

Numerical modeling and inverse design of complex nanophotonic systems

by

Joshua Stuart Johannes Baxter

A thesis submitted to the University of Ottawa
in partial fulfillment of the requirement for the
Doctor of Philosophy degree in Physics

© Joshua Stuart Johannes Baxter, Ottawa, Canada, 2024

Examining Committee

The following served on the Examining Committee for this thesis.

External Examiner:	Sean Molesky Professor, Dept. of Engineering Physics, Polytechnique Montréal
Internal Member(s):	Pierre Berini Professor, School of Electrical Engineering and Computer Science, University of Ottawa Ebrahim Karimi Professor, Dept. of Physics, University of Ottawa John Armitage Professor, Dept. of Physics, Carleton University
Supervisor:	Lora Ramunno Professor, Dept. of Physics, University of Ottawa

Abstract

Nanophotonics is the study and technological application of the interaction of electromagnetic waves (light) and matter at the nanometer scale. The field's extensive research focuses on generating, detecting, and controlling light using nanoscale features such as nanoparticles, waveguides, resonators, nanoantennas, and more. Exploration in the field is highly dependent on computational methods, which simulate how light will interact with matter in specific situations. However, as nanophotonics advances, so must the computational techniques. In this thesis, I present my work in various numerical studies in nanophotonics, sorted into three categories; plasmonics, inverse design, and deep learning. In plasmonics, I have developed methods for solving advanced material models (including nonlinearities) for small metallic and epsilon-near-zero features and validated them with other theoretical and experimental results. For inverse design, I introduce new methods for designing optical pulse shapes and metalenses for focusing high-harmonic generation. Finally, I used deep learning to model plasmonic colour generation from structured metal surfaces and to predict plasmonic nanoparticle multipolar responses.

Acknowledgements

I want to thank my Mom and Dad for their encouragement and caring support over this past decade of university studies. I want to thank my partner, Maria, who has put up with me since early in my graduate studies and has helped me get through many low points. I'd like to offer my deepest thanks to my thesis advisor, Prof. Lora Ramunno, whose compassion and guidance enabled my success and for giving me the opportunity to freely explore this diverse field of study. I'd like to thank Prof. Antonio Calà Lesina, who I jokingly call my thesis co-advisor, for passing on your knowledge and intuition. I'd like to thank Prof. Arnaud Weck, Prof. Béla Joós, Francine Ford, and Emily Zhang for your letters of recommendation that helped me win the Vanier scholarship. I'd like to thank my thesis committee, Prof. Pierre Berini, Prof. Ebrahim Karimi, Prof. John Armitage, and Prof. Sean Molesky. Finally, I'd like to thank all my collaborators, my fellow members and executives of the Physics Graduate Student Association (PGSA), my research teammates, the staff of the physics department, and the Faculty of Science graduate student office.

Table of Contents

List of Tables	viii
List of Figures	viii
List of Abbreviations	xii
A toast to Heaviside	xiii
1 Introduction	1
1.1 Nonlinear and nonlocal plasmonics	5
1.1.1 Introduction to Plasmonics	5
1.1.2 Nonlinear plasmonics	7
1.1.3 Nonlocal plasmonics	8
1.2 Inverse design in nanophotonics	11
1.3 Deep learning for nanophotonics	14
2 Methods	18
2.1 The FDTD method	18
2.1.1 The FDTD algorithm	20
2.1.2 High Performance Computing	26
2.2 Material Models	28
2.2.1 The Drude Model	29

2.2.2	The Lorentz Model	33
2.2.3	The Drude-Lorentz model	35
2.2.4	The critical points model	37
2.2.5	Numerical Solution of Drude and Lorentz models	38
2.2.6	Numerical solution of the critical points model	41
2.2.7	Hydrodynamic plasma model	42
2.2.8	Numerical solution of the hydrodynamic plasma model	46
2.3	Inverse Design	49
2.3.1	Adjoint sensitivity analysis for topology optimization	51
2.3.2	Adjoint topology optimization implementation	54
2.4	Deep Learning	58
2.4.1	How does the machine learn	61
3	Parallel FDTD Modeling of Nonlocality in Plasmonics	63
4	Hyperpolarizability of Plasmonic Meta-Atoms in Metasurfaces	78
5	Dynamic nanophotonics in epsilon-near-zero conductive oxide films and metasurfaces: a quantitative, nonlinear, computational model	89
6	Inverse Design of pulse shapes	109
7	Plasmonic colours predicted by deep learning	136
8	Unfinished projects	147
8.1	Deep learning for the prediction of multipoles	147
8.1.1	Introduction	147
8.1.2	Methods	148
8.1.3	Results	152
8.1.4	Discussion	154

8.2	Inverse design of a metalens for high harmonic generation	155
8.2.1	Introduction	155
8.2.2	Methods	156
8.2.3	Results	161
9	Conclusion	171
	References	175

List of Figures

1.1	Top figure: Circular gold nanoparticle in a time-varying electromagnetic field. The force between the free charges in the gold particle and the electric field causes the negative electrons to gather on one surface, creating a net positive charge on the other. The charge oscillates at the same frequency as the field resulting in a dipolar emission pattern. Inset: The charge accumulation at the particle interface can lead to large field enhancements (electromagnetic hot spots) at the particle interface, especially when the frequency matches the localized surface plasmon resonance.	6
2.1	Summary of the FDTD algorithm. a) The leap-frog update scheme where the electric (blue x) and magnetic fields (red circles) are solved at integer and half-integer timesteps, respectively. b) Table summarizing the locations of the vector field components on the FDTD staggered grid. c) $y - z$ cross-section of the FDTD grid illustrating the electric field components and locations (green and blue arrows) required for the H_x magnetic field component (red circle). Grid lines are placed at every half-step in each direction.	23
2.2	Real (blue) and imaginary (red) components of the frequency-dependent permittivity of bulk gold. In computational nanophotonics, this strong dispersion is simulated using material models. The numerical data is collected from the database of Ref. [137], but originates from Ref. [138].	29
2.3	Real (blue) and imaginary (red) components of the gold permittivity spectrum (solid lines) with a Drude model fit to this data (circles) where the fitting parameters are $\varepsilon_\infty = 7.79$, $\omega_p = 1.33e16$ rad/s, and $\Gamma = 1.14e14$ rad/s.	32
2.4	Real (blue) and imaginary (red) components of the gold permittivity spectrum (solid lines) with a Lorentz model fit to this data (circles) where the fitting parameters are $\varepsilon_\infty = 1$, $\Omega = 4.69e16$ rad/s, $\omega_0 = 4.14e16$ rad/s, and $\Gamma = 1e16$ rad/s.	35

2.5	Real (blue) and imaginary (red) components of the gold permittivity spectrum (solid lines) with a combined Drude + 2 Lorentz model where the fitting parameters are given by, $\varepsilon_\infty = 0.174$, $\omega_p = 1.33e16$ rad/s, $\gamma = 5.36e13$ rad/s, $\Omega_1 = 5.13e16$ rad/s, $\omega_{0_1} = 2.48e16$ rad/s, $\Gamma_1 = 3.97e16$ rad/s, $\Omega_2 = 1.06e16$ rad/s, $\omega_{0_2} = 5.84e15$ rad/s, $\Gamma_2 = 4.24e15$ rad/s.	36
2.6	Real (blue) and imaginary (red) components of the gold permittivity spectrum (solid lines) with a Drude + 2 critical points model fit to this data (circles) where the fitting parameters are $\varepsilon_\infty = 0.005$, $\omega_p = 1.31e16$ rad/s, $\gamma = 1.40e14$ rad/s, $\Omega_1 = 1.34e16$ rad/s, $A_1 = 0.45$, $\Gamma_1 = 0.0$ rad/s, $\Omega_2 = 3.95e15$ rad/s, $A_2 = 3.16$, $\Gamma_2 = 1.76e15$ rad/s.	39
2.7	Algorithm for nanophotonic inverse design. The main optimization loop is highlighted in green, and pre and post-processing are highlighted in blue. .	57
2.8	A simple neural network illustrating the three primary layers; the input layer receives numerical data, the hidden layers operate on the data, and the output layer combines the data from the last hidden layer and returns the value(s) to the user.	59
2.9	Artificial neurons in a hidden layer and the math it performs.	60
8.1	An example of a randomly generated nanoparticle from a Fourier series with randomly generated weights and truncation index. In the out-of-plane direction y , the particle has a uniform thickness of 40 nm. The incident electric field is polarized along z and travels along y . The output of the simulation is the multipole spectrum $a_{l,m}(\lambda)$ of the particle.	150
8.2	Diagram of a CNN, where an image is transformed by a sequence of convolutional and pooling layers to a flattened vector. This becomes the input of a dense DNN that predicts a vector of multipole coefficients for each wavelength. Red boxes represent convolution kernels, and blue boxes represent the pooling sub-arrays.	152
8.3	Selected data points from the test set, unseen by the CNN in training. The spectra show the magnitude of the first-order electric multipole (dipole) coefficient as calculated by FDTD (blue line) and predicted by the DNN (red x 's). The adjacent image shows the cross-section of the test particle. .	153
8.4	Comparison of the FDTD-calculated values of the first-order electric multipole (dipole) coefficient and the DNN-prediction for all data points in the test set (blue dots). A linear function ($y = x$) is included for convenience (red line).	154

8.5	FDTD simulation setup for an HHG-focusing metalens used for inverse design; a) is a model metalens where the grey matter is MgO, and the ambient medium is air, b) a Gaussian source plane is placed inside the bulk MGO and light propagates along y and is polarized in the z direction, c) a discrete Fourier transform ($x - z$) plane monitor is placed at the bulk MgO-air interface, d) another plane monitor is placed at the top nanostructured MgO-air interface, e) a $y - z$ cross-section of the simulation setup where the green and purple arrows represent the first and second propagation steps respectively, and the yellow x 's represent non-interface regions where fields are set to zero before propagating.	159
8.6	Inverse design of an amplitude plate to maximize focusing at the seventh harmonic $\lambda \sim 114$ nm via automatic differentiation of the ASM. a) Initial condition of the inverse design is chosen to be a Fresnel zone plate tuned for the seventh harmonic. Yellow regions represent unity transmittance, and purple regions represent zero transmittance. b) Final design optimized to maximize the seventh harmonic at the focal plane. c) Seventh harmonic intensity spectrum at the focal point, normalized to the incident intensity for the initial FZP (yellow line) and the inverse-designed device (blue line). d) Electric field intensity at the focal plane.	163
8.7	Inverse design of an amplitude plate to maximize focusing at the seventh harmonic $\lambda \sim 114$ nm and minimize the fundamental $\lambda = 800$ nm via automatic differentiation of the ASM. a) Initial condition of the inverse design is chosen to be a random grid of pixels ranging from fully transmitting (yellow) to total blocking (purple). b) Final design optimized to maximize the seventh harmonic and minimize the fundamental at the focal plane. c) Seventh harmonic intensity spectrum at the focal point, normalized to the incident intensity for the FZP (yellow line) and the inverse-designed device (blue line). d) Log plot of the fundamental intensity spectrum at the focal point, normalized to the incident intensity for the FZP (yellow line) and the inverse-designed device (blue line).	165

8.8	Broadband inverse design of a phase plate to maximize focusing around the seventh harmonic $\lambda \sim 114$ nm via automatic differentiation of the ASM. a) Initial condition of the inverse design is chosen to be a Fresnel zone plate tuned for the seventh harmonic. Yellow regions represent a π phase shift, and purple regions represent a zero phase shift. b) Final design optimized to maximize several wavelengths around the seventh harmonic at the focal plane. This device was not binarized. c) Seventh harmonic intensity spectrum at the focal point, normalized to the incident intensity for the initial FZP (yellow line) and the inverse-designed device (blue line). d) Electric field intensity at the focal plane, demonstrating a large focal spot.	166
8.9	Broadband inverse design of a phase plate to maximize focusing around the seventh harmonic $\lambda \sim 114$ nm and minimize around the fundamental $\lambda = 800$ nm via automatic differentiation of the ASM. a) Initial condition of the inverse design is chosen to be a random grid of pixels ranging from π phase shift (yellow) to zero phase shift (purple). b) Final design optimized to maximize several wavelengths around the seventh harmonic and minimize the fundamental at the focal plane. c) Seventh harmonic intensity spectrum at the focal point, normalized to the incident intensity for the FZP (yellow line) and the inverse-designed device (blue line). d) Log plot of the fundamental intensity spectrum at the focal point, normalized to the incident intensity for the FZP (yellow line) and the inverse-designed device (blue line).	167
8.10	Inverse design of HHG metalens using finite difference gradients. a,c,e) Diagrams showing the optimal location and thickness of rings for 10, 6, and 4 rings, respectively. b,d,f) The evolution of the objective function during the gradient descent process for 10, 6, and 4 rings, respectively.	169

List of Abbreviations

CNN: Convolutional Neural Network
CPU: Central Processing Unit
DNN: Deep Neural Network
ENZ: Epsilon-Near-Zero
FDTD: Finite-Difference Time-Domain
GNOR: Generalized Nonlocal Optical Response (model)
GPU: Graphics Processing Unit
HHG: High-Harmonic Generation
HPC: High-Performance Computing
HRS: Hyper-Rayleigh Scattering
JIT: Just-In-Time (compilation)
LRA: Local Response Approximation
LSP: Localised Surface Plasmon
LSPR: Localised Surface Plasmon Resonance
MPI: Message Passing Interface
SHG: Second-Harmonic Generation
TCO: Transparent Conductive Oxides
TPU: Tensor Processor Unit
TTM: Two-Temperature Model

A toast to Heaviside

James Clerk Maxwell is widely regarded as the father of electromagnetic theory because of his work integrating the discoveries of many before him, including Ampere, Faraday, Neumann, and Lenz, to name a famous few. However, in his 1865 publication *A dynamical theory of the electromagnetic field* [1], Maxwell published 20 equations, unrecognizable to the modern reader. Thus, I believe the title of *father* must be shared with a name rarely taught to students in an electromagnetism course.

Indeed, the four famous equations were introduced to the world through an (at the time) relatively unknown contributor to *The Electrician*, a London-based weekly trade journal. It was Oliver Heaviside who, through a series of publications beginning in 1884, introduced the world to Maxwell's equations in vector form [2], as we know them today (written here in modern notation, including magnetic conduction current and magnetic charge ¹)

$$\nabla \cdot \mathbf{D}(\mathbf{r}, t) = \rho_e, \quad (1)$$

$$\nabla \cdot \mathbf{B}(\mathbf{r}, t) = \rho_m, \quad (2)$$

$$-\nabla \times \mathbf{E}(\mathbf{r}, t) = \mathbf{G}(\mathbf{r}, t) + \frac{\partial \mathbf{B}}{\partial t}, \quad (3)$$

$$\nabla \times \mathbf{H}(\mathbf{r}, t) = \mathbf{J}(\mathbf{r}, t) + \frac{\partial \mathbf{D}}{\partial t}. \quad (4)$$

Maxwell is a rock star, but only because Heaviside was an excellent producer. After a decade of studying physics, I still get excited when I see Maxwell's equations listed at the

¹Heaviside acknowledged that there are probably no magnetic charge/conduction currents, however, insisted on their presence in the equations for completeness and symmetry [3]. I present them here as an homage.

beginning of a paper; now, that's art.

With this knowledge, I now owe Heaviside a debt of gratitude. My fascination with computational nanophotonics and electromagnetism generally hinges on the fact that it's just four memorable equations. If you can write software to solve these equations, you can simulate light and its interactions. With such power, it's hard not to become impassioned.

Chapter 1

Introduction

Photonics is a branch of optics interested in generating, detecting, and manipulating light [4]. This definition is broad, and the field may even seem broader from within. Photonics is more colloquially seen as the practical application of optics, using light for various applications, including microscopy, spectroscopy, holography, fibre optics communications, lasers and photon sources, optical computing, photovoltaics (solar energy), medicine, virtual and augmented reality, to name a humble few. The name *photonics* is a not-so-uncommon misnomer in physics, as photons are not always involved. Much of photonics research and development assumes the wave nature of light, which pre-dates quantum mechanics, as is the case in most of this thesis.

A new branch of photonics research has grown in popularity over the last few decades as fabrication and imaging technologies have matured enough to explore light-matter interaction at extreme scales. *Nanophotonics* is the study of photonics at the nanoscale, interested in the interaction of light with material structures whose length scales are less than the wavelength of optical light ($\lambda \sim 500$ nanometers). Using scanning probe techniques, optical tweezers, high-resolution electron microscopes and lithography tools, focused ion-beam milling systems and more, we can measure, fabricate, and manipulate nanometer-scale structures with unprecedented precision [5]. However, interest in nanooptics has been around for a while. As described in the upcoming sections, many pioneering works of nanophotonics, especially regarding thin-film optics, plasmonics, and quantum confinement research, were conducted as far back as the 1960s. Moreover, principles underlying nanophotonics, such as the scattering of light waves from particles and near-field microscopy, date back to 1908 (Mie) [6], and 1928 (Synge) [5] respectively.

As discovered by Ernst Abbe in 1873 and Rayleigh in 1879, an upper bound is set on the resolution of any optical system due to diffraction; this is the so-called diffraction limit, which is typically given as the wavelength divided by 2 (though this value varies depending on how it is derived and depending on the numerical aperture of the system). As light travels through space, it loses information through evanescent waves. As a result, we lose information regarding the smallest features of the light source, details whose sizes are below the diffraction limit. A similar limit exists for focusing light as well.

Thus, some of the primary interests of nanophotonics are the generation, detection, and focusing of light at the nanoscale. Enhancing these processes involves light confinement, achieved via dielectric waveguides (wires for light), resonators, surface plasmons (electron waves on a metal surface), localized surface plasmons (electron waves on a metal nanostructure surface), nanoscale apertures (nano-holes), to name a few avenues. Waveguides, for example, guide light using the principle of total internal reflection, similar to how a copper wire guides an electronic current. Waveguides thus enable the creation of circuits to perform computations with light (i.e. photonic neural networks). As another example, surface plasmons confine electromagnetic energy into nanoscale regions at the interface between metals and dielectrics. The intensity of the confined field is usually much stronger than the incident electric field resulting in electromagnetic hotspots that enhance the efficiency of optical processes like absorption and emission. Plasmonics becomes useful for quantum emitters and detectors (e.g. fluorescent molecules and quantum dots), which are much more likely to emit/absorb photons if placed in the hotspot of a suitable plasmonic nanostructure.

In nanophotonics, the wavelength of light and the features it interacts with have comparable length scales. Because of this similarity in length scale, the geometry of a nanophotonic system, the distribution of materials that make up the system, determines how the light will respond to it. For example, in the case of a dielectric ring resonator, the circumference determines the resonant wavelengths. The resonance requires constructive interference of the wave in the ring, and thus resonant wavelengths must be the circumference divided by an integer, as in a whispering gallery. As another example, photonic crystals consist of a periodic arrangement of nanostructures whose collective scattering properties prevent the transmission of frequency bands of light (analogous to electronic band gaps in semiconductors) and can be used as mirrors to confine light to nanoscale geometries. The shape, size, and distribution of the meta-atoms (nanostructures) that make up the photonic crystal determine which frequency bands are effectively confined. The strong geometric dependence of nanophotonic systems is the driving force behind a new branch of

study called *inverse design*, which seeks the optimal geometry of an optical system given the desired objective. Instead of using human intuition to design optical systems from simple components (e.g. a ring resonator), inverse design uses advanced algorithms (including artificial intelligence) to design compact, non-intuitive nanophotonic devices that perform the desired function.

Plasmonic particles (small metal particles) also have geometry-dependent optical resonances such that the light will be much more confined to the particle surface at the resonant wavelengths. The resonant wavelengths depend on the shape and size of the particle. If a particle's size is increased by a small amount, its resonant wavelength is also expected to increase. A spectacular demonstration of this physics is *plasmonic colouring*, where metal nanostructures generate colours by absorbing selected wavelengths of white light. The colour generated depends on the geometry of the nanostructures involved. Thus, plasmonic colours are tuneable; we can design plasmonic systems to generate a desired colour. An example of this technology is the Roman Lycurgus cup (created around the 4th century CE), made up of glass infused with gold and silver nanoparticles giving it a dichroic effect; the glass colour appears red if the light transmits from within the glass and green if the light reflects from the glass. Likely without knowing it, the Romans had found an early application of nanophotonics.

The ability to control light at the nanoscale has many applications. As discussed earlier, the diffraction limit restricts the resolution of far-field microscopy (the classic lens-based systems). Nanophotonics has enabled the detection of optical near-fields using plasmonic metal tips, thus capturing fields and features that would be lost to propagation. On a larger scale, by arranging nanoscale features cleverly on a flat surface, we get metasurfaces - flat optical devices (< 1000 nm thick) that can behave as bulk optical devices such as lenses and gratings but in a much more compact form factor (i.e. metalenses and meta-gratings). Nanophotonic computing is also promising to replace electronic computers for simple tasks. Using integrated photonic circuits which confine light to waveguides, light travels through a series of on-chip photonic elements (couplers, resonators, modulators, etc.) that solve a mathematical problem. Nanophotonics also lends itself well to applications in sensing. Along with geometry, nanophotonic systems, especially resonant ones, are sensitive to changes in the ambient refractive index and can thus be used to sense changes in analyte concentrations.

Evident from this discussion is the necessity of sophisticated computational methods in nanophotonics. In much of computational nanophotonics (including this thesis), sophisti-

cated full-wave electromagnetic simulations are required. These simulations directly solve Maxwell’s equations without simplification for arbitrary nanophotonic systems, enabling the exploration of non-trivial geometries. For example, the scattering properties of non-spheroidal particles generally do not have a simple solution. Full wave simulations enable researchers to simulate electromagnetic waves interacting with anything resolvable on the computation mesh (the discrete solution points). These simulations also enable inverse design, as they allow for Maxwell’s equations to be solved for such non-intuitive structures and, thus, the extraction of the electromagnetic gradients. As a final example of the use of computational nanophotonics, simulations are useful in determining the geometric composition of sub-diffraction-limited systems. The diffraction limit hinders the optical characterization of nanoscale objects, but reverse-engineering the far-field spectral data using simulations can present invaluable insight.

One of the most (if not the most) used Maxwell’s equation solver is the finite-difference time-domain (FDTD) method which solves the equations on a rectangular grid at discrete points in time using the data from the previous *timestep*. FDTD’s rectangular grid and time-stepping nature make its implementation on large-scale computing platforms relatively simple and ease the investigation of complex material models, which can be solved on the same grid with the same time-stepping algorithm in parallel. The FDTD method is thus the primary computational tool in this thesis.

These paragraphs have given a glimpse into the world of nanophotonics with a less intensive light before exploring the work conducted in my graduate studies. As is evident above, there is a lot to explore in this diverse and growing field, and I have been lucky in the past several years to have been able to touch many different aspects of nanophotonics from a theoretical and computational perspective. I wish to write all I have learned, but not all projects made it through the investigation phase and thus are omitted. The following sections introduce the subjects behind the research presented in later chapters.

The outline of the thesis is as follows. In the remainder of Chapter 1, we introduce the topics covered in the thesis by dividing them into three categories, plasmonics, inverse design, and deep learning, and review the relevant literature in each. After each topic review, I briefly discuss my work and its relevance in the context of the literature. Chapter 2 provides an overview of the essential methods underlying the thesis work, including an introduction to the finite-difference time-domain method, computational modelling for linear and nonlinear optical materials, inverse design methods, and deep learning. This chapter is meant to provide a detailed basis for the work covered in the subsequent chapters

and provides resources for readers interested in replicating the subject material. Chapters 3 through 7 constitute the body of the thesis, each containing a self-authored, peer-reviewed and published article. Each chapter begins with a high-level introduction to the paper, the inspiration/story behind the work, and a brief discussion of possible future work. Chapter 8 contains unpublished work that, although not formally peer-reviewed, has been presented at conferences. Finally, concluding remarks are provided in Chapter 9.

1.1 Nonlinear and nonlocal plasmonics

1.1.1 Introduction to Plasmonics

When light of a given frequency is incident on a metallic nanostructure (embedded in a dielectric), the free electrons oscillate back and forth at the same frequency due to the force caused by the electric field. This process is illustrated in Figure 1.1, where on top, we illustrate the time evolution of the charge distribution in a circular gold particle irradiated by an electric field $E_0(t)$. At the peak of a half cycle, a net negative charge forms at the structure's surface (blue shading), and on the opposing surface, a net positive charge is induced (red shading). In the next half cycle, these surface charges switch. This charge oscillation results in a (primarily) dipolar scattered field. The buildup of free charges at the structure's interface can result in near-fields that are much more intense than the incident field; these electromagnetic hot spots are illustrated in the inset of Figure 1.1.

The coherent oscillation of electrons at the metal-dielectric interface (more accurately between a material with a negative permittivity and a material with a positive permittivity [5]) is called a *surface plasmon* [7]. A *localized surface plasmon* (LSP) refers to a surface plasmon occurring on the surface of a localized structure (a nanostructure) as illustrated in Figure 1.1. Although surface plasmons need not be localized, the plasmonics work presented in this thesis focuses on LSPs which can be excited directly with light, as opposed to planar surface plasmons, which require momentum-matching to be optically excited [7].

LSPs are resonant phenomena, and the local field enhancement is maximized at the LSP resonance (LSPR) frequency which depends on the nanostructure's shape, size, material, and ambient medium (see chapter 12.1.1 of Ref. [5] for an intuitive explanation). For this reason, plasmonic particles are typically referred to as tuneable as we can design them to resonate at a desired frequency via their shape/size [8,9] or by changing the envi-

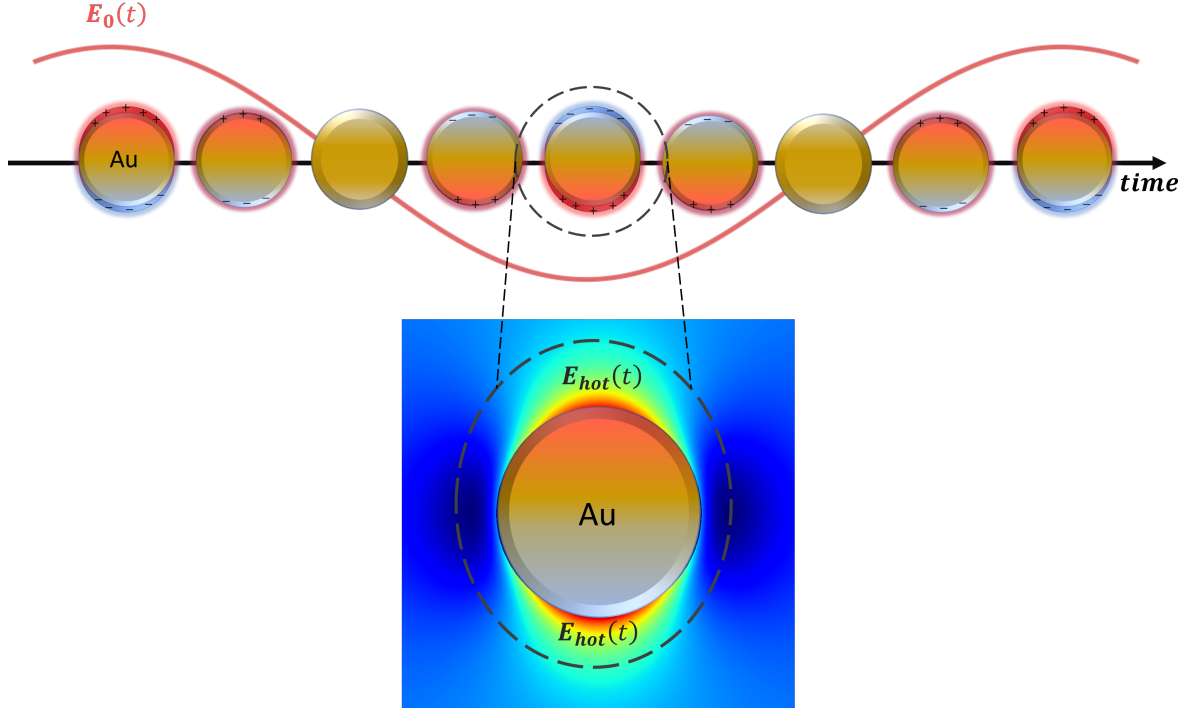


Figure 1.1: Top figure: Circular gold nanoparticle in a time-varying electromagnetic field. The force between the free charges in the gold particle and the electric field causes the negative electrons to gather on one surface, creating a net positive charge on the other. The charge oscillates at the same frequency as the field resulting in a dipolar emission pattern. Inset: The charge accumulation at the particle interface can lead to large field enhancements (electromagnetic hot spots) at the particle interface, especially when the frequency matches the localized surface plasmon resonance.

ronment [10, 11].

The strong near-fields generated at the interface of plasmonic nanostructures near their LSPR have seen applications in several fields (see Ref. [12] for a comprehensive review). Of significant interest to this thesis are applications in nonlinear optics where plasmonic structures are used primarily in two ways, 1.) to enhance external nonlinear optical processes, and 2.) to generate optical nonlinearities [13].

1.1.2 Nonlinear plasmonics

Nonlinear optical processes are inherently weak and require intense optical fields to produce measurable effects. In the perturbative nonlinear optical description (for sufficiently weak fields), the n^{th} order nonlinearly generated field strength $E_{nonlin}^{(n)}$ scales with the n^{th} power of the incident electric field strength, $E_{nonlin}^{(n)} \sim (E_{inc})^n$. By adding plasmonic particles to a weak nonlinear medium, we can generate strong fields E_{hot} inside the nonlinear medium, thus enhancing the nonlinear generation in the hotspots of the particles. It is typical to expect the LSPR hotspot to have $\sim 10\times$ the intensity of the incident beam, $(E_{hot})^2 \sim 10(E_{inc})^2$. In the perturbative description, the n^{th} - order nonlinear intensity (generated in the hotspot) would be expected to be $\sim 10^n$ times stronger than without the plasmonic particle present, $(E_{nonlin,hot}^{(n)})^2 \sim (E_{hot})^{2n} \sim 10^n(E_{inc})^{2n}$.

In 1981, Chen *et al.* demonstrated this nonlinear enhancement by measuring second harmonic generation (SHG) from smooth and rough (nanostructured) metal films. The nanostructured films generated an SHG signal up to four orders of magnitude higher than the smooth films [14]. More recently, Vampa *et al.* used gold monopole nanoantennas to enhance high harmonic generation (HHG) from silicon surfaces [15]. Although the detected signal was enhanced by a factor of only 5 - 10 compared to non-nanostructured silicon, they claim that the HHG emission density is 3 - 4 orders of magnitude greater in the plasmonic hotspots than elsewhere on the surface.

On the other hand, metals exhibit strong optical nonlinearities; thus, metal particles are commonly used as the source of optical nonlinearities with a self-enhancement effect [16]. Furthermore, for odd-order nonlinearities, the bulk nonlinear susceptibility of metals is much larger than traditional nonlinear crystals [17]. As for even-order nonlinearities, they are said to be absent in centrosymmetric media (like plasmonic metals); however, due to

symmetry breaking at the surface of metallic nanostructures, and non-dipolar contributions to the bulk nonlinear optical response, strong even-ordered nonlinearities can also be generated from metals [18]. Furthermore, nonlinear emission from plasmonic nanoparticles can be self-enhanced by designing particles to be asymmetric [19–21], maximizing the surface area [22], and engineering the LSPR to coincide with the pump and (or) nonlinear wavelengths [23, 24].

That said, to design nonlinear plasmonic systems, we have to be able to model them. For this, we turn to the hydrodynamic plasma model, where the free electron system in a metal is treated as a continuous fluid. The hydrodynamic description of free electrons dates back to the work of Felix Bloch in 1933 [25] and was used by Rufus Ritchie in his theoretical prediction of surface plasmons in 1957 [26]. The model was finally applied to theoretically study SHG from metal surfaces in 1980 [18]. Since then it has become the standard classical model for simulating free electron nonlinearities in metals [13, 27–36].

Chapter 4 uses the hydrodynamic plasma model to validate the experimental and theoretical study of SHG emission from metasurfaces of asymmetric gold nano-prisms on glass. To the best of our knowledge, this paper shows the first-ever quantitative comparison between the numerical solution of the hydrodynamic model and experimental measurements, with remarkable agreement. Finally, in Chapter 5, we derive a model for the optical properties of epsilon-near-zero (ENZ) conductive oxide films. These materials have metallic optical properties; however, their high doping concentration results in the electronic temperature dependence of the permittivity, which changes rapidly under ultrafast pulse irradiation. The near-zero permittivity of the conductive oxide film enhances this nonlinearity when TM-polarized light is incident on the film due to the interface conditions of Maxwell’s equations. This permittivity shift can change the ENZ medium from behaving as a reflective metal to a transparent dielectric. These so-called time-varying materials drive the novel field of time-varying/active photonics and the popular concept of *time crystals*.

1.1.3 Nonlocal plasmonics

Not only has it found use in nonlinear plasmonics, but the hydrodynamic plasma model is also applied to linear optics in the case of extreme nanoscale plasmonics (i.e. for structures with size ≤ 10 nm). On such scales, we can no longer assume that the charge is restricted to the surface (as a surface charge density) but must account for charge-smearing effects

near the surface (as a volume charge density). As a result, the optical response of electrons in such small structures depends on the electromagnetic field distribution rather than the local field; our classical plasmonic material models need to be revised at such scales. Models that only account for the local field use the local response approximation (LRA), analogous to the dipole approximation. Plasmonics with extreme nanoscale features is thus the realm of nonlocal plasmonics [37] due to the nonlocal response of the free electrons.

This sub-field of plasmonics began with the prediction of optically-excited longitudinal plasmon resonances in thin metallic films in the late 1960s [38] and experimental realization in the early 1970s [39, 40]. In these studies, the primary characteristic noticed is the existence of optically excited resonances above the plasma frequency, attributed to longitudinal modes in the thin films. These optically excited longitudinal modes are not supported in the LRA, where charges are restricted to the film surface resulting in a strictly transverse electric field. Another late 1960s discovery involved the seeming change in the refractive index of small silver spheres attributed to the change in electron mean-free path in such small particles [41]. In 1973, Ruppin theoretically studied the optical properties of small plasma spheres (an extension of Mie scattering theory [6, 42]) by allowing for longitudinal modes inside the bulk, again demonstrating the existence of longitudinal resonances above the plasma frequency [43, 44]. Part of the discovery of Ref. [41] was further validated in the 1990s when it was shown that LSP resonances of small plasmonic particles blueshifted and the linewidths increased with decreasing particle size [45, 46], an effect that has been further demonstrated in the 2000s using optical and electron energy loss spectroscopic measurements [47, 48]. This blueshift and increasing linewidth with decreasing particle size are inconsistent with our classical understanding since particles whose sizes fall within the quasistatic limit (i.e. the particle size is much smaller than the wavelength) have size-independent resonance frequencies.

These discoveries all point to an unaccounted-for effect in our classical modelling of plasmonics when it comes to the extreme nanoscale. Longitudinal modes mean a non-zero volume charge density inside the bulk. Indeed, most of the nonlocal plasmonic discoveries are accounted for by the existence of the electron pressure of the hydrodynamic model, which depends on the volumetric charge density distribution. Initially, this was dealt with by introducing nonlocal permittivity functions which account for spatial dispersion as introduced by Hopfield and Thomas in 1963 [38, 49–53] or through the solution of the hydrodynamic model for simple systems [54–56].

The early-2000s saw an emergence of interest in nonlocal plasmonics, mainly due to

increased capabilities in fabrication technology and the ability to numerically simulate plasmonic systems with unprecedented resolution; we can solve the hydrodynamic model in plasmonic nanostructures with small enough features to host nonlocal effects. For example, in 2009, McMahon *et al.* were able to observe nonlocal plasmonic effects in 2D cylinder and triangular nanowires using the hydrodynamic plasma model in FDTD [57], extending this to 3D the next year [58, 59]. Several numerical studies followed this foray [47, 60–66] where Maxwell’s equations and the hydrodynamic model are solved primarily using finite element methods. The papers using the hydrodynamic model were able to show the hallmarks of nonlocal plasmonics, including the blue shift of the LSPR [47] and longitudinal resonances [60]. However, the increased resonance linewidth was not accounted for beyond phenomenological corrections to the damping coefficient [41].

That was until the development of the Generalized Nonlocal Optical Response (GNOR) model by Mortensen *et al.* in 2014 [67]. This model was derived from the hydrodynamic model but accounts for diffusion dynamics of the induced charge density and electronic pressure and demonstrates the effects of nonlocality, including the increased resonance linewidth. Since then, the model was applied to study plasmonic particle ensembles [68], plasmonic enhancement of quantum emitters [69, 70], hybrid plasmonic-quantum dot systems [71], plasmonic lasers [72], Förster resonance energy transfer in the vicinity of metallic nanospheres [73], and highly confined gap plasmon modes [74], to name a few examples. However, until 2021, this model had yet to be implemented in FDTD.

Furthermore, the FDTD implementation of the hydrodynamic plasma model by McMahon *et al.* (mentioned above) was based on an erroneous mathematical approximation which caused spurious resonances both above *and below* the plasma frequency [60]. Indeed, until our 2021 paper [75] nobody had reported a correct FDTD implementation of the nonlocal hydrodynamic model, nor the GNOR model due to the computational load involved for 3D structures. Moreover, as electronic smearing occurs within a Fermi wavelength of the surface (~ 0.5 nm for plasmonic metals), significantly high resolutions are required to resolve these effects. Thus, these simulations require prohibitive amounts of memory and time [76].

The uOttawa Computational Nanophotonics research group specializes in large-scale electrodynamics simulations due to our state-of-the-art FDTD software capable of scaling linearly on thousands of CPUs [77]. Our FDTD implementation of the nonlocal hydrodynamic and GNOR models is the subject of Chapter 4 of this thesis.

1.2 Inverse design in nanophotonics

Typically in nanophotonics, designs for systems (i.e. metasurface unit cells, photonic crystal cavities, nanostructure geometry, etc.) are conceived using intuition-based design, or *forward design*. In this strategy, a potential device is typically composed of highly symmetric structures with known optical responses. A small number of parameters are tuned to achieve the desired functionality. This forward design method is simple in principle but limited in scope. Indeed, there is no reason the optimal device comprises simple, human-chosen components. Recently we have seen an interest in expanding the feasible design space using computational methods. Some of these methods seek to find a design that maximizes an objective function, allowing an optimization algorithm to perform the design process. We refer to these objective-first methods as *inverse design*. In forward design, the human designer is restricted to their imagination and time; how many arrangements of how many shapes can be tested in a reasonable amount of time? In *inverse design*, we are limited only by the methods we use and the restrictions we place on the problem; are we using a gradient-based or evolutionary approach, how do we parameterize our design, and what feature scales are permitted?

In nanophotonics, with its strong geometric dependence, we deeply understand the interest in inverse methods for geometric design. By allowing a computer to optimize the sub-wavelength geometry of a nanophotonic system, we can achieve unprecedented control over the confinement and flow of electromagnetic waves using previously unimaginable devices.

We, therefore, see it applied even in the nascent years of modern nanophotonics research (the late 90s), seemingly as a continuation of the use of optimization algorithms to optimize diffractive optics [78–82]. To the best of my knowledge the first nanophotonics application of inverse design was in 1998 by Spühler *et al.* who used an evolutionary algorithm to design a spot-size converter (a.k.a a fibre-waveguide coupler) [83]. They discretized their design region (along the waveguide direction) into equal segments (of length $L \sim 3 \mu\text{m}$) and optimized the width of the silica waveguide in each. These segments naturally allowed for a grating-like structure with a compact form factor. A year later, Cox *et al.* used a gradient descent algorithm to maximize photonic bandgaps in 2D photonic crystals using an early form of density optimization [84, 85]. Felici *et al.* used a similar approach as Spühler to design waveguide tapers (i.e. discretizing the taper in the propagation direction and tuning the width of each segment), using numerical finite differences and gradient descent and applying regularization to reduce the solution space to only stable solutions [86].

As a final example of early nanophotonics inverse design, in 2005, Preble *et al.* used an evolutionary algorithm to design 2D photonic crystals with large photonic bandgaps [87] by tuning the individual pixels of a 32x32 square grid. This method of grid-based devices foreshadowed the current state-of-the-art inverse design – adjoint topology optimization.

In topology optimization, device design determines the optimum binary distribution of matter to accomplish a desired task. For example, consider designing a bike frame by determining the optimal distribution of aluminum and air that minimizes weight and air resistance. In nanophotonics, this typically means finding a binary permittivity distribution $\varepsilon(\mathbf{r}) = [\varepsilon_1 \text{ or } \varepsilon_2]$ to optimize some electromagnetic quantity. Adjoint sensitivity analysis (colloquially referred to as the adjoint method) provides a way to calculate how sensitive an objective function is to a perturbation $\delta\varepsilon(\mathbf{r})$. This sensitivity (the gradient of the objective function) provides a map to guide the optimization algorithm. In most implementations, $\varepsilon(\mathbf{r})$ is initially allowed to take on any value between ε_1 and ε_2 and is gradually forced towards a binary distribution during the optimization process. Chapter 2.3 of this thesis covers the mathematics and details behind adjoint topology optimization.

Adjoint topology optimization was introduced to nanophotonics in 2004 through the works of Jensen, Sigmund *et al.*, where the adjoint method is used to maximize transmission through photonic crystal waveguide bends [88,89]. In the same year, Burger *et al.* wrote a review on inverse methods for photonic crystals, and they introduced the idea of level set optimization for photonic crystals [90]. The level set method is a special adjoint topology optimization method that parameterizes the binary permittivity distribution $\varepsilon(\mathbf{r}) = [\varepsilon_1 \text{ or } \varepsilon_2]$ as a continuous function known as the level set function $L(\mathbf{r})$. The function is positive for one medium $L(\mathbf{r}) > 0 \implies \varepsilon(\mathbf{r}) = \varepsilon_1$, and negative for the other $L(\mathbf{r}) < 0 \implies \varepsilon(\mathbf{r}) = \varepsilon_2$, and thus interfaces are the zeros of the function $L(\mathbf{r}_{interface}) = 0$ [91,92]. The gradient of the objective function with respect to the level set function is calculated using the adjoint method and used to optimize the level set function. The benefit of such a method is that the distribution $\varepsilon(\mathbf{r})$ remains binary while being optimized.

Jensen, Sigmund *et al.* continued their work with adjoint topology optimization, using it to inverse design a low-loss photonic crystal waveguide T-junction, introducing the concept of filtering to avoid fine feature sizes [93]. In addition, the co-authors of Ref. [90] used the level set method to inverse design 2D band structures with unforeseen bandgaps [94]. At the same time, Jiao *et al.* also presented an adjoint-based approach for designing photonic crystal structures, using Wannier basis functions to expand their fields, and thus providing an efficient, semi-analytical method for extracting the gradient of the objective

function with respect to perturbations in the photonic crystal structure [95, 96]. A review of early adjoint-based design works in nanophotonics is found in Ref. [97].

The early 2010s saw inverse design applications in integrated silicon photonic circuits. These devices require the use of vectorial full-wave simulations to determine the objective functions as well as the sensitivities. In the case of these waveguide-fed devices, objective functions can (usually) be parameterized by the mode coefficients, and thus the adjoint method is well suited.

In 2013 Lu and Vučković introduced their generalized inverse design method, using a mixed procedure of continuous topology optimization followed by further tuning via a level-set approach [98]. They used a finite-difference frequency-domain electromagnetic solver to evaluate the electromagnetic problem at each optimization iteration. They demonstrated their method by designing many on-chip mode and wavelength-splitting devices. In the next two years, they were successful in designing and fabricating a demultiplexing grating coupler [99], an on-chip wavelength demultiplexer [100], and a 3-channel wavelength multiplexer [101]. Also in 2013, Lalau-Keraly *et al.* demonstrated level-set optimization using Lumerical FDTD, the first demonstration of this type of inverse design using a commercial FDTD solver (to the best of my knowledge) [102]. They demonstrated this by designing a Y-branch silicon waveguide splitter, an example used by Lumerical to this day [103]. Finally, Shen *et al.* used topology optimization to inverse design and fabricate the (at the time) smallest polarization beam splitter [104].

Inverse design has recently been applied to designing metasurfaces and flat optics. For example, in 2017, Sell *et al.* demonstrated the use of adjoint topology optimization on multi-functional, periodic dielectric metasurfaces [105, 106]. In the following year, Pestourie *et al.* demonstrated the inverse design of aperiodic metasurfaces using a series of approximations that simplify the forward problem to a fast semi-analytic procedure [107]. Backer generalized this procedure to cascaded meta-optic systems using Fourier optics to propagate light between each metasurface [108]. Finally, in 2022, Skarda *et al.* used a GPU-parallelized transition matrix method to design millimetre-sized, nanophotonic metasurfaces [109].

Inspired by some of the later publications, I slowly learned how the adjoint method worked and the intricacies involved throughout my graduate studies, implementing the approach in our in-house FDTD software. In my final years of study, I found potential

applications to inverse design in my research. In Chapter 6, we apply the adjoint method to the problem of designing optical pulse shapes for propagation through time-varying media. These materials’ transmission/reflection properties depend on the temporal shape of the incident optical pulse. Thus computational design can be applied to maximize the transmittance (for example) through such materials via pulse shaping. To the best of our knowledge, this is the first time the adjoint method has been applied to design pulse shapes. Finally, in Chapter 9.2, I show my unfinished work in a project that seeks to inverse design metalenses that focus light generated by solid-state HHG. We consider three approaches; using finite differences to approximate sensitivities, using automatic differentiation in the simulations for the automatic extraction of sensitivities, and using the adjoint method with a slight extension to account for the optical nonlinearity.

1.3 Deep learning for nanophotonics

The last decade has witnessed a paradigm shift in the computational sciences. Machine learning is the field of study devoted to understanding and developing algorithms that learn from data [110]. These algorithms seem mystical from a user perspective, but they are simply mathematical manipulations of the input data, resulting in some output. For example, the image of a dog (an array of numbers representing the colour of each pixel) can be fed into such an algorithm, which will apply algebraic manipulations to the image and output the label “dog.” This extraordinary ability to find complex relationships between input-output data makes (supervised) machine learning an excellent tool for advanced *curve fitting*, which remains one of the simplest and ubiquitous applications of machine learning in modern physics research. That said, the creativity of researchers is boundless when applying new tools to their problems, and machine learning and its underlying machinery are no exceptions.

Machine learning covers a vast landscape of methods, from simple linear regression to self-learning agents. In general, however, machine learning involves gradually adjusting the fitting parameters of a model to reduce an error function. In this way, the computer gradually learns the data – the algorithm finds the parameters that minimize the difference between the output and the original data. The word *learn* is being used loosely here. However, the term becomes slightly less loose when we focus on algorithms that attempt to model the brain or, at least, neural connections.

Deep learning is a subfield of machine learning where the model is a multi-layered (deep) artificial neural network. These networks are mathematical models based on biological neurons and their connections. Numerical data is fed into the network, and each artificial neuron performs a nonlinear operation on the data it receives, passing it on to the next artificial neuron in the net. The output of the network is thus the sum of many nonlinear manipulations on the data. The learning process involves tuning the parameters in each artificial neuron to match the network’s output to a desired result.

The use of artificial neural networks in photonics dates back to the 1990s during the second surge of AI research (the first surge peaking in the 1960s, and the third surge is occurring now), where researchers applied it to tasks such as fitting effective permittivity of microstrip lines [111], spectroscopic data reduction, correction and interpretation [112], modelling and optimization of microwave integrated circuit components [113, 114], and direction finding for phased arrays [115], to name a few.

The most recent surge in AI research started in the early 2000s, with the maturing of computational technologies and advanced algorithms that made training deeper networks more efficient [116]. At this time, the importance of the DNN depth became realized; deeper networks are generally more accurate than their shallower counterparts. Simple networks worked for microwave photonics due to the abundance of available data (the physics is well known, and computational methods were quite mature) and the ability to decompose device models into a handful of variables [117]. As a result, shallow networks provided a flexible curve-fitting tool to replace less-efficient computations.

On the other hand, nanophotonics was in its infancy. The growth of nanophotonics theory depended on the growth of computational power due to the complex nature of light-matter interaction at the nanoscale. At the nanoscale, devices feature high dispersion and sub-wavelength features, and the computational power required to model these devices accurately only came in the early 2000s. Therefore, we expect networks required to model such complex systems to be deeper and more extensive than those used in the 1990s.

It took two decades for computational power to reach the level where we could train DNNs with millions of fitting parameters compounded by the development of robust parallel computing architectures in modern graphics processing units (GPUs), enabling orders of magnitude speed-up in computations. Finally, AI was open to the broader scientific community via open-source code packages enabling high-level development of DNNs without

knowing the low-level computational details. For example, most of these packages have backpropagation and optimization algorithms built in, so the user only needs to define the network architecture and call a function to train it. These packages include (but are certainly not limited to) Tensorflow (2015), Pytorch (2016), Keras (2015), OpenNN (date unknown) and Scikit Learn (2010 – though DNN support came later).

With the metaphorical stars aligning, DNN applications to nanophotonics research began exploding around 2018. One ground-breaking piece came from Malkiel *et al.* who demonstrated a bidirectional DNN trained to, in one way, predict the reflection spectrum of a metasurface composed of H-shaped metallic nanostructures and, in the other way, generate an H-shaped metallic particle for a desired reflection spectrum [118]. Another example comes from Liu *et al.*, who tackled the inverse design uniqueness problem by training an inverse design network using a pre-trained forward network. They applied their architecture to thin film stacks wherein the transmittance of a stack can be predicted (forward), or the film-stack parameters can be generated for a desired transmittance spectrum (inverse) [119]. Ma *et al.* also used bidirectional networks trained to predict the spectral response of chiral plasmonic metasurfaces under irradiation from three circular polarization states and solved the inverse problem [120].

As a final example of high-impact contributions from 2018, Peurifoy *et al.* used neural networks to predict the scattering cross-section spectrum of multilayer dielectric nanospheres [121]. As one might expect, given the trend, the inverse problem is also solved; however, this time, not by more neural networks. DNNs are differentiable; we can calculate the gradient of the network’s output with respect to all the fitting parameters. Similarly, the derivative of the neural network’s output can be backpropagated to the neural network’s input layer, an underrated revelation. This is the gradient of the output of the DNN with respect to the input. Given this gradient, one can use gradient-based optimization algorithms to optimize the width of the dielectric layers given a desired spectrum. This paper is specifically highlighted here because (to the best of my knowledge) it is the first paper using an automatic differentiation scheme for inverse design, although not explicitly mentioned.

A common theme in these papers is the size, depth, and complexity of the networks presented in these publications. These vast networks further illustrate the complex relationships in light-matter interaction at the nanoscale. The amount of data required to train the network accurately is tremendous, requiring thousands to tens of thousands of simulations to generate the necessary data.

These early papers predicted two of the main applications of DNNs in nanophotonics even to this day, simulation replacement and inverse design. Even now, 3D full-wave simulations can take hours or days, whereas a pre-trained DNN typically produces results within a second. Furthermore, conventional inverse design algorithms can require hundreds or thousands of simulations to converge on a result that is not guaranteed to be a global optimum nor to provide desired performance. Although DNNs do not provide these guarantees, the data-based approach makes it more likely (we are not traversing the parameter space blind), and failure to produce a desirable result takes orders of magnitude less time to realize.

In 2019, Sajedian *et al.* published two papers illustrating the motion of the DNN-nanophotonics field toward more complicated deep-learning algorithms [122, 123]. In Ref. [122], deep Q-learning (reinforcement learning) was used to optimize the geometry of silicon nanodisks on a Si₃N₄ substrate to produce a desired colour when illuminated by white light. At the time, deep Q-learning had become a popular topic in deep learning communities because of its ability to self-learn (the DNN learns what steps to take to accomplish a task). It was commonly applied to train bots to beat video games. In Ref. [123], the authors used convolutional neural networks to predict the absorption spectrum of silver plasmonic particles on glass. This time, instead of parameterizing the system into a handful of values, an image of the nanoparticle was fed to the network, with the output vector being the spectrum.

In this thesis, I present two projects that harness deep learning for nanophotonic design. In Chapter 8, we train DNNs to predict reflected colour from random plasmonic metasurfaces and solve the inverse problem. This work, published in 2019, was part of the initial AI boom in nanophotonics research, where DNNs were used to make predictions on a parameterized dataset. In Chapter 9.1, we take this further and train convolutional neural networks to predict the multipole scattering coefficients of plasmonic particles from pictures of their geometric cross-sections. This work remains unpublished outside of conference publications [124].

Chapter 2

Methods

This chapter covers the methods used to accomplish the work in this thesis. These include the FDTD method, plasmonic material models and their implementation in FDTD, the hydrodynamic plasma model and its solution, inverse design, adjoint sensitivity analysis, and deep learning. References are provided to fill in details that are not relevant and thus not included in the text. Furthermore, the subsequent chapters each include unique method details, and they may be excluded here.

2.1 The FDTD method

In 1966, Kane Yee revolutionized the study of electromagnetism through his famous *Numerical Solution of Initial Boundary Value Problems Involving Maxwell's Equations in Isotropic Media* [125]. The algorithm he introduced is now called the finite-difference time-domain (FDTD) method and was the first numerical solver of Maxwell's equations in the time domain. I specify *time-domain* as the method of moments was introduced in the same year [126].

FDTD enabled the ability to solve Maxwell's equations for arbitrary systems within computational limits. Unfortunately, these limits were quite restricting, as remarked by Yee himself, who kept the examples to 2D. As a result, FDTD-related publications remained sparse until the mid-1980s to early 1990s [127]; the acronym itself only being introduced in 1980 by arguably the fun step-father (or favourite uncle) of FDTD, Alan Taflove [128].

In 1995, Taflovie published the first edition of the most comprehensive text on the FDTD method, revising it every five years. In the third edition (2005), Taflovie introduces an invited chapter on hardware acceleration for FDTD, indicating a turning point for FDTD use in photonics (see chapter 20 of Ref. [129]). As discussed by the author, at this time in history, personal computing (PC) technology had reached a point where FDTD software packages were being written for them rather than supercomputers, with this chapter advising relatively low-cost acceleration solutions. This transition to PCs coincides with the founding of electromagnetic software companies like Lumerical (founded 2003) and Optiwave (incorporated 2005), offering their own FDTD software.

FDTD remains one of the most widely used tools for full-wave electromagnetic simulations primarily due to its ease of implementation and use. The ability to directly access the electric and magnetic fields at a given point in space and time presents the user with a *playground* of data. Typical uses of the software include Fourier transforming the fields to the frequency domain to extract spectral response data, compiling the field data into time-domain movies for visualizing the optical processes *in real time*, propagating fields using Green's functions to predict far-field radiation patterns, experimenting with complex material models, and even determining objective function gradients in the case of inverse design.

The limitations of FDTD today are the same faced by Yee, only at a much different scale. With rapid computing power and memory capacity advancements, 3D simulations of realistic optical systems are now possible on mediocre PC hardware, and 3D nanophotonic scattering problems are now solvable on the laptops carried by most university students (with limited resolution).

Furthermore, FDTD lends itself well to high-performance computing. Its rectangular grid and the fact that only nearest-neighbour points are required to evolve the fields means that it can be parallelized on distributed-memory systems (such as compute clusters) with relative ease [77]. The explicit, grid-based nature of the method also lends itself well to shared-memory systems like graphics processing units (GPUs) [130].

The near future of the FDTD method (and other such PDE solvers) is one of new platforms, both hardware and software. For example, new programming libraries such as *TensorFlow* [131], and *JAX* [132] allow users to run numerical calculations on different processor types (including central, graphics and tensor processing units - CPUs, GPUs, and TPUs) with only minor adjustments to the code, enabling hardware acceleration of code

without the prohibitive overhead. Such platforms are beginning to enable advanced functionality, including just-in-time compilation (JIT), where functions written in high-level interpreter languages (like Python) can be compiled at runtime, and automatic differentiation, where the gradient of an arbitrary function with respect to its inputs is provided to the user with a simple function call. These resources allow researchers to write simulation software in high-level languages (e.g. Python, Julia) with their stereotypical accessibility and rich, open-source feature sets without risking speed and scalability. Automatic differentiation is a significant development for scientific computing and machine learning, allowing coders to optimize systems using gradient-based methods without explicitly calculating the gradients or relying on finite-difference approximations. Although the memory requirements present limitations at the moment, an automatic-differentiated full-wave electromagnetics solver may be in the future of inverse design.

On the other hand, the advancement in hardware accelerators, such as GPUs and TPUs, hints at chip-level acceleration. Simply put, electronic (and photonic) chips are being developed for specified tasks (such as matrix multiplication in TPUs) that enable them to perform computations at speeds that are orders of magnitude faster than CPUs in PCs. This is the promise of Tidy3D [133], a new competitor in the computational electromagnetism market, which advertises its use of “emerging computing chips” to perform simulations much faster than other (CPU-based) solvers. This technology promises to enable the simulation and design of complete meta-optical devices (and possibly photonic circuits) in FDTD [134].

2.1.1 The FDTD algorithm

Now that the FDTD method has been introduced, in this section, we will derive the FDTD algorithm, a numerical method for solving Maxwell’s equations, given by

$$\nabla \cdot \mathbf{D}(\mathbf{r}, t) = \rho_e \quad (2.1)$$

$$\nabla \cdot \mathbf{B}(\mathbf{r}, t) = 0 \quad (2.2)$$

$$\nabla \times \mathbf{E}(\mathbf{r}, t) = -\frac{\partial \mathbf{B}}{\partial t} \quad (2.3)$$

$$\nabla \times \mathbf{H}(\mathbf{r}, t) = \mathbf{J}(\mathbf{r}, t) + \frac{\partial \mathbf{D}}{\partial t}, \quad (2.4)$$

where $\mathbf{D}$ is the electric displacement field, ρ_e is the free electron charge density, $\mathbf{B}$ is the magnetic flux density, $\mathbf{E}$ is the electric field, $\mathbf{H}$ is the magnetic field, and $\mathbf{J}$ is the free electron current density. In this text, we use $\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}$, where $\mathbf{P}$ is the bound-charge polarization density and ε_0 is the permittivity of free space. We also assume the absence of magnetic materials so that $\mathbf{B} = \mu_0 \mathbf{H}$, where μ_0 is the permeability of free space, so Maxwell's equations become,

$$\varepsilon_0 \nabla \cdot \mathbf{E}(\mathbf{r}, t) + \nabla \cdot \mathbf{P} = \rho_e \quad (2.5)$$

$$\nabla \cdot \mathbf{H}(\mathbf{r}, t) = 0 \quad (2.6)$$

$$\nabla \times \mathbf{E}(\mathbf{r}, t) = -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \quad (2.7)$$

$$\nabla \times \mathbf{H}(\mathbf{r}, t) = \mathbf{J}(\mathbf{r}, t) + \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}}{\partial t}. \quad (2.8)$$

The FDTD method uses the finite difference method to solve the latter two equations (the so-called *curl equations*). Simply put, the finite difference method approximates derivatives using a difference equation,

$$\left. \frac{\partial f}{\partial t} \right|_{t=t_0} \approx \frac{f(t_0 + \Delta t/2) - f(t_0 - \Delta t/2)}{\Delta t}, \quad (2.9)$$

where Δt is generally chosen to be as large as reasonably (or physically) possible while still accurately representing the derivative. Notice that the derivative is evaluated at $t = t_0$, but the function f is evaluated at half-steps $\pm \Delta t/2$ on either side of t_0 ; this is called a centred difference with second-order accuracy as the error scales with Δt^2 .

Yee's algorithm allows for both $\mathbf{E}$ and $\mathbf{H}$ to be solved using central differences (with second-order accuracy) where the fields are updated at discrete timesteps, using a leap-frog scheme; the electric field and magnetic fields are solved at offset-times. The electric field is solved at integer-multiples of the timestep; $\mathbf{E}(t) = \mathbf{E}(n\Delta t)$ where n is an integer, and Δt is the discrete timestep chosen for the calculation (the choice of which is discussed later). The magnetic field is solved at half-integer multiples of the timestep; $\mathbf{H}(t) = \mathbf{H}((n + 1/2)\Delta t)$.

In order to keep the notation simplified, from now on, we will replace time with the timestep; for example, $\mathbf{E}(n\Delta t)$ is labelled as $\mathbf{E}^n$. Using the central finite difference scheme, we rewrite Maxwell's curl equations as

$$\nabla \times \mathbf{E}^n = -\mu_0 \frac{\mathbf{H}^{n+1/2} - \mathbf{H}^{n-1/2}}{\Delta t} \quad (2.10)$$

$$\nabla \times \mathbf{H}^{n+1/2} - \mathbf{J}^{n+1/2} = \varepsilon_0 \frac{\mathbf{E}^{n+1} - \mathbf{E}^n}{\Delta t} + \frac{\mathbf{P}^{n+1} - \mathbf{P}^n}{\Delta t}, \quad (2.11)$$

where we can see the benefit of the leap-frog (half-step) scheme. On the right side of both equations are time derivatives of a field, and on the left side, we have the curl of the opposite field at the center time. Rearranging, we achieve the time part of the Yee algorithm,

$$\mathbf{H}^{n+1/2} = \mathbf{H}^{n-1/2} - \frac{\Delta t}{\mu_0} \nabla \times \mathbf{E}^n \quad (2.12)$$

$$\mathbf{E}^{n+1} = \mathbf{E}^n + \frac{\Delta t}{\varepsilon_0} \nabla \times \mathbf{H}^{n+1/2} - \frac{\Delta t}{\varepsilon_0} \mathbf{J}^{n+1/2} - \frac{1}{\varepsilon_0} (\mathbf{P}^{n+1} - \mathbf{P}^n). \quad (2.13)$$

The leap-frog scheme is illustrated in Figure 2.1a. At timestep n , we first solve the magnetic field $\mathbf{H}^{n+1/2}$ using the electric field from the previous timestep, $\mathbf{E}^n$. Next, we solve the electric field $\mathbf{E}^{n+1}$ using the magnetic field $\mathbf{H}^{n+1/2}$, the current density $\mathbf{J}^{n+1/2}$, and polarization density $\mathbf{P}^{n+1}$ fields. The treatment of $\mathbf{P}^{n+1}$ and $\mathbf{J}^{n+1/2}$ field will be discussed later.

Like in time, space must also be discretized using whole and half steps so that no two $\mathbf{E}$ or $\mathbf{H}$ field components are collocated. Again, use index notation (for example, the point $(i+1/2, j, k)$ represents the real-space location $((i+1/2)\Delta x, j\Delta y, k\Delta z)$ where, as with the timestep Δt , the $\Delta x, \Delta y$, and Δz are the space step-sizes in the x, y , and z directions.

