations) | Interpretable Machine Learning. Christophm.github.io (2020). at <https://christophm.github.io/interpretable-ml-book/shap.html#kernelshap>

[30] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. arXiv e-prints (2019). arXiv:1802.03888.

[31] Phan, T. et al. High-efficiency, large-area, topology-optimized metasurfaces. Light Sci. Appl. 8, 1-9 (2019).

[32] Bayati, E. et al. Inverse Designed Metalenses with Extended Depth of Focus. ACS Photonics 7, 873-878 (2020).

[33] Sajedian, I., Kim, J. & Rho, J. Finding the optical properties of plasmonic structures by image processing using a combination of convolutional neural networks and recurrent neural networks. Microsystems Nanoeng. 5, 1-8 (2019).

[34] Wiecha, P. R. & Muskens, O. L. Deep Learning Meets Nanophotonics: A Generalized Accurate Predictor for near Fields and Far Fields of Arbitrary 3D Nanostructures. Nano Lett. 20, 329-338 (2019).

[35] Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should i trust you?' Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144 (2016).

[36] Roscher, R., Bohn, B., Duarte, M. F. & Garcke, J. Explainable Machine Learning for Scientific Insights and Discoveries. IEEE Access 8, 42200 (2020).

[37] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García,S., Gil-López, S., Molina, D., Benjamins, R., et al. (2019). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82-115 (2019).

[38] Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. CoRR (2014). arXiv:1312.6034

[39] Montavon, G., Lapuschkin, S., Binder, A., Samek, W. & Müller, K. R. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit. 65, 211-222 (2017).

[40] Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. Comput. Vision–ECCV 2014, 818-833 (2014).

[41] Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One (2015). doi:10.1371/journal.pone.0130140.

[42] Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat. Biomed. Eng. 2, 749-760 (2018).

[43] Ghosal, S. et al. An explainable deep machine vision framework for plant stress phenotyping. Proc. Natl. Acad. Sci. U. S. A. 115, 4613-4618 (2018).

[44] Groth, O., Fuchs, F. B., Posner, I. & Vedaldi, A. ShapeStacks: Learning Vision-Based Physical Intuition for Generalised Object Stacking. Comput. Vision–ECCV 2018, 724-739 (2018).

[45] L. von Rueden, S. Mayer, and et al. Informed machine learning – a taxonomy and survey of integrating knowledge into learning systems. arXiv preprint (2020). arXiv:1903.12394v2.

[46] Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. 34th International Conference on Machine Learning, ICML 70, 3145-3153 (2017).

5. Enhancing Adjoint Optimization-based Photonics Inverse Design with Explainable Machine Learning

5.1 Introduction

Effectively optimizing nanophotonic structures is key to their use in a broad range of optical applications. For example, photonic integrated circuits, metasurfaces, and guided-wave systems can be geometrically manipulated at subwavelength scales to deliver a wide range of functionalities [1-5]. However, a large design space must be rapidly explored in order to optimize the geometry for a particular application. To effectively navigate such a design space, gradient-based optimization algorithms such as the adjoint variables method have been widely adopted to design non-intuitive or irregularly-shaped electromagnetic structures that are highly efficient at accomplishing a particular goal. By calculating the shape derivatives at all points in space using only two electromagnetic simulations per iteration [6], adjoint optimizations are orders of magnitude more computationally efficient than alternative optimization methods and capable of achieving state-of-the-art performance [6-9].

Although adjoint optimization-based methods have been successfully applied to a variety of photonic systems [10-13], the method's reliance on gradient-based information means that the method is local in nature, and therefore bounded by the corresponding limitations. Specifically, since the design space for electromagnetic structures is predominantly non-convex, adjoint optimizations (or gradient-based optimization algorithms in general) are susceptible to getting stuck in local minima valleys or saddle points (hereon collectively referred to as local minima) [14,35]. Thus, unless a region of high-performance devices is known in advance, multiple optimization runs are needed (typically by using random starting points) to arrive at a single optimization target [15]. To overcome these limitations, recent efforts have combined machine learning (ML) with adjoint optimization. For example, population-based inverse design was demonstrated using global topology-optimization networks, or GLOnets [16], which integrate the adjoint method directly into the training process. Alternative strategies also include the integration of ML and adjoint optimization as a two-step process, where the ML-component performs an initial global-search approximation, then the optimization improves design performance further [16-18,41]. Although both approaches can improve upon the algorithm's performance, the underlying issue of local minima trapping remains unaddressed, since the integration and use of a gradient-based optimizer inherently indicates that the issue is still present. In this regard, metaheuristic techniques such as simulated annealing have been proposed to escape local minima in the search process [19,20]. This method may directly address the issue of local minima trapping through neighbor-based exploration, but its applications in photonics shape and topology optimization have been severely limited due to relatively low computational efficiency (on the order of 1,000 iterations) [21,22].

To comprehensively address the issue of local minima trapping, we seek to identify the root of the problem and ask: what caused the algorithm to get trapped in the first place? In the context of the optical structures being optimized, arriving at certain geometric elements and their resulting electromagnetic response must be responsible for (or contribute to) guiding the optimization towards suboptimal results. To discover the geometric features responsible for local minima trapping, and to then overcome them, we employ an explainable artificial intelligence (XAI) based approach. XAI serves as a promising candidate for addressing local minima trapping due to its well-known ability to reveal a model's decision-making process as well as the contributing factors thereof (*i.e.*, addressing the black box problem) [23-25]. For example, XAI can reveal the spatial regions of a nanophotonic structure that contribute to the presence or lack of an absorption peak [26]. Thus, to explain the causes of local minima trapping in gradient-based adjoint optimization, and subsequently use this information to prevent the optimization from converging onto suboptimal states, we present an XAI-based framework that utilizes the relationships between device efficiency and nanoscale structuring to increase optimization performance.

5.2 Results and Discussion

We demonstrate our optimization framework in the context of Y-splitter waveguide design (Figure 1a), where the objective is to optimize the shape of the silicon-oxide interface to maximize power transmission efficiency from an input port to two output ports of the same width. We represent the objective function as a decreasing figure-of-merit (FOM) which ranges from 1 to 0, where 0 represents ideal performance. With this definition in place, adjoint shape optimizations are performed on an initial Y-splitter design to minimize the FOM (Figure 1b) across a range of target wavelengths in the telecommunications range (1.3 to 1.8 μm). The design and FOM information from the optimizations are used in conjunction with neural architecture search to automate the training of an ML model (AutoML). The model learns the relationships between device structure and performance by accurately predicting the FOM and target wavelength of an input design (Figure 1c). We then use a suite of XAI algorithms, SHapley Additive exPlanations or SHAP (a post hoc explanation technique based on game theory [27]), on the model to extract the structure-performance relationships as “feature explanation” heatmaps (Figure 1d). By interrogating our trained ML model, the explanations here inform the structural features that contribute to the FOM of interest. Using this information, we devise a boundary extraction algorithm that takes the explanations and makes guided design changes (Figure 1e). These design changes then provide a new starting point for the local adjoint optimization method, which allows the method to reach lower FOMs than before (at multiple target wavelengths).

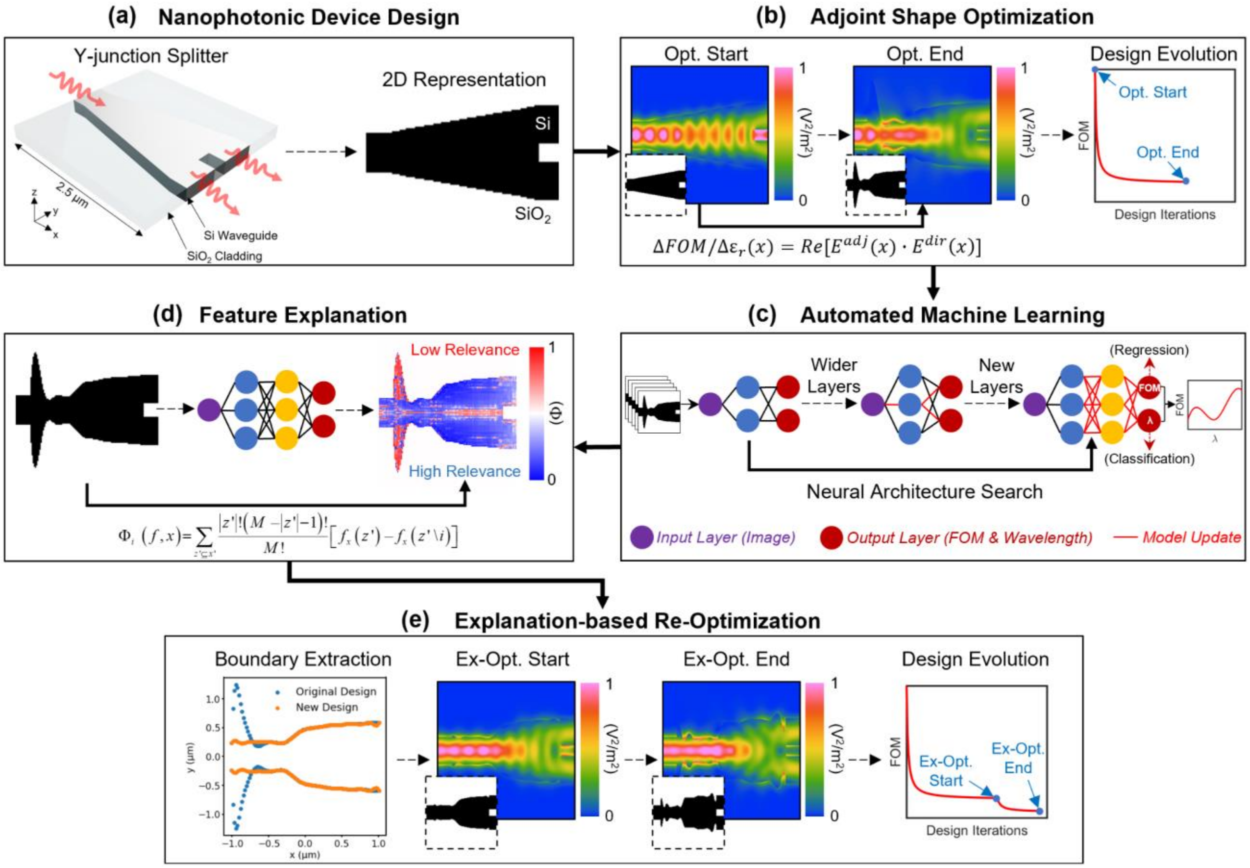


Figure 1. (a) Nanophotonic device optimization: silicon-on-insulator Y-junction splitter for telecom applications. (b) Multiple adjoint optimization runs are applied to the Y-splitter design at various target wavelengths. (c) Results of the optimizations are used as training data in automated machine learning (AutoML) to train a neural network, where the inputs are images and the outputs are device figure-of-merits (FOM) and target wavelengths. (d) Explainable AI algorithms are used on the neural network to capture feature explanations, (e) which are used to optimize device performance further by allowing the algorithm to escape its local minima.

Adjoint Optimization and Convolutional Neural Network Training

We first developed our training dataset by performing multiple adjoint shape optimization runs on a starting Y-splitter design (Figure 2a). We applied a widely-adopted implementation of the adjoint method that is integrated with a commercial finite-difference time-domain (FDTD) solver [28]. The 2D cross-sections of the Y-splitter designs are represented as black and white images, where the black and white pixels represent the permittivity of silicon and SiO_2, respectively. In our configuration of the adjoint method, the optimizable region is the area between the input and output ports, while the port sizes remain fixed. The optimizable geometry within this region is defined using the level set method and cubic spline interpolations [28,29]. Each optimization run was performed on randomized starting designs (waveguide structures with 25%, 35%, 40%, and 50% fill fractions; collectively shown in Figure S2) using different operating wavelengths as optimization objectives (1.3 to 1.8 µm in 0.1 µm steps) to produce a collection of device designs with gradual performance improvements. Performance improvement is indicated by a decreasing FOM (as design iteration increases), until a plateau is reached. As shown in Figure 2a (here, the 35% fill fraction starting design), each design iteration consists of a forward and adjoint (*i.e.*, time-reversed) simulation, which calculates the shape derivative over the entire optimizable region and modifies the geometry (per iteration) in proportion to the FOM gradient [6]. At the final iteration "N" (which may vary across each optimization run), device geometry is tailored to achieve maximum attainable performance with respect to the sought target. In our application of the adjoint method, the FOM represents the power coupling of guided modes, and is defined as:

$$FOM = \frac{1}{P(\lambda)}\int |T_0(\lambda) - T(\lambda)|\, d\lambda, \tag{1}$$

where λ is the evaluated wavelength, T is the actual power transmission through the output ports, T_0 is the ideal power transmission, and P is the source power (in Watts). Thus, the FOM is the difference between the input and output transmission normalized by the power injected by the source, which results in maximum performance at 0. Figure 2b shows the results of each optimization run, where the collective FOM information and corresponding designs (at each iteration; not including the starting design) are used as training data for deep learning.

From the optimized structures shown in Figure 2, unique geometries are obtained for each target wavelength, which in turn yield a range of FOM values. Therefore, the FOM and target wavelength are coupled with one another, and both are dependent on the waveguide structure. Thus, to ensure that our model simultaneously learns both of these properties, which in turn captures more information regarding the structure than models trained on the properties individually, we designed a single neural network that takes the Y-splitter geometries as inputs (here, 128×64 pixel images, or 2.5×1.25 μm^2 domains) and outputs both FOM and target wavelength. In the particular design space we explored, over 600 input and output pairs were generated for the neural network. For ease of training, target wavelengths were converted into categorical labels, where a position-specific output node value of 1 represents the wavelength of a specific design, while the other positional nodes equal 0. For example, a target wavelength (T_λ) of 1.3 µm is represented as $T_{1.3} = [1,0,0,0,0,0]$, 1.4 µm is $T_{1.4} = [0,1,0,0,0,0]$, and this pattern is repeated up to 1.8 µm. Alternatively, *argmax*(T_λ), or the index of the maximum value along the vector, represents the target wavelength. Combined with a floating-point value (ranging from 0 to 1) to serve as the FOM, we devise a training data structure that is amenable to both classification and regression-based tasks. With this input-output relationship defined, as well as a 90:10 training-validation data split, we use neural architecture search (AutoKeras [30]) with image blocks to automate the deep learning process by testing different model variants across multiple trials. We observe that the optimal architecture was identified after 12 trials, which had a validation loss of

9.1×10^{-5}. The final training and validation losses of each trial are presented in Figure S1a, and the evolution of the convolutional neural network (CNN) architecture from the first trial to the last is shown in Figure S1b. The optimized CNN possesses five convolutional blocks (with 512, 256, 128, 64, and 32 filters, respectively) followed by a dense layer. Each block contains Leaky ReLu, batch normalization, and max pooling layers. Training progression of the optimized architecture is shown in Figure 2c, where a strong convergence between the training and validation losses can be observed. We further verified our model's performance through cross validation and overfitting analyses (found in Table S1, S2 and Figures S12, S13 of the Supporting Information).

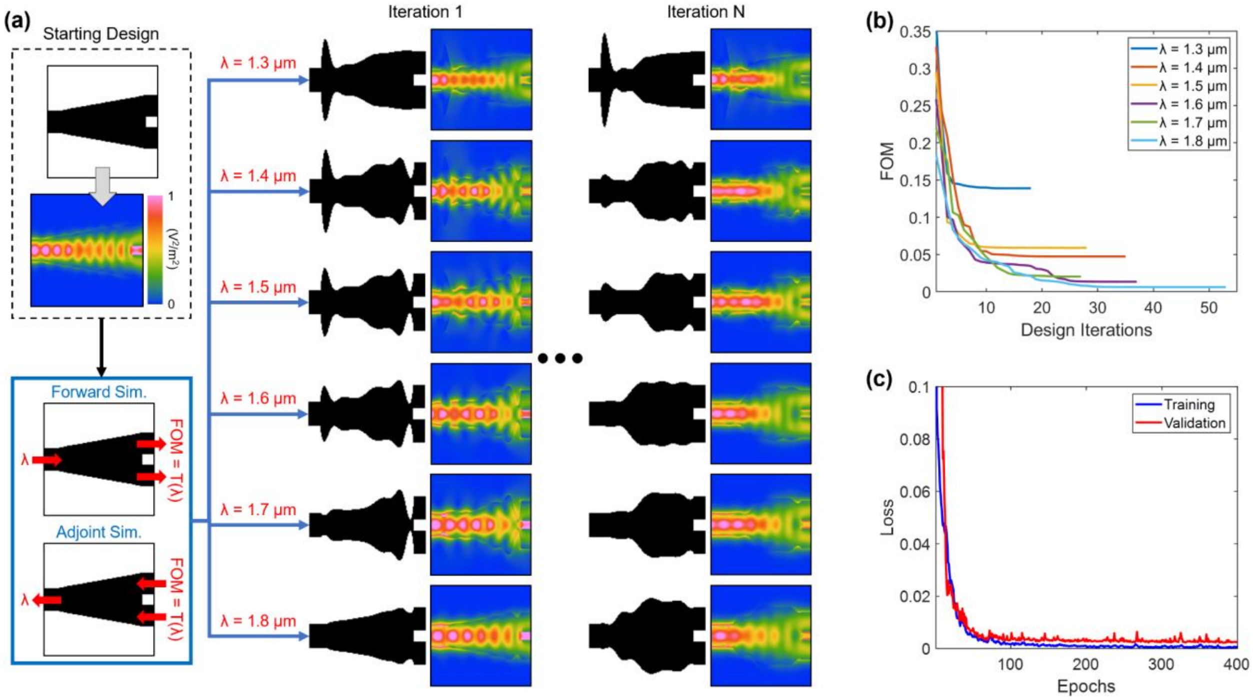


Figure 2. (a) Training data generation. Adjoint optimization runs are performed on random starting designs (35% fill fraction starting design shown) across target wavelengths ranging from 1.3 μm to 1.8 μm to produce high-performance devices in the telecom window. (b) FOM vs. design iterations across each optimization run. (c) Training and validation losses for the AutoML-optimized neural network shows high training accuracy and model convergence.

