Elucidating the Design and Behavior of Nanophotonic Structures through Explainable Convolutional Neural Networks

Christopher Yeung1†, Ju-Ming Tsai1†, Yusaku Kawagoe1,2, Brian King1, David Ho1, and Aaswath P. Raman1*

1 Department of Materials Science and Engineering, University of California, Los Angeles, CA
2 Department of Mechanical Engineering, Nagoya University, Nagoya, Japan

† These authors contributed equally to this work.
* Corresponding Author: Aaswath P. Raman (aaswath@ucla.edu)

Abstract

A central challenge in the development of nanophotonic structures and metamaterials is identifying the optimal design for a sought target functionality, and understanding the physical mechanisms that enable the optimized device’s capabilities. To this end, previously investigated design methods for nanophotonic structures have encompassed both conventional forward and inverse optimization approaches as well as nascent machine learning (ML) strategies. While in principle more computationally efficient than optimization processes, ML-based methods that are capable of generating complex nanophotonic structures are still ‘black boxes’ that lack explanations for their predictions. Motivated by this challenge, in this article we demonstrate that convolutional neural networks (CNN) trained to be highly accurate at forward design, can be explained to derive physics-driven insights by revealing the underlying light-matter relationships learned by network. We trained a CNN model with 10,000 images representative of a class of metal-dielectric-metal metamaterial resonators and their corresponding absorption spectra obtained from simulations. The trained CNN predicted the spectra of new and unknown designs with over 95% accuracy. We then applied the Shapley Additive Explanations (SHAP) algorithm to the trained model to determine features that made positive or negative contributions towards specific spectral points, thereby informing which features to create or eliminate in order to meet a new target spectrum. Our results reveal that the physical relationships between a nanophotonic structure and its electromagnetic response can be obtained — and new designs can be achieved — by exposing the valuable information hidden within a machine learning algorithm.
Introduction

Nanophotonic materials and structures have enabled a wide range of transformative technologies including photonic integrated circuits for optical communication\textsuperscript{1-3}, and metasurfaces that compactly control the propagation of electromagnetic waves\textsuperscript{4-6}. There is significant potential for further advancements in these and other applications, but current photonic design strategies remain a limiting factor. The most common approach to designing nanophotonic structures and devices is via numerical simulations based on physical laws (e.g., Maxwell’s Equations). This design technique, which we here refer to as ‘forward design’\textsuperscript{7}, is well established, but the solution space is restricted by human intuition, and requires a computationally expensive trial-and-error process to obtain target functionalities. As the functionalities being sought from photonic devices increase in complexity, so has the realization that a library of traditional template structures and design intuition will be unable to meet the demand for high performance and highly tailored functionality\textsuperscript{8}.

To address the limitations of forward design, ‘inverse design’ methods have been developed to generate nanophotonic structures that meet predefined targets\textsuperscript{8}. Methodologies such as topology\textsuperscript{9-11} and adjoint-based optimization\textsuperscript{12,13} have shown success in designing complex structures which can deliberately interact with electromagnetic fields, often at sub-wavelength scales, to enable a desired response. However, while inverse design algorithms allow for the discovery of unintuitive and sometimes intricate structures, they can potentially miss globally optimal designs\textsuperscript{7,8}, produce unstable results\textsuperscript{14}, and are often constrained by long runtimes. Additionally, conventional inverse design methods 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 an alternate strategy for the inverse design of nanophotonic devices and metamaterials. Tandem neural networks have been used to design multilayer thin films based on target transmission spectra\textsuperscript{15}, and for spatially-complex geometries, Generative Adversarial Networks (GANs) have been shown to produce images of novel structures, given an input of desired spectral properties. In particular, GANs have been used to design diffractive metagratings\textsuperscript{16}, sub-wavelength antenna\textsuperscript{17}, and two-dimensional metamaterials\textsuperscript{18} with highly tailored spectral properties. Despite recent interest in ML-based inverse design, ML-based forward design remains largely unexplored. ML-based forward design, if implemented, could offer the same advantages as its inverse design counterpart, but for characterizing the electromagnetic response of a given nanophotonic structure rather than devising structures based on a target function. Moreover, ML-based forward design holds the potential to dramatically outperform conventional electromagnetic simulation tools in terms of speed, by accurately and efficiently correlating geometric features to desired electromagnetic variables such as spectral response.