Although the convention differs for different FDTD codes, we will use the following simple pattern here. The location of an electric field polarization component is offset by a half step in the polarization direction. For example, E_x is offset by $1/2\Delta x$. The opposite is true for the magnetic fields. For example, H_x is offset by half steps in y and z , but remains at integer steps of Δx . The polarization and current density fields are usually collocated with the electric field. This is summarized in the table of Figure 2.1b.

The reason behind the choice of the polarization component offsets becomes obvious when we replace the curls of Eq. 2.12 with centred finite differences as written in Eqs. 2.14 to 2.16 for all magnetic field components,

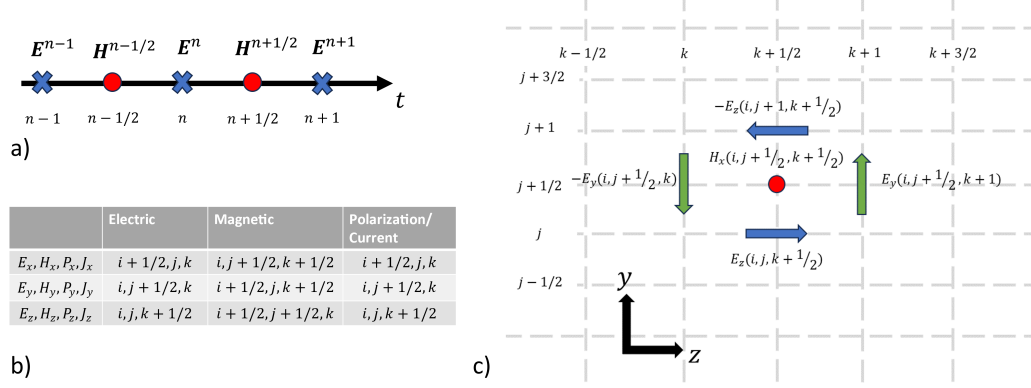


Figure 2.1: Summary of the FDTD algorithm. a) The leap-frog update scheme where the electric (blue x) and magnetic fields (red circles) are solved at integer and half-integer timesteps, respectively. b) Table summarizing the locations of the vector field components on the FDTD staggered grid. c) $y-z$ cross-section of the FDTD grid illustrating the electric field components and locations (green and blue arrows) required for the H_x magnetic field component (red circle). Grid lines are placed at every half-step in each direction.

$$\begin{aligned}
 H_x^{n+1/2}(i, j + 1/2, k + 1/2) &= H_x^{n-1/2}(i, j + 1/2, k + 1/2) \\
 &- \frac{\Delta t}{\mu_0} \left[\frac{E_z^n(i, j + 1, k + 1/2) - E_z^n(i, j, k + 1/2)}{\Delta y} \right. \\
 &\quad \left. - \frac{E_y^n(i, j + 1/2, k + 1) - E_y^n(i, j + 1/2, k)}{\Delta z} \right] \quad (2.14)
 \end{aligned}$$

$$\begin{aligned}
 H_y^{n+1/2}(i + 1/2, j, k + 1/2) &= H_y^{n-1/2}(i + 1/2, j, k + 1/2) \\
 &- \frac{\Delta t}{\mu_0} \left[\frac{E_x^n(i + 1/2, j, k + 1) - E_x^n(i + 1/2, j, k)}{\Delta z} \right. \\
 &\quad \left. - \frac{E_z^n(i + 1, j, k + 1/2) - E_z^n(i, j, k + 1/2)}{\Delta x} \right] \quad (2.15)
 \end{aligned}$$

$$\begin{aligned}
H_z^{n+1/2}(i+1/2, j+1/2, k) = & H_z^{n-1/2}(i+1/2, j+1/2, k) \\
& - \frac{\Delta t}{\mu_0} \left[\frac{E_y^n(i+1, j+1/2, k) - E_y^n(i, j+1/2, k)}{\Delta x} \right. \\
& \left. - \frac{E_x^n(i+1/2, j+1, k) - E_x^n(i+1/2, j, k)}{\Delta y} \right]. \quad (2.16)
\end{aligned}$$

It is clear now why the staggered grid is chosen; it allows us to calculate the curls using central differences as with the time derivatives. The H_x field update Eq. 2.14 is illustrated in Figure 2.1c where the red circle represents the location of $H_x(i, j+1/2, k+1/2)$ and the green/blue arrows represent the E_y/E_z electric field components and the directions they point (based on their sign in Eq. 2.14). The FDTD mesh demonstrates the curl operation as the arrows in Figure 2.1c rotate around the update location. The choice of the magnetic field locations also has the same benefit for the electric field update equations,

$$\begin{aligned}
E_x^{n+1}(i+1/2, j, k) = & E_x^n(i+1/2, j, k) \\
& + \frac{\Delta t}{\varepsilon_0} \left[\frac{H_z^{n+1/2}(i+1/2, j+1/2, k) - H_z^{n+1/2}(i+1/2, j-1/2, k)}{\Delta y} \right. \\
& - \frac{H_y^{n+1/2}(i+1/2, j, k+1/2) - H_y^{n+1/2}(i+1/2, j, k-1/2)}{\Delta z} \\
& \left. - J_x^{n+1/2}(i+1/2, j, k) \right] - \frac{1}{\varepsilon_0} [P_x^{n+1}(i+1/2, j, k) - P_x^n(i+1/2, j, k)] \quad (2.17)
\end{aligned}$$

$$\begin{aligned}
E_y^{n+1}(i, j+1/2, k) = & E_y^n(i, j+1/2, k) \\
& + \frac{\Delta t}{\varepsilon_0} \left[\frac{H_x^{n+1/2}(i, j+1/2, k+1/2) - H_x^{n+1/2}(i, j+1/2, k-1/2)}{\Delta z} \right. \\
& - \frac{H_z^{n+1/2}(i+1/2, j+1/2, k) - H_z^{n+1/2}(i-1/2, j+1/2, k)}{\Delta x} \\
& \left. - J_y^{n+1/2}(i, j+1/2, k) \right] - \frac{1}{\varepsilon_0} [P_y^{n+1}(i, j+1/2, k) - P_y^n(i, j+1/2, k)] \quad (2.18)
\end{aligned}$$

$$\begin{aligned}
E_z^{n+1}(i, j, k + 1/2) = & E_y^n(i, j, k + 1/2) \\
& + \frac{\Delta t}{\varepsilon_0} \left[\frac{H_y^{n+1/2}(i + 1/2, j, k + 1/2) - H_y^{n+1/2}(i - 1/2, j, k + 1/2)}{\Delta x} \right. \\
& - \frac{H_x^{n+1/2}(i, j + 1/2, k + 1/2) - H_x^{n+1/2}(i, j - 1/2, k + 1/2)}{\Delta y} \\
& \left. - J_z^{n+1/2}(i, j, k + 1/2) \right] - \frac{1}{\varepsilon_0} [P_z^{n+1}(i, j, k + 1/2) - P_z^n(i, j, k + 1/2)] \quad (2.19)
\end{aligned}$$

These update equations and the boundary conditions are the essential elements of a working FDTD simulation (although, without a wave source or monitors, it will not be beneficial). The boundary conditions are usually chosen to be Dirichlet (i.e. $\mathbf{E} = \mathbf{H} = 0$ at the boundaries), complemented by PMLs to minimize reflections from the simulation boundaries, or periodic (i.e. $\mathbf{E}(\text{near boundary}) = \mathbf{E}(\text{far boundary})$). For further information on PMLs and periodic boundary conditions, see Chapters 7 and 13, respectively, of Ref. [129]. For information on the implementation of wave sources in FDTD, a thorough discussion is provided in Chapter 5 of Ref. [129]. Finally, FDTD monitors are usually implemented as discrete Fourier transform (DFT) monitors. The usual implementation uses a running Fourier transform, where, in the monitor regions, the transform-integral over time is evaluated (as a discrete sum) in conjunction with the simulation rather than storing the fields as a function of time (which would require far more memory).

Prior to the simulation, all fields are allocated and set to zero. As described before, the fields are then updated in the order $\mathbf{H} \rightarrow \mathbf{J} \rightarrow \mathbf{P} \rightarrow \mathbf{E}$ using the above update equations (and the models required for $\mathbf{J}$ and $\mathbf{P}$ as discussed in the next section).

The step-size $\Delta x, y, z$ is commonly set using intuition; however, a few practicalities need to be considered. In FDTD, the dispersion relation of a wave travelling in a uniform medium depends on the step size [129]. It converges to the physical dispersion relation as $\Delta x, y, z \rightarrow 0$. A rule of thumb to reasonably reduce *numerical dispersion* is to set the step size less than the minimum wavelength of interest divided by twenty, $\Delta x, y, z \leq \lambda/20$, which is on the order of 20 nm for optical simulations. Especially in the case of nanophotonics, the step size should also be small enough to resolve the relevant features of the simulated system. For example, in Chapter 4, metal nanoparticles of sizes < 10 nm are simulated,

unresolved with a step size of 20 nm. A similar rule of thumb may be $\Delta x, y, z \leq a/20$, where a is the smallest feature size. That said, even smaller step sizes are required for accurate plasmonic simulations [77].

The timestep size Δt has a strict upper-bound set by the stability criteria,

$$\Delta t \leq \frac{1}{c_{max} \sqrt{\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2} + \frac{1}{\Delta z^2}}}, \quad (2.20)$$

where c_{max} is the maximum speed of light in the system (which can be assumed as the speed of light in vacuum c_0). In the case of a cubic grid ($\Delta x = \Delta y = \Delta z$), this condition becomes $\Delta t \leq \Delta x/(\sqrt{3}c_0)$.

2.1.2 High Performance Computing

A significant drawback of the FDTD method is that the computational cost scales on the fourth order of the step size since the number of computations per iteration scales with the space step sizes, and the number of iterations scales with the timestep, $\Delta x \Delta y \Delta z \Delta t \sim \Delta x^4$. If the step size is halved, the computational load is $2^4 = 16$ times larger. Unfortunately, the poor scalability frequently results in practitioners using step sizes that disobey the aforementioned rules of thumb and result in less-accurate simulations. Fortunately, FDTD lends itself well to high-performance computing (HPC). HPC refers to using non-standard computing hardware to speed up calculations, such as supercomputers or GPUs for numerical simulations. In the works published in this thesis, two high-performance FDTD codes were used - one for supercomputers and the other for a GPU. In this section, I will briefly highlight how both enhance FDTD performance.

The primary simulation software used in this thesis is our in-house 3D-FDTD software built for supercomputers using a domain decomposition approach wherein the simulation space is divided among N processors, which solve Maxwell's equations only in their respective subspaces. Because FDTD uses a rectangular mesh and field updates require only nearest neighbour cells, it lends itself well to domain-decomposition schemes. In the absence of overhead and bottlenecks, this should reduce the computation time by a factor of $1/N$, which is called *linear scaling* when achieved. It also solves the memory problem for extensive calculations, as each processor only needs to maintain the memory for fields

in its subdomain. However, electromagnetic field data must be exchanged between subdomains at each simulation timestep, made easier by the nearest neighbour scheme. For example, the magnetic field solved at the edge of subspace 1 requires the electric field from the adjacent edge of subspace 2. This data communication between processors is the purview of the Message Passing Interface (MPI), a standardized means of exchanging data between these processors. It consists of a helpful set of routines for communication and coordination between processors for coders who wish to harness the power of large, multiprocessor systems. For example, an MPI command sends an electric field array from one subdomain to an adjacent one. Other commands enable combining data from across multiple subdomains, which in our case, helps combine electric field data spanning several subdomains for electric field plots or other post-processing purposes.

The domain decomposition scheme used in our 3D-FDTD code is the subject of detailed discussion in Chapter 3. This discussion is in the context of the nonlocal plasmonic model, which demands more communication between processors than the standard (local) FDTD method.

A secondary code was developed during my graduate research, built to run on a GPU. A GPU also uses parallelization; instead of on subdomains, it is cell-by-cell. A GPU has thousands of cores available for parallel computation and is a shared-memory system wherein all cores have access to the same, limited memory. Beneficially, this means no interprocess communication protocols are required (on the single-GPU level). A GPU is well suited for repetitive, explicit calculations as it can launch thousands of threads to solve each calculation instance in parallel. That said, a single GPU thread is much slower than a CPU thread, so high performance requires liberal use of the GPU cores. For example, in the FDTD code, the electric and magnetic field update for all points in the mesh (Eq. 2.13) are calculated in parallel instead of looping through the domain as on a CPU.

This FDTD program was written in *C* and used the CUDA platform for GPU programming. This program was instrumental in validating the nonlocal Drude and hydrodynamic plasma models before implementation in the main MPI-based code, a much more tedious endeavour. Implementing material models on the GPU code is no more challenging than on a single-CPU code, and the results are obtained in shorter times for small test systems. For this reason, this code was also used for initial testing of other features, such as Bloch boundary conditions (used in Chapter 5), multipole decomposition (used in Chapter 9.1), and the adjoint method, before implementation in the main code. Finally, we make liberal use of this program to calculate the multipoles from the randomly generated plasmonic

nanoparticles as training data for a convolutional deep neural network in Chapter 9.1.

2.2 Material Models

Maxwell’s equations describe the propagation of electromagnetic fields and the interaction of these fields with matter. These interactions are accounted for via the polarization field $\mathbf{P}$ for bound/valence electrons and the current density $\mathbf{J}$ for free/conduction electrons. This is a matter of convention as the free-electron current is commonly described using a polarization field via $\mathbf{J} = \partial\mathbf{P}/\partial t$. This, however, is an approximation.

Indeed, the full current density, if derived microscopically, is described by a multipole expansion [135, 136],

$$\mathbf{J}_{total} = \mathbf{J}_{free} + \frac{\partial\mathbf{P}}{\partial t} - \frac{\partial}{\partial t}[\nabla \cdot \overleftrightarrow{Q}] + \nabla \times \mathbf{M} + \dots, \quad (2.21)$$

where $\mathbf{P}$ represents the macroscopic polarization density, $Q_{\alpha,\beta}$ represents the macroscopic quadrupole density tensor, and $\mathbf{M}$ is the macroscopic magnetization density. Higher order terms (octupole, hexadecapole, etc.) can also be derived.

The multipole expansion was highlighted here to point out that the response of a charge distribution to an external field, $\mathbf{J}_{total}$ is in general nonlocal – it depends on the distribution (space derivative) of $Q_{\alpha,\beta}$, $\mathbf{M}$ and higher order multipole terms. It is common and reasonable to make the “dipole approximation” where only the first two terms of Eq. 2.21 are kept, leaving us with the approximation that began this discussion,

$$\mathbf{J}_{total} = \mathbf{J}_{free} + \frac{\partial\mathbf{P}}{\partial t}, \quad (2.22)$$

The dipole approximation assumes that the field deviates negligibly over atomic length scales, which is generally reasonable since the wavelengths of light of interest are much larger than the molecules (atoms) that make up optical materials. This approximation has proven accurate in modelling structures with sizes much larger than the atomic length scales but seems to break down for structures < 10 nm in size, as discussed in Chapter 3.

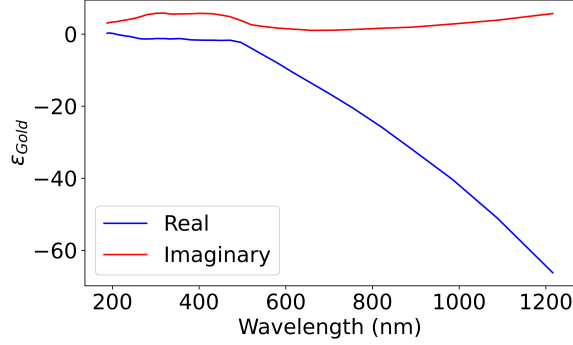


Figure 2.2: Real (blue) and imaginary (red) components of the frequency-dependent permittivity of bulk gold. In computational nanophotonics, this strong dispersion is simulated using material models. The numerical data is collected from the database of Ref. [137], but originates from Ref. [138].

Optical media are generally dispersive; their permittivity is frequency-dependent $\varepsilon(\omega)$ and complex. For example, we plot the relative permittivity of gold [137, 138] in Figure 2.2 over the visible spectrum. Note that the negative real component of the permittivity enables the plasmonic behaviour of gold [5]. To account for this complex material dispersion in the time domain, we must use Fourier-transformable mathematical models from the frequency domain (Figure 2.2) to the time domain.

In this section, we introduce three material models integral to theoretical and computational electromagnetism; 1.) the Drude model, 2.) the Lorentz model, and 3.) the N -critical points model. The Drude model is commonly used to model the response of free (conduction) electrons to an external field, whereas the latter two are commonly used for bound (valence) electrons. These models are considered *local* or *dipole* models as the material response ($\mathbf{J}$ and $\mathbf{P}$) is dependent only on the strength of the local electric field, $\mathbf{P}(\mathbf{r}, t) \sim \chi(\mathbf{r})\mathbf{E}(\mathbf{r}, t)$ rather than the formal $\mathbf{P}(\mathbf{r}, t) \sim \int \chi(\mathbf{r} - \mathbf{r}')\mathbf{E}(\mathbf{r}', t)d\mathbf{r}'$.

2.2.1 The Drude Model

The Drude model can be formally derived from the Boltzmann transport equation via a few approximations (see the supplementary information of Chapter 5) and from the Hamiltonian for an inhomogeneous electron gas [139]. Here then, we will derive the model from

physical arguments.

The goal is to describe the motion of free electrons in a time-varying electromagnetic field. We can derive the velocity $\mathbf{v}$ of a free electron with effective mass m^* from Newton's second law,

$$\frac{\mathbf{F}_{ext}}{m^*} = \frac{\partial \mathbf{v}}{\partial t}. \quad (2.23)$$

The external force $\mathbf{F}_{ext}$ needs to account for the effects of the electromagnetic field (the Lorentz force), and the effect of collisions (including electron–electron, electron–ion, and electron – impurity). The collisions can be roughly accounted for via an average collision rate γ . Physically, this represents the fact that the velocity of the electron (in the absence of an electromagnetic field) should drop to zero on time scales on the order of $\sim 1/\gamma$. This is a simple but effective model of a collision.

With the above considerations, Newton's second law becomes,

$$-\frac{e}{m^*}\mathbf{E} - \gamma\mathbf{v} = \frac{\partial \mathbf{v}}{\partial t}, \quad (2.24)$$

where $e = |e|$ is the magnitude of the charge of an electron, and $\mathbf{E}$ is the local electric field (including the incident and scattered fields). Note that we have ignored the magnetic field component of the Lorentz force, as it is negligible for low-intensity light (in the linear optics regime). Also notice what happens when $\mathbf{E} \rightarrow 0$; a decaying function describes the velocity field, $\mathbf{v} \sim \exp(-\gamma t)$ which, as previously mentioned, decays to zero on the order of $t \sim 1/\gamma$.

To convert this to a macroscopic optical model, we must account for the motion of the free electron ensemble. To do this, we use the current density, which is related to the velocity field via $\mathbf{J} = -en\mathbf{v}$ where n is the number density of the free electrons. Note that this relation ignores the effects of free carrier diffusion, a reasonable approximation on macroscopic scales where the free carriers are restricted to metal-dielectric interfaces. However, Chapter 3 discusses the generalized nonlocal optical response (GNOR) model for extremely small plasmonic particles, which accounts for this diffusion since the free carrier density is no longer negligible.

Our equation of motion becomes,

$$\frac{\partial \mathbf{J}}{\partial t} + \gamma \mathbf{J} = \frac{e^2 n}{m^*} \mathbf{E}. \quad (2.25)$$

Finally, we use the definition of the plasma frequency (derived in Refs. [5, 140, 141]) $\omega_p^2 = e^2 n / (m^* \varepsilon_0)$ to achieve the well-known time domain Drude model,

$$\frac{\partial \mathbf{J}}{\partial t} + \gamma \mathbf{J} = \varepsilon_0 \omega_p^2 \mathbf{E}, \quad (2.26)$$

or using the polarization field,

$$\frac{\partial^2 \mathbf{P}}{\partial t^2} + \gamma \frac{\partial \mathbf{P}}{\partial t} = \varepsilon_0 \omega_p^2 \mathbf{E}. \quad (2.27)$$

Both Eqs. 2.26 and 2.27 can be solved numerically in a time-domain electromagnetic simulator. This is covered in Section 2.2.5.

A significant and, to some, confusing approximation was made in the step between Eq. 2.24 and Eq. 2.25 when we transitioned from a single electron to the ensemble. We have claimed that the motion of a single electron is unaffected by the motion of the 'sea' of electrons around it (ignoring the ad-hoc collision term). However, one might expect that additional considerations are required for the high electron densities of noble metals $n \sim 10^{28} \text{ m}^{-3}$.

The reason we can ignore interaction is due to charge-screening effects. The electric potential ϕ of a single charge in a plasma is highest in its local environment but decays exponentially $\phi(\mathbf{r}) \sim \exp(-|\mathbf{r}|/\lambda_{TF})$ where λ_{TF} is the Thomas-Fermi wavelength (see Chapter 12.1.3 of Ref. [5]). The potential of a given electron will not be felt by charges at distances $|\mathbf{r}| \gg \lambda_{TF}$. For noble metals like gold, $\lambda_{TF} \sim 0.6 \text{ \AA}$ is smaller than the mean distance between electrons as defined by the electron density $r_{mean} \sim 1.6 \text{ \AA}$. It is because of this that we can ignore the effects of electron-electron interactions in our model.

Finally, a Fourier transform can convert the Drude model to the frequency domain. However, since the model is linear, an equivalent and simpler method involves assuming the fields are monochromatic; $\mathbf{E}(t) = \mathbf{E}(\omega)e^{-i\omega t}$, and $\mathbf{P}(t) = \mathbf{P}(\omega)e^{-i\omega t}$. Plugging these into Eq. 2.27 gives us

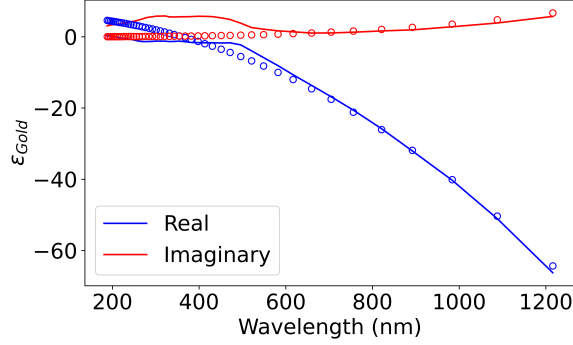


Figure 2.3: Real (blue) and imaginary (red) components of the gold permittivity spectrum (solid lines) with a Drude model fit to this data (circles) where the fitting parameters are $\varepsilon_\infty = 7.79$, $\omega_p = 1.33e16$ rad/s, and $\Gamma = 1.14e14$ rad/s.

$$\mathbf{P}(\omega) = -\varepsilon_0 \frac{\omega_p^2}{\omega^2 + i\gamma\omega} \mathbf{E}(\omega), \quad (2.28)$$

which provides an expression for the susceptibility,

$$\chi_{Drude}(\omega) = -\frac{\omega_p^2}{\omega^2 + i\gamma\omega}, \quad (2.29)$$

from which the permittivity can be derived,

$$\varepsilon_{Drude}(\omega) = \varepsilon_\infty - \frac{\omega_p^2}{\omega^2 + i\gamma\omega}, \quad (2.30)$$

where we have introduced a final parameter, ε_∞ , the high-frequency permittivity. In theory, we expect $\varepsilon_\infty = 1$; however, this and the following models are very simple and generally require extra tuning.

In Figure 2.3, we fit experimental data for gold (from Figure 2.2) to Eq. 2.30 to extract values for the plasma frequency ω_p , the damping coefficient γ , and the high-frequency permittivity ε_∞ . These values (provided in the caption) can be used in Eqs. 2.26 or 2.27 to simulate real materials in FDTD. Fitting for this section and the following was accomplished by minimizing the difference between the model and authentic permittivity spectra

using a gradient-based minimizer [142] with automatic differentiation [132] providing the gradient.

From Figure 2.3, we can see that the Drude model accurately represents the permittivity of gold for wavelengths $\lambda > 600$ nm. This bandwidth is the intraband regime since the conduction electrons dominate the optical response, and there are no interband transitions. However, in the higher frequencies (lower wavelengths), electrons in the valence band may absorb a photon and join the conduction band. The Drude model does not account for this resonant behaviour; thus, we must look to resonator models next.

2.2.2 The Lorentz Model

Wherein the Drude model represents the motion of free electrons, the Lorentz model represents the motion of bound electrons. The derivation covered in many student-level texts (see Refs. [136, 143, 144], for example) is usually based on a harmonic-oscillator model where the bound electron cloud of the atom is the oscillator, attached to a stationary object (the nucleus) by a restoring force (the Coulomb force). However, in this section, we will adapt the Drude model using simple physical arguments again.

First, we will return to the model for the velocity of the electron in a time-varying field (Eq. 2.24) and replace the velocity $\mathbf{v}$ with the derivative of the electron's position $\partial\mathbf{x}/\partial t$ to get,

$$-\frac{e}{m_b^*}\mathbf{E} - \Gamma\frac{\partial\mathbf{x}}{\partial t} = \frac{\partial^2\mathbf{x}}{\partial t^2}, \quad (2.31)$$

where we differentiate between the free electron γ and m^* and the bound electron Γ and m_b^* .

From a classical mechanics point of view, the essential difference between the free and bound electrons is the presence of a restoring force due to a non-negligible force between the nucleus and the bound electrons. As bound electrons respond resonantly, the restoring force is $\mathbf{F}_{restoring}/m_b^* = -\omega_0^2\mathbf{x}(t)$ where ω_0 is the resonance frequency of an electron resonance (interband transitions).

The equation of motion then becomes,

$$\frac{\partial^2\mathbf{x}}{\partial t^2} + \Gamma\frac{\partial\mathbf{x}}{\partial t} + \omega_0^2\mathbf{x}(t) = -\frac{e}{m_b^*}\mathbf{E}. \quad (2.32)$$

This is the equation of a damped, driven harmonic oscillator.

Again, we must convert from the single electron – atom picture to that of a collection of atomic dipoles. We do this by first recognizing that the dipole moment of a single electron-ion system is $\mathbf{p}(t) = -e\mathbf{x}(t)$, and thus the polarization density $\mathbf{P}$ is the number density of atomic dipoles multiplied by the dipole moment of a single electronic dipole, $P = fN\mathbf{p} = -feN\mathbf{x}(t)$ where N is the molecular/atomic number density, and f is the number of electrons in each molecule/atom with the same resonance frequency. We again use Coulomb screening to justify ignorance of electron-electron interaction. The Lorentz model for the bound electron polarization density is,

$$\frac{\partial^2 \mathbf{P}}{\partial t^2} + \Gamma \frac{\partial \mathbf{P}}{\partial t} + \omega_0^2 \mathbf{P}(t) = f \frac{e^2 N}{m_b^*} \mathbf{E}. \quad (2.33)$$

Eq. 2.33 is solved numerically in FDTD programs – typically to simulate dispersion in dielectrics or to model the bound electron subsystem in conductors and semiconductors. Using the same harmonic approximation as in the previous section, we can obtain the frequency-dependent susceptibility function,

$$\chi_{Lorentz}(\omega) = \frac{e^2 N}{m_b^*} \frac{f}{\omega_0^2 - \omega^2 - i\Gamma\omega}. \quad (2.34)$$

Of course, molecules and atoms contain multiple electronic resonances, each at frequencies ω_{0_j} with linewidth Γ_j , shared by f_j electrons. In general,

$$\chi_{Lorentz}(\omega) = \frac{e^2 N}{m_b^*} \sum_j \frac{f_j}{\omega_{0_j}^2 - \omega^2 - i\Gamma_j\omega}. \quad (2.35)$$

As with the Drude model, the parameters in the Lorentz model can be fit to experimental data and can then be used in an FDTD program to simulate the dispersive material accurately. It is helpful to define a frequency $\Omega_j^2 = e^2 N f_j / (\epsilon_0 m_b^*)$ to reduce the number of fitting parameters.

In Figure 2.4, we fit the low-wavelength permittivity spectrum to a single-resonance model, which we can see captures some of the activity occurring in the low wavelength regions $\lambda < 500$ nm, thus called the interband regime. The interband transition corresponds to a wavelength of $\lambda \approx 500$ nm, which our model captures. A single Lorentz resonance is

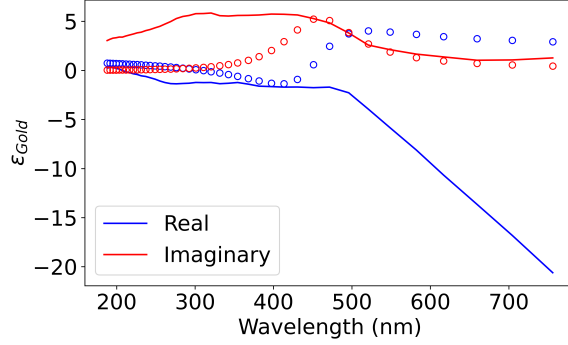


Figure 2.4: Real (blue) and imaginary (red) components of the gold permittivity spectrum (solid lines) with a Lorentz model fit to this data (circles) where the fitting parameters are $\varepsilon_\infty = 1$, $\Omega = 4.69e16$ rad/s, $\omega_0 = 4.14e16$ rad/s, and $\Gamma = 1e16$ rad/s.

insufficient; however, in the next section, we will combine two to better fit the interband transition regime.

2.2.3 The Drude-Lorentz model

As shown in the previous two sections (Figures 2.3 and 2.4), the optical response of conductors and semiconductors have contributions from both free and bound electrons. For such materials, it is most common to use a combined Drude – Lorentz model, where,

$$\chi_{DL}(\omega) = -\frac{\omega_p^2}{\omega^2 + i\gamma\omega} + \sum_j \frac{\Omega_j^2}{\omega_{0j}^2 - \omega^2 - i\Gamma_j\omega}. \quad (2.36)$$

Finally, the relative permittivity becomes,

$$\varepsilon_{DL}(\omega) = \varepsilon_\infty - \frac{\omega_p^2}{\omega^2 + i\gamma\omega} + \sum_j \frac{\Omega_j^2}{\omega_{0j}^2 - \omega^2 - i\Gamma_j\omega}. \quad (2.37)$$

We can treat this model in the time domain using the two PDEs introduced in the previous sections. Combining the Drude, Lorentz, and Maxwell's curl equations, we can self-consistently simulate the interaction of light with a non-magnetic, linear, dispersive material,

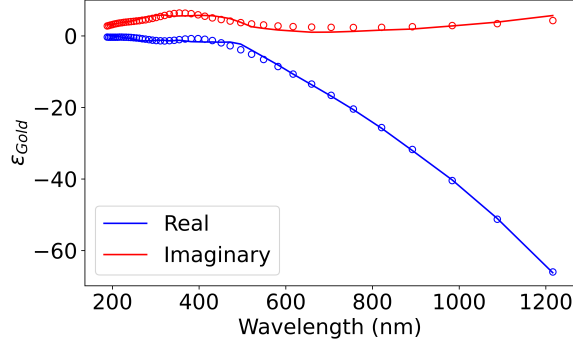


Figure 2.5: Real (blue) and imaginary (red) components of the gold permittivity spectrum (solid lines) with a combined Drude + 2 Lorentz model where the fitting parameters are given by, $\varepsilon_\infty = 0.174$, $\omega_p = 1.33e16$ rad/s, $\gamma = 5.36e13$ rad/s, $\Omega_1 = 5.13e16$ rad/s, $\omega_{0_1} = 2.48e16$ rad/s, $\Gamma_1 = 3.97e16$ rad/s, $\Omega_2 = 1.06e16$ rad/s, $\omega_{0_2} = 5.84e15$ rad/s, $\Gamma_2 = 4.24e15$ rad/s.

$$\frac{\partial^2 \mathbf{P}_{L_j}}{\partial t^2} + \Gamma_j \frac{\partial \mathbf{P}_{L_j}}{\partial t} + \omega_{0_j}^2 \mathbf{P}_{L_j}(t) = \varepsilon_0 \Omega_j^2 \mathbf{E}(t), \quad (2.38)$$

$$\frac{\partial^2 \mathbf{P}_D}{\partial t^2} + \gamma \frac{\partial \mathbf{P}_D}{\partial t} = \varepsilon_0 \omega_p^2 \mathbf{E}(t), \quad (2.39)$$

$$\varepsilon_0 \varepsilon_\infty \frac{\partial \mathbf{E}}{\partial t} = \nabla \times \mathbf{H} - \frac{\partial \mathbf{P}_D}{\partial t} - \sum_j \frac{\partial \mathbf{P}_{L_j}}{\partial t}, \quad (2.40)$$

$$\mu_0 \frac{\partial \mathbf{H}}{\partial t} = -\nabla \times \mathbf{E}, \quad (2.41)$$

where we have separated the polarization density into the Drude and multiple Lorentz components, $\mathbf{P} = \mathbf{P}_D + \sum_j \mathbf{P}_{L_j}$, and incorporate ε_∞ directly into Maxwell's equation.

In Figure 2.5, we fit experimental data from Figure 2.2 to Eq. 2.37 with two Lorentz terms to extract the fitting parameters. The value of these parameters (provided in the caption) can then be used in Eqs. 2.38 and 2.39 to accurately simulate gold in FDTD.

From Figure 2.5, we can see that the Drude + 2 Lorentz model accurately represents the permittivity of gold over the bandwidth of interest, thus illustrating how we must account for both inter and intra-band contributions in plasmonic metals.

2.2.4 The critical points model

The previous section dealt with the derivation of the Lorentz model for bound-electron dispersion. Although the Lorentz model is physically well-grounded, when applied to highly-dispersive materials (with many intraband transitions to account for), its lack of flexibility becomes a problem. This section presents a model better suited for highly-dispersive materials (e.g. plasmonic materials) called the critical points model. Introduced as a more efficient way to describe the dielectric response of Silicon-based compounds, the critical points model has no physical basis besides the Lorentz-like form [145]. However, it has been shown to provide a more accurate fit than a general (N-term) Lorentz model, especially for plasmonic materials, when combined with the Drude model [146–149].

This model was originally derived in the frequency domain and is given by,

$$\chi_{CP}(\omega) = \sum_j A_j \Omega_j \left(\frac{e^{i\phi_j}}{\Omega_j - \omega - i\Gamma_j} + \frac{e^{-i\phi_j}}{\Omega_j + \omega + i\Gamma_j} \right) \quad (2.42)$$

where each term of the sum is called a critical point. This function allows for much more flexible fitting when compared to the Lorentz model, and it respects the Kramers-Kronig relations [145].

The application of this model to the time domain requires the Fourier transform of the following equation,

$$\mathbf{P}_j(\omega) = \varepsilon_0 A_j \Omega_j \left(\frac{e^{i\phi_j}}{\Omega_j - \omega - i\Gamma_j} + \frac{e^{-i\phi_j}}{\Omega_j + \omega + i\Gamma_j} \right) \mathbf{E}(\omega). \quad (2.43)$$

To do this, we first combine the two terms,

$$\mathbf{P}_j(\omega) = 2\varepsilon_0 A_j \Omega_j \left(\frac{\Omega_j \cos(\phi_j) - \Gamma_j \sin(\phi_j) + i\omega \sin(\phi_j)}{\Omega_j^2 - \omega^2 + \Gamma_j^2 - 2i\omega\Gamma_j} \right) \mathbf{E}(\omega), \quad (2.44)$$

and rearrange,

$$(\Omega_j^2 - \omega^2 + \Gamma_j^2 - 2i\omega\Gamma_j) \mathbf{P}_j(\omega) = 2\varepsilon_0 A_j \Omega_j (\Omega_j \cos(\phi_j) - \Gamma_j \sin(\phi_j) + i\omega \sin(\phi_j)) \mathbf{E}(\omega). \quad (2.45)$$

We convert this to the time domain by using the Fourier transform rules $-i\omega \xrightarrow{\text{FT}} \partial/\partial t$ and $-\omega^2 \xrightarrow{\text{FT}} \partial^2/\partial t^2$, which gives us the following time-domain model,

$$\begin{aligned} \frac{\partial^2 \mathbf{P}_j}{\partial t^2} + 2\Gamma_j \frac{\partial \mathbf{P}_j}{\partial t} + (\Omega_j^2 + \Gamma_j^2) \mathbf{P}_j(t) \\ = 2\varepsilon_0 A_j \Omega_j (\Omega_j \cos(\phi_j) - \Gamma_j \sin(\phi_j)) \mathbf{E}(t) - 2\varepsilon_0 A_j \Omega_j \sin(\phi_j) \frac{\partial \mathbf{E}}{\partial t}. \end{aligned} \quad (2.46)$$

The primary change offered by the critical points model in comparison to the Lorentz model is the time derivative of the electric field on the RHS. Note that there is no physical explanation for this additional term aside from an apparent ad-hoc higher-order correction,

$$\frac{\partial^2 \mathbf{P}_j}{\partial t^2} + \hat{\Gamma}_j \frac{\partial \mathbf{P}_j}{\partial t} + \hat{\omega}_{0j}^2 \mathbf{P}_j(t) = \varepsilon_0 \hat{\Omega}_j^2 \mathbf{E}(t) + C_j \frac{\partial \mathbf{E}}{\partial t}, \quad (2.47)$$

where C_j is the correction constant, and the hatted quantities are Lorentz model parameters (calculated via comparison of Eqs. 2.46 and 2.47). Furthermore, C_j is only non-zero when ϕ_j is non-zero. Eq. 2.42 is a Lorentz model with an extra fitting parameter that physically manifests as a correction to the driving/restoring force.

For gold, we use a Drude + 2 critical points model whose permittivity can be constructed by replacing the Lorentz term in Eq. 2.37 with Eq. 2.42 where $j \in [1, 2]$. As in previous sections, we fit the Gold permittivity data to the Drude + 2 critical points model plotted in Figure 2.6, and the fitting parameters are provided in the caption. The mean absolute error between the model fit and the experimental data is $2\times$ less than from the Drude + 2 Lorentz model in Figure 2.5.

2.2.5 Numerical Solution of Drude and Lorentz models

This section shows how the Drude and Lorentz models are solved in our FDTD simulators. Note that there are a few methods for solving material models in FDTD (see Chapter 9 of Ref. [129]). However, we use the explicit auxiliary differential equation (ADE) method, which uses finite differences to solve the time-domain differential equations.

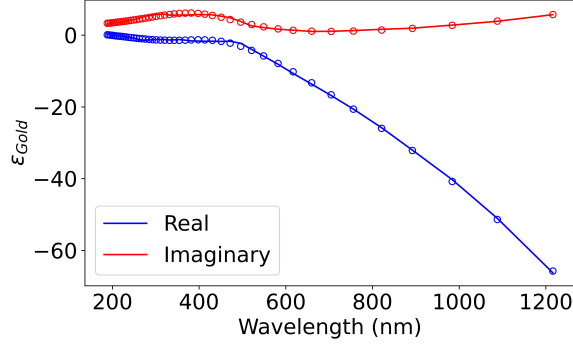


Figure 2.6: Real (blue) and imaginary (red) components of the gold permittivity spectrum (solid lines) with a Drude + 2 critical points model fit to this data (circles) where the fitting parameters are $\varepsilon_\infty = 0.005$, $\omega_p = 1.31e16$ rad/s, $\gamma = 1.40e14$ rad/s, $\Omega_1 = 1.34e16$ rad/s, $A_1 = 0.45$, $\Gamma_1 = 0.0$ rad/s, $\Omega_2 = 3.95e15$ rad/s, $A_2 = 3.16$, $\Gamma_2 = 1.76e15$ rad/s.

Recall that the polarization density fields $\mathbf{P}_L$, $\mathbf{P}_D$, are collocated in space and time with the local electric field $\mathbf{E}$. The polarization fields are thus updated at the integer $n + 1$ timesteps. Applying finite differences to the Lorentz model gives us,

$$\frac{\mathbf{P}_j^{n+1} - 2\mathbf{P}_j^n + \mathbf{P}_j^{n-1}}{\Delta t^2} + \Gamma_j \frac{\mathbf{P}_j^{n+1} - \mathbf{P}_j^{n-1}}{2\Delta t} + \omega_{0j}^2 \mathbf{P}_j^n = \varepsilon_0 \Omega_j^2 \mathbf{E}^n, \quad (2.48)$$

where we ignore the L subscript for simplicity of notation. This equation is solved for $\mathbf{P}_j^{n+1}$, thus providing an explicit, second-order, update equation for the bound electron polarization density field,

$$\mathbf{P}_j^{n+1} \left(\frac{\Gamma_j}{2\Delta t} + \frac{1}{\Delta t^2} \right) = \mathbf{P}_j^n \left(\frac{2}{\Delta t^2} - \omega_{0j}^2 \right) + \mathbf{P}_j^{n-1} \left(\frac{\Gamma_j}{2\Delta t} - \frac{1}{\Delta t^2} \right) + \varepsilon_0 \Omega_j^2 \mathbf{E}^n. \quad (2.49)$$

Similarly, the Drude model update equation can be derived by setting $\omega_{0j}^2 = 0$ and changing the notation,

$$\mathbf{P}_D^{n+1} \left(\frac{\gamma}{2\Delta t} + \frac{1}{\Delta t^2} \right) = \mathbf{P}_D^n \left(\frac{2}{\Delta t^2} \right) + \mathbf{P}_D^{n-1} \left(\frac{\gamma}{2\Delta t} - \frac{1}{\Delta t^2} \right) + \varepsilon_0 \omega_p^2 \mathbf{E}^n. \quad (2.50)$$

These, combined with the electric and magnetic field update equations (Eqs. 2.12 and 2.13), provide the complete algorithm for simulating dispersive materials in FDTD. As

mentioned previously, the solution order is $\mathbf{H}^{n+1/2} \rightarrow \mathbf{P}^{n+1} \rightarrow \mathbf{E}^{n+1}$. It is necessary to store the $\mathbf{P}^{n+1}$ and $\mathbf{P}^n$ field arrays. The $\mathbf{P}^{n-1}$ array can be overwritten by $\mathbf{P}^{n+1}$.

The main downside of this naïve but effective implementation is that it is not necessarily numerically stable, as discussed in Ref. [150]. An implementation that is stable within the FDTD stability limit involves weighted centred averaging where non-derivative terms are replaced by $\mathbf{A}^n = (\mathbf{A}^{n+1} + 2\mathbf{A}^n + \mathbf{A}^{n-1})/4$. The Lorentz model in finite difference form, Eq. 2.48, becomes,

$$\begin{aligned} \frac{\mathbf{P}_j^{n+1} - 2\mathbf{P}_j^n + \mathbf{P}_j^{n-1}}{\Delta t^2} + \Gamma_j \frac{\mathbf{P}_j^{n+1} - \mathbf{P}_j^{n-1}}{2\Delta t} + \frac{\omega_{0j}^2}{4} (\mathbf{P}_j^{n+1} + 2\mathbf{P}_j^n + \mathbf{P}_j^{n-1}) \\ = \frac{\varepsilon_0 \Omega_j^2}{4} (\mathbf{E}^{n+1} + 2\mathbf{E}^n + \mathbf{E}^{n-1}). \end{aligned} \quad (2.51)$$

The right-hand side shows that we now require $\mathbf{E}^{n+1}$, seemingly making our algorithm implicit. However, through some algebra, we can work around it. We first derive our update equation for $\mathbf{P}_j^{n+1}$,

$$\begin{aligned} \mathbf{P}_j^{n+1} \left(\frac{\Gamma_j}{2\Delta t} + \frac{1}{\Delta t^2} + \frac{\omega_{0j}^2}{4} \right) = \mathbf{P}_j^n \left(\frac{2}{\Delta t^2} - \frac{\omega_{0j}^2}{2} \right) \\ + \mathbf{P}_j^{n-1} \left(\frac{\Gamma_j}{2\Delta t} - \frac{1}{\Delta t^2} - \frac{\omega_{0j}^2}{4} \right) + \frac{\varepsilon_0 \Omega_j^2}{4} (\mathbf{E}^{n+1} + 2\mathbf{E}^n + \mathbf{E}^{n-1}), \end{aligned} \quad (2.52)$$

which can be refactored as,

$$\mathbf{P}_j^{n+1} = C_{0,j} \mathbf{P}_j^n + C_{1,j} \mathbf{P}_j^{n-1} + C_{2,j} (\mathbf{E}^{n+1} + 2\mathbf{E}^n + \mathbf{E}^{n-1}), \quad (2.53)$$

where the $C_{i,j}$ parameters are trivial to derive by comparing Eq. 2.52 to Eq. 2.53. Next, we substitute the RHS of Eq. 2.53 into Eq. 2.13,

$$\begin{aligned} \nabla \times \mathbf{H}^{n+1/2} = \varepsilon_0 \varepsilon_\infty \frac{\mathbf{E}^{n+1} - \mathbf{E}^n}{\Delta t} + \mathbf{J}^{n+1/2} \\ + \sum_j \frac{C_{0,j} \mathbf{P}_j^n + C_{1,j} \mathbf{P}_j^{n-1} + C_{2,j} (\mathbf{E}^{n+1} + 2\mathbf{E}^n + \mathbf{E}^{n-1}) - \mathbf{P}_j^n}{\Delta t}, \end{aligned} \quad (2.54)$$

and solve for $\mathbf{E}^{n+1}$,

$$\begin{aligned} \mathbf{E}^{n+1} \left(\frac{\varepsilon_0 \varepsilon_\infty}{\Delta t} + \sum_j \frac{C_{2,j}}{\Delta t} \right) &= \mathbf{E}^n \left(\frac{\varepsilon_0 \varepsilon_\infty}{\Delta t} - 2 \sum_j \frac{C_{2,j}}{\Delta t} \right) \\ &- \mathbf{E}^{n-1} \sum_j \left(\frac{C_{2,j}}{\Delta t} \right) + \nabla \times \mathbf{H}^{n+1/2} + \sum_j \mathbf{P}_j^n \left(\frac{1}{\Delta t} - \frac{C_{0,j}}{\Delta t} \right) - \sum_j \mathbf{P}_j^{n-1} \left(\frac{C_{1,j}}{\Delta t} \right). \end{aligned} \quad (2.55)$$

In this case, the solution order changes $\mathbf{H}^{n+1/2} \rightarrow \mathbf{E}^{n+1} \rightarrow \mathbf{P}^{n+1}$. Note that the above can be applied to the Drude model as well by setting $\omega_0 = 0$, $\Gamma \rightarrow \gamma$, and $\Omega \rightarrow \omega_p$ in the calculation of the $C_{i,j}$ terms.

2.2.6 Numerical solution of the critical points model

Eq. 2.47 can be discretized for a finite difference solution,

$$\begin{aligned} \frac{\mathbf{P}_j^{n+1} - 2\mathbf{P}_j^n + \mathbf{P}_j^{n-1}}{\Delta t^2} + \hat{\Gamma}_j \frac{\mathbf{P}_j^{n+1} - \mathbf{P}_j^{n-1}}{2\Delta t} + \frac{\hat{\omega}_{0j}^2}{4} (\mathbf{P}_j^{n+1} + 2\mathbf{P}_j^n + \mathbf{P}_j^{n-1}) \\ = \frac{\varepsilon_0 \hat{\Omega}_j^2}{4} (\mathbf{E}^{n+1} + 2\mathbf{E}^n + \mathbf{E}^{n-1}) + C_j \frac{\mathbf{E}^{n+1} - \mathbf{E}^{n-1}}{2\Delta t}, \end{aligned} \quad (2.56)$$

and solved for $\mathbf{P}_j^{n+1}$,

$$\begin{aligned} \mathbf{P}_j^{n+1} \left(\frac{\hat{\Gamma}_j}{2\Delta t} + \frac{1}{\Delta t^2} + \frac{\hat{\omega}_{0j}^2}{4} \right) &= \mathbf{P}_j^n \left(\frac{2}{\Delta t^2} - \frac{\hat{\omega}_{0j}^2}{2} \right) + \mathbf{P}_j^{n-1} \left(\frac{\hat{\Gamma}_j}{2\Delta t} - \frac{1}{\Delta t^2} - \frac{\hat{\omega}_{0j}^2}{4} \right) \\ &+ \mathbf{E}^{n+1} \left(\frac{\varepsilon_0 \hat{\Omega}_j^2}{4} + \frac{C_j}{2\Delta t} \right) + \mathbf{E}^n \left(\frac{\varepsilon_0 \hat{\Omega}_j^2}{2} \right) + \mathbf{E}^{n-1} \left(\frac{\varepsilon_0 \hat{\Omega}_j^2}{4} - \frac{C_j}{2\Delta t} \right) \end{aligned} \quad (2.57)$$

As with the Drude and Lorentz models, we can simplify this by introducing constant factors,

$$\mathbf{P}_j^{n+1} = C_{0,j} \mathbf{P}_j^n + C_{1,j} \mathbf{P}_j^{n-1} + C_{2,j} \mathbf{E}^{n+1} + C_{3,j} \mathbf{E}^n + C_{4,j} \mathbf{E}^{n-1}. \quad (2.58)$$

This model is updated in the same way as the weighted centred averaged Drude and Lorentz models in the previous section, where the RHS of Eq. 2.58 is substituted into the electric-field update Eq. 2.13.

For plasmonic materials, the critical points model is solved instead of a Lorentz model (for bound electrons), along with the Drude model, known as the Drude + N critical points model. In Ref. [148], a table of numerical values are provided for the fitting parameters of the Drude + 2 critical points (D2CP) model for silver, gold, aluminum, and chromium. The D2CP model complemented with the table of Ref. [148] is used in our in-house FDTD program – the primary tool behind all of the work in this thesis.

2.2.7 Hydrodynamic plasma model

In the previous sections, we introduced three linear models that utilize the dipole (local response) approximation. However, more advanced models depend also on the distribution of the local electric field, thus including higher-order multipoles. This section introduces one such model, the hydrodynamic plasma model. This model treats the free-electron subsystem as a compressible fluid, the velocity field $\mathbf{v}$ of which obeys Euler’s fluid equations for ideal (non-viscous) flow [151],

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} = \mathbf{a} - \frac{1}{\rho} \nabla p, \quad (2.59)$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \quad (2.60)$$

where $\mathbf{a}$ is the external acceleration (usually due to gravity $\mathbf{g}$ in fluid dynamics), ρ is the fluid density and p is the fluid pressure. The first equation is Newton’s second law applied to continuous matter, and the second is the continuity equation for the conservation of mass.

I will briefly discuss the meaning of the left side of Eq. 2.59. In continuous media, the total time derivative of a quantity c is called the material time derivative, $dc/dt = \partial c/\partial t + (\mathbf{v} \cdot \nabla)c$ where the second term is called the advective rate of change of c . It results from the fact that we are describing how the fluid particles (the electrons, in our case) experience the change in c . As the fluid particles are moving, they observe $c(\mathbf{x}, t)$ to change according to the chain rule as $dc(\mathbf{x}, t)/dt = \partial c/\partial t + (\partial \mathbf{x}/\partial t \cdot \partial c/\partial \mathbf{x})$ where $\partial \mathbf{x}/\partial t$

is the velocity of the fluid particle making the measurement.

As an example, the velocity field of a river at an exact point $\mathbf{x}_0$ might not show any local acceleration $\partial\mathbf{v}/\partial t|_{\mathbf{x}_0} = 0$ as measured from a stationary position. However, if we measure the velocity field at different positions around $\mathbf{x}_0$, we may find that the velocity field is not constant, $\nabla\mathbf{v}|_{\mathbf{x}_0} \neq 0$. Thus, a water molecule passing through $\mathbf{x}_0$ will experience a change in velocity due to its change in position, despite $\partial\mathbf{v}/\partial t|_{\mathbf{x}_0} = 0$ remaining true. This is the advective rate of change, and when applied to the velocity field $(\mathbf{v} \cdot \nabla)\mathbf{v}$, the advective rate is also called the convective rate of change.

The applicability of this equation to the free electrons is not necessarily obvious; however, it can be derived from the Boltzmann transport equation and the Hamiltonian for an incompressible electron gas [139, 152] and has proven itself to reproduce experimental observations accurately [13] including in Chapter 4. It can predict a myriad of nonlinear effects from plasmonic systems, namely harmonic generation.

To apply this model to plasmonic (electromagnetic) systems, we must first convert the velocity field to a current density via $\mathbf{J} = -en\mathbf{v}$ where n is the free electron number density. The mass density ρ is converted to a number density $n = \rho/m^*$ where m^* is the effective electron mass. The acceleration term is proportional to the Lorentz force, and a resistance term accounts for damping,

$$\mathbf{a} = -\frac{e}{m^*} \left(\mathbf{E} - \frac{\mu_0}{ne} \mathbf{J} \times \mathbf{H} \right) + \frac{\gamma}{en} \mathbf{J}. \quad (2.61)$$

Finally, the pressure is taken to be the electron degeneracy pressure (a.k.a the Fermi pressure), $p = 2/5 n_0 E_F (n/n_0)^{5/3}$ where E_F is the Fermi energy, and $n_0 = n(\mathbf{r}, t \rightarrow -\infty)$ is the equilibrium electron number density, assumed to be constant. Although this assumption has been challenged for small particles [153, 154], we maintain it in our software to avoid pre-calculation of $n_0(\mathbf{r})$ which typically results in sub-step-size (sub-nanometer) features.

With these adaptations, Euler's equations become,

$$\begin{aligned} \frac{\partial \mathbf{J}}{\partial t} + \gamma \mathbf{J} - \frac{\mathbf{J}}{en} (\nabla \cdot \mathbf{J}) - \frac{1}{en} (\mathbf{J} \cdot \nabla) \mathbf{J} - \frac{\mathbf{J}}{en^2} (\mathbf{J} \cdot \nabla) n \\ = \frac{ne^2}{m^*} \left(\mathbf{E} - \frac{\mu_0}{ne} \mathbf{J} \times \mathbf{H} \right) + \frac{env_F^2}{3} \left(\frac{n}{n_0} \right)^{2/3} \nabla n, \end{aligned} \quad (2.62)$$

$$\frac{\partial n}{\partial t} - \frac{1}{e} \nabla \cdot \mathbf{J} = 0, \quad (2.63)$$

where $v_F = \sqrt{2E_F/m^*}$ is the Fermi velocity.

This is the full hydrodynamic model, which can be solved directly using finite differences, as discussed later. However, some approximations can be made to reduce the computational complexity, especially in cases where specific nonlinear effects are being targeted.

From here, I will use the fact that $n(\mathbf{r}, t) = n_0 + \eta(\mathbf{r}, t)$ where $\eta \ll n_0$, to apply a first-order expansion to the $1/n$ terms,

$$\begin{aligned} \frac{\partial \mathbf{J}}{\partial t} + \gamma \mathbf{J} - \frac{\mathbf{J}}{en_0} \left[1 - \frac{\eta}{n_0} \right] (\nabla \cdot \mathbf{J}) - \frac{1}{en_0} \left[1 - \frac{\eta}{n_0} \right] (\mathbf{J} \cdot \nabla) \mathbf{J} - \frac{\mathbf{J}}{en_0^2} \left[1 - 2\frac{\eta}{n_0} \right] (\mathbf{J} \cdot \nabla) \eta \\ = \frac{[n_0 + \eta]e^2}{m^*} \mathbf{E} - \frac{e\mu_0}{m^*} \mathbf{J} \times \mathbf{H} + \frac{e[n_0 + \eta]v_F^2}{3} \left[1 + \frac{2}{3} \frac{\eta}{n_0} \right] \nabla \eta. \end{aligned} \quad (2.64)$$

Although closed-form solutions for the hydrodynamic model do not exist, we can treat it using a perturbative approach in this form. Here, the current density field and electron density perturbations are given by,

$$\mathbf{J} = \lambda \mathbf{J}_1 + \lambda^2 \mathbf{J}_2 + \dots, \quad (2.65)$$

$$\eta = \lambda \eta_1 + \lambda^2 \eta_2 + \dots, \quad (2.66)$$

where λ is the perturbation strength. For now, we will assume that the electric and magnetic fields are only first-order $\mathbf{E} = \lambda \mathbf{E}_1$ and $\mathbf{H} = \lambda \mathbf{H}_1$. Plugging this into Eq. 2.64 and separating into terms of the same power of λ , we calculate the first two perturbation orders to be,

$$\frac{\partial \mathbf{J}_1}{\partial t} + \gamma \mathbf{J}_1 = \varepsilon_0 \omega_p^2 \mathbf{E}_1, \quad (2.67)$$

$$\frac{\partial \mathbf{J}_2}{\partial t} + \gamma \mathbf{J}_2 - \frac{1}{en_0} [(\nabla \cdot \mathbf{J}_1) \mathbf{J}_1 + (\mathbf{J}_1 \cdot \nabla) \mathbf{J}_1] = \frac{\eta_1 e^2}{m^*} \mathbf{E}_1 - \frac{e\mu_0}{m^*} \mathbf{J}_1 \times \mathbf{H}_1. \quad (2.68)$$

The first-order solution is the Drude model, discussed in Section 2.2.1. Note that for the sake of this discussion, we are assuming that $\nabla\eta$ is negligible – a generally true statement in bulk media, except in cases of minimal feature sizes and within atomic layers of metal-dielectric interfaces. Accounting for the pressure term, $\nabla\eta$, is the subject of Chapter 3. In Chapter 4, we use Eqs. 2.67 and 2.68 to quantify SHG from gold nanoantennas. Eq. 2.67 can be solved for $\mathbf{J}_1$ via the Fourier transform,

$$\mathbf{J}_1(\omega) = \frac{\varepsilon_0 \omega_p^2}{-i\omega + \gamma} \mathbf{E}_1(\omega). \quad (2.69)$$

The nonlinearity in Eq. 2.68 prevents a simple Fourier-based solution. However, a solution for $\mathbf{J}_2(2\omega)$ is possible if we assume the primary driving fields are monochromatic, $\mathbf{E}_1(t) = \mathbf{E}_1(\omega)e^{-i\omega t}$ (same for η_1 , $\mathbf{J}_1$, and $\mathbf{H}_1$). We will use the fact that the continuity equation now tells us that $\eta_1 = -1/(i\omega e)\nabla \cdot \mathbf{J}_1$. The second-order equation can be solved with the ansatz $\mathbf{J}_2(t) = \mathbf{J}_2(2\omega)e^{-i2\omega t}$,

$$\mathbf{J}_2(2\omega)(-2i\omega + \gamma) = -\frac{e}{i\omega m^*}(\nabla \cdot \mathbf{J}_1)\mathbf{E}_1 + \frac{1}{en_0}[(\nabla \cdot \mathbf{J}_1)\mathbf{J}_1 + (\mathbf{J}_1 \cdot \nabla)\mathbf{J}_1] - \frac{e\mu_0}{m^*}\mathbf{J}_1 \times \mathbf{H}_1. \quad (2.70)$$

Using Eq. 2.67, $\mathbf{H}_1 = 1/(i\omega\mu_0)\nabla \times \mathbf{E}_1$, and the vector identity $\mathbf{E}_1 \times (\nabla \times \mathbf{E}_1) = 1/2\nabla(\mathbf{E}_1 \cdot \mathbf{E}_1) - (\mathbf{E}_1 \cdot \nabla)\mathbf{E}_1$ we find after some algebra that,

$$\begin{aligned} \mathbf{J}_2(2\omega) = \frac{\varepsilon_0 e \omega_p^2}{i\omega(-2i\omega + \gamma)(-i\omega + \gamma)} & \left[(\nabla \cdot \mathbf{E}_1)\mathbf{E}_1 \left[\frac{2i\omega - \gamma}{-i\omega + \gamma} \right] \right. \\ & \left. + (\mathbf{E}_1 \cdot \nabla)\mathbf{E}_1 \left[\frac{\gamma}{-i\omega + \gamma} \right] - \frac{1}{2}\nabla(\mathbf{E}_1 \cdot \mathbf{E}_1) \right]. \quad (2.71) \end{aligned}$$

This equation is the first-order approximation of the steady-state, second-harmonic generation in a free-electron metal. It is commonly stated that centrosymmetric crystals

cannot generate even-order harmonics, but this is true only in the dipole approximation. Eq. 2.71 shows the three primary bulk contributions to second harmonic generation in metals (as confirmed in Ref. [16]). In the brackets (RHS), the first term represents the interaction of the charge density field with the electric field. The second term is the advection term from earlier. The third term represents the ponderomotive force. None of these terms exists within the dipole approximation as they depend on the field distribution. Each term in the brackets of Eq. 2.71 can be decomposed into electric quadrupole and magnetic dipole terms [135].

2.2.8 Numerical solution of the hydrodynamic plasma model

In this section, we derive the update equations for the hydrodynamic plasma model given by Eq. 2.62 and supplemented by either Gauss's Law (Eq. 2.1) or the continuity equation (Eq. 2.63). The current density formulation is used here rather than the polarization density as in the implementations for the previous linear models to avoid the addition of unnecessary time derivatives. Our general electric field update Eq. 2.13 shows that the current density is at half timesteps as with the magnetic field. The current density vector components are collocated in space with the electric field components. Using the finite difference approximation on the time derivative in Eq. 2.62 gives,

$$\begin{aligned} \frac{\mathbf{J}^{n+1/2} - \mathbf{J}^{n-1/2}}{\Delta t} + \gamma \mathbf{J}^n - \frac{\mathbf{J}^n}{en^n} (\nabla \cdot \mathbf{J}^n) - \frac{1}{en^n} (\mathbf{J}^n \cdot \nabla) \mathbf{J}^n - \frac{\mathbf{J}^n}{e(n^n)^2} (\mathbf{J}^n \cdot \nabla) n^n \\ = \frac{n^n e^2}{m^*} \left(\mathbf{E}^n - \frac{\mu_0}{n^n e} \mathbf{J}^n \times \mathbf{H}^n \right) + \frac{en v_F^2}{3} \left(\frac{n^n}{n_0} \right)^{2/3} \nabla n^n, \quad (2.72) \end{aligned}$$

where all terms are centred at timestep n . Here, we require current density terms centred at timestep n . This is typically calculated by taking the first-order approximation, $\mathbf{J}^n \approx 1/2(\mathbf{J}^{n+1/2} + \mathbf{J}^{n-1/2})$; however, this is not possible here due to the nonlinear terms. The simplest approach is to take the zeroth order approximation $\mathbf{J}^n \approx \mathbf{J}^{n-1/2}$. Although simple, this breaks the second-order accuracy of the FDTD method; however, this may not be a significant issue if using small step sizes. Another approach is to take a double finite difference step so that all terms are centred at $n - 1/2$,

$$\left. \frac{\partial \mathbf{J}}{\partial t} \right|_{n-1/2} \approx \frac{\mathbf{J}^{n+1/2} - \mathbf{J}^{n-3/2}}{2\Delta t}, \quad (2.73)$$

which can be explicitly updated.