Structure Explanation and Re-Optimization

After training our machine learning model, we next sought to explain the relationship between the overall shape and FOM such that this information can be leveraged to potentially further optimize the devices, and overcome any local minima the adjoint method may have arrived at. To verify that the model properly learned the structure-FOM relationship, we passed the final design iterations (of each target wavelength) into the trained model and compared the ground truth outputs to the model's predictions. The comparison is shown in Figure 3a, where we observe a strong match (over 90% accuracy) between the predictions (blue points) and ground truths (orange points). The inset images in Figure 3a are model inputs. From this result, we can infer that the model accurately learned the key features on the optimized structures which contribute to the target wavelength-specific FOM values. Therefore, we can utilize XAI to reveal the structure-performance relationships of each device. Specifically, we employed an explanation strategy for photonics design – using SHAP – to highlight the device feature contributions to their respective FOM [26]. These feature contribution heatmaps (represented as SHAP values, $\Phi(x,y)$, ranging from -1 to 1) are illustrated in Figure 3b, where the blue and red pixels indicate positive and negative contributions towards the FOM, respectively. We note that this is the reverse of conventional SHAP definitions due to our desired FOM being minimized. We then leverage the information captured by the XAI algorithm, and manipulate the structure accordingly, to assess its effect on device performance.

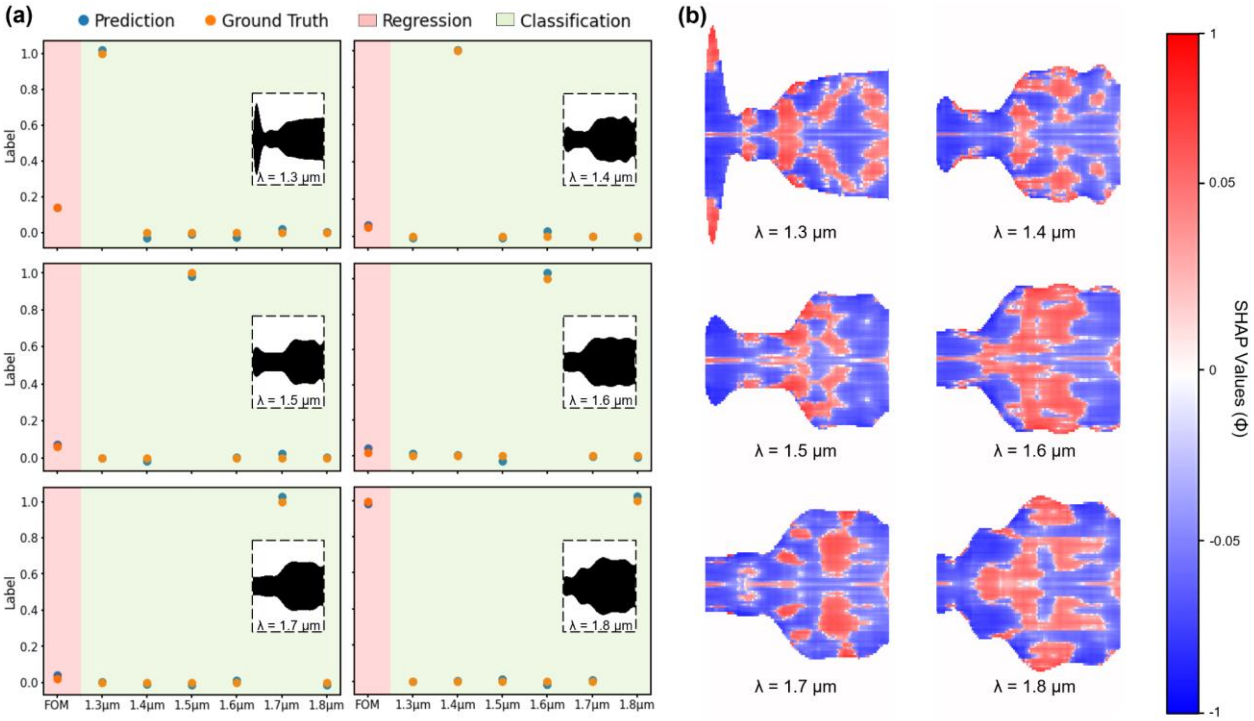


Figure 3. (a) Comparisons between model predictions and ground truths, for FOM (regression) and target wavelength (classification) values, show that the model accurately learned the relationship between device structure and performance. Inset images show the model inputs, which are adjoint-optimized devices. (b) SHAP explanation heatmaps of the optimized devices reveal the structural features that contribute positively (blue) or negatively (red) towards optimal device performance. Note that this is the reverse of conventional SHAP definitions due to our desired FOM being minimized.

To determine how to practically use the SHAP values (represented as red/blue heatmap pixels), we first note that high concentrations of blue pixels are located throughout a majority of each structure, while the center of the structures and select portions of the outer boundaries contain large regions of red pixels. In this regard, since the training data solely consists of geometries with varying degrees of shape changes at the SOI boundary, and no geometry change is introduced within the structure (*i.e.*, no material subtractions or white pixels are inside the main island of black pixels), we focus our analysis on the SHAP values located at the structure boundary rather than the center. Following the aforementioned principles of positive and negative contributions, we define the red regions along the structure boundaries as negative contributions towards device performance that should be removed from the design. Accordingly, we devised a boundary extraction algorithm to systematically adjust the shape of the adjoint-optimized devices using the explanation heatmaps. The algorithm, conceptually illustrated in Figure 4a, consists of an initial filtering procedure, which identifies the red-to-blue transition points along the structure boundary. In this procedure, a binarization function is applied to the SHAP values that converts the structure into existing and non-existing elements (shown in the center of Figure 4a as white and black pixels, respectively). Thresholds for binarization ($\rho(x,y)$) are given by the following step function:

$$\rho(x,y) = \begin{cases} 1 & \text{for} \quad \Phi(\mathrm{x},\mathrm{y}) \leq 0, \\ 0 & \text{for} \quad \Phi(\mathrm{x},\mathrm{y}) > 0, \end{cases} \tag{2}$$

where $\rho(x,y)$=1 and $\rho(x,y)$=0 indicates existing and non-existing elements, respectively. A median filter is applied to the binarization to reduce noise. Next, we "draw" a new boundary around the existing elements (Figure 4a, right) by capturing an array of points $\eta(x,y)=[X_\mathrm{i};Y_\mathrm{i}]$, in which $X_\mathrm{i}=[x_1,x_2,...,x_\mathrm{i}]$ and $Y_\mathrm{i}=[y_1,y_2,...,y_\mathrm{i}]$ are vectors of length *i*. X_i is an evenly spaced set of x-coordinate values from the left to the right of the image. Since *i* ultimately determines the resolution of the shape, we set its value to 20 points (matching the interval used in the initial optimization runs) to ensure that the optimizable geometry is within feasible fabrication range. This interval equates to 100 nm spacing along the x-axis, which is well within CMOS lithography resolutions. Each point on the spline can range from 0 to 1.25 µm in 20 nm steps, thus the number of parameter permutations describing the design are on the order of 1×10^{30}. To find the y-coordinate values in Y_i, we apply Algorithm 1 (detailed in the Supporting Information).

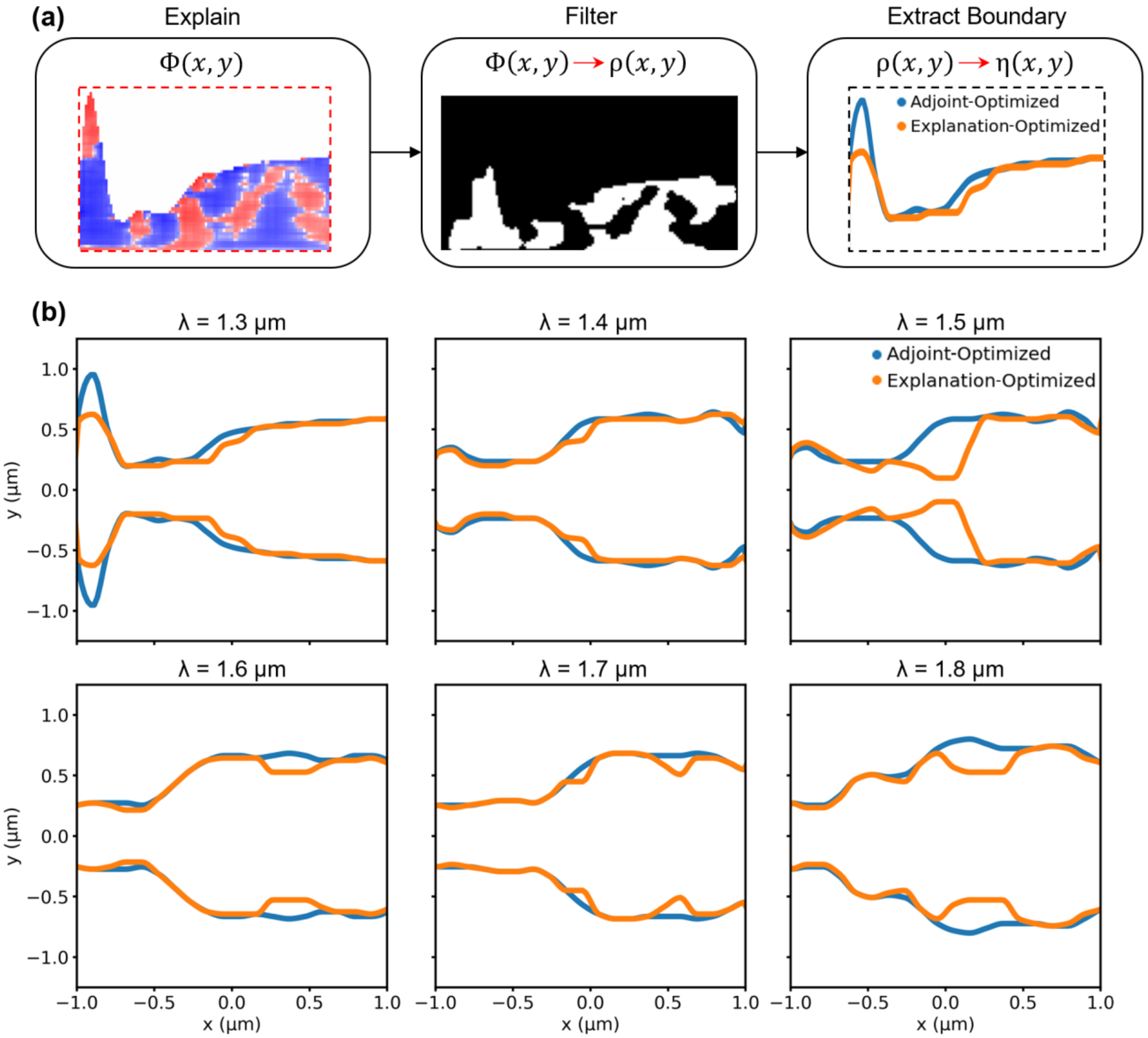


Figure 4. (a) Schematic representation of our explanation-optimization algorithm and workflow, consisting of explanation, filtering, and boundary extraction steps. (b) Comparison between the adjoint-optimized and explanation-optimized geometries.

Using Algorithm 1, we raster the image (from top to bottom) across all values in X_i and find the points where existing elements are found (indicated by $P(x,y)=1$), then mark these points for Y_i. For quality purposes, we add the α hyperparameter to enhance robustness by reducing sharp changes in the structure as a result of filtering or noise from the explanations. We apply this workflow to each wavelength-specific adjoint-optimized structure from the previous step and present the new "explanation-optimized" results in Figure 4b. As an example of our method's application, for the 1.3 μm target design, we note that the explanation algorithm deemed the large vertical spike near the input port as a negative (red) contribution. After applying our explanation-based boundary extraction process, the height of the spike was reduced.

To assess whether the explanation (or SHAP value) based modifications to the optimized structures (*e.g.*, the spike reduction) yielded meaningful or effective contributions, we simulated the explanation-optimized designs and used them as new starting points for a second stage of adjoint optimization runs. In Figure 5, we show the FOM evolutions over the entire optimization cycle of the 35% fill fraction starting design. The explanations and optimization cycles of the remaining starting designs can be found in Figures S2-S5. The red arrows indicate the end of the first optimization stage and the beginning of the second explanation-based re-optimization stage. Further observation revealed that in the second stage of the 1.3 μm target design, reducing the vertical spike immediately reduced the FOM from 0.139 to 0.090 (a 35% improvement) at the first iteration, while the end of the optimization resulted in a final FOM of 0.050 (a 64% improvement compared to the end of the first stage). This result indicates that the explanation-based modifications overcame a saddle point in the original adjoint optimization process. In some of the other examples (*e.g.*, 1.4-1.8 μm), the first step of the second stage did not always result in an immediate FOM reduction, particularly when the FOM value was already exceedingly low (<0.075). We validate in Figure S6 that this increase in the FOM is due to the optimization getting stuck in a local minima valley instead of a saddle point, since the FOM must first increase before the algorithm can identify a lower global minima, particularly when modifying the design from where the optimization algorithm ended. However, across all the optimization targets, the end of every second-stage optimization consistently resulted in a lower final FOM than the first-stage

FOM (a 39% decrease on average). Moreover, an increase in FOM followed by a further global reduction is indicative of an objective function that was previously stuck in a local minimum [14]. Thus, we demonstrate that our explanation-based re-optimization technique is capable of enhancing the performance of the adjoint optimization algorithm by allowing the FOM to escape its local minima for various optimization targets and performance ranges. We note that this entire workflow only used two optimization runs per target: one for feature contribution learning and the other for local minima escape or global FOM reduction. As previously mentioned, alternative methods at identifying lower minima typically involve repeated optimizations at random starting points or metaheuristic approaches, which can scale well-beyond two optimization runs per target (on the order of 1,000 iterations in the case of simulated annealing) [21,22].

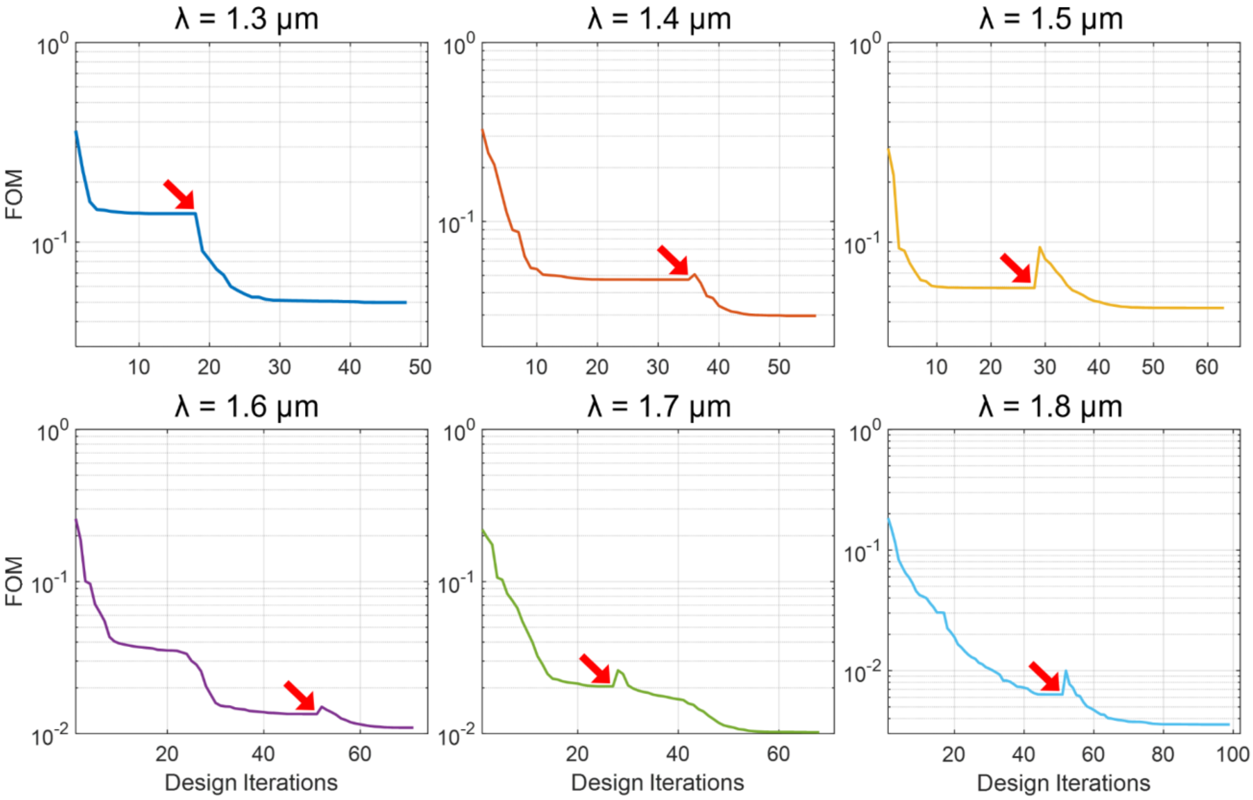


Figure 5. Two-stage optimization of SOI waveguide designs, across target wavelengths ranging from 1.3 μm to 1.8 μm, using the 35% fill fraction starting design. Red arrows indicate the end of

the first adjoint optimization stage and the beginning of the second explanation-based re-optimization stage. Final FOM values are improved by 39%, on average, across all target wavelengths.

In the proceeding sections, we further assess the performance of the presented optimization scheme by conducting a number of additional tests, including: 1) an evaluation of the ability for SHAP to immediately improve the FOM (if the adjoint optimization is stopped prematurely) to determine the link between structure modifications and FOM improvements (*i.e.*, an “early stop” analysis), 2) the applicability of our approach on a smaller dataset and 3) different material systems, and 4) comparisons of our approach against arbitrary perturbations to the adjoint-optimized structure (*i.e.,* a “random change” analysis). First, to verify that the model is actually learning how to modify the structure, in our “early stop” analysis, we removed the portion of the training data where the adjoint optimization reached the local minima, retrained our model, and repeated our explanation-based modification. We observe in Figure S7 that across all target wavelengths, the final FOM obtained through the SHAP explanations is lower than the best available design, thereby confirming that the model is learning how to modify the structure to improve the FOM (based on the information it was given to learn).

Since there is substantial precedence in existing literature on using explainable artificial intelligence with small training datasets (on the order of several dozens to hundreds of data points), particularly to reduce the burden of data collection or computation costs [36-40], we have repeated our study on a reduced dataset using only the optimization results from a single starting design (35% fill fraction). Results of this analysis are shown in Figures S8 and S9, where we also observe performance improvements (43% on average) for all target wavelengths, which suggests that the proposed approach is applicable (to a degree) to smaller datasets.