However, the internal decision models built by the machine learning algorithm for either forward- or inverse-design are not well-understood; their contents are widely regarded as ‘black boxes’\textsuperscript{19,20}. This challenge emerges from the fact that supervised ML algorithms, including neural networks, learn by optimizing hundreds of thousands to millions of internal variables (weights and biases) to fit the training data\textsuperscript{21}. Consequently, it is challenging to explain \textit{why} a machine
learning algorithm makes one prediction over another. Indeed, this lack of explainability is a key limitation for both ML-based methods as well as conventional inverse design strategies.

Motivated by these long-standing challenges, in this article, we uncover what a particular class of machine learning algorithms, convolutional neural networks (CNNs), have learned regarding the underlying physical principles which govern specific classes of nanophotonic structures when trained to accurately perform forward design. CNNs are deep neural networks widely used for image analysis and classification\textsuperscript{22}. We first demonstrate that a CNN can accurately perform forward design by learning the relationship between a two-dimensional image representing the geometry of a metal-dielectric-metal metamaterial, and its absorption spectrum over mid-infrared wavelengths. We then use the Deep SHapley Additive exPlanations (SHAP) framework\textsuperscript{24}, to explain the CNN’s predictions by calculating contribution values for each feature in the input image. These explanations reveal that the CNN has learned important physical behaviors of the class of metamaterials studied, including the relationship between resonator length and resonance wavelength. We then leverage these explanations to identify changes to the original structure that would produce a new target spectrum with a desired resonant wavelength absorption peak. Overall, the presented strategy demystifies some of the relationships that enable accurate machine-learning model predictions for photonic devices in general, while also identifying limitations in their capabilities.
Forward Design Convolutional Neural Network (CNN) Development and Evaluation

Figure 1: Utilizing a CNN’s explanations to understand a structure’s response and design transformation. (a) Schematic of the explanation and optimization approach to elucidating the underlying physics learned by a CNN. A CNN is trained on nanophotonic metal-insulator-metal structures and their corresponding absorption spectra over mid-infrared wavelengths. The relationships between structural features and absorption peaks are explained with the SHAP algorithm, and then used to construct new designs with new target resonant wavelengths. (b) The process of converting 3D models into 2D representations for the training data set.

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\textsuperscript{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. Figure 1b illustrates the process of creating training data for the CNN. The simulated structures, previously demonstrated in literature to possess selective thermal emissivity over a large bandwidth\textsuperscript{23}, consist of a 100 nm gold bottom layer, a 200 nm Al\textsubscript{2}O\textsubscript{3} dielectric middle layer, and a 100 nm gold resonator top layer (within a 3.2 \(\mu\)m \(\times\) 3.2\(\mu\)m 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 Fig. S3). 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 \(\mu\)m to 12 \(\mu\)m). Periodic boundary conditions were applied along the x- and y-planes. Each image was resized to 40 \(\times\) 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. 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 \(\times\) 3 filters, numbering in 8, 16, and 32 in each subsequent layer. The pooling layer used 2 \(\times\) 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).
Figure 2 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. Based on the comparison between the simulations and the CNN predictions, we observe that the CNN 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 $0.270 \pm 0.043$ seconds ($n = 10$), while each simulation took approximately 15 minutes. The results here demonstrate that the CNN can successfully perform forward design-based tasks with high accuracy and is orders of magnitude faster than conventional numerical simulations.

**Figure 2: Simulating nanophotonic material response with CNNs.** 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.
Explaining CNN performance with Shapley Additive Explanations (SHAP)

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. To explain the CNN’s behavior and draw useful conclusions from the network’s internal model, we utilized the Shapley Additive Explanations (SHAP) method. This empirical method calculates a ‘SHAP value’ for each pixel that provides a measure of how important that pixel is to the CNN’s overall prediction for a given output node. With the SHAP values, we can thus explain the contribution of a given geometrical 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:

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

1

where, $\Phi_i$ 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 $f_{x'}(z' \setminus i)$ is a model trained with the feature withheld. The SHAP algorithm captures the effect of withholding a feature, then iterates the computation across all possible subsets ($z' \subseteq x'$).