That said, the nonlinear plasmonics work presented in this thesis is restricted to second-order nonlinearities for which a better scheme can be utilized based on Eqs. 2.67 and 2.68. All fields are split into linear and second-order (nonlinear) components, including the electric, magnetic, and polarization fields. Maxwell's equations and the associated models become,

$$\varepsilon_0 \nabla \cdot \mathbf{E}_1 + \nabla \cdot \mathbf{P}_1 = -e\eta_1 \quad (2.74)$$

$$\nabla \cdot \mathbf{H}_1 = 0 \quad (2.75)$$

$$\nabla \times \mathbf{E}_1 = -\mu_0 \frac{\partial \mathbf{H}_1}{\partial t} \quad (2.76)$$

$$\nabla \times \mathbf{H}_1 = \mathbf{J}_1 + \varepsilon_0 \frac{\partial \mathbf{E}_1}{\partial t} + \frac{\partial \mathbf{P}_1}{\partial t} \quad (2.77)$$

$$\frac{\partial \mathbf{J}_1}{\partial t} + \gamma \mathbf{J}_1 = \varepsilon_0 \omega_p^2 \mathbf{E}_1, \quad (2.78)$$

and

$$\varepsilon_0 \nabla \cdot \mathbf{E}_2 + \nabla \cdot \mathbf{P}_2 = 0 \quad (2.79)$$

$$\nabla \cdot \mathbf{H}_2 = 0 \quad (2.80)$$

$$\nabla \times \mathbf{E}_2 = -\mu_0 \frac{\partial \mathbf{H}_2}{\partial t} \quad (2.81)$$

$$\nabla \times \mathbf{H}_2 = \mathbf{J}_2 + \varepsilon_0 \frac{\partial \mathbf{E}_2}{\partial t} + \frac{\partial \mathbf{P}_2}{\partial t} \quad (2.82)$$

$$\frac{\partial \mathbf{J}_2}{\partial t} + \gamma \mathbf{J}_2 = \frac{1}{en_0} [(\nabla \cdot \mathbf{J}_1) \mathbf{J}_1 + (\mathbf{J}_1 \cdot \nabla) \mathbf{J}_1] + \varepsilon_0 \omega_p^2 \mathbf{E}_2 + \frac{\eta_1 e^2}{m^*} \mathbf{E}_1 - \frac{e\mu_0}{m^*} \mathbf{J}_1 \times \mathbf{H}_1, \quad (2.83)$$

where we ignore the effects of the electron pressure and the second-order charge perturbation (e.g. $\eta_2 = 0$). Although not explicitly displayed, $\mathbf{P}_1$ and $\mathbf{P}_2$ can be calculated using a Lorentz or critical points model. In Chapter 4, we use the two critical points model for gold.

Numerically we can solve all of these equations explicitly. The linear current density $\mathbf{J}_1^{n+1/2}$ is updated by numerically solving Eq. 2.78 using finite-differences and linear interpolation,

$$\frac{\mathbf{J}_1^{n+1/2} - \mathbf{J}_1^{n-1/2}}{\Delta t} + \gamma \frac{\mathbf{J}_1^{n+1/2} + \mathbf{J}_1^{n-1/2}}{2} = \varepsilon_0 \omega_p^2 \mathbf{E}_1^n, \quad (2.84)$$

which is rearranged into an update equation,

$$\mathbf{J}_1^{n+1/2} \left(\frac{1}{\Delta t} + \frac{\gamma}{2} \right) = \mathbf{J}_1^{n-1/2} \left(\frac{1}{\Delta t} - \frac{\gamma}{2} \right) + \varepsilon_0 \omega_p^2 \mathbf{E}_1^n. \quad (2.85)$$

This equation is clearly explicit as $\mathbf{E}_1^n$ and $\mathbf{J}_1^{n-1/2}$ will be known from the prior timestep.

The second-order nonlinear current density $\mathbf{J}_2^{n+1/2}$ is also updated by numerically solving Eq. 2.83 using finite-differences where,

$$\begin{aligned} \mathbf{J}_2^{n+1/2} \left(\frac{1}{\Delta t} + \frac{\gamma}{2} \right) &= \mathbf{J}_2^{n-1/2} \left(\frac{1}{\Delta t} - \frac{\gamma}{2} \right) + \varepsilon_0 \omega_p^2 \mathbf{E}_2^n \\ &+ \frac{1}{en_0} [(\nabla \cdot \mathbf{J}_1^n) \mathbf{J}_1^n + (\mathbf{J}_1^n \cdot \nabla) \mathbf{J}_1^n] + \frac{\eta_1 e^2}{m^*} \mathbf{E}_1^n - \frac{e\mu_0}{m^*} \mathbf{J}_1^n \times \mathbf{H}_1^n, \end{aligned} \quad (2.86)$$

where $\mathbf{J}_1^n$ and $\mathbf{H}_1^n$ are evaluated using linear interpolation. Some spatial derivatives require a double-step finite difference (as in Eq. 2.73) to maintain a centred difference scheme. Finally, η_1^n can be updated by numerical differentiation of Gauss's law Eq. 2.74, which is evaluated at the center of the Yee grid,

$$\begin{aligned} \eta_1^n(i, j, k) &= -\frac{\varepsilon_0}{e} \left(\frac{E_{x,1}^n(i+1/2, j, k) - E_{x,1}^n(i-1/2, j, k)}{\Delta x} \right. \\ &+ \frac{E_{y,1}^n(i, j+1/2, k) - E_{y,1}^n(i, j-1/2, k)}{\Delta y} + \frac{E_{z,1}^n(i, j, k+1/2) - E_{z,1}^n(i, j, k-1/2)}{\Delta z} \Big) \\ &\quad \left. - \frac{1}{e} \nabla \cdot \mathbf{P}_1^n, \end{aligned} \quad (2.87)$$

where the finite difference scheme is repeated for $\nabla \cdot \mathbf{P}_1^n$.

From here, an explicit update scheme is possible. First $\mathbf{H}_1^{n+1/2}$ and $\mathbf{H}_2^{n+1/2}$ are updated via Eq. 2.12. Next, $\mathbf{J}_1^{n+1/2}$ is updated via Eq. 2.85, followed by $\mathbf{J}_2^{n+1/2}$ with Eq. 2.86. Finally $\mathbf{E}_1^{n+1}$ and $\mathbf{E}_2^{n+1}$ can be evaluated via Eq. 2.13 with $\mathbf{P}_1^{n+1}$ and $\mathbf{P}_2^{n+1}$ being updated by their respective update equations.

2.3 Inverse Design

This section will introduce the methods used in this thesis for nanophotonic inverse design. In inverse design, we wish to minimize/maximize some objective function that is an arbitrary function of the electric and magnetic fields $F(\mathbf{E}, \mathbf{H})$ by tuning a set of parameters $\mathbf{p} = [p_1, \dots, p_i, \dots, p_N]$. For example, suppose we wish to maximize the transmittance through a metasurface at a specific frequency ω . We may define F as the energy flux through a transmission plane, $F = \int \text{Re}(1/2\mathbf{E}^*(\omega) \times \mathbf{H}(\omega)) \cdot \hat{\mathbf{n}} dA$, and $\mathbf{p}$ as a vector consisting of the positions and sizes of nanostructures on a metasurface.

In a gradient-based design approach, the optimization procedure is guided by the gradient $\nabla_{\mathbf{p}}F = [\partial F/\partial p_1, \dots, \partial F/\partial p_i, \dots, \partial F/\partial p_N]$, whose calculation becomes the primary objective. Once calculated, a gradient-guided optimizer such as gradient descent, ADAM, method of moving asymptotes, etc., can evolve $\mathbf{p}$ through many iterations. For example, using gradient descent, a new value of $\mathbf{p}$ can be evaluated via $\mathbf{p}^{new} = \mathbf{p}^{prev} - \gamma \nabla_{\mathbf{p}}F$, where γ is a scalar step size. The gradient-based optimization procedure for nanophotonics is generally simple. First, an initial condition is chosen to start the optimization procedure $\mathbf{p}^{init}$. Next a calculation/experiment is run to evaluate $F(\mathbf{E}, \mathbf{H}, \mathbf{p})$ and $\nabla_{\mathbf{p}}F$. Finally, the optimizer generates $\mathbf{p}^{new}$, and the procedure repeats until a stopping criterion is reached.

There are three primary methods for obtaining the electromagnetic gradients, 1.) finite difference approximation, 2.) automatic differentiation, and 3.) adjoint sensitivity analysis. All three methods were applied in the research period and will be reviewed in the following.

For problems where the number of tuneable parameters, N , is small, using a finite-difference approximation may be the most economical and simple solution for obtaining the gradient. The gradient is obtained by multiple evaluations of the objective function F for different perturbations of $\mathbf{p}$. A centred difference (see Eq. 2.9) would require $2N$ evaluations to obtain the full gradient, one for each $p_i + \delta_i$ and $p_i - \delta_i$ where δ_i is the finite-difference stepsize chosen for element i . However, in cases where an evaluation of F is computationally expensive (i.e., requiring full-wave simulations), one may wish to sacrifice accuracy for a smaller evaluation number as provided by a forward (or backward) difference approximation,

$$\frac{\partial f}{\partial p_i} \approx \frac{f(p_i + \delta_i) - f(p_i)}{\delta_i}, \quad (2.88)$$

which requires only $N + 1$ evaluations since $f(p_i)$ is only calculated once. For example, finding the ideal spacing and size of nanostructures on a periodic metasurface is well-suited for a finite-difference form of gradient-based optimization.

For problems with a medium or large N , but where evaluating F is relatively inexpensive (i.e. for semi-analytic problems such as Fourier-propagation or transfer-matrix methods), automatic differentiation may be considered. Automatic differentiation is a set of techniques that enable the evaluation of gradients of arbitrary programming functions. Although popularly applied to machine learning and deep learning for optimizing fitting parameters, the applicability of *autodiff* extends into any computational optimization problem. For example, in photonics, a common design problem is the optimization of thin film stacks, whose optical properties are evaluated using the transfer-matrix method. The input of a transfer-matrix method function is the data concerning each layer’s refractive index and thickness, and the output is the transmittance and reflectance. Suppose the program is written using an automatic differentiation framework (for example, the JAX python module [132]). In that case, the gradient of the transmittance with respect to the stack properties is provided to the user via a function (called *grad* in JAX) which evaluates the gradient in the backend. This dramatically simplifies design processes where compact models are available to describe the system. Even in more general design problems where full-wave simulations are mandatory, autodiff can be applied to subproblems, such as evaluating the gradient of filtering functions in topology optimization as discussed in Chapter 2.3.2 and used in Chapter 6 to include filter functions that control the optical pulse width and bandwidth in the pulse-shape inverse design workflow.

Finally, for generalized inverse design, we wish the optimizer to design the permittivity distribution $\varepsilon(\mathbf{r})$ to minimize or maximize F . In this case, $\mathbf{p} = \varepsilon(\mathbf{r})$ and N becomes quite large, and evaluating F is expensive as it requires full-wave calculations. These optimization problems are called *topology optimization* as the final device is a solid structure with holes etched in. Topology optimization has been applied to several nanophotonics problems; however, today, it is commonly applied to design integrated photonic circuit components, including mode or demultiplexing splitters and combiners, which demand unintuitive geometries for the correct interferometric effect. For these problems, we turn to adjoint sensitivity analysis.

2.3.1 Adjoint sensitivity analysis for topology optimization

In this section, we derive the *adjoint method* for inverse design in nanophotonics where our design variables $\mathbf{p}$ are the permittivity at discrete locations in space, $p_i = \varepsilon(\mathbf{r}_i)$. Although not used in any published works, a significant amount of time was spent implementing the adjoint method in the MPI and GPU - parallelized FDTD programs. A plan to extend this method to the inverse design of metalenses for focusing high-harmonic generated photons is discussed in Chapter 8.2. Furthermore, in Chapter 6, we use adjoint sensitivity analysis to optimize pulse shapes for interaction with time-varying nanophotonic devices. The content of this section is an excellent primer to understanding the underlying math and is the inspiration behind the work of Chapter 6.

To optimize the objective function F by tuning the permittivity distribution $\varepsilon(\mathbf{r})$, we must find how sensitive F is to $\varepsilon(\mathbf{r}_i)$ for all positions $\mathbf{r}_i$ in the design region. We can do this using calculus,

$$\frac{dF}{d\varepsilon(\mathbf{r}_i)} = \frac{\partial F}{\partial \mathbf{E}} \cdot \frac{\partial \mathbf{E}}{\partial \varepsilon(\mathbf{r}_i)} + \frac{\partial F}{\partial \mathbf{H}} \cdot \frac{\partial \mathbf{H}}{\partial \varepsilon(\mathbf{r}_i)}, \quad (2.89)$$

where it's assumed that F has no explicit dependence on $\varepsilon(\mathbf{r}_i)$. Since the functional form of F is known (defined for the desired application), the vectors $\partial F/\partial \mathbf{E}$ and $\partial F/\partial \mathbf{H}$ are trivially calculated, although, in some implementations, autodiff is used for better user experience [155]. Evaluation of the vectors $\partial \mathbf{E}/\partial \varepsilon(\mathbf{r}_i)$ and $\partial \mathbf{H}/\partial \varepsilon(\mathbf{r}_i)$ requires Maxwell's equations.

Maxwell's curl equations in the frequency domain for inhomogeneous, linear, non-magnetic, isotropic systems can be written in the matrix form,

$$\begin{bmatrix} i\omega\varepsilon_0\varepsilon(\mathbf{r},\omega)\mathbf{I} & \nabla \times \\ -\nabla \times & i\omega\mu_0\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{E} \\ \mathbf{H} \end{bmatrix} = \begin{bmatrix} \mathbf{J} \\ 0 \end{bmatrix}, \quad (2.90)$$

where $\mathbf{I}$ is the identity operator, and all other terms have been explained in previous sections. Here we define $\mathbf{J}$ as a source current unrelated to the material dispersion or loss.

Eq. 2.90 can also be thought of in a numerical sense, assuming a finite-difference approach. In this interpretation, we imagine our spatial grid discretized using the FDTD scheme where the simulation domain contains $N^3 = N_x \times N_y \times N_z$ cells, where N_α is

the number of Yee cells in the α -direction. To maintain the matrix formulation, the spatial grid is flattened so that all functions of space are flattened to a vector, $f(\mathbf{r}) \rightarrow [f(\mathbf{r}_0), f(\mathbf{r}_1), f(\mathbf{r}_2), \dots, f(\mathbf{r}_{N^3})]^\top$ where $\mathbf{r}_\beta$ is the β^{th} grid point in the 3D grid. Thus, $[\mathbf{E}, \mathbf{H}]^\top$ is a vector of length $6N^3$, $i\omega\varepsilon_0\mathbf{I}$ and $i\omega\mu_0\mathbf{I}$ are constant diagonal matrices of size $3N^3 \times 3N^3$, $\nabla \times$ is a 3×3 block matrix, consisting of $N^3 \times N^3$ finite difference matrices, and $\varepsilon(\mathbf{r}, \omega)$ is a $3N^3 \times 3N^3$ matrix with the permittivity values flattened (as before) along the diagonal three times (for each $\mathbf{E}$ field component).

To simplify the notation, we will use the more compact form, $A\mathbf{x} = \mathbf{b}$ where A is the matrix in Eq. 2.90, $\mathbf{x}$ is the electromagnetic field vector and $\mathbf{b}$ is the source current vector.

Continuing with the adjoint formulation, we differentiate Eq. 2.90 with respect to $\varepsilon(\mathbf{r}_i)$,

$$\frac{\partial A}{\partial \varepsilon(\mathbf{r}_i)}\mathbf{x} + A\frac{\partial \mathbf{x}}{\partial \varepsilon(\mathbf{r}_i)} = \frac{\partial \mathbf{b}}{\partial \varepsilon(\mathbf{r}_i)}, \quad (2.91)$$

where typically the current source is external to the device, and thus,

$$A\frac{\partial \mathbf{x}}{\partial \varepsilon(\mathbf{r}_i)} = -\frac{\partial A}{\partial \varepsilon(\mathbf{r}_i)}\mathbf{x}. \quad (2.92)$$

Although $\mathbf{b}$ is independent of ε , in Chapter 6, we deal with the case where the source $\mathbf{b}$ depends on the tuneable parameters $\mathbf{p}$.

Next, we left - multiply both sides by an adjoint vector $\mathbf{y}$,

$$\mathbf{y}^\top A\frac{\partial \mathbf{x}}{\partial \varepsilon(\mathbf{r}_i)} = -\mathbf{y}^\top \frac{\partial A}{\partial \varepsilon(\mathbf{r}_i)}\mathbf{x}, \quad (2.93)$$

and we define a new vector equation that at first seems arbitrary,

$$\mathbf{y}^\top A = \left[\frac{\partial F}{\partial \mathbf{E}}, \frac{\partial F}{\partial \mathbf{H}} \right]. \quad (2.94)$$

When the RHS of Eq. 2.94 is plugged into the LHS of Eq. 2.93 we get

$$\mathbf{y}^\top A\frac{\partial \mathbf{x}}{\partial \varepsilon(\mathbf{r}_i)} = \left[\frac{\partial F}{\partial \mathbf{E}}, \frac{\partial F}{\partial \mathbf{H}} \right] \cdot \frac{\partial \mathbf{x}}{\partial \varepsilon(\mathbf{r}_i)} = \frac{\partial F}{\partial \mathbf{E}} \cdot \frac{\partial \mathbf{E}}{\partial \varepsilon(\mathbf{r}_i)} + \frac{\partial F}{\partial \mathbf{H}} \cdot \frac{\partial \mathbf{H}}{\partial \varepsilon(\mathbf{r}_i)} = \frac{dF}{d\varepsilon(\mathbf{r}_i)}, \quad (2.95)$$

where we used $\mathbf{x} = [\mathbf{E}, \mathbf{H}]^T$. Finally, we see that the RHS of Eq. 2.93 becomes the gradient,

$$\frac{dF}{d\varepsilon(\mathbf{r}_i)} = -\mathbf{y}^\top \frac{\partial A}{\partial \varepsilon(\mathbf{r}_i)} \mathbf{x}. \quad (2.96)$$

The *arbitrary* Eq. 2.94 can be transposed to yield,

$$A\mathbf{y} = \left[\frac{\partial F}{\partial \mathbf{E}}, \frac{\partial F}{\partial \mathbf{H}} \right]^\top, \quad (2.97)$$

where we used the fact that $A^\top = A$ (which can be shown by analysis of the curl operators), and therefore, $\mathbf{y} = [\mathbf{E}^{adj}, \mathbf{H}^{adj}]^\top$ is another solution to Maxwell's equation where the source is given by the derivative of the objective function $[\partial F/\partial \mathbf{E}, \partial F/\partial \mathbf{H}]^\top$,

$$\begin{bmatrix} i\omega\varepsilon_0\varepsilon(\mathbf{r},\omega)\mathbf{I} & \nabla \times \\ -\nabla \times & i\omega\mu_0\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{E}^{adj} \\ \mathbf{H}^{adj} \end{bmatrix} = \begin{bmatrix} \partial F/\partial \mathbf{E} \\ \partial F/\partial \mathbf{H} \end{bmatrix}. \quad (2.98)$$

The derivatives $[\partial F/\partial \mathbf{E}, \partial F/\partial \mathbf{H}]^\top$ are evaluated using the solution of $A\mathbf{x} = \mathbf{b}$. Therefore, the location where the objective function is evaluated becomes the location of the source in the adjoint calculation. For this reason, it is sometimes called the *backwards* simulation.

From Eq. 2.96, the gradient finally becomes,

$$\frac{dF}{d\varepsilon(\mathbf{r}_i)} = -i\omega\varepsilon_0 \mathbf{E}^{adj} \cdot \mathbf{E}^{for} \delta(\mathbf{r} - \mathbf{r}_i), \quad (2.99)$$

where we used the fact $\partial A/\partial \varepsilon(\mathbf{r}_i)$ is only non-zero at the point $\mathbf{r}_i$ in the upper-left block of the matrix. We also added a superscript to $\mathbf{E}$ to clarify that it is the electric field from the *forward* calculation.

By solving Maxwell's equations twice for the same system (with different sources), we can find the gradient of our objective function with respect to the permittivity everywhere. This seemingly simple trick of algebra and calculus has become a powerful tool in nanophotonics research and development.

2.3.2 Adjoint topology optimization implementation

Before optimizing, one needs to set up the simulation's layout and details, including where the sources, design region and objective function are located. We also need to choose the permittivities of the final design. Typically, these designs are binary, where we want our final permittivity distribution to be $\varepsilon(\mathbf{r}) = \varepsilon_1$ or ε_2 . However, to get to this point, we will allow the permittivity distribution to be continuously deformed during the optimization procedure, $\varepsilon(\mathbf{r}) \in [\varepsilon_1, \varepsilon_2]$.

To do this, we parameterize the permittivity distribution using a density function $\rho(\mathbf{r}) \in [0, 1]$ where,

$$\varepsilon(\mathbf{r}) = \varepsilon_1(1 - \rho(\mathbf{r})) + \varepsilon_2\rho(\mathbf{r}). \quad (2.100)$$

Since we are now optimizing for $\rho(\mathbf{r})$, we can use the chain rule to derive our new gradient from Eqs. 2.99 and 2.100,

$$\frac{dF}{d\rho(\mathbf{r})} = \frac{dF}{d\varepsilon(\mathbf{r})} \frac{d\varepsilon(\mathbf{r})}{d\rho(\mathbf{r})} = -i\omega\varepsilon_0(\varepsilon_2 - \varepsilon_1)\mathbf{E}^{adj}(\mathbf{r}) \cdot \mathbf{E}^{for}(\mathbf{r}). \quad (2.101)$$

Filters are typically used to decrease the likelihood of feature sizes that are too small in order to ensure that devices are fabricable. One simple filter is the Gaussian convolution filter, where,

$$\rho_{filtered}(\mathbf{r}) = \frac{1}{C} \int_{|\mathbf{r}-\mathbf{r}'| < 2a} \rho(\mathbf{r}') \exp\left(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{a^2}\right) d\mathbf{r}', \quad (2.102)$$

where a is the desired minimum feature size and $C = \int_{|\mathbf{r}-\mathbf{r}'| < 2a} \exp\left(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{a^2}\right) d\mathbf{r}'$ is a normalization factor. Note that this is generally written as a volume integral; however, in most applications of nanophotonic inverse design, design is done in 2D, with the third dimension having no variation within the devices. For example, a metalens who focuses light along the z -direction only needs to be designed in the $x - y$ plane so that $\rho(x, y, z) = \rho(x, y)$, $\forall z \in [z_{min}, z_{max}]$ where z_{min} is the bottom of the metalens design region (the substrate), and z_{max} is the top of the design region (the lens-air interface).

Other feature size filter examples include the Helmholtz-type partial differential equation filter [156] where $\rho_{filtered}$ is the solution of a PDE with the unfiltered $\rho(\mathbf{r})$ being the

PDE source, or the FFT filter where $\rho(\mathbf{r})$ has its high spatial frequencies removed [157]. All these filters are smooth (theoretically) and thus differentiable in their domain. The derivative $\partial\rho_{filtered}(\mathbf{r})/\partial\rho(\mathbf{r}')$ (for two different positions) is calculable, which is necessary for calculating the gradient,

$$\frac{dF}{d\rho(\mathbf{r})} = \int_{\text{design space}} \frac{dF}{d\varepsilon(\mathbf{r}')} \frac{d\varepsilon(\mathbf{r}')}{d\rho_{filtered}(\mathbf{r}')} \frac{\partial\rho_{filtered}(\mathbf{r}')}{\partial\rho(\mathbf{r})} d\mathbf{r}'. \quad (2.103)$$

These filters act as smoothing filters which push $\rho(\mathbf{r})$ away from a binary structure. Thus, a thresholding function is usually applied to force $\rho(\mathbf{r})$ towards a binary distribution. A Heaviside function is a naïve choice as it would push intermediate values of $\rho(\mathbf{r})$ to 0 or 1, but its trivial derivative inhibits gradient-based optimization. A commonly-used smooth Heaviside function is [155, 158]

$$\rho_{threshold}(\mathbf{r}) = \frac{\tanh(\beta\eta) + \tanh(\beta(\rho_{filtered} - \eta))}{\tanh(\beta\eta) + \tanh(\beta(1 - \eta))} \quad (2.104)$$

where $\beta \in [1, \infty)$ controls the steepness of the step, and $\eta \in [0, 1]$ controls the threshold value. In the limit $\beta \rightarrow \infty$, the thresholded field becomes binary, $\rho_{threshold}(\mathbf{r}) = 0$ or 1. The common strategy is to set $\eta \sim 0.5$ as constant and slowly increase β throughout the optimization procedure (say every ~ 20 iterations) from a small scalar (say $\beta_{init} \sim 2$) to a large scalar (say $\beta_{stop} \sim 10^5$) using a logarithmic scale. After including thresholding, the gradient becomes

$$\frac{dF}{d\rho(\mathbf{r})} = \int_{\text{design space}} \frac{dF}{d\varepsilon(\mathbf{r}')} \frac{d\varepsilon(\mathbf{r}')}{d\rho_{threshold}(\mathbf{r}')} \frac{\partial\rho_{threshold}(\mathbf{r}')}{\partial\rho_{filtered}(\mathbf{r}')} \frac{\partial\rho_{filtered}(\mathbf{r}')}{\partial\rho(\mathbf{r})} d\mathbf{r}'. \quad (2.105)$$

Additional filters can be used for various reasons as long as their gradients are accounted for via the chain rule.

Note that the first two derivatives $\frac{\partial F}{\partial\varepsilon(\mathbf{r})} \frac{\partial\varepsilon(\mathbf{r})}{\partial\rho_{threshold}(\mathbf{r})}$ are calculated using Eq. 2.101. The filter derivatives can be tricky, and it is now common practice to use automatic differentiation [155]. In this case, a global filter function can be programmed using an automatic differentiation programming library, and the full filter gradient is calculated in the back-end, enabling generalized filtering schemes without requiring tedious math, assuming all elements are differentiable.

Figure 2.7 summarizes the adjoint topology inverse design method in 11 steps. The main optimization loop is highlighted in green and the pre- and post-processing in blue.

Note that the gradient descent algorithm is simple to implement, especially when testing and debugging the algorithm. However, there are minimization algorithms that are much better suited to photonic topology optimization, including the method of moving asymptotes (MMA) [158] and the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS) [159] which are available in open-source libraries such as NLOpt [160] and Scipy [142].

The source of the adjoint simulation is given by the derivative of the objective function with respect to the electromagnetic field of the forward simulation, $[\partial F/\partial \mathbf{E}, \partial F/\partial \mathbf{H}]$. In the frequency domain, these are complex numbers. For example, a simple objective function may be the z -component of the electric field amplitude at some position $\mathbf{r}_F$. We want to maximize

$$F = |E_z(\mathbf{r}_F, \omega)|^2. \quad (2.106)$$

Since $E_z(\mathbf{r}_F, \omega)$ is a complex number, the derivative is also complex $\partial F/\partial E_z = E_z^*(\mathbf{r}_F, \omega)$. In the frequency domain, this is not a problem. However, care must be taken in time-domain simulations (where the fields are Fourier transformed), where it is best to represent the complex number as,

$$\frac{\partial F}{\partial \mathbf{E}} = E_z^*(\mathbf{r}_F, \omega) = |E_z^*(\mathbf{r}_F, \omega)| \exp(i\phi), \quad (2.107)$$

where ϕ is the phase of $E_z^*(\mathbf{r}_F, \omega)$. Thus the time-domain source can be any function with amplitude $|E_z^*(\mathbf{r}_F, \omega)|$ and phase-delay ϕ . For example,

$$J^{adj}(t) = |E_z^*(\mathbf{r}_F, \omega)| \cos(\omega t - \phi), \quad (2.108)$$

is a naïve choice. As a better choice, one may find an envelope function $A(t)$, such that,

$$J^{adj}(t) = |E_z^*(\mathbf{r}_F, \omega)| A(t) \cos(\omega t - \phi), \quad (2.109)$$

Fourier transforms to Eq. 2.107.

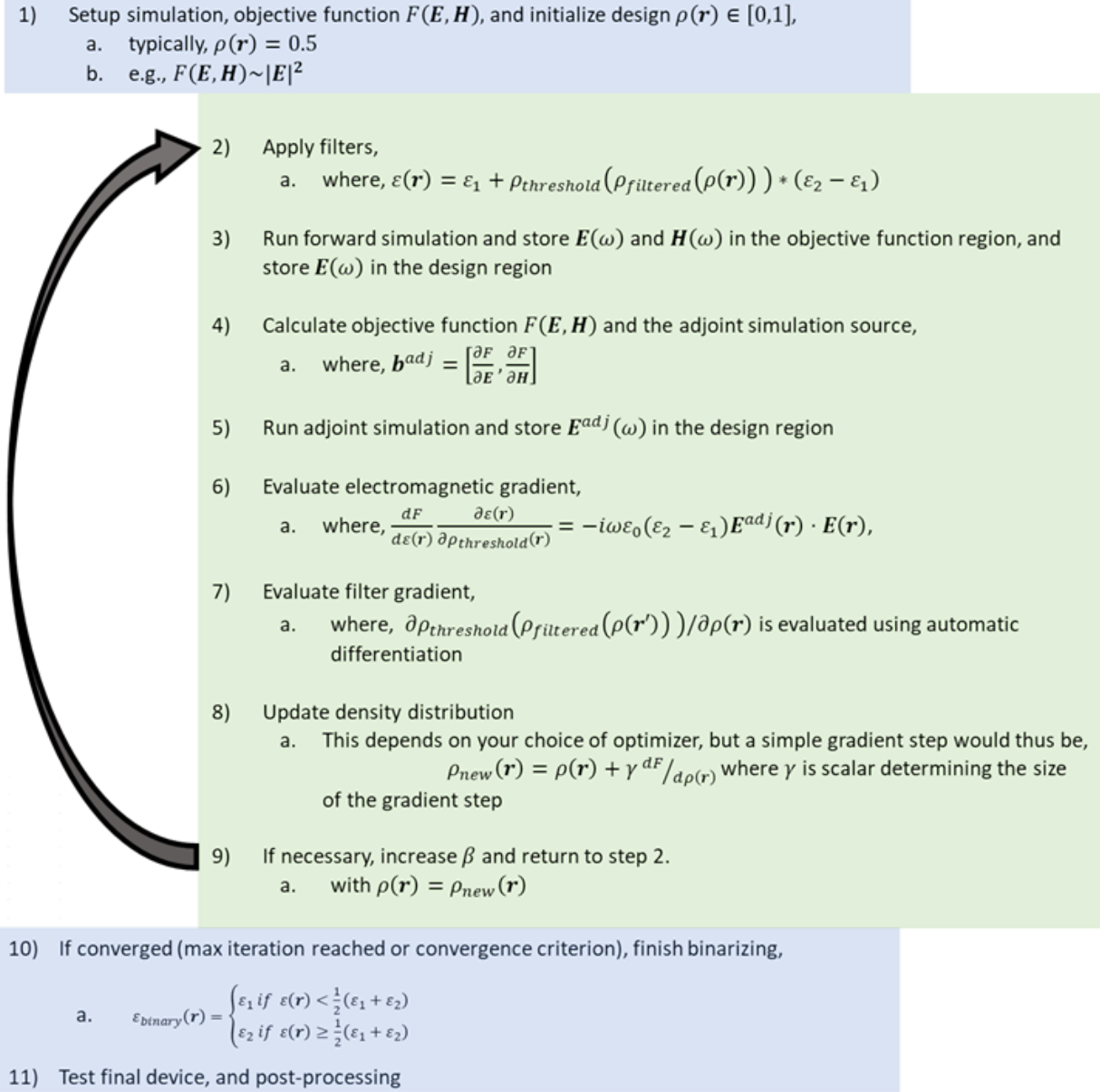


Figure 2.7: Algorithm for nanophotonic inverse design. The main optimization loop is highlighted in green, and pre and post-processing are highlighted in blue.

For broadband objective functions, $F(\mathbf{E}(\omega_1), \dots, \mathbf{E}(\omega_M), \mathbf{H}(\omega_1), \dots, \mathbf{H}(\omega_M))$, more care must be taken. Although using a sum of Gaussians (each centred at the discrete frequencies, ω_i) is tempting, each Gaussian must be narrow-band to avoid spectral overlap, forcing longer simulation times. In Ref. [155], a novel method is introduced for calculating a short pulse for $J^{adj}(t)$ for cases where the objective function contains multiple, discrete frequencies.

2.4 Deep Learning

Deep learning is a subfield of machine learning where deep (multi-layered) artificial neural networks are used to make predictions based on data. This section will summarize the mathematics underlying deep learning and deep neural networks (DNNs) and briefly explain how they are trained.

Figure 2.8 illustrates an artificial neural network consisting of a sequence of algebraic manipulations summarized here before using math. First, the input numerical data is fed into the input layer (green dots) – in this case, we expect only two input values. Next, the numbers are sent to the first (in this case, only) *hidden layer* (blue dots), where they are linearly combined, and the *activation function* is applied. The weights for the linear combination are different for each hidden layer node. The nonlinear activation function imitates the action potential in real neurons; they only transfer an electronic signal if the given stimulus is above a certain threshold, thus explaining why the hidden layer nodes are commonly referred to as artificial neurons. The resulting number from each node of the first hidden layer is sent next to each node of the second hidden layer, where they are again linearly combined and acted on by the activation function. Finally, numbers from the last hidden layer are linearly combined and sent to the output layer (purple dot), the final result. Note that a neural network is considered *deep* when it has more than one hidden layer.

In the following, we will discuss the mathematics behind DNNs, starting from the data input from which we wish to make a prediction. The input layer of the network receives a vector of numerical data $\mathbf{x}^{(0)} = [x_0^{(0)}, x_1^{(0)}, \dots, x_m^{(0)}]$, where the superscript denotes the layer (the input is the 0th layer). Each value $x_i^{(0)}$ is considered a *node* of the input layer.

The vector $\mathbf{x}^{(0)}$ is sent to each node in the first hidden layer. The i^{th} node of the first hidden layer will perform a linear combination of the values of $\mathbf{x}^{(0)}$,

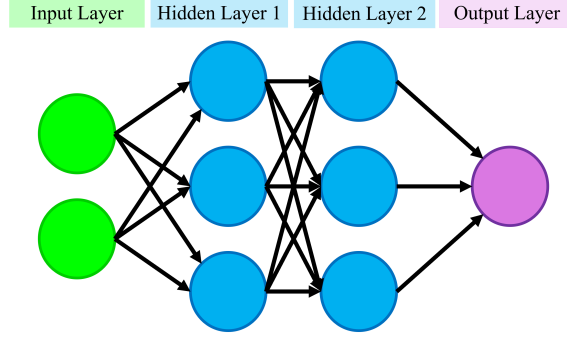


Figure 2.8: A simple neural network illustrating the three primary layers; the input layer receives numerical data, the hidden layers operate on the data, and the output layer combines the data from the last hidden layer and returns the value(s) to the user.

$$z_i^{(1)} = b_i^{(1)} + \sum_j w_{i,j}^{(1)} x_j^{(0)}, \quad (2.110)$$

where the parameters $w_{i,j}^{(1)}$ are the weights of the first hidden layer, and $b_i^{(1)}$ are the biases. This equation can be written as a matrix-algebra problem,

$$\mathbf{z}^{(1)} = \mathbf{b}^{(1)} + \hat{W}^{(1)} \mathbf{x}^{(0)}, \quad (2.111)$$

where the bolded terms are vectors and $\hat{W}^{(1)}$ is the matrix containing the values $w_{i,j}^{(1)}$. At the output of the hidden layer, the activation function is applied

$$\mathbf{x}^{(1)} = f_{act}(\mathbf{z}^{(1)}), \quad (2.112)$$

and $\mathbf{x}^{(1)}$ becomes the input for the next layer. The math behind an individual neuron in a hidden layer is summarized in Figure 2.9. This sequence of manipulations continues until the output layer,

$$\mathbf{y}^{(out)} = f_{act}(\mathbf{b}^{(out)} + \hat{W}^{(out)} \mathbf{x}^{(out-1)}), \quad (2.113)$$

which is generally a vector but, in simple cases, may be a scalar. For example, vector outputs are commonly used in classification problems where the length of $\mathbf{y}^{(out)}$ is the number

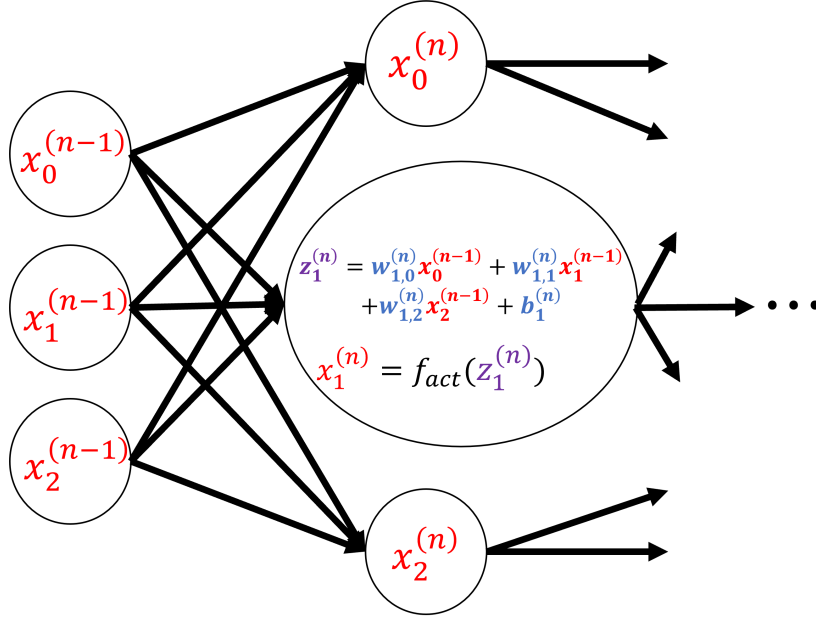


Figure 2.9: Artificial neurons in a hidden layer and the math it performs.

of classes, and the value of $y_i^{(out)}$ represents the probability of that $\mathbf{x}^{(0)}$ is the i^{th} class. For example, $\mathbf{x}^{(0)}$ may be a picture of an animal, and $y_{i=3}^{(out)}$ is the probability that $\mathbf{x}^{(0)}$ is a picture of a dog. In photonics applications, vector outputs may represent the spectral response of a device characterized by the values in $\mathbf{x}^{(0)}$.

It is the activation function that provides nonlinearity to the neural network. Indeed, without the activation function (or if $f_{act}(x) = x$), a neural network becomes a linear regression algorithm, regardless of the number of hidden layers. There are many candidate activation functions to choose from, and common ones include the sigmoid and tanh functions which behave like continuous, differentiable step functions, providing an accurate model of the activation potential. However, for regression problems where we expect output values to be continuous, the rectified linear unit (ReLU) function, a thresholded linear function, is commonly applied (or one of its many variations).

2.4.1 How does the machine learn

The previous section overviewed how a DNN works in the forward direction. However, the true underlying complexity involves how DNNs learn. There are many strategies for training networks, each based on collecting large amounts of data to feed through the network. Here will only discuss supervised learning which assumes the data is separated into inputs and outputs. In the image classification example, the inputs may be pictures of animals, and the outputs are labels of the animal type. Alternatively, in the spectral example, the input is the device parameters, and the output is the spectrum.

Training a DNN involves tuning the weights $\hat{W}^{(n)}$ and biases $\mathbf{b}^{(n)}$ to minimize the absolute difference between the DNN output and the true output from the dataset. The difference is given by an error function $E(\mathbf{y}^{(out)})$. As with activation functions, there are many error functions to choose from, but a common example is a mean-squared error,

$$E(\mathbf{y}^{(out)}) = \frac{1}{D} \sum_{d=1}^D \left| \mathbf{y}_d^{(out)} - \mathbf{y}_d^{(dataset)} \right|^2, \quad (2.114)$$

where $\mathbf{y}_d^{(out)}$ is the output vector of the DNN when given the d^{th} input data point $\mathbf{x}_d^{(0)}$, and $\mathbf{y}_d^{(dataset)}$ is the expected output from the dataset.

Thus training a DNN becomes an optimization problem; minimize $E(\mathbf{y}^{(out)})$ using the weights $\hat{W}^{(n)}$ and biases $\mathbf{b}^{(n)}$. Now it is worth pointing out that a DNN is completely differentiable with respect to the weights and biases. In other words,

$$\frac{\partial E}{\partial w_{i,j}^{(n)}} = \frac{dE}{d\mathbf{y}^{(out)}} \cdot \frac{\partial \mathbf{y}^{(out)}}{\partial w_{i,j}^{(n)}}, \quad (2.115)$$

is calculable through repeated use of the chain rule. First $\partial \mathbf{y}^{(out)} / \partial w_{i,j}^{(out)}$ can be easily computed via direct differentiation of Eq. 2.113. Next, the Jacobian $\partial \mathbf{y}^{(out)} / \partial \mathbf{x}^{(out-1)}$ can also be evaluated by differentiation of Eq. 2.113, allowing for the calculation of

$$\frac{\partial E}{\partial w_{i,j}^{(out-1)}} = \frac{\partial \mathbf{y}^{(out)}}{\partial \mathbf{x}^{(out-1)}} \cdot \frac{\partial \mathbf{x}^{(out-1)}}{\partial w_{i,j}^{(out-1)}}, \quad (2.116)$$

where $\partial \mathbf{x}^{(out-1)} / \partial w_{i,j}^{(out-1)}$ can also be computed in the same way as $\partial \mathbf{y}^{(out)} / \partial w_{i,j}^{(out)}$.

This process of compounding chain rules is called *backpropagation*, which can be applied to find the sensitivity of $E(\mathbf{y}^{(out)})$ with respect to all weights $\hat{W}^{(n)}$ and biases $b^{(n)}$ for all hidden layers n .

This requirement for backpropagation inspired the development of auto-differentiation programming libraries. Using autodiff, users can calculate $E(\mathbf{y}^{(out)})$ using a forward-pass of the network, and the gradients, $\partial E / \partial w_{i,j}^{(n)}$ and $\partial E / \partial b_i^{(n)}$ are evaluated in the backend and are fed to a gradient-based optimization algorithm to optimally tune the weights and biases to best model the dataset. Automatic differentiation is usually accomplished by storing the computation graph of the forward-pass calculation and then applying the chain rule in a backward-pass of the same graph. Thus, in the view of automatic differentiation, optimizing parameterized problems can be treated as training a neural network where the tuneable parameters become the fitting parameters of the network.

Chapter 3

Parallel FDTD Modeling of Nonlocality in Plasmonics

In this chapter, we apply the content of Chapters 1.1, 2.1, and 2.2 to the problem of computational plasmonics at the extreme nanoscale, with features of sizes ≤ 10 nm. I am the lead author of the work presented in this chapter, published in the journal *IEEE Transactions on Antennas and Propagation* in July 2021.

My first project as a graduate student, the initial goal was to implement the hydrodynamic plasma model in our in-house FDTD software, built with plasmonic simulations in mind [77]. In the meantime, we found that a proper treatment of *linear*, nonlocal plasmonics was lacking in the literature, and the project was steered in that direction. This lack of implementation is clearly explained in Ref. [76]; the memory and computational requirements are unwieldy because the resolution must be fine enough to resolve the electron perturbations near the structure surface. Our 3D-FDTD software was designed for distributed computing systems, and its computation time scales linearly with resources [75, 77], providing the ideal tool for this investigation.

In this chapter, we derive the nonlocal Drude model and the generalized nonlocal optical response (GNOR) model in the time domain. We then derive auxiliary differential equations for the free-electron polarization fields and update equations for the FDTD solver. We dedicate a section to how the FDTD program is parallelized using the Message Passing Interface (MPI) and how this is adapted for nonlocal modelling. With the essential ingredients in place, the FDTD implementations of the nonlocal Drude and GNOR models

are tested against nonlocal Mie theory (including longitudinal modes) for light scattering from nanometer-scale particles with exquisite accuracy and experimental results from the literature.

As an unforeseen bonus, the model (necessarily) reduces staircasing artifacts at the metal-dielectric interface that plagues FDTD simulations of plasmonic particles (even at high resolutions). Without the electronic pressure term in the Drude model, the charge can build up infinitely in stair-casing regions with sharp corners, resulting in artificial hotspots that distort the near fields. The electronic pressure ensures such charge build-up is reduced (in agreement with the Pauli exclusion principle), and the fields reflect this charge-smoothing effect.

Although remaining unexplored, this model was believed to explain the finding of blue-coloured structured gold metasurfaces; however, exploration of this was limited, and the project fell through. Studying the effects of nonlocality on plasmonic colouring remains a potential future application of this work. Similarly, we find that in FDTD, convergence issues arise if two metallic structures are placed too close to each other. Nonlocality may present the solution to this problem via the near-field smoothing effect; however, this also needs to be explored.

Supporting media in the form of time-domain movies is provided with this article and can be found in a zip folder at ieeexplore.ieee.org/document/9301363/media.

Parallel FDTD Modeling of Nonlocality in Plasmonics

Joshua Baxter[✉], Antonio Calà Lesina[✉], and Lora Ramunno

Abstract—As nanofabrication techniques become more precise, with ever smaller feature sizes, the ability to model nonlocal effects in plasmonics becomes increasingly important. Although nonlocal models based on hydrodynamics have been implemented using various computational electromagnetics techniques, the finite-difference time-domain (FDTD) version has remained elusive. Here, we present a comprehensive FDTD implementation of nonlocal hydrodynamics, including parallel computing. As a subnanometer step size is required to resolve nonlocal effects, a parallel implementation makes the computational cost of nonlocal FDTD more affordable. We first validate our algorithms for small spherical metallic particles, and find that nonlocality smears out staircasing artifacts at metal surfaces, increasing the accuracy over local models. We find this also for a larger nanostructure with sharp extrusions. The large size of this simulation, where nonlocal effects are clearly present, highlights the importance and impact of a parallel implementation in FDTD.

Index Terms—Finite-difference time-domain (FDTD), generalized nonlocal optical response (GNOR), hydrodynamic plasma model, nonlocality, parallel computing, plasmonics.

I. INTRODUCTION

FABRICATING objects with nanoscale precision is possible due to significant progress in nanofabrication techniques over the last few decades [1]. As a result, plasmonic nanostructures and metamaterials [2] are having tremendous impact on many fields including biosensing [3], quantum cryptography [4], nonlinear optics [5], photovoltaics [6], light emitting devices [7], and precision medicine [8].

Numerical modeling in plasmonics is typically based on optical models for bulk permittivity. These models—such as the Drude model for free-electron response, and the Lorentz and critical points models for bound electron response [9], [10]—are based on the local response approximation (LRA), which assumes that the induced polarization or current at a given location depends only on the electromagnetic field at

that same location. While appropriate for many applications, the Drude model is insufficient for modeling plasmonic structures smaller than 10 nm [11]–[14], as well as those containing sharp features or nanoscale gaps [15], [16]. The charge density near the surface of plasmonic objects is spread over a finite thickness on the order of a few angstroms [16], [17], and thus for small features, cannot be treated as localized at the surface, as is implicit in LRA models.

The hydrodynamic plasma model treats the conduction band electrons as a free electron gas [5], [12], [16]–[18], accounting for free-electron fluctuations via a pressure term. Unlike the Drude model, it does not make an LRA and thus nonlocality is incorporated. It has been found to correctly predict the expected spread of charge density near plasmonic surfaces as well as predict the blueshift in plasmon resonance with decreasing particle size [13], [17], [19]. More recently, the effects of electron diffusion were incorporated within a generalized nonlocal optical response (GNOR) model, correctly predicting a broadened line shape that was observed experimentally [20].

Nonlocal hydrodynamic models have been implemented using several computational electromagnetic methods including the finite element method [21], the discontinuous Galerkin time-domain (DGTD) method [22], and the boundary element method [23]. Due to their inhomogeneous mesh, finite element methods, including DGTD, offer computationally efficient calculations for systems containing sharp features, or complex geometry, especially when compared to finite difference methods. This has been demonstrated for dielectric resonator simulations [24]. However, the complexity of implementation of such methods, and a dearth of reliable open source code, may present a hurdle for researchers in the field.

A finite-difference time-domain (FDTD) implementation of nonlocal plasmonic models would be advantageous due to the widespread use of FDTD in the photonics community, its relative ease of implementation, the availability of reliable open source codes [25], its broadband capabilities, its ability to produce time-domain movies, and its near-linear scalability when run on many processors [26]. Few attempts have been made [27]–[29], however none of them consider high-performance computing and parallel implementations, nor include electron diffusion. In addition, one uses an erroneous approximation, and another deals with terahertz plasmonics. A correct and comprehensive FDTD implementation for plasmonics within and near the visible regime has remained unreported.

In this article, we present a parallel FDTD implementation of nonlocal hydrodynamics, including a GNOR implementation. A parallel implementation in FDTD is especially

Manuscript received May 20, 2020; revised October 31, 2020; accepted November 27, 2020. Date of publication December 21, 2020; date of current version July 7, 2021. This work was supported in part by the SOSCIP, in part by the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs Program, and in part by Vanier Canada Graduate Scholarships. (Corresponding author: Joshua Baxter.)

Joshua Baxter and Lora Ramunno are with the Department of Physics and Center for Research in Photonics, University of Ottawa, Ottawa, ON K1N 6N5, Canada (e-mail: jbx089@uottawa.ca).

Antonio Calà Lesina is with the Hannover Centre for Optical Technologies, Cluster of Excellence PhoenixD (Photonics, Optics, and Engineering – Innovation Across Disciplines), and Fakultät für Maschinenbau (Institut für Transport- und Automatisierungstechnik), Leibniz Universität Hannover, 30167 Hannover, Germany.

This article has supplementary material provided by the authors and color versions of one or more figures available at <https://doi.org/10.1109/TAP.2020.3044579>.

Digital Object Identifier 10.1109/TAP.2020.3044579

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see <https://creativecommons.org/licenses/by/4.0/>

important for simulating nonlocality, as the grid cell size needs to be smaller than the Fermi wavelength $\lambda_F \sim 0.5$ nm to capture the spread of electron density [30]. This can require a large amount of memory and can present a prohibitive computational load. High-performance computing represents a viable solution. The FDTD rectangular-meshing scheme, where a field update at one location requires fields only in adjacent Yee-cells [31], lends itself well to parallel computing; indeed, message passing interface (MPI)-based FDTD solvers are well reported for LRA models [26]. Furthermore, high-performance computing is becoming more accessible with cloud computing, computing consortiums, increased high-performance computing investments, and new, higher-level parallel programming languages such as Chapel [32].

The structure of this article is as follows. We review nonlocal hydrodynamics in Section II, including the most common version without electron diffusion, as well as GNOR, which does include electron diffusion. In Section III, we derive from the nonlocal models FDTD update equations for the polarization field via the auxiliary differential equation (ADE) method. In Section IV, we discuss the implementation of the nonlocal FDTD update equations for parallel computing within an MPI framework. In Section V, we test our implementations by simulating the optical response of small metallic nanospheres and comparing our results to analytic solutions and experimental results. We find an unexpected benefit of nonlocal versus LRA modeling: a marked decrease in staircasing artifacts at the metal boundary. Due to the rectangular discretization inherent in most FDTD approaches, fields can build up in an unphysical manner at plasmonic surfaces, and this can be particularly problematic in applications that rely on plasmonic near-field enhancement [33]. The incorporation of nonlocality significantly decreases this unphysical field build-up. In Section VI, we simulate the response from a spherical nanoparticle-containing sharp extrusions as a demonstration of a large-scale simulation that requires both parallel computing and nonlocal modeling. We find that despite the nanoparticle being larger, the sharp extrusions exhibit plasmonic features not accessible to the LRA. Finally, in Section VII, we give concluding remarks.

II. HYDRODYNAMIC MODELS FOR NONLOCALITY

The most general response of a linear optical material to incident radiation is described by

$$\mathbf{D}(\mathbf{r}, \omega) = \epsilon_0 \int \epsilon(\mathbf{r}, \mathbf{r}', \omega) \mathbf{E}(\mathbf{r}', \omega) d\mathbf{r}' \quad (1)$$

where the dielectric function of the material, $\epsilon(\mathbf{r}, \mathbf{r}', \omega)$, is non-local when the displacement field $\mathbf{D}$ at one location depends on the electric field $\mathbf{E}$ at other locations. In many applications, it is appropriate to use a LRA, wherein $\epsilon(\mathbf{r}, \mathbf{r}', \omega) = \delta(\mathbf{r} - \mathbf{r}') \epsilon(\omega)$. In plasmonic modeling, the interaction of the electric field and the free electron plasma is typically described by the Drude model, which employs the LRA, where the dielectric function reduces to [9]

$$\epsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2 + i\gamma\omega} \quad (2)$$

where ω_p is the plasma frequency, γ is a collisional damping rate, and i is the imaginary unit.

The hydrodynamic plasma model [17] goes beyond the LRA, more accurately describing spatial-temporal free electron dynamics via

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} = -\frac{e}{m} (\mathbf{E} + \mathbf{v} \times \mathbf{B}) - \gamma \mathbf{v} - \frac{1}{m} \nabla \frac{\delta G[n]}{\delta n} \quad (3)$$

where $\mathbf{v}$ is the velocity field of the free electron plasma, $\mathbf{E}$ and $\mathbf{B}$ are the electric and magnetic fields, and n is the free electron density. The energy functional $G[n]$ considers the internal kinetic energy of the electron gas and is usually taken to be the Thomas-Fermi functional, giving

$$\frac{\delta G[n]}{\delta n} = \frac{h^2}{2m} \left(\frac{3}{8\pi} n \right)^{\frac{2}{3}} \quad (4)$$

where h is Planck's constant. As we are only considering the linear nonlocal response in this article, we neglect the nonlinear terms (magnetic-Lorentz and convection) in (3) giving

$$\frac{\partial \mathbf{v}}{\partial t} + \gamma \mathbf{v} = -\frac{e}{m} \mathbf{E} - \frac{\beta^2}{n} \nabla n \quad (5)$$

where $\beta^2 = (1/3) v_F^2$ and v_F is the Fermi velocity. The free electron density fluctuations are accounted for in the last term of (5). This is often referred to as the pressure term, and is responsible for the known blue shift in the surface plasmon resonance with decreasing particle size [13], [17]. Along with this formula for the velocity field, we require the continuity equation given by

$$\frac{\partial n}{\partial t} = -\nabla \cdot (n\mathbf{v}). \quad (6)$$

From here, we consider two approaches. The first is the most widely used and assumes a current density given by $\mathbf{J} = -en\mathbf{v}$. As we use a polarization field formulation in this article, we set $\mathbf{J} = (\partial \mathbf{P}_{NL} / \partial t)$, where we have defined $\mathbf{P}_{NL}$ to be the (nonlocal) free electron polarization field. The velocity field is thus $\mathbf{v} = -(1/ne) (\partial \mathbf{P}_{NL} / \partial t)$, and (6) becomes

$$n = n_0 + \frac{1}{e} \nabla \cdot \mathbf{P}_{NL} \quad (7)$$

where n_0 is the equilibrium free electron density. From (5), we then obtain

$$\frac{\partial^2 \mathbf{P}_{NL}}{\partial t^2} + \gamma \frac{\partial \mathbf{P}_{NL}}{\partial t} = \epsilon_0 \omega_p^2 \mathbf{E} + \beta^2 \nabla (\nabla \cdot \mathbf{P}_{NL}) \quad (8)$$

where $\omega_p = \sqrt{(e^2 n_0 / m \epsilon_0)}$. We call (8) the “nonlocal Drude model” because it consists of the classical LRA Drude model plus one additional term that gives rise to nonlocality (i.e., the term proportional to β^2).

The model represented by (8) differs from those used in the three previous nonlocal FDTD works. In Ref. [27], the gradient-divergence term was simplified to a Laplacian, and this resulted in spurious, nonphysical resonances [18]. In Ref. [23], a current density formulation is used. A benefit to using a polarization field formulation is that it gives ready access to the free electron density via Gauss's Law, as we demonstrate in Section V. In Ref. [29], the thermal kinetic

energy is used instead of the Fermi energy in the pressure term. Because we are interested in plasmonics within and near the visible regime with noble metals, we need to use the Fermi energy since $E_F \approx 215 K_B T$ at room temperature. It is worth noting that Ref. [34] introduces an explicit scheme to solve (3) via FDTD for nonlinear plasmonics without the Thomas–Fermi functional, which is the only nonlocal term that is relevant for linear simulations.

The second approach we consider includes electron diffusion, which was also found to play an important role in the optical response of small metallic particles [20]. This has been described by the GNOR model, where diffusion is considered by modifying the expression for the velocity field via $\mathbf{v} = (-1/en)((\partial \mathbf{P}_G / \partial t) + \mathcal{D} \nabla (\nabla \cdot \mathbf{P}_G))$, where $\mathcal{D}$ is the diffusion coefficient and where we have denoted the GNOR free electron polarization field by $\mathbf{P}_G$.

Using in (5) and (6) the GNOR definition for the velocity field, we obtain the following time-domain nonlocal-diffusive hydrodynamics model for the polarization field, which we hereafter refer to as the “GNOR” model:

$$\frac{\partial^2 \mathbf{P}_G}{\partial t^2} + \gamma \frac{\partial \mathbf{P}_G}{\partial t} = \epsilon_0 \omega_p^2 \mathbf{E} + \eta \nabla (\nabla \cdot \mathbf{P}_G) + \mathcal{D} \frac{\partial}{\partial t} \nabla (\nabla \cdot \mathbf{P}_G) \quad (9)$$

where $\eta = \beta^2 + \mathcal{D}\gamma$. GNOR correctly predicts both the blueshift in the plasmon resonance frequency as well as the broadening of the absorption peak with decreasing particle size. The broadening is the result of the diffusion coefficient, but it is worth noting that this effect can be introduced through other models. For example, one can increase the Drude damping coefficient phenomenologically as a function of nanoparticle size to account for electron scattering from the particle surface [35]. One may also introduce a stress tensor term into the hydrodynamic model [36]. In this article, we will consider only the nonlocal Drude and GNOR models (both of which are thoroughly reviewed in Ref. [17]).

It is from (8) and (9) that we derive in Section III, our FDTD update equations for implementing the two different models of the free electron response of plasmonic material, one accounting for nonlocality only (8), and the other nonlocality with diffusion (9).

To properly model the optical response of many plasmonic materials, one must also include the contribution of interband transitions through bound electrons. We employ the LRA-based N -critical points model that assumes a susceptibility of the form

$$\chi_{CP}(\omega) = (\epsilon_\infty - 1) + \sum_{p=1}^N A_p \Omega_p \left(\frac{e^{i\phi_p}}{\Omega_p - \omega - i\Gamma_p} + \frac{e^{-i\phi_p}}{\Omega_p + \omega + i\Gamma_p} \right) \quad (10)$$

where ϵ_∞ is the infinite frequency permittivity, and A_p, Ω_p, Γ_p , and ϕ_p are fitting parameters. This model can be readily transformed to the time-domain for FDTD implementation, as described in detail in Ref. [37].

III. UPDATE EQUATIONS FOR NONLOCAL FDTD

We derive in this section the FDTD update equations for the time-domain nonlocal Drude and GNOR models

TABLE I
YEE CELL POSITIONS OF THE ELECTRIC, MAGNETIC,
AND POLARIZATION FIELDS

	Electric	Magnetic	Polarization
E_x, H_x, P_x	i+1/2, j, k	i, j+1/2, k+1/2	i+1/2, j, k
E_y, H_y, P_y	i, j+1/2, k	i+1/2, j, k+1/2	i, j+1/2, k
E_z, H_z, P_z	i, j, k+1/2	i+1/2, j+1/2, k	i, j, k+1/2

[(8) and (9), respectively], using the ADE method. We also present the update equations for the N -critical points model (10) to account for interband transitions which is derived elsewhere [37].

We discretize our domain via the Yee cell [31] where the electric fields are collocated with the polarization fields in time and space; the Yee cell positions of the electric, magnetic, and polarization fields used in this article are listed in Table I. We denote the free-electron polarization field by $\mathbf{P}_f$ (which signifies either $\mathbf{P}_{NL}$ or $\mathbf{P}_G$), and the bound electron polarization field by $\mathbf{P}_{CP}$.

The FDTD update algorithm at time $= n\Delta t$, where Δt is the time step size, consists of updating (in order) the:

- 1) magnetic field $\mathbf{H}^{n+1/2} = f_1(\mathbf{H}^{n-1/2}, \mathbf{E}^n)$,
- 2) electric field $\mathbf{E}^{n+1} = f_2(\mathbf{E}^n, \mathbf{P}_{NL}^n, \mathbf{P}_{CP}^n, \mathbf{H}^{n+1/2})$,
- 3) bound charge polarization $\mathbf{P}_{CP}^{n+1} = f_3(\mathbf{E}^{n+1}, \mathbf{P}_{CP}^n)$,
- 4) free charge polarization $\mathbf{P}_f^{n+1} = f_4(\mathbf{E}^{n+1}, \mathbf{P}_f^n)$,

via update equations f_1, f_2, f_3 , and f_4 , whose form and required inputs depend on the model from which they are derived. The equation f_1 is the magnetic field update derived through discretization of the Maxwell–Faraday equation [38]; we do not derive this here because in plasmonic simulations it is typically unchanged from the vacuum equation. In what follows, we present the electric field update equation f_2 (derived from the Maxwell–Ampère law), the bound charge polarization update equation f_3 [derived from (10)], and the free charge polarization update f_4 (derived from (8) for the nonlocal Drude model, and (9) for the GNOR model). Though we present f_4 for two nonlocal models, one must choose which to use—they cannot be used simultaneously.