Next, we evaluated the generalizability of the proposed framework by applying it to other contemporary nanophotonics design challenges. We note that prior studies have successfully applied XAI to alternative nanophotonic structures, such as metasurfaces, and demonstrated performance enhancements in the form of spectral property tuning [26] (though no optimization algorithm integration was employed). Thus, we focused this generalizability analysis on material alternatives. In this regard, over the past few years, Si_3N_4 has emerged as a promising alternative to silicon in photonic systems. Compared to silicon, Si_3N_4 has lower propagation losses and does not exhibit two-photon absorption in the telecommunications range [31-34] (among other pros and cons). Accordingly, we performed the same two-stage optimization study on a Si_3N_4 waveguide (for the same Y-splitter starting geometry) and found that the second-round optimizations were also able to surpass the results of the first at every target wavelength (Figure S10). Across all the test cases, an average FOM improvement of 74% was achieved.

Lastly, to show that the achieved performance enhancements were not simply obtained through arbitrary perturbations to the optimized structure, we performed an additional "random change" analysis where we randomly modified the first-stage structures, repeated the second stage of optimizations, and compared the results. This comparison is presented in Figure S11, where five random modifications (defined in the Supporting Information) were made to each structure. Across the 30 tests performed on the six optimized designs, all of the randomly modified structures possess higher FOM values (*i.e.,* lower performance) than those of the explainability-optimized devices, while only two "random change" results fall within 25% of the explainability-optimized device performance. Additionally, 28 tests produced higher final FOM values than the initial optimized designs. Therefore, not only are the random changes ineffective in terms of escaping the local minima, but they can also inadvertently push the optimization into a worse state than where the optimization started at. As such, we demonstrate that our XAI-based approach is not stochastic in nature, but can deterministically tune a structure in order to maximize performance. Through the preceding series of tests, we show that the presented approach is generally applicable to numerous applications of adjoint optimization for electromagnetic design, including those with different constituent materials, structures, and optimization targets.

5.3 Conclusions

In summary, we present an inverse design framework that extends the capabilities of gradient-based shape or topology optimization algorithms for photonic inverse design, while elucidating the relationships between device performance and nanoscale structuring. Our framework combines adjoint optimization, AutoML, and XAI to enhance device performance beyond that which is obtainable through the optimization algorithm alone. We applied our method to SOI waveguide design and showed that the optimization algorithm initially reaches a performance plateau (*i.e.*, local minima). After utilizing XAI to reveal the device's structural contributions towards a designated FOM (where 0 represents ideal performance), we leveraged this information (in conjunction with a boundary extraction algorithm) to push the optimization out of its local minima and reduce the FOM further. Across a range of performance-plateaued devices optimized for various wavelengths within the 1.3 μm to 1.8 μm telecom window, our method was able to improve device performance by an average of 39%. The entire procedure only requires two optimization runs per optimization target, which is potentially more computationally efficient than alternative approaches, particularly those that rely on multiple optimization runs and random starting points. Additionally, generalizability tests performed on Si_3N_4 waveguides showed an average of 74% device performance improvement. Thus, we demonstrate that our XAI-based approach provides an automated and systematic solution for an electromagnetic optimization algorithm to escape local minima and achieve greater device performance. Looking ahead, integrating conventional optimization and data-driven machine learning will likely prove a fruitful direction for inverse design and physics discovery in photonic systems.

5.3 Supporting Information

AutoML Training and Execution

In this work, we use neural architecture search (AutoKeras) with image blocks as well as an early stopping callback to automate the deep learning process by testing different model variants across multiple trials. We observe that the optimal architecture was identified after 12 trials, which has a validation loss of 9.1×10^{-5}. The final training and validation losses of each trial are presented in Figure S1a, and the evolution of the convolutional neural network (CNN) architecture from the first trial to the last is shown in Figure S1b. The first trial initially begins with a two-block architecture and a validation loss of 3.8×10^{-2}. At the final trial, the optimized CNN possesses five convolutional blocks (with 512, 256, 128, 64, and 32 filters, respectively) followed by a dense layer. Each block contains Leaky ReLu, batch normalization, and max pooling layers. The final model possesses approximately 1.5 million training parameters. We note that the number of parameters are attributed to the size of the input images (128×64 pixels).

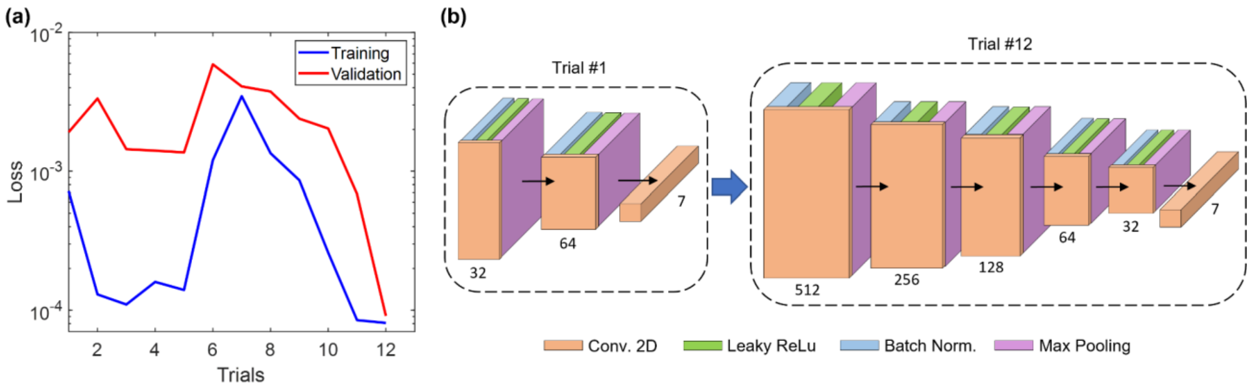


Figure S1. (a) Training and validation loss progression across the AutoML trials. (b) Schematic of the model architecture changes between the first and final trials.

Boundary Extraction Algorithm

As described in the main text, we devised a boundary extraction algorithm to systematically adjust the shape of the adjoint-optimized devices using the explanation heatmaps. This algorithm is represented as:

Algorithm 1: Boundary Extraction

```
Input: Filtered image array ($P$) of $\rho(x,y)$ values at every image pixel.
for $x \in X_\mathrm{i}$ do
  for $y = 1:L$, where $1$ is the top of the image and $L$ is the height or bottom of the image, do
    if $P(x,y) = 1$ and $y_i - y_{i-1} < \alpha$, where $\alpha$ is a hyperparameter for enhancing robustness, then
      $Y_\mathrm{i}.append(y_i)$
    else
      $Y_\mathrm{i}.append(y_{i-1})$
Return: $Y_\mathrm{i}$
```

As shown in Algorithm 1, we raster the image (from top to bottom) across all values in X_i and find the points where existing elements are found (indicated by $P(x,y)=1$), then mark these points for Y_i. For quality purposes, we add the α hyperparameter (with a default value of 5 points or 100 nm in device length) to enhance robustness by reducing sharp changes in the structure as a result of filtering or noise from the explanations.

Random Starting Designs and Explanation Results

We generated the data for our model by performing adjoint optimization runs on randomized starting designs. Figure S2 below shows all the starting designs, where S2a, S2b, S2c, and S2d (first column) correspond to 25%, 35%, 40%, and 50% fill fractions, respectively. The optimization runs on the presented starting designs (at the designated target wavelengths) yielded a training dataset of over 600 input-output pairs. After training our model on this larger dataset, SHAP explanations were captured for each optimized structure (second column of Figure S2) and re-optimized following the same procedure described in the main text (third column of Figure S2).

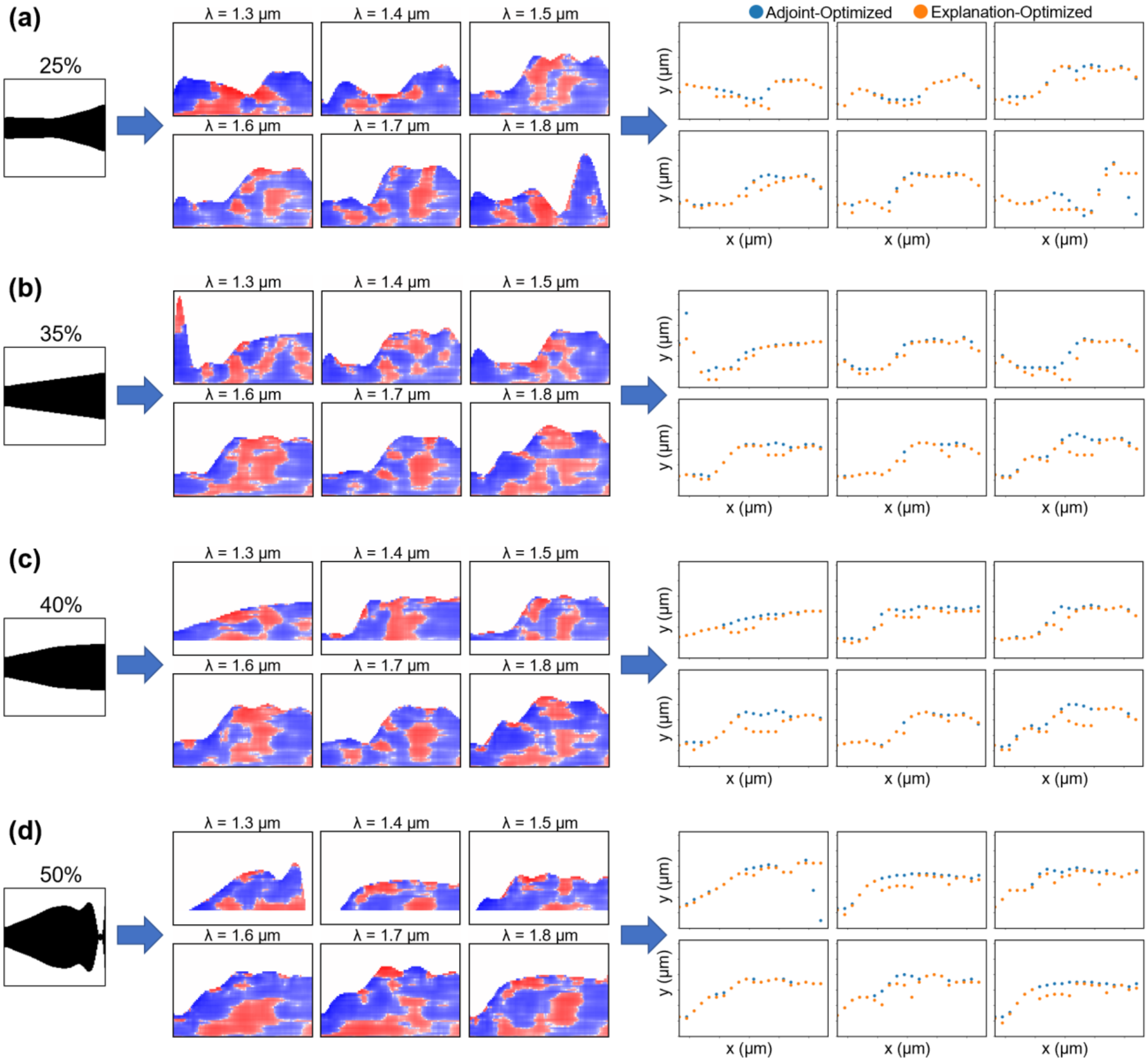


Figure S2. Randomized adjoint optimization starting designs with (a) 25%, (b) 35%, (c) 40%, and (d) 50% fill fractions. Device layouts (first column) and their corresponding target wavelength-specific adjoint-optimized device explanations (second column) are illustrated. Top halves of the structures are presented for ease of visualization. Explanations are used to identify new starting points for optimization, which can escape local minima (third column). Optimization results of the presented designs are shown in Figures 5 and S3-S5.

In Figures S3, 5, S4, and S5, we present the optimization cycles of the 25%, 35%, 40%, and 50% fill fraction starting designs, respectively. The red arrow indicates the end of the first optimization stage and the beginning of the second explanation-based re-optimization stage. Across each starting design, the end of every second-stage optimization consistently resulted in a lower final FOM than the first-stage FOM. Specifically, the 25% fill fraction starting design ended with a 47% average improvement across all target wavelengths, while the 35%, 40%, and 50% fill fractions resulted in 39%, 45%, and 45% improvements, respectively.

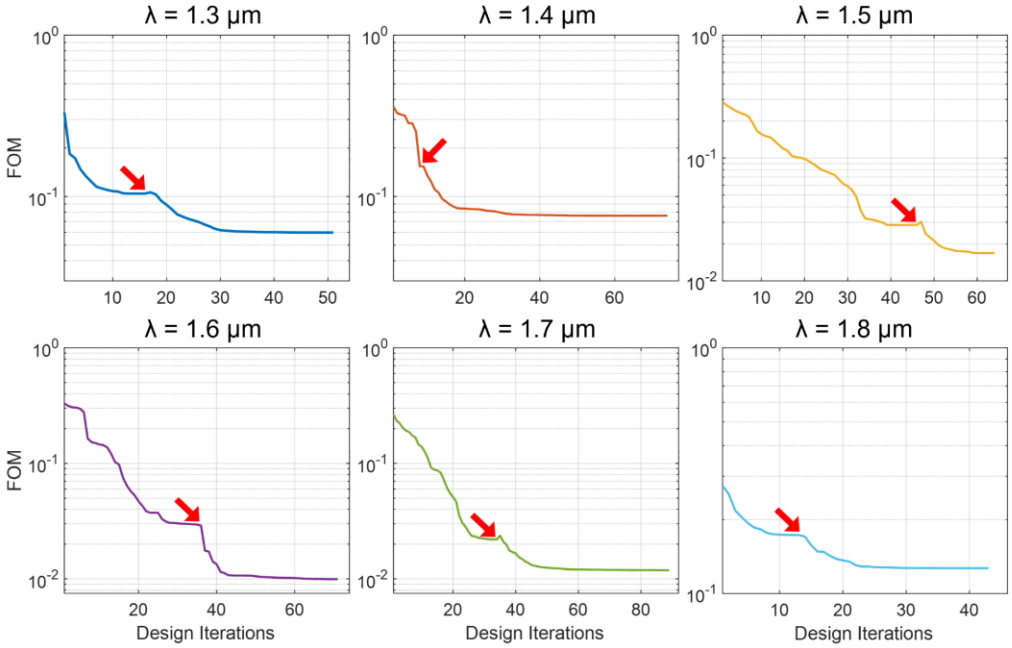


Figure S3. Two-stage optimization of SOI waveguide designs, across target wavelengths ranging from 1.3 μm to 1.8 μm, using the 25% fill fraction starting design shown from Figure S2. Red arrows indicate the end of the first adjoint optimization stage and the beginning of the second explanation-based re-optimization stage. Final FOM values are improved by 47%, on average, across all target wavelengths.

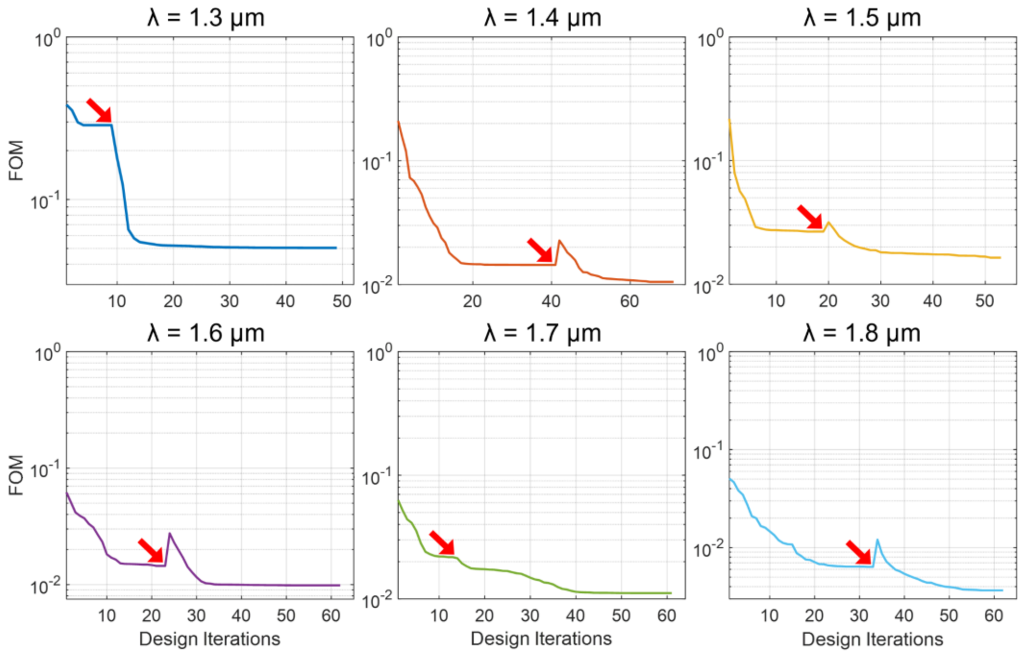


Figure S4. Two-stage optimization of SOI waveguide designs, across target wavelengths ranging from 1.3 μm to 1.8 μm, using the 40% fill fraction starting design shown from Figure S2. Red arrows indicate the end of the first adjoint optimization stage and the beginning of the second explanation-based re-optimization stage. Final FOM values are improved by 45%, on average, across all target wavelengths.

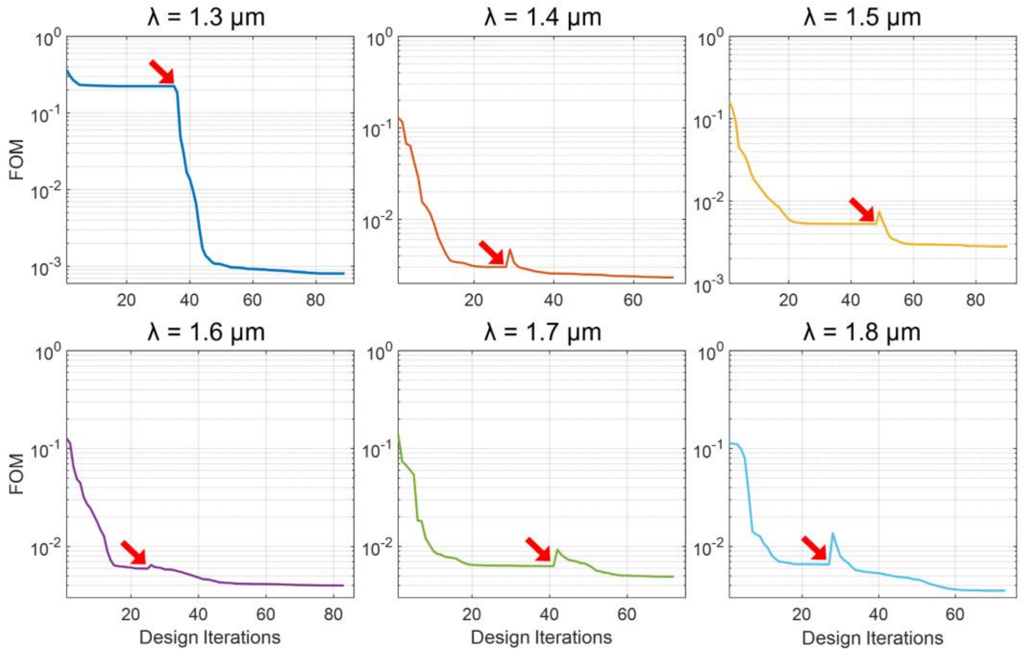


Figure S5. Two-stage optimization of SOI waveguide designs, across target wavelengths ranging from 1.3 μm to 1.8 μm, using the 50% fill fraction starting design shown from Figure S2. Red

arrows indicate the end of the first adjoint optimization stage and the beginning of the second explanation-based re-optimization stage. Final FOM values are improved by 45%, on average, across all target wavelengths.