We performed SHAP explanations on the CNN model trained on 10,000 images of nanophotonic structures previously described. The explanation is typically represented as a heatmap, where red pixels represent positive contributions towards the model’s prediction, and
blue pixels represent negative contributions. As shown in Fig. 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 Methods section), 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 (negative contributions) 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, Fig. 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 resonant 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 weights each arm differently in determining the resonant 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, which dictates the features that contribute the least or the most to a
resonant wavelength. CNNs extract information from images by applying a hierarchy of filters to the input image\textsuperscript{28}. The filters are optimized such that the error is minimized when comparing the CNN’s output to the target output. CNNs have a tendency to develop edge detection filters, since non-edge patterns (\textit{e.g.}, a patch of black pixels) do not typically provide sufficient information to differentiate discrete objects\textsuperscript{22}. Therefore, our CNN was tasked with creating the minimum set of filters that captures the most important features and distinctions (\textit{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 address the model uncertainty by using the SHAP explanations to observe the most prominent sections of the structure contributing to resonance as well as the sections that were disregarded; two pieces of information that provide insight on what the CNN did and did not learn, respectively. 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.
Figure 3: SHAP explanations for nanophotonic structure classification. 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 resonant wavelengths, revealing the CNN learned that the cross-arms must increase to achieve resonance at longer wavelengths and vice versa.

Using SHAP Explanations for Targeted Design Transformations

To further confirm the physical relationship learned by the CNN, and to assess the possibility of using the SHAP explanations for designing nanophotonic structures with alternate target functionalities, we used the SHAP value heatmaps from the previous section to transform the base image to generate resonances at alternate wavelengths. These transformed designs were then compared with the corresponding FDTD simulated background images (as shown in Fig.
S1a) to ensure that the CNN learned the relationship between cross-arm length and resonance wavelength. Transformation was performed by converting all the blue pixels (and the first 1% of pixels greater than 0 to account for noise) in the heatmaps to black pixels on the base image. Figure S1b and S1c show the spectra of the transformed structures and the original FDTD simulated structures, respectively.

We validated the SHAP explanations, and the design transformations derived from them, against a standard antenna-based analytical relationship between the MIM resonator arm lengths and resonant wavelengths:

\[ \lambda = (2n_{\text{eff}}) L + C. \]  

Here, \( \lambda \) is the resonant wavelength, \( n_{\text{eff}} \) is the effective index of the transverse electric (TE) mode, \( L \) is the length of the resonator, and \( C \) is a correction phase term\(^{25-27} \). In Fig. S1d, 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 \( n_{\text{eff}} \) of 1.13 and \( C \) of 2.21, while the SHAP-transformed structures display an \( n_{\text{eff}} \) of 1.15 and \( C \) of 2.10, yielding an \( n_{\text{eff}} \) error of 1.8% and a \( C \) error of 4.9%. 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, and that the extracted data can be utilized to make reasonably accurate targeted design transformations.

**Deriving Physical Insights by Explaining Complex Shapes and Spectral Patterns**

To demonstrate the applicability of this explanation and design transformation method to more complex nanophotonic structures, we generated additional explanations for more complex
shapes from the previously described CNN. 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 Fig. 4a). The SHAP explanation heatmaps at the designated wavelengths are shown in Fig. 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 Fig. 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 Fig. 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.

Furthermore, the explanations give us 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 resonant wavelengths (Fig. S2). 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 5.2 µm, the center of the arm contributes more to the resonance, which aligns with the nature of the E-field distribution for this resonance (Fig. S2).

**Figure 4:** Explanations of a dual-peak structure and a single-peak structure at various wavelengths. (a) Absorption spectra of a single-peak I-shaped resonator and a dual-peak L-shaped resonator. Red circles indicate the resonant wavelengths. (b) SHAP explanations of the resonators at the previously identified resonant wavelengths. Red dashed boxes indicate the explanations for obtaining new target resonant 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.