We start by deriving f_4 from the nonlocal Drude model. Using central differencing, we discretize (8) centered at time $n\Delta t$, to obtain

$$\frac{\mathbf{P}_{NL}^{n+1} - 2\mathbf{P}_{NL}^n + \mathbf{P}_{NL}^{n-1}}{\Delta t^2} + \gamma \frac{\mathbf{P}_{NL}^{n+1} - \mathbf{P}_{NL}^{n-1}}{2\Delta t} = \epsilon_0 \omega_p^2 \mathbf{E}^n + \beta^2 \nabla (\nabla \cdot \mathbf{P}_{NL}^n). \quad (11)$$

Furthermore, using weighted central averaging as discussed in Ref. [33], we set $\mathbf{E}^n = (\mathbf{E}^{n-1} + 2\mathbf{E}^n + \mathbf{E}^{n+1})/4$ to obtain

$$\mathbf{P}_{NL}^{n+1} = D_1 \mathbf{P}_{NL}^n + D_2 \mathbf{P}_{NL}^{n-1} + D_3 (\mathbf{E}^{n-1} + 2\mathbf{E}^n + \mathbf{E}^{n+1}) + D_{NL} \nabla (\nabla \cdot \mathbf{P}_{NL}^n) \quad (12)$$

where

$$D_1 = \frac{2}{D_d \Delta t^2} \quad (13)$$

$$D_2 = \frac{1}{D_d} \left(\frac{\gamma}{2\Delta t} - \frac{1}{\Delta t^2} \right) \quad (14)$$

$$D_3 = \frac{\varepsilon_0 \omega_p^2}{4D_d} \quad (15)$$

$$D_{NL} = \frac{\beta^2}{D_d} \quad (16)$$

$$D_d = \left(\frac{\gamma}{2\Delta t} + \frac{1}{\Delta t^2} \right). \quad (17)$$

As the x , y , and z components of $\nabla(\nabla \cdot \mathbf{P}_{NL}^n)$ must be collocated with the x , y , and z components of $\mathbf{P}_{NL}$, we have

$$\begin{aligned} \nabla(\nabla \cdot \mathbf{P}_{NL}^n)|_x^{i+1/2,j,k} &= \frac{\partial}{\partial x} \left(\frac{\partial P_x}{\partial x} + \frac{\partial P_y}{\partial y} + \frac{\partial P_z}{\partial z} \right)_x^{i+1/2,j,k} \\ &\approx \frac{P_x^{i+3/2,j,k} - 2P_x^{i+1/2,j,k} + P_x^{i-1/2,j,k}}{\Delta x^2} \\ &\quad + \frac{P_y^{i+1,j+1/2,k} - P_y^{i+1,j-1/2,k} - P_y^{i,j+1/2,k} + P_y^{i,j-1/2,k}}{\Delta x \Delta y} \\ &\quad + \frac{P_z^{i+1,j,k+1/2} - P_z^{i+1,j,k-1/2} - P_z^{i,j,k+1/2} + P_z^{i,j,k-1/2}}{\Delta x \Delta z} \end{aligned} \quad (18a)$$

$$\begin{aligned} \nabla(\nabla \cdot \mathbf{P}_{NL}^n)|_y^{i,j+1/2,k} &= \frac{\partial}{\partial y} \left(\frac{\partial P_x}{\partial x} + \frac{\partial P_y}{\partial y} + \frac{\partial P_z}{\partial z} \right)_y^{i,j+1/2,k} \\ &\approx \frac{P_y^{i,j+3/2,k} - 2P_y^{i,j+1/2,k} + P_y^{i,j-1/2,k}}{\Delta y^2} \\ &\quad + \frac{P_x^{i+1/2,j+1,k} - P_x^{i-1/2,j+1,k} - P_x^{i+1/2,j,k} + P_x^{i-1/2,j,k}}{\Delta x \Delta y} \\ &\quad + \frac{P_z^{i,j+1,k+1/2} - P_z^{i,j+1,k-1/2} - P_z^{i,j,k+1/2} + P_z^{i,j,k-1/2}}{\Delta y \Delta z} \end{aligned} \quad (18b)$$

$$\begin{aligned} \nabla(\nabla \cdot \mathbf{P}_{NL}^n)|_z^{i,j,k+1/2} &= \frac{\partial}{\partial z} \left(\frac{\partial P_x}{\partial x} + \frac{\partial P_y}{\partial y} + \frac{\partial P_z}{\partial z} \right)_z^{i,j,k+1/2} \\ &\approx \frac{P_z^{i,j,k+3/2} - 2P_z^{i,j,k+1/2} + P_z^{i,j,k-1/2}}{\Delta z^2} \\ &\quad + \frac{P_x^{i+1/2,j,k+1} - P_x^{i-1/2,j,k+1} - P_x^{i+1/2,j,k} + P_x^{i-1/2,j,k}}{\Delta x \Delta z} \\ &\quad + \frac{P_y^{i,j+1/2,k+1} - P_y^{i,j-1/2,k+1} - P_y^{i,j+1/2,k} + P_y^{i,j-1/2,k}}{\Delta y \Delta z} \end{aligned} \quad (18c)$$

where, for brevity, the components of $\mathbf{P}_{NL}$ are written as P_a for $a = x, y, z$.

For the GNOR model, the f_4 update equation includes (12)–(18) (with β^2 replaced by η) along with a suitable discretization for the last term of (9), the only term that does not appear in the nonlocal Drude model. As this term cannot be achieved via central differencing, we use an alternate second order backward finite difference scheme given in Table 3 of Ref. [39]

$$\frac{\partial}{\partial t} \nabla(\nabla \cdot \mathbf{P}_G^n) \approx \frac{3\nabla(\nabla \cdot \mathbf{P}_G^n) - 4\nabla(\nabla \cdot \mathbf{P}_G^{n-1}) + \nabla(\nabla \cdot \mathbf{P}_G^{n-2})}{2\Delta t}. \quad (19)$$

The GNOR f_4 update then becomes

$$\begin{aligned} \mathbf{P}_G^{n+1} &= D_1 \mathbf{P}_G^n + D_2 \mathbf{P}_G^{n-1} + D_3 (\mathbf{E}^{n-1} + 2\mathbf{E}^n + \mathbf{E}^{n+1}) \\ &\quad + D_{G1} \nabla(\nabla \cdot \mathbf{P}_G^n) \\ &\quad + D_{G2} (3\nabla(\nabla \cdot \mathbf{P}_G^n) - 4\nabla(\nabla \cdot \mathbf{P}_G^{n-1}) + \nabla(\nabla \cdot \mathbf{P}_G^{n-2})) \end{aligned} \quad (20)$$

where

$$D_{G1} = \frac{\eta}{D_d} \quad (21)$$

$$D_{G2} = \frac{\mathcal{D}}{2\Delta t D_d}. \quad (22)$$

The f_3 update equation for the bound charge polarization field is derived from the critical points model (10) in Ref. [37]. We state it here

$$\mathbf{P}_{CP,p}^{n+1} = C_{1p} \mathbf{P}_{CP,p}^n + C_{2p} \mathbf{P}_{CP,p}^{n-1} + C_{3p} \mathbf{E}^{n+1} + C_{4p} \mathbf{E}^n + C_{5p} \mathbf{E}^{n-1} \quad (23)$$

where

$$C_{1p} = \frac{1}{C_p} \left(\frac{2}{\Delta t^2} - \frac{\Omega_p^2 + \Gamma_p^2}{2} \right) \quad (24)$$

$$C_{2p} = \frac{1}{C_p} \left(\frac{\Gamma_p}{\Delta t} - \frac{1}{\Delta t^2} - \frac{\Omega_p^2 + \Gamma_p^2}{4} \right) \quad (25)$$

$$C_{3p} = \frac{C_{4p}}{2} - \frac{\varepsilon_0 A_p \Omega_p \sin \phi_p}{2\Delta t C_p} \quad (26)$$

$$C_{4p} = \frac{\varepsilon_0 A_p \Omega_p (\Omega_p \cos \phi_p - \Gamma_p \sin \phi_p)}{C_p} \quad (27)$$

$$C_{5p} = \frac{C_{4p}}{2} + \frac{\varepsilon_0 A_p \Omega_p \sin \phi_p}{2\Delta t C_p} \quad (28)$$

$$C_p = \frac{\Gamma_p}{\Delta t} + \frac{1}{\Delta t^2} + \frac{\Omega_p^2 + \Gamma_p^2}{4} \quad (29)$$

where p denotes the p th critical point.

Next we turn to the f_2 update equation for $\mathbf{E}^{n+1}$. Ampère's law is discretized in time to give

$$\begin{aligned} \varepsilon_0 \varepsilon_\infty \frac{\mathbf{E}^{n+1} - \mathbf{E}^n}{\Delta t} + \frac{\mathbf{P}_f^{n+1} - \mathbf{P}_f^n}{\Delta t} \\ + \sum_{p=1}^N \frac{(\mathbf{P}_{CP,p}^{n+1} - \mathbf{P}_{CP,p}^n)}{\Delta t} = \nabla \times \mathbf{H}^{n+1/2}. \end{aligned} \quad (30)$$

For the nonlocal Drude model we set $\mathbf{P}_f = \mathbf{P}_{NL}$, and plug (12) into (30) to obtain

$$\begin{aligned} \mathbf{E}^{n+1} &\left(\varepsilon_0 \varepsilon_\infty + D_3 + \sum_{p=1}^N C_{3p} \right) \\ &= \Delta t (\nabla \times \mathbf{H}^{n+1/2}) \\ &\quad + \left(\varepsilon_0 \varepsilon_\infty - 2D_3 + \sum_{p=1}^N C_{4p} \right) \mathbf{E}^n \\ &\quad - \left(D_3 + \sum_{p=1}^N C_{5p} \right) \mathbf{E}^{n-1} - (D_1 - 1) \mathbf{P}_{NL}^n \end{aligned}$$

$$\begin{aligned}
& - \sum_{p=1}^N (C_{1p} - 1) \mathbf{P}_{CP,p}^n - D_2 \mathbf{P}_{NL}^{n-1} \\
& - \sum_{p=1}^N C_{2p} \mathbf{P}_{CP,p}^{n-1} - D_{NL} \nabla (\nabla \cdot \mathbf{P}_{NL}^n). \quad (31)
\end{aligned}$$

For the GNOR model we set $\mathbf{P}_f = \mathbf{P}_G$, and plug (20) into (30) to obtain

$$\begin{aligned}
& \mathbf{E}^{n+1} \left(\varepsilon_0 \varepsilon_\infty + D_3 + \sum_{p=1}^N C_{3p} \right) \\
& = \Delta t (\nabla \times \mathbf{H}^{n+1/2}) + \left(\varepsilon_0 \varepsilon_\infty - 2D_3 + \sum_{p=1}^N C_{4p} \right) \mathbf{E}^n \\
& - \left(D_3 + \sum_{p=1}^N C_{5p} \right) \mathbf{E}^{n-1} \\
& - (D_1 - 1) \mathbf{P}_G^n - \sum_{p=1}^N (C_{1p} - 1) \mathbf{P}_{CP,p}^n - D_2 \mathbf{P}_G^{n-1} \\
& - \sum_{p=1}^N C_{2p} \mathbf{P}_{CP,p}^{n-1} - D_{G1} \nabla (\nabla \cdot \mathbf{P}_G^n) \\
& - D_{G2} (3 \nabla (\nabla \cdot \mathbf{P}_G^n) - 4 \nabla (\nabla \cdot \mathbf{P}_G^{n-1}) + \nabla (\nabla \cdot \mathbf{P}_G^{n-2})). \quad (32)
\end{aligned}$$

For the GNOR update, to avoid recalculation of $\nabla (\nabla \cdot \mathbf{P}_G^{n-1})$, and $\nabla (\nabla \cdot \mathbf{P}_G^{n-2})$, we store them in arrays; therefore GNOR requires additional memory.

The update equations derived above must be implemented in all cells in which the plasmonic material exists. However, care needs to be taken if the nonlocal material extends to the boundary of the simulation domain. Terminating a nonlocal material with a perfectly matched layer (PML) may result in instability and convergence issues. Within a total field/scattered field (TF/SF) framework—which is applicable to many plasmonic nanostructure scattering problems—this can be overcome by using the nonlocal model only in the total field region. If TF/SF cannot be used, such as for geometries that include a plasmonic substrate, one may use the LRA Drude model in the PML region, and the nonlocal models everywhere else.

Finally, we would like to highlight that since the polarization fields are nonlocal, an additional boundary condition is required at the interface between the plasmonic and external media. In the FDTD implementation, material boundary conditions are implicitly imposed by not applying the plasmonic material update equations outside of the plasmonic material. Thus, the free electron density necessarily vanishes just outside the material boundaries. This implies that $\mathbf{P}_f \cdot \mathbf{n} = 0$ at the material boundary, where $\mathbf{n}$ is the unit vector normal to the boundary; however, the tangential component of $\mathbf{P}_f$ does not vanish at the material boundary [18]. We have confirmed numerically that our code implementation indeed implicitly imposes this additional boundary condition.

TABLE II
SUBDOMAIN UPDATE SCHEME FOR DIFFERENT FIELD COMPONENTS

Component	x dimension	y dimension	z dimension
E_x and P_x	$0 \rightarrow N_x - 1$	$1 \rightarrow N_y$	$1 \rightarrow N_z$
E_y and P_y	$1 \rightarrow N_x$	$0 \rightarrow N_y - 1$	$1 \rightarrow N_z$
E_z and P_z	$1 \rightarrow N_x$	$1 \rightarrow N_y$	$0 \rightarrow N_z - 1$
H_x	$1 \rightarrow N_x$	$0 \rightarrow N_y - 1$	$0 \rightarrow N_z - 1$
H_y	$0 \rightarrow N_x - 1$	$1 \rightarrow N_y$	$0 \rightarrow N_z - 1$
H_z	$0 \rightarrow N_x - 1$	$0 \rightarrow N_y - 1$	$1 \rightarrow N_z$

IV. PARALLEL IMPLEMENTATION OF NONLOCAL FDTD

In this section, we describe a parallel FDTD scheme using the MPI framework. Nonlocal simulations require a lot of memory since the step-size Δx needs to be smaller than the Fermi wavelength. In turn, significant computation time is required because the Courant–Friedrichs–Lewy condition restricts the time-step Δt according to Δx . Parallel computing thus becomes essential.

The simulation domain is decomposed into $n_x \times n_y \times n_z$ MPI processes where n_d is the number of MPI processes in the d direction, with $d = x, y, z$. Each MPI process is identified by a vector (m_x, m_y, m_z) which gives its relative spatial position within the simulation domain, where $0 \leq m_d < n_d$.

Each process performs field updates within its own subdomain (i.e., its section of the simulation domain) defined according to a local grid with $(N_x + 1) \times (N_y + 1) \times (N_z + 1)$ points. The vector (i, j, k) identifies an individual grid cell within the local grid, where i ranges from 0 to N_x , j from 0 to N_y , and k from 0 to N_z , inclusively. The electric and magnetic field components at (i, j, k) correspond to different locations in physical space within the grid cell, according to the Yee cell shown in Table I. For example, $E_x(i, j, k)$ refers to $E_x(i \Delta x + \Delta x/2, j \Delta y, k \Delta z)$, whereas $E_y(i, j, k)$ refers to $E_y(i \Delta x, j \Delta y + \Delta y/2, k \Delta z)$.

The use of domain decomposition requires that data from the boundaries of subdomains be transferred to other subdomains at each time step. For subsequent updates to be executed efficiently, an overlap of information is required between adjacent subdomains. For example, the cells (N_x, j, k) in subdomain (m_x, m_y, m_z) represents the same physical locations as the cells $(0, j, k)$ in subdomain $(m_x + 1, m_y, m_z)$. The update scheme we describe below guarantees that the update equations for a given field component at a given physical location are only applied once.

The components of the magnetic and electric fields in a subdomain are updated via f_1 and f_2 , respectively, however, each is updated for different ranges of indices (i, j, k) as summarized in Table II. The notation $0 \rightarrow N_d$ means we update from index 0 to index N_d , inclusively. Bound and free charge polarization fields are updated via f_3 and f_4 , respectively, for the same ranges of indices as for the electric field, as listed in Table II. To understand this visually, the field update regions for the x and y components of all fields within a single subdomain are illustrated in Fig. 1.

While the components of the magnetic, electric, and bound charge polarization fields in a subdomain all require $(N_x + 1) \times (N_y + 1) \times (N_z + 1)$ values to be stored, the free

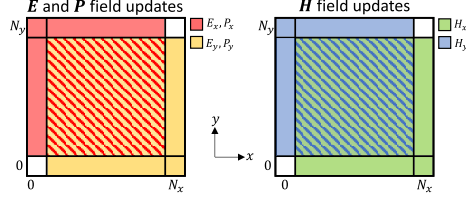


Fig. 1. 2-D representation of the field update regions within a subdomain for the electric/polarization fields (left) and the magnetic fields (right). Different colors are chosen to represent the different field components. The hatched regions indicate where both field components are updated.

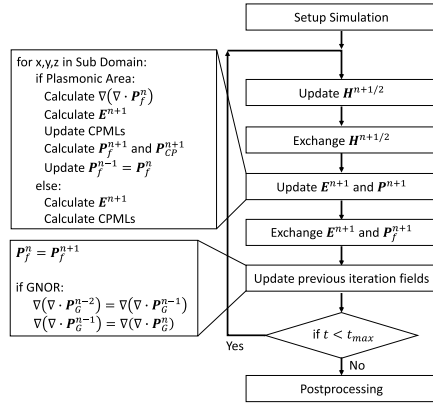


Fig. 2. FDTD algorithm for the nonlocal Drude and GNOR models in an MPI framework.

charge polarization fields requires $(N_x + 2) \times (N_y + 2) \times (N_z + 2)$. This is because the free charge polarization field updates via (18) requires information from additional cells. For example, to update $P_{x,NL}(0, j, k)$ via (18a), we need $P_{x,NL}(-1, j, k)$; in general, we need an extra cell in each dimension to store the -1 index. However, we do not calculate updates at this index, as $(-1, j, k)$ in subdomain (m_x, m_y, m_z) obtain their values from a transfer of data from cells $(N_x - 1, j, k)$ in subdomain $(m_x - 1, m_y, m_z)$.

We now discuss the flow of the FDTD algorithm for both the nonlocal Drude and GNOR models, including what data needs to be communicated, and when. This is summarized in Fig. 2, and we go through each step in detail in the following.

The first update to execute after the setup of the simulation is that of the magnetic field via f_1 . Once this has been completed in each subdomain, a subset of the magnetic field values at the subdomain boundaries must be communicated to adjacent subdomains. The necessary communications for H_x are given in the first row of Table III; those for the other components of $\mathbf{H}$ can be obtained from this table by an ordered permutation of $x \rightarrow y \rightarrow z$. The intersubdomain communications in the x and y directions are illustrated in the top part of Fig. 3. Note that the data transfers of the magnetic

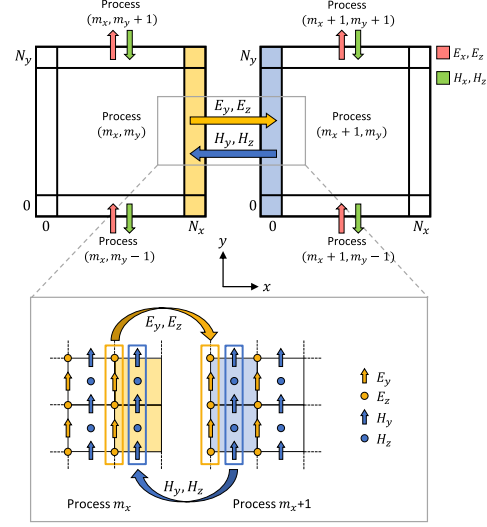


Fig. 3. 2-D representation of magnetic and electric field data transfers between adjacent subdomains. The different colors represent the different field components as indicated. The inset illustrates all field component transfers along x , and their relative positions in the Yee cell.

fields are all made in the “backward” direction. The inset in Fig. 3 details all data transfers along x and their relative Yee cell positions. For example, $H_y(0, j, k)$ updated locally in subdomain $(m_x + 1, m_y, m_z)$ is passed to $H_y(N_x, j, k)$ in subdomain (m_x, m_y, m_z) .

After the magnetic field data transfer, the electric field is updated in each subdomain via f_2 , after which the bound and free charge polarization fields are updated in each subdomain via f_3 and f_4 , respectively. Subsequently, a subset of their values at the subdomain boundaries needs to be communicated to adjacent subdomains. The necessary communications for E_x are given in the second row of Table III; again, those for the other components are obtained by an ordered permutation of $x \rightarrow y \rightarrow z$. The intersubdomain communications for the electric field in the x and y directions are also illustrated in the top part of Fig. 3, with further detail given in the inset. Unlike the magnetic field data transfers, those for the electric field are made in the “forward” direction. For example, $E_y(N_x, j, k)$ updated in subdomain (m_x, m_y, m_z) , is passed to $E_y(0, j, k)$ in subdomain $(m_x + 1, m_y, m_z)$. Note that only the $\mathbf{E}$ and $\mathbf{H}$ components tangential to the direction of the data transfer need to be exchanged.

In general, FDTD updates for LRA polarization models only require the collocated electric and polarization field values so that the polarization field values need not be communicated to other subdomains. Thus, no communication is necessary for the bound charge polarization field $\mathbf{P}_{CP}$. This is not true for the nonlocal models. The necessary communications for the x component of $\mathbf{P}_f$ are given in the third row of Table III and visualized in the top part Fig. 4. The communication for the y

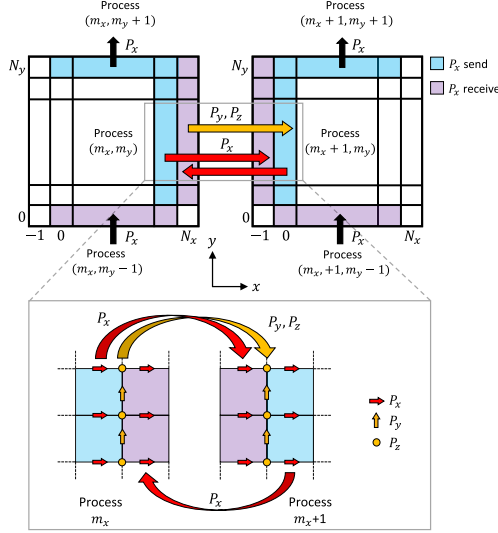


Fig. 4. 2-D representation of the data transfers between adjacent subdomains for the x component of the free charge polarization field. The $j = -1$ row in the y -direction is not shown or required for P_x , but would be required for P_y . The blue shade represents regions where P_x is locally calculated and sent to adjacent processes, and the purple represents regions where P_x is received from adjacent processes. The inset illustrates all free charge polarization field component transfers along x , and their relative positions in the Yee cell.

and z field components are obtained from Table III again via an ordered permutation of $x \rightarrow y \rightarrow z$. The inset of Fig. 4 shows the data transfers along x for all field components. Note that there are now four data transfers for every $\mathbf{P}_f$ component—three “forward” and one “backward”—due to the increased data required for updates via (18). Thus, the inclusion of nonlocality doubles the number of communications required at each time-step. This has an effect on performance and scalability and is discussed further in Section V.

It is worth noting that MPI is not the only solution for parallel computing as new higher-level languages are being introduced for this purpose. Chapel, a language produced by Cray [32], allows for algorithm implementation on a distributed system without the challenges of MPI. For example, [40] presents a finite difference implementation of Poisson’s equation in Chapel. One may also use a shared memory implementation where the memory is shared amongst the processes (or threads) and therefore no data need be communicated. This can be implemented via OpenMP [41] for multithreading on CPUs or via CUDA [42] or OpenCL [43] on graphics processing units (GPUs). Indeed, GPU-FDTD implementations are well reported in literature [44]. Since GPUs can launch thousands of parallel threads that all have access to shared memory, a GPU-based implementation of nonlocal FDTD requires no special treatment beyond what was presented in Section III. While GPUs do suffer from memory constraints, they can still be useful for smaller nonlocal plasmonic simulations.

TABLE III
DATA TRANSFER PROTOCOL FOR THE x -COMPONENT OF THE FIELDS

H_x	Transfer $y = 0$ plane backward to adjacent $y = N_y$ plane Transfer $z = 0$ plane backward to adjacent $z = N_z$ plane
E_x	Transfer $y = N_y$ plane forward to adjacent $y = 0$ plane Transfer $z = N_z$ plane forward to adjacent $z = 0$ plane
$P_{f,x}$	Transfer $x = 0$ plane backward to adjacent $x = N_x$ plane Transfer $x = N_x - 1$ plane forward to adjacent $x = -1$ plane Transfer $y = N_y$ plane forward to adjacent $y = 0$ plane Transfer $z = N_z$ plane forward to adjacent $z = 0$ plane

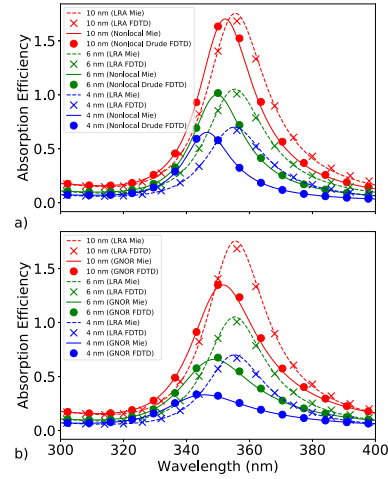


Fig. 5. Absorption cross sections for silver spheres of diameter 4 nm (blue), 6 nm (green), and 10 nm (red) using (a) nonlocal Drude FDTD (filled circles) and (b) GNOR FDTD (filled circles). The solid lines represent the nonlocal Mie theory solutions (a) without diffusion and (b) with diffusion. The dashed lines in both plots represent the LRA Mie theory solutions and the x’s represent the LRA FDTD solutions.

V. NONLOCAL FDTD APPLIED TO SMALL SPHERES

In this section, we test and validate our FDTD implementations of the nonlocal models by using them to simulate the optical response of small silver nanospheres. We compare our results to: 1) analytic solutions based on Mie theory, 2) LRA FDTD plasmonic simulations that employ the LRA Drude model for free-electron response, and 3) experimental results from the literature. Not only do our nonlocal FDTD simulations agree well with analytic and experimental results, but we also find an unexpected benefit over the LRA approach: a pronounced reduction of staircasing artifacts.

In Fig. 5, we plot the absorption efficiencies for silver nanospheres in vacuum for three different diameters—4 nm (blue), 6 nm (green) and 10 nm (red)—and different free charge polarization models. To model silver, we use the fitting parameters reported in [45] for ω_p and γ in (8) and (9), and for all critical points model parameters in (10). We set $\beta^2 = 1/3v_F^2$, where $v_F = 1.39 \cdot 10^6$ m/s [46].

The FDTD domain is a 200 cell $\times$ 200 cell $\times$ 200 cell box truncated by a convolutional perfectly matched

layer (CPML) [38] consisting of 20 additional cells at each boundary. We use a uniform step-size of $\Delta x = \Delta y = \Delta z = 0.04, 0.06$, and 0.1 nm for the 4, 6, and 10 nm particles, respectively, which equates to $D/100$, where D is the diameter of the sphere. For the particle sizes of interest here, this guarantees both that the step-size is less than the Fermi wavelength, and that the spherical shape is sufficiently resolved. Convergence tests confirmed that these step-sizes were sufficient. The total number of iterations varies with step-size and therefore particle size. For a 10 nm diameter particle, $3 \cdot 10^5$ iterations are used to reach convergence; this is scaled appropriately for the other particle sizes.

In Fig. 5(a), we compare our nonlocal Drude FDTD calculations (filled circles) with LRA FDTD calculations (x's), classical (LRA) Mie theory solutions (dashed lines), and with nonlocal Mie theory solutions (solid lines). Nonlocal Mie theory [47]–[49] is a modified version of Mie theory [50] that allows for longitudinal modes, which permits a nonzero free charge density within the nanoparticle. It has been successfully validated with experimental results for small particles larger than several nanometers. For example, a quasistatic version predicted absorption peaks that agree quantitatively with experiment for particle diameters down to 10 nm with a further qualitative agreement down to 2 nm [19].

We find in Fig. 5(a) excellent agreement between nonlocal Drude FDTD and nonlocal Mie theory, with less than 2% mean error for all three particle sizes. Doubling the FDTD step-size increases the mean error for the 10 nm particle case to 6.4%. Halving the step-size decreases it to 1.3%. Thus, reasonable convergence is reached with reasonable simulation domain sizes. As expected, we see an increasing blue-shift in the absorption peak with decreasing nanoparticle size with respect to the LRA model.

In Fig. 5(b) we compare our GNOR-FDTD calculations (filled circles) with LRA FDTD calculations (x's), LRA Mie theory solutions (dashed lines), and nonlocal diffusive Mie theory solutions (solid lines), where diffusion is accounted for by substituting $\beta^2 + D\gamma - iD\omega$ in nonlocal Mie theory [17]. The free electron diffusion coefficient in silver is taken as $D = 3.61 \times 10^{-4}$ m²/s [17]. We again see excellent agreement, with less than 1.6% mean error in GNOR-FDTD relative to nonlocal diffusive Mie theory for all three particle sizes. As expected the resonance positions predicted by the nonlocal Drude FDTD in Fig. 5(a) and GNOR FDTD in Fig. 5(b) are the same, with an increased linewidth for GNOR FDTD.

For the simulations of Fig. 5, we also tracked the CPU time. For the 4 nm sphere simulations (256 cores) we report 155, 166, and 135 thousand time steps per hour for the LRA, nonlocal Drude, and GNOR FDTD simulations, respectively. Similarly, we report 278, 102, and 91 thousand time steps per hour for the 6 nm sphere (128 cores) and 200, 63, and 47 thousand time steps per hour for the 10 nm sphere (64 cores). Note that the supercomputer used for these simulations was comprised of cores that were not all the same speed, which may explain why the time step rate of the 4 nm LRA FDTD simulation is proportionally lower than it should be. In general, the LRA model is faster than the

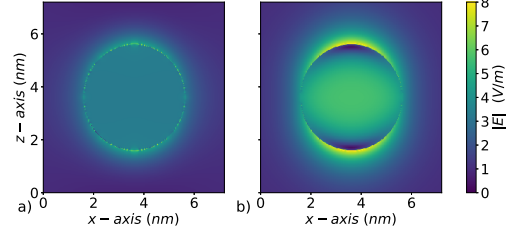


Fig. 6. Electric field amplitude distribution of a 4 nm diameter silver sphere in vacuum at $\lambda = 343$ nm using (a) LRA FDTD and (b) nonlocal Drude FDTD.

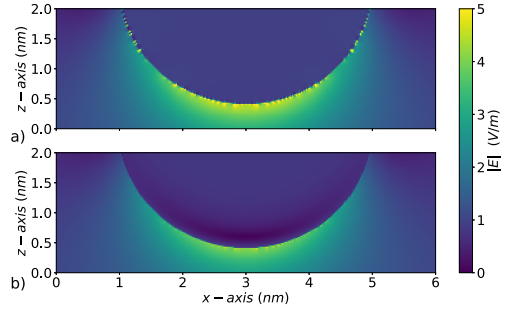


Fig. 7. Close-up of the electric field amplitude distribution of a 4 nm diameter silver sphere in vacuum at $\lambda = 425$ nm. Staircasing artifacts are much more prominent in (a) for LRA FDTD than in (b) for nonlocal Drude FDTD.

nonlocal models and GNOR is slower than the nonlocal Drude model.

We now turn to examine the near-field and free electron density distributions produced by nonlocal Drude FDTD, comparing to those produced by LRA FDTD. In Fig. 6 we show the electric field amplitude distribution for the simulations of the 4 nm diameter silver nanoparticle produced by (a) LRA FDTD and (b) nonlocal Drude FDTD, for the wavelength corresponding to the peak of the (nonlocal) absorption spectrum ($\lambda = 343$ nm). Shown are cuts in the xz -plane through the center of the particle, where the incident plane wave is z -polarized and propagates along the y -axis. The field amplitudes are normalized, corresponding to an input field of 1 V/m. For higher quality images, we used a halved step size of $\Delta x = D/200$. LRA FDTD produces a constant electric field within the sphere, as expected, while nonlocal Drude FDTD produces a field gradient due to the nonzero free charge distribution.

We see in Fig. 6 a significant difference in the appearance of the fields at the particle boundary. While the effects of staircasing in LRA FDTD are clearly visible at the edges of the sphere in Fig. 6(a), they appear smoothed out for nonlocal Drude FDTD in Fig. 6(b). This is even more evident in Fig. 7 which shows the normalized electric field amplitude distribution produced by (a) LRA FDTD and (b) nonlocal

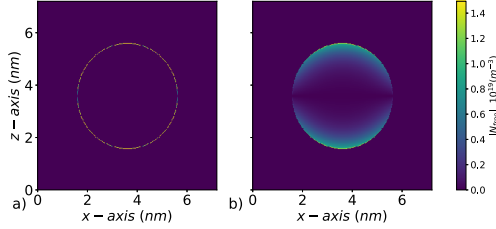


Fig. 8. Free electron density within a 4 nm diameter silver sphere in vacuum at $\lambda = 343$ nm using (a) LRA FDTD and (b) nonlocal Drude FDTD.

Drude FDTD at $\lambda = 425$ nm, a wavelength where the absorption efficiency and near fields for both approaches are almost identical, except for dramatic differences at the sphere boundary. The staircasing-induced boundary artifacts seen for LRA FDTD are clearly reduced with nonlocal Drude FDTD. Interestingly, a smoothing effect in the presence of rough boundaries has been observed in nonlocal plasmonic FEM calculations [51]. In our case, however, the “roughness” is not due to a true physical boundary, but rather one that is artificially imposed on the structure due to the rectangular grid.

Artifacts at the particle boundary are especially problematic for calculations that involve fields just outside the particle, such as, for example, determining the plasmonic near field enhancement of fluorescence, or engineering the spontaneous emission lifetime of fluorophosphores [52]. This could also be important for calculations of plasmonics-enhanced nonlinear optics, where enhanced near fields close to plasmonic boundaries can be harnessed to not only enhance nonlinear optical processes by orders of magnitude, but also shape nonlinear optical fields [53]–[55]. With nonlocal FDTD, the reduction of staircasing artifacts would result in more reliable calculations.

In Fig. 8, we plot the free electron density distribution corresponding to the 4 nm silver particle simulations of Fig. 6. Shown are cuts in the xz plane through the center of the sphere for $\lambda = 343$ nm, as produced by (a) LRA FDTD and (b) nonlocal Drude FDTD. As expected, nonlocal Drude FDTD allows for the spread of the charge distribution near the particle boundary, which we see is on the order of the Fermi wavelength for silver, $\lambda_F = 0.5$ nm. In contrast, for LRA FDTD the charge is bound to the surface.

We present in the Supplementary Information (SI) time-domain movies for the 4 nm silver sphere simulations corresponding to Figs. 6–8 where the incident plane wave pulse function is a normalized raised cosine $f(t) = [(1 - \cos(\omega_{\max}t))/2]^3$, where $\omega_{\max} = 1.26 \cdot 10^{16}$ rad/s is the maximum frequency of interest (corresponding to $\lambda = 150$ nm). Included in the SI are movies of the electric field amplitude dynamics for nonlocal Drude FDTD in the xz and yz planes, *movie1*, and *movie2*, respectively, where both planes cut through the center of the particle. The comparable movies for LRA FDTD are *movie3* and *movie4*. The nonlocal FDTD movies show a radially propagating wave inside the sphere

whereas the LRA FDTD movies do not; in the latter, the field is almost always constant across the sphere (as expected). Furthermore, while staircasing artifacts in the LRA FDTD movies are quite pronounced, they are hardly visible in the nonlocal Drude FDTD movies. The corresponding free electron density movie for nonlocal Drude FDTD is *movie5*. Since the free electron density does not need to be tracked within the LRA FDTD implementation, and all charge is strictly localized to the surface, we do not include a time-domain movie for this case.

We now turn to further validate our FDTD implementations by comparing our simulated results to those of the experiment. In Ref. [13], electron energy loss spectroscopy measurements are presented for silver nanosphere diameters ranging from 2 to 24 nm on a carbon film (3 nm) and compared to Mie theory calculations that use size-dependent permittivities derived quantum mechanically. Using GNOR FDTD, we calculate the absorption spectra of silver nanospheres in a $n = 1.3$ dielectric background (as used in the calculations of Ref. [13]) with diameters ranging from 2 to 24 nm, and plot these in Fig. 9(a). Using the versatility inherent in FDTD, it is quite straightforward to include the effects of the carbon film, unlike what was done with Mie theory. We plot in Fig. 9(b) the absorption spectra of silver nanospheres on a 3 nm film of carbon ($n = 2$ according to Ref. [13]), where agreement with the experimental results was obtained only when a thin dielectric shell (radius of silver sphere/10, $n = 1.5$) around the silver particles was included, which could indicate that an oxide layer had formed. Thus, such FDTD simulations can be useful in helping determine the actual composition of the materials under experimental investigation.

We find the same trend as presented in Ref. [13], showing that GNOR FDTD is consistent with experimental measurements and calculations using quantum-based permittivities, with quantitative agreement down to 10 nm diameter, and qualitative agreement to 2 nm. As discussed in [17], for diameters less than 10 nm, the nonlocal model predicts a resonance shift that is not as large as determined by experimental measurements, which is consistent with our results. This may be due to more complicated phenomena occurring in silver, such as inhomogeneous equilibrium electron density, or spill-out effects, that are not included in the GNOR model we have implemented.

Finally, we turn to a discussion of computational resources and scalability. To give an example, for the 10, 6, and 4 nm sphere simulations presented in Fig. 5, we used 64, 128, and 256 cores, respectively, on the Graham cluster operated by Compute Canada [56]. Recall that the workload is heavier for the smallest particle since our step size Δx (and thus time step size) is proportional to sphere diameter, and thus more iterations in time are required to reach convergence for the smallest particles.

In general, when running simulations on a large number of cores, as we do for this article, it is imperative to investigate the implementation’s scalability, that is, the performance enhancement obtained by increasing the number of CPUs. In the ideal case, the scaling is linear, meaning when one doubles the number of processes, the computation time is

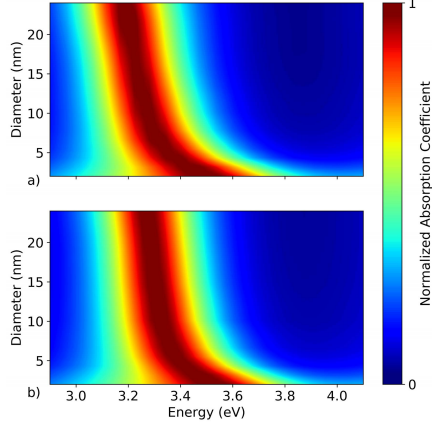


Fig. 9. Contour plot of the absorption cross section of silver spheres (a) embedded in an $n = 1.3$ homogeneous dielectric background and (b) on a 3 nm carbon film with a thin $n = 1.5$ dielectric shell, as a function of incident plane wave photon energy (horizontal axis) and sphere diameter (vertical axis).

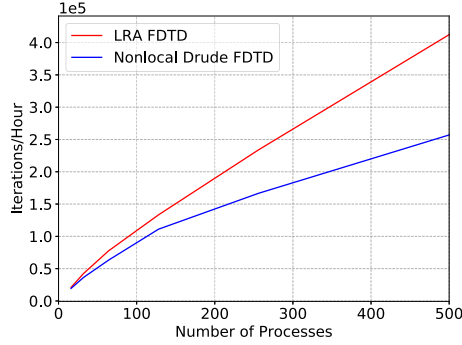


Fig. 10. Scalability of the nonlocal Drude FDTD (blue line) compared to that of LRA FDTD (red line). Plotted is the number of FDTD time iterations computed per hour versus the number of processes used.

halved; this is often not achieved due to overhead, including interprocessor communications.

In Fig. 10, we plot the number of FDTD time iterations calculated per hour as a function of the number of processes used in the nonlocal Drude FDTD (blue line) and LRA FDTD (red line) simulations. The simulation set up is the same as for Fig. 5, where we vary the number of processes while keeping the total number of cells constant. For ≤ 128 processes, the scalability of nonlocal Drude FDTD is nearly linear and comparable to LRA FDTD. While the scalability of LRA FDTD remains mostly linear for a larger number of processes, that of nonlocal Drude FDTD does not, likely due to the doubled interprocess communication required, as discussed in Section IV. As one increases the number of processes while maintaining the same domain size, the subdomain surface-to-volume ratio increases, and therefore the inter-process communication time eventually becomes comparable

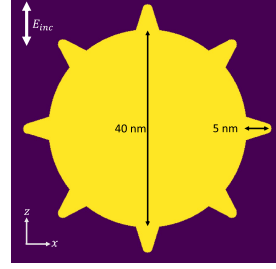


Fig. 11. Cross section in the xz plane of the star-shaped silver nanoparticle, consisting of a sphere with 40 nm diameter and triangular nanoprism extrusions extending 5 nm from the sphere surface. The incident plane wave used to illuminate this structure is z -polarized and propagates along the y axis.

to the computation time within a time step. Thus, if we were to consider a larger domain size, we expect the near-linear scalability to extend to larger process numbers.

VI. NONLOCAL FDTD APPLIED TO LARGER, COMPLEX NANOPARTICLES

In this section, we demonstrate the benefit of our parallel FDTD implementation even further, by considering a much larger plasmonic structure containing small, sharp features. Such a structure would still be expected to exhibit nonlocal effects, and thus simulating it via FDTD would still require a small grid cell size, small enough to resolve electron dynamics within the sharp features. We consider as an example a structure inspired by the star-shaped nanoparticles synthesized for Ref. [57], which are spherical particles around 50 nm in diameter with small 5 nm extrusions from their surfaces that come to a sharp tip.

The structure we simulate is a silver nanosphere with triangular nanoprisms extruding from the equator in the xz plane, perpendicular to the incident plane wave propagation axis, as illustrated in Fig. 11. The sphere is 40 nm in diameter and the nanoprisms are embedded into the sphere so that the effective prism length is 5 nm. The tips of the nanoprisms are rounded with a radius of 1 nm. The FDTD step-size is uniform with $\Delta x = 0.1$ nm and the domain size is $600 \times 600 \times 600$ Yee cells (not including CPMLs). The simulations are run for $4 \cdot 10^5$ iterations over 1000 processes ($10 \times 10 \times 10$).

The absorption spectrum of this particle is shown in Fig. 12, calculated using both LRA FDTD (red) and nonlocal Drude FDTD (blue). There are two resonances attributed to the star-shaped particle. The peaks near 355 nm correspond to the plasmonic resonance of the sphere, and these completely overlap for the two models. This is expected, as nonlocal effects should be negligible for particles larger than 20 nm. The peaks near 575 nm, however, correspond to the resonance of the nanoprisms, specifically the ones aligned parallel to the incident field polarization. Here, we do see that nonlocal effects become important, as the peak wavelength predicted by nonlocal Drude FDTD is 10 nm blue-shifted from that predicted by LRA FDTD.

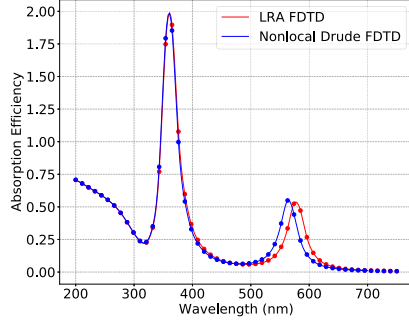


Fig. 12. Absorption efficiency of the silver star-shaped nanoparticle calculated using LRA FDTD (red dots) and nonlocal Drude FDTD (blue dots); the lines are interpolations of the FDTD data.

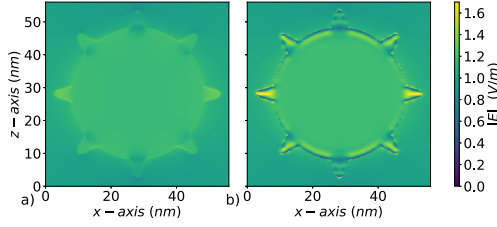


Fig. 13. Electric field amplitude distribution of a star-shaped nanoparticle in vacuum at $\lambda = 320$ nm using (a) LRA FDTD and (b) nonlocal Drude FDTD.

Additional differences between the two models manifest in the field amplitude distributions for wavelengths below the interband transition wavelength $\lambda_{IB} \approx 330$ nm. The electric field amplitude distributions at $\lambda = 320$ nm are shown in Fig. 13 for (a) LRA FDTD and (b) nonlocal Drude FDTD. Standing waves inside the vertical triangles, with a wavevector in the z -direction, are visible in Fig. 13(b) for nonlocal Drude FDTD. This is a longitudinal mode [18] and is expected since its wavelength lies below the epsilon-near-zero wavelength, which in silver is approximately equal to λ_{IB} . In the LRA, these modes cannot be excited by an incident transverse wave, and we see in Fig. 13(a) that they are not. However, with nonlocality, such excitation can occur [18] due to the additional material boundary conditions, discussed in Section III.

These standing waves become more pronounced for isolated nanotriangles as shown in Fig. 14(b) where nonlocal Drude FDTD was used, whereas they are absent in Fig. 14(a), where LRA FDTD was used. The standing waves found for nonlocal FDTD are damped by interband transitions for wavelengths below 300 nm.

We now turn to examine the effect of staircasing for the star-shaped nanoparticle. We show in Fig. 15 a zoomed-in view of the electric field amplitude distribution at $\lambda = 475$ nm, a wavelength where there is more contrast at the particle boundary than for the wavelength considered in Fig. 13. Despite a very fine mesh of $\Delta x = 0.1$ nm, one can clearly see staircasing artifacts in the electric field amplitude distribution

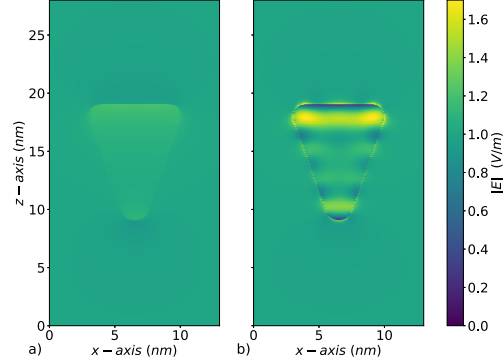


Fig. 14. Electric field amplitude distribution of isolated nanotriangles in vacuum at $\lambda = 320$ nm using (a) LRA FDTD and (b) the nonlocal Drude FDTD.

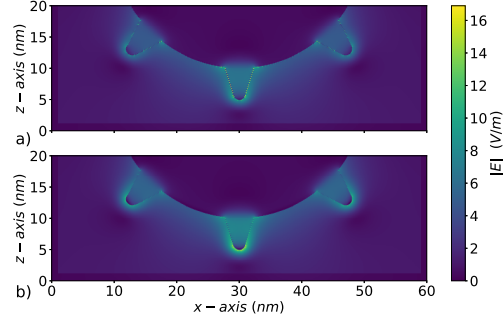


Fig. 15. Close up of the electric field distribution of the star-shaped particle at $\lambda = 475$ nm. The staircasing artifacts are much more prominent in (a) for LRA FDTD than in (b) for nonlocal Drude FDTD.

for LRA FDTD in Fig. 13(a). These are notably reduced in Fig. 13(b), where nonlocal Drude FDTD was used.

This reduction in staircasing artifacts is further illustrated by the time-domain movies of the star-shaped nanoparticle simulations presented in the SI. These depict the time evolution of the electric field amplitude and free electron density in the xz - and yz -planes through the center of the particle, where the incident plane wave is polarized in z and propagates along y . The electric field evolution movies for nonlocal Drude FDTD are *movie6* and *movie7*, while those for the corresponding free electron density evolution are *movie8* and *movie9*. The electric field evolution movies for LRA FDTD are *movie10* and *movie11*.

VII. CONCLUSION

We have introduced a parallel FDTD implementation for modeling nonlocality in plasmonics. We used an ADE approach to model nonlocality with and without electron diffusion and described in detail how to use the MPI framework for parallel computation via domain decomposition. After validating our implementation via comparisons with

analytical and experimental results for small nanospheres, we demonstrated the robustness of our parallel implementation for larger particles with sharp nanoscale features. We find that the inclusion of nonlocality within FDTD significantly reduces staircasing artifacts that can plague standard plasmonic FDTD modeling based on the LRA. This suggests that beyond its importance for modeling nonlocality, nonlocal FDTD might be advantageous for calculations that require precise values for the fields at plasmonic boundaries. This includes, for example, determining plasmonic fluorescence enhancement, plasmonics-mediated fluorescent lifetime engineering, and plasmonics enhanced nonlinear optics.

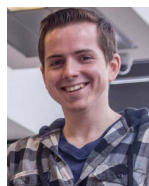
ACKNOWLEDGMENT

The authors would like to thank Compute Canada and Scinet for computational resources and SOSCIP for their computational resources. We would also like to thank Dr. Jean-Michel Guay for his discussion on core-shell particles. A.C.L. acknowledges the Bundesministerium für Bildung und Forschung (German Federal Ministry of Education and Research) under the Tenure-Track Program, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC 2122, Project ID 390833453).

REFERENCES

- [1] V.-C. Su, C. H. Chu, G. Sun, and D. P. Tsai, "Advances in optical metasurfaces: Fabrication and applications [Invited]," *Opt. Exp.*, vol. 26, no. 10, pp. 13148–13182, May 2018. [Online]. Available: <https://www.osapublishing.org/oe/abstract.cfm?uri=oe-26-10-13148>
- [2] J. A. Schuller, E. S. Barnard, W. Cai, Y. C. Jun, J. S. White, and M. L. Brongersma, "Plasmonics for extreme light concentration and manipulation," *Nature Mater.*, vol. 9, no. 3, pp. 193–204, Mar. 2010. [Online]. Available: <https://www.nature.com/articles/nmat2630>
- [3] J. R. Mejía-Salazar and O. N. Oliveira, "Plasmonic biosensing," *Chem. Rev.*, vol. 118, no. 20, pp. 10617–10625, Oct. 2018. [Online]. Available: <https://doi.org/10.1021/acs.chemrev.8b00359>
- [4] S. I. Bozhevolnyi and N. A. Mortensen, "Plasmonics for emerging quantum technologies," *Nanophotonics*, vol. 6, no. 5, pp. 1185–1188, Jan. 2017. [Online]. Available: <https://www.degruyter.com/view/j/nanoph.2017.6.issue-5/nanoph-2016-0179/nanoph-2016-0179.xml>
- [5] M. Kauranen and A. V. Zayats, "Nonlinear plasmonics," *Nature Photon.*, vol. 6, no. 11, pp. 737–748, Nov. 2012. [Online]. Available: <http://www.nature.com/articles/nphoton.2012.244>
- [6] F. Enrichi, A. Quandt, and G. C. Righini, "Plasmonic enhanced solar cells: Summary of possible strategies and recent results," *Renew. Sustain. Energy Rev.*, vol. 82, pp. 2433–2439, Feb. 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S136403211731239X>
- [7] N.-Y. Kim *et al.*, "Localized surface plasmon-enhanced green quantum dot light-emitting diodes using gold nanoparticles," *RSC Adv.*, vol. 5, no. 25, pp. 19624–19629, Feb. 2015. [Online]. Available: <https://pubs.rsc.org/en/content/articlelanding/2015/ra/c4ra15585h>
- [8] M. Sharifi *et al.*, "Plasmonic gold nanoparticles: Optical manipulation, imaging, drug delivery and therapy," *J. Controlled Release*, vols. 311–312, pp. 170–189, Oct. 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0168365919305164>
- [9] S. A. Maier, *Plasmonics: Fundamentals Application*. New York, NY, USA: Springer, 2007.
- [10] P. G. Etchegoin, E. C. Le Ru, and M. Meyer, "An analytic model for the optical properties of gold," *J. Chem. Phys.*, vol. 125, no. 16, Oct. 2006, Art. no. 164705. [Online]. Available: <http://aip.scitation.org/doi/10.1063/1.2360270>
- [11] U. Kreibitz and L. Genzel, "Optical absorption of small metallic particles," *Surf. Sci.*, vol. 156, pp. 678–700, Jun. 1985. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/0039602885902390>
- [12] S. Raza, W. Yan, N. Stenger, M. Wubs, and N. A. Mortensen, "Blueshift of the surface plasmon resonance in silver nanoparticles: Substrate effects," *Opt. Exp.*, vol. 21, no. 22, pp. 27344–27355, Nov. 2013. [Online]. Available: <https://www.osapublishing.org/oe/abstract.cfm?uri=oe-21-22-27344>
- [13] J. A. Scholl, A. L. Koh, and J. A. Dionne, "Quantum plasmon resonances of individual metallic nanoparticles," *Nature*, vol. 483, no. 7390, pp. 421–427, Mar. 2012. [Online]. Available: <http://www.nature.com/articles/nature10904>
- [14] H. Baida *et al.*, "Quantitative determination of the size dependence of surface plasmon resonance damping in single AgSiO₂ nanoparticles," *Nano Lett.*, vol. 9, no. 10, pp. 3463–3469, Oct. 2009. [Online]. Available: <https://pubs.acs.org/doi/10.1021/nl901672b>
- [15] W. Zhu *et al.*, "Quantum mechanical effects in plasmonic structures with subnanometre gaps," *Nature Commun.*, vol. 7, no. 1, Sep. 2016, Art. no. 011495. [Online]. Available: <http://www.nature.com/articles/ncomms11495>
- [16] C. Ciraci *et al.*, "Probing the ultimate limits of plasmonic enhancement," *Science*, vol. 337, no. 6098, pp. 1072–1074, Aug. 2012. [Online]. Available: <http://www.sciencemag.org/cgi/doi/10.1126/science.1224823>
- [17] S. Raza, S. I. Bozhevolnyi, M. Wubs, and N. A. Mortensen, "Nonlocal optical response in metallic nanostructures," *J. Phys., Condens. Matter*, vol. 27, no. 18, May 2015, Art. no. 183204. [Online]. Available: <http://stacks.iop.org/0953-8984/27/i=18/a=183204?key=crossref.8c354f8034f59793e8ac363e22b99808>
- [18] S. Raza, G. Toscano, A.-P. Jauho, M. Wubs, and N. A. Mortensen, "Unusual resonances in nanoplasmonic structures due to nonlocal response," *Phys. Rev. B, Condens. Matter*, vol. 84, no. 12, Sep. 2011, Art. no. 121412. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevB.84.121412>
- [19] S. Raza *et al.*, "Blueshift of the surface plasmon resonance in silver nanoparticles studied with EELS," *Nanophotonics*, vol. 2, no. 2, pp. 131–138, 2013. [Online]. Available: <https://www.degruyter.com/view/j/nanoph.2013.2.issue-2/nanoph-2012-0032/nanoph-2012-0032.xml>
- [20] N. A. Mortensen, S. Raza, M. Wubs, T. Søndergaard, and S. I. Bozhevolnyi, "A generalized non-local optical response theory for plasmonic nanostructures," *Nature Commun.*, vol. 5, no. 1, p. 3809, Sep. 2014. [Online]. Available: <http://www.nature.com/articles/ncomms4809>
- [21] K. R. Hirmath, L. Zschiedrich, and F. Schmidt, "Numerical solution of nonlocal hydrodynamic Drude model for arbitrary shaped nano-plasmonic structures using Nédélec finite elements," *J. Comput. Phys.*, vol. 231, no. 17, pp. 5890–5896, Jul. 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0021999112002550>
- [22] N. Schmitt, J. Viquerat, C. Scheid, S. Lanteri, M. Moferdt, and K. Busch, "A 3D discontinuous Galerkin time-domain method for nano plasmonics with a nonlocal dispersion model," in *Proc. Prog. Electromagn. Res. Symp.-Spring (PIERS)*, May 2017, pp. 2792–2799.
- [23] X. Zheng, M. Kupresak, R. Mittra, and G. A. E. Vandenbosch, "A boundary integral equation scheme for simulating the nonlocal hydrodynamic response of metallic antennas at deep-nanometer scales," *IEEE Trans. Antennas Propag.*, vol. 66, no. 9, pp. 4759–4771, Sep. 2018. [Online]. Available: <https://ieeexplore.ieee.org/document/8399847>
- [24] J. Niegemann, W. Pernice, and K. Busch, "Simulation of optical resonators using DGTD and FDTD," *J. Opt. A, Pure Appl. Opt.*, vol. 11, no. 11, Nov. 2009, Art. no. 114015. [Online]. Available: <https://iopscience.iop.org/article/10.1088/1464-4258/11/11/114015>
- [25] A. F. Oskooi, D. Roundy, M. Ibanescu, P. Bermel, J. D. Joannopoulos, and S. G. Johnson, "MEEP: A flexible free-software package for electromagnetic simulations by the FDTD method," *Comput. Phys. Commun.*, vol. 181, no. 3, pp. 687–702, Mar. 2010. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S001046550900383X>
- [26] A. C. Lesina, A. Vaccari, P. Berini, and L. Ramunno, "On the convergence and accuracy of the FDTD method for nanoplasmonics," *Opt. Exp.*, vol. 23, no. 8, pp. 10481–10497, Apr. 2015. [Online]. Available: <https://www.osapublishing.org/oe/abstract.cfm?uri=oe-23-8-10481>
- [27] J. M. McMahon, *Topics in Theoretical and Computational Nanoscience: From Controlling Light at the Nanoscale to Calculating Quantum Effects With Classical Electrodynamics* (Springer Theses). New York, NY, USA: Springer-Verlag, 2011. [Online]. Available: <https://www.springer.com/gp/book/9781441982483>
- [28] L. Zhang, "FDTD algorithm for plasmonic nanoparticles with spatial dispersion," M.S. thesis, Dept. Elect. Comput. Eng., Ohio State Univ., Columbus, OH, USA, 2016. [Online]. Available: https://etd.ohiolink.edu/pg_1070:NO:10:P10_ACCESSION_NUM:osu1452174003

- [29] S. Bhardwaj, F. L. Teixeira, and J. L. Volakis, "Fast modeling of terahertz plasma-wave devices using unconditionally stable FDTD methods," *IEEE J. Multiscale Multiphys. Comput. Techn.*, vol. 3, pp. 29–36, 2018. [Online]. Available: <https://ieeexplore.ieee.org/document/8338149/>
- [30] B. Gallinet, J. Butet, and O. J. F. Martin, "Numerical methods for nanophotonics: Standard problems and future challenges," *Laser Photon. Rev.*, vol. 9, no. 6, pp. 577–603, Nov. 2015. [Online]. Available: <http://doi.wiley.com/10.1002/lpor.201500122>
- [31] K. Yee, "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media," *IEEE Trans. Antennas Propag.*, vol. AP-14, no. 3, pp. 302–307, May 1966.
- [32] *Chapel: Productive Parallel Programming*. Accessed: Sep. 4, 2019. [Online]. Available: <https://chapel-lang.org/>
- [33] W. H. P. Pernice, "Finite-difference time-domain methods and material models for the simulation of metallic and plasmonic structures," *J. Comput. Theor. Nanosci.*, vol. 7, no. 1, pp. 1–14, Jan. 2010. [Online]. Available: <http://openurl.ingenta.com/content/cref?genre=article&issn=1546-1955&vol=7&issue=1&page=1>
- [34] M. Fang, Z. Huang, W. E. I. Sha, and X. Wu, "Maxwell-hydrodynamic model for simulating nonlinear terahertz generation from plasmonic metasurfaces," *IEEE J. Multiscale Multiphys. Comput. Techn.*, vol. 2, pp. 194–201, 2017. [Online]. Available: <http://ieeexplore.ieee.org/document/8031967/>
- [35] U. Kreibitz and C. V. Fragstein, "The limitation of electron mean free path in small silver particles," *Zeitschrift für Physik*, vol. 224, no. 4, pp. 307–323, Aug. 1969. [Online]. Available: <http://link.springer.com/10.1007/BF01393059>
- [36] I. Tokatly and O. Pankratov, "Hydrodynamic theory of an electron gas," *Phys. Rev. B, Condens. Matter*, vol. 60, no. 23, pp. 15550–15553, Dec. 1999. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevB.60.15550>
- [37] K. P. Prokolidis and D. C. Zografopoulos, "A unified FDTD/PML scheme based on critical points for accurate studies of plasmonic structures," *J. Lightw. Technol.*, vol. 31, no. 15, pp. 2467–2476, Aug. 2013.
- [38] A. Taflov and S. C. Hagness, *Computational Electrodynamics: The Finite-Difference Time-Domain Method* (Artech House Antennas and Propagation Library), 3rd ed. Boston, MA, USA: Artech House, 2005.
- [39] B. Fornberg, "Generation of finite difference formulas on arbitrarily spaced grids," *Math. Comput.*, vol. 51, no. 184, pp. 699–706, Jan. 1988. [Online]. Available: <http://www.ams.org/jourcgi/jour-getitem?pii=S0025-5718-1988-0935077-0>
- [40] R. Barrett, P. Roth, and S. Poole, "Finite difference stencils implemented using chapel," Future Technol. Group, Comput. Sci. Math. Division, Oak Ridge Nat. Lab., Oak Ridge, TN, USA, Tech. Rep. TM-2007/119, 2007.
- [41] OpenMP, *en-GB OpenMP*. Accessed: Sep. 4, 2019. [Online]. Available: <https://www.openmp.org/>
- [42] NVIDIA Developer. (Mar. 2014). *CUDA*. [Online]. Available: <https://developer.nvidia.com/about-cuda>
- [43] The Khronos Group. (Jul. 2013). *OpenCL*. [Online]. Available: <https://www.khronos.org/opencl/>
- [44] V. Demir and A. Z. Elsherbeni, "Compute unified device architecture (CUDA) based finite-difference time-domain (FDTD) implementation," *ACES J. Appl. Comput. Electromagn. Soc.*, vol. 25, no. 4, p. 303, 2010.
- [45] A. Vial, T. Laroche, M. Dridi, and L. Le Cunff, "A new model of dispersion for metals leading to a more accurate modeling of plasmonic structures using the FDTD method," *Appl. Phys. A, Solids Surf.*, vol. 103, no. 3, pp. 849–853, Jun. 2011. [Online]. Available: <http://link.springer.com/10.1007/s00339-010-6224-9>
- [46] C. Kittel, *Introduction to Solid State Physics*, 8th ed. Hoboken, NJ, USA: Wiley, 2005.
- [47] R. Ruppin, "Optical properties of a plasma sphere," *Phys. Rev. Lett.*, vol. 31, no. 24, pp. 1434–1437, Dec. 1973. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevLett.31.1434>
- [48] R. Ruppin, "Optical properties of small metal spheres," *Phys. Rev. B, Condens. Matter*, vol. 11, no. 8, pp. 2871–2876, Apr. 1975. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevB.11.2871>
- [49] A. Pack, M. Hietschold, and R. Wannemacher, "Failure of local Mie theory: Optical spectra of colloidal aggregates," *Opt. Commun.*, vol. 194, nos. 4–6, pp. 277–287, Jul. 2001. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0030401801013104>
- [50] H. C. Van der Hulst, *Light Scattering by Small Particles*. New York, NY, USA: Dover, 1981.
- [51] A. Wiener, A. I. Fernández-Domínguez, A. P. Horsfield, J. B. Pendry, and S. A. Maier, "Nonlocal effects in the nanofocusing performance of plasmonic tips," *Nano Lett.*, vol. 12, no. 6, pp. 3308–3314, Jun. 2012. [Online]. Available: <https://pubs.acs.org/doi/10.1021/nl301478n>
- [52] D. V. Guzatov, S. V. Vaschenko, V. V. Stankevich, A. Y. Lunevich, Y. F. Glukhov, and S. V. Gaponenko, "Plasmonic enhancement of molecular fluorescence near silver nanoparticles: Theory, modeling, and experiment," *J. Phys. Chem. C*, vol. 116, no. 19, pp. 10723–10733, May 2012. [Online]. Available: <http://pubs.acs.org/doi/10.1021/jp301598w>
- [53] G. Li, S. Zhang, and T. Zentgraf, "Nonlinear photonic metasurfaces," *Nature Reviews Materials*, vol. 2, no. 5, Mar. 2017, Art. no. 017010. [Online]. Available: <https://www.nature.com/articles/natrevmats201710>
- [54] A. Krasnok, M. Tymchenko, and A. Alù, "Nonlinear metasurfaces: A paradigm shift in nonlinear optics," *Mater. Today*, vol. 21, no. 1, pp. 8–21, Jan./Feb. 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S136970211730233X>
- [55] S. Keren-Zur, L. Michaeli, H. Suchowski, and T. Ellenbogen, "Shaping light with nonlinear metasurfaces," *Adv. Opt. Photon.*, vol. 10, no. 1, pp. 309–353, Mar. 2018. [Online]. Available: <https://www.osapublishing.org/aop/abstract.cfm?uri=aop-10-1-309>
- [56] *Graham—CC Doc*. Accessed: Sep. 4, 2019. [Online]. Available: <https://docs.computeCanada.ca/wiki/Graham>
- [57] L. Minati, F. Benetti, A. Chiappini, and G. Speranza, "One-step synthesis of star-shaped gold nanoparticles," *Colloids Surf. A, PhysicoChem. Eng. Aspects*, vol. 441, pp. 623–628, Jan. 2014. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0927775713007796>



Joshua Baxter received the B.Sc. degree in physics/mathematics from the University of Ottawa, Ottawa, ON, Canada, in 2017 and started the M.Sc. degree in the same year. He transferred to the Ph.D. degree in 2019.