Local Minima Analysis

We demonstrate that the adjoint-optimized designs are trapped in a local minimum by attempting to improve a corresponding design further through exhaustive exploration. Here, we focus our analysis on the waveguide structure from the main text (Figure 3) that is optimized for power transmission at 1.8 μm. Using this structure, we individually modify each of the 20 parameters that make up the overall shape of the waveguide (conceptually illustrated in Figure S6a). In Figure S6b, we observe that each of these changes (40 new simulations total) yield a higher FOM than the original adjoint-optimized design. Thus, the design here has reached a local minimum and cannot be improved through gradient-based optimization alone. However, since the adjoint-optimized design was obtained after more than 50 iterations, an alternative optimization route may be identified by utilizing an earlier design iteration as the base of the analysis, but this would require a large number of additional simulations. Furthermore, if multiple interdependent parameters or points are considered, this can exponentially increase computation time and requirements. Therefore, the presented XAI approach simply augments this process by rapidly guiding the optimization towards a location where the adjoint method or sensitivity analysis performs better. This is indicated by the dashed lines in Figure S6b, where we show that the FOM of the explanation-optimized design is lower than that obtained through adjoint optimization.

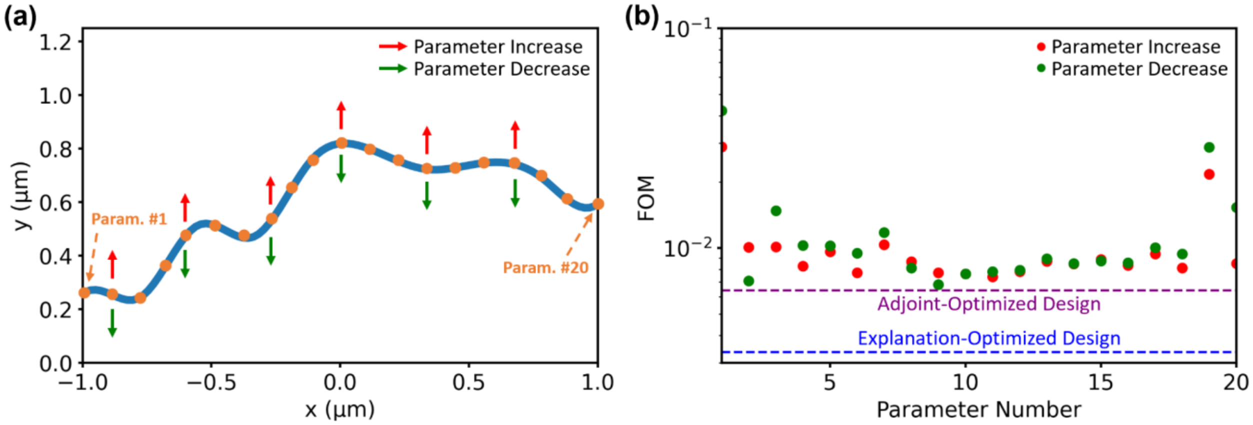


Figure S6. Local minimum analysis of an adjoint-optimized design. (a) Schematic illustration of the analysis approach, where each shape parameter is individually modified by either increasing (red) or decreasing (green) the size or spatial coordinate of the structure by 50 nm, while all other parameters remain fixed. (b) The resulting FOM values of the corresponding parameter changes show that the adjoint-optimized design is stuck in a local minimum, while the explanation-optimized design can improve the design further by finding a new optimization route.

Early Stop Analysis

To verify that the model is actually learning how to modify the structure, we evaluate the ability for SHAP to immediately improve the FOM to determine the link between structure modifications and FOM improvements. Since in the main text, the SHAP modifications either yield an increased or decreased FOM (due to the presence of either a saddle point or local minimum valley), this link is not apparent. To perform this analysis, we removed the portion of the training data where the adjoint optimization reached the local minima, retrained our model, and repeated the explanation-based modification. In this “early stop” analysis (shown in Figure S7), the solid lines represent the optimization data that is included in the training dataset, and the (*) marker represents the FOM of the modified structure. We note that across all target wavelengths, the final FOM obtained through the SHAP explanations is lower than the best available design, thereby confirming that the model is learning how to modify the structure to improve the FOM (based on the information it was given to learn).

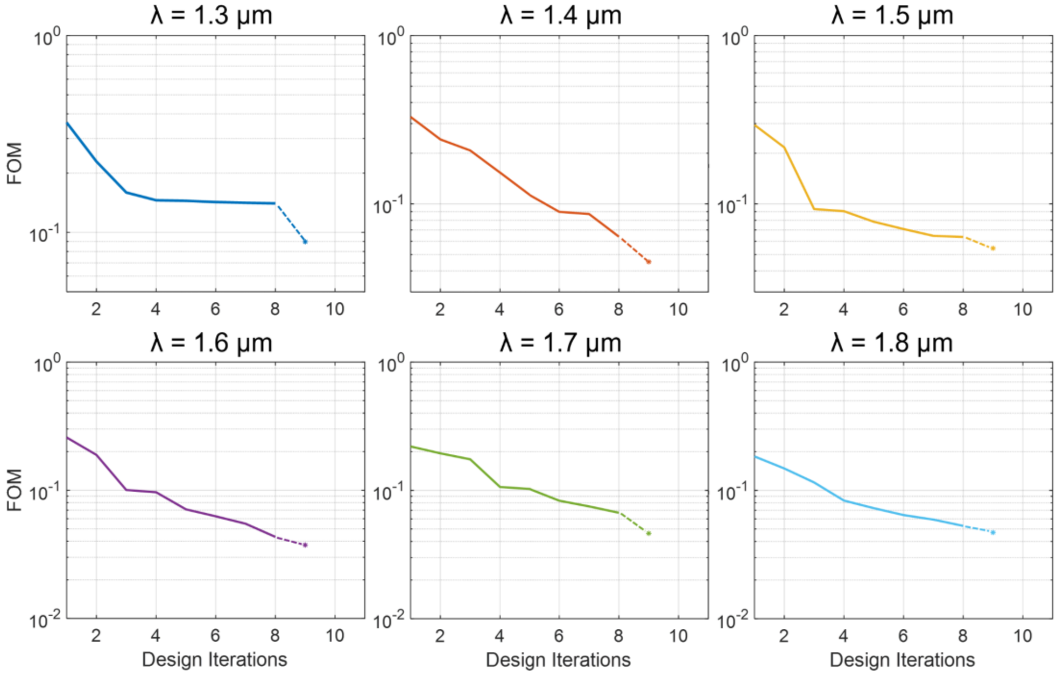


Figure S7. Explanation-modified structures, indicated by the dashed lines leading to the "*" markers, when the adjoint optimization is stopped prematurely. Modified structures possess lower FOM values than the best structure from the training dataset, which further indicate that explanation-based modifications are linked to FOM improvements.

Reduced Dataset Analysis

Since there is substantial precedence in existing literature on using explainable artificial intelligence with small training datasets (on the order of several dozens to hundreds of data points), particularly to reduce the burden of data collection or computation costs, we have repeated our study on a reduced dataset using only the optimization results from a single starting design (35% fill fraction), which yielded 192 input-output pairs. Using this reduced dataset, AutoML identified the optimal model architecture after 17 trials, which has a validation loss of 8.6×10^{-5}. Similar to the main text, Figure S8 compares the model predictions (blue points) with the ground truths (orange points) using this reduced dataset. From this result, we can infer that the model accurately learned the key features on the optimized structures which contribute to the target wavelength-specific FOM values. Therefore, we can utilize XAI to reveal the structure-performance relationships of each device.

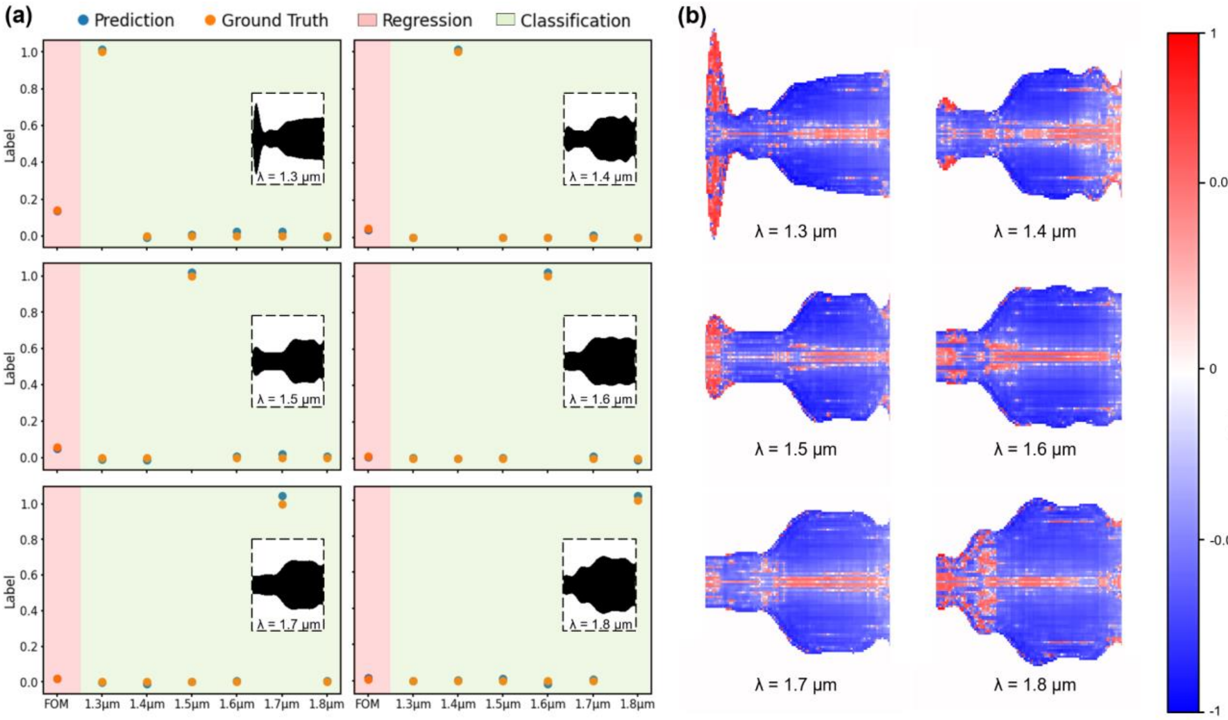


Figure S8. (a) Comparisons between model predictions and ground truths, for FOM (regression) and target wavelength (classification) values, show that the model accurately learned the relationship between device structure and performance. Inset images show the model inputs, which are adjoint-optimized devices. (b) SHAP explanation heatmaps of the optimized devices reveal the structural features that contribute positively (blue) or negatively (red) towards optimal device performance.

In Figure S9, we show the FOM evolutions over the entire optimization cycle of the 35% fill fraction starting design, and similarly observe that the end of every second-stage optimization consistently resulted in a lower final FOM than the first-stage FOM (by 43% on average). We note that the reduced dataset (with only the 35% fill fraction starting design) performed slightly better than the larger-data model for this particular starting design (a 39% improvement for the larger dataset in comparison to a 43% improvement for the reduced dataset). Thus, we demonstrate that the presented approach can yield performance improvements in a wide range of randomized starting points. However, the degree of performance improvement may be linked to the size and breadth of the training data, which we aim to investigate in future works.

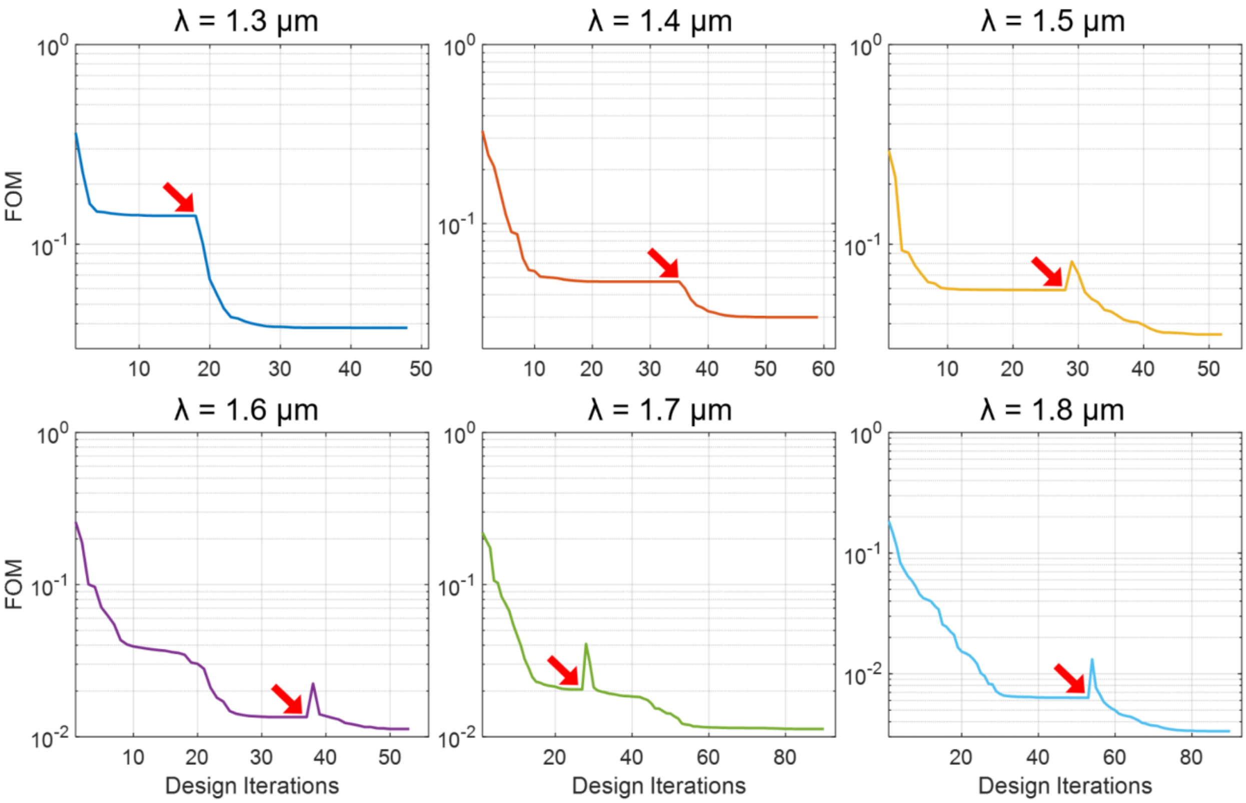


Figure S9. Two-stage optimization of SOI waveguide designs across target wavelengths ranging from 1.3 μm to 1.8 μm. Red arrows indicate the end of the first adjoint optimization stage and the beginning of the second explanation-based re-optimization stage. Final FOM values are improved by 43%, on average, across all target wavelengths.

Random Change Comparison

To show that the achieved performance enhancements were not simply obtained through arbitrary perturbations to the optimized structure, we performed an additional "random change" analysis where we randomly modified the first-stage structures, repeated the second stage of optimizations, and compared the results. This comparison is presented in Figure S10, where five random modifications were made to each structure. Random modifications were introduced by adding 300 nm to y_n and y_{n+1} in the optimizable parameter vector Y_i , where n is varied from 1 to i-1 (where i is the length of the vector). Across the 30 tests performed on the six optimized designs, all of the randomly modified structures possess higher FOM values (*i.e.*, lower performance) than the explainability-optimized devices, while only two results fall within 25% of the latter's final device performance. Additionally, 28 tests yield higher final FOM values than the initial optimized design. Therefore, not only are the random changes ineffective in terms of escaping the local minima, but they can also inadvertently push the optimization into a worse state than where the optimization started at.

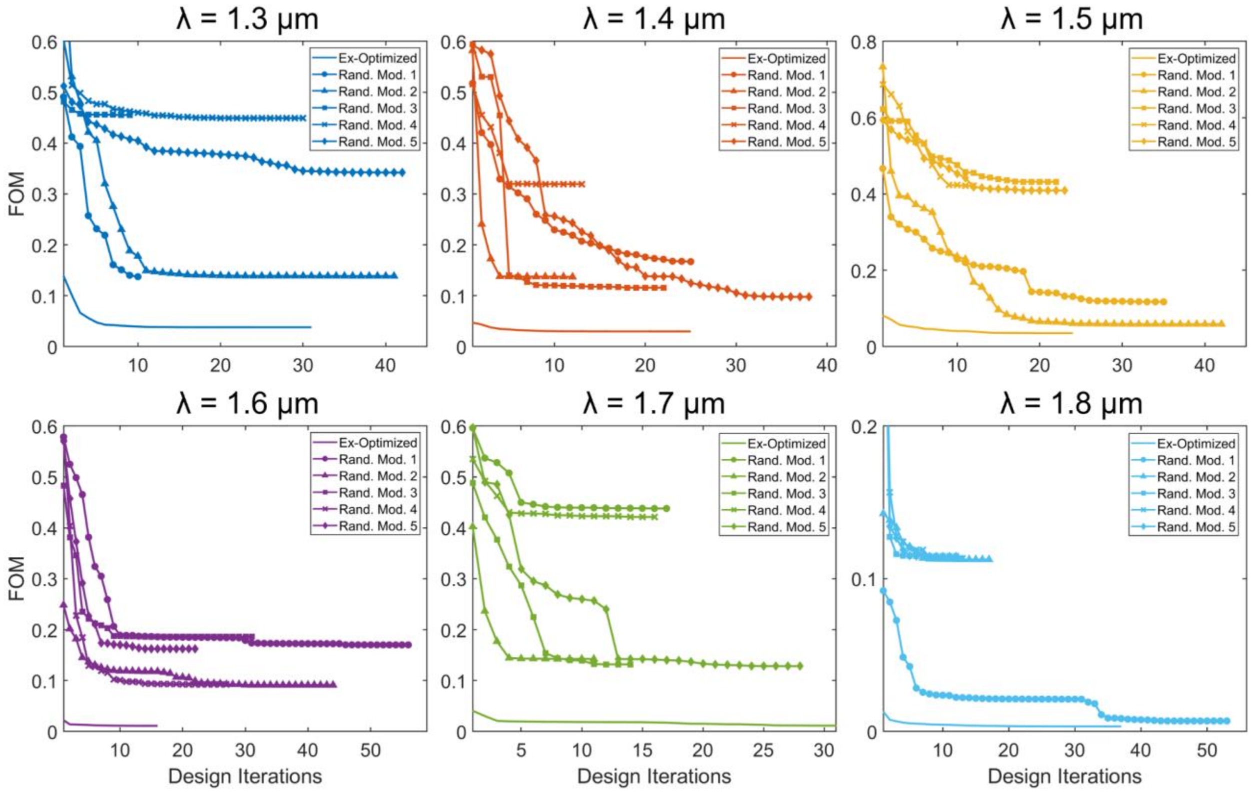


Figure S10. Comparison between the explanation-optimized and randomly modified structures. Random modifications either produce FOM values that are higher (lower performance) than the explanation-optimized structures, or higher than the starting point of optimization.