We performed the same design transformation study from the previous section to assess its applicability to complex, multi-peak transformations. In Fig. 5a, the L-shaped structure was transformed 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 (Fig. 5b). We leveraged the explanation from one of the dual-peak wavelengths (as either wavelength resulted in negligible differences) and applied it to the design transformation process. The transformed structure possessed an absorption peak of ~0.6 at 4.8 µm and ~0.48 at 6.9 µm.

The design transformation 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. However, the resonant wavelengths of the transformed structures deviate from the target wavelengths by ~9% (relative to the evaluated wavelength range), indicating that there are limitations associated with this approach. Such limitations may be attributed to one of the most prominent shortcomings of SHAP: its inability to account for feature dependence\textsuperscript{29,30}, which may have inhibited the identification of key structural features required for resonance. Despite this discrepancy between the target and resulting resonant wavelengths, the general patterns identified by the transformed designed structures still offer significant insights into the critical features which contribute to resonance; a critical element which was not accessible in previous ML studies pertaining to nanophotonic structures.
Figure 5: Targeted design transformation for complex spectral responses. Transformation of a (a) dual-peak structure to a single-peak structure and a (b) single-peak structure to a dual-peak structure by utilizing the SHAP values at targeted resonant wavelengths for image conversion.

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. We further used an explanation algorithm (Shapley Additive Explanations, or SHAP) to identify the contributions of individual features of an image (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 (~15 minutes). By examining the SHAP explanations, both qualitative and quantitative relationships between structure and spectra can be obtained (i.e., resonator arm length vs resonant wavelength), and the explanations themselves can be used to make targeted design transformations. Importantly, the explanations also revealed what the CNN did not learn, thus exposing potential limitations and risks associated with the trained model. As a result, the presented explanation and design
transformation method shows that the patterns and principles encoded within the ML model can be extracted to derive valuable insights into nanophotonic device physics. 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 will 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.
Methods

Numerical simulations were performed in Lumerical FDTD. To generate the absorption spectra, nanophotonic structures with 3.2 μm × 3.2 μm periodic arrays were simulated. Each structure contained a 100 nm gold bottom layer, a 200 nm Al₂O₃ 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.

The CNN was implemented using TensorFlow and Keras, and trained on one Intel Core i5-8600T CPU for 300 epochs. In addition to the model information presented in Table 1, the CNN was trained with a learning rate of 0.001, beta1 of 0.9, beta2 of 0.999, and test dataset of 10%. The DeepExplainer module from the SHAP Python library was used to explain the predictions of the CNN. Image analysis and conversion was performed in Python.

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 observe the change in the model output. To minimize the noise that was generated by the SHAP explanations, which enabled the optimal conditions for design transformation, 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 \( \mu \text{m} \)), and captured the explanation at this peak wavelength. This process was repeated, as necessary, for all target wavelengths.

**Author contributions**

C.Y., J.T., and A.R. conceived and designed the study; C.Y. and J.T. led the experiments; Y.K., B.K., and D.H. contributed to data analysis and interpretation; C.Y., J.T., B.K., and A.R. drafted the manuscript and all authors provided feedback. A.R. supervised the study.

**Conflicts of interest**

There are no conflicts to declare.

**Acknowledgements**

This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group. This work was partially supported by the Sloan Research Fellowship from the Alfred P. Sloan Foundation. We also acknowledge feedback and assistance from Dr. Mark Knight and the Northrop Grumman Corporation.
References