His work has involved plasmonic coloring with ultrafast lasers, simulations of laser heating, and computational nanophotonics. His current work focuses on deep learning applications to nanophotonics, nonlinear plasmonics via hydrodynamics, and active metasurfaces.

Mr. Baxter was awarded a Vanier Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council in 2020.



Antonio Calà Lesina received the B.Sc. degree in electronics engineering in 2006 and the M.Sc. degree in telecommunications engineering in 2009 from the University of Catania, Catania, Italy, and the Ph.D. degree in information and communication technologies (computational electromagnetics) from the University of Trento, Trento, Italy, in 2013.

In 2009, he joined the Fondazione Bruno Kessler and the University of Trento. From 2013 to 2020, he worked at the University of Ottawa, Ottawa, ON, Canada, in the groups of Professors Lora Ramunno and Pierre Berini, where he conducted his research on computational nanophotonics, plasmonic coloring, photonics design, nonlinear metasurfaces and optical phased arrays. He is a Professor (W2, tenure-track) with Leibniz Universität Hannover, Hannover, Germany, since 2020.



Lora Ramunno received the Ph.D. degree in physics from the University of Toronto, Toronto, ON, Canada, in 2002.

After a year in a high-tech startup, she returned to academia as a Post-Doctoral Fellow with University of Ottawa (UOttawa), Ottawa, ON. Since 2007, she has been a Faculty Member in Physics with UOttawa, where she is currently a Full Professor. Her theoretical/computational research focuses on nonlinear and nano-optics, including intense laser-matter interaction, nanophotonics, nonlinear optical microscopy, and photonic devices.

Dr. Ramunno held a Canada Research Chair in Computational Nanophotonics from 2007 to 2017 and was the recipient of an Early Researcher Award in 2007. She also has several successful industrial collaborative projects. She is a fellow of the Max-Planck-UOttawa Centre for Extreme and Quantum Photonics, a fellow of the NRC-UOttawa Joint Centre for Extreme Photonics, and a member of the UOttawa Centre for Research in Photonics.

Chapter 4

Hyperpolarizability of Plasmonic Meta-Atoms in Metasurfaces

In this chapter, we apply the theory included in Chapters [1.1](#), [2.1](#), and [2.2](#) to the problem of simulating nonlinear optics generated from plasmonic nanoparticles. As a collaborative effort, I am the second author of this work, published in *ACS Nano Letters* in January 2021.

The work presented in this chapter focuses on the derivation of a simple model for the second-harmonic hyperpolarizability $\beta^{(2)}(2\omega)$ of plasmonic particles. The usefulness of such a model lies in the ability to predict the nonlinear emission from a collection of these particles forming a nonlinear plasmonic metasurface, as demonstrated in the article. The model is applied to the case of plasmonic triangular prisms, chosen due to their simple yet non-symmetric shapes (along the long axis), ideal for second harmonic generation. Validation is performed via three tests; 1.) experimental measurements, 2.) hydrodynamic-FDTD simulations, and 3.) nonlinear scattering theory. The four sets of data are found to be in quantitative agreement after a free parameter a in the $\beta^{(2)}(2\omega)$ model is derived.

Although we were approached for hydrodynamic FDTD simulations, my work extended to fine-tuning the model and deriving the aforementioned free parameter a . For this project, I had to implement the hydrodynamic plasma model in our in-house 3D-FDTD program, a continuation of the work presented in Chapter [3](#). The hydrodynamic model was implemented using both a perturbative and non-perturbative method (introduced in Chapter [2.2](#)). These methods produced equivalent numerical results at the second harmonic; however, the perturbed model allowed for the separation of linear and nonlinear fields, which

was found to be advantageous in producing results for this work.

The physical derivation and meaning of the free parameter are found by recognizing that the first and second-order components of a perturbative expansion in the voltage across the particle should be similar when the induced surface charge q_{ind} is high-enough, $V(q_{ind}) \approx V^{(1)} + aV^{(2)}(q_{ind}) \approx 2V^{(1)}$. A reasonable guess for this critical induced surface charge is the equilibrium free electron charge found at the surface. In other words, we assume $V^{(1)} \approx aV^{(2)}(q_{eq})$ when the induced surface charge density is equal to the equilibrium free electron density $q_{ind} = q_{eq}$. This assumption provides a value for a such that the hyperpolarizability model agrees quantitatively with the results of the validation tests.

The model derived in this work assumes a dipolar nonlinear optical response which further assumes that the wavelength is much larger than the particle, a truth at the fundamental wavelength. However, at the second harmonic, this is only partially accurate. It was found later that the second harmonic radiation from one such triangle contains significant quadrupolar components. Future work for this project may investigate how the simple dipole models can be adapted for higher-order multipoles. For example, perhaps the decomposition of the second harmonic current density into terms dependent on $\nabla \mathbf{E}$ in Chapter 2.2 may be used to derive a quadrupole response model for the nanoparticles. Furthermore, the model parameters (the resistance, inductance and capacitance) used in this work were roughly approximated using physical arguments. However, as found in Ref. [161], a more rigorous treatment can be applied to calculating these parameters in the future.

Supplementary information for this manuscript is available at <https://pubs.acs.org/doi/abs/10.1021/acs.nanolett.0c02991>; however, I do not contribute to the content within, and it will not be included here.

Hyperpolarizability of Plasmonic Meta-Atoms in Metasurfaces

M. Saad Bin-Alam, Joshua Baxter, Kashif M. Awan, Antti Kiviniemi, Yaryna Mamchur, Antonio Calà Lesina, Kosmas L. Tsakmakidis, Mikko J. Huttunen, Lora Ramunno, and Ksenia Dolgaleva*

Cite This: *Nano Lett.* 2021, 21, 51–59

Read Online

ACCESS |

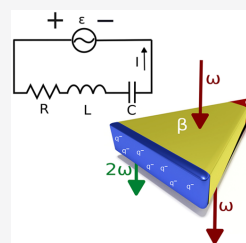
Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: Plasmonic metasurfaces are promising as enablers of nanoscale nonlinear optics and flat nonlinear optical components. Nonlinear optical responses of such metasurfaces are determined by the nonlinear optical properties of individual plasmonic meta-atoms. Unfortunately, no simple methods exist to determine the nonlinear optical properties (hyperpolarizabilities) of the meta-atoms hindering the design of nonlinear metasurfaces. Here, we develop the equivalent *RLC* circuit (resistor, inductor, capacitor) model of such meta-atoms to estimate their second-order nonlinear optical properties, that is, the first-order hyperpolarizability in the optical spectral range. In parallel, we extract from second-harmonic generation experiments the first-order hyperpolarizabilities of individual meta-atoms consisting of asymmetrically shaped (elongated) plasmonic nanoprisms, verified with detailed calculations using both nonlinear hydrodynamic-FDTD and nonlinear scattering theory. All three approaches, analytical, experimental, and computational, yield results that agree very well. Our empirical *RLC* model can thus be used as a simple tool to enable an efficient design of nonlinear plasmonic metasurfaces.

KEYWORDS: Plasmonics, Meta-atoms, Nonlinear Optics, Hyperpolarizability, Anharmonic *RLC* oscillator



Downloaded via UNIV OF NEW MEXICO on May 15, 2021 at 20:46:59 (UTC).
See <https://pubs.acs.org/sharingguidelines> for options on how to legitimately share published articles.

Photonic metamaterials are artificial structures exhibiting optical properties that in natural materials are either very weak or entirely lacking. Among these properties are phase mismatched light-propagation, optical magnetism, strong chirality, and epsilon-near-zero behavior.^{1–5} There is a growing interest in understanding and harnessing the nonlinear optical responses of metasurfaces.^{6–14} Many photonic applications including frequency conversion, THz generation, photon-pair generation, ultrashort-pulse generation, frequency-comb generation, and all-optical switching^{15–19} rely on nonlinear optics occurring in large bulky devices where one must contend with phase mismatching. In contrast, the small footprint of metasurfaces virtually guarantees phase matching, and moreover, the nonlinear emission can be precisely controlled.^{2,3,20,21}

Plasmonic metasurfaces have recently emerged as a promising candidate for enabling nanoscale nonlinear optics.⁶ The optical responses of plasmonic meta-atoms serving as unit cells of metasurfaces are dictated by the collective movement of the conduction electrons giving rise to localized surface plasmons (LSPs). Therefore, it is imperative to investigate the conduction electron dynamics and the nonlinear response of the constituting meta-atoms. Such investigations can be performed, for example, by using the hydrodynamic plasma model^{22–25} and the nonlinear scattering theory.²⁶ However, relying on computational tools is not always convenient due to their complexity and the large amount of computational resources they often require.

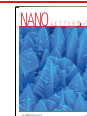
Linear and nonlinear optical responses of plasmonic meta-atoms can also be predicted by equivalent *RLC* circuit

(resistor, inductor, capacitor) theory.^{27–31} This approach has correctly described the nonlinear and magnetic responses of split-ring resonators operating at microwave frequencies.³² Nevertheless, it remained unclear whether the *RLC* approach could describe conduction electron dynamics in plasmonic meta-atoms adequately enough to accurately predict their nonlinear responses at optical wavelengths as well. In this Letter, we derive the first-order hyperpolarizability of individual plasmonic meta-atoms by adapting the equivalent *RLC* model that we find can, indeed, be used to predict collective second-order nonlinear responses of large metasurfaces. First, we derive an expression for the first-order hyperpolarizability with an unknown nonlinear coefficient *a* that represents the strength of the nonlinear charge oscillation in an individual meta-atom. Our goal is to determine the value of *a* through physical arguments and to validate our approach through rigorous experiments and numerical calculations. Next, we describe our second-harmonic generation (SHG) experiments, where we measured SHG emission from metasurfaces consisting of randomly arranged gold elongated nanoprisms. From these measurements, we extract the

Received: July 21, 2020

Revised: December 17, 2020

Published: December 24, 2020



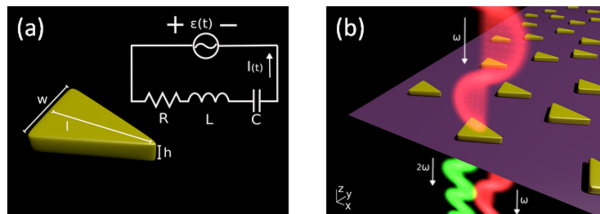


Figure 1. An artist's depiction of (a) an equivalent RLC circuit diagram for the investigated nanoprisms, and (b) metasurface consisting of a random array of elongated gold nanoprisms.

spectrum of their first-order hyperpolarizability. This result is then validated by two sets of finite-difference time domain (FDTD) calculations. In the first, the material nonlinearities are directly implemented within the code via a nonlinear hydrodynamic plasma model. The second is based on nonlinear scattering theory where the experimentally determined second-order nonlinear permittivity of gold is used.

Finally, we present a simple empirical derivation for the unknown RLC -model nonlinear coefficient a , which we then use to perform an order-of-magnitude estimate of the first-order hyperpolarizability. Our intuitive physical interpretation of a is such that the RLC circuit model we present can be extended to describe metasurfaces with different meta-atoms without requiring a priori experimental validation. This simple analytical tool, capable of predicting the hyperpolarizabilities of meta-atoms based on their shapes, dimensions, and material compositions, will prove indispensable in the realization of metasurfaces with tailored nonlinear optical responses.

Because of its resonant nature, a plasmonic meta-atom behaves as an RLC circuit, illustrated in Figure 1a, whose dynamics can be described via³²

$$L\dot{\tilde{I}} + R\tilde{I} + \tilde{V}_C(\tilde{q}) = \tilde{\varepsilon}(t) \quad (1)$$

where $\tilde{I}$ is the free charge current, L is the distributed inductance, R is the distributed resistance, and $\tilde{V}_C$ is the induced voltage due to the effective capacitance C of the circuit. We use a tilde to represent time-varying quantities and an overdot to represent time derivatives.

The free charges of a rodlike plasmonic meta-atom, in the presence of an incident field $\tilde{E}_{\text{inc}} = E_0 \exp(-i\omega t)$ polarized along its length l , will accumulate at a facet resulting in a charge separation and an effective capacitance C . The incident field drives the RLC circuit by creating an electromotive force, which in the dipole approximation, takes the form $\tilde{\varepsilon} = \tilde{E}_{\text{inc}} l$. The dynamics of the free charge is described by the current $\tilde{q} = \tilde{I}$, where $\pm q$ is the induced charge at opposing facets of the nanostructure. We assume that the nonlinearity is sufficiently weak, allowing us to write the nonlinear voltage as³²

$$V_C(\tilde{q}) = (\tilde{q} + a\tilde{q}^2)/C \quad (2)$$

where a quantifies the strength of the nonlinearity. Note that the nonlinear perturbation we assumed here is valid at least up to ~ 10 GW/cm² incident field intensities.³³ Inserting these expressions into eq 1, we obtain

$$\tilde{q} + 2\gamma\dot{\tilde{q}} + \omega_0^2\tilde{q} + a\omega_0^2\tilde{q}^2 = C\omega_0^2\tilde{E}_{\text{inc}} \quad (3)$$

where $\omega_0 = 1/\sqrt{LC}$ is the resonance frequency of the circuit and $\gamma = R/2L$ is the free-electron damping constant of

metals.³² This model accounts for the dispersive nature of metals by implicitly considering a Drude model with γ in eq 3 and ω_p in eq 10b which are given for gold in Methods.³² The RLC model is valid in the same wavelength ranges as the Drude model. Typically for plasmonics, this is when the pump wavelength is in the infrared.

When the excitation wavelengths are close to the LSP resonance of the meta-atom, the conduction electron dynamics are well described using the RLC approach.²⁸ A steady-state solution to eq 3 is found by implementing perturbation theory³⁴

$$\tilde{q} = \frac{C\omega_0^2 l}{D(\omega)} E_{\text{inc}} e^{-i\omega t} - \frac{aC^2\omega_0^6 l^2}{D^2(\omega)D(2\omega)} E_{\text{inc}}^2 e^{-i2\omega t} \quad (4)$$

where $D(\omega') = (\omega_0^2 - \omega'^2 - 2i\gamma\omega')$. We write the total induced dipole moment as $\tilde{p} = \tilde{q}l$, and recall that the dipole moment can be expressed in the frequency domain as³⁴

$$\tilde{p} = \epsilon_0\alpha E_{\text{inc}} + \epsilon_0\beta E_{\text{inc}}^2 \quad (5)$$

where α is the linear polarizability and β is the first-order hyperpolarizability of the meta-atom. Combining eqs 4 and 5, we obtain

$$\alpha(\omega) = -\frac{C\omega_0^2 l^2}{\epsilon_0 D(\omega)} \quad (6a)$$

$$\beta(2\omega; \omega) = -a \frac{C^2\omega_0^6 l^3}{\epsilon_0 D(2\omega)D^2(\omega)} \quad (6b)$$

As the lengths of the meta-atoms can be a significant fraction of the exciting wavelength, higher-order multipoles can be excited, and one must consider this possibility in general, especially for nonlinear emission at shorter wavelengths.³⁵ However, in this study our goal is to create a model that gives an order of magnitude estimate of SHG emission, and we therefore consider only the electric dipole term. The hyperpolarizability in eq 6b can be considered an "effective" hyperpolarizability that helps us to achieve this goal. In what follows, we use experimentally and computationally derived hyperpolarizability spectra to determine a quantitative value for the nonlinear coefficient a .

We first need to consider a geometry for our meta-atom. SHG is a coherent second-order nonlinear process, and therefore, very sensitive to the symmetry of the object under investigation.³⁴ In fact, centrosymmetric-shaped structures exhibit very weak electric-dipole SHG responses³⁶ (see also Figure S1 in Supporting Information Section S1). Nanostructures with low symmetry, such as split-ring resonators, L-

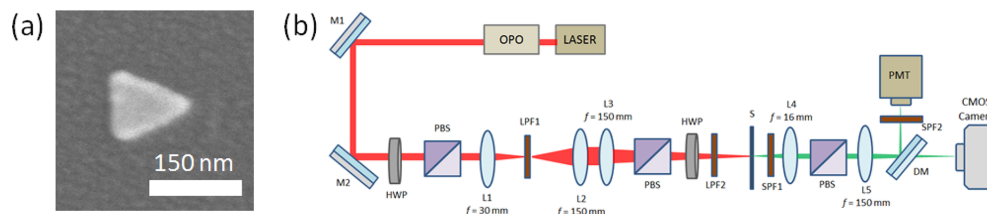


Figure 2. (a) Representative scanning electron micrograph of a fabricated gold nanoprism. (b) Schematic of the nonlinear SHG experimental setup (details in ref 36).

shapes, or nanoprisms exhibit a stronger second-order response, and consequently are especially interesting in studies of even-order nonlinear optical effects.^{36–38} We therefore study elongated triangular nanoprisms due to their low symmetry and simple geometrical shapes.

We now determine the hyperpolarizability through experiments. Four metasurfaces containing gold nanoprisms with widths of $w = 100$ nm, thicknesses of $h = 20$ nm, and lengths of $l = 145, 156, 167$, and 178 nm were fabricated on a fused silica substrate using electron beam lithography, thermal evaporation, and a standard metal lift-off procedure.³⁹ In order to minimize interparticle coupling effects occurring in periodic arrays,^{40–46} each metasurface consisted of 10 000 identical, randomly positioned nanoprisms (oriented in the same direction) deposited into an area of $200 \times 200 \mu\text{m}^2$. This allowed us to investigate ensemble responses that have spectral features identical to the responses of individual meta-atoms.⁴⁷ The arrangement is schematically represented in Figure 1b, while a representative scanning electron micrograph of an individual nanoprism is shown in Figure 2a. Alternatively, individual meta-atoms could be investigated by using nonlinear microscopy.⁴⁸

The transmission spectra of the four metasurfaces are shown in Figure 3a. To verify successful fabrication of the metasurfaces, we compared the measured spectra with those obtained from FDTD (see Methods) [see Figure 3b]. In order for the simulated nanoprism to closely resemble the actual fabricated meta-atoms, we rounded its corners with a circle of 15 nm radius. The LSP resonances of simulated nanoprisms for $l = 145, 156, 167$, and 178 nm peaked at $1030, 1080, 1120$, and 1170 nm, respectively, which are in good agreement with the measurements.

We performed SHG experiments using the setup shown in Figure 2b (described in detail elsewhere³⁶). A laser beam originating from an optical parametric oscillator (Chameleon Compact) was used to illuminate the sample metasurfaces. The optical parametric oscillator was pumped with a Ti:sapphire laser (Chameleon Vision II), generating 200 fs long pulses with a repetition rate of 82 MHz. The average power of the signal beam was kept at 8 mW to avoid potential sample damage via accumulative heating. The SHG emission from the metasurfaces was detected as a function of fundamental wavelength ranging from 1000 – 1300 nm using a power-calibrated photomultiplier tube [see Figure 3c]. The input-beam polarization was set to be linear and aligned with the long axes of the nanoprisms.

We verified the calibration of our setup to provide order-of-magnitude estimates for the first-order hyperpolarizabilities β . This was achieved by measuring SHG emission from a 0.5

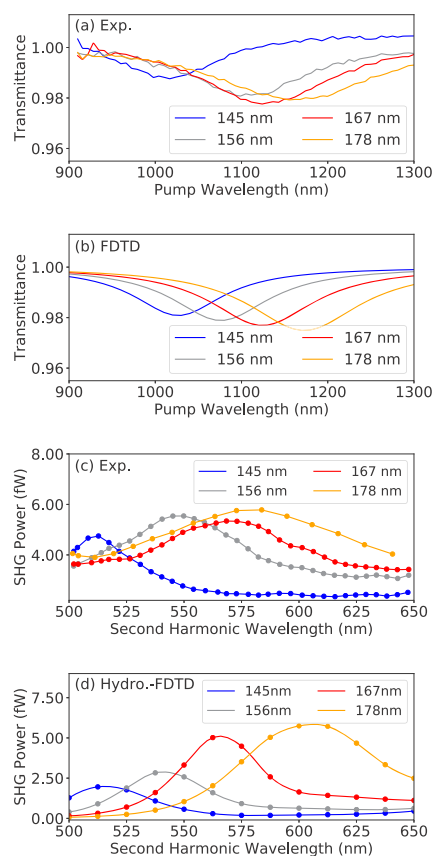


Figure 3. Measured (a) and calculated (b) transmission spectra of random arrays of elongated gold nanoprisms of different lengths l along the direction of the light polarization ($l = 145, 156, 167$, and 178 nm). Measured (c) and calculated (d) SHG emission intensity as a function of second-harmonic wavelength for the same four metasurfaces.

mm-thick Y-cut quartz crystal and using the model described in ref 49 to estimate the second-order susceptibility $\chi^{(2)}$ value

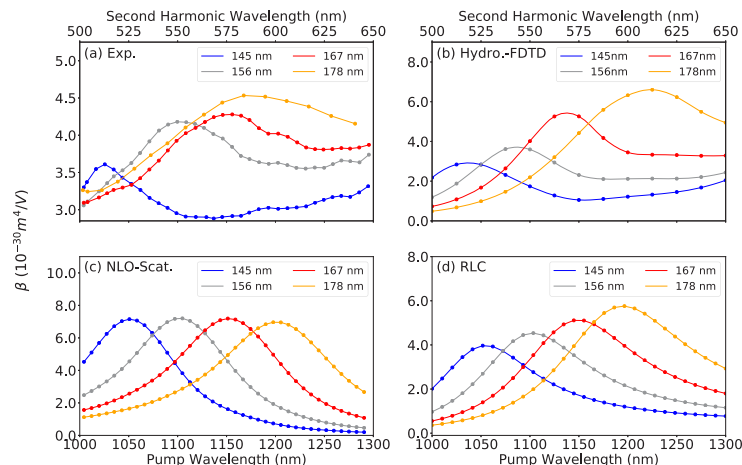


Figure 4. (a) Experimentally extracted values of β for the investigated nanoprism samples. Corresponding values of β , predicted by (b) hydrodynamic-FDTD, (c) nonlinear scattering theory, and (d) the introduced nonlinear RLC model.

for the quartz crystal. Our estimate ($\chi_{\text{xxx}}^{(2)} = 0.53 \text{ pm/V}$) is in excellent agreement with the literature.³⁴

To estimate the values of β for the nanoprisms from the experimental data, we calculated the macroscopic $\chi^{(2)}$ using the same approach as we used for quartz.⁴⁹ We estimated the meta-atom number density of the metasurfaces to be $n = 10\,000/(200 \mu\text{m} \times 200 \mu\text{m} \times 20 \text{ nm}) = 1.25 \times 10^{19} \text{ m}^{-3}$, then linked the detected SHG intensities to the hyperpolarizabilities using the relation $\beta = \chi^{(2)}/n$. The extracted values of β as a function of the incident fundamental wavelength are plotted for the four investigated nanoprisms in Figure 4a.

Next we wish to determine the hyperpolarizability using numerical computations via the hydrodynamic plasma model.^{22–25} This was used in conjunction with the two-critical-points model⁵⁰ for gold in an in-house 3D-FDTD code.⁵¹ As the metasurface contains randomly positioned nanoprisms, periodic boundary conditions could not be used to simplify the calculation. Further, the metasurfaces were too large for a single FDTD simulation. We therefore developed a new method for quantitative approximation of the power. First, we calculated the nonlinear scattered power in the forward direction of a single nanoprism resulting from a Gaussian pulse with duration and peak intensity as in the experiments. For each nanoprism in the random metasurface, the forward-scattered power was scaled to the laser field incident on that particular nanoprism. The individual powers from all nanoprisms in the metasurface were then incoherently summed. The calculated second-harmonic power spectrum, shown in Figure 3d shows excellent quantitative agreement with experimental measurements.

To estimate β using the hydrodynamic plasma model, we first calculate the nonlinear scattered power from a single nanoprism by integrating the nonlinear scattering spectrum from a single pulse and multiplying it by the repetition rate of the laser. If the nonlinear scattered power P_{NL} was purely from a dipole, we can approximate the magnitude of nonlinear dipole moment $|\mathbf{p}|$ by $|\mathbf{p}|^2 = 12\pi P_{\text{NL}}/(nc_0^2 Z_0 k_0^4)$,⁵² where c_0 is the speed of light in vacuum, k_0 is the vacuum wavenumber, Z_0 is

the vacuum impedance, and n is the refractive index of the surrounding medium. We calculate β using eq 5 and plot it in Figure 4b. One can see excellent quantitative and qualitative agreement between the simulation and experimental results.

Finally we use another numerical approach based on the nonlinear scattering theory^{53,54} to determine the nonlinear coefficient a . Though not as rigorous as the hydrodynamic plasma model, it is a simpler and more computationally efficient technique for calculating nonlinear emission from nanostructures. Because the SHG emission was detected only in one direction, and the detector was in the far-field, we were able to use the Lorentz reciprocity theorem.^{26,55} First, we obtained the local field distributions for both the fundamental and SHG wavelengths of interest via linear simulations. The excitation field was assumed to be a normally incident plane wave polarized along the long axis of the nanoprism. The dimensions of the simulated nanoprism were matched with the experimental values, and the sharp corners of the meta-atoms were again rounded.

Then we calculated the generated nonlinear source polarization present on the surface of the meta-atom, where we used the experimentally extracted local nonlinear susceptibility values for gold.⁵⁶ The local field was transformed into the surface coordinate system by reconstructing the surface of the meta-atom by using a Delaunay triangulation mesh and performing subsequent field interpolation onto that surface. Once the nonlinear surface polarization was calculated, we used the Lorentz reciprocity theorem to predict the emitted SHG field from the meta-atom (in the forward direction) by calculating the mode overlap integral between the SHG source polarization and the local field distribution at the SHG wavelength.^{26,57} The last step was to use eq 5 to calculate β . The values of β are plotted in Figure 4c and agree in resonance position and in the order-of-magnitude values with the experimental and hydrodynamic plasma model results.

Now that we have obtained agreement for the values of β from the experimental data and the two numerical approaches, we can estimate the value of the nonlinear coefficient a

required for the *RLC* model from eq 6b. We find that $a \approx 10^{14} \text{ C}^{-1}$, and plot eq 6b, our *RLC* model prediction, in Figure 4d.

While we are confident in the order of magnitude of a , it was quite labor intensive to obtain. We now present a physically intuitive explanation for the order of magnitude of a that we have obtained. In the discussion surrounding Miller's rule presented by Boyd,³⁴ it is argued that for a bulk material, the linear and nonlinear restoring forces felt by an electron are comparable in magnitude when the charge displacement is approximately equal to the interatomic spacing d . This is then used to estimate the strength of the nonlinear restoring force.

In contrast, the polarization of a meta-atom results from an accumulation of free charge on the surface. The charge buildup occurs at the facet within a thin layer with thickness on the order of the Fermi wavelength ($\lambda_F \approx 0.5 \text{ nm}$ for gold). In the absence of external fields, the density of free electrons in this layer would be n_0 , the equilibrium free electron density, and thus there is no excess charge at the surface.

In the presence of an external field, it is reasonable to assume that at most two electrons (with different spins) can reside within λ_F of each other. This would lead to a surface charge density of

$$\sigma \approx \pm en_0 \lambda_F \quad (7)$$

where the sign depends on which side of the meta-atom is being considered. Under this condition one would expect the linear and nonlinear responses to be comparable, in a similar spirit to ref 34; in fact, $\lambda_F \approx d$, the interatomic spacing in gold. From eq 2, the linear and nonlinear terms of V_C are equal when $\tilde{q} = a\tilde{q}^2$, giving $a = 1/\tilde{q}$, where $\tilde{q} = \sigma A$ and A is the surface area containing σ on one side of the meta-atom.

For the nanoprisms considered in this paper, our simulations showed that the region over which there is significant charge density covers a perimeter of approximately 100 nm both at the tip end and at the base end. Given the height of the nanoprism is 20 nm, we obtain

$$a = \frac{1}{en_0 \lambda_F} = 1.1 \times 10^{14} \text{ C}^{-1} \quad (8)$$

where $n_0 = 5.9 \times 10^{28} \text{ m}^{-3}$ in gold. This is remarkably close to what the detailed simulations and measurements predict.

It is also worth mentioning that from the *RLC* model we can obtain the ratio

$$\frac{\beta(2\omega, \omega)}{\alpha^2(\omega)\alpha(2\omega)} = \frac{ae_0^2}{Cl^3} \quad (9)$$

which is, in essence, a form of Miller's rule for meta-atoms. It is frequency independent and only depends on the geometry of the particle. This is similar to the anharmonic oscillator model of ref 34 where the ratios of the susceptibility functions are also only dependent on the effective mass and the number density of the electrons.

From Figure 4, we see that all four approaches yield values of β within the same order of magnitude $\sim 10^{-30} \text{ m}^4/\text{V}$. This agreement is very encouraging because it has been notoriously difficult to make quantitative predictions of nonlinear optical processes occurring in plasmonic materials.^{9,36} The simpler methods (*RLC* and nonlinear scattering theory), though predicting the correct spectral peak positions, are missing some features that are visible in the experimental results and hydrodynamic calculation, such as the oscillations at longer wavelengths. These features are believed to be caused by a

secondary resonance of the nanoprisms and interband transitions in gold.

Despite the great amount of previous work on nonlinear metasurfaces, only a handful of investigations have provided order-of-magnitude estimates of meta-atom's hyperpolarizabilities.^{35,58–61} Our values of hyperpolarizability are 2–3 orders of magnitude smaller than previous estimates of somewhat similar meta-atoms; this discrepancy is not unexpected, because the experimental setups, the wavelength ranges considered, and the investigated meta-atoms have all been different. While we investigated the coherent SHG emission, earlier investigations measured incoherent hyper-Rayleigh scattering (HRS) signals, and estimated the values of β indirectly by comparing the HRS signals from meta-atoms with HRS signals measured from known solvents. Furthermore, the earlier HRS experiments performed at shorter excitation wavelengths than what we considered. Because the interband transitions of gold start playing a role at wavelengths shorter than 550 nm,⁶² these earlier-extracted hyperpolarizability values may have contained an additional contribution arising from the interband transitions.³⁴

Our results also demonstrate the usefulness of the nonlinear *RLC* approach. Although simple and intuitive, it describes the dynamics of conduction electrons adequately enough to predict the strength of nonlinear optical responses of plasmonic meta-atoms. Nonlinear *RLC* model had earlier been found to accurately describe nonlinear responses of metamaterials at the microwave wavelengths.³² We show here that it can be adapted for optical wavelengths as well. This model can also be applied to meta-atoms of any other geometry made of any other plasmonic material, which will be demonstrated in an upcoming publication.

To conclude, we have derived a nonlinear equivalent *RLC* circuit model that can be used to quickly and accurately predict the nonlinear optical responses of meta-atoms in the visible and near-IR spectral ranges. We fabricated four metasurfaces consisting of randomly positioned gold nanoprisms and characterized their second-harmonic generation emissions. We compared the experimental results with the predictions based on the hydrodynamic plasma model, nonlinear scattering theory, and our nonlinear *RLC* model. All of the results were found to be in good agreement. Our *RLC* approach provides new insights into understanding the nonlinear responses of meta-atoms and opens new possibilities for application-oriented efficient nonlinear metasurface design.^{36,38,63–66}

METHODS

Parameters for the Nonlinear *RLC* Model. In order to find the expressions for the capacitance C and inductance L for an elongated nanoprism [shown in Figure 1a], we first calculated the capacitance C and inductance L of a rectangular nanobar from its geometrical dimensions using the following formulas

$$C = \pi\epsilon_0\epsilon_r r_0 \quad (10a)$$

$$L = \frac{\mu_0 l}{2\pi} \log \frac{4l}{h} + \frac{\mu_0 l}{\frac{\omega_p^2}{c_0^2} wh} \quad (10b)$$

In eq 10a, the capacitance C can be defined as the electric charge divided by the potential difference between the two facets of a rectangular nanobar (similar to ref 28 for a plasmonic cylindrical nanorod).

To keep similarity with the formula for C given in ref 28, we assumed $r_0 \approx 10^{-9}$ nm in eq 10a as a constant parameter mimicking the radius of a cylinder encircling the nanobar. However, due to the geometrical difference between the cylinder and rectangular nanobar, we slightly modified eq 10b for the inductance L to match the LSPR spectral position with the simulation and experimental linear transmission data.

In eq 10b, the self-inductance is given by

$$L_{\text{self}} = \frac{\mu_0 l}{2\pi} \log \frac{4l}{h} \quad (11)$$

the kinetic inductance is given by

$$L_{\text{kinetic}} = \frac{\mu_0 l}{\frac{\omega_p^2}{c_0^2} wh} \quad (12)$$

ϵ_0 is the vacuum permittivity (8.85×10^{-12} F/m), the relative permittivity of glass is $\epsilon_r \approx 3.9$, μ_0 is the vacuum permeability (1.257×10^{-6} H/m), $\omega_p = 13.8 \times 10^{15}$ rad/s is the plasma frequency for gold in the optical regime,⁶⁷ and $c_0 = 3 \times 10^8$ m/s is the speed of light in vacuum.

Simulations. Linear FDTD Simulations. Linear FDTD simulations were performed to calculate the linear transmission spectra of the metasurface using an in-house FDTD solver.⁵¹ The spectra were calculated by subtracting the absorption and backscattering cross sections of all the meta-atoms on the metasurface from the total area of the metasurface. The transmission spectrum is then $T = (A_{\text{metasurface}} - (A_{\text{back-scatt}} + A_{\text{abs}}))/A_{\text{metasurface}}$. The cross-sectional data is calculated from a single meta-atom and scaled by the number of meta-atoms on the metasurface. A standard total-field/scattered-field layout is used to calculate the cross sections and the simulation domain is truncated by convolutionally perfectly matched layers. The linear Drude +2 critical points model⁵⁰ is used for the optical properties of gold and accounts for contributions from the conduction electrons and interband transitions. A broadband raised cosine pulse⁵¹ is used as a source excitation.

Hydrodynamic FDTD Simulations. Hydrodynamic FDTD calculations were conducted using the same in-house FDTD solver. The simulation setup is identical to that used in the linear transmission spectra, except that the hydrodynamic model (solved via centered finite differences) replaces the Drude model (in the Drude +2 critical points model), and the source excitation is replaced with a 200 fs Gaussian pulse centered at wavelengths ranging from 1000 to 1300 nm. All simulations using the in-house FDTD solver were run on the Graham cluster operated by Compute Canada.⁶⁸

Nonlinear Scattering Theory. The nonlinear response of a meta-atom was estimated also by using calculations based on the nonlinear scattering theory and the Lorentz reciprocity theorem.^{26,57} The strength of the SHG emission in the direction of interest was evaluated by calculating a mode overlap integral over the fundamental excitation and SHG emission modes. The relevant field profiles for the fundamental and SHG fields were calculated using the Lumerical FDTD software, and the mode overlap integrals were calculated numerically using Matlab. In the FDTD simulations the optical constant of gold was taken from ref 67. The fields on the surface of the meta-atom were estimated by using Delaunay triangulation, and only the surface contributions were considered when calculating the nonlinear response of the gold.⁵⁶

Fabrication. We used $2 \text{ cm} \times 2 \text{ cm}$ fused silica chips as substrates. The chips were coated with bilayer electron-beam resist, consisting of 50 nm thick PMMA with a molecular weight of 495 k as the bottom layer and 25 nm thick PMMA with a molecular weight of 950 k as the top resist layer. The plasmonic nanostructures were then patterned using 30 kV Raith electron-beam lithography system (CRPuO, uOttawa) with a dose of $550 \mu\text{C}/\text{cm}^2$. The patterned resist was then developed for 2 min in 3:1 MIBK/IPA (methyl isobutyl ketone/isopropyl alcohol), followed by depositing a 20 nm layer of gold by electron-beam evaporation and, finally, a lift-off by immersion in acetone. A computer-aided layout of a randomly arranged elongated nanoprism array that was used to create the resist mask and a schematic of the fabrication process flow are shown in Figure S2 and Figure S3, respectively, in Supporting Information Sections S2 and S3. More details are described in ref 39.

Characterization. Linear Characterization. Linear transmission spectra of the samples were measured using a collimated tungsten-halogen light source (experimental setup is shown in Figure S4 in the Supporting Information Section S4). The incident polarization was controlled using a broadband linear polarizing filter. The entire sample was illuminated, and the transmission from a single device was measured by first using a lens to image the sample plane into an intermediate image plane. The transmission from the correct device was then selected by a translating a variable aperture in this image plane, and by using a second lens to guide the transmitted light into the spectrometer.

Nonlinear Characterization. Signal beam from an optical parametric oscillator (Chameleon Compact) was used to illuminate the sample metasurfaces using a spectral SHG setup described in Supporting Information Section S5 (shown in Figure S5). The optical parametric oscillator was pumped using a Ti:sapphire laser (Chameleon Vision II) generating 200 fs long pulses with a repetition rate of 82 MHz. The average power of the signal beam was kept at 8 mW to avoid potential sample damage via accumulative heating. The SHG emission from the metasurfaces was detected as a function of the fundamental wavelength ranging between 1000 and 1300 nm using a power-calibrated photomultiplier tube (see Figure S5 in the Supporting Information Section S5). The input beam polarization was set to be linear and aligned with the long axes of the nanoprisms.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.nanolett.0c02991>.

Symmetric and nonsymmetric nanostructures' SHG response, computer-aided design of the layout of the randomly oriented elongated nanoprisms, plasmonic metasurface fabrication process, experimental setup for linear transmittance measurement, details of nonlinear experimental setup (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Ksenia Dolgaleva — School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada; Department of Physics, University of

Ottawa, Ottawa, Ontario K1N 6N5, Canada;
Email: ksenia.dolgaleva@uottawa.ca

Authors

- M. Saad Bin-Alam** – School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada; orcid.org/0000-0002-0208-4004
- Joshua Baxter** – Department of Physics, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada; orcid.org/0000-0002-1498-9361
- Kashif M. Awan** – Stewart Blusson Quantum Matter Institute, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada; School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada
- Antti Kiviniemi** – Laboratory of Photonics, Tampere University, FI-33014 Tampere, Finland
- Yaryna Mamchur** – National Technical University of Ukraine, "Igor Sikorsky Kyiv Polytechnic Institute", 03056 Kyiv, Ukraine; School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada
- Antonio Calà Lesina** – Hannover Centre for Optical Technologies, Cluster of Excellence PhoenixD (Photonics, Optics, and Engineering – Innovation Across Disciplines), and Fakultät für Maschinenbau (Institut für Transport- und Automatisierungstechnik), Leibniz Universität Hannover, 30167 Hannover, Germany; orcid.org/0000-0002-9384-6245
- Kosmas L. Tsakmakidis** – Section of Condensed Matter Physics, Department of Physics, National and Kapodistrian University of Athens, Panepistimiopolis, GR-157 84 Athens, Greece
- Mikko J. Huttunen** – Laboratory of Photonics, Tampere University, FI-33014 Tampere, Finland
- Lora Ramunno** – Department of Physics, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acs.nanolett.0c02991>

Author Contributions

K.D. conceived the basic idea for the whole work. K.L.T. conceived the idea of implementing the RLC model. M.S.B.-A. derived the analytical nonlinear RLC model. K.M.A. fabricated the metasurface substrate. M.J.H., M.S.B.-A., and A.K. carried out the measurements. J.B. performed the linear and nonlinear hydrodynamic-FDTD model simulations. J.B. and L.R. derived an expression for the nonlinear coefficient a using intuitive arguments. Y.M. made the illustration. K.D., L.R., M.J.H., K.L.T., and A.C.L. supervised the research and the development of the manuscript. M.S.B.A., J.B., and M.J.H. wrote the first draft of the manuscript. All coauthors subsequently took part in the revision process and approved the final copy of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Robert W. Boyd, Ekaterina Poutrina, John E. Sipe, and Gerd Leuchs for their valuable feedback and suggestions. The authors also thank Martti Kauranen for providing his lab at Tampere University in Finland to perform

the experiment. K.D. and L.R. acknowledge support from the Canada Research Chairs (CRC) Program. K.D. acknowledges the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery program. M.S.B.-A. acknowledges the support of the Ontario Graduate Scholarship (OGS), the University of Ottawa Excellence Scholarship, and the University of Ottawa International Experience Scholarship. J.B. acknowledges the financial support of NSERC Canada Graduate Scholarship-Master's (CGSM) Program and the University of Ottawa Excellence Scholarship, and the computational resources of Compute Canada. M.J.H. acknowledges the support of the Academy of Finland (Grant 308596) and the Flagship of Photonics Research and Innovation (PREIN) funded by the Academy of Finland (Grant 320165). K.L.T. acknowledges support from the General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI) under Grant 1819. A.C.L. acknowledges the Bundesministerium für Bildung und Forschung (German Federal Ministry of Education and Research) under the Tenure-Track Programme, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC 2122, Project ID 390833453).

REFERENCES

- (1) Kildishev, A. V.; Boltasseva, A.; Shalae, V. M. Planar photonics with metasurfaces. *Science* **2013**, *339*, 1232009.
- (2) Yu, N.; Genevet, P.; Kats, M. A.; Aieta, F.; Tietjen, J.-P.; Capasso, F.; Gaburro, Z. Light propagation with phase discontinuities: generalized laws of reflection and refraction. *Science* **2011**, *334*, 333–337.
- (3) Alù, A.; Silveirinha, M. G.; Salandrino, A.; Engheta, N. Epsilon-near-zero metamaterials and electromagnetic sources: Tailoring the radiation phase pattern. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *75*, 155410.
- (4) Zhang, S.; Park, Y.-S.; Li, J.; Lu, X.; Zhang, W.; Zhang, X. Negative index in chiral metamaterials. *Phys. Rev. Lett.* **2009**, *102*, 023901.
- (5) Soukoulis, C. M.; Wegener, M. Past achievements and future challenges in the development of three-dimensional photonic metamaterials. *Nat. Photonics* **2011**, *5*, 523.
- (6) Kauranen, M.; Zayats, A. V. Nonlinear plasmonics. *Nat. Photonics* **2012**, *6*, 737–748.
- (7) Lapine, M.; Shadrivov, I. V.; Kivshar, Y. S. Colloquium: Nonlinear metamaterials. *Rev. Mod. Phys.* **2014**, *86*, 1093–1123.
- (8) Minovich, A. E.; Miroshnichenko, A. E.; Bykov, A. Y.; Murzina, T. V.; Neshev, D. N.; Kivshar, Y. S. Functional and nonlinear optical metasurfaces. *Laser & Photonics Reviews* **2015**, *9*, 195–213.
- (9) Butet, J.; Brevet, P. F.; Martin, O. J. Optical Second Harmonic Generation in Plasmonic Nanostructures: From Fundamental Principles to Advanced Applications. *ACS Nano* **2015**, *9*, 10545–10562.
- (10) Li, G.; Zhang, S.; Zentgraf, T. Nonlinear photonic metasurfaces. *Nat. Rev. Mater.* **2017**, *2*, 1–14.
- (11) Rahimi, E.; Gordon, R. Nonlinear Plasmonic Metasurfaces. *Adv. Opt. Mater.* **2018**, *6*, 1800274.
- (12) Keren-Zur, S.; Michaeli, L.; Suchowski, H.; Ellenbogen, T. Shaping light with nonlinear metasurfaces. *Adv. Opt. Photonics* **2018**, *10*, 309–353.
- (13) Pertsch, T.; Kivshar, Y. Nonlinear optics with resonant metasurfaces. *MRS Bull.* **2020**, *45*, 210–220.
- (14) Lesina, A. C.; Berini, P.; Ramunno, L. Origin of third harmonic generation in plasmonic nanoantennas. *Optical Materials Express*. *Opt. Mater. Express* **2017**, *7*, 1575–1580.

- (15) Kwiat, P. G.; Mattle, K.; Weinfurter, H.; Zeilinger, A.; Sergienko, A. V.; Shih, Y. New high-intensity source of polarization-entangled photon pairs. *Phys. Rev. Lett.* **1995**, *75*, 4337.
- (16) Rasekh, P.; Saliminabi, M.; Yildirim, M.; Boyd, R. W.; Ménard, J.-M.; Dolgaleva, K. Propagation of broadband THz pulses: effects of dispersion, diffraction and time-varying nonlinear refraction. *Opt. Express* **2020**, *28*, 3237–3248.
- (17) Brabec, T.; Krausz, F. Intense few-cycle laser fields: Frontiers of nonlinear optics. *Rev. Mod. Phys.* **2000**, *72*, 545.
- (18) Kippenberg, T. J.; Holzwarth, R.; Diddams, S. A. Microresonator-Based Optical Frequency Combs. *Science* **2011**, *332*, 555–560.
- (19) Shcherbakov, M. R.; Vabishchevich, P. P.; Shorokhov, A. S.; Chong, K. E.; Choi, D. Y.; Staude, I.; Miroshnichenko, A. E.; Neshev, D. N.; Fedyanin, A. A.; Kivshar, Y. S. Ultrafast All-Optical Switching with Magnetic Resonances in Nonlinear Dielectric Nanostructures. *Nano Lett.* **2015**, *15*, 6985–6990.
- (20) Minerbi, E.; Keren-Zur, S.; Ellenbogen, T. Nonlinear metasurface fresnel zone plates for terahertz generation and manipulation. *Nano Lett.* **2019**, *19*, 6072–6077.
- (21) Lin, Z.; Huang, L.; Xu, Z. T.; Li, X.; Zentgraf, T.; Wang, Y. Four-Wave Mixing Holographic Multiplexing Based on Nonlinear Metasurfaces. *Adv. Opt. Mater.* **2019**, *7*, 1900782.
- (22) Sipe, J. E.; So, V. C. Y.; Fukui, M.; Stegeman, G. I. Analysis of second-harmonic generation at metal surface. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1980**, *21*, 4389.
- (23) Scalora, M.; Vincenti, M. A.; De Ceglia, D.; Roppo, V.; Centini, M.; Akozbek, N.; Bloemer, M. J. Second- and third-harmonic generation in metal-based structures. *Phys. Rev. A: At, Mol, Opt. Phys.* **2010**, *82*, 1–14.
- (24) Ciraci, C.; Poutirina, E.; Scalora, M.; Smith, D. R. Second-harmonic generation in metallic nanoparticles: Clarification of the role of the surface. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2012**, *86*, 1–10.
- (25) Ginzburg, P.; Krasavin, A. V.; Wurtz, G. A.; Zayats, A. V. Nonperturbative hydrodynamic model for multiple harmonics generation in metallic nanostructures. *ACS Photonics* **2015**, *2*, 8–13.
- (26) O'Brien, K.; Suchowski, H.; Rho, J.; Salandrino, A.; Kante, B.; Yin, X.; Zhang, X. Predicting nonlinear properties of metamaterials from the linear response. *Nat. Mater.* **2015**, *14*, 379–383.
- (27) Engheta, N.; Salandrino, A.; Alu, A. Circuit Elements at Optical Frequencies: Nanoinductors, Nanocapacitors, and Nanoresistors. *Phys. Rev. Lett.* **2005**, *95*, 095504.
- (28) Huang, C.-p.; Yin, X.-g.; Huang, H.; Zhu, Y.-y. Study of plasmon resonance in a gold nanorod with an LC circuit model. *Opt. Express* **2009**, *17*, 6407.
- (29) Zhou, J.; Koschny, T.; Kafesaki, M.; Economou, E. N.; Pendry, J. B.; Soukoulis, C. M. Saturation of the Magnetic Response of Split-Ring Resonators at Optical Frequencies. *Phys. Rev. Lett.* **2005**, *95*, 223902.
- (30) Tretyakov, S. On geometrical scaling of split-ring and double-bar resonators at optical frequencies. *Metamaterials* **2007**, *1*, 40–43.
- (31) Staffaroni, M.; Conway, J.; Vedantam, S.; Tang, J.; Yablonovitch, E. Circuit analysis in metal-optics. *Photonics Nanostructures - Fundam. Appl.* **2012**, *10*, 166–176.
- (32) Poutirina, E.; Huang, D.; Smith, D. R. Analysis of nonlinear electromagnetic metamaterials. *New J. Phys.* **2010**, *12*, 093010.
- (33) Lippitz, M.; Van Dijk, M. A.; Orrit, M. Third-harmonic generation from single gold nanoparticles. *Nano Lett.* **2005**, *5*, 799–802.
- (34) Boyd, R. W. *Nonlinear optics*, 4th ed.; Academic Press: San Diego, 2020.
- (35) Butet, J.; Duboiset, J.; Bachelier, G.; Russier-Antoine, I.; Benichou, E.; Jonin, C.; Brevet, P. F. Optical second harmonic generation of single metallic nanoparticles embedded in a homogeneous medium. *Nano Lett.* **2010**, *10*, 1717–1721.
- (36) Czaplicki, R.; Kiviniemi, A.; Huttunen, M. J.; Zang, X.; Stolt, T.; Vartiainen, I.; Butet, J.; Kuittinen, M.; Martin, O. J. F.; Kauranen, M. Less is more – enhancement of second-harmonic generation from metasurfaces by reduced nanoparticle density. *Nano Lett.* **2018**, *18*, 7709–7714.
- (37) Klein, M. W.; Enkrich, C.; Wegener, M.; Linden, S. Second-harmonic generation from magnetic metamaterials. *Science* **2006**, *313*, 502–504.
- (38) Huttunen, M. J.; Rasekh, P.; Boyd, R. W.; Dolgaleva, K. Using surface lattice resonances to engineer nonlinear optical processes in metal nanoparticle arrays. *Phys. Rev. A: At, Mol, Opt. Phys.* **2018**, *97*, 053817.
- (39) Awan, K. *Fabrication of III-V Integrated Photonic Devices*. Ph.D. Thesis, Université d'Ottawa/University of Ottawa, Ottawa, 2018.
- (40) Auguie, B.; Barnes, W. L. Collective resonances in gold nanoparticle arrays. *Phys. Rev. Lett.* **2008**, *101*, 143902.
- (41) Meinzer, N.; Barnes, W. L.; Hooper, I. R. Plasmonic meta-atoms and metasurfaces. *Nat. Photonics* **2014**, *8*, 889–898.
- (42) Kataja, M.; Hakala, T. K.; Julku, A.; Huttunen, M. J.; van Dijken, S.; Törmä, P. Surface lattice resonances and magneto-optical response in magnetic nanoparticle arrays. *Nat. Commun.* **2015**, *6*, 7072.
- (43) Huttunen, M. J.; Dolgaleva, K.; Törmä, P.; Boyd, R. W. Ultra-strong polarization dependence of surface lattice resonances with out-of-plane plasmon oscillations. *Opt. Express* **2016**, *24*, 28279.
- (44) Kravets, V. G.; Kabashin, A. V.; Barnes, W. L.; Grigorenko, A. N. Plasmonic Surface Lattice Resonances: A Review of Properties and Applications. *Chem. Rev.* **2018**, *118*, 5912–5951.
- (45) Reshef, O.; Saad-Bin-Alam, M.; Huttunen, G.; Carlow, M.; Sullivan, B. T.; Ménard, J.-M.; Dolgaleva, K.; Boyd, R. W. Multiresonant High-Q Plasmonic Metasurfaces. *Nano Lett.* **2019**, *19*, 6429–6434.
- (46) Bin-Alam, M. S.; Reshef, O.; Mamchur, Y.; Alam, M. Z.; Carlow, G.; Upham, J.; Sullivan, B. T.; Ménard, J.-M.; Huttunen, M. J.; Boyd, R. W.; Dolgaleva, K. Ultra-high-Q resonances in plasmonic metasurfaces. *arXiv* **2020**, <https://arxiv.org/abs/2004.05202v2> (accessed July 10, 2020).
- (47) Shi, L.; Hakala, T. K.; Rekola, H. T.; Martikainen, J.-P.; Moerland, R. J.; Torma, P. Spatial Coherence Properties of Organic Molecules Coupled to Plasmonic Surface Lattice Resonances in the Weak and Strong Coupling Regimes. *Phys. Rev. Lett.* **2014**, *112*, 153002.
- (48) Bautista, G.; Huttunen, M. J.; Mäkitalo, J.; Kontio, J. M.; Simonen, J.; Kauranen, M. Second-harmonic generation imaging of metal nano-objects with cylindrical vector beams. *Nano Lett.* **2012**, *12*, 3207–3212.
- (49) Herman, W. N.; Hayden, L. M. Maker fringes revisited: second-harmonic generation from birefringent or absorbing materials. *J. Opt. Soc. Am. B* **1995**, *12*, 416.
- (50) Prokopydis, K. P.; Zografopoulos, D. C. A Unified FDTD/PML Scheme Based on Critical Points for Accurate Studies of Plasmonic Structures. *J. Lightwave Technol.* **2013**, *31*, 2467–2476.
- (51) Lesina, A. C.; Vaccari, A.; Berini, P.; Ramunno, L. On the convergence and accuracy of the FDTD method for nanoplasmonics. *Opt. Express* **2015**, *23*, 10481–10497.
- (52) Jackson, J. D. *Classical electrodynamics*, 3rd ed.; John Wiley: New York, 1999.
- (53) de Beer, A. G. F.; Roke, S.; Dadap, J. I. Theory of optical second-harmonic and sum-frequency scattering from arbitrarily shaped particles. *J. Opt. Soc. Am. B* **2011**, *28*, 1374–1384.
- (54) Mäkitalo, J.; Suuriniemi, S.; Kauranen, M. Boundary element method for surface nonlinear optics of nanoparticles: erratum. *Opt. Express* **2013**, *21*, 10205.
- (55) Roke, S.; Bonn, M.; Petukhov, A. V. Nonlinear optical scattering: The concept of effective susceptibility. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2004**, *70*, 1–10.
- (56) Wang, F. X.; Rodríguez, F. J.; Albers, W. M.; Ahorinta, R.; Sipe, J. E.; Kauranen, M. Surface and bulk contributions to the second-order nonlinear optical response of a gold film. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2009**, *80*, 4–7.

- (57) De Beer, A. G.; Roke, S. Nonlinear Mie theory for second-harmonic and sum-frequency scattering. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2009**, *79*, 1–9.
- (58) Vance, F. W.; Lemon, B. I.; Hupp, J. T. Enormous Hyper-Rayleigh Scattering from Nanocrystalline Gold Particle Suspensions. *J. Phys. Chem. B* **1998**, *102*, 10091–10093.
- (59) Russier-antoine, I.; Jonin, C.; Nappa, J.; Benichou, E.; Brevet, P. F. Wavelength dependence of the hyper Rayleigh scattering response from gold nanoparticles. *J. Chem. Phys.* **2004**, *120*, 10748.
- (60) Nappa, J.; Revillod, G.; Russier-Antoine, I.; Benichou, E.; Jonin, C.; Brevet, P. F. Electric dipole origin of the second harmonic generation of small metallic particles. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2005**, *71*, 165407.
- (61) Duboisset, J.; Russier-Antoine, I.; Benichou, E.; Bachelier, G.; Jonin, C.; Brevet, P. F. Single Metallic Nanoparticle Sensitivity with Hyper Rayleigh Scattering. *J. Phys. Chem. C* **2009**, *113*, 13477–13481.
- (62) Novotny, L.; Hecht, B. *Principles of nano-optics*, 1st ed.; Cambridge University Press: New York, 2006.
- (63) Michaeli, L.; Keren-Zur, S.; Avayu, O.; Suchowski, H.; Ellenbogen, T. Nonlinear Surface Lattice Resonance in Plasmonic Nanoparticle Arrays. *Phys. Rev. Lett.* **2017**, *118*, 970.
- (64) Huttunen, M. J.; Reshef, O.; Stolt, T.; Dolgaleva, K.; Boyd, R. W.; Kauranen, M. Efficient nonlinear metasurfaces by using multiresonant high-Q plasmonic arrays. *Journal of the Optical Society of America B* **2019**, *36*, 118.
- (65) Chen, S.; Reineke, B.; Li, G.; Zentgraf, T.; Zhang, S. Strong nonlinear optical activity induced by lattice surface modes on plasmonic metasurface. *Nano Lett.* **2019**, *19*, 6278–6283.
- (66) Hooper, D. C.; Kuppe, C.; Wang, D.; Wang, W.; Guan, J.; Odom, T. W.; Valev, V. K. Second harmonic spectroscopy of surface lattice resonances. *Nano Lett.* **2019**, *19*, 165–172.
- (67) Johnson, P. B.; Christy, R. W. Optical Constants of the Noble Metals. *Phys. Rev. B - Condens. Matter Mater. Phys.* **1972**, *6*, 4370.
- (68) Graham - CC Doc. <https://docs.computecanada.ca/wiki/Graham> (accessed Sept. 4, 2019).

Chapter 5

Dynamic nanophotonics in epsilon-near-zero conductive oxide films and metasurfaces: a quantitative, nonlinear, computational model

In this chapter, we apply the knowledge introduced in Chapters [1.1](#), [2.1](#), and [2.2](#) to the problem of modelling the spatiotemporal nonlinear optical response of transparent conducting oxides. Another collaborative effort, I am the first author of this work which was published in *Advanced Photonics Research* in January 2023. Our work in Chapter [4](#) attracted the attention of our collaborator, Prof. Israel De Leon, who was interested in self-consistently modelling epsilon-near-zero (ENZ) nonlinearities in transparent conducting oxides, namely indium tin oxide (ITO).

Due to its high doping concentration, ITO's optical properties depend highly on the conduction electrons' temperature (the electronic energy distribution). It exhibits strong, ultrafast permittivity perturbations when optically excited at high intensities, especially in the bandwidth where its permittivity is near zero. ITO is thus an intriguing material in the realm of time-varying photonics, a new and explosive field of research interested in light wave propagation through materials with time-varying permittivities. However, a self-consistent model would require the coupled solution of the material's electron tem-

perature, the conduction electron dynamics, and the electromagnetic propagation. Our high-performance code and electromagnetic modelling expertise matched perfectly with the research background of our collaborators.

The theory underlying the model is rigorously derived in the thesis of our co-author Luis Cortés Herrera [162], who also inspired the use of the Runge-Kutta method for solving the two-temperature model. My contribution involved implementing the model into the FDTD software and the subsequent fine-tuning, testing and validation. For example, I was able to show that the effects of athermal electrons were negligible, greatly simplifying the model and number of free parameters (see the supplemental information). Furthermore, using the simulations, we tune the model by accounting for increased collisions at high temperatures and adjusting the electron-phonon coupling rate to match experimental pump-probe measurements.

Our results agree remarkably with previously published experimental results, especially when determining the angle and wavelength dependence of the nonlinearity in ITO thin films. We also validated our model in the time domain by comparing our results to published experiments demonstrating frequency translation in ITO; frequency translation (or time-refraction) occurs when the frequency of light changes in a time-varying optical material.

The time-varying nature of ITO's optical properties inspired a question regarding tailored pulses for traversing time-varying media. Evidently, the work of this chapter inspired the work of Chapter 6, using inverse design to design pulse shapes to maximize transmittance through time-varying media. An idea of a future project involves demonstrating pulse shaping in an ITO-based metasurface, which can further enhance the strength of the permittivity change. This model may also be applied to many projects involving active photonics; ITO can be used to enhance modulators, optical switches, and all-optical beam-steering, and this model may be the tool to investigate.

Supplementary information is included in this chapter as the content within is self-written and derived. Furthermore, supporting media in the form of time-domain movies is provided with this article and can be found in a zip folder at <https://onlinelibrary.wiley.com/doi/full/10.1002/adpr.202200280>.

Dynamic Nanophotonics in Epsilon-Near-Zero Conductive Oxide Films and Metasurfaces: A Quantitative, Nonlinear, Computational Model

Joshua Baxter, Adriana Pérez-Casanova, Luis Cortes-Herrera, Antonio Calà Lesina, Israel De Leon, and Lora Ramunno*

The promise of dynamic nanophotonic technologies relies on the confinement and spatiotemporal control of light at the nanoscale. Confinement via plasmonics, dielectric resonators, and waveguides can be complemented with materials whose optical properties can be controlled using nonlinear effects. Transparent conducting oxides (TCOs) exhibit strong optical nonlinearities in their near-zero permittivity spectral region, on the femtosecond timescale. Harnessing full spatiotemporal control over the nonlinear response requires a deeper understanding of the process. To achieve this, a self-consistent multiphysics time-domain model for the nonlinear optical response of TCOs is developed and implemented into a 3D finite-difference time-domain code. Simulations are compared and tuned against recently published experimental results for intense laser irradiation of thin indium tin oxide (ITO) films, achieving good quantitative agreement; the time-domain dynamics of the nonlinear response and the phenomenon of time-refraction are also explored. Finally, by simulating intense laser irradiation of a plasmonic particle on an ITO film, the applicability of the approach to time-varying metasurfaces is demonstrated. As expected, significant enhancement of the nonlinear response of an ITO-based metasurface over bare ITO thin films is found. This work thus enables quantitative nanophotonics design with conductive oxides in their epsilon-near-zero region.

1. Introduction

The nonlinear optical response of materials enables a broad range of capabilities within nanophotonics and plasmonic devices, such as refractive index tuning and frequency conversion.^[1–3] These capabilities are of utmost importance for applications such as all optical data processing,^[4] optical switching,^[5–7] active photonic metasurfaces,^[8] and dynamically reconfigurable nanophotonic systems.^[9–11] However, the efficiency and scalability of nonlinear optical devices is typically limited by the weak nonlinear response of most natural materials. This results in bulky devices that require long interaction lengths and/or large optical intensities, making it difficult to scale up and integrate nonlinear optical processes in photonic devices with a small footprint.