Generalizability or Material Alternative Assessment

We evaluated the generalizability of the proposed framework by applying it to other contemporary nanophotonics design challenges. In this regard, we performed the same two-stage optimization study on a Si_3N_4 waveguide (for the same Y-splitter starting geometry). Similar to the results presented in the main text (Figure 5), here we observe that the second stage (red-boxed region) of each target wavelength-specific optimization produces a lower overall FOM than the final FOM obtained through adjoint optimization alone (indicated by the red arrows). In particular, as shown in Figure S11, we observe a minimum of 31% FOM improvement (for the 1.6 µm design) and a maximum of 90% improvement (for the 1.3 µm design), for an average of 74% improvement across all the test cases. As a result, we show that the presented approach is generally applicable to numerous applications of adjoint optimization for electromagnetic design, including those with different constituent materials, structures, and optimization targets.

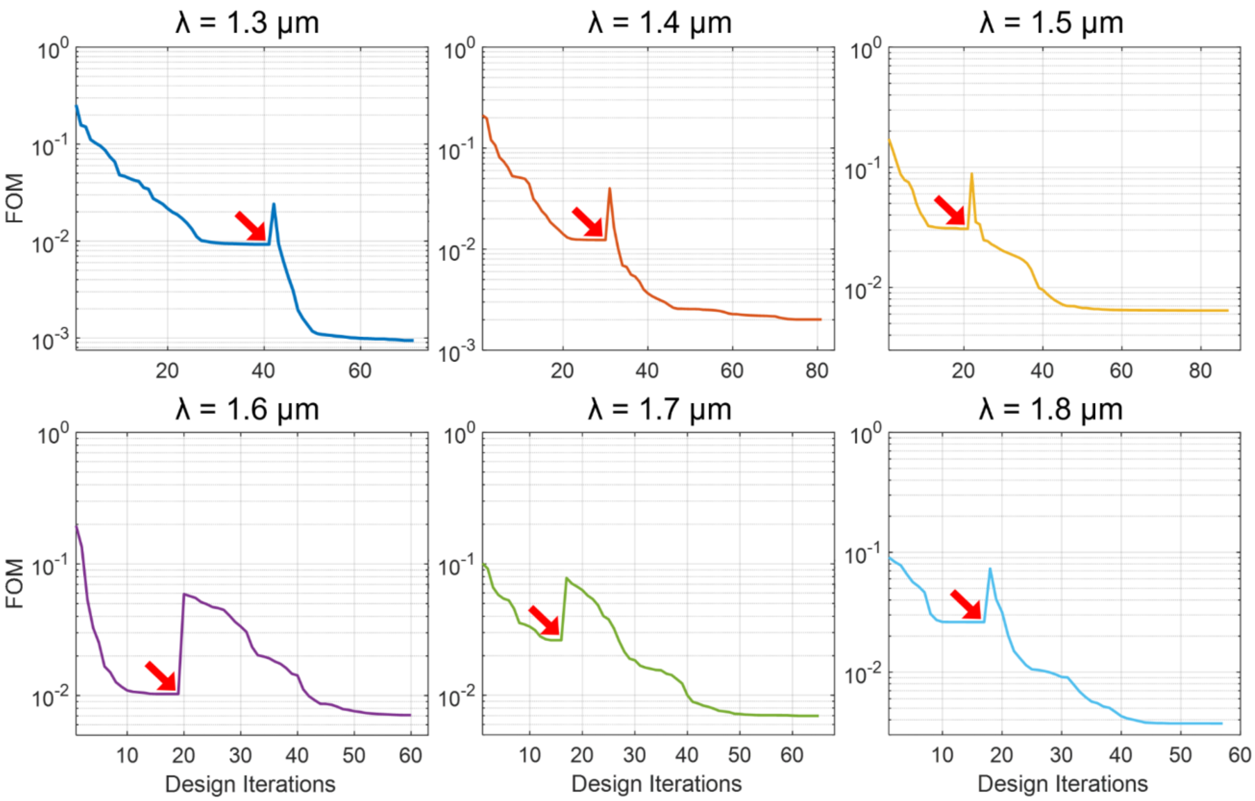


Figure S11. Two-stage optimization of Si_3N_4 waveguide designs across target wavelengths ranging from 1.3 μm to 1.8 μm. Red arrows indicate the end of the first adjoint optimization stage and the beginning of the second explanation-based re-optimization stage. Final FOM values are improved by 74%, on average, across all target wavelengths.

Cross Validation and Overfitting Analysis

To ensure that our model did overfit during or after training, we performed a number of additional tests. First, we performed k-fold cross validation. Here, the dataset was split into ‘k’ consecutive folds, and each fold was used once as validation while the remaining ‘k-1’ folds formed the training data (at each fold number), thus the entire dataset was used to validate our model. For a more thorough analysis, we performed 10-fold and 5-fold cross validations on our optimized model, and presented the results in Tables S1 and S2, respectively. 10-fold cross validation (where 90% of the data was used for training and 10% for validation, at each fold, number after shuffling the data) resulted in an average MSE of 2.1×10^{-4} with a standard deviation of 0.01%. On the other hand, 5-fold cross validation (where 80% of the data was used for training and 20% for validation, at each fold, number after shuffling the data) resulted in an average MSE of 2.9×10^{-4} with a standard deviation of 0.01%. Since the MSE values derived from the 5-fold validation were considerably larger than those obtained from the 10-fold validation, it can be inferred that an increased amount of training data (90% vs. 80%) contributed to increased model performance. Additionally, across both cross-validation procedures, the validation losses were consistently low across each fold (evident from the 0.01% standard deviation between folds), which indicates that our model is not overfitting against a particular validation dataset.

Table S1. K-fold cross validation results (k=10).

Fold Number	**Validation Loss (MSE)**
1	1.2019×10^{-4}
2	3.0729×10^{-4}
3	3.7655×10^{-4}
4	9.0292×10^{-5}
5	2.3388×10^{-4}
6	1.4062×10^{-4}
7	1.5254×10^{-4}
8	3.8326×10^{-4}
9	1.9175×10^{-4}
10	1.3702×10^{-4}
Average MSE	2.1334×10^{-4}
Standard Deviation (%)	0.010742

Table S2. K-fold cross validation results (k=5).

Fold Number	**Validation Loss (MSE)**
1	2.0566×10^{-4}
2	1.8427×10^{-4}
3	2.5280×10^{-4}
4	4.1815×10^{-5}
5	4.0167×10^{-4}
Average MSE	2.9251×10^{-4}
Standard Deviation (%)	0.0110

For another detailed look at model performance, we performed a goodness of fit analysis on our model and included a new test set that was derived from the previous training and validation datasets. In this new analysis, we used 80% of the data for training, 10% for validation, and 10% for testing. To adhere to standard machine learning practices, we note that this test set was used to evaluate the model *after* training, whereas the validation set was used to evaluate the model *during* training. Losses for the training, validation, and test sets are 2.0×10^{-4}, 2.5×10^{-4}, and 2.6×10^{-4}, respectively. We believe this result shows that the test set, previously not seen by the neural network in either training or validation, has almost exactly the same loss value as the validation set, which is a strong indication of generalization. Additionally, Figure S12 shows the ground truth and prediction comparisons for each datapoint, where we observe that our model's predictions are accurate for both high and low performance designs across each dataset (training, validation and test), thus further indicating that the model is not overfitting or skewed towards near-optimal devices.

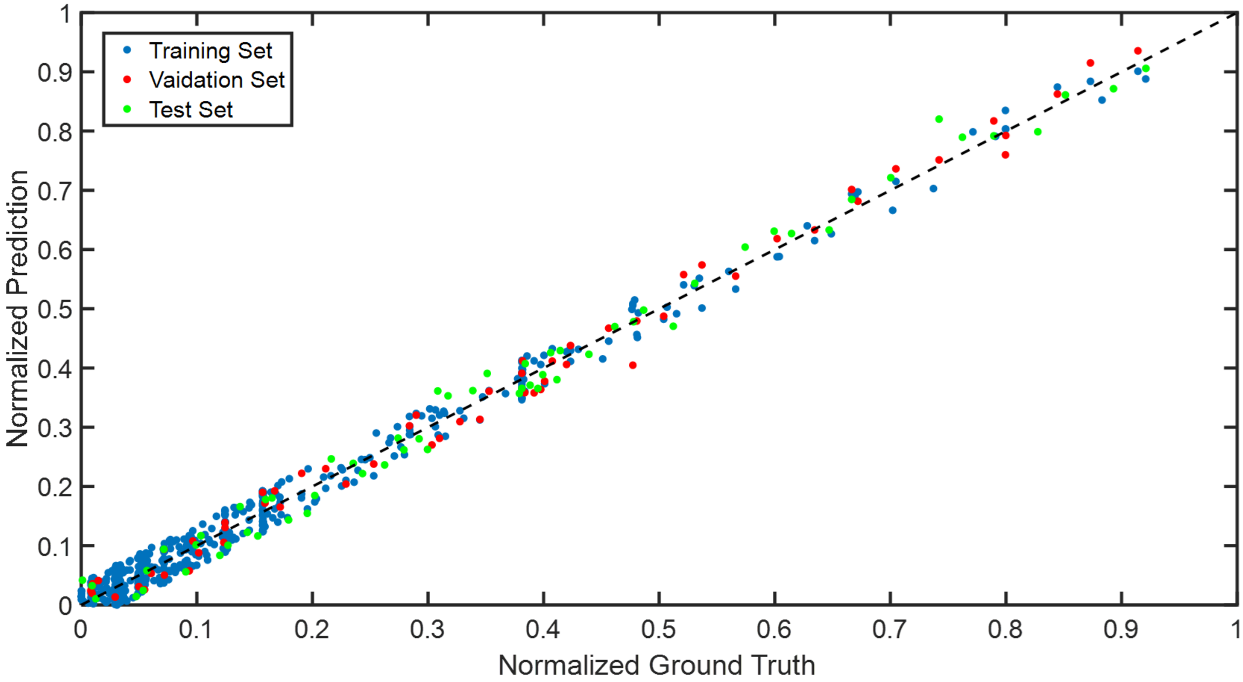


Figure S12. Goodness of fit analysis on model predictions vs. ground truth (normalized) in the training (80%), validation (10%), and test (10%) sets. Losses for the training, validation, and test sets are 2.0×10^{-4}, 2.5×10^{-4}, and 2.6×10^{-4}, respectively.

In the original training process, we included an early stopping callback function to ensure that training automatically terminates once the validation loss shows no further improvement. However, to determine whether our model will eventually overfit during training (such that we can confirm that our model did not overfit over the allotted epochs), we removed the early stop callback and retrained the model up to 1000 epochs. In Figure S13 below, we observe that the training and validation losses begin to diverge at approximately 400 epochs (where the training was previously terminated), which indicates that the model is memorizing the training data after 400 epochs.

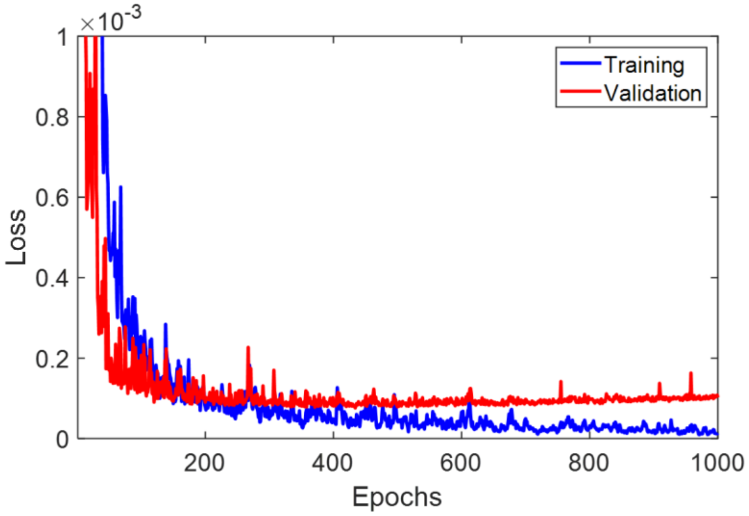


Figure S13. Model training without early stopping. Training and validation losses begin to diverge at approximately 400 epochs, thus indicating that the final model did not overfit the training data.

5.5 References

[1] Sandborn, P.; Quack N.; Hoghooghi, N.;Chou, J. B.; Ferrara, J.; Gambini, S.; Behroozpour, B.; Zhu, L.; Boser, B.; Chang-Hasnain, C., and Wu, M. C. Linear frequency chirp generation employing optoelectronic feedback loop and integrated silicon photonics. *CLEO* **2013**, 1-2.

[2] Meng, Y., Chen, Y., Lu, L., Ding, Y., Cusano, A., Fan, J. A., Hu, Q., Wang, K., Xie, Z., Liu, Z., Yang, Y., Liu, Q., Gong, M., Xiao, Q., Sun, S., Zhang, M., Yuan, X., & Ni, X. Optical meta-waveguides for integrated photonics and beyond. *Light: Science & Applications* **2021**, *10*, 1–44.

[3] Krutova, I. A.; Saygin, M. Yu.; Dyakonov, I. V.; Kulik, S. P. Optimized low-loss integrated photonics silicon-nitride Y-branch splitter. *AIP Conference Proceedings* **2020**, *2241*, 020027.

[4] Shen, Y.; Harris, N. C.; Skirlo, S.; Prabhu, M.; Baehr-Jones, T.; Hochberg, M.; Sun, X.; Zhao, S.; Larochelle, H.; Englund, D.; Soljacic, M. Deep learning with coherent nanophotonic circuits. *Nature Photon.* **2017**, *11*, 441.

[5] Harris, N. C.; Carolan, J.; Bunandar, D.; Prabhu, M.; Hochberg, M.; Baehr-Jones, T.; Fanto, M. L.; Smith, A. M.; Tison, C. C.; Alsing, P. M.; Englund, D. Linear programmable nanophotonic processors. *Optica* **2018** *5*, 1623.

[6] Miller, O.D. Photonic Design: From Fundamental Solar Cell Physics to Computational Inverse Design. *University of California, Berkeley* **2012**, 1-137.

[7] Li, W.; Meng, F.; Chen, Y.; Li, Y.; Huang, X. Topology optimization of photonic and phononic crystals and metamaterials: a review. *Adv. Theory Simul.* **2019**, *2*, 1900017.

[8] Campbell, S. D.; Sell, D.; Jenkins, R. P.; Whiting, E. B.; Fan, J. A.; Werner, D. H. Review of numerical optimization techniques for meta-device design. *Opt. Mater. Express* **2019**, *9*, 1842-1863.

[9] Molesky, S.; Lin, Z.; Piggott, A. Y.; Jin, W.; Vucković, J.; Rodriguez, A. W. Inverse design in nanophotonics. *Nature Photon.* **2018**, *12*, 659–670.

[10] Sell, D.; Yang, J.; Doshay, S.; Yang, R.; Fan, J. A. Large-angle, multifunctional metagratings based on freeform multimode geometries. *Nano Lett.* **2017**, *17*, 3752–3757.

[11] Matzen, R.; Jensen, J. S.; Sigmund, O. Topology optimization for transient response of photonic crystal structures. *J. Opt. Soc. Am. B* **2010**, *27*, 2040-2050.

[12] Deng, Y.; Korvink, J. G.. Topology optimization for three-dimensional electromagnetic waves using an edge element-based finite-element method. *Proc. R. Soc. A* **2016**, *472*, 20150835.

[13] Shen, B.; Wang, P.; Polson, R.; R. Menon. Integrated metamaterials for efficient and compact free-space-to-waveguide coupling. *Opt. Express* **2014**, *22*, 27175–27182.

[14] Jiang, J.; Chen, M.; Fan, J. A. Deep neural networks for the evaluation and design of photonic devices. *Nat Rev Mater* **2021**, *6*, 679–700.

[15] Fan, J. A. Freeform metasurface design based on topology optimization. *MRS Bulletin* **2020**, *45*, 196–201.

[16] Jiang, J.; Fan, J. A. Multiobjective and categorical global optimization of photonic structures based on ResNet generative neural networks. *Nanophotonics* **2021**, *10*, 361-369.

[17] Yeung, C.; Tsai, R.; Pham, B.; King, B.; Kawagoe, Y.; Ho, D.; Liang, J.; Knight, M. W.; Raman, A. P. Global Inverse Design across Multiple Photonic Structure Classes Using Generative Deep Learning. *Adv. Optical Mater.* **2021**, *9*, 2100548.

[18] Ma, W.; Liu, Z.; Kudyshev, Z. A.; Boltasseva, A.; Cai, W.; Liu, Y.. Deep learning for the design of photonic structures. *Nature Photon.* **2021**, *15*, 77–90.

[19] Nalep, J. Smart Delivery Systems: Solving Complex Vehicle Routing Problems. *Elsevier* **2020**, 1-276.

[20] Kim, W. J.; O'Brien, J. D.. Optimization of a two-dimensional photonic-crystal waveguide branch by simulated annealing and the finite-element method. *J. Opt. Soc. Am. B* **2004**, *21*, 289-295.