1. Chen, R. et al. Nanophotonic integrated circuits from nanoresonators grown on silicon. Nat. Commun. (2014) doi:10.1038/ncomms5325.
2. Pelton, M. Modified spontaneous emission in nanophotonic structures. Nature Photonics (2015) doi:10.1038/nphoton.2015.103.
3. Odebo Länk, N., Verre, R., Johansson, P. & Käll, M. Large-Scale Silicon Nanophotonic Metasurfaces with Polarization Independent Near-Perfect Absorption. Nano Lett. (2017) doi:10.1021/acs.nanolett.7b00416.
4. Pralle, M. U. et al. Photonic crystal enhanced narrow-band infrared emitters. Appl. Phys. Lett. (2003) doi:10.1063/1.1526919.
5. Lin, S. Y., Moreno, J. & Fleming, J. G. Three-dimensional photonic-crystal emitter for thermal photovoltaic power generation. Appl. Phys. Lett. (2003) doi:10.1063/1.1592614.
6. Laroche, M., Carminati, R. & Greffët, J. J. Coherent thermal antenna using a photonic crystal slab. Phys. Rev. Lett. (2006) doi:10.1103/PhysRevLett.96.123903.
7. Yao, K., et al. Intelligent nanophotonics: Merging photonics and artificial intelligence at the nanoscale. Nanophotonics (2019).
8. Molesky, S. et al. Inverse design in nanophotonics. Nature Photonics (2018) doi:10.1038/s41566-018-0246-9.
9. Lin, Z., Groever, B., Capasso, F., Rodriguez, A. W. & Lončar, M. Topology-Optimized Multilayered Metaoptics. Phys. Rev. Appl. (2018) doi:10.1103/PhysRevApplied.9.044030.
10. Sell, D., Yang, J., Doshay, S. & Fan, J. A. Periodic Dielectric Metasurfaces with High-Efficiency, Multiwavelength Functionalities. Adv. Opt. Mater. (2017) doi:10.1002/adom.201700645.
11. Yang, J. & Fan, J. A. Topology-optimized metasurfaces: impact of initial geometric layout. Opt. Lett. (2017) doi:10.1364/ol.42.003161.
12. Jensen, J. S. & Sigmund, O. Topology optimization for nano-photonics. Laser and Photonics Reviews (2011) doi:10.1002/lpor.201000014.
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* (2016) doi:10.1021/acsphotonics.6b00066.

14. Burger, M., Osher, S. J. & Yablonovitch, E. Inverse Problem Techniques for the Design of Photonic Crystals. *IEICE Transactions on Electronics* (2004).

15. Liu, D., Tan, Y., Khoram, E. & Yu, Z. Training Deep Neural Networks for the Inverse Design of Nanophotonic Structures. *ACS Photonics* (2018) doi:10.1021/acsphotonics.7b01377.

16. Jiang, J. et al. Free-form diffractive metagrating design based on generative adversarial networks. *ACS Nano* (2019) doi:10.1021/acsnano.9b02371.

17. So, S. & Rho, J. Designing nanophotonic structures using conditional deep convolutional generative adversarial networks. *Nanophotonics* (2019) doi:10.1515/nanoph-2019-0117.

18. Liu, Z., Zhu, D., Rodrigues, S. P., Lee, K. & Cai, W. A Generative Model for Inverse Design of Metamaterials. *arXiv* (2018) doi:arXiv:1805.10181v1.

19. Olden, J. D. & Jackson, D. A. Illuminating the ‘black box’: A randomization approach for understanding variable contributions in artificial neural networks. *Ecol. Modell.* (2002) doi:10.1016/S0304-3800(02)00064-9.

20. Qiu, F. & Jensen, J. R. Opening the black box of neural networks for remote sensing image classification. *Int. J. Remote Sens.* (2004) doi:10.1080/01431160310001618798.

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

22. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* (2017) doi:10.1145/3065386.

23. Liu, X. et al. Taming the blackbody with infrared metamaterials as selective thermal emitters. *Phys. Rev. Lett.* (2011) doi:10.1103/PhysRevLett.107.045901.

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

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.* (2017) doi:10.1063/1.4995515.

27. Chen, K., Adato, R. & Altug, H. Dual-band perfect absorber for multispectral
plasmon-enhanced infrared spectroscopy. *ACS Nano* (2012) doi:10.1021/nn3026468.

28. Brachmann, A. & Redies, C. Using convolutional neural network filters to measure
left-right mirror symmetry in images. *Symmetry (Basel).* (2016)
doi:10.3390/sym8120144.

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, art. arXiv:1802.03888, Feb 2018.