There has been much effort toward the development and understanding of materials with a large nonlinear optical response.^[12–15] In particular, natural and artificial materials that exhibit a vanishingly small electric permittivity at optical frequencies, known as epsilon-near-zero (ENZ) materials, promise to be an efficient platform for enhancing nonlinear optical

to be an efficient platform for enhancing nonlinear optical

J. Baxter, L. Ramunno
Department of Physics and Centre for Research in Photonics
University of Ottawa
Ottawa K1N 6N5, Canada
E-mail: lora.ramunno@uOttawa.ca

A. Pérez-Casanova, L. Cortes-Herrera, I. De Leon
School of Engineering and Sciences
Tecnológico de Monterrey
Monterrey, NL 64849, Mexico

L. Cortes-Herrera
Institute of Optics
University of Rochester
Rochester, NY 14627, USA

A. Calà Lesina
Hannover Centre for Optical Technologies
Cluster of Excellence PhoenixD, and Faculty of Mechanical Engineering
(Institute for Transport and Automation Technology)
Leibniz University Hannover
Hannover 30167, Germany

I. De Leon, L. Ramunno
Max Planck–University of Ottawa Centre for Extreme and Quantum
Photonics
Ottawa K1N 6N5, Canada

I. De Leon
School of Electrical Engineering and Computer Science
University of Ottawa
Ottawa K1N 6N5, Canada

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/adpr.202200280>.

© 2022 The Authors. Advanced Photonics Research published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/adpr.202200280

interactions.^[16–19] Enhanced nonlinear processes such as intensity-dependent refractive index^[20–25] and harmonic generation^[26–29] have been observed in both thin-film and nanostructured ENZ materials. A widely studied subset of ENZ materials includes transparent conductive oxides (TCOs), which are degenerately doped semiconductors such as indium tin oxide (ITO) and aluminum-doped zinc oxide (AZO). Their enhanced nonlinear response,^[20,22,23,26,27] the modification of their linear optical properties through doping,^[30] and their suitability for nanoscale integration through complementary metal–oxide–semiconductor (CMOS) fabrication^[31] have positioned these materials as a promising option for enabling nonlinear nanophotonic and nanoplasmonic devices. Furthermore, their ultrafast response times make them an intriguing platform for active nanophotonics.

The nonlinear optical response of TCOs has been extensively studied in their ENZ spectral ranges,^[20,23] where a large nonlinear response has been experimentally observed. This is attributed to intraband transitions occurring in the nonparabolic region of the conduction band of these materials.^[32,33] For the case of ITO, different models accounting for the temperature dependence of the plasma frequency,^[20,21,34,35] the effect of free-carrier scattering,^[36] other carrier kinetics effects,^[37] and nonlocality^[38,39] have been proposed. However, the temporal nature of the nonlinearity is either not accounted for or, when it has been considered, the modeling of the transient response of electron dynamics is separate from the electrodynamics modeling of optical propagation.^[20,21] Currently, no model can account for the temporal dynamics of the electron distribution and the optical properties simultaneously and self-consistently as the incident pulse propagates through and interacts with the material. This poses a limitation to understanding the nonlinear response of complex time-varying nanophotonic structures that include ENZ materials and limits their design.

In this work, a self-consistent multiphysics numerical model is presented that accounts for the temporal modification of the optical properties and electron distribution in TCO-based ENZ materials, as well as fully accounting for electrodynamical propagation. We use a two-temperature model (TTM) to describe the electron excitation and thermalization and account for its effect on the material's optical properties with a temperature-dependent plasma frequency. The implementation of this numerical model in a finite-difference time-domain (FDTD) electrodynamics solver provides access to local optical properties in the time domain. This enables the simulation of the material's nonlinear optical response in space and time and the design and optimization of complex nonlinear time-varying photonic structures incorporating TCO-based ENZ materials. We validate this model against several experimental measurements, make predictions, and gain insight into several phenomena including the temporal dynamics of the intensity-dependent refractive index, time refraction in transmission and reflection, and the nonlinear response of plasmonic metasurfaces interacting with TCO-based ENZ media.

The structure of the article is as follows. We introduce our spatiotemporal, multiphysics material model for the optical nonlinearity of TCOs in Section 2 and provide an intuitive explanation of the physical processes behind the time-varying optical properties. In Section 3, we refine and validate our model by

reproducing experimental results from experiments published in ref. [20], exploring the effect of the pump pulse's central wavelength, incident angle, and intensity on the nonlinear optical response of ITO thin films. Comparisons between the simulation results and experimental data demonstrate quantitative agreement. In Section 4, we investigate the time-domain nature of the nonlinearity and compare our findings to pump-probe experiments published in ref. [20]. There, we also present movies of our simulations that show how the time-dependent, nonlinear response of the ENZ material modifies the propagation of the electric field amplitude. In Section 5, we use our computational approach to demonstrate time refraction, the change in the central frequency of a pulse while propagating through a medium with time-varying permittivity.^[40] Our results predict a larger frequency shift for reflected light than for transmitted light; only the latter has been measured in experiment.^[2] In Section 6, we simulate an ITO-based time-varying plasmonic metasurface, demonstrating the versatility and power of the FDTD implementation of our model. We find that the presence of the plasmonic particle enhances the effective refractive index change by a factor of 5 over a bare ITO thin film. Finally, in Section 7, we conclude our work.

2. Model for TCO Time-Varying Nonlinear Optical Response

TCOs are wide-bandgap, degenerate semiconductors that possess a metallic optical response in the NIR due to an abundance of free electrons in the conduction band. Therefore, the frequency-domain permittivity of TCOs is well described in the NIR spectral range by the Drude model.^[35]

$$\epsilon(\omega) = \epsilon_{\infty} - \frac{\omega_p^2}{\omega^2 + i\gamma\omega} \quad (1)$$

where $\omega_p = \sqrt{Ne^2/\epsilon_0 m^*}$ is the plasma frequency, γ is the damping coefficient, ϵ_{∞} is the high-frequency permittivity, ω is the angular frequency of the electromagnetic field, m^* is the effective mass of the conduction electrons, e is the charge of an electron, N is the free electron density, and ϵ_0 is the permittivity of free space. According to Equation (1), the wavelength at which the real part of the permittivity equals zero (the zero-crossing wavelength) is given by $\lambda_{ZC} = 2\pi\sqrt{\epsilon_{\infty}c/\omega_p}$, and this identifies the ENZ region.

The properties of the conduction electrons are dictated by the conduction-electron energy band. Typically, the parabolic band approximation can be applied, such that m^* does not change upon interaction with light. This is generally not the case in TCOs exposed to intense, ultrafast laser pulses. Due to their high doping concentration (and thus high carrier density), a correction term is required in the dispersion relation^[41]

$$\frac{\hbar^2 k^2}{2m_0^*} = E + CE^2 \quad (2)$$

where E is the electron energy, $\hbar k$ is the magnitude of the crystal momentum, $\hbar$ is the reduced Planck's constant, m_0^* is the effective mass at the conduction band minimum, and C is a nonparabolicity factor. This nonparabolic band structure results in a

plasma frequency that depends on conduction band electron temperature, T_e . An expression for $\omega_p(T_e)$ is derived in Section S1, Supporting Information, where we use the semiclassical Boltzmann equation in the relaxation-time approximation, assuming an isotropic conduction band. The resulting formula is

$$\omega_p^2(T_e) = \frac{e^2}{3m_0^* \epsilon_0 \pi^2} \left(\frac{2m_0^*}{\hbar^2} \right)^{3/2} \int_0^\infty dE (E + CE^2)^{3/2} \times (1 + 2CE)^{-1} \left(-\frac{\partial f_{FD}(E, T_e)}{\partial E} \right) \quad (3)$$

where $f_{FD} = [\exp((E - \mu)/k_B T_e) + 1]^{-1}$ is the Fermi–Dirac distribution, μ is the chemical potential (discussed later), and k_B is the Boltzmann constant. This temperature dependence of the plasma frequency is a consequence of the energy dependence of the electronic effective mass in the conduction band.^[35] An alternate derivation of Equation (3) is given in ref. [33], resulting in the same equation.

In Figure 1, we plot both the plasma frequency (evaluated via numerical integration of Equation (3)) and the resulting real component of the Drude permittivity (Equation (1)) as a function of electron temperature (T_e) for several wavelengths, using parameters given in Section 3. The plasma frequency decreases monotonically with T_e , leading to a dramatic increase in the relative permittivity. For wavelengths between 1200 and 1300 nm, room-temperature (300 K) ITO is an ENZ material. In fact, for $\lambda > 1240$ nm, room-temperature ITO behaves as a metal since $\text{Re}(\epsilon) < 0$, but as the ITO heats, $\text{Re}(\epsilon)$ becomes positive, and it transitions to a dielectric. We discuss the dynamics of this metal-to-dielectric transition further in Section 4.

Before introducing our model for the electron thermalization dynamics, we first qualitatively describe the effect on the conduction band electron distribution upon irradiation by an intense pulse with a center angular frequency of ω_{pump} . As with other degenerate semiconductors, the Fermi energy E_F is dependent on the doping concentration, though it typically lies in the conduction band. Defining the zero-energy level to be at the

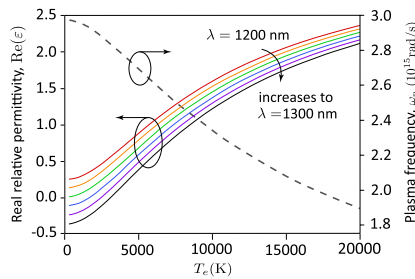


Figure 1. Left vertical axis: Real component of the relative permittivity of ITO plotted as a function of the conduction band electron temperature for wavelengths ranging from 1200 nm (top solid curve) to 1300 nm (bottom solid curve), calculated using Equation (1). The parameters used for this plot are given in Section 3. Right vertical axis: Plasma frequency plotted as a function of electron temperature (dashed line), calculated via numerical integration of Equation (3).

conduction band edge, the Fermi level of ITO is on the order of 1 eV inside the conduction band. As illustrated in Figure 2, when a pump pulse is incident on TCOs initially in equilibrium at an ambient temperature $T_{e,0}$, conduction band electrons with energies E between $E_F - \hbar\omega_{\text{pump}}$ and E_F are excited above the Fermi level, leading to an athermal (nonequilibrium) distribution of electrons (Figure 2b) that is perturbed from the original Fermi–Dirac distribution (Figure 2a).^[42] Subsequent electron–electron scattering results in the athermal energy being rapidly redistributed in the electron gas, resulting in a heated Fermi distribution with temperature $T_{e,1} > T_{e,0}$ (Figure 2c).

Now we turn to deriving our quantitative model. The thermalized electron and lattice subsystems are treated using an extended TTM^[15,43] given by

$$\frac{\partial U(t, \mathbf{r})}{\partial t} = -\frac{U(t, \mathbf{r})}{\tau_{ee}(T_e)} - \frac{U(t, \mathbf{r})}{\tau_{ep}(T_e, T_l)} + P(\mathbf{r}, t) \quad (4a)$$

$$C_e(T_e) \frac{\partial T_e(t, \mathbf{r})}{\partial t} = -g_{ep}(T_e, T_l)(T_e(t, \mathbf{r}) - T_l(t, \mathbf{r})) + \frac{U(t, \mathbf{r})}{\tau_{ee}(T_e)} \quad (4b)$$

$$C_l(T_l) \frac{\partial T_l(t, \mathbf{r})}{\partial t} = g_{ep}(T_e, T_l)(T_e(t, \mathbf{r}) - T_l(t, \mathbf{r})) + \frac{U(t, \mathbf{r})}{\tau_{ep}(T_e, T_l)} \quad (4c)$$

where T_l is the lattice temperature, C_e is the electron heat capacity, C_l is the lattice heat capacity, g_{ep} is the electron–phonon coupling coefficient, τ_{ee} is the electron–electron scattering time, τ_{ep} is the electron–phonon scattering time, and P is the electromagnetic power absorbed by the conduction electrons. In Equation (4), the electron and lattice subsystems are treated separately, and energy is exchanged between the electron and lattice subsystems via electron–phonon coupling g_{ep} . This version is referred to as the extended TTM since it accounts for the athermal electron energy density $U(\mathbf{r}, t)$ (J m^{-3}), a scalar field which acts as a thermal energy source to the conduction electrons.

Due to the relatively low Fermi energy of ITO, the electron–electron scattering time τ_{ee} is on the order of a few femtoseconds (becoming subfemtosecond for a hot electron gas) as we describe in Section S2, Supporting Information. This is advantageous in active nanophotonics as it means that the material response is ultrafast and limited mainly by the exciting pulse width. Because time scales of interest are typically on the order of hundreds of femtoseconds (much larger than τ_{ee}), the effect of including U in our modeling is negligible. For simplicity we thus use a simplified version of the TTM.

$$C_e(T_e) \frac{\partial T_e(t, \mathbf{r})}{\partial t} = -g_{ep}(T_e, T_l)(T_e(t, \mathbf{r}) - T_l(t, \mathbf{r})) + P(\mathbf{r}, t) \quad (5a)$$

$$C_l(T_l) \frac{\partial T_l(t, \mathbf{r})}{\partial t} = g_{ep}(T_e, T_l)(T_e(t, \mathbf{r}) - T_l(t, \mathbf{r})) \quad (5b)$$

The parameters C_e , C_l , and g_{ep} are all temperature dependent; expressions for calculating them are given in Section S3, Supporting Information, along with a method for calculating the temperature-dependent chemical potential μ .

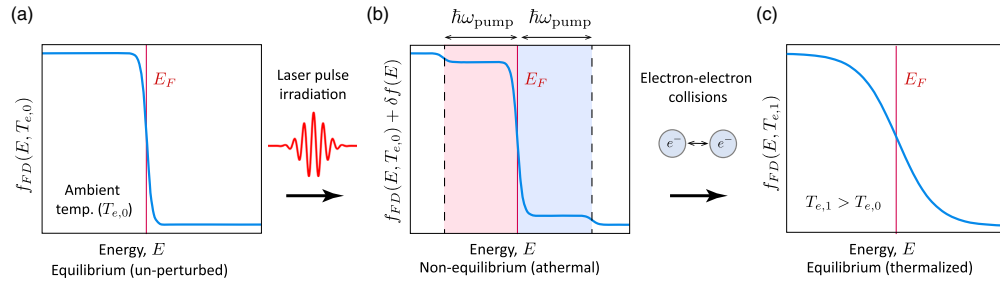


Figure 2. Qualitative illustration of the electron distribution thermalization process. a) Initially the conduction electrons are in thermal equilibrium at temperature $T_{e,0}$ and their energy distribution is described by a Fermi-Dirac distribution $f_{FD}(E, T_{e,0})$. b) The pump pulse acts by exciting electrons below the Fermi level in the range from $(E_F - \hbar\omega_{pump})$ to E_F (red shading) to states above the Fermi level, from E_F to $(E_F + \hbar\omega_{pump})$ (blue shading). This can be described as a perturbation added to the Fermi-Dirac distribution $\delta f(E)$. c) Through electron-electron collisions, the distribution thermalizes to a hot Fermi-Dirac distribution with temperature $T_{e,1} > T_{e,0}$.

In order to complete our model, Maxwell's equations are solved in the time domain concurrently with the time-varying, temperature-dependent Drude model via

$$\mu_0 \frac{\partial \mathbf{H}}{\partial t} = -\nabla \times \mathbf{E} \quad (6)$$

$$\epsilon_\infty \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \nabla \times \mathbf{H} - \mathbf{J} \quad (7)$$

$$\frac{\partial \mathbf{J}}{\partial t} = -\gamma \mathbf{J} + \epsilon_0 \omega_p^2(T_e) \mathbf{E} \quad (8)$$

where $\mathbf{E}$ is the electric field, $\mathbf{H}$ is the magnetic field, and $\mathbf{J}$ is the conduction electron current density that accounts for intraband transitions. For constant ω_p , Equation (8) is equivalent to the standard Drude model (Equation (1)). As there are no interband transitions in the NIR for ITO, we use constant permittivity to describe the effects of the bound electrons that are absorbed into ϵ_∞ , though this can be easily generalized.

The TTM and electrodynamics equations are coupled via two terms: 1) the absorbed power $P = \mathbf{J} \cdot \mathbf{E}$, given by Poynting's theorem, and used as the source in Equation (5a), and 2) the plasma frequency $\omega_p(T_e)$, calculated via Equation (3) and (5), and used as the source of the nonlinearity in the electrodynamics equations. Solved together, Equation (3), (5), (9)–(8) describe the spatiotemporal, thermally driven optical nonlinearity in TCOs.

We focus the implementation of our numerical model to account for the third-order nonlinearity responsible for the temperature-dependent permittivity. A more accurate treatment, especially for geometries with length scales on the order of 10 nm or less, would involve including hydrodynamic terms in Equation (8) and a nonlinear model for the bound electrons as in refs. [38,39]. These extra terms would broaden the functionality of our model by taking into account even-ordered nonlinearities such as second-harmonic generation. After some numerical experiments using our existing hydrodynamics solver,^[44] we found that including hydrodynamic terms has very little effect on the permittivity change when solved in conjunction with the TTM. In fact the hydrodynamic terms are nonlocal, adding

an unnecessary computational load,^[45] and thus we chose to ignore these terms for the present work.

We use an in-house parallel 3D FDTD code^[45,46] to solve Equation (9)–(8). In the same code, we also solve the TTM (Equation (5)) using the Runge-Kutta method.^[47] At each step in the FDTD simulation, the electric field, magnetic field, current density, and the electron and lattice temperatures are updated self-consistently.

3. Model Validation and Refinement in ITO Thin Films

To validate the model presented in the previous section, we compare it against experimental results obtained by femtosecond laser irradiation of ITO thin films. Our choice of ITO is due to the wide availability of experimental studies on its nonlinear optical response under various scenarios and the string of recent interest on this material as an ENZ medium. Indeed, our model can be applied to any highly doped semiconductor with an ultrafast thermalization time, as long as the model parameters are known.

Despite the many experimental studies in recent years,^[2,20,21,32,35,37,38] due to the sensitivity of ITO's optical properties to manufacturing specifications, we focus first on reproducing the results of the article by Alam, De Leon, and Boyd,^[20] in which the intensity-dependent complex refractive index of an ITO thin film in the ENZ spectral region is investigated using several tests. First, the authors measure the effect of the pump central wavelength and incident angle on the nonlinear refractive index of ITO. Second, they show how the reflection, transmission, and absorption shift as a function of the pump intensity. Third, they measure the time-domain dynamics of the nonlinearity via pump-probe measurements of the transmission. We use our model to reproduce the first two of these tests in this section and tackle the third (time-domain dynamics) in the next section. Through these comparisons, we are able to further refine our model to most accurately match this particular set of experimental data. We also perform an optimization procedure

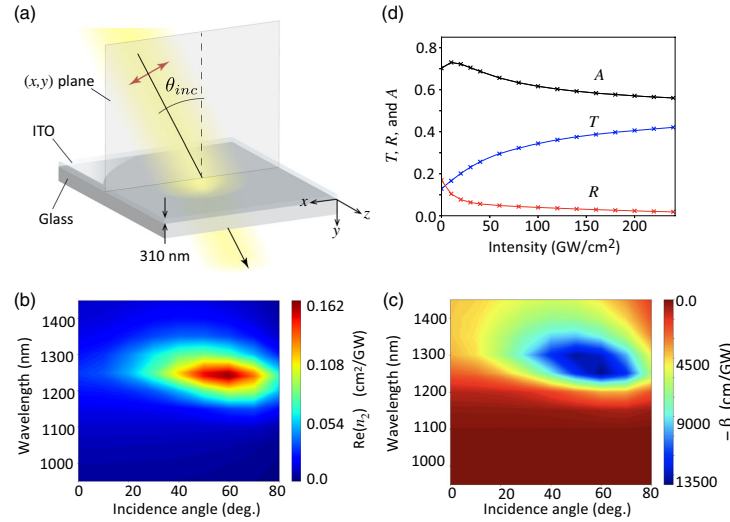


Figure 3. a) Simulations of a 310 nm-thick ITO film illuminated by an intense *p*-polarized light pulse with varying angle of incidence θ_{inc} ; the polarization direction is indicated by the red arrow. b,c) the real part of the nonlinear refractive index (n_2) and the nonlinear absorption constant (β) as a function of θ_{inc} and the central wavelength of the incident light pulse, respectively. d) Intensity-dependent transmission (T), reflection (R), and absorptance (A) obtained for $\theta_{inc} = 30^\circ$ and a central wavelength of 1240 nm. For all cases, the pulse duration is 150 fs.

to determine optimal pulse parameters for maximizing the nonlinearity.

We consider the structure shown in **Figure 3a**. It consists of a 310 nm-thick ITO layer on a glass substrate, bounded by air on top, as in Ref. [20]. The structure is illuminated from the air side with *p*-polarized light incident at angle θ_{inc} . To model the 1D thin-film geometry (along *y*), we apply Bloch periodic boundary conditions to the *x* and *z* directions of our 3D simulation domain enabling us to use oblique-incident pulses.^[48,49] We use the room-temperature Drude model parameters for ITO reported in Ref. [20], given by $\omega_p = 2\pi \times 473$ THz, $\gamma = 0.0468 \omega_p$, and $\epsilon_\infty = 3.8055$, resulting in a zero-crossing wavelength of $\lambda_{ZC} = 1237$ nm.

Ref. [20] quantifies the nonlinear response by assuming it is Kerr-like with an intensity-dependent refractive index $n = n_0 + n_2 I$. The real part of the effective nonlinear index, $\text{Re}(n_2)$, and the effective nonlinear attenuation constant, $\beta = (4\pi/\lambda) \text{Im}(n_2)$, are extracted via *z*-scan measurements. It was observed that the maximum $\text{Re}(n_2)$ and minimum β are found when the pump wavelength is centered at 1240 nm and the angle of incidence is 60° with respect to the ITO surface normal.

Treating the thermal nonlinearity described in Section 2 as Kerr-like is a rough approximation for two primary reasons: 1) there is no reason to assume that higher odd-order terms do not contribute to the nonlinearity and 2) the nonlinearity is not instantaneous. Although our numerical implementation allows us to calculate the nonlinear polarization (or current density) field using more formal definitions, to compare with the

experimental results, we calculate an effective n_2 from the simulation data. Though there are several ways to go about this, we use the following computationally efficient method. We calculate the mean conduction electron temperature ($\bar{T}_e$) within the ITO over space and time. This is used to obtain an effective plasma frequency and thus, by applying Equation (1), an effective, temperature-dependent, refractive index $n(\bar{T}_e) = \text{Re}(\sqrt{\epsilon})$, which differs from the room-temperature refractive index n_0 by $\Delta n = n(\bar{T}_e) - n_0$. We then extract an effective n_2 using

$$n_2 = \frac{\Delta n}{\bar{I}} \quad (9)$$

where $\bar{I}$ is the time-averaged pulse intensity. The true refractive index $n(\mathbf{r}, t)$ will vary in space and time during pulse irradiation and therefore $\Delta n(\mathbf{r}, t) = n(\mathbf{r}, t) - n_0$ also varies in space and time. While the time and space dependence of n_2 is not measured in Ref. [20], we do explicitly calculate $\Delta n(\mathbf{r}, t)$ in Section 6 and Section S4, Supporting Information.

We carry out a 2D parameter sweep of pump incident angle (in 10° increments) and center wavelength (in 50 nm increments) and plot the effective n_2 and β values in Figure 3b,c, respectively. Our results indicate that n_2 is maximized at a wavelength of 1250 nm and an angle of 60° , in agreement with the experimental results. Further, our values of $n_2 = 0.16 \text{ cm}^2/\text{GW}$ and $\beta = -13\,000 \text{ cm/GW}$ are in reasonable quantitative agreement with the values obtained in Ref. [20], which are $n_2 = 0.11 \text{ cm}^2 \text{ GW}^{-1}$ and $\beta = -7500 \text{ cm GW}^{-1}$.

Next, we determine the effect of the incident pulse intensity on the reflectance, transmittance, and absorptance for a central pump wavelength of 1240 nm and an incident angle of 30° as in Ref. [20]. We show the results of our calculations in Figure 3d where we see an increase in the transmittance from $T = 0.13$ to $T = 0.4$ as the intensity increases from 0 to 250 GW cm^{-2} and a corresponding decrease in the reflectance from $R = 0.19$ to $R = 0.02$. Our results agree quantitatively with the experimental measurements,^[20] which show an increase in the transmittance from $T = 0.12$ to $T = 0.4$ and a corresponding decrease in reflectance from $R = 0.2$ to $R = 0.04$ for the same intensity range.

A feature of note in Figure 3d is the saturation of the nonlinearity for high intensities. For simulations in which we assume that the Drude damping coefficient γ is temperature independent, this saturation occurs for intensities larger than 250 GW cm^{-2} . In order to produce Figure 3d, and reach agreement with experiments, we needed to consider a model wherein the damping coefficient increases with temperature. As a first-order approximation, we use

$$\gamma(T_e) = \gamma_0 \left(1 + \frac{T_e}{T_0} \right) \quad (10)$$

where γ_0 is the low-intensity Drude damping coefficient and $T_0 = 15\,000 \text{ K}$ is a fitting parameter with units of temperature. The increase in the Drude damping coefficient is reasonable because we would indeed expect a decrease in electron mobility due to an increase in electron-phonon, electron-impurity, and grain boundary collisions. This effect was also found experimentally in Ref. [35], where an almost-linear relationship was found between an effective γ and the pulse intensity (and thus, indirectly temperature), by fitting the measured spectra to the Drude model. The fact that γ depends on the temperature highlights the benefit of time-domain modeling as we can extract simple, yet effective models for optical parameters that otherwise would require semiclassical or quantum mechanical calculations.

Finally, we optimize properties of the pump pulse (center wavelength, incident angle, and linear chirp for a given intensity)

in order to maximize the change in the refractive index $\Delta n(\mathbf{r}, t)$. Using a Nelder Mead algorithm, we find that for lower intensities, Δn is largest when the center wavelength is tuned to the ENZ wavelength. Interestingly, for higher intensities, a larger center wavelength is required to maximize Δn . For all intensities, the optimal incident angle is $\theta_{\text{inc}} = 30^\circ$ and the linear chirp has negligible effect. These results are presented and further explained in Section S4, Supporting Information.

4. Analysis of Time-Domain Dynamics

In the previous section, we used spectral results to validate our model. However, part of the interest in ITO's strong nonlinearity stems from its ultrafast response time, limited mainly by the pulse width (since the athermal relaxation time is small), allowing precise control of the instantaneous refractive index. Thus, we are particularly interested in modeling the time-varying optical response of ITO.

Degenerate pump-probe measurements in ITO thin films have shown that the rise time is less than 200 fs for a 150 fs laser pulse and the fall time was $\approx 360 \text{ fs}$.^[20] Here we present corresponding pump-probe simulations using a two-step numerical process. A primary FDTD-TTM simulation computes the temperature fields as a function of space and time as the material is excited by a pump pulse. These temperature fields are stored and subsequently used in separate probe simulations, where it is assumed that the probe intensity is low enough that it has no effect on the time- and space-varying temperature field.

We plot the transmission at $\lambda = 1240 \text{ nm}$ as a function of probe pulse delay in Figure 4, which shows a rise time of 170 fs and a fall time of 385 fs. As in Ref. [20] these are defined as the time it takes for the transmittance to rise (fall) from 10% (90%) to 90% (10%) of its maximum change. In order to obtain Figure 4, we needed to adapt a model for the electron-phonon coupling strength g_{ep} that was originally derived for metals (Equation S34, Supporting Information) and contains a free parameter. Since the fall time of the optical response is controlled by g_{ep} , we compared numerical calculations to the experimental

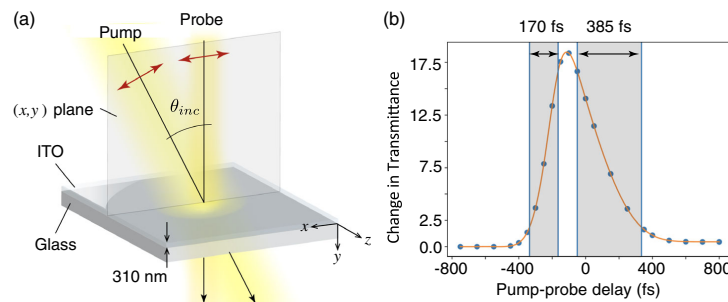


Figure 4. Degenerate pump-probe simulations. a) The simulated structure is the same as that in Figure 3a. It is illuminated with an intense *p*-polarized light pulse (pump) at incidence angle $\theta_{\text{inc}} = 30^\circ$; a weak normally-incident light pulse with the same wavelength (probe) is used to detect changes in transmittance. Both pump and probe pulses have a duration of 150 fs and central wavelength of 1240 nm. The polarization direction of the beams is indicated by the red arrows. b) Transmittance of the probe pulse as a function of pump-probe delay obtained from the FDTD-TMM model.

results to obtain the free parameter; this is presented in Section S3, Supporting Information.

In Figure 5a, we show the corresponding effective refractive index of the ITO film as a function of time, which is extracted using an effective medium theory where we consider the thin film to be an effective (uniform) medium.^[50] We see that the real part of the refractive index increases by ≈ 0.5 and the imaginary part decreases by ≈ 0.2 . Such a dramatic refractive index change suggests a transition from metallic ($\text{Re}[\epsilon] < 0$) to dielectric

($\text{Re}[\epsilon] > 0$) behavior in the epsilon-near-zero region, which also explains the large increase in transmittance in Figure 4.

We illustrate this further in Figure 5b,c, which shows snapshots of the electric field amplitude during the simulation of a 150 fs pump-pulse incident on an ITO film on glass. We see in Figure 5b that, early in the simulation, a large amount of the incident light is reflected, as evidenced by the interference patterns, indicating that ITO behaves like a metal. Figure 5c shows the electric field amplitude later in the same pulse where the ITO layer is hotter and there is much less reflection (no discernible interference pattern), indicating that ITO behaves as a dielectric.

Our results show that the coupled FDTD-TTM model for ITO agrees remarkably well with experiments, indicating the strength of the FDTD implementation. Because our modeling is in the time domain, we can track the nonlinear dynamics as a function of both space and time. In the Supporting Information, we present 10 sets of movies showing the evolution of the pump electric field amplitude for different incident angles, center wavelengths, and intensities. Also included are the corresponding videos of the electronic temperature field evolution. These are briefly discussed in Section S5, Supporting Information.

Here we highlight one example set of movies. Movie 1 shows the electric field amplitude evolution of a 250 GW cm^{-2} intensity pump, centered at 1240 nm, incident at 30° , and Movie 1-T shows the corresponding electronic temperature field evolution. In Movie 1, we see that at low temperatures (before the pump peak), the ITO film exhibits metallic properties as it reflects much of the incident pulse and absorbs much of the transmitted light, resulting in an evanescent wave in the ITO layer. Indeed, the angle of transmission in the ITO is small due to the low refractive index. As the pulse grows in intensity (and T_e increases), we see the skin depth of the ITO increase ($\text{Im}(\Delta n) < 0$), the angle of transmission in the ITO increase ($\text{Re}(\Delta n) > 0$), and a corresponding increase in transmission into the glass substrate. At this point ITO behaves like a dielectric. Finally, near the end of the pulse, we see the interference pattern in vacuum diminish significantly due to decreased reflectance.

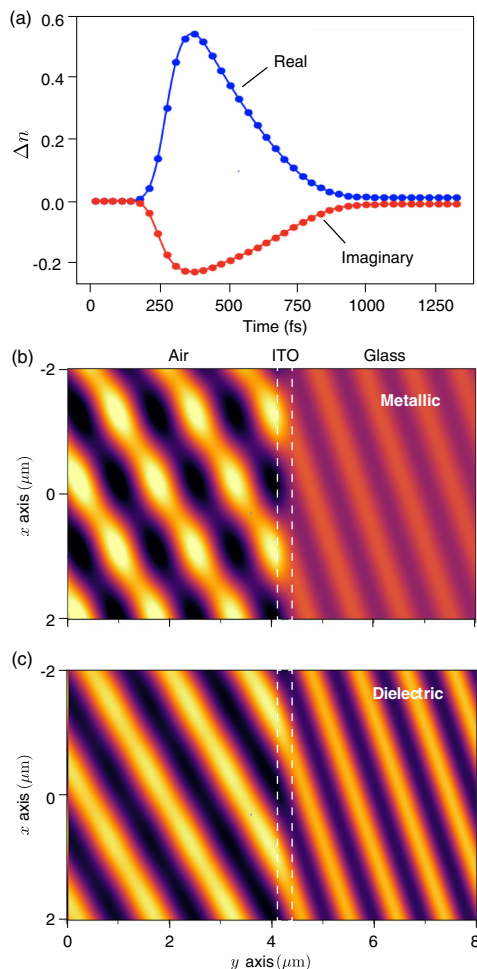


Figure 5. a) Change in the effective refractive index of a 310 nm ITO film for the same simulation as Figure 4. b) Snapshot of electric field amplitude early in the pulse when the ITO film is metallic and c) later in the pulse when the ITO has become dielectric. The incident pulse propagates from vacuum (left) through the ITO film into the glass substrate (right).

5. Time Refraction in ITO Thin Films

In the previous sections, we refined and validated our model using experimental measurements published in Ref. [20]. We demonstrated how intense optical pulses can actively change the refractive index of ITO. However, the reverse is also true. A recent publication cleverly investigated the elusive time-domain response of ITO by harnessing a little-known optical effect known as adiabatic frequency translation or more colloquially, time refraction.^[2]

Adiabatic frequency translation occurs as the result of a wave traveling through a medium with time-varying permittivity.^[3,40] The center frequency of an optical pulse leaving the time-varying medium is shifted from that of the incident pulse. Its colloquial title is inspired by space refraction where the wavevector changes due to spatial variations of the permittivity. Here, the frequency (i.e., photon energy) changes due to temporal variations in the permittivity.

It was recently demonstrated experimentally that ITO produces measurable frequency translation when pumped with a high-intensity pulse.^[2] Degenerate pump-probe experiments,

with a pulse width of 120 fs, were used to demonstrate time refraction in a 620 nm film of ITO bounded on either side by glass. By changing the pump-probe delay, they demonstrated a time-dependent shift in the probe's center frequency. Using our self-consistent modeling approach, we observe the experimental adiabatic frequency shifting in transmission of Ref. [2]. Moreover, we predict a stronger effect in reflection, where we find that, for certain pump-probe delays, two distinct reflected pulses are created.

We show in Figure 6b,c the probe pulse transmission and reflection spectra, respectively, as a function of pump-probe delay for three different pump intensities, where we find that the reflection spectra show a much more pronounced frequency translation. For all pump simulations, we used the same excitation conditions reported in Ref. [2], that is, the pump field is

p-polarized with an incidence angle of 15° . Normal incidence was used for the broadband, probe simulations.

We found that the probe pulse must be down chirped to show pronounced frequency translation in our simulations. An unchirped 120 fs pulse was found to contain too few frequency components to show appreciable translation. Requiring chirp is not explicitly mentioned in the main article of Ref. [2]. However, in the accompanying Supplementary Material, the frequency-resolved optical gating (FROG) measurement of their pulse is plotted and is shown to be greatly down chirped. In our numerical experiments, we find that a chirp parameter (as defined in Ref. [51]) of $a = -1.5$ (where the negative indicates a down chirp) best replicates the experimental data. We believe this insight presents the chirp as a new degree of freedom to explore in future works.

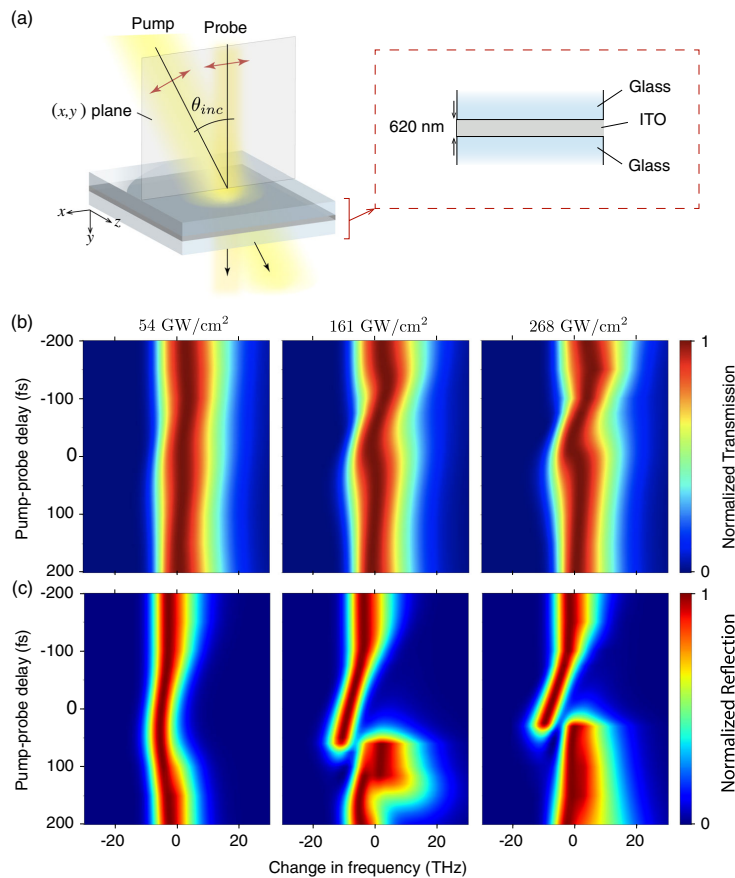


Figure 6. Frequency translation results obtained through pump-probe simulations as a function of frequency shift (*x*-axis) and pump-probe delay (*y*-axis) for three pump intensities as indicated. a) The material structure consists of a 620 nm film of ITO bounded on both sides by glass. The pump field is *p*-polarized with an incidence angle of $\theta_{inc} = 15^\circ$. b) Transmission and c) reflection spectra for varying pump-probe delay.

As the reflected light was not measured in the experiments, our calculation that shows enhanced frequency translation in reflection over transmitted light represents a novel prediction. We see that the reflection spectrum is composed of two distinct peaks for some values of pump-probe delay. To explore this further, we plot in **Figure 7** the time domain reflected (blue) and transmitted (red) electric fields for a pump-probe delay of 25 fs and an incident pump intensity of 268 GW cm^{-2} (right panel of Figure 6c). Such pump intensity has been used in experiments without material breakdown.^[20,23] We observe that the

reflected field is composed of two distinct envelopes in time, unlike the transmitted field.

Taking a windowed Fourier transform of the reflected time-domain field centered at different times, we obtain an effective spectrum as a function of time, and we plot its evolution as a contour plot in Figure 7b. We see that the spectral content of each distinct time-domain lobe is different. In Figure 7c, we plot the spectra at the times corresponding to the peaks of the two lobes (indicated by the vertical dotted lines in Figure 7b); we see that the two spectra are indeed shifted.

The shift itself is not surprising, since we used a chirped pulse in the simulation as described earlier. Indeed, when we repeated this simulation for the low-intensity limit where a constant temperature and thus constant plasma frequency is maintained (not shown), we also see a shift but in a single lobe. As the peak instantaneous wavelength shifts from 1300 to 1200 nm, the real part of the relative permittivity at the peak instantaneous wavelength increases from ≈ -0.2 to $+0.4$ as it crosses through the ENZ region; this is seen in Figure 1 for the case of low temperature.

The two-lobe reflected pulse only occurs for high intensity, as a direct consequence of the nonlinearity. As the temperature increases, the plasma frequency decreases, causing a corresponding increase of the relative permittivity up to a value of 2 or more. During the intense pulse irradiation, the effective index of ITO briefly comes to match that of the surrounding medium (SiO_2), leading to a null in reflection that does not occur in transmission.

In sum, the ability of this model to quantitatively replicate experimental results in the time and frequency domain gives us confidence in its accuracy and the novel phenomena it predicts. This allows us to explore even more complex scenarios, to which we now turn.

6. Time-Varying Plasmonic Metasurface on ITO

In the previous sections, we refined and validated our model using experimental measurements published in Refs. [2,20] using both spectral and time-domain results. It was found that the pump's wavelength and angle of incidence have a strong effect on the magnitude of the refractive index change. Intuitively, one might suppose that the strength of the nonlinear optical response can be further increased by enhancing the near fields within ITO films by including nanostructures. Plasmonic nanoparticles have been used to enhance optical processes of all sorts, including, nonlinear optical effects.^[52] As our model is built into FDTD, we are not restricted to simulations of the simple planar geometries presented above, but we can simulate the refractive index shift in ITO in the presence of nanostructuring, including nearby plasmonic particles, where the 3D spatial structure and dispersion of the plasmonic material are taken into account self consistently.

As an illustrative example, we consider the plasmonic metasurface depicted in **Figure 8a**. It consists of a squared array of gold rectangular nanoparticles on a 310 nm-thick layer of ITO on glass. The array is characterized by a $600 \text{ nm} \times 600 \text{ nm}$ unit cell. The rectangular nanoparticle has a length of 500 nm, width of 300 nm, and thickness of 40 nm. This geometry was chosen because it has a very large absorption peak of $A = 0.9$ within the

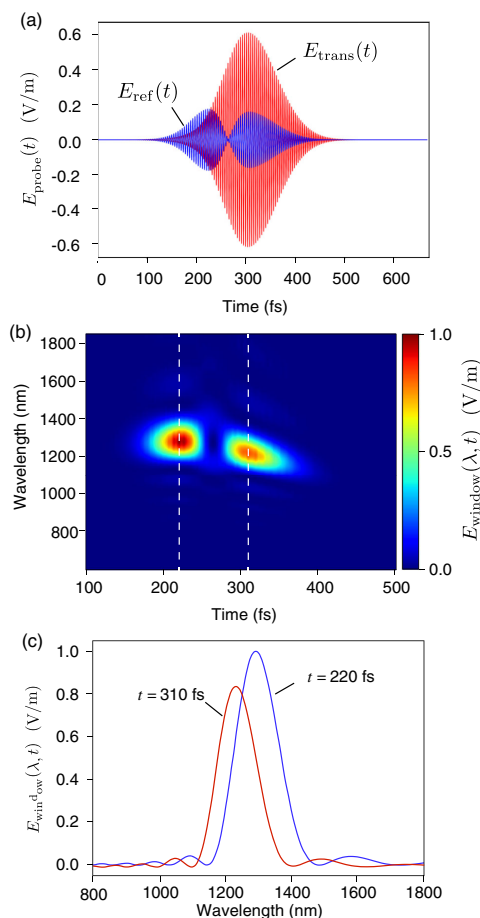


Figure 7. a) Time-domain reflected (blue) and transmitted (red) electric field from the simulations of the right panels of Figure 6b,c, with a pump-probe delay of 25 fs and an incident pulse intensity of 268 GW cm^{-2} . b) Spectral evolution over time corresponding to the reflected fields of (a), calculated via a windowed Fourier transform of width 33 fs; the time axis in the plot represents the center of each time window. c) Spectra at the two times indicated by the vertical dotted lines in plot (b).

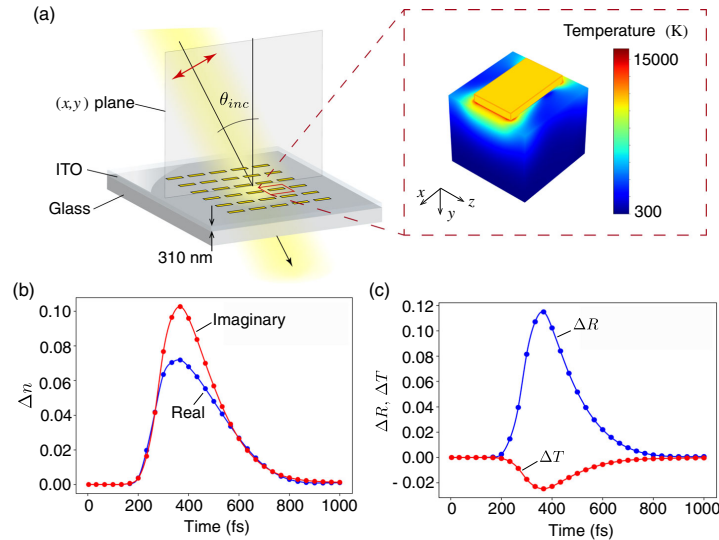


Figure 8. Intensity-dependent refractive index and intensity-dependent transmission/reflection of a plasmonic metasurface on ITO. a) The simulated structure is an array of gold rectangular nanoparticles on top of an ITO/glass system with the same properties as in Figure 3a. The structure is illuminated with a *p*-polarized optical field incident at an angle of $\theta_{inc} = 30^\circ$. The inset shows a 3D time-domain snapshot of the temperature field in the ITO film at the peak of the laser pulse, within a unit cell of the plasmonic particle array. b,c) The change in refractive index and the change in reflectance and transmittance, obtained using the FDTD-TMM model, respectively.

wavelength range of interest; this peak occurs at $\lambda = 1350$ nm. We use Bloch boundary conditions in our FDTD simulations, where a plane wave that is linearly polarized along the length of the particle is incident on an infinitely periodic metasurface at an angle of 30° . As in Section S4, Supporting Information, we use a 2 GW cm^{-2} pump pulse for ease of comparison to our results there. We save 3D temperature field snapshots within the ITO at various times during the pump pulse simulation—the one at the pump pulse peak is shown in the inset of Figure 8a. For each temperature field snapshot, we perform separate probe simulations. From the resulting probe reflection and transmission coefficient spectra, we use effective medium theory^[50] to extract the effective refractive index of the metasurface as a function of wavelength and pump pulse simulation time.

We plot in Figure 8b the effective index change of the metasurface as a function of pump simulation time at $\lambda = 1350$ nm. We find that the maximum value of Δn is five times larger than for the bare ITO thin film (Section S4, Figure S4b, Supporting Information) irradiated at the same incident pump intensity. In Figure 8c we plot the corresponding change in reflectance and transmittance, ΔR and ΔT , respectively. We find that the transmittance decreases during the pulse irradiation ($\Delta T < 0$), in contrast to the increase that was demonstrated for bare ITO thin films ($\Delta T > 0$) previously (Figure 4 and S4c, Supporting Information). In addition to this change in sign, the magnitude of ΔT is also two times larger than in Figure S4c, Supporting Information. The large enhancement in ΔT here is due to the plasmonic particle that creates

electromagnetic hotspots when excited near the localized surface plasmon resonance. As their names imply, these hotspots are regions of enhanced Joule heating, leading to a stronger nonlinearity being induced in the ITO film (see Figure 8a). By comparing Figure 8c and S4c, Supporting Information, we also observe a longer transient nonlinear response when the plasmonic particle is present. This occurs because the decay time of the plasmon resonance is longer than the decay time of the ITO nonlinear response.

Effective medium theory does not only output an effective refractive index $n = \sqrt{\epsilon\mu}$ but also an effective admittance $Y = \sqrt{\epsilon/\mu}$. Typically, for sufficiently small feature sizes, we expect these values to be equal as long as the material is nonmagnetic ($\mu = 1$). We have verified that this indeed remains true for bare ITO films considered in the previous sections. However, this is not the case for the metasurface investigated in this section, where applying effective medium theory suggests an effective magnetic response of the metasurface. In fact, we can see this effect in Figure 8 where we see an increased imaginary component of the effective index (Figure 8b) coupled with a decrease in the absorptance (Figure 8c); this can only occur if $\mu \neq 1$. This effective magnetic behavior is due to complex response of the plasmonic particle, and will fundamentally change the ENZ behavior, because we must now take μ into account. This will be investigated in future work.

Previous experimental work has investigated ultrafast irradiation of ITO-based time-varying plasmonic metasurfaces.^[21] There, it was also found that the transmittance decreases during

pulse irradiation for the metasurface but increases (with a smaller line-width) for a bare ITO film. This further validates the results of this (and the previous) section. The magnitudes of ΔT and Δn here are not directly comparable to those of Ref. [21], as their ITO sample was different (with different film thicknesses and optical properties), and they applied a different pump intensity. However, the enhancement of Δn when a plasmonic particle is present versus a bare ITO film is comparable.

We also present movies showing the time evolution of the electric field amplitude and temperature field of the plasmonic metasurface on ITO while being irradiated by an oblique incidence pump pulse. In Movie 11, we see the incident pulse exciting the plasmonic particle and creating electromagnetic hotspots at the particle edges. These result in electronic temperature hotspots, as shown in Movie 11-T and in the inset of Figure 8a. As the pump pulse peaks, we start to see strong interference between the incident and reflected pulse in the vacuum region due to the increased reflectance (as also indicated by the ΔR plot in Figure 8c). There is very little decrease in the transmittance to balance the reflectance (ΔT plot in Figure 8c). That, in Movie 11, is evident by the increased penetration of the field into the ITO film, suggesting an increased skin depth during intense field irradiation.

7. Conclusion

We introduced a self-consistent, multiphysics, spatiotemporal numerical model for the intensity-dependent, time-varying refractive index of TCO-based ENZ materials. The nonlinear optical response of TCOs is described by Drude-like permittivity where the plasma frequency is dependent upon the conduction band electron temperature. A TTM describing the electron temperature distribution in time and space as induced by irradiation with an ultrafast intense laser pulse was implemented directly into a 3D FDTD electrodynamics solver. Our method captures the full nonlinear spatiotemporal optical response of TCO-based ENZ materials, including those within complex geometries containing, for example, plasmonic nanostructures in close proximity. We validated our implementation via comparisons with experimental results for a bare ITO film, where we obtained quantitative agreement, as well as an ITO film decorated with plasmonic nanoantennas. We also used our model to demonstrate time refraction wherein the time-varying permittivity of the ITO film results in a shifting of the central frequency of an optical pulse. We show that the reflected light is predicted to have a larger and more noticeable frequency change than the transmitted light. The ability to model and understand the complex nonlinear response of TCOs will be instrumental for designing sophisticated nonlinear nanophotonic devices with femtosecond-scale response times, such as ultrafast reconfigurable metasurfaces for active switching, modulation or beam manipulation.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors would like to thank Compute Canada and Scinet for computational resources. J.B. and L.R. acknowledge the Natural Sciences and Engineering Research Council (NSERC) Vanier Canada Graduate Scholarships program and the Canada Research Chairs program for financial support. I.D.L., A.P.C., and L.C.H. acknowledge financial support from CONACyT (Ciencia Básica) grant no. 286150. I.D.L. acknowledges the support of the Federico Baur Endowed Chair in Nanotechnology. A.C.L. acknowledges the Bundesministerium für Bildung und Forschung (German Federal Ministry of Education and Research) under the Tenure-Track Program and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC 2122, Project ID390833453). J.B. thanks Dr. M. Zahirul Alam for his useful input. The authors also thank the Max Planck Institute for the Science of Light for enabling this collaborative effort, through funding the travel of photonics researchers and their annual meeting.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

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

Keywords

active nanophotonics, dynamic nanophotonics, epsilon near zero, finite-difference time-domain, transparent conductive oxides

Received: November 9, 2022

Published online:

- [1] R. W. Boyd, in *Nonlinear Optics*, 4 ed., Academic Press is an imprint of Elsevier, San Diego 2019.
- [2] Y. Zhou, M. Z. Alam, M. Karimi, J. Upham, O. Reshef, C. Liu, A. E. Willner, R. W. Boyd, *Nat. Commun.* **2020**, 11, 2180.
- [3] N. Karl, P. P. Vabishchevich, M. R. Shcherbakov, S. Liu, M. B. Sinclair, G. Shvets, I. Brener, *Nano Lett.* **2020**, 20, 7052.
- [4] D. Cotter, R. J. Manning, K. J. Blow, A. D. Ellis, A. E. Kelly, D. Nasset, I. D. Phillips, A. J. Poustie, D. C. Rogers, *Science* **1999**, 286, 1523.
- [5] Q. Guo, Y. Cui, Y. Yao, Y. Ye, Y. Yang, X. Liu, S. Zhang, X. Liu, J. Qiu, H. Hosono, *Adv. Mater.* **2017**, 29, 1700754.
- [6] X. Jiang, H. Lu, Q. Li, H. Zhou, S. Zhang, H. Zhang, *Nanophotonics* **2018**, 7, 1835.
- [7] T. Daghooghi, M. Soroosh, K. Ansari-Asl, *Appl. Opt.* **2018**, 57, 2250.
- [8] A. M. Shaltout, V. M. Shalae, M. L. Brongersma, *Science* **2019**, 364, eaat3100.
- [9] G. P. Zograf, M. I. Petrov, S. V. Makarov, Y. S. Kivshar, *Adv. Opt. Photonics* **2021**, 13, 643.
- [10] E. A. A. Pogna, M. Celebrano, A. Mazzanti, L. Ghirardini, L. Carletti, G. Marino, A. Schirato, D. Viola, P. Laporta, C. De Angelis, G. Leo, G. Cerullo, M. Finazzi, G. Della Valle, *ACS Nano* **2021**, 15, 11150.
- [11] M. Celebrano, D. Rocco, M. Gandolfi, A. Zilli, F. Rusconi, A. Tognazzi, A. Mazzanti, L. Ghirardini, E. A. A. Pogna, L. Carletti, C. Baratto, G. Marino, C. Gigli, P. Biagioni, L. Duò, G. Cerullo, G. Leo, G. D. Valle, M. Finazzi, C. D. Angelis, *Opt. Lett.* **2021**, 46, 2453.
- [12] J. Leuthold, C. Koos, W. Freude, *Nat. Photonics* **2010**, 4, 535.

- [13] A. V. Z. M. Kauranen, *Nat. Photonics* **2012**, 6, 5.
- [14] G. Li, S. Zhang, T. Zentgraf, *Nat. Rev. Mater.* **2017**, 2.
- [15] M. Conforti, G. Della Valle, *Phys. Rev. B* **2012**, 85, 245423.
- [16] V. G. Naik, V. M. Shalaev, A. Boltasseva, *Adv. Mater.* **2013**, 25, 3264.
- [17] O. Reshef, I. De Leon, M. Z. Alam, R. W. Boyd, *Nat. Rev. Mater.* **2019**, 4, 535.
- [18] N. Kinsey, J. Khurgin, *Opt. Mater. Express* **2019**, 9, 2793.
- [19] N. Kinsey, C. DeVault, A. Boltasseva, V. M. Shalaev, *Nat. Rev. Mater.* **2019**, 4, 742.
- [20] M. Z. Alam, I. De Leon, R. W. Boyd, *Science* **2016**, 352, 795.
- [21] M. Z. Alam, S. A. Schulz, J. Upham, I. De Leon, R. W. Boyd, *Nat. Photonics* **2018**, 12, 79.
- [22] Y. U. Lee, E. Garoni, H. Kita, K. Kamada, B. H. Woo, Y. C. Jun, S. M. Chae, H. J. Kim, K. J. Lee, S. Yoon, E. Choi, F. Mathevet, I. Ozerov, J. C. Ribierre, J. W. Wu, A. D'Aléo, *Adv. Opt. Mater.* **2018**, 6, 1701400.
- [23] L. Caspani, R. P. Kaipurath, M. Clerici, M. Ferrera, T. Roger, J. Kim, N. Kinsey, M. Pietrzyk, A. Di Falco, V. M. Shalaev, A. Boltasseva, D. Faccio, *Phys. Rev. Lett.* **2016**, 116, 233901.
- [24] E. G. Carnemolla, L. Caspani, C. DeVault, M. Clerici, S. Vezzoli, V. Bruno, V. M. Shalaev, D. Faccio, A. Boltasseva, M. Ferrera, *Opt. Mater. Express* **2018**, 8, 3392.
- [25] N. Kinsey, C. DeVault, J. Kim, M. Ferrera, V. M. Shalaev, A. Boltasseva, *Optica* **2015**, 2, 616.
- [26] T. S. Luk, D. de Ceglia, S. Liu, G. A. Keeler, R. P. Prasankumar, M. A. Vincenti, M. Scalora, M. B. Sinclair, S. Campione, *Appl. Phys. Lett.* **2015**, 106, 151103.
- [27] A. Capretti, Y. Wang, N. Engheta, L. D. Negro, *Opt. Lett.* **2015**, 40, 1500.
- [28] A. Capretti, Y. Wang, N. Engheta, L. Dal Negro, *ACS Photonics* **2015**, 2, 1584.
- [29] M. A. Vincenti, M. Kamandi, D. de Ceglia, C. Guclu, M. Scalora, F. Capolino, *Phys. Rev. B* **2017**, 96, 45438.
- [30] M. A. Noginov, L. Gu, J. Livenere, G. Zhu, A. K. Pradhan, R. Mundle, M. Bahoura, Y. A. Barnakov, V. A. Podolskiy, *Appl. Phys. Lett.* **2011**, 99, 12.
- [31] V. E. Babicheva, A. Boltasseva, V. A. Lavrinenko, *Nanophotonics* **2015**, 4, 165.
- [32] X. Liu, J. Park, J.-H. Kang, H. Yuan, Y. Cui, H. Y. Hwang, M. L. Brongersma, *Appl. Phys. Lett.* **2014**, 105, 181117.
- [33] P. Guo, R. D. Schaller, J. B. Ketterson, R. P. H. Chang, *Nat. Photonics* **2016**, 10, 267.
- [34] P. Guo, R. D. Schaller, L. E. Ocola, B. T. Diroll, J. B. Ketterson, R. P. Chang, *Nat. Commun.* **2016**, 7.
- [35] H. Wang, K. Du, C. Jiang, Z. Yang, L. Ren, W. Zhang, S. J. Chua, T. Mei, *Phys. Rev. Appl.* **2019**, 11, 064062.
- [36] T. Mei, H. Wang, K. Du, X. Dai, W. Zhang, R. Liu, S. J. Chua, *Nanophotonics* **2020**, 9, 4287.
- [37] R. Secondo, J. Khurgin, N. Kinsey, *Opt. Mater. Express* **2020**, 10, 1545.
- [38] L. Rodríguez-Suné, M. Scalora, A. S. Johnson, C. Cojocar, N. Akozbek, Z. J. Coppens, D. Perez-Salinas, S. Wall, J. Trull, *APL Photonics* **2020**, 5, 010801.
- [39] M. Scalora, J. Trull, D. de Ceglia, M. A. Vincenti, N. Akozbek, Z. Coppens, L. Rodríguez-Suné, C. Cojocar, *Phys. Rev. A* **2020**, 101, 053828.
- [40] Y. Xiao, G. P. Agrawal, D. N. Maywar, *Opt. Lett.* **2011**, 36, 505.
- [41] T. Pisarkiewicz, A. Kolodziej, *Phys. Status Solidi B* **1990**, 158, K5.
- [42] C. Voisin, N. Del Fatti, D. Christofilos, F. Vallée, *J. Phys. Chem. B* **2001**, 105, 2264.
- [43] E. Carpene, *Phys. Rev. B* **2006**, 74, 024301.
- [44] M. S. Bin-Alam, J. Baxter, K. M. Awan, A. Kiviniemi, Y. Mamchur, A. C. Lesina, K. L. Tsakmakidis, M. J. Huttunen, L. Ramunno, K. Dolgaleva, *Nano Lett.* **2021**, 21, 51.
- [45] J. Baxter, A. C. Lesina, L. Ramunno, *IEEE Trans. Antennas Propag.* **2021**, 69, 3982.
- [46] A. C. Lesina, A. Vaccari, P. Berini, L. Ramunno, *Opt. Express* **2015**, 23, 10481.
- [47] P. L. DeVries, J. E. Hasbun, P. L. DeVries, *A First Course in Computational Physics*, 2nd ed., Jones and Bartlett Publishers, Sudbury, Mass, **2011**.
- [48] F. Yang, J. Chen, R. Qiang, A. Elsherbeni, *Radio Sci.* **2007**, 42, 4.
- [49] A. Aminian, Y. Rahmat-Samii, *IEEE Trans. Antennas Propag.* **2006**, 54, 1818.
- [50] D. R. Smith, S. Schultz, P. Markoš, C. M. Soukoulis, *Phys. Rev. B* **2002**, 65, 195104.
- [51] B. E. A. Saleh, M. C. Teich, *Fundamentals of Photonics*, 2nd ed., Wiley Series in Pure and Applied Optics, Wiley Interscience, Hoboken, NJ **2007**.
- [52] S. V. Gaponenko, D. V. Guzatov, *Proc. IEEE* **2020**, 108, 704.

Supplemental Information

Section 1: Temperature dependence of plasma frequency

In this section, we derive a model for the temperature dependence of the plasma frequency. We begin with the Boltzmann transport equation, which describes the time evolution of the conduction band non-equilibrium distribution function $f(\mathbf{r}, \mathbf{k}, t)$, where $\mathbf{r}$ and $\mathbf{k}$ are the three-dimensional position vector and wave-vector respectively, and where $f(\mathbf{r}, \mathbf{k}, t)d\mathbf{r}d\mathbf{k}/4\pi^3$ is the number of conduction electrons in phase space from $\mathbf{r}$ to $\mathbf{r}+d\mathbf{r}$ and $\mathbf{k}$ to $\mathbf{k}+d\mathbf{k}$. As we will see, our final result is equivalent to that given by Ref. [1].

In equilibrium, $f(\mathbf{r}, \mathbf{k}, t)$ is given by the Fermi-Dirac distribution f_{FD} and thus deviations from equilibrium are quantified by the athermal distribution

$$\delta f(\mathbf{r}, \mathbf{k}, t) = f(\mathbf{r}, \mathbf{k}, t) - f_{FD}, \quad (1)$$

whose dynamics in an external force $\mathbf{F}$ are described by the semi-classical Boltzmann transport equation [2],

$$\frac{\partial \delta f}{\partial t} + \mathbf{v}_{\mathbf{k}} \cdot \frac{\partial \delta f}{\partial \mathbf{x}} + \mathbf{F} \cdot \frac{1}{\hbar} \frac{\partial f}{\partial \mathbf{k}} = \left(\frac{\partial \delta f}{\partial t} \right)_{\text{collisions}}, \quad (2)$$

where $\mathbf{v}_{\mathbf{k}} = \hbar^{-1} \partial E / \partial \mathbf{k}$ is the group velocity of a Bloch electron in the conduction band (where E is its energy), and where the last term describes the effects of electronic collisions. For the latter, we use the relaxation-time approximation

$$\left(\frac{\partial \delta f}{\partial t} \right)_{\text{collisions}} = -\frac{\delta f}{\tau}, \quad (3)$$

where τ is the inverse scattering rate. To simplify the equation further, we will assume that $\partial f / \partial \mathbf{k} = \partial f_{FD} / \partial \mathbf{k}$ in the third term on the left hand side of Eq. 2.