[21] Zhao, C.; Zhang, J. Binary plasmonics: launching surface plasmon polaritons to a desired pattern. *Opt. Lett.* **2009**, *34*, 2417-2419.

[22] Fayyaz, Z.; Mohammadian, N.; Salimi, F.; Fatima, A.; Tabar, M. R. R.; Avanaki, M. R. N. Simulated annealing optimization in wavefront shaping controlled transmission. *Appl. Opt.* **2018**, *57*, 6233-6242.

[23] Vilone, G.; Longo, L. Explainable Artificial Intelligence: a Systematic Review. *arXiv:2006.00093 [cs]*, May 29, **2020**. https://arxiv.org/abs/2006.00093 (accessed 2020-05-15).

[24] Lundberg, S. M.; Nair, B.; Vavilala, M. S.; Horibe, M.; Eisses, M. J.; Adams, T.; Liston, D. E.; Low, D. K.; Newman, S.; Kim, J.; Lee, S. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2018**, *2*, 749–760.

[25] Arrieta, A. B.; Díaz-Rodríguez, N.; Ser, J. D.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; Herrera, F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Info. Fus.* **2020**, *58*, 82-115.

[26] Yeung, C.; Tsai, J.; King, B.; Kawagoe, Y.; Ho, D.; Knight, M. W.; Raman, A. P. Elucidating the Behavior of Nanophotonic Structures through Explainable Machine Learning Algorithms. *ACS Photo.* **2020**, *7*, 2309–2318.

[27] Lundberg, S.; Lee, S. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs]*, May 22, **2017**. https://arxiv.org/abs/1705.07874 (accessed 2020-06-11).

[28] Lalau-Keraly, C. M.; Bhargava, S.; Miller, O. D.; Yablonovitch, E. Adjoint shape optimization applied to electromagnetic design. *Opt. Express* **2013**, *21*, 21693-21701.

[29] Osher, S.; Sethian, J. A. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *J. Comput. Phys.* **1988**, *79*, 12–49.

[30] Jin, H.; Song, Q.; Hu, X. Auto-Keras: An Efficient Neural Architecture Search System. *arXiv:1806.10282 [cs]*, June 27, **2018**. https://arxiv.org/abs/1806.10282 (accessed 2020-06-11).

[31] Zhang, Y.; Husko, C.; Lefrancois, S.; Rey, I. H.; Krauss, T. F.; Schröder, J.; Eggleton, B. J. Non-degenerate two-photon absorption in silicon waveguides: analytical and experimental study. *Opt. Express* **2015**, *23*, 17101-17110.

[32] Krutova, I. A.; Saygin, M. Y.; Dyakonov, I. V.; Kulik, S. P. Optimized low-loss integrated photonics silicon-nitride Y-branch splitter. *AIP Conf. Proc.* **2020**, *2241*, 020027.

[33] Dirani, H. E.; Youssef, L.; Petit-Etienne, C.; Kerdiles, S.; Grosse, P.; Monat, C.; Pargon, E.; Sciancalepore, C. Ultralow-loss tightly confining Si3N4 waveguides and high-Q microresonators. *Opt. Express* **2019**, *27*, 30726-30740.

[34] Baets, R.; Subramanian, A. Z.; Clemmen, S.; Kuyken, B.; Bienstman, P.; Le Thomas, N.; Roelkens, G.; Van Thourhout, D.; Helin, P.; Severi, S. Silicon Photonics: silicon nitride versus silicon-on-insulator. *Opt. Fiber Comm. Conf.* **2016**, 1-3.

[35] Jin, C.; Ge, R.; Netrapalli, P.; Kakade, S. M.; Jordan, M. I. How to escape saddle points efficiently. *ICML* **2017**, 1-9.

[36] Mangalathu, S.; Hwang, S.; Jeon, J. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. *Eng. Str.* **2020**, *219*, 110927.

[37] Ogami, C.; Tsuji, Y.; Seki, H.; Kawano, H.; To, H.; Matsumoto, Y.; Hosono, H. An artificial neural network−pharmacokinetic model and its interpretation using Shapley additive explanations. *CPT Pharmacometrics Syst. Pharmacol*. **2021**, 760–768.

[38] Zhang, Y.; Zhang, X.; Zhu, W. ANC: Attention Network for COVID-19 Explainable Diagnosis Based on Convolutional Block Attention Module. *Comp. Mod. in Eng. & Sci*. **2021**, *127*, 1037–1058.

[39] Sharma, V.; Dyreson, C. COVID-19 Screening Using Residual Attention Network an Artificial Intelligence Approach. *arXiv:2006.16106 [eess]*, June 26, **2020**. https://arxiv.org/abs/2006.16106 (accessed 2020-07-25).

[40] Grewal, M.; Srivastava, M. M.; Kumar, P.; Varadarajan, S. RADNET: Radiologist Level Accuracy Using Deep Learning for Hemorrhage Detection in CT Scans. *arXiv.1710.04934 [cs]*, Oct. 13, **2017**. https://arxiv.org/abs/1710.04934 (accessed 2020-08-05).

[41] Yeung, C.; Tsai, J. M.; King, B.; Pham, B.; Ho, D.; Liang, J.; Knight, M. W.; Raman, A. P. Multiplexed supercell metasurface design and optimization with tandem residual networks. *Nanophotonics* **2021**, *10*, 1133–1143.

6. DeepAdjoint: All-in-One Hybrid Global Photonics Inverse Design Framework Combining Machine Learning with Electromagnetic Optimization Algorithms

6.1 Introduction

Photonic structures and materials are the driving forces for critical developments in various domains, including information technology, life sciences, and renewable energy. In particular, advancements in photonic devices have led to the creation of plasmonic waveguides for photonic integrated circuits [1,2], optical filters for spectroscopy and super-resolution imaging [3,4], and metasurfaces or metamaterials for holography and solar energy harvesting [5,6]. However, due to the rising demands in nanophotonic device performance and functionality, nanophotonics design is becoming increasingly complex and computationally intensive [7]. For example, subwavelength dielectric and metallic nanostructured materials can form complex geometric configurations that scatter, localize, and/or tailor electromagnetic fields to achieve new modalities in light-matter interactions [8]. Due to the wide choice of materials and the spatial degrees-of-freedom available for design however, the solution space is highly nonlinear and non-convex (*i.e.;* contains many local optima), and thus extremely challenging and time-consuming to solve even for experienced researchers.

As a result, over the past several years, machine learning (ML) or deep learning methods based on neural networks have demonstrated tremendous contributions towards addressing the aforementioned challenges in photonics design. Neural networks are capable of capturing, interpolating, and optimizing nonlinear data-based and physics-based relationships, including those found in nanophotonic systems. Neural networks achieve such capabilities by building an implicit relationship between input and output responses, which for nanophotonics inverse design are the optical responses and geometric/material parameters, respectively. A trained neural network is orders of magnitude faster than typical full-wave simulations and can generate non-intuitive physical structures in response to desired optical properties [9]. Accordingly, a substantial number of studies have employed neural networks for designing a broad range of photonic systems, including: metasurfaces [10,11,36], photonic crystals [12,13], and plasmonic nanostructures [14,15]. However, despite numerous advancements, it is well-known that neural networks cannot generalize too far beyond the information available in the training dataset [16-18]. Due to these limitations, hybrid algorithms combining deep learning and conventional optimization methods have emerged as a new class of efficient inverse design methodologies [19,20].

Recent studies have combined different types of neural networks and optimization schemes for photonics inverse design. Early works have paired artificial neural networks (ANN) with particle swarm optimization (PSO) [21] and evolutionary algorithms [22,23]. These studies successfully showed that the ANN can perform a rough estimate of the desired solution (*i.e.;* a global search), while the iterative optimization algorithm carries out an additional refinement step (*i.e.;* a local search). Since conventional optimization algorithms need an ideal initial condition in order to obtain the optimal result, and the neural network is restricted by its training data, the combination of both techniques can simultaneously overcome their individual limitations [20,24]. Recent hybrid ML-optimization approaches have also employed more advanced neural networks and optimization algorithms. For example, generative adversarial networks (GANs) [25] and variational autoencoders (VAEs) [26] were used together with the adjoint variables method for photonic design. Adjoint-based optimization is one of the most widely-used algorithms for photonics inverse design because regardless of the number of elements in the design space, the algorithm can determine the shape or topology gradient using only a forward and adjoint (time-reversed) simulation at each iteration [27-29]. GANs and VAEs can design complex topological structures, while the adjoint method can efficiently push performance further [30]. Additionally, generative models trained on physics-informed losses (or using the adjoint method within the training process) have also benefited from a subsequent optimization-based refinement step [31]. Thus, a number of deep learning models have been trained across various photonic device types, and a precedent has been established for hybrid ML-optimization algorithms as the next generation of inverse design methods. However, an ML-optimization strategy that simultaneously optimizes across materials and geometries has yet to be realized, and the integration between neural networks and optimization algorithms typically involve elaborate and highly-specialized procedures. For example, to establish a link between neural networks and conventional optimization methods, intermediate steps are required to introduce robustness, convert file formats, and/or to ensure that the network outputs can be adapted to the algorithm of interest.

To streamline the ML-optimization process, here we introduce an “all-in-one” global inverse design application framework which seamlessly combines generative networks with adjoint-based optimization algorithms to simultaneously optimize across materials and geometries. “Global” in this context refers to the network’s ability to perform a global search within the surveyed design space, which includes material properties and freeform topology, but the network does not guarantee that the final generated device is globally optimal. Schematically illustrated in Figure 1, our framework, DeepAdjoint, allows a researcher to specify an arbitrary spectral target (labeled “1” in Figure 1) and pass the target directly into a pre-trained generative network (whether the network is data-driven, physics-driven, or a combination thereof). We note that the increasing number of deep learning models being generated for photonics design (which we expect will continue to grow exponentially in the near future) reinforces the need for a design process that can integrate pre-trained models, particularly when practices such as network sharing and model serving are expanding within the machine learning community [32-34]. Accordingly, as a proof of concept, we employed a global inverse design GAN model with the ability to simultaneously predict device class, material properties (*e.g.;* refractive index and Drude plasma frequency), and nanoscale geometric structuring (including planar topology and layer thickness) for metal-insulator-metal (MIM) metasurfaces [9]. After passing the target into the GAN (labeled “2” in Figure 1), DeepAdjoint then validates the GAN-generated design using full-wave numerical simulations (labeled “3” in Figure 1). As a default simulation tool, DeepAdjoint integrates directly with a commercial finite-difference time-domain solver (Lumerical FDTD). The GAN-generated design can then be further augmented by converting the design into an adjoint optimization procedure (labeled “4” in Figure 1), after which the final design can yield even greater accuracy or performance by extending beyond the model’s limitations (labeled “5” in Figure 1). We demonstrate this end-to-end workflow for a range of optical device targets, including single- and multi-resonance responses, for infrared-controlled MIM metasurfaces.

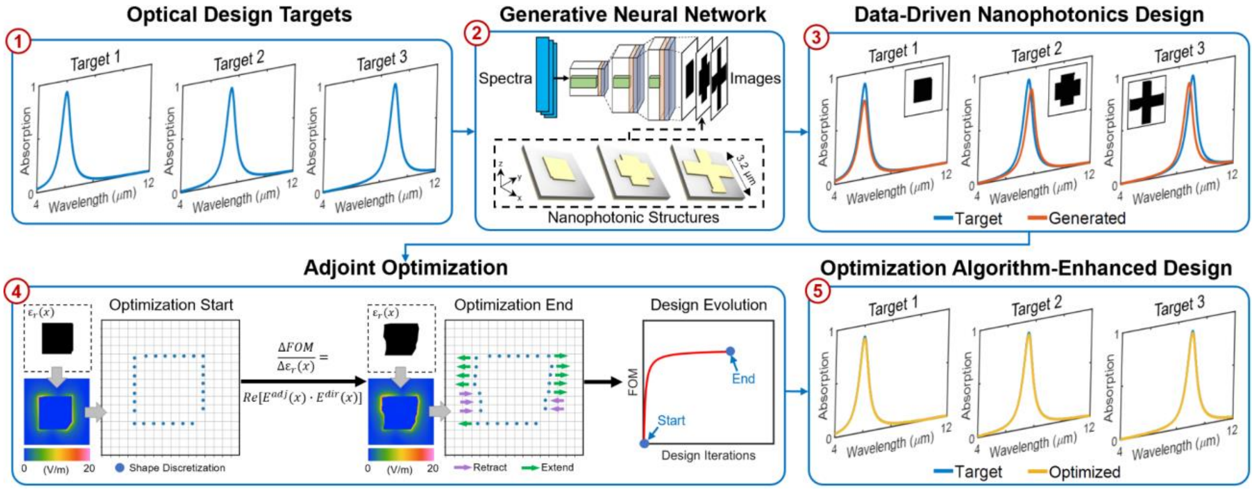


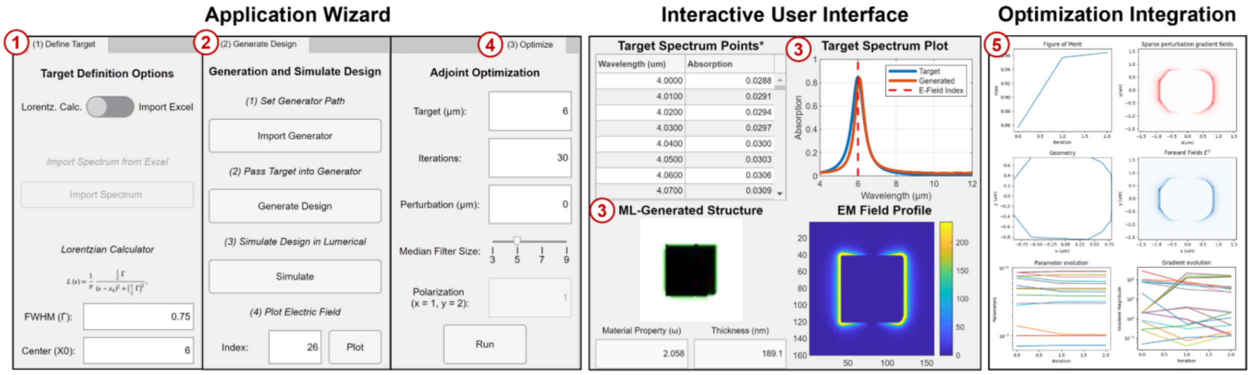
Figure 1. DeepAdjoint: an photonics inverse design framework schematic combining deep learning and adjoint optimization. (1) An arbitrary optical design target can be specified and (2) passed into a pre-trained neural network to generate a nanophotonic structure that is (3) validated

using full-wave FDTD simulations. (4) The network's design can then be automatically converted into an adjoint optimization procedure, which can (5) yield device accuracy or performance that extends beyond the network's potential limitations.

To democratize the hybridization of deep learning with electromagnetic optimization, and to make our framework easily-accessible to a wide range of practitioners, we deployed and packaged DeepAdjoint as a standalone application with a user-guided interface. Figure 2 presents the details of the application, where each step in Figure 1 can be executed (and the results can be observed) within a single user-friendly environment. As an example step-by-step procedure for designing MIM metasurfaces, DeepAdjoint first allows the user to specify an input target absorption spectrum (labeled "1" in Figure 2). Here, a Lorentzian function with a center wavelength of 6 μm and full width half maximum (FWHM) of 0.75 μm is defined and shown within the built-in visualization tool (blue curve). Next, the user simply imports the generative model (here, the aforementioned conditional GAN [9]), then generates the design (in ~500 ms) with a single button press at the step labeled "2" in Figure 2. We note that the direct output of the GAN is a set of matrix values and must be converted into a simulation model for numerical analysis. Accordingly, with the press of another button, DeepAdjoint converts the GAN's output into an FDTD model of the metasurface, runs the simulation, then reports the results back into the user interface for comparison (labeled "3" in Figure 2). Following this step, the FDTD-validated absorption spectrum (orange curve) and corresponding electric field profiles can be observed directly on the application interface.

Next, the GAN's design can be enhanced by applying the adjoint optimization method (labeled "4" in Figure 2), where an optimization target wavelength can be specified that the algorithm aims to maximize. To execute the adjoint optimization procedure, we implemented a customized version of the LumOpt module [37] (a Python wrapper for Lumerical FDTD). In this particular implementation, a number of enhancements were made to the base module in order to support free-space reflective metasurface design and optimization, which we summarize in Figure S3 of the Supporting Information. We also note that our current demonstration of DeepAdjoint leverages a deep learning model which is trained on polarization-dependent designs. Accordingly, the following adjoint-optimized structures are optimized specifically for single-polarization performance at normal incidence. However, the presented methodology is amenable to polarization-independent structures as well – should the integrated network be trained with the corresponding designs.

To configure the GAN's design for adjoint optimization, an automatic multistep process is performed (shown in Figure S1 of the Supporting Information), where the GAN's output is refined (by removing voids and defects) and transformed into a set of discretized polygon points at the meta-atom or resonator boundary. In doing so, the polygon points (*i.e.;* optimizable parameters) are compatible with the adjoint shape optimization process. Then, as the adjoint optimization iteratively progresses, the coordinates of the polygon points gradually change in the direction of figure-of-merit (FOM) improvements (labeled "5" in Figure 2). Moreover, since the presented metasurface designs operate in a reflective manner at normal incidence, the typical forward and adjoint simulations required are identical here and can be reduced to a single simulation at each iteration. Thus, we note that our particular implementation of the adjoint optimization algorithm has increased computational efficiency for metasurface design. Additionally, our framework allows the user to specify minimum feature sizes and fabrication tolerances without sacrificing device performance (shown in Figure S2 of the Supporting Information).



(1) Define Target

Target Definition Options

Lorentzian Calculator

$$L(x) = \frac{1}{\pi}\frac{\frac{1}{2}\Gamma}{(x-x_0)^2+(\frac{1}{2}\Gamma)^2}$$

FWHM (Γ): 0.75

Center (X0): 6

(2) Generate Design

Generation and Simulate Design

(1) Set Generator Path

(2) Pass Target into Generator

(3) Simulate Design in Lumerical

(4) Plot Electric Field

Index: 26

(3) Optimize

Adjoint Optimization

Target (μm): 6

Iterations: 30

Perturbation (μm): 0

Median Filter Size: 3 5 7 9

Polarization (x = 1, y = 2): 1

Target Spectrum Points*

Wavelength (um)	Absorption
4.0000	0.0288
4.0100	0.0291
4.0200	0.0294
4.0300	0.0297
4.0400	0.0300
4.0500	0.0303
4.0600	0.0306
4.0700	0.0309

Target Spectrum Plot

ML-Generated Structure

Material Property (ω): 2.058

Thickness (nm): 189.1

Figure 2. DeepAdjoint application interface and step-by-step workflow. Users can (1) define targets, (2) generate designs, (3) validate designs, (4) run adjoint optimizations, and (5) monitor optimization results.