31. Explainers. *Explainers - SHAP latest documentation* Available at:
https://shap.readthedocs.io/en/latest/. (Accessed: 20th February 2020)
Figure S1: Design transformations with a forward-trained CNN and SHAP. (a) Images of the transformed and FDTD simulated structures. The absorption spectra for the corresponding (b) transformed and (c) simulated structures. Image 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. The antenna arm lengths in the transformed structures exhibited resonant wavelengths with an $n_{eff}$ error of 1.8% and a C error of 4.9%, in comparison to the ground truths (linear fit of plots shown with $R^2=0.998$).
Figure S2: Electric field simulations of MIM resonators at various resonant wavelengths. The electric field profile of (a) an L-shaped resonator and (b) an I-shaped resonator at 5.5 µm, 6.3 µm, and 8.2 µm.

Figure S3: Nanophotonic structures used for CNN training. 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.
| Model 1 | Model 2 | Model 3 |
|--------|--------|--------|
| Layers | Param. | Options | Layers | Param. | Options | Layers | Param. | Options |
| conv2d | 3x3,8  | sgd m  | conv2d | 3x3,16 | sgd m  | conv2d | 3x3,16 | sgd m  |
| ReLU   | 2x2, 2 | 256 minibatch | ReLU   | 2x2, 2 | leakyReLU | ReLU   | 2x2, 2 | leakyReLU |
| avgPool| 3x3,16 | 100 epochs | avgPool| 3x3,32 | avgPool | avgPool| 3x3,32 |
| ReLU   | 2x2, 2 | ReLU    | conv2d | 2x2, 2 | leakyReLU | conv2d | 2x2, 2 |
| avgPool| 3x3,32 | ReLU    | conv2d | 3x3,64 | leakyReLU | conv2d | 3x3,64 |
| ReLU   | 2x2, 2 | ReLU    | conv2d | 2x2, 2 | leakyReLU | conv2d | 2x2, 2 |
| 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 | 3x3,8  | sgd m  | conv2d | 3x3,8  | adam   | conv2d | 3x3,8  | sgd m  |
| ReLU   | 2x2, 2 | 256 minibatch | ReLU   | 2x2, 2 | maxPool | ReLU   | 2x2, 2 | maxPool |
| avgPool| 3x3,16 | 100 epochs | avgPool| 3x3,16 | ReLU    | avgPool| 3x3,16 |
| ReLU   | 2x2, 2 | ReLU    | conv2d | 2x2, 2 | ReLU    | conv2d | 2x2, 2 |
| avgPool| 3x3,32 | ReLU    | conv2d | 3x3,32 | ReLU    | conv2d | 3x3,32 |
| ReLU   | 2x2, 2 | ReLU    | conv2d | 2x2, 2 | ReLU    | conv2d | 2x2, 2 |
| avgPool| 3x3,64 | ReLU    | conv2d | 3x3,64 | ReLU    | conv2d | 3x3,64 |
| conv2d | 2x2, 2 | ReLU    | conv2d | 3x3,128| leakyReLU| conv2d | 3x3,128|
| ReLU   | 3x3,128| ReLU    | conv2d | leakyReLU| ReLU    | conv2d | leakyReLU|
| 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 | 3x3,8  | sgd m  | conv2d | 3x3,8  | sgd m  | conv2d | 3x3,16 | adam   |
| ReLU   | 2x2, 2 | 256 minibatch | ReLU   | 2x2, 2 | leakyReLU | ReLU   | 2x2, 2 | 128 minibatch |
| avgPool| 3x3,16 | 300 epochs | avgPool| 3x3,16 | leakyReLU | avgPool| 3x3,32 |
| ReLU   | 2x2, 2 | ReLU    | conv2d | 2x2, 2 | leakyReLU | conv2d | 2x2, 2 |
| avgPool| 3x3,32 | ReLU    | conv2d | 3x3,32 | leakyReLU | conv2d | 3x3,64 |
| ReLU   | 2x2, 2 | ReLU    | conv2d | 2x2, 2 | leakyReLU | conv2d | 3x3,128|
| RMSE   | 0.097562| RMSE | 0.14086| RMSE   | 0.07709|
| Time   | 229 min | Time    | 42 min | Time   | 340 min |

Table S1: CNN hyperparameter optimization. Table of trained CNN architectures and corresponding RMSE values.