We take the external force to be $\mathbf{F} = -e\mathbf{E}$, where $\mathbf{E}$ is the local electric field. Since the optical wavelength is much larger than the lattice spacing, we disregard the spatial variation of δf . Further, using, $\frac{\partial f_{FD}}{\partial \mathbf{k}} = \hbar \mathbf{v}_{\mathbf{k}} \frac{\partial f_{FD}}{\partial E}$, and setting $\gamma = 1/\tau$, Eq. 2 becomes

$$\frac{\partial \delta f}{\partial t} - e\mathbf{E} \cdot \mathbf{v}_{\mathbf{k}} \frac{\partial f_{FD}}{\partial E} = -\gamma \delta f, \quad (4)$$

which in the frequency domain gives

$$(-i\omega + \gamma)\delta f = e\mathbf{E} \cdot \mathbf{v}_{\mathbf{k}} \frac{\partial f_{FD}}{\partial E}. \quad (5)$$

We calculate the current density $\mathbf{J}$ by integrating $e\mathbf{v}_{\mathbf{k}}$ over all $\mathbf{k}$ states weighted by $\delta f(\mathbf{k})/(4\pi^3)$ [2]

$$\mathbf{J} = -\frac{e}{4\pi^3} \int d^3k \delta f(\mathbf{k}) \mathbf{v}_{\mathbf{k}} = -\frac{e^2}{4\pi^3} \int d^3k \mathbf{v}_{\mathbf{k}} \left(\mathbf{E} \cdot \mathbf{v}_{\mathbf{k}} \frac{\partial f_{FD}}{\partial E} \right) (-i\omega + \gamma)^{-1}, \quad (6)$$

from which we identify the conductivity tensor

$$\sigma = -\frac{1}{-i\omega + \gamma} \frac{e^2}{4\pi^3} \int d^3k \mathbf{v}_{\mathbf{k}} \mathbf{v}_{\mathbf{k}} \frac{\partial f_{FD}}{\partial E}, \quad (7)$$

where $\mathbf{v}_{\mathbf{k}} \mathbf{v}_{\mathbf{k}}$ is a dyadic tensor defined by $(\mathbf{v}_{\mathbf{k}} \mathbf{v}_{\mathbf{k}})_{ij} = (\mathbf{v}_{\mathbf{k}})_i (\mathbf{v}_{\mathbf{k}})_j$, where i and j refer to Cartesian coordinates.

Under the assumption of an isotropic band structure, $\mathbf{v}_{\mathbf{k}} = \hbar^{-1} \nabla_{\mathbf{k}} E$ is parallel to $\mathbf{k}$ and therefore, in spherical coordinates $\mathbf{v}_{\mathbf{k}} = (\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta) v_k$, where v_k is the magnitude of the velocity vector, thus giving

$$\mathbf{v}_{\mathbf{k}} \mathbf{v}_{\mathbf{k}} = \begin{pmatrix} \sin^2 \theta \cos^2 \phi & \sin^2 \theta \cos \phi \sin \phi & \sin \theta \cos \theta \sin \phi \\ \sin^2 \theta \cos \phi \sin \phi & \sin^2 \theta \sin^2 \phi & \sin \theta \cos \theta \sin \phi \\ \sin \theta \cos \theta \cos \phi & \sin \theta \cos \theta \sin \phi & \cos^2 \theta \end{pmatrix} v_k^2 = \mathbf{M}(\theta, \phi) v_k^2. \quad (8)$$

Rewriting the integral in Eq. 7 in spherical coordinates, we obtain

$$\sigma = -\frac{1}{-i\omega + \gamma} \frac{e^2}{4\pi^3} \int_0^{2\pi} \int_0^\pi \mathbf{M}(\theta, \phi) \sin \theta d\theta d\phi \int dk k^2 v_k^2 \frac{\partial f_{FD}}{\partial E}. \quad (9)$$

By integrating over θ and ϕ , we find that the non-diagonal terms vanish, and the diagonal terms are all equal to $\frac{4}{3}\pi$, thus confirming that the conductivity is isotropic, given by

$$\sigma = -\frac{1}{-i\omega + \gamma} \frac{e^2}{3\pi^2} \int dk k^2 v_k^2 \frac{\partial f_{FD}}{\partial E}. \quad (10)$$

Changing the integration variable from momentum magnitude k to energy E , using Eq. 2 in the main text, we obtain

$$\sigma = -\frac{1}{-i\omega + \gamma} \frac{e^2}{3m_0^* \pi^2} \left(\frac{2m_0^*}{\hbar^2} \right)^{3/2} \int dE (E + CE)^{3/2} (1 + 2EC)^{-1} \frac{\partial f_{FD}}{\partial E}. \quad (11)$$

Comparing Eq. 11 to the form of the Drude model conductivity

$$\sigma = -\frac{\varepsilon_0 \omega_p^2}{-i\omega + \gamma}, \quad (12)$$

we find that the plasma frequency is then, as given by Eq. 3 in the main text,

$$\omega_p^2 = \frac{e^2}{3m_0^* \varepsilon_0 \pi^2} \left(\frac{2m_0^*}{\hbar^2} \right)^{3/2} \int dE (E + CE)^{3/2} (1 + 2EC)^{-1} \frac{\partial f_{FD}}{\partial E}. \quad (13)$$

Eq. 13 is also derived in Ref. [1] using a different method, but with the same conclusion. When $C = 0$, this expression becomes the traditional expression for the plasma frequency,

$$\omega_p^2 = \frac{e^2 n}{\varepsilon_0 m_0^*}. \quad (14)$$

For ITO the non-parabolicity factor was found to be $C = 0.4191$ (eV⁻¹) [3] and the effective mass was found to be $m_0^* = 0.4m_e$ [4], where m_e is the free-electron mass.

Section 2: Validation for Neglecting the Athermal Electron Distribution

In this section, we justify why the athermal electrons do not need to be accounted for explicitly in the two-temperature model for ITO (Eq. 4) that we introduced in the main text. Our theoretical analysis follows Ref. [5], and we find that since the Fermi energy of ITO and the photon energy of our laser pulses are approximately the same, $E_F \approx \hbar\omega_{pulse} \approx 1$ eV, the thermalization time is negligible compared to the pump pulse duration, and therefore we can neglect the athermal electrons in our model. This is not the case for materials with higher Fermi energies, like gold, to which the extended TTM is commonly applied (for example see Refs. [6, 7, 5]).

Recall that the electron distribution function can be written as a sum of the thermal $f_{FD}(T_e, t)$ and athermal $\delta f(E, t)$ components as in Eq. 1. The athermal electrons are the result of the ultrafast pulse perturbing the conduction electrons out of thermal equilibrium. As the pulse irradiates the sample, the electronic distribution deviates further from a thermal distribution. The electrons eventually become thermalized via collisions resulting in a Fermi-Dirac distribution at an increased temperature. The thermalization process is thus described by $\delta f(E, t) \rightarrow 0$ and $f_{FD}(T_{e,0}, t = 0) \rightarrow f_{FD}(T_{e,1} > T_{e,0}, t)$.

We begin by assuming that the lifetime of an excited electron (of energy E) in a degenerate system due to both elastic and inelastic electron-electron collisions is given by Fermi-liquid theory under the random-phase approximation as [5, 8]

$$\tau = \frac{128}{\pi^2 \sqrt{3} \omega_p} \left(\frac{E_F}{E - E_F} \right)^2 \approx \frac{7.5}{\omega_p} \left(\frac{E_F}{E - E_F} \right)^2, \quad (15)$$

where ω_p is the plasma frequency. In the relaxation time approximation, we have

$$\frac{\partial \delta f}{\partial t} \approx \frac{\delta f}{\tau} \approx \delta f \frac{\omega_p}{7.5} \frac{(E - E_F)^2}{E_F^2}, \quad (16)$$

and therefore

$$\delta f(E, t) = \delta f(E, t = 0) \exp \left[-\frac{\omega_p}{7.5} \frac{(E - E_F)^2}{E_F^2} t \right], \quad (17)$$

where $\delta f(E, t = 0)$ is the initial athermal distribution. Here we assume that the perturbation $\delta f(E, t = 0)$ is instantaneous which, physically, is not the case. The perturbation will occur gradually as the pulse propagates through the sample. However, this assumption should not affect the conclusion of the following analysis, as we are only interested in how long it takes for conduction electrons to thermalize (on average).

The initial distribution is described by the simplified model described in the main text (see discussion accompanying Fig. 2), where electrons below the Fermi energy ($E_F - E_{pump} \rightarrow E_F$) are excited above the Fermi energy ($E_F \rightarrow E_F + E_{pump}$) and is thus given by [7]

$$\delta f(E, t = 0) = A(f_{FD}(E - E_{pump})[1 - f_{FD}(E)] - f_{FD}(E)[1 - f_{FD}(E + E_{pump})]), \quad (18)$$

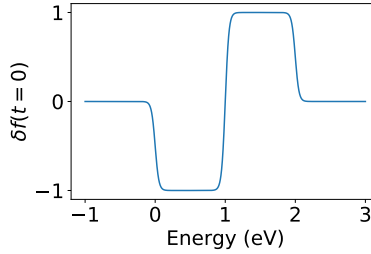


Figure S1: Initial athermal electron distribution as a function of energy (above conduction band minimum) given by Eq. 18.

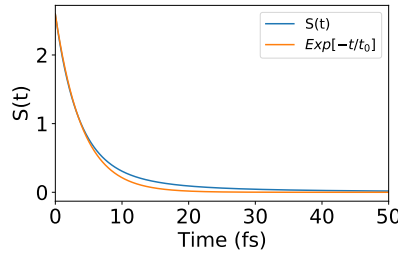


Figure S2: Athermal electron energy density source function $S(t)$ for ITO. The blue line is the numerical integration of Eq. 22 and the yellow line is an exponential fit with $t_0 = 4.2$ fs.

where A is a constant amplitude which depends on the optical fluence of the pump pulse, and where we use $T_{e,0} = 300$ K for the evaluation of $f_{FD}(E)$ which, as we will see, gives an upper bound for the athermal decay time. This function is plotted in Fig. S1 for $E_F = E_{pump} = 1.0$ eV and $A = 1.0$. We also plot $f_{FD} + \delta f(t=0)$ in Fig. 2b of the main text. The energy density stored in the initial athermal distribution is given by

$$U = \int_0^\infty D(E) \delta f(E, t=0) E dE, \quad (19)$$

where $D(E)$ is the density of states, which for non-parabolic bands can be derived using Eq. 2, and is given by [4]

$$D(E) = \frac{1}{2\pi^2} \left(\frac{2m_0^*}{\hbar^2} \right)^{3/2} (E + CE^2)^{1/2} (1 + 2CE). \quad (20)$$

Energy lost from the athermal electrons will be passed primarily to the thermal electron subsystem with (time-dependent) temperature T_e and therefore we write Eq. 5a of the main text as

$$C_e \frac{\partial T_e}{\partial t} = -G(T_e - T_l) + \frac{\partial U}{\partial t}, \quad (21)$$

where the last term is the energy source of the thermalized electrons which will henceforth be denoted by

$$S(t) = \frac{\partial U}{\partial t} = \frac{\partial}{\partial t} \int_0^\infty D(E) \delta f(E, t=0) E dE. \quad (22)$$

We numerically integrate Eq. 22 (using Eq. 18), which we plot in Fig. S2 where the blue line is the numerical integration, and the yellow line is an exponential fit with decay constant $t_0 = 4.2$ fs. This means that the athermal electrons decay with an effective decay time of approximately 4 fs and therefore the characteristic time of this process is negligible when compared to pulse duration (on the order of 100 fs).

As the thermalized electronic temperature T_e increases, the thermalization time decreases even further. This is because there are more available high-energy states into which the athermal electrons can thermalize. In fact, for temperatures in the thousands of Kelvins, the thermalization time becomes sub-femtosecond.

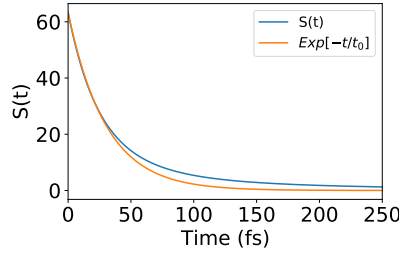


Figure S3: Athermal electron energy density source function $S(t)$ for gold. The blue line is the numerical integration of Eq. 22 and the yellow line is an exponential fit with $t_0 = 30$ fs.

For reference, we repeat this calculation for gold with a room temperature Fermi energy of 5.53 eV and a plasma frequency of $\omega_p \approx 1.3 \times 10^{16}$ rad/s using the same pulse energy of 1 eV. $S(t)$, for this case, is plotted in Fig S3 (blue) along with an exponential fit (yellow). Here we find the decay constant of the exponential fit to be $t_0 = 30$ fs, which is comparable to typical ultrafast pulse widths of interest. Therefore, for gold (and indeed any other materials with $E_F \gg E_{pump}$), the athermal electron distribution must be considered.

Section 3: Temperature Dependent Parameters of the Two Temperature Model

In this section we give the formulas used for calculating the temperature-dependent parameters used in the two-temperature model Eqs. 5, including the heat capacities and the electron-phonon coupling coefficient.

The electron heat capacity C_e is calculated using [9]

$$C_e(T_e) = \int_0^\infty \frac{\partial f_{FD}(E, \mu, T_e)}{\partial T_e} D(E) E dE, \quad (23)$$

where $D(E)$ is the density of states given in Eq. 20, and μ is the chemical potential (that will be discussed later). The lattice heat capacity is assumed constant at $C_l = 2.54 \times 10^6$ Jm⁻³K⁻¹ due to the small change in lattice temperature during the simulation [10].

The electron-phonon coupling coefficient is calculated using [11, 10]

$$g_{ep}(T_e) = \frac{A_{ep} \pi \hbar k_B}{D(E_F)} \int_{-\infty}^\infty D^2(E) \frac{\partial f_{FD}(E, \mu, T_e)}{\partial T_e} dE, \quad (24)$$

where A_{ep} is the product of the electron-phonon mass enhancement parameter and the second moment of the phonon spectrum. As with Ref. [10], we take this to be a free parameter with $A_{ep} = 5.25 \times 10^{-4}$ eV².

In the above formulas, we specify the dependence of the Fermi-Dirac distribution on the chemical potential μ , which is also a temperature-dependent quantity. Numerically, its value can be extracted using the fact that the conduction band electron density is independent of temperature (in the absence of interband transitions), i.e., $n(T_e) = n(T_e = 0)$ with

$$n(T_e) = \int_0^\infty f_{FD}(E, T_e) D(E) dE, \quad (25)$$

and

$$n(T_e = 0) = \int_0^{E_F} D(E) dE, \quad (26)$$

where E_F is the Fermi energy ($E_F \approx 1$ eV for ITO). Thus we extract $\mu(T_e)$ by minimizing the function $|n(\mu(T_e)) - n(T_e = 0)|$ using a minimization algorithm.

Eqs. 23, 24, and 13 can be integrated numerically. For our application, we fit these functions of T_e to fifth-order, single-variable polynomials and use those in the TTM-FDTD code.

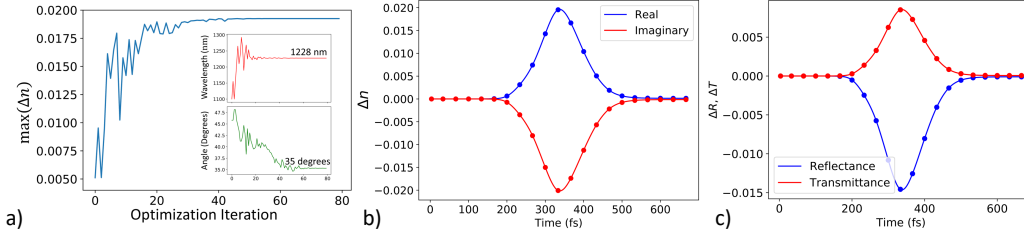


Figure S4: a) Optimization of refractive index change by tuning pump pulse wavelength and incident angle using Nelder-Mead. The horizontal axis is the optimization iteration, and the vertical axis is the maximum change in the refractive index. The insets show how the central wavelength of the pump pulse and its angle of incidence (with respect to the interface normal) evolve during the optimization process. b) Complex refractive index change and c) reflectance and transmittance change as a function of time for the optimized pump pulse.

Section 4: Optimizing pump parameters for maximum index change

The main text was dedicated to validating and refining our hybrid TTM-FDTD approach. In doing so, we found that the properties of the incident excitation have a large effect on the response of ITO, a fact that is not necessarily surprising, but one that can be harnessed. In this section we use a Nelder-Mead optimizer to find the maximum change in refractive index, $\max[\Delta n(\mathbf{r}, t)]^a$, using three of the pump's parameters: incident angle, central wavelength, and linear chirp parameter. Through previous trials, we find that by increasing the pulse width and laser intensity, we also increase the nonlinearity since this results in a larger energy density being deposited into the ITO. We therefore keep these values constant with a peak intensity of $I_{\text{pump}} = 2 \text{ GW/cm}^2$ and a full-width at half-maximum pulse width of $\tau_{\text{pump}} = 150 \text{ fs}$.

The optimization results are found in Fig. S4a, where the main plot shows the evolution of $\max[\Delta n]$ as the optimization progresses, and the insets show the progression of the wavelength (red line) and angle (green line). The optimization algorithm was run several times with different initial conditions so we are confident that we have found a global maximum. We find that the maximum change in the refractive index is on the order of 0.02 and occurs when $\lambda_0 = 1228 \text{ nm}$, $\theta_{\text{inc}} = 35 \text{ degrees}$, and with negligible chirp (not shown). We found that the chirp of the pulse had very little effect on the nonlinearity.

The results in Fig. S4b shows the time dependence of the optical response of the ITO film. We see that the real part of the refractive index increases and the imaginary part decreases by approximately 0.02. In Fig. S4c we plot the corresponding change in reflectance and transmittance, ΔR and ΔT , respectively, as functions of time where we see that the transmittance increases during the pulse irradiation. This is not surprising as it is expected there would be less overall absorption in the ITO layer.

We repeated this optimization for higher pulse intensities (up to 110 GW/cm^2). The results are summarized in Fig. S5, where we show (a) the optimal central wavelength, (b) incident angle, and (c) the maximum of Δn as functions of pump intensity. As expected, we see an increase in the maximum of Δn with increasing intensity. We also see that the optimal incident angle remains relatively constant with a slight decrease with increasing intensity. Interestingly, we see an increase in the optimal wavelength with increasing intensity, which can be easily explained. As shown in Fig. 1 of the main text, as the temperature is increased, the ENZ wavelength red-shifts. Therefore, in order for the pump pulse to remain in the ENZ spectral range, and thus to maximize the nonlinearity, we must tune our pulse to higher wavelengths for increased intensities.

Section 5: Time-domain movies

In Section 4 of the main text, we discuss the time-domain results of our model, and we introduce movies that can be found in the Supplementary Materials. We describe *Movie1* and *Movie1-T* in detail in Section 4, and we will briefly describe the others here.

Movie2, *Movie3*, and *Movie4* are also for a 1240 nm pulse, incident at 30 degrees, for intensities of 150, 50, and 10 GW/cm^2 , respectively. In *Movie2* and *Movie3*, we see the aforementioned decreasing skin depth and change in transmission angle (in the ITO layer), however the reflectance remains high enough that the interference pattern is observed long after the pulse peak. In *Movie4* there is no perceivable change in the optical response due to the low pump intensity.

^aNote that Δn presented in Fig. S4a and n_2 presented in Fig. 3 of the main text are different values. The nonlinear refractive index n_2 is a space and time average of Δn and is normalized by the intensity of the pump in the ITO layer, whereas $\Delta n(\mathbf{r}, t)$ is the instantaneous change of the refractive index.

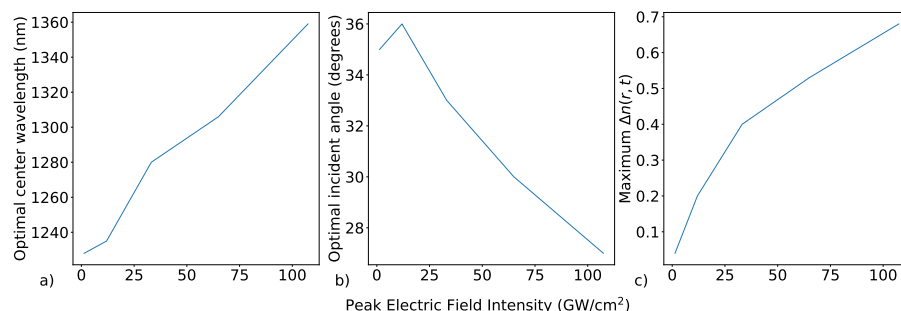


Figure S5: The a) center wavelength, b) incident angle and c) resultant change in the refractive index of the optimized pump pulses found via Nelder Mead optimization, plotted here as a function of peak electric field intensity of the pulse. The pulses were modulated Gaussians with a pulse width of 150 fs.

Movie5 to *Movie8* have the same intensity and incident angle as *Movie1*, but with central wavelengths of 1160, 1200, 1280, and 1320 nm, respectively. We see the same effects as mentioned previously (angle shift and change in skin depth), however, we observe that the initial transmission angles are all different, which is because the initial permittivities are all different for the different wavelengths (as illustrated in Fig. 1 of the main text). From these videos, we cannot observe a perceptible difference in the strength of the nonlinearity, however, we know from spectral results (see Fig. 3 of the main text) that they do exist.

Finally, *Movie9* and *Movie10* have the same intensity and center wavelength as *Movie1* but with incident angles of 0 and 60 degrees, respectively. In *Movie9*, we see negligible change in the optical response as expected from Fig. 3 of the main text (for 0 degrees). In *Movie10*, the reflectance is strong throughout the simulation due to the large incident angle. However a perceptible shift in the transmitted angle in the ITO is visible.

References

- [1] P. Guo, R. D. Schaller, J. B. Ketterson, R. P. H. Chang, *Nature Photonics* **2016**, *10*, 4 267, number: 4 publisher: Nature Publishing Group.
- [2] N. W. Ashcroft, N. D. Mermin, *Solid state physics*, Holt, Rinehart and Winston, New York, **1976**.
- [3] X. Liu, J. Park, J.-H. Kang, H. Yuan, Y. Cui, H. Y. Hwang, M. L. Brongersma, *Applied Physics Letters* **2014**, *105*, 18 181117.
- [4] H. Wang, K. Du, C. Jiang, Z. Yang, L. Ren, W. Zhang, S. J. Chua, T. Mei, *Physical Review Applied* **2019**, *11*, 6 064062.
- [5] W. S. Fann, R. Storz, H. W. K. Tom, J. Bokor, *Physical Review B* **1992**, *46*, 20 13592.
- [6] C. Voisin, N. Del Fatti, D. Christofilos, F. Vallée, *The Journal of Physical Chemistry B* **2001**, *105*, 12 2264.
- [7] E. Carpene, *Physical Review B* **2006**, *74*, 2 024301.
- [8] P. Nozières, D. Pines, *The theory of quantum liquids*, Advanced book classics. Perseus Books, Cambridge, Mass, **1999**.
- [9] C. Kittel, *Introduction to solid state physics*, Wiley, Hoboken, NJ, 8th ed edition, **2005**.
- [10] M. Z. Alam, S. A. Schulz, J. Upham, I. De Leon, R. W. Boyd, *Nature Photonics* **2018**, *12*, 2 79.
- [11] Z. Lin, L. V. Zhigilei, V. Celli, *Physical Review B* **2008**, *77*, 7 075133.

Chapter 6

Inverse Design of pulse shapes

In this chapter, we apply the knowledge introduced in Chapters [1.2](#), [2.1](#), and [2.3](#) to the problem of inverse designing optical pulse shapes for time-varying nanophotonic systems. This project is not a collaboration but a completely novel idea inspired by the results of Chapter 5 and was published in *Optics Express* in June 2023.

Research into time-varying photonics has blown up in recent years due to the discoveries of both natural and metamaterials with high optical nonlinearities that present large permittivity modulations. Energy is not necessarily conserved in these materials because time-translational symmetry is broken, and effects such as gain and reciprocity-breaking are possible. However, to maximize these effects, we need to enhance the light-matter interaction involved, which is typically accomplished by structuring the optical material; however, in this work, we chose a different approach.

The optical response of static optical materials (those with a non-time-varying permittivity) is independent of the pulse shape of the probing light source. The incident pulse can be broken down into frequency components and solved exactly for each. However, for a material whose permittivity rapidly changes in time (for example, ITO from Chapter 5), the incident pulse will experience different permittivities as it propagates, and thus the pulse shape becomes important.

Around the time of working on Chapter 5, I was also working on implementing adjoint topology optimization (see Chapter [2.3](#)) into our in-house FDTD program. We realized the adjoint method could also be applied to pulse optimization; the gradient calculation

includes the source current of Maxwell's equations. Thus we can use the adjoint method to calculate the gradient of an objective function (say, the transmittance) with respect to the source itself. This concept is abstract until we parameterize the source, which we do in two ways,

1. dividing the source pulse into discrete timesteps and the gradient is with respect to the value of the pulse at each step, and
2. decomposing the source pulse into its frequency components, and the gradient is with respect to the phase of each frequency component.

Once the gradient is obtained, gradient-based optimization can be applied to obtain a pulse that maximizes the objective. Simply by shaping the optical pulse, we show that it is possible to achieve gain in these time-varying systems; the shaped pulse gains energy from the system.

In this paper, we apply the pulse shaping method to toy model examples of thin films with a sinusoidal modulation and a realistic example of ITO (using the model of Chapter 5). Future work may include pulse shaping to enhance other optical nonlinearities, including the Kerr effect and high-harmonic generation, or with non-thin film systems like nonlinear metasurfaces. Finally, this work shows that a dual space-time inverse design scheme is possible where the geometry and pulse shape can be co-optimized; however, this has yet to be implemented.

Supplemental materials are included with this article in the form of additional information, which will follow the main article and Python scripts to replicate each example, which can be found in a zip folder at https://opticapublishing.figshare.com/articles/software/Code_zip/21905262.



Inverse design of optical pulse shapes for time-varying photonics

JOSHUA BAXTER^{1,2} AND LORA RAMUNNO^{1,*}

¹Department of Physics, and Nexus for Quantum Technologies Institute, University of Ottawa, Ottawa, Canada

²jbaxt089@uottawa.ca

*Lora.Ramunno@uottawa.ca

Abstract: Recent advancements in materials and metamaterials with strong, time-varying, nonlinear optical responses have spurred a surge of interest in time-varying photonics. This opens the door to novel optical phenomena including reciprocity breaking, frequency translation, and amplification that can be further optimized by improving the light-matter interaction. Although there has been recent interest in applying topology-based inverse design to this problem, we propose a novel approach in this article. We introduce a method for the inverse design of optical pulse shapes to enhance their interaction with time-varying media. We validate our objective-first approach by maximizing the transmittance of optical pulses of equal intensity through time-varying media. Through this approach, we achieve large, broadband enhancements in pulse energy transmission, including gain, without altering the incident pulse energy. As a final test, we maximize pulse transmission through thin films of indium tin oxide, a time-varying medium when strongly pumped in its ENZ band. Our work presents a new degree of freedom for the exploration, application, and design of time-varying systems and we hope it inspires further research in this direction.

© 2023 Optica Publishing Group under the terms of the [Optica Open Access Publishing Agreement](#)

1. Introduction

Until recently, nanophotonics research has been mainly focused on the spatial confinement of light via plasmonics, dielectric resonators, and waveguides. The space of possible devices is virtually infinite, inspiring novel techniques to explore this parameter space in non-intuitive ways. One such method is adjoint sensitivity analysis, which has gained recent interest with the successful design of integrated optical devices including wavelength demultiplexers [1], demultiplexing grating couplers [2], y-splitters [3], and reflectors/resonators [4]. As these methods become more accessible via open-source libraries [5,6], textbooks [7], and tutorials [8], we should expect non-intuitive device design to become the norm in nanophotonics. Adjoint sensitivity analysis has been developed for the design of two- and three-dimensional geometries, however, there is a fourth dimension that has yet to be exploited: time.

In recent years, a new branch of photonics has emerged that investigates wave propagation through time-varying media [9–12] driven by developments and promises of 4D metamaterials [13–15], and recent discoveries of materials with optically-driven, ultrafast permittivity modulations [16–18]. Time-varying materials have been shown to host unique physical effects including frequency translation [19], reciprocity/time-reversal symmetry breaking [14] and have applications in optical isolation [20], all-optical switching [21], and beam steering [22], to name a few.

The design of modulation patterns has been the focus of many recent articles where typically analytic methods are proposed to calculate the scattering of light in unbounded, temporally structured (modulated) media. These methods are used to design homogeneous time-domain

structures such as temporal tapers/anti-reflection filters [23–25], analogue filters [26], differentiators [26–28], vortex beam generators [27], and amplifiers [25]. These papers demonstrate control over the propagation of light without the requirement of geometric structuring.

Furthermore, there is interest in applying inverse design techniques for geometric optimization in time-varying photonic media [29]. Indeed, as derived in this article, it is possible to inverse design time-varying photonic devices using time-domain adjoint sensitivity analysis [30–32] with minor adjustments. These exciting works and prospect are focused on tuning the 4D properties of the medium through which light is propagating. This is only half of the story when it comes to controlling light - matter interaction; we can also control the light.

In static, linear materials, the shape of the incident pulse is irrelevant to the device output as it can always be decomposed into its Fourier components. In time-varying media, this is no longer the case. As a trivial example, two identical pulses delayed with respect to each-other (a simple form of pulse shaping) and incident onto a time-varying thin film may experience different refractive indices and thus each pulse will have a different transmittance/reflectance spectrum, despite being spectrally equivalent. As a less trivial example, one may conceive of a pulse chirped so that its central frequency in the time-varying film will follow the transmittance peak. In the same way one may tailor the modulation signal for a desired effect, we can also tailor the pulse shape of the incident light.

In this paper, we introduce a new paradigm in inverse design to optimize the shape of light pulses based on a new application of adjoint sensitivity analysis. We present protocols for the inverse design of the shapes of pulses incident on time-varying optical materials that can be used, for example, to enhance or minimize the amount of light that is transmitted through such a time-varying medium. This is done via two approaches: 1) optimizing the pulse in time, where the field of the pulse at each (discretized) time-step is tuned, and 2) optimizing the pulse in frequency, where the phase of each (discretized) frequency component is tuned as in 4f pulse shaping. The first approach allows for the full exploration of parameter space in time, presenting us with non-intuitive pulse shapes that are not restricted to the frequency spectrum of the initial pulse shape. The second approach conserves the frequency spectrum of the initial pulse. In Section 2, we present the adjoint sensitivity analysis for time-varying materials. We outline how they are implemented for pulse shaping in time (Section 3) and in frequency (Section 4), where we test both approaches through a series of toy-model examples and analyze the behaviour of the optimized pulse. Finally, in Section 5, we apply these two methods to the realistic example of light propagation through a thin film of indium tin oxide (ITO), a material synonymous with time-varying photonics due to its high refractive index perturbations when excited near its epsilon-near-zero (ENZ) band [16,17].

2. Adjoint method for time-varying materials

To optimize any sort of design, be it spatial or temporal, it is useful to calculate the gradient of the objective function F with respect to the tunable parameters. Once this gradient is known, there are many iterative, gradient-based optimization algorithms that can be used for the design process. Adjoint sensitivity analysis provides a very convenient method for obtaining this gradient with just two simulations, for an arbitrarily large number of tuning parameters, and accordingly has been investigated extensively for topology optimization of static optical systems.

Since we wish to shape pulses incident on time-varying optical materials, we employ a time-domain adjoint sensitivity analysis, wherein the objective function F is defined as the integral over a time-dependent objective function ψ , which in general depends on a set of M tunable parameters $\mathbf{p} = [p_0, \dots, p_i, \dots, p_M]$, as well as the time- and space-varying electric ($\mathbf{E}$) and magnetic ($\mathbf{H}$) fields, and possibly additional auxiliary fields [32]. The gradient of F is then

described by [7]

$$\frac{dF}{d\mathbf{p}} = \left[\frac{dF}{dp_0}, \dots, \frac{dF}{dp_i}, \dots, \frac{dF}{dp_M} \right] = \frac{d}{d\mathbf{p}} \int_0^{T_m} \psi(\mathbf{p}, \mathbf{E}, \mathbf{H}, \dots) dt, \quad (1)$$

where T_m is the maximum simulation time, and the fields are assumed to be zero before $t = 0$ and after $t = T_m$. The functional form of ψ is known and set according to what we desire to optimize.

In the [Supplement 1](#) (SI) Section A we outline an adjoint sensitivity analysis for two classes of time-varying materials, one that can be described by a "dispersionless" time-varying permittivity $\varepsilon(t)$ (SI Section A.1), and another that can be described by time-varying dispersive models for the current density or polarization fields (SI Section A.2). For the latter case, we derive explicitly the equations for the Drude model with a time-varying plasma frequency $\omega_p(t)$.

Our aim is to tune the time signal of the source current $\mathbf{J}_s(t)$ in order to maximize some quantity (e.g. the transmittance), rather than tuning the geometric topology. This is done via two simulations: a *forward* simulation that is a direct solution of Maxwell's equations with a current source $\mathbf{J}_s(t)$, and an *adjoint* simulation that is a solution to an adapted form of Maxwell's equations with a current source (see Eq. S13)

$$\mathbf{J}_s^{adj} = -\frac{\partial \psi}{\partial \mathbf{E}}(T_m - t) \quad (2)$$

placed at the spatial location where the objective function is defined. Note that $\mathbf{J}_s^{adj}$ is time-reversed for reasons made clear in its derivation.

For both time-varying permittivity and time-varying dispersion, we have a prescription for calculating the full gradient $\partial F / \partial \mathbf{p}$ with only two simulations. The steps involved are

- (1) compute the forward problem (Eq. S10 or S18) to find the forward fields, and evaluate the external current density source of the adjoint problem (Eq. (2)),
- (2) compute the adjoint problem (Eq. S12 or S19) to find the adjoint fields, with minor modifications made to the simulation software if this is not the same as the forward problem,
- (3) compute the gradient (Eq. S14 or S20) using the time-reversed adjoint fields and the residue vector calculated from the forward fields (Eq. S15 or S21).

Our pulse shaping inverse design methods are implemented in a one-dimensional (1D) finite - difference time - domain (FDTD) code [33], an example of which can be found in [Code 1](#) (Ref. [34]). The python scripts for all examples of Sections 3 – 5 are also made available in Ref. [34] with titles corresponding to the section numbers.

We consider 1D examples for the sake of simplicity in introducing this method, and for simple interpretation of the code we are making available. In fact, 2D and 3D implementations of this method are entirely possible, depending upon the spatial extent of the objective function. However, 1D is also useful in itself because much current work in active nanophotonic has involved planar geometries like thin-films [16,35] and metasurfaces [11,17], which can be described using bulk (for thin-films) or effective [36,37] (for metasurfaces) refractive indices. This enables the problem to be scaled to 1D wherein our provided pulse shaping code can be directly applied.

In our code, we take the propagation direction to be along y, and our incident current density $\mathbf{J}_s$ to be linearly polarized in the x-direction. We define our instantaneous objective function $\psi(t)$ as the square of the x-component of $\mathbf{E}$ at a single location y_{trans} at least a few wavelengths away

from the thin film. We then set the objective function as

$$F = \frac{1}{2} \int_0^{T_m} E_x^2(y_{trans}, t) dt, \quad (3)$$

which is proportional to the transmitted pulse energy, and the factor of 2 is included for future convenience.

The source current of the adjoint simulation is thus applied only at the single location y_{trans} , and using Eq. (2) we find it is given by

$$J_s^{adj}(t) = -E_x(y_{trans}, T_m - t), \quad (4)$$

where $E_x(y_{trans}, t)$ is stored from the forward simulation. In order to accomplish pulse shaping, we allow the external source of our forward simulation, $J_s(t)$, to vary with respect to a set of tuning parameters, p_i . Assuming $J_s(t)$ is applied at a single point y_s , the gradient of the objective function (see Eqs. S14 or S20) becomes

$$\frac{\partial F}{\partial p_i} = - \int_0^{T_m} E_x^{adj}(y_s, T_m - t) \frac{\partial J_s(t)}{\partial p_i} dt. \quad (5)$$

In the sections that follow, we introduce two strategies for parameterizing $J_s(t)$. In Section 3 we take the set p_i to be the actual time-domain values of the (normalized) pulse signal at discrete times t_i . In Section 4 we take p_i to be a phase factor applied to the i^{th} frequency component of the pulse, such as would be introduced in a 4f pulse shaping setup.

3. Pulse shaping in time

In this section, we optimize our objective function by directly tuning the pulse shape as a function of time. We take the values of the pulse at each discretized moment in time t_i as our tuning parameters p_i though other methods are possible (for example, restricting the pulse to a particular functional form with the fitting parameters as the p_i). To conserve energy of the pulse throughout the optimization process, however, the pulse must be normalized at each optimization step. Our external current density source is then taken as

$$J_s(t) = J_0 \frac{p(t)}{\bar{p}}, \quad (6)$$

where $p(t)$ is the unitless pulse shape that we are optimizing, and $\bar{p}^2 = 1/T_m \int_0^{T_m} p^2(t) dt \approx 1/N_T \sum_{i=0}^{N_T} p_i^2$ is an energy normalization factor, where N_T is the maximum number of time steps in the simulation. Our optimization parameters are then $p_i = p(t_i)$ for discrete values of time t_i . The current density amplitude is J_0 , which is chosen such that $\int J_s^2(t) dt = 1$.

Assuming fixed geometrical topology, and a source located at y_s in our 1D forward simulation, the gradient according to Eq. (5) becomes

$$\frac{\partial F}{\partial p_i} = -E_x^{adj}(y_s, T_m - t_i) \frac{J_0}{\bar{p}} \left[1 - \frac{1}{N_T} \frac{p_i^2}{\bar{p}^2} \right]. \quad (7)$$

This expression gives us the sensitivity of the objective function F (proportional to the transmitted pulse energy) with respect to p_i (the incident pulse at each time step). With this gradient we can now tune each p_i , and thus the incident pulse shape, in order to maximize the F and thus the transmitted pulse energy.

Though we have ensured that the incident pulse energy remains constant, we have made no restrictions on its frequency components or bandwidth. In fact, tuning the pulse using this

gradient can and will result in the creation of new frequency components. While this is not so useful for optimizing a given pulse source with a given bandwidth (a scenario we return to in Section 4), it could be useful for exploring, for example, what kind of source with what bandwidth would be ideal for a given geometry and material response. It also presents some physically interesting results that we will describe below.

In the following, we use our pulse shaping method to maximize the transmitted pulse energy through time-varying dielectric (Section 3.1), and metallic (Section 3.2) thin films. We will restrict ourselves to test-models commonly used in time-varying photonics: sinusoidal varying permittivities [9,38,39], including with frequency dispersion. In SI Section B, for completeness, we demonstrate the effect of pulse shaping when tested on static (time-invariant) thin films where it is found that the pulse changes frequency to match the transmittance resonances of the thin films. In Section 5 we will demonstrate our method on a physical example, ITO pumped by high intensity light in its ENZ spectral region.

3.1. Test case 1: time-varying permittivity

As the first test case, we consider a 200 nm dispersionless thin film, with a slowly-varying sinusoidal modulation of the permittivity $\varepsilon(t)$, as plotted in Fig. 1(a) (dashed green line – right axis) alongside the initial pulse (solid blue line – left axis) that is taken to be a modulated Gaussian with a central wavelength of $\lambda_0 = 2 \mu\text{m}$ and duration $\tau = 13$ fs. The optimized pulse after 60 iterations is plotted in Fig. 1(b) (red line - left axis). Note that in Figs. 1(a) and (b), only a portion of the pulses are shown for a small window of time in the larger simulation.

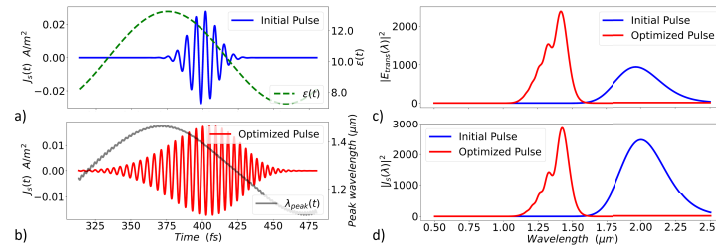


Fig. 1. a) Initial Gaussian input pulse, with duration 13.3 fs centered at $2 \mu\text{m}$ (blue line - left vertical axis) and the time-varying permittivity (green, dashed line - right vertical axis). b) Input pulse after optimization (red line - left vertical axis) and its instantaneous peak wavelength as a function of time (black line - right vertical axis). c) Transmission spectrum of the initial (blue line) and optimized (red line) pulses. d) Spectrum of the initial (blue line) and optimized (red line) pulses.

Also plotted in Fig. 1(b) is the instantaneous peak wavelength of the optimized pulse as a function of time, calculated via windowed Fourier transforms (black line – right axis). It is clear that the optimized pulse is chirped in such a way that the instantaneous peak wavelength is following the time-varying permittivity of the thin-film, which would consequently have a time-varying transmission resonance.

The transmitted energy spectra for the initial (blue line) and optimized (red line) pulses are plotted in Fig. 1(c), where we see that the transmitted pulse after optimization has a peak that is $3\times$ higher than the initial pulse. The enhancement of the transmitted pulse energy after optimization (and thus the objective function F) is found by taking the ratio of the integral over the initial and optimized spectra, and is found to be larger than $2.5\times$. In Fig. 1(d) we plot the spectra of the initial (blue line) and optimized (red line) pulses where we see that the optimized pulse is

blue shifted, and the bandwidth is reduced. The spectrum is no longer a Gaussian, but includes additional frequency components. Note that the values on the left-hand axis for Figs. 1(c) and 1(d) are plotted such that they have the same units, that is, $J_s(\lambda)$ is scaled to have the same units as the electric field. This remains true in future sections, and makes physical interpretation easier. Indeed, comparing Fig. 1(c) and Fig. 1(d), we find that the new pulse is almost completely transmitted, whereas the initial pulse is highly attenuated.

3.2. Test case 2: time-varying plasma frequency

For our second test case, we consider a 200 nm Drude metal thin film with a rapid sinusoidal time-dependent plasma frequency $\omega_p(t) \sim \sin(2\omega_0 t)$, where ω_0 is the central frequency of the initial pulse corresponding to $\lambda_0 = 2 \mu\text{m}$. The initial pulse is plotted in Fig. 2(a) for a cropped time window, and is the same initial pulse shape as used in the previous section. The green dotted line shows the time-varying plasma frequency. After applying our optimization algorithm to the initial pulse in this time-varying medium, we obtain the optimized pulse shown in red in Fig. 2(b) (in the same cropped time window as for panel a). In Fig. 2(c) we plot the transmission spectrum of the initial (blue line) and optimized (red line) pulses; the inset shows the same in log scale. The transmitted pulse energy from the optimized pulse has increased by a factor of 4.5 over that of the initial, unshaped pulse. Finally in Fig. 2(d) we plot the pulse spectrum of the initial and optimized pulses, again with a log scaled-inset.

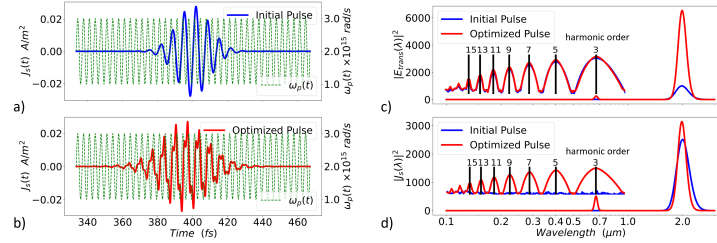


Fig. 2. a) Initial Gaussian input pulse with duration 13.3 fs centered at $2 \mu\text{m}$ (blue line - left vertical axis) and the time-varying plasma frequency (green, dashed line - right vertical axis). b) Input pulse after optimization (red line - left vertical axis) and the time-varying plasma frequency, shown again here for convenience (green, dashed line - right vertical axis). c) Transmission spectrum of the initial (blue line) and optimized (red line) pulses. The inset shows a log plot at the lower wavelengths, demonstrating the generated odd harmonic orders, whose locations are indicated with black vertical lines. d) Spectrum of the initial (blue line) and the optimized (red line) pulses. The inset shows a log plot at the lower wavelengths, demonstrating the odd harmonic components in the optimized pulse only.

The first effect of optimization on the input pulse is the increased pulse duration (see Fig. 2(b)) and corresponding decrease in bandwidth near the central wavelength (see Fig. 2(d)) illustrating that transmittance enhancement only occurs within a narrow bandwidth as further demonstrated in SI Section C. The second effect is the appearance in Fig. 2(b) of higher frequency components that modulate the pulse in time, that appear to sync with the modulated plasma frequency. From the optimized pulse spectrum in Fig. 2(d) we see that these are odd harmonics. Indeed, it can be shown that odd harmonics are generated because the plasma frequency oscillates at $2\omega_0$ and are a demonstration of Floquet harmonics [38]. Since these harmonics are transmitted out of the thin film (see the inset in Fig. 2(c)), and because the thin film has a high transmittance at low

wavelengths (a property of the Drude model), it is not a surprise that our optimization algorithm added these harmonics to our pulse (see Fig. 2(d)).

As mentioned in Section 3.3, the scale of the vertical axes of Figs. 2(c) and (d) are in the same units, and we can see that the optimized transmitted pulse has experienced significant amplification. Indeed, due to the time-varying nature of these films, the energy of the input pulse need not be conserved, and gain is an expected consequence. It can be shown that parametric amplification occurs in periodic time-varying systems (time crystals) with a modulation frequency $\omega_{mod} = 2\omega_0$, where ω_0 is the frequency of light [9,38,40] due to the coherent sum of the forward scattered waves from the time- boundaries. As such, the phase relationship between the optical pulse and the sinusoidal modulation is critical to amplification as demonstrated in SI Section C. In SI Section D, we show via Poynting's theorem that gain is accounted for by negative energy lost to the free electron system ($\mathbf{J} \cdot \mathbf{E}$) and that the total energy of the system is conserved.

Our pulse shaping method was able to find an optimized pulse that achieves broadband gain upon transmission through this time-varying Drude-metal. After integrating over both the initial and optimized transmission spectra, we find there is $\sim 2\times$ more energy being transmitted in the optimized pulse.

4. Pulse shaping in frequency

In this section we will introduce and demonstrate pulse shaping in frequency, wherein the phase of the frequency components is optimized, as would be in a phase-based 4f pulse shaper. Unlike the pulse shaping in time method of the previous section, this method does not generate new frequency components so would be ideal for optimizing a pre-existing laboratory pulse of a set bandwidth. This method is naturally energy preserving, and the frequency spectrum remains the same for all possible pulse shapes (up to potentially small numerical errors arising from the discretization of the Fourier transform). If one were to consider amplitude-based pulse shaping (which would also be possible with our formalism, but that we do not consider here) then, of course, energy would not be conserved.

Consider a current density $\tilde{J}_s(t)$ that would represent the input to a 4f pulse shaper. The j^{th} frequency component ω_j of this current source is found via a discrete Fourier transform of $\tilde{J}_s(t)$,

$$\tilde{J}_s(\omega_j) = \sum_n \tilde{J}_s(t_n) \exp(i\omega_j t_n). \quad (8)$$

As in 4f pulse shaping, we allow each term in this sum to experience a distinct phase shift. We take these phase shifts $p_j = \phi(\omega_j) = \phi_j$ to be the tunable parameters in our optimization algorithm. We set the shaped pulse in time, and thus the source of our forward problems (Eqs. S10 or S18) to be

$$J_s(t_n) = \frac{1}{2} \left[\sum_j \tilde{J}_s(\omega_j) \exp(-i\omega_j t_n + i\phi_j) + c.c. \right] \quad (9)$$

where c.c. means complex conjugate, which we have added to ensure $J_s(t_n)$ is real-valued. The gradient of the objective function F with respect to the tuning parameters $p_j = \phi_j$ (Eq. (5)) now becomes

$$\frac{\partial F}{\partial \phi_j} = -\Delta t \sum_{n=1}^{N_T} E_x^{adj}(y_s, T_m - t_n) \text{Im}[\tilde{J}_s(\omega_j) \exp(-i\omega_j t_n + i\phi_j)] \quad (10)$$

where N_T is the number of time iterations of our simulations. As in Section 3, we set the time step $\Delta t = t_{i+1} - t_i \forall i$ as uniform. As before, only two simulations are required to calculate the gradient, which allows us to optimize F with respect to the phase shifts ϕ_j , and thereby tune the incident pulse.

In the following subsections, we will use our pulse shaping in frequency method to maximize the transmitted energy through time-varying dielectric and metallic thin films, similar to the previous section.

4.1. Test case 1: Time-varying permittivity

Here we test our pulse shaping in frequency method on a similar problem to that considered in Section 3.3, a 200 nm dispersionless thin film with a slowly-varying sinusoidal modulation $\varepsilon(t)$ as plotted in Fig. 3(a) (green, dashed line – right axis) alongside the initial pulse (solid blue line – left axis). The initial pulse is identical to that used in Section 3.1. We use a numerical discrete Fourier transform to construct the pulse in the frequency domain via Eq. (8) with 300 discrete frequencies sampled evenly between $0.5\omega_0$ and $2\omega_0$, where ω_0 is the pulse center frequency.

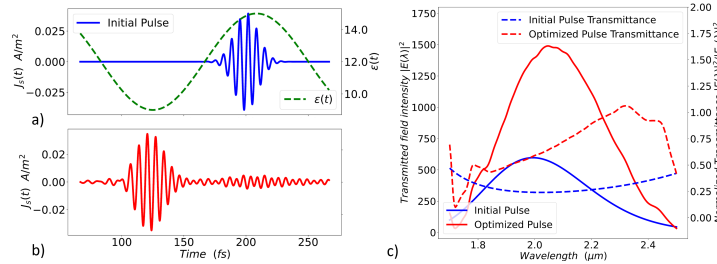


Fig. 3. a) Initial Gaussian input pulse with duration 13.3 fs centered at $2 \mu\text{m}$ (blue line - left vertical axis) and the time-varying permittivity oscillations (green, dashed line - right vertical axis). b) Optimized input pulse. c) Left vertical axis: transmission spectrum of the initial (blue line) and optimized (red line) pulses. Right vertical axis: normalized transmittance of the initial (blue, dashed line) and optimized (red, dashed line) pulses.

The optimized pulse after 60 iterations is plotted in Fig. 3(b). The main effect of the optimization is to create a delay such that the majority of the pulse has been shifted to where the permittivity is lowest (around 125 fs) as the instantaneous transmittance would be highest at this permittivity minima. In fact, near the minima of the permittivity where $\varepsilon \sim 9$, the thin film has a resonance around $\lambda \sim 2400 \text{ nm}$. As the permittivity increases, the resonance wavelength will further redshift from the bandwidth of the pulse (which, recall, is centred at $2 \mu\text{m}$). Indeed, the best course of action for the optimizer was to move the pulse to a lower permittivity, where the resonance overlaps with the pulse bandwidth.

In Fig. 3(c) we plot the transmitted intensity (left vertical axis – solid lines) and the normalized transmittance (ratio of transmitted to input intensity, right vertical axis – dashed lines) for the initial (blue line) and optimized (red line) pulses. We have achieved more than $2.4 \times$ enhancement of the transmitted intensity at the peak wavelength, and $2.2 \times$ enhancement in transmitted energy (that is, integrated across the spectrum). Here we see further evidence for our physical interpretation of the action of the optimization on the pulse. We see that enhancement in transmittance (comparing the dashed red and blue lines) is highly biased towards the higher (resonant) wavelengths. A notable feature is the presence of gain at around 2320 nm. This is most likely due to frequency translation owing to the time-varying permittivity [41]. This frequency translation causes the pulse frequency components to shift while traversing a time-varying material. In spectral regions of high transmittance, frequency translation can result in a transmittance > 1 .

From this simple model we can see that by selectively delaying the frequency components of the pulse, we can achieve a broadband transmittance enhancement across the spectrum of the pulse without changing the amplitude of the pulse spectral components.

4.2. Test case 2: time-varying plasma frequency

In this example, we use a similar setup as in Section 3.2. We use the same initial pulse (plotted in Fig. 4(a), blue line) incident on a 200 nm Drude-metal film with the same time-varying plasma frequency as before (plotted in Figs. 4(a) and (b), green dashed line)

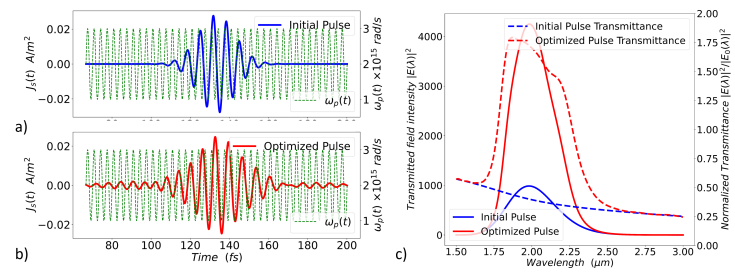


Fig. 4. a) Initial Gaussian input pulse with duration 13.3 fs centered at $2 \mu\text{m}$ (blue line - left vertical axis) and the time-varying plasma frequency (green, dashed line - right vertical axis). b) Input pulse after optimization and time-varying plasma frequency shown again for convenience (green dashed line - right vertical axis). c) Left vertical axis: transmission spectrum of the initial (blue line) and optimized (red line) pulses. Right vertical axis: normalized transmittance of the initial (blue, dashed line) and optimized (red, dashed line) pulses.

The optimized pulse after 30 iterations is plotted in Fig. 4(b). We see the optimized pulse (red line) is largely unchanged, but slightly delayed relative to the initial pulse (Fig. 4(a), blue line). This delay works to align the pulse sub-cycle peaks with the plasma frequency troughs (green-dashed line in Figs. 4(a) and (b)) near the middle of the pulse. The optimized pulse is also stretched in time.

In Fig. 4(c) we plot the transmission spectrum (left vertical axis – solid lines) and the normalized transmittance (right vertical axis – dashed lines) of the initial (blue line) and optimized (red line) pulses. As in Section 3.2, we see that our new pulse has again achieved gain due to parametric amplification. We achieve a $3.5\times$ increase in the transmitted energy (that is, integrated across the transmission spectrum). Once again we show how pulse shaping can be used to achieve broadband gain via transmission through time-varying media, only this time through the controlled delay of the pulse frequency components.

5. Pulse shaping for a strongly pumped, ENZ material

In the previous section, we explored our pulse shaping inverse design processes by maximizing the energy transmitted through time-varying materials based on toy-models. In this section, we will demonstrate both our time and frequency pulse shaping methods for a realistic time-varying medium by maximizing the transmitted energy of a probe pulse through a pumped ITO thin film.

The permittivity of ITO is highly dependent on the temperature of the conduction band electrons, and as such, it is a time-varying material under ultrafast pulse irradiation, especially near its ENZ band [16,17,19,42]. Because it exhibits strong, and fast permittivity perturbations, ITO is a material of high interest in the field of active nanophotonics. Its nonlinear optical

properties are well studied and can be modelled using a self-consistent multiphysics model that couples electrodynamics and thermodynamics introduced in Ref. [35].

5.1. Pump simulation

Our goal will be to optimize a probe pulse incident on an ITO thin film after the film has been irradiated by an ultrafast intense light pulse, which we call the pump pulse. This pump pulse creates a time-varying medium through a temperature dependent plasma frequency, that we simulate by implementing the model of Ref. [35] into a 1D-FDTD solver (with the code provided in Ref. [34]). We simulate a modulated Gaussian pump pulse with peak intensity $I_{peak} = 13.3$ TW/cm², pulse duration $\tau = 100$ fs, and center wavelength $\lambda_0 = 1.23$ μ m incident on a 320 nm thin film of ITO. This center wavelength corresponds to the ENZ wavelength of the ITO film. We store the time and spatially dependent plasma frequency $\omega_p(\mathbf{r}, t)$ at each time-step and position in a text file for future use; this could be cumbersome for 2 and 3D geometries if a large number of points are required. We plot $\omega_p(\mathbf{r}, t)$ in Fig. 7 where the vertical axis represents the depth in the ITO film (where $\mathbf{r} = y\hat{y}$), and the horizontal axis represents time (which is cropped to highlight the important time window). The unpumped ITO film has a plasma frequency everywhere of $\omega_p = 2.97 \times 10^{15}$ rad/s. Upon irradiation with the pump pulse, the plasma frequency in the ITO decreases as the excited conduction electrons occupy higher energy levels [16,35,43]. In Fig. 5, the plasma frequency drops to half of its initial value through most of the ITO layer in the period of ~ 200 fs and it's return to equilibrium will take several picoseconds (not shown in figure).

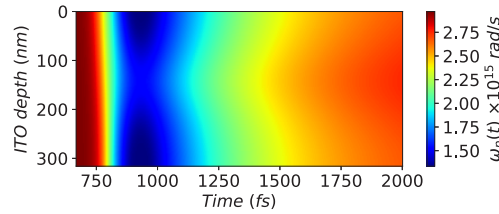


Fig. 5. Colour contour plot of the space (vertical axis) and time (horizontal axis)-varying plasma frequency $\omega_p(y, t)$ which is extracted from the multiphysics modelling of an ITO thin film under high intensity pulse irradiation [35].

With the spatiotemporal plasma frequency of the pumped ITO film stored, and the time-varying medium thus defined, we can now apply our time and frequency pulse shaping techniques to maximize the transmitted energy of a probe pulse incident on the pumped ITO film. We turn to this in the next two subsections.

5.2. Probe optimization: pulse shaping in time

In this subsection, we will use our pulse shaping in time method to optimize the transmitted energy of the probe pulse incident on the pumped ITO film. The initial probe pulse is a Gaussian centered at $\lambda_0 = 2$ μ m with pulse duration $\tau = 67$ fs. It is plotted in Fig. 6(a) (blue line) alongside the space-averaged plasma frequency from Fig. 5 for reference (dashed, green line). The optimized pulse after 300 iterations is plotted in Fig. 6(b).

Additional filtering was required to keep the pulse spectrum in the near-infrared as the optimizer prefers to introduce frequency components as high as possible to capitalize on the high-frequency transparency of metals. Prior to the forward simulation, the (pre-filtered) pulse $p(t)$ is band-pass filtered via fast-Fourier transforms. As in Section 3, the pulse is normalized to ensure no energy is added or removed from the filtered pulse $J_s(t)$. This filtering procedure is differentiable via

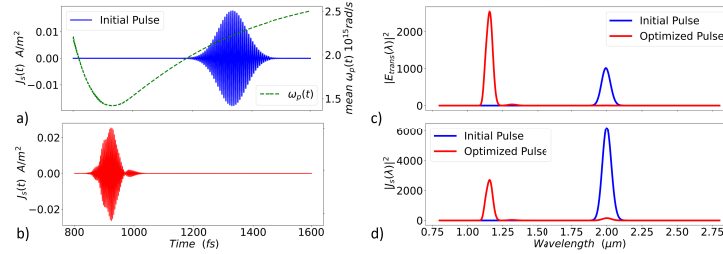


Fig. 6. a) Initial Gaussian probe pulse with duration 67 fs centered at 2 μm (blue line) and, b) optimized probe pulse (red line). In a) we overlay the spatially-averaged plasma frequency (dashed green line). c) Transmission spectrum of the initial (blue line) and optimized (red line) pulses. d) Spectrum of the initial (blue line) and optimized (red line) pulses.

automatic differentiation [44]. The derivative $\frac{\partial J_s(t)}{\partial p_i}$ in Eq. (5) now accounts for the band-pass filtering and normalization, where p_i is the i^{th} timestep of the pre-filtered pulse $p(t)$ that we are optimizing. Although not necessary, automatic differentiation can also be applied to the pulse normalization of Section 3.

The major effect of the optimization is to shift the pulse in time, such that most of the energy is near the minimum of the plasma frequency. As the plasma frequency decreases, the imaginary component of the permittivity $\text{Im}(\epsilon(\omega))$ also decreases for $\omega > \omega_p$, thus reducing the loss, and the real component $\text{Re}(\epsilon(\omega))$ increases. The film exhibits dielectric behaviour, and it is understandable why the optimizer chose to delay the pulse as it did, and reduce the center wavelength. Furthermore, as plotted in Fig. 6(c), the optimized pulse is centered at $\lambda_{0,\text{opt}} = 1177$ nm corresponding to a transmittance maximum for a thin film with a static plasma frequency of $\omega_p = 1.5 \times 10^{15}$ rad/s. The transmitted spectra are plotted in Fig. 6(d) where the energy of the optimized pulse is 92% transmitted, 5.5 $\times$ higher than the initial pulse.

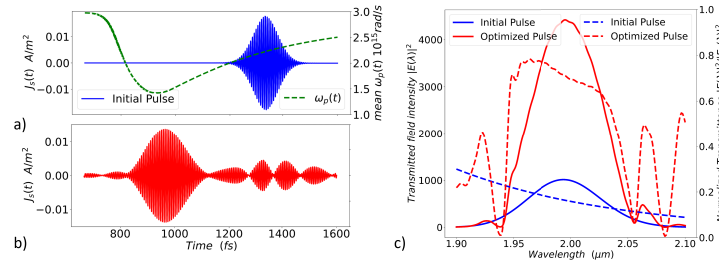


Fig. 7. a) Initial Gaussian probe pulse with duration 67 fs centered at 2 μm (blue line) and, b) optimized probe pulse (red line). In a) we overlay the spatially-averaged plasma frequency (dashed green line) for comparison. c) Left vertical axis: transmission spectrum of the initial pulse (blue line) and the optimized pulse (red line). Right vertical axis: Normalized transmittance of the initial (blue, dashed line) and optimized (red, dashed line) pulses.

5.3. Probe optimization: pulse shaping in frequency

Like in the previous section, here we are optimizing the transmitted energy of the probe pulse traversing the pumped ITO film simulated in Section 5.1, only this time using our pulse shaping in frequency method. The initial probe pulse is again a Gaussian centered at $\lambda_0 = 2 \mu\text{m}$ with duration $\tau = 67 \text{ fs}$, as plotted in Fig. 7(a) (blue line) along with the space-averaged plasma frequency from the pump simulation (dashed, green line) calculated from Fig. 5.

After 30 iterations of pulse shaping, we obtain the optimized pulse plotted in Fig. 7(b). Once again, we see the pulse being time-shifted such that it overlaps in time with the plasma frequency minimum. In Fig. 7(c) we plot the transmitted field intensity (left vertical axis – solid lines) and the normalized transmittance (right vertical axis – dashed lines) of the initial (blue line) and optimized (red line) pulses. The transmitted energy is increased by a factor of $4\times$ after optimization.

6. Conclusion

We have introduced a method for the inverse design of optical pulse shapes for time-varying nanophotonic systems, opening a new paradigm for control over light-matter interaction. We derive the sensitivity of an objective function to the pulse shape in two ways. First, we introduce pulse shaping in time, where the gradient of the objective function with respect to the pulse amplitude at a given time can be extracted using two FDTD simulations, allowing for pulses of arbitrary frequency components (but same pulse energy) to be designed. Second, we develop pulse shaping in frequency, where the phase of discrete frequencies in the pulse are tuned, replicating a 4f pulse shaping setup. We demonstrate these methods in time varying materials, including the optimization of a probe pulse in intensely irradiated ITO thin films. This is an unconventional approach to computational-design in nanophotonics, but one that is likely to be important in the near-future given the current interest in active nanophotonics.