6.2 Results and Discussion

To highlight the advantages of the proposed framework, we first investigated the performance of the adjoint optimization algorithm in relation to the algorithm's initial designs (for the particular MIM structure design space we evaluated in this work). In Figure 3, three adjoint optimization runs were executed at three different target wavelengths (indicated by the dashed red lines): 6, 7, and 8 μm, which are presented in Figures 3a, 3b, and 3c, respectively. Each optimization was performed using randomized starting designs (orange lines), and the objective was to maximize horizontal polarization (θ=0) absorption at the designated target wavelength. At the end of the optimization runs, we observe that the final designs (yellow lines) typically exhibited higher absorption values/peaks than the initial designs. Center and right columns of Figure 3 show symmetric and asymmetric starting designs, respectively. We note that different starting designs yielded different degrees of performance improvements (*i.e.;* different absorption peak amplitudes). Moreover, it can be observed that several optimized designs possess extra absorption peaks (beyond the target wavelengths) that were originally unintended. In several instances, as shown in the left column of Figure 3, a poor starting design can also cause the adjoint optimization to fail entirely by not finding any improvements to the initial structure. Thus, a deep learning algorithm that can provide the optimization with an ideal starting design would not only save computation time, but also allow the optimization to succeed and reach an optimal solution without any excess optical behaviors.

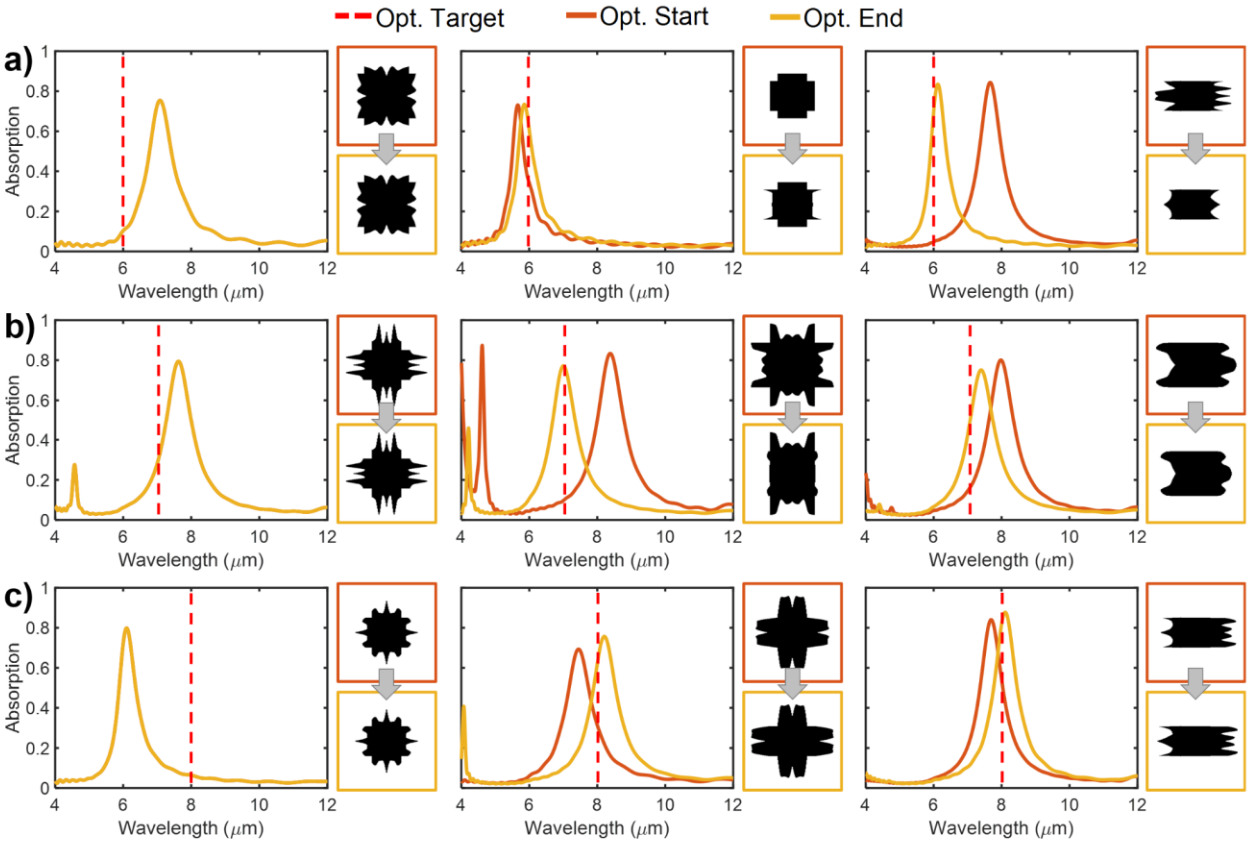


Figure 3. Metasurface designs created via adjoint optimization with randomized initial designs (orange lines). Optimization objectives include maximizing horizontal polarization (θ=0) absorption at (a) 6, (b) 7, and (c) 8 μm target wavelengths (red dashed lines). Optimized structures and corresponding spectra (yellow lines) possess various degrees of performance improvements and several extra unintended absorption peaks due to the random, suboptimal nature of the starting designs. Left column shows instances where the adjoint optimization fails to improve the starting design. Center and right columns show symmetric and asymmetric starting designs, respectively.

Next, we applied our DeepAdjoint framework to the optimization of metasurfaces with single-resonance absorption peaks. Figure 4 presents a series of optimized designs, generated through DeepAdjoint, using a range of input absorption spectra (blue lines) with “hand-drawn” Lorentzian-shaped peaks from 5 to 9 μm. Here, we observe that the GAN’s designs and simulated spectra (orange lines) are close matches to the input targets. However, several designs possess off-centered peaks or lower amplitudes in comparison to the original target. After using the GAN-generated designs as the starting points for subsequent adjoint optimization runs (with the optimization targets marked by the red dashed lines), it can be observed that the off-centered peaks are rectified and the low-amplitude peaks are increased by up to 75% (compared to the starting spectra). Moreover, the final absorption peaks of DeepAdjoint’s designs are approximately 10% higher than those achieved through the adjoint optimization algorithm alone with random starting designs (from Figure 3). Therefore, for conditional photonics inverse design with a wide range of input targets, we demonstrate that the hybridization of generative networks with the adjoint optimization algorithm offers a number of advantages, including: superior device performance in comparison to each standalone method, increased computational efficiency, and eliminating reliance on randomized starting designs.

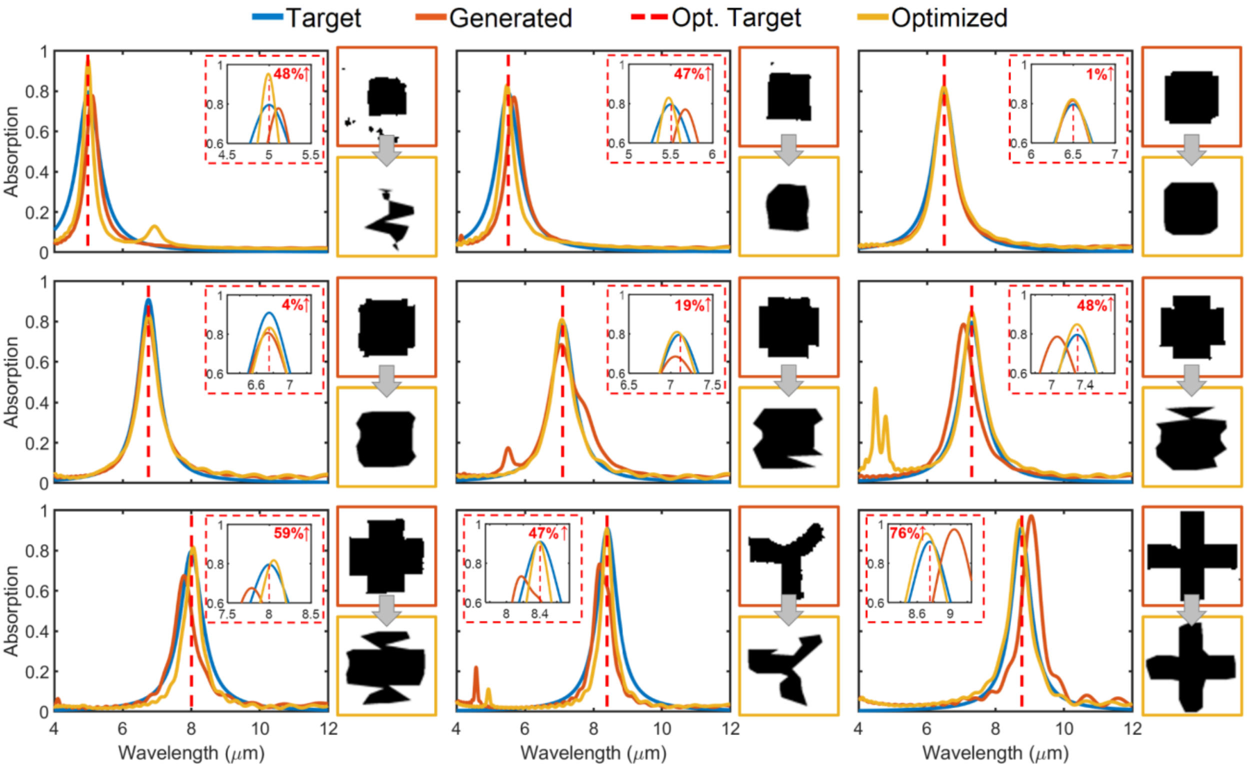


Figure 4. Single-objective metasurface designs (one absorption peak) created via DeepAdjoint. Target spectra (blue lines) are passed into the generative model (GAN) to produce starting designs (orange lines) for adjoint optimization (red dashed lines). Optimized designs (yellow lines) exhibit up to 75% performance enhancements in comparison to GAN-generated designs (shown in the inset images), indicating the hybrid approach exceeds the performance of each individual method.

Because meta-structures with simple, single-resonator periodic unit cells may only offer limited capabilities [23], we next demonstrate the versatility of our ML-optimization framework by applying it to multi-objective supercell design, where the goal is to design compound meta-atoms with multiple resonant behaviors. We note that designing such supercell structures is particularly challenging using conventional approaches, since adjacent elements may exhibit coupling and interference [35]. Furthermore, the increased number of parameters in the supercell naturally results in additional optimization complexity and computational costs. Accordingly, using DeepAdjoint, we address these challenges by first specifying the individual target resonance peaks within the supercell structure (as shown in the blue lines of Figure 5). This in turn generates the individual unit cells which contribute to the target absorption peaks (as previously demonstrated). When the individual unit cells are merged into supercell structures, it can be observed that the final structures (orange lines) produce fairly close matches in comparison to the input targets. However, compared to the single unit cell designs, the supercells have lower absorption peaks as a result of cross-element coupling. Thus, designing a supercell is not as simple as generating and combining the individual components, though this can provide a decent approximation. In this regard, a multi-objective adjoint optimization procedure can be applied to the generated supercell structures, which simultaneously maximizes multiple absorption peaks while accounting for the optical behaviors produced by the entire supercell (including cross-element coupling).

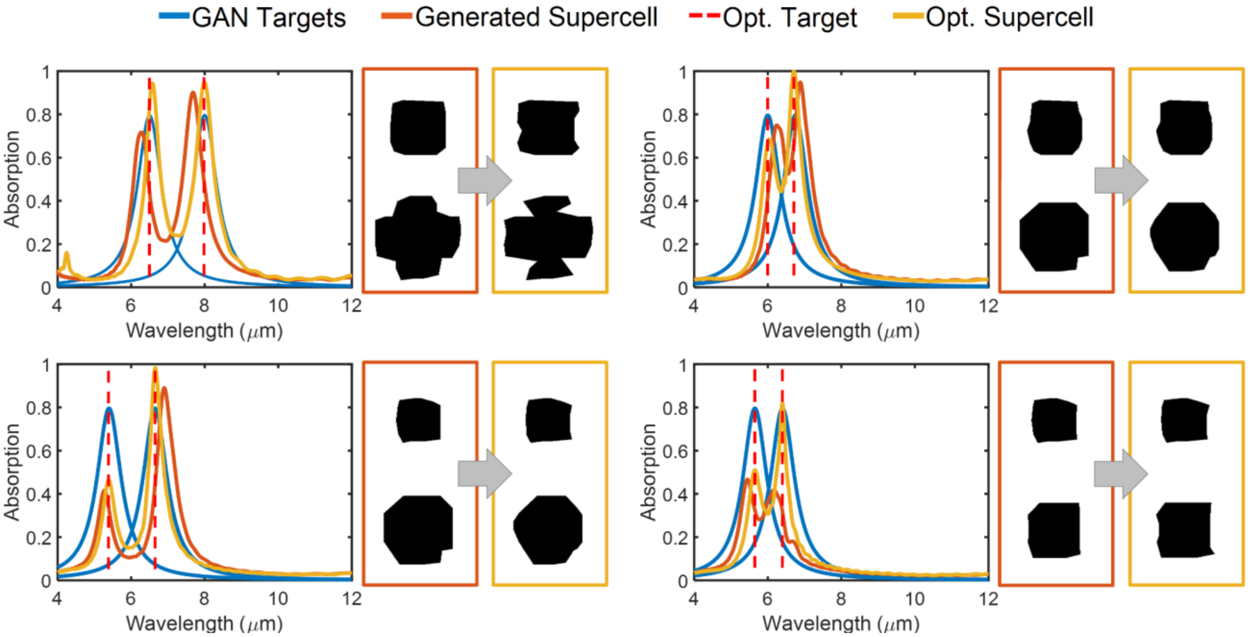


Figure 5. Multi-objective metasurface designs (multiple absorption peaks) created via DeepAdjoint. Target spectra (blue lines) are passed into the generative model (GAN) to produce starting designs (orange lines) for adjoint optimization (red dashed lines). Optimized designs (yellow lines) exhibit up to 50% performance enhancements in comparison to GAN-generated supercell designs (shown in the inset images).

In Figure 5, the results of multiple supercell optimization runs are presented (at the optimization targets indicated by the red dashed lines). Here, we observe that the optimized supercells (yellow lines) all yield up to 50% higher absorption peaks than the initial designs, though the degree of absorption enhancement appears to be peak-dependent (possibly due to different coupling mechanisms induced by particular elements). In addition to increasing the target absorption peaks, Figure 5 also shows that the adjoint optimization procedure can rectify or recenter off-target peaks within the supercell. As a result, we show that our hybrid ML-optimization framework can be used to design and achieve a wide range of optical behaviors, including periodic unit cell structures with single resonances and complex supercell structures with multiple resonances or broadband characteristics.

6.3 Conclusions

In summary, we present DeepAdjoint, a general-purpose, open-source, and multi-objective "all-in-one" global photonics inverse design application framework which integrates pre-trained deep generative networks with state-of-the-art electromagnetic optimization algorithms such as the adjoint variables method. DeepAdjoint allows a designer to specify an arbitrary optical design target, then obtain a photonic structure that is robust to fabrication tolerances and possesses the sought optical properties – all within a single user-guided workflow and application interface. As a proof of concept, we demonstrated our framework for the design and optimization of infrared-controlled metasurfaces, and showed that a wide range of structures and absorption spectra can be achieved, including single- and multi-resonance behavior through single- and supercell-class structures, respectively. By specifying an input target spectrum, a global inverse design generative neural network serves as a rapid global approximation search step (~500 ms) and produces a nanophotonic structure with material properties, layer thicknesses, and planar geometry defined. Afterwards, the generated design can be sent through an adjoint optimization procedure, which serves as a local search step to increase performance further. As a result, the limitations of training data restriction and starting point dependency for deep learning and conventional optimization, respectively, can be simultaneously overcome. Our proposed framework is thus an important step towards the systematic unification of machine learning and optimization algorithms for photonics inverse design. Original contributions of our work include: a streamlined framework for integrating neural networks with conventional optimization algorithms, algorithmic improvements to the adjoint method to enhance computational efficiency for free-space metasurfaces, a demonstration of multi-objective supercell design using ML-optimization techniques, minimum feature size control, and a user-friendly application interface for non-experts and multidisciplinary research.

6.4 Supporting Information

Generative Adversarial Network (GAN) to Adjoint Optimization Configuration

In this work, we show that a GAN's design can be enhanced through the application of the adjoint variables method. To easily-support the adjoint variables method for shape optimization, which requires the calculation of a structure's shape derivative, the shape can be described by a 2D array of polygon points that are sorted in a counterclockwise fashion. Accordingly, we implemented an automatic multistep process (shown in Figure S1) where the GAN's image-formatted output data is refined and transformed into a set of discretized polygon points. In this process, a median filter is first applied to the GAN's output shape to remove voids, defects, and image artifacts. Then, the Canny edge detection algorithm is used to extract the meta-atom or resonator boundary. Afterwards, Hough line transformation is applied to the extracted edge in order to convert the resonator boundary into line coordinates, which we then interpolate to produce the required polygon points for shape optimization. We note that the interpolation spacing between the polygon points can be user-specified (and is 100 nm by default), which determines the resolution of the adjoint-optimized nanostructure.

Minimum Feature Size Analysis

During the GAN to adjoint optimization configuration process, the median filter can serve as a means to remove image artifacts and to specify minimum feature sizes. In our implementation of DeepAdjoint, we set the size of the filter as a user-defined value, such that a user can enhance the fabricability of the ML-optimization generated designs. In Figure S2, we show that the minimum feature size of 50 nm can be increased to 250 nm without significantly affecting the absorption spectra of the GAN-generated designs. As a result, designers can ensure their designs are robust to fabrication tolerances without having to retrain the entire model, although retraining a model using larger feature sizes is also a valid option that is supported by the model-loading capabilities of DeepAdjoint.

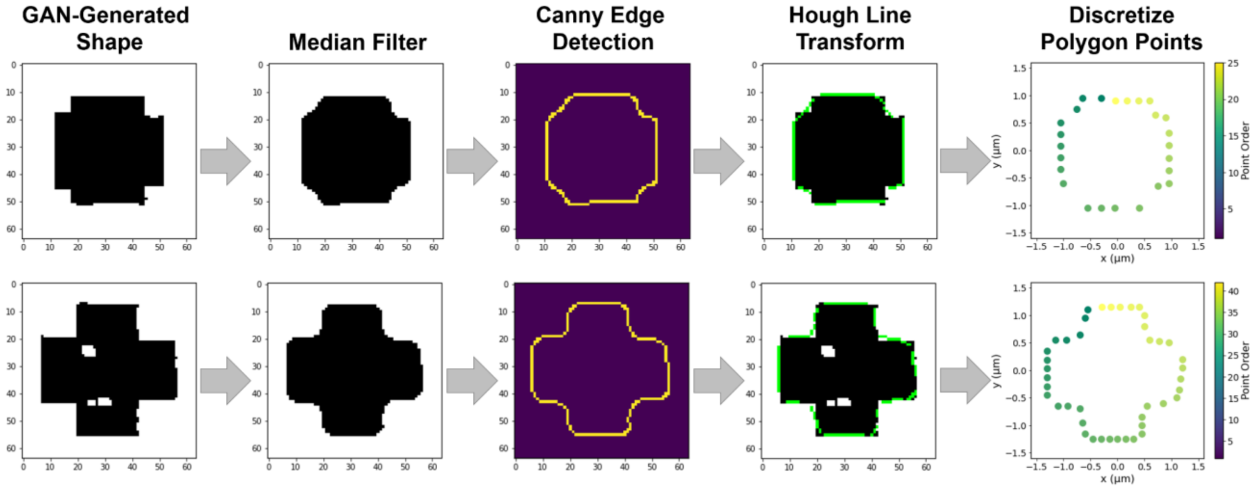


Figure S1. GAN to adjoint optimization configuration procedure. The GAN's output image is filtered, then passed through the Canny edge detection and Hough transform algorithms, to produce a set of discretized polygon points that are amenable to adjoint shape optimization.

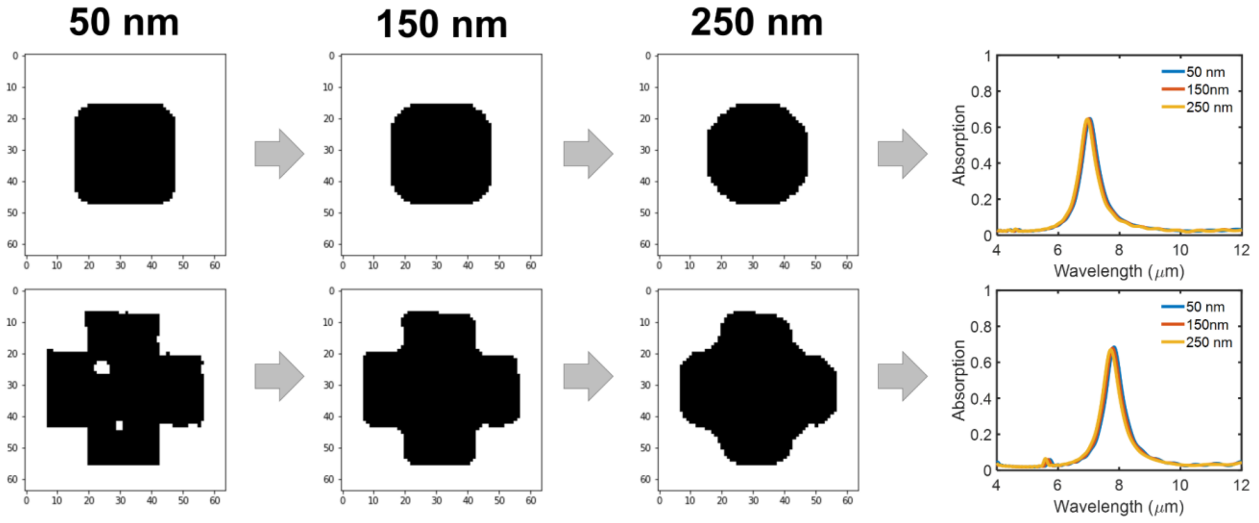
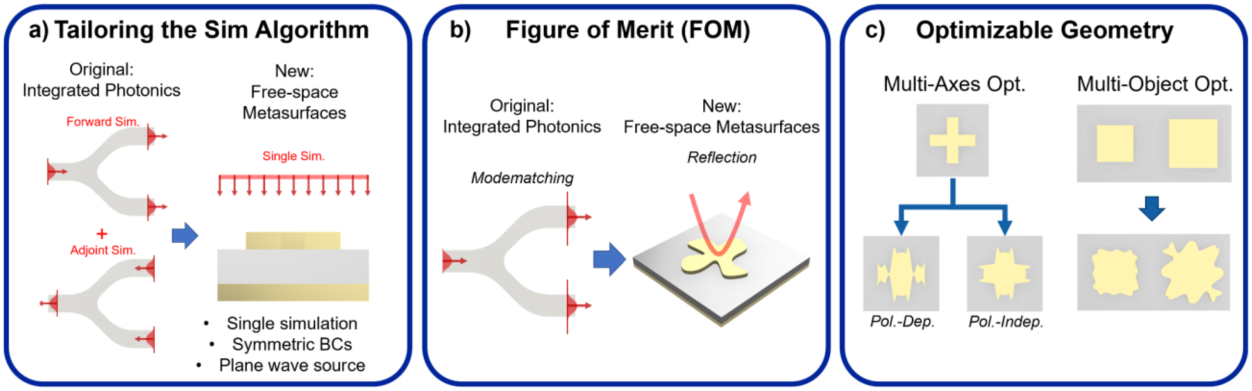


Figure S2. Effect of changing the median filter or minimum feature size. The minimum feature size (50 nm default) can be set to 250 nm without significantly affecting the absorption spectra.

Adjoint Optimization Implementation Details and Features

In our particular implementation of the adjoint method, we employed a customized version of a widely-used continuous adjoint optimization Python wrapper for FDTD simulations (LumOpt; referenced in the main manuscript). Here, we note several key enhancements and modifications that were made to the base LumOpt module (which was primarily designed for integrated photonics) in order to enable free-space metasurface design and optimization. First, as shown in Figure S3a, we tailored the simulation setup and boundary conditions by applying a plane wave source instead of a mode source. In addition, since the presented metasurface designs operate in a reflective manner at normal incidence, the typical forward and adjoint simulations required are identical and were reduced to a single simulation at each iteration. Thus, we note that our particular implementation of the adjoint optimization algorithm has increased computational efficiency for metasurface design. Symmetric boundary conditions across the x- and y-planes (one plane for two-fold symmetry or two planes for four-fold symmetric) are also supported to further reduce simulation time.



- Single simulation
- Symmetric BCs
- Plane wave source

Figure S3. Custom implementation of the adjoint method based on numerous enhancements and modifications to the LumOpt module. Enhancements include: (a) changes to the simulation setup and boundary conditions to support free-space metasurface optimization, (b) new FOM definitions to capture metasurface reflectivity and other spectral properties, and (c) several user-defined features to enable more degrees of freedom for the optimizable geometry.

Additional modifications were made to the FOM specification, shown in Figure S3b, to capture metasurface reflection through a plane in free-space (and absorption by extension) as an optimization metric instead of modematching to a fundamental TE or TM mode in the original

LumOpt implementation. Lastly, a number of new user-defined features were incorporated to enable more degrees of freedom for the optimizable geometry (shown in Figure S3c). These new features include: multi-axis optimization for both polarization-dependent and polarization-independent designs, and the ability to simultaneously optimize multiple objects to facilitate supercell structures.

6.5 References

[1] Meng, Y.; Chen, Y.; Lu, L.; Ding, Y.; Cusano, A.; Fan, J. A.; Hu, Q.; Wang, K.; Xie, Z.; Liu, Z.; Yang, Y.; Liu, Q.; Gong, M.; Xiao, Q.; Sun, S.; Zhang, M.; Yuan, X.; Ni, X. Optical meta-waveguides for integrated photonics and beyond. *Light: Science Applications* **2021**, *10*, 1–44.

[2] Fang, Y.; Sun, M. Nanoplasmonic waveguides: towards applications in integrated nanophotonic circuits. *Light: Science Applications* **2015**, *4* (6), 294.

[3] Zhao, H.; Xie, J.; Liu, J. Optical and acoustic super-resolution imaging in a Stampfli-type photonic quasi-crystal flat lens. *Results in Physics* **2021**, *27*, 104537.

[4] Dhama, R.; Yan, B.; Palego, C.; Wang, Z. Super-Resolution Imaging by Dielectric Superlenses: TiO2 Metamaterial Superlens versus BaTiO3 Superlens. *Photonics* **2021**, *8* (6), 222.

[5] Zhang, Q.; Liu, X.; Chaker, M.; Ma, D. Advancing Graphitic Carbon Nitride-Based Photocatalysts toward Broadband Solar Energy Harvesting. *ACS Materials Letters* **2021**, *3* (6), 663–697.

[6] Ahmed, S.; Li, Z.; Javed, M. S.; Ma, T. A review on the integration of radiative cooling and solar energy harvesting. *Materials Today Energy* **2021**, *21*, 100776.

[7] Hegde, R. S. Deep learning: a new tool for photonic nanostructure design. *Nanoscale Advance* **2020**, *2* (3), 1007–1023.

[8] Jiang, J.; Chen, M.; Fan, J. A. Deep neural networks for the evaluation and design of photonic devices. *Nature Reviews Materials* **2020** *6* (8), 679–700.

[9] Yeung, C.; Tsai, R.; Pham, B.; King, B.; Kawagoe, Y.; Ho, D.; Liang, J.; Knight, M. W.; Raman, A. P. Global Inverse Design across Multiple Photonic Structure Classes Using Generative Deep Learning. *Advanced Optical Materials* **2021**, *9* (20), 2100548.

[10] Jiang, L.; Li, X.; Wu, Q.; Wang, L.; Gao, L. Neural network enabled metasurface design for phase manipulation. *Optics Express* **2021**, *29* (2), 2521–2528.

[11] Ghorbani, F.; Beyraghi, S.; Shabanpour, J.; Oraizi, H.; Soleimani, H.; Soleimani, M. Deep neural network-based automatic metasurface design with a wide frequency range. *Scientific Reports* **2021**, *11* (1), 1–8.

[12] Zhan, T.; Liu, Q. S.; Sun, Y. J.; Qiu, L.; Wen, T.; Zhang, R. A general machine learning-based approach for inverse design of one-dimensional photonic crystals toward targeted visible light reflection spectrum. *Optics Communications* **2022**, *510*, 127920.

[13] Qiu, C.; Wu, X.; Luo, Z.; Yang, H.; Wang, G.; Liu, N.; Huang, B. Simultaneous inverse design continuous and discrete parameters of nanophotonic structures via back-propagation inverse neural network. *Optics Communications* **2021**, *483*, 126641.

[14] Wu, Q.; Li, X.; Jiang, L.; Xu, X.; Fang, D.; Zhang, J.; Song, C.; Yu, Z.; Wang, L.; Gao, L. Deep neural network for designing near- and far-field properties in plasmonic antennas. *Optical Materials Express* **2021**, *11* (7), 1907–1917.

[15] Verma, S.; Chugh, S.; Ghosh, S.; Azizur Rahman, B. M. Artificial Neural Network Modeling for Optimizing the Optical Parameters of Plasmonic Paired Nanostructures. *Nanomaterials* **2022**, *12* (1), 170.

[16] Novak, R.; Bahri, Y.; Abolafia, D. A.; Pennington, J.; Sohl-Dickstein, J. Sensitivity and Generalization in Neural Networks: An Empirical Study *arXiv* **2018**.

[17] Fort, S.; Nowak, P. K.; Jastrzebski, S.; Narayanan, S. Stiffness: A New Perspective on Generalization in Neural Networks, *arXiv* **2020**.

[18] Lenaerts, J.; Pinson, H.; Ginis, V. Artificial neural networks for inverse design of resonant nanophotonic components with oscillatory loss landscapes. *Nanophotonics* **2020**, *10* (1), 385–392.

[19] Xu, Y.; Zhang, X.; Fu, Y.; Liu, Y. Interfacing photonics with artificial intelligence: an innovative design strategy for photonic structures and devices based on artificial neural networks. *Photonics Research* **2021**, *9* (4), 135–152.

[20] Wiecha, P. R.; Arbouet, A.; Girard, C.; Muskens, O. L. Deep learning in nano-photonics: inverse design and beyond. *Photonics Research* **2021**, *9* (5), B182–B200.

[21] Ma, Z.; Li, Y. U. Parameter extraction and inverse design of semiconductor lasers based on the deep learning and particle swarm optimization method. *Optics Express* **2020**, *28* (15), 21971–21981.

[22] Hegde, R. S. Photonics inverse design: Pairing deep neural networks with evolutionary algorithms. *IEEE Journal of Selected Topics in Quantum Electronics* **2020**, *26* (1).

[23] Liu, Z.; Zhu, D.; Lee, K. T.; Kim, A. S.; Raju, L.; Cai, W. Compounding Meta-Atoms into Metamolecules with Hybrid Artificial Intelligence Techniques. *Advanced Materials* **2020**, *32* (6), 1904790.

[24] Ma, W.; Liu, Z.; Kudyshev, Z. A.; Boltasseva, A.; Cai, W.; Liu, Y. Deep learning for the design of photonic structures. *Nature Photonics* **2020**, *15* (2), 77–90.

[25] Jiang, J.; Sell, D.; Hoyer, S.; Hickey, J.; Yang, J.; Fan, J. A. Free-form diffractive metagrating design based on generative adversarial networks. *ACS Nano* **2019**, *13* (8), 8872–8878.

[26] Kudyshev, Z. A.; Kildishev, A. v.; Shalaev, V. M.; Boltasseva, A. Machine learning-assisted global optimization of photonic devices. *Nanophotonics* **2020**, *10* (1), 371–383.

[27] Piggott, A. Y.; Petykiewicz, J.; Su, L.; Vučković, J. Fabrication-constrained nanophotonic inverse design. *Scientific Reports* **2017**, *7* (1), 1–7.

[28] Mansouree, M.; McClung, A.; Samudrala, S.; Arbabi, A. Large-Scale Parametrized Metasurface Design Using Adjoint Optimization. *ACS Photonics* **2021**, *8* (2), 455–463.

[29] Molesky, S.; Lin, Z.; Piggott, A. Y.; Jin, W.; Vucković, J.; Rodriguez, A. W. Inverse design in nanophotonics. *Nature Photonics* **2018**, *12* (11), 659–670.

[30] Wang, K.; Ren, X.; Chang, W.; Lu, L.; Liu, D.; Zhang, A. M. Inverse design of digital nanophotonic devices using the adjoint method. *Photonics Research* **2020**, *8* (4), 528–533.

[31] Hooten, S.; Beausoleil, R. G.; van Vaerenbergh, T. Inverse design of grating couplers using the policy gradient method from reinforcement learning. *Nanophotonics* **2021**, *10* (15), 3843–3856.

[32] Wu, H.; Wang, C.; Yin, J.; Lu, K.; Zhu, L. *Sharing Deep Neural Network Models with Interpretation* **2018**, 11.

[33] Kaissis, G.; Ziller, A.; Passerat-Palmbach, J.; Ryffel, T.; Usynin, D.; Trask, A.; Lima, I.; Mancuso, J.; Jungmann, F.; Steinborn, M. M.; Saleh, A.; Makowski, M.; Rueckert, D.; Braren, R. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence* **2021**, *3* (6), 473–484.

[34] Olston, C.; Fiedel, N.; Gorovoy, K.; Harmsen, J.; Lao, L.; Li, F.; Rajashekhar, V.; Ramesh, S.; Soyke, J. TensorFlow-Serving: Flexible, High-Performance ML Serving *arXiv* **2022**.

[35] Yeung, C.; Tsai, J. M.; King, B.; Pham, B.; Ho, D.; Liang, J.; Knight, M. W.; Raman, A. P. Multiplexed supercell metasurface design and optimization with tandem residual networks. *Nanophotonics* **2021**, *10* (3), 1133–1143.

[36] Yeung, C.; Tsai, J. M.; King, B.; Kawagoe, Y.; Ho, D.; Knight, M. W.; Raman, A. P. Elucidating the Behavior of Nanophotonic Structures through Explainable Machine Learning Algorithms. *ACS Photonics* **2020**.

[37] Lalau-Keraly, C. M.; Yablonovitch, E.; Miller, O. D.; Bhargava, S. Adjoint shape optimization applied to electromagnetic design. *Optics Express* **2013**, *21* (18), 21693–21701.

7. Conclusion and Future Work

In conclusion, deep learning, or more generally machine learning and artificial intelligence, possesses the ability to design and characterize nanophotonic materials and structures with remarkable accuracy and precision. In this work, convolutional neural networks, tandem neural networks, generative adversarial networks, and explainable AI were demonstrated as capable methods and tools for the forward and inverse design of nanophotonic materials and structures. Furthermore, hybridized methods can be employed to bridge the limitations of data-driven machine learning. In particular, cDCGANs can generate arbitrarily-defined optical targets, then pair with electromagnetic optimization algorithms to perform high-performance local searches. In further synergy between ML with conventional optimization techniques, XAI can be applied to the adjoint optimization algorithm to elucidate the limitations of the algorithm itself. It is worth noting that in a number of the presented works, a majority of the materials property remained fixed while (for fabricability purposes) the primary optimizable parameters are physical structuring (particularly in Chapters 3, 4, 5, and 6). However, the presented methods and approaches can easily be generalized or adapted to capture additional material parameters or properties so long as the corresponding information can be represented within the training dataset (as demonstrated in Chapter 2).

Despite significant progress in this field, a number of challenges still remain. Specifically, machine learning algorithms require significant amounts of data to succeed, which may prohibit the practical application of ML towards solving distinct, “one-time” problems. In addition, though an ML model may predict an optimal or novel design for a particular application, the design must be fabricable using large-scale, economic-friendly methods. In this regard, physics-enhanced loss functions and neural networks, as well as exploration-based ML methods such as reinforcement learning, serve as promising candidates as the next generation of ML techniques that can reduce training data dependence. Hybrid methods combining ML with analytical methods and optimization algorithms also hold great potential in improving model generalizability and diversity of application, where the strengths of particular techniques can be used to address the weaknesses of others. General speaking, there is a common misconception that ML can be applied to any arbitrary problem and is the best tool for solving any materials-related optimization problem. However, this is far from the truth and the nature of the problem must be carefully considered before identifying which ML method to apply, or whether ML should be applied at all. To this end, one must be cognizant of the strengths and limitations of ML, such that improved methods and techniques may be developed in the near future which can accelerate the development and discovery of novel materials and structures in a wide range of physical systems.