Funding. Compute Canada; Natural Sciences and Engineering Research Council of Canada.

Acknowledgments. The authors would like to thank Prof. Antonio Calà Lesina, Prof. Israel De León, and Dr. Orad Reshef for helpful discussions.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper can be generated via the python scripts in [Code 1](#) [34].

Supplemental document. See [Supplement 1](#) for supporting content.

References

1. A. Y. Piggott, J. Lu, K. G. Lagoudakis, J. Petykiewicz, T. M. Babinec, and J. Vuckovic, "Inverse design and demonstration of a compact and broadband on-chip wavelength demultiplexer," *Nat. Photonics* **9**(6), 374–377 (2015).
2. A. Y. Piggott, J. Lu, T. M. Babinec, K. G. Lagoudakis, J. Petykiewicz, and J. Vuckovic, "Inverse design and implementation of a wavelength demultiplexing grating coupler," *Sci. Rep.* **4**(1), 7210 (2014).
3. C. M. Lalau-Keraly, S. Bhargava, O. D. Miller, and E. Yablonovitch, "Adjoint shape optimization applied to electromagnetic design," *Opt. Express* **21**(18), 21693 (2013).
4. G. H. Ahn, K. Y. Yang, R. Trivedi, A. D. White, L. Su, J. Skarda, and J. Vuckovic, "Photonic Inverse Design of On-Chip Microresonators," *ACS Photonics* **9**(6), 1875–1881 (2022).
5. L. Su, D. Vercauteren, J. Skarda, N. V. Sapra, J. A. Petykiewicz, and J. Vuckovic, "Nanophotonic inverse design with SPINS: Software architecture and practical considerations," *Appl. Phys. Rev.* **7**(1), 011407 (2020).
6. A. M. Hammond, A. Oskooi, M. Chen, Z. Lin, S. G. Johnson, and S. E. Ralph, "High-performance hybrid time/frequency-domain topology optimization for large-scale photonics inverse design," *Opt. Express* **30**(3), 4467 (2022).
7. M. Bakr, A. Elsherbeni, and V. Demir, *Adjoint Sensitivity Analysis of High Frequency Structures with MATLAB®* (Institution of Engineering and Technology, 2017).
8. R. E. Christiansen, R. E. Christiansen, O. Sigmund, and O. Sigmund, "Inverse design in photonics by topology optimization: tutorial," *J. Opt. Soc. Am. B* **38**(2), 496–509 (2021).
9. E. Galiffi, R. Tirole, S. Yin, H. Li, S. Vezzoli, P. A. Huidobro, M. G. Silveirinha, R. Sapienza, A. Alù, and J. B. Pendry, "Photonics of time-varying media," *Adv. Photonics* **4**(01), 1 (2022).

10. M. Ferrera, N. Kinsey, A. Shaltout, C. DeVault, V. Shalaev, and A. Boltasseva, "Dynamic nanophotonics [Invited]," *J. Opt. Soc. Am. B* **34**(1), 95–103 (2017).
11. A. M. Shaltout, V. M. Shalaev, and M. L. Brongersma, "Spatiotemporal light control with active metasurfaces," *Science* **364**(6441), eaat3100 (2019).
12. T. Badloe, J. Lee, J. Seong, and J. Rho, "Tunable Metasurfaces: The Path to Fully Active Nanophotonics," *Adv. Photonics Res.* **2**(9), 2000205 (2021).
13. N. Engheta, "Metamaterials with high degrees of freedom: space, time, and more," *Nanophotonics* **10**(1), 639–642 (2020).
14. A. Shaltout, A. Kildishev, and V. Shalaev, "Time-varying metasurfaces and Lorentz non-reciprocity," *Opt. Mater. Express* **5**(11), 2459 (2015).
15. K. Fan, R. D. Averitt, and W. J. Padilla, "Active and tunable nanophotonic metamaterials," *Nanophotonics* **11**(17), 3769–3803 (2022).
16. M. Z. Alam, I. De Leon, and R. W. Boyd, "Large optical nonlinearity of indium tin oxide in its epsilon-near-zero region," *Science* **352**(6287), 795–797 (2016).
17. M. Z. Alam, S. A. Schulz, J. Upham, I. De Leon, and R. W. Boyd, "Large optical nonlinearity of nanoantennas coupled to an epsilon-near-zero material," *Nat. Photonics* **12**(2), 79–83 (2018).
18. M. Ferrera and E. G. Carnemolla, "Ultra-fast transient plasmonics using transparent conductive oxides," *J. Opt.* **20**(2), 024007 (2018).
19. Y. Zhou, M. Z. Alam, M. Karimi, J. Upham, O. Reshef, C. Liu, A. E. Willner, and R. W. Boyd, "Broadband frequency translation through time refraction in an epsilon-near-zero material," *Nat. Commun.* **11**(1), 2180 (2020).
20. A. Mock, "Optical Isolation using Compact Time-modulated Cavity Array," in *2020 International Applied Computational Electromagnetics Society Symposium (ACES)*, (2020), pp. 1–2.
21. J. Bohn, T. S. Luk, C. Tollerton, S. W. Hutchings, I. Brener, S. Horsley, W. L. Barnes, and E. Hendry, "All-optical switching of an epsilon-near-zero plasmon resonance in indium tin oxide," *Nat. Commun.* **12**(1), 1017 (2021).
22. R. Sabri, M. M. Salary, and H. Mosallaei, "Broadband continuous beam-steering with time-modulated metasurfaces in the near-infrared spectral regime," *APL Photonics* **6**(8), 086109 (2021).
23. E. Galiffi, S. Yin, and A. Alù, "Tapered photonic switching," *Nanophotonics* **11**(16), 3575–3581 (2022).
24. G. Castaldi, V. Pacheco-Peña, M. Moccia, N. Engheta, and V. Galdi, "Exploiting space-time duality in the synthesis of impedance transformers via temporal metamaterials," *Nanophotonics* **10**(14), 3687–3699 (2021).
25. D. Ramaccia, A. Alù, A. Toscano, and F. Bilotti, "Temporal multilayer structures for designing higher-order transfer functions using time-varying metamaterials," *Appl. Phys. Lett.* **118**(10), 101901 (2021).
26. O. Silbiger and Y. Hadad, "Optimization-Free Filter and Matched-Filter Design through Spatial and Temporal Soft Switching of the Dielectric Constant," *Phys. Rev. Appl.* **19**(1), 014047 (2023).
27. C. Rizza, G. Castaldi, and V. Galdi, "Spin-controlled photonics via temporal anisotropy," *Nanophotonics* (2023).
28. C. Rizza, G. Castaldi, and V. Galdi, "Short-Pulsed Metamaterials," *Phys. Rev. Lett.* **128**(25), 257402 (2022).
29. Z. Li, R. Pestourie, Z. Lin, S. G. Johnson, and F. Capasso, "Empowering Metasurfaces with Inverse Design: Principles and Applications," *ACS Photonics* **9**(7), 2178–2192 (2022).
30. M. Bakr and N. Nikolova, "An Adjoint Variable Method for Time-Domain Transmission-Line Modeling With Fixed Structured Grids," *IEEE Trans. Microwave Theory Tech.* **52**(2), 554–559 (2004).
31. M. Bakr, A. Elsherbeni, and V. Demir, "An FDTD-based adjoint sensitivity approach," in *2016 IEEE International Symposium on Antennas and Propagation (APSURSI)*, (IEEE, Fajardo, PR, USA, 2016), pp. 2021–2022.
32. E. Hassan, E. Hassan, E. Hassan, E. Hassan, A. C. Lesina, and A. C. Lesina, "Topology optimization of dispersive plasmonic nanostructures in the time-domain," *Opt. Express* **30**(11), 19557–19572 (2022).
33. Computational Nanophotonics Videos, "Write your own 1D - FDTD program with python," (2021).
34. J. Baxter, "Code.zip," figshare 2023 <https://doi.org/10.6084/m9.figshare.21905262>.
35. J. Baxter, A. Pérez-Casanova, L. Cortes-Herrera, A. Calà Lesina, I. De Leon, and L. Ramunno, "Dynamic Nanophotonics in Epsilon-Near-Zero Conductive Oxide Films and Metasurfaces: A Quantitative, Nonlinear, Computational Model," *Advanced Photonics Research* **n/a**, 2200280. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/adpr.202200280>.
36. D. R. Smith, S. Schultz, P. Markoš, and C. M. Soukoulis, "Determination of effective permittivity and permeability of metamaterials from reflection and transmission coefficients," *Phys. Rev. B* **65**(19), 195104 (2002).
37. D. R. Smith, D. C. Vier, T. Koschny, and C. M. Soukoulis, "Electromagnetic parameter retrieval from inhomogeneous metamaterials," *Phys. Rev. E* **71**(3), 036617 (2005).
38. T. T. Koutserimpas, A. Alù, and R. Fleury, "Parametric amplification and bidirectional invisibility in PT-symmetric time-Floquet systems," *Phys. Rev. A* **97**(1), 013839 (2018).
39. J. B. Pendry, E. Galiffi, and P. A. Huidobro, "Gain in time-dependent media—a new mechanism," *J. Opt. Soc. Am. B* **38**(11), 3360–3366 (2021).
40. J. T. Mendonça, A. M. Martins, and A. Guerreiro, "Temporal beam splitter and temporal interference," *Phys. Rev. A* **68**(4), 043801 (2003).
41. Y. Xiao, G. P. Agrawal, and D. N. Maywar, "Spectral and temporal changes of optical pulses propagating through time-varying linear media," *Opt. Lett.* **36**(4), 505 (2011).
42. O. Reshef, E. Giese, M. Z. Alam, I. D. Leon, J. Upham, and R. W. Boyd, "Beyond the perturbative description of the nonlinear optical response of low-index materials," *Opt. Lett.* **42**(16), 3225–3228 (2017).

43. P. Guo, R. D. Schaller, J. B. Ketterson, and R. P. H. Chang, "Ultrafast switching of tunable infrared plasmons in indium tin oxide nanorod arrays with large absolute amplitude," *Nat. Photonics* **10**(4), 267–273 (2016).
44. "Autograd," (2022). Original-date: 2014-11-24T15:50:23Z.

Inverse design of optical pulse shapes for time-varying photonics: supplement

JOSHUA BAXTER^{1,2}  AND LORA RAMUNNO^{1,*}

¹Department of Physics, and Nexus for Quantum Technologies Institute, University of Ottawa, Ottawa, Canada

²jbaxt089@uottawa.ca

*Lora.Ramunno@uottawa.ca

This supplement published with Optica Publishing Group on 22 June 2023 by The Authors under the terms of the [Creative Commons Attribution 4.0 License](#) in the format provided by the authors and unedited. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Supplement DOI: <https://doi.org/10.6084/m9.figshare.22701109>

Parent Article DOI: <https://doi.org/10.1364/OE.485902>

Inverse design of optical pulse shapes for time-varying photonics: supplemental document

JOSHUA BAXTER, AND LORA RAMUNNO

In this document we derive adjoint sensitivity analysis in the time-domain and apply it to time-varying photonic systems with and without dispersion. The final two sections are demonstrations of our pulse shaping in time method applied to static materials.

A. Adjoint method in the time domain

In this section, we provide a derivation of the adjoint method in the time domain for completeness, following Ref. [1]. We begin with Maxwell's equations, which can be written as,

$$A\mathbf{x} + B\mathbf{x} = \mathbf{s} \quad (\text{S1})$$

where A and B are the system matrices, $\mathbf{x}$ is the electromagnetic field vector, and $\mathbf{s}$ is the source vector. In order to obtain the gradient, we differentiate Eq. S1 with respect to some tunable parameter p_i ,

$$\frac{\partial A}{\partial p_i}\mathbf{x} + A\frac{\partial \mathbf{x}}{\partial p_i} + \frac{\partial B}{\partial p_i}\mathbf{x} + B\frac{\partial \mathbf{x}}{\partial p_i} = \frac{\partial \mathbf{s}}{\partial p_i}. \quad (\text{S2})$$

In order to scalarize this equation, we left-multiply by an adjoint electromagnetic field $\mathbf{x}_{adj}$ vector (to be specified later),

$$\mathbf{x}_{adj}^T A \frac{\partial \mathbf{x}}{\partial p_i} + \mathbf{x}_{adj}^T B \frac{\partial \mathbf{x}}{\partial p_i} = \mathbf{x}_{adj}^T \mathbf{R}_i, \quad (\text{S3})$$

where we have introduced the residue vector

$$\mathbf{R}_i = \frac{\partial \mathbf{s}}{\partial p_i} - \frac{\partial A}{\partial p_i}\mathbf{x} - \frac{\partial B}{\partial p_i}\mathbf{x}. \quad (\text{S4})$$

Finally, we time-integrate Eq. S3 over the simulation ($t = 0$ to $t = T_m$),

$$\int_0^{T_m} (-\mathbf{x}_{adj}^T A + \mathbf{x}_{adj}^T B) \frac{\partial \mathbf{x}}{\partial p_i} dt = \int_0^{T_m} \mathbf{x}_{adj}^T \mathbf{R}_i dt. \quad (\text{S5})$$

making the assumption that $\mathbf{x}$ and $\mathbf{x}_{adj}$ are negligible outside this time window. The bracketed expression within the left-hand integrand resembles the left hand side Maxwell's equation S1, but here for the adjoint field. Taking the transpose of the bracketed term, and setting it equal to $\partial\psi/\partial\mathbf{x}$, we obtain

$$-A^T \mathbf{x}_{adj} + B^T \mathbf{x}_{adj} = \frac{\partial\psi}{\partial\mathbf{x}}, \quad (\text{S6})$$

then Eq. S5 becomes

$$\int_0^{T_m} \left(\frac{\partial\psi}{\partial\mathbf{x}} \right)^T \frac{\partial \mathbf{x}}{\partial p_i} dt = \int_0^{T_m} \mathbf{x}_{adj}^T \mathbf{R}_i dt. \quad (\text{S7})$$

From Eq. S7, we see that the integrand on the left hand side is simply the chain-rule expansion of $\partial\psi/\partial p_i$ and thus we have our gradient of the objective function defined in Eq. 1 of the main text,

$$\frac{\partial F}{\partial p_i} = \frac{\partial}{\partial p_i} \int_0^{T_m} \psi(\mathbf{x}, t) dt = \int_0^{T_m} \mathbf{x}_{adj}^T \mathbf{R}_i dt. \quad (\text{S8})$$

We can calculate $\mathbf{R}_i$ using forward simulation data obtained by simulating S1. The adjoint field is calculated using the time reversal of Eq. S6

$$A^T \mathbf{x}_{adj} + B^T \mathbf{x}_{adj} = \frac{\partial\psi}{\partial\mathbf{x}}(T_m - t), \quad (\text{S9})$$

which ensures the sign of each term has the same sign as Eq. S1. Note that if A and B are time-varying they must also be time-reversed.

Eq. S9 is the general formula for the adjoint simulations. For most static, optical modelling, A is diagonal and B is symmetric, and thus the adjoint simulator is identical to the forward simulator with a different source. However, for more complicated problems (as we shall see in the following subsections), the matrices are not symmetric and the simulation software must be adapted.

A.1. Time-varying permittivity

Let us consider a non-magnetic medium with a permittivity that can vary in both space and time but is otherwise dispersionless. In this case, Maxwell's equations (Eq. S1) can be written as

$$\begin{bmatrix} \epsilon_0 \epsilon(t) & 0 \\ 0 & -\mu_0 \end{bmatrix} \frac{\partial}{\partial t} \begin{bmatrix} \mathbf{E} \\ \mathbf{H} \end{bmatrix} + \begin{bmatrix} \epsilon_0 \dot{\epsilon}(t) & -\nabla \times \\ -\nabla \times & 0 \end{bmatrix} \begin{bmatrix} \mathbf{E} \\ \mathbf{H} \end{bmatrix} = \begin{bmatrix} -\mathbf{J}_s(t) \\ \mathbf{M}_s(t) \end{bmatrix}, \quad (\text{S10})$$

Implicit in this algebraic notation is the assumption that Eq. S10 will be solved on a discretized spatial grid with, say, N cells. Then the electric ($\mathbf{E}$) and magnetic ($\mathbf{H}$) fields are themselves column vectors of length $3N$, where the factor of 3 comes from the three components of $\mathbf{E}$ and $\mathbf{H}$; the matrix elements (ie. $\epsilon_0 \epsilon(t)$, $\epsilon_0 \dot{\epsilon}(t)$, $-\mu_0$, 0, and $-\nabla \times$) are themselves $3N \times 3N$ sub-matrices. Since we are considering non-magnetic materials, the $-\mu_0$ sub-matrix is diagonal, with each diagonal element having the value of the (negative of the) permeability of free space. The permittivity sub-matrix $\epsilon_0 \epsilon(t)$ need not be diagonal, but in our examples below we assume isotropic materials, so it is taken as diagonal. The sub-matrix $\nabla \times$ is of the form

$$\nabla \times \mathbf{E} = \begin{bmatrix} 0 & -\partial/\partial z & \partial/\partial y \\ \partial/\partial z & 0 & -\partial/\partial x \\ -\partial/\partial y & \partial/\partial x & 0 \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix}, \quad (\text{S11})$$

where each partial derivative is a $N \times N$ finite difference matrix. The $\epsilon_0 \dot{\epsilon}(t)$ sub-matrix arises naturally in Maxwell's equations for a time-dependent permittivity.

The quantities that define the input pulse are $\mathbf{J}_s$ and $\mathbf{M}_s$, the time-dependent external current density and magnetization sources, respectively, that are both represented as column vectors of length $3N$. It is the shape of $\mathbf{J}_s$ and/or $\mathbf{M}_s$ that we seek to tune in order to maximize our objective. Note that in the following, we will assume the external source is non-magnetic, so we set $\mathbf{M}_s = 0$.

The corresponding adjoint matrix equation (Eq. S9) is given by

$$\begin{bmatrix} \epsilon_0 \epsilon(T_m - t) & 0 \\ 0 & -\mu_0 \end{bmatrix} \frac{\partial}{\partial t} \begin{bmatrix} \mathbf{E}^{adj} \\ \mathbf{H}^{adj} \end{bmatrix} + \begin{bmatrix} -\epsilon_0 \dot{\epsilon}(T_m - t) & -\nabla \times \\ -\nabla \times & 0 \end{bmatrix} \begin{bmatrix} \mathbf{E}^{adj} \\ \mathbf{H}^{adj} \end{bmatrix} = \begin{bmatrix} -\mathbf{J}_s^{adj}(t) \\ 0 \end{bmatrix}, \quad (\text{S12})$$

where the external source is taken to be

$$\mathbf{J}_s^{adj} = -\frac{\partial \psi}{\partial \mathbf{E}}(T_m - t). \quad (\text{S13})$$

Here we have assumed that ψ depends only on $\mathbf{E}$. Since we choose the functional form of $\psi(\mathbf{E})$, we can derive an analytic expression for $\partial \psi / \partial \mathbf{E}$ as a function of $\mathbf{E}$, where $\mathbf{E}$ is to be calculated from Eq. S10 (the forward simulation). For general objective functions, ψ could also be a function of $\mathbf{H}$ and thus the adjoint problem may also contain a magnetic source $\mathbf{M}_s^{adj} = \frac{\partial \psi}{\partial \mathbf{H}}(T_m - t)$, though we do not consider this here. Recent implementations of adjoint sensitivity analysis use automatic differentiation to calculate $\partial \psi / \partial \mathbf{E}$ and $\partial \psi / \partial \mathbf{H}$ for user-defined objective functions [2, 3].

For the case where ϵ is not time-varying, $\dot{\epsilon} = 0$, and the adjoint problem defined by Eq. S12 is identical to the forward problem of Eq. S10, the only difference being the external current density source in the adjoint simulation is given by S13. However, in time-varying materials, the adjoint

problem is not the same as the forward one, as $\varepsilon(t)$ is time-reversed and $\dot{\varepsilon}$ is time-reversed and changes sign. This requires a small but straightforward change to the simulation software.

Once the forward and adjoint problems are computed, one can use the forward and adjoint fields to compute the gradient, Eq. S8 via

$$\frac{\partial F}{\partial p_i} = \int_0^{T_m} \begin{bmatrix} \mathbf{E}^{adj}(T_m - t) \\ \mathbf{H}^{adj}(T_m - t) \end{bmatrix} \cdot \mathbf{R}_i(t) dt, \quad (\text{S14})$$

where the integrand is the scalar product of the time-reversed adjoint fields and the residue vector

$$\mathbf{R}_i(t) = - \left(\frac{\partial}{\partial p_i} \begin{bmatrix} \varepsilon_0 \varepsilon(t) & 0 \\ 0 & -\mu_0 \end{bmatrix} \right) \frac{\partial}{\partial t} \begin{bmatrix} \mathbf{E}(t) \\ \mathbf{H}(t) \end{bmatrix} + \left(\frac{\partial}{\partial p_i} \begin{bmatrix} -\varepsilon_0 \dot{\varepsilon}(t) & \nabla \times \\ \nabla \times & 0 \end{bmatrix} \right) \begin{bmatrix} \mathbf{E}(t) \\ \mathbf{H}(t) \end{bmatrix} + \frac{\partial}{\partial p_i} \begin{bmatrix} -\mathbf{J}_s(t) \\ 0 \end{bmatrix}. \quad (\text{S15})$$

Note that this formalism allows for simultaneous optimization of both input pulse shape and topology (the latter being somewhat computationally expensive in the time-domain). For topology optimization, where we tune ε over many spatial grid cells, only the first term is required for a static ε , but both the first and second terms are required for a time-dependent ε . However, for the optimization of pulse shape for a given topology, we need only consider the third term. As this is our interest here, we thus set

$$\mathbf{R}_i(t) = \frac{\partial}{\partial p_i} \begin{bmatrix} -\mathbf{J}_s(t) \\ 0 \end{bmatrix}, \quad (\text{S16})$$

so that no forward fields are required to obtain the residue, unlike for topology optimization. However, the adjoint fields (in particular, the electric adjoint field) are required at the spatial locations from which the source $\mathbf{J}_s(t)$ is injected.

In time-domain topology optimization, the time-domain fields must be kept at all spatial locations over which topology optimization is to occur, typically resulting in large memory and I/O requirements that can present a limitation, especially for simulations involving long time-domain signals. The time-domain fields are required for the calculation of the first two terms of the residue in Eq. S15, and are in general over a 3D volume [4].

For pulse shaping, in contrast, fields are not required for the residue $\mathbf{R}_i(t)$ calculation. Rather, $\mathbf{R}_i(t)$ only depends on the external source, which can be further parameterized and is known analytically. Thus, the only field data that absolutely needs to be kept from the forward simulation are the fields at the locations at which the objective function is defined. If this happens to be only a single or few points, then the memory and I/O requirements are, in fact, modest. If the objective function is defined over a significant number of points, then memory and I/O requirements may again start to present a limitation, though not as severely as for 3D time-domain topology optimization; typically objective functions are defined on at most a 2D sub-space, such as a transmission plane.

A.2. Time-varying dispersion

Materials with time-varying optical properties can also be described by dispersive models, such as the Drude and Lorentz models, and many others [5, 6], described by auxiliary differential equations for additional vector fields, such as the current density or polarization fields. While our general method can apply to many such time-domain dispersive models, we consider explicitly here only the Drude model given by

$$\frac{\partial \mathbf{J}}{\partial t} + \gamma \mathbf{J} - \varepsilon_0 \omega_p^2(t) \mathbf{E} = 0, \quad (\text{S17})$$

where $\mathbf{J}$ is the free electron current density inside the medium, γ is the Drude damping coefficient, and ω_p is the plasma frequency. By allowing the plasma-frequency to be time-varying (see, for

example, Ref. [7]), we are allowing for the material to have a time-varying dispersion.

For our forward simulation, Eq. S17 is solved numerically in the time-domain along with Maxwell's equations (Eq. S1):

$$\begin{bmatrix} \varepsilon_0 \varepsilon & 0 & 0 \\ 0 & -\mu_0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \frac{\partial}{\partial t} \begin{bmatrix} \mathbf{E} \\ \mathbf{H} \\ \mathbf{J} \end{bmatrix} + \begin{bmatrix} 0 & -\nabla \times & 1 \\ -\nabla \times & 0 & 0 \\ -\varepsilon_0 \omega_p^2(t) & 0 & \gamma \end{bmatrix} \begin{bmatrix} \mathbf{E} \\ \mathbf{H} \\ \mathbf{J} \end{bmatrix} = \begin{bmatrix} -\mathbf{J}_s(t) \\ 0 \\ 0 \end{bmatrix}, \quad (\text{S18})$$

The topology is defined not only in ε (which we take to be static here, but it need not be), but also in the sub-matrices arising from the Drude model (bottom row of the two system matrices), which vanish outside the dispersive medium. In other words, the Drude model is not solved outside the dispersive medium.

The time-domain dispersive adjoint system equation (from Eq. S9) is

$$\begin{bmatrix} \varepsilon_0 \varepsilon & 0 & 0 \\ 0 & -\mu_0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \frac{\partial}{\partial t} \begin{bmatrix} \mathbf{E}^{adj} \\ \mathbf{H}^{adj} \\ \mathbf{J}^{adj} \end{bmatrix} + \begin{bmatrix} 0 & -\nabla \times & \omega_p^2(T_m - t)/\omega_{p0}^2 \\ -\nabla \times & 0 & 0 \\ -\varepsilon_0 \omega_{p0}^2 & 0 & \gamma \end{bmatrix} \begin{bmatrix} \mathbf{E}^{adj} \\ \mathbf{H}^{adj} \\ \mathbf{J}^{adj} \end{bmatrix} = \begin{bmatrix} -\mathbf{J}_s^{adj}(t) \\ 0 \\ 0 \end{bmatrix}, \quad (\text{S19})$$

where, as before, $\mathbf{J}_s^{adj} = -\frac{\partial \psi}{\partial \mathbf{E}}(T_m - t)$ and where we defined the quantity ω_{p0} to be a reference plasma frequency, for example, taken at the beginning of the forward simulation $\omega_p(t=0)$; note that we used this quantity to define $\mathbf{J}^{adj}$ such that it has units of current density.

For static dispersion, the adjoint problem in Eq. S19 is very similar to the forward problem in Eq. S18; its implications in topology optimization have been recently investigated [4]. For time-varying dispersion, the second system matrix of the adjoint problem in Eq. S19 is fundamentally different than that of the forward problem of Eq. S18. As in the previous section, performing the adjoint simulation requires a small modification in the simulation software. The time-varying plasma frequency of the forward simulation must also be time-reversed in the adjoint simulation.

Once the forward and adjoint problems are computed, one can use the forward and adjoint fields to compute the gradient of the objective function via

$$\frac{\partial F}{\partial p_i} = \int_0^{T_m} \begin{bmatrix} \mathbf{E}^{adj}(T_m - t) \\ \mathbf{H}^{adj}(T_m - t) \\ \mathbf{J}^{adj}(T_m - t)/(\varepsilon_0 \omega_{p0}^2) \end{bmatrix} \cdot \mathbf{R}_i(t) dt, \quad (\text{S20})$$

where the integrand is the scalar product between the time-reversed adjoint fields and the residue vector, now given by

$$\mathbf{R}_i(t) = - \left(\frac{\partial}{\partial p_i} \begin{bmatrix} \varepsilon_0 \varepsilon & 0 & 0 \\ 0 & -\mu_0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right) \frac{\partial}{\partial t} \begin{bmatrix} \mathbf{E}(t) \\ \mathbf{H}(t) \\ \mathbf{J}(t) \end{bmatrix} + \left(\frac{\partial}{\partial p_i} \begin{bmatrix} 0 & \nabla \times & -1 \\ \nabla \times & 0 & 0 \\ \varepsilon_0 \omega_p^2(t) & 0 & \gamma \end{bmatrix} \right) \begin{bmatrix} \mathbf{E}(t) \\ \mathbf{H}(t) \\ \mathbf{J}(t) \end{bmatrix} + \frac{\partial}{\partial p_i} \begin{bmatrix} -\mathbf{J}_s(t) \\ 0 \\ 0 \end{bmatrix}, \quad (\text{S21})$$

Again, for pulse shape optimization, only the third term is considered, so the residue is given by Eq. S16. This means that, again, no forward fields are required; $\mathbf{E}^{adj}$ is required only at the spatial locations where $\mathbf{J}_s(t)$ is non-zero.

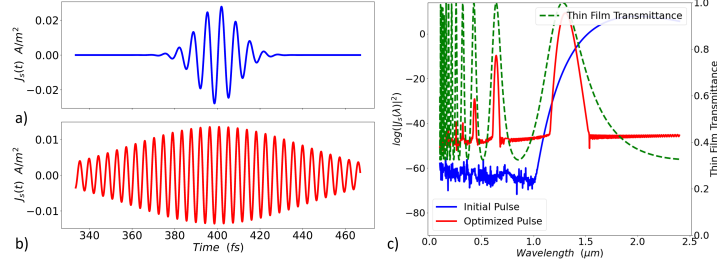


Fig. S1. a) Initial Gaussian input pulse with duration 13 fs centered at $2 \mu\text{m}$. b) Optimized pulse after 90 iterations. c) Log plot of pulse spectrum for the initial Gaussian pulse (blue line - left vertical axis), and the optimized pulse (red line - left vertical axis), overlayed with the transmittance spectrum of the dielectric thin film (green dashed line - right vertical axis).

B. Pulse shaping in static materials

In this section we apply our pulse shaping in time method from Section 3 to maximize the transmitted energy through static dielectric and metallic thin films.

B.1. Test case 1: Static, dispersionless thin film

As a test case, we consider a 200 nm dielectric thin film with static permittivity $\epsilon = 10.25$. As plotted in Fig. S1a, our initial pulse is a Gaussian with a central wavelength of $\lambda_0 = 2 \mu\text{m}$ and pulse duration $\tau = 13$ fs. The center wavelength is purposefully chosen to be in a spectral range of low transmittance through the dielectric film. The evolution of the objective function during the optimization is plotted in Fig. S2, showing a $2.6\times$ enhancement in transmitted energy after only 90 iterations of our pulse shaping algorithm.

The optimized pulse is plotted in Fig. S1b, and has a much larger pulse width than the initial pulse (only a portion of the pulse is shown here, in the same time window as the initial pulse in Fig. S1a). In Fig. S1c, the spectrum of the initial pulse (solid blue line), and optimized pulse (solid red line) are plotted corresponding to the left-vertical axis, with the transmittance spectrum of the dielectric thin film (dashed green line) corresponding to the right-vertical axis. Here we can see that, indeed, the spectrum of the initial pulse is in a bandwidth of low transmittance. The pulse shaping algorithm added frequency components centered at the Fabry-Perot resonances of the thin film, thus achieving the goal of increasing the transmitted energy without increasing the pulse energy. The large width of the optimized pulse is due to the optimizer keeping the bandwidth narrow within each resonance to ensure a higher transmission.

B.2. Test case 2: Static, Drude-metal thin film

As a second test case, we will demonstrate pulse shaping in time with a 200nm metallic thin film with an optical response described by the Drude model (Eq. S17), where $\omega_p = 3 \times 10^{15}$ rad/s is the static plasma frequency, $\gamma = 2 \times 10^{14}$ rad/s, and with a background permittivity of $\epsilon = 4$. As plotted in Fig. S3a, our initial pulse is a Gaussian with a central wavelength of $\lambda_0 = 2 \mu\text{m}$ and duration $\tau = 13$ fs. Once again, the center wavelength is purposefully chosen to be in a spectral range of low transmittance through the metal film. After 40 iterations of pulse shaping in time, we achieve a transmitted energy that is $12\times$ higher than for the initial pulse. The optimized pulse is plotted in Fig. S3b.

Fig. S3c shows a log plot of the frequency spectra of the initial pulse (solid blue line) and the optimized pulse (solid red line), which correspond to the left-vertical axis. The green, dashed line is the transmittance of the thin, metallic film. We again see that the new pulse has gained frequency components at the Fabry-Perot resonances of the thin film.

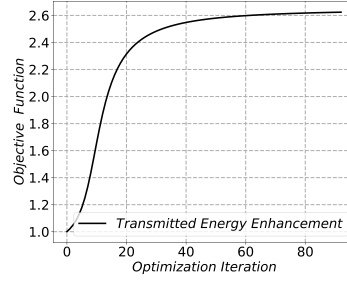


Fig. S2. Objective function evolution during optimization procedure for a static, dispersionless thin film.

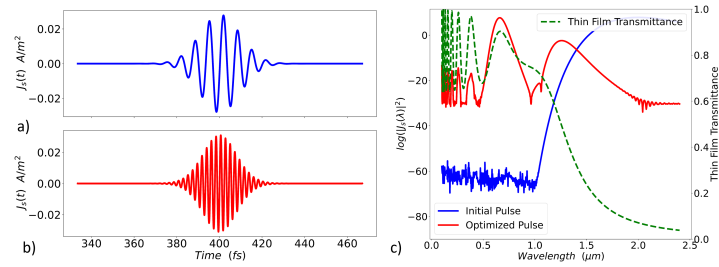


Fig. S3. a) Initial Gaussian input pulse with duration 13 fs centered at 2 μm . b) Optimized pulse after 40 iterations. c) Log plot pulse spectrum of the initial Gaussian (blue line - left vertical axis), the optimized pulse (red line - right vertical axis) overlaid with the transmittance spectrum of the metallic thin film (green dashed line - right vertical axis).

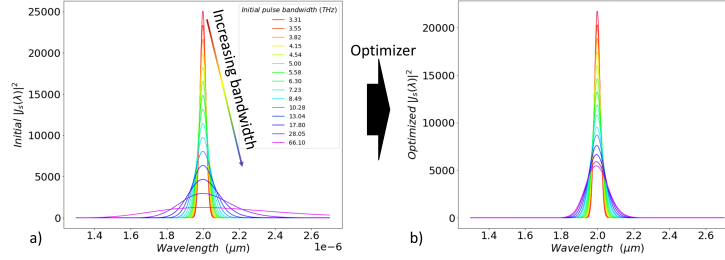


Fig. S4. Study on the effect of initial bandwidth. a) Initial pulse intensity spectra with increasing bandwidths (labelled in the legend). b) Optimized pulse spectra for the corresponding initial pulses (same color) in (a).

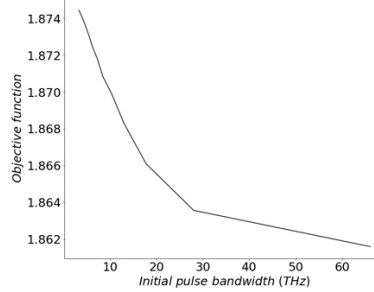


Fig. S5. Final objective function (transmitted pulse energy) vs. initial pulse bandwidth. Despite the differing bandwidths, each optimized pulse gains approximately the same amount of energy. The optimizer had little to gain from further compressing the bandwidth.

C. Effect of pulse bandwidth and phase on amplification

In Section 3.2 of the main text we show how a shaped pulse experiences gain when the modulation frequency is twice that of the incident pulse. In this section we show that the pulse bandwidth found by the optimizer aligns with the amplification bandwidth of the system, and we show how the phase relationship between the pulse and modulation plays an important role in amplification.

We determined the amplification bandwidth by optimizing the pulse shape for several different initial pulses with increasing bandwidths as plotted in Fig. S4a for the system described in Section 3.2 of the main text. The optimized pulse spectra plotted in Fig. S4b are all found within the same bandwidth ± 150 nm of the center wavelength, indicating the amplification bandwidth. While narrower bandwidth initial pulses result in narrower bandwidth optimized pulses, the larger bandwidth initial pulses get significantly narrower, eventually converging to a fixed bandwidth. Despite the variance in optimized pulse bandwidth, the final objective function (the normalized transmitted pulse energy) is negligibly affected by the initial bandwidth, as plotted in Fig. S5. This confirms that all pulses plotted in Fig. S4b have approximately the same energy gain and thus lie within the amplification bandwidth.

Next, we investigate the phase relationship between the input pulse and the sinusoidal permittivity modulation for the system described Section 3.2 of the main text. We run several simulations, each with a different phase ϕ of the modulation signal $\omega_p(t) \sim \sin(2\omega_0 + \phi)$, which is varied from 0 to 2π . We used the same initial and optimized pulses as for Section 3.2, where the optimized pulse was calculated for a permittivity modulation phase of zero. The effect of ϕ

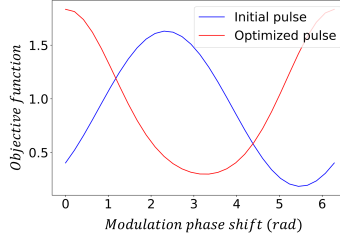


Fig. S6. The effect of modulation phase on pulse transmission through a time-varying film.

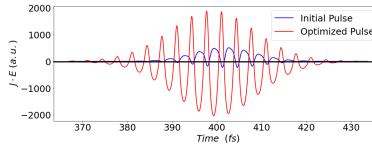


Fig. S7. Rate of energy transferred from the optical pulse to the free electron system for the initial (blue line) and optimized (red line) pulses. Clearly the optimized pulse gains more energy from the free electrons than it gives. The opposite is observed for the initial pulse.

on the final objective function (transmitted pulse energy) is plotted in Fig. S6 for both the initial pulse (blue line) and optimized pulse (red line). The results show that, for both the initial and optimized incident pulses, there is a periodic relationship between the modulation phase and the transmitted energy. Notably, the optimized pulse reaches a higher maximum gain (at a phase shift of zero, where it was optimized) than the on-phase un-optimized pulse, which shows its maximum gain at a phase shift of around 2.2 radians. The major effect of pulse shaping here is to correct the phase, but clearly the pulse shaping had a more complicated effect as the maximum gain of the optimized pulse has increased.

D. Energy conservation in pulse amplification

In this section we confirm that energy in the system of Section 3.2 of the main text is conserved. In a dispersionless time-varying system with permittivity $\epsilon(t)$, energy gained by the pulse (or increased loss) can be attributed to an additional term found in Poynting's theorem proportional to $\partial\epsilon(t)/\partial t$ [8].

In our case, we found it more convenient to assess energy conservation by studying the $\mathbf{J} \cdot \mathbf{E}$ term in Poynting's theorem as both the current density and electric field are immediately available to us from the FDTD simulations. This quantity describes the rate at which fields do work on the charges in a volume, that is, the rate of electromagnetic energy density absorbed by the free electrons that constitute the current $\mathbf{J}$ at time t and at position $\mathbf{x}$. By integrating over the volume, we get the rate of energy absorbed by the electrons at a given time,

$$w_{abs} = \int \mathbf{J}(\mathbf{x}, t) \cdot \mathbf{E}(\mathbf{x}, t) d\mathbf{x}. \quad (\text{S22})$$

In a lossy system, this is expected to be positive. We plot this quantity in Fig. S7 for the initial (blue line) and optimized (red line) pulses. We can see that the initial pulse finds itself losing energy to the free electrons for most of its existence ($w_{abs}(t) > 0$), whereas the optimized pulse is shaped in such a way that the pulse is gaining energy from the free electrons for most of its existence ($w_{abs}(t) < 0$). This agrees with our amplification result.

To demonstrate energy conservation, we must show that the energy of the reflected, transmitted,

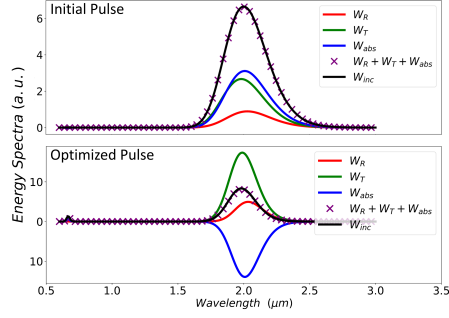


Fig. S8. Energy spectrum for the reflected (red line), transmitted (green line), and absorbed (blue line) energy. These energies are totalled (purple x) and are shown to be equivalent to the initial pulse energy spectrum (black line) demonstrating energy conservation of the overall system.

and absorbed fields are equal to the energy of the optical pulse. We accomplish this in the frequency domain by Fourier transforming both the electric and current density fields. The total absorbed power (in the frequency domain) is given by [9],

$$W_{abs} = \frac{1}{2} \int \text{Re}(\mathbf{J}^*(\mathbf{x}, \omega) \cdot \mathbf{E}(\mathbf{x}, \omega)) d\mathbf{x}, \quad (\text{S23})$$

where the $*$ represents the complex conjugate. The total reflected and transmitted power are given by the frequency-domain Poynting vector, which simplifies to

$$W_{R/T} = \frac{1}{Z_0} |\mathbf{E}(\omega, \mathbf{x}_{R/T})|^2, \quad (\text{S24})$$

where the subscript R/T represents reflected/transmitted waves. The incident energy spectrum is calculated similarly. We plot these energy spectra in Fig. S8 for the initial (top plot) and optimized (bottom plot) pulses. First of all, in both situations it is clear that energy is conserved, $W_R + W_T + W_{abs} = W_{inc}$. Secondly, we can see now that the energy gain is fully accounted for in the negative absorbed energy (energy given to the pulse by free electrons).

REFERENCES

1. M. Bakr, A. Elsherbeni, and V. Demir, *Adjoint Sensitivity Analysis of High Frequency Structures with MATLAB®* (Institution of Engineering and Technology, 2017).
2. A. M. Hammond, A. Oskooi, M. Chen, Z. Lin, S. G. Johnson, and S. E. Ralph, "High-performance hybrid time/frequency-domain topology optimization for large-scale photonics inverse design," *Opt. Express* **30**, 4467 (2022).
3. T. W. Hughes, I. A. D. Williamson, M. Minkov, and S. Fan, "Forward-Mode Differentiation of Maxwell's Equations," *ACS Photonics* **6**, 3010–3016 (2019). Publisher: American Chemical Society.
4. E. Hassan, E. Hassan, E. Hassan, E. Hassan, A. C. Lesina, and A. C. Lesina, "Topology optimization of dispersive plasmonic nanostructures in the time-domain," *Opt. Express* **30**, 19557–19572 (2022). Publisher: Optica Publishing Group.
5. A. Taflov and S. C. Hagness, *Computational electrodynamics: the finite-difference time-domain method*, Artech House antennas and propagation library (Artech House, Boston, 2005), 3rd ed.
6. A. Taflov, A. Oskooi, and S. G. Johnson, eds., *Advances in FDTD computational electrodynamics: photonics and nanotechnology*, Artech House antennas and propagation series (Artech House, Boston, 2013). OCLC: ocn811964793.
7. J. Baxter, A. Pérez-Casanova, L. Cortes-Herrera, A. Calà Lesina, I. De Leon, and L. Ramunno, "Dynamic Nanophotonics in Epsilon-Near-Zero Conductive Oxide Films and Metasurfaces:

A Quantitative, Nonlinear, Computational Model," Adv. Photonics Res. **n/a**, 2200280. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/adpr.202200280>.

8. T. T. Koutserimpas and R. Fleury, "Electromagnetic fields in a time-varying medium: Exceptional points and operator symmetries," IEEE Transactions on Antennas Propag. **66**, 5300–5307 (2018). ArXiv:2003.03188 [physics].
9. J. D. Jackson, *Classical electrodynamics* (Wiley, New York, 1999), 3rd ed.

Chapter 7

Plasmonic colours predicted by deep learning

In this chapter, we apply the content introduced in Chapters [1.3](#) and [2.4](#) to the problem of predicting colours reflected from pulsed laser-generated plasmonic metasurfaces using deep learning. The inverse problem is also solved using deep learning to predict the conditions required to generate a specific colour. This paper was spawned by the work of Dr. Jean-Michel Guay (third author), who used pulsed laser ablation and re-deposition to generate plasmonic colours on noble metal surfaces [[163](#), [164](#)].

As an undergraduate, my first research position was assisting the research of (at the time) Ph.D. candidate Jean-Michel Guay investigating laser-induced plasmonic colours on blank silver coins. The colours are the result of plasmonic particle distributions that were induced by laser irradiation. Much of the work was done in the labs of the Royal Canadian Mint, where we measured the gamut of available colours by sweeping the pulsed laser settings (discussed in the article). The colours were measured via a colour meter which quantifies the *lightness*, *chroma*, and *hue* (LCH). As the colours were reproducible, this data was stored in a spreadsheet as a reference.

The reproducibility of the colours led Jean-Michel to collaborate with the Computational Nanophotonics group, which conducted a numerical study that mapped the nanoparticle distributions to the resulting colours. Therefore, from my undergraduate work, I was familiar with two related sets of complex data that can be used to train a supervised learning system; the experimental set relating laser settings to colour and the computational

set relating nanoparticle geometries to colour.

At the SOSCIP Impact Conference in May 2017, the idea of a machine learning-related project was discussed in a conversation with (now) Prof. Antonio Calá Lesina. SOSCIP, being a computing consortium, hosted many speakers presenting work applying machine learning to solve scientific problems. Although the memory is vague, it was here that the idea for this project was conceived. It would take several months for me to realize the ease of coding neural networks with frameworks like *TensorFlow*. Around the beginning of my graduate studies (2018 - 2019), the subject of deep learning was exploding in the field of nanophotonics, with the general theme centring around predicting the optical response of parameterized nanophotonic/plasmonic systems and solving the inverse problem using deep neural networks.

By the end of 2018, I had run a study on the prediction of colours from both datasets using various machine learning techniques, demonstrating that neural networks were vastly more accurate than other popular methods (such as linear/polynomial regression, K-nearest-neighbors, etc). Inspired by other papers at the time, I became interested in solving the inverse problem. I was unaware of numerical optimization and opted for a more ad-hoc approach, as outlined in the article. In hindsight, there are much more robust techniques for solving the inverse problem.

This paper piqued the interest of Dr. Sergey Gaponenko, whose research interests involve plasmonics applications to enhance optical processes such as luminescence, scattering, and optical nonlinearities. As mentioned in Ref. [12], he was interested in applying deep learning to predict the enhancement factors in complex plasmonic systems, citing this paper. The enhancement of photo-luminescence, for example, in the presence of plasmonic particles, depends on the enhancement of both the local electric field and the local density of photon states. These enhancement factors depend on many parameters, including the ambient environment, proximity of the emitter and plasmonic structure, etc. Furthermore, beyond simple systems like spherical plasmonic particles, or planar metallic surfaces, the calculation of the enhancement factors requires numerical calculations. A deep learning model can be a pseudo-model to predict plasmonic enhancement factors for complex systems such as plasmonic dimers and trimers. The collaboration was cut short due to circumstances on our collaborator's end.

SCIENTIFIC REPORTS

OPEN

Plasmonic colours predicted by deep learning

Joshua Baxter^{1,2}, Antonino Calà Lesina^{1,2,3}, Jean-Michel Guay^{1,2,4}, Arnaud Weck^{1,2,4}, Pierre Berini^{1,2,3} & Lora Ramunno^{1,2}

Received: 12 March 2019

Accepted: 17 May 2019

Published online: 30 May 2019

Picosecond laser pulses have been used as a surface colouring technique for noble metals, where the colours result from plasmonic resonances in the metallic nanoparticles created and redeposited on the surface by ablation and deposition processes. This technology provides two datasets which we use to train artificial neural networks, data from the experiment itself (laser parameters vs. colours) and data from the corresponding numerical simulations (geometric parameters vs. colours). We apply deep learning to predict the colour in both cases. We also propose a method for the solution of the inverse problem – wherein the geometric parameters and the laser parameters are predicted from colour – using an iterative multivariable inverse design method.

Deep learning (DL) refers to a sub-field of Machine Learning and Artificial Intelligence where multi-layered artificial neural networks, or deep neural networks (DNNs), are used for high accuracy predictions and classifications. Deep learning has been used in speech recognition, image classification, vision, and pattern extraction, to name a few examples¹. Its application in photonics is recent. For example, DL has been used to classify scanning electron microscope (SEM) images of nanostructures², to increase the resolution of microscopy³, and for the design of nanophotonic structures^{4–10}. DL models can be used to find patterns in complicated datasets and are particularly useful when large pre-calculated or pre-measured datasets are available for training. Both simulations and experiments in nanophotonics can provide the large datasets required for DL. For example, geometric parameter sweeps are used in plasmonics to optimize the design of nanostructures, while parameters are tuned in experiments to produce the best nanofabrication results. The exploration of the parameter space based on trials is time-consuming. Using data from parameter sweeps, one can train a DL model to predict the experimental or simulation outcome for a new set of input parameters, thus avoiding the need to run many trials of the actual experiment or computation. Further, in an industrial setting where the environment is not well controlled, environmental parameters can be tracked and can become additional inputs of a deep learning application.

An application which may benefit from DL is the plasmonic colouration of metal surfaces via laser machining. By controlling different laser parameters such as scanning speed, hatch spacing, burst number, etc., laser-machined surfaces (Fig. 1(a)) with a large range of colours (Fig. 1(b)) were created^{11,12}. Finite-difference time-domain (FDTD) simulations were performed based on the geometries inspired by the experimental SEM images and the measured nanoparticle distributions, and these simulations were able to explain experimental trends. This includes finding that the colour transition from blue to yellow occurs when the inter-particle spacing for nanoparticles of radius $R = 35$ nm is progressively increased. The nanoparticle distribution is modelled as a periodic hexagonal lattice of single-sized nanoparticles embedded into a metal surface (Fig. 1(c)), and the output of the simulation is a reflection spectrum from which the colour is rendered to the Red-Green-Blue (RGB) colour scale¹³ (Fig. 1(d)). Both experiments and simulations contain challenges which can be addressed by DL. On the experimental side, the colours generated by a new set of laser parameters must be determined empirically through trial and error. On the computational side, the requirement for accuracy means that a large-scale FDTD simulation has to be run in order to find the colour corresponding to a new set of geometric parameters, and this is time-consuming.

In this paper we apply DL to the direct prediction of new colours using both the experimental and simulation data. We predict colours from laser parameters by leveraging the dataset linking laser parameters to real, physical colours. We also predict colours from nanoparticle geometries by using the dataset linking the simulation

¹University of Ottawa, Department of Physics, Ottawa, K1N 6N5, Canada. ²University of Ottawa, Centre for Research in Photonics, Ottawa, K1N 6N5, Canada. ³University of Ottawa, School of Electrical Engineering and Computer Science, Ottawa, K1N 6N5, Canada. ⁴University of Ottawa, Department of Mechanical Engineering, Ottawa, K1N 6N5, Canada. Correspondence and requests for materials should be addressed to J.B. (email: jbaxt089@uottawa.ca) or A.C.L. (email: antonino.calalesina@uottawa.ca)

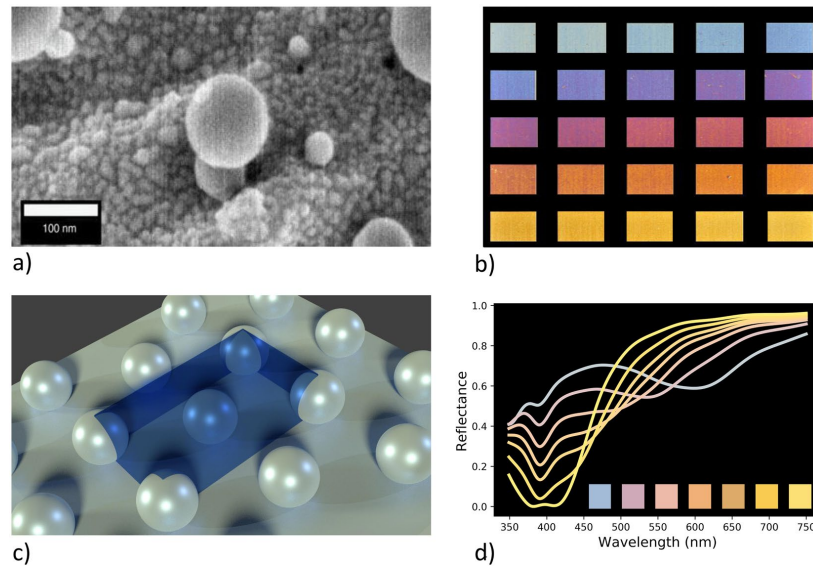


Figure 1. (a) High-magnification SEM image of a laser-machined silver surface. (b) Colour palette obtained by the laser-machining of silver surfaces¹¹. (c) Simplified model of the silver surface used in FDTD simulations (where the blue box is the unit cell). (d) Simulated reflectance spectra as a function of inter-nanoparticle distance. (a,b) Are used and modified (stretched and removed labels from b) with permission under the Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>).

geometries to computed colours. We then apply DNNs to aid in the solution of the inverse design problem, whereby inputting colours, the model returns appropriate geometric or laser parameters. In solving the inverse design problem, we use an iterative multivariable inverse design method in which the inverse solution is found by iteratively solving for individual outputs. To the best of our knowledge, this is the first time this method has been reported. Traditional optimization and inverse design techniques (e.g. genetic algorithms or gradient-based methods) have been applied to photonics applications¹⁴. These methods would be time consuming for the large-scale simulations, and costly to implement in the lab. This is why we highlight the importance of using the data that has already been accumulated to train efficient DL models.

Direct Prediction of Colours

Using simulation data. In this section we use DNNs for the prediction of colours from nanoparticle geometries using the FDTD dataset. The FDTD simulations were run to calculate the reflectance spectra from single-sized silver nanoparticle distributions embedded into flat silver surfaces.

Statistical analysis of SEM images of laser-coloured silver surfaces show a bimodal distribution of nanoparticle sizes, which are referred to as medium nanoparticles ($10 < R_m < 75$ nm) with an average radius of ≈ 35 nm, and small nanoparticles ($R_s < 10$ nm)^{11,12}. Simulations have shown that the effect of the small nanoparticles averages out, thus they do not contribute significantly to the colours observed in experiments^{11,12}. Indeed, the experimental palette (see Fig. 1(b)) can be reproduced with single-sized nanoparticles of radius 35 nm (see Fig. 1(d)). For this reason, we use the single-sized nanoparticle data for this study.

The particles are arranged in a hexagonal lattice in the x - z plane as shown in Fig. 2, where only a unit cell is simulated and periodic boundaries are placed in the x - and z - directions. We excited the system with a broadband plane wave pulsed signal which allows us to capture the full visible spectrum (broadband simulation) with a single run of the code. The incoming pulse was z -polarized and propagated in the y -direction, which is truncated by Convolutional Perfectly Matched Layers (CPMLs). Simulations are run with particle radii R ranging from 5 nm to 100 nm. The embedding depth R_{emb} represents the percentage of the radius embedded into the silver surface, and it ranges from 30 to 100%, where an embedding of 100% means that we simulate a hemisphere. The inter-particle spacing D_{cc} describes the center-to-center spacing between the nanoparticles. To remove correlation between spacing and radius, we also define an edge-to-edge spacing $D_{ee} = D_{cc} - 2R$ which ranges from 5 nm to 80 nm. The simulation domain was discretized using a uniform space step and a cubic Yee cell¹⁵ ($dx = dy = dz$). We used a space-step of $dx = 0.125$ nm for $R = 5$ to 10 nm, $dx = 0.25$ nm for $R = 15$ to 20 nm, $dx = 0.5$ nm for $R = 25$ to 45 nm and $dx = 1$ nm for $R = 50$ to 100 nm. The total number of Yee cells depends on the space step and geometry but is on the order of 10^8 . We run convergence tests to ensure sufficient accuracy¹⁶. The optical parameters of silver were modelled by the Drude + 2 Critical Points model^{17,18}. To simulate the metal substrate as a half plane, we used a

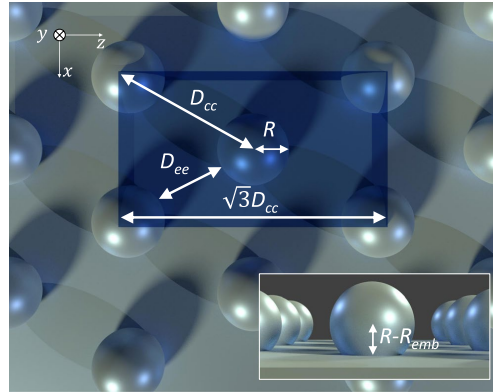


Figure 2. Layout of the nanoparticle geometry in the x - z plane. The blue tinted box is the unit cell used in the FDTD simulations.

100 nm thick silver layer, at the end of which there was zero transmittance; this layer was terminated by a CPML. The length of the y -dimension was chosen such that the reflectance is computed in the far-field region, which is at least 100 nm from the top of the nanoparticles. The reflectance spectra were computed over the optical wavelength range of 350–750 nm and were converted to RGB values. Simulations were performed using an in-house 3D-FDTD parallel code¹⁶ running on the IBM BlueGene/Q supercomputer (64 k cores)¹⁹ operated by SOSCHIP at Scinet, University of Toronto, Canada.

For the direct prediction DL application, the input is the simulation geometries which are parameterized by nanoparticle radius R , embedding depth R_{emb} , and edge-to-edge nanoparticle spacing D_{ee} . The output is one of the colour values R, G, or B. Our total simulation dataset consists of 770 datapoints.

To optimize the use of DL, we need first to find the most ideal DNN architecture: the number hidden layers (HLs) and the number of nodes in each HL that best models our dataset. N-fold Cross Validation (CV)²⁰ is a process where the data is separated into N subsets. A DNN is trained using $N-1$ of the subsets. The N th subset is used for testing and the error is recorded. This process is repeated N times so that each subset is used for testing. The error of a given DNN architecture is the average error from the N DNNs on the test sets.

Ten-fold CV is applied to DNNs with HL number ranging from 1 to 10, and nodes per HL ranging from 4 to 72. We simplify this process by only making DNNs with equal numbers of nodes per HL. Seeing as each DNN outputs only one value (be it R, G, or B), this process is repeated for each colour component. To find an ideal DNN architecture we decided to find one with a minimum sum of normalized Mean Square Errors (MSEs):

$$\min \left(\frac{R_{MSE}}{\min(R_{MSE})} + \frac{B_{MSE}}{\min(B_{MSE})} + \frac{G_{MSE}}{\min(G_{MSE})} \right) \quad (1)$$

where R_{MSE} , G_{MSE} , and B_{MSE} are the mean squared error for the R, G and B values respectively. We also want a DNN that uses a minimal number of HLs and nodes, in order to avoid over-fitting and long training times.

Because of the large number of DNNs trained in the CV process, we made use of the SOSCHIP GPU cluster²¹ where we could train 4 DNNs consecutively on a single node consisting of 4 GPUs. The DNN codes were written using TensorFlow²² and the weights of the DNNs were optimized using an Adam Optimizer²³.

Figure 3 shows the normalized MSE sum as a function of nodes per HL for all of HLs considered. We see that the error plateaus for five or more HLs, after 40 nodes per HL. Beyond this point, the DNNs all have similar accuracies. Because of this, the five HL DNN with 40 nodes per HL, as sketched in Fig. 4(a), was chosen to model the simulation dataset.

A node in a hidden layer will receive as inputs the outputs from a previous layer, as shown in Fig. 4(b). The node performs a linear transformation on the inputs:

$$f(\mathbf{x}^{(n-1)}) = b_j^{(n)} + \sum_i a_{i,j}^{(n)} x_i^{(n-1)} \quad (2)$$

where $a_{i,j}^{(n)}$ is the weight of the j^{th} node in the n^{th} hidden layer applied to the i^{th} input, and $b_j^{(n)}$ is the bias of the j^{th} node in the n^{th} hidden layer. The output of this node is given by the rectified linear activation function:

$$x_j^{(n)} = \max(f(\mathbf{x}^{(n-1)}), 0) \quad (3)$$

The training process involves adjusting the weights $a_{i,j}^{(n)}$ and the biases $b_j^{(n)}$ to minimize the error on the training set. This DNN was trained using 90% of the available data. The remaining randomly selected 10% not used in the training process was used for testing.

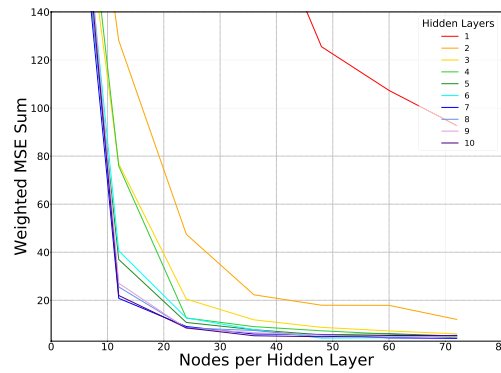


Figure 3. Weighted sum of the MSE for different DNN architectures for the simulation dataset.

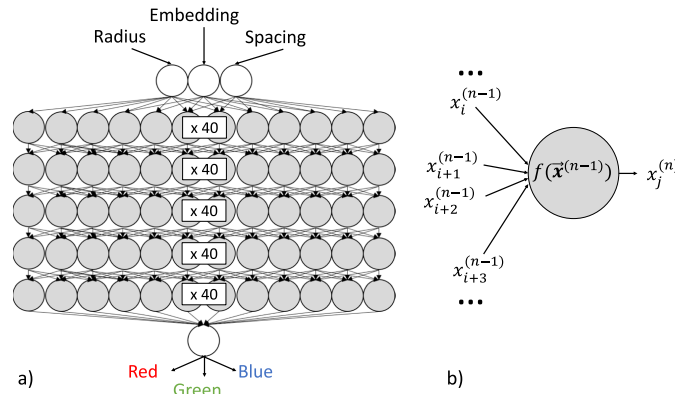


Figure 4. (a) Sketch of the DNN used on the simulation dataset. (b) The j^{th} hidden layer node of the n^{th} hidden layer.

Figure 5(a) shows a palette of some of the colours used in the test set, where the calculated colours from the FDTD simulations are shown on the left and the DNN predicted colours are shown on the right for the same geometric parameters. To the right of the colour pair in Fig. 5(a) we report the corresponding ΔE value²⁴ which is dimensionless and provides a quantitative measure of the difference between the two colours:

$$\Delta E = \sqrt{(L_1 - L_2)^2 + (a_1 - a_2)^2 + (b_1 - b_2)^2} \quad (4)$$

where L_i , a_i , and b_i ($i = 1, 2$) are the colour coordinates in an alternate colour space, calculated from the RGB values¹³. For reference, a ΔE of 1 is the threshold of a colour difference that is just perceptible to the human eye, and 6–7 (as a subjective threshold) is often deemed acceptable in commercial practice. In the test set of 77 colours, only 8 surpass a ΔE of 7. For example, the penultimate block in Fig. 5(a) yields $\Delta E = 7.73$, yet the colour difference is only somewhat noticeable. The mean ΔE of the entire test set is 2.61. The colour palette and corresponding ΔE demonstrate the predictive accuracy of our model.

Figure 5(b) compares the RGB values of FDTD vs DNN for the entire test set. Each circle represents a single datapoint, where the x -axis is the R, G, or B value of the FDTD calculation, and the y -axis is the corresponding R, G, or B value predicted by the DNN. The colour of the circle is either red, green or blue, corresponding to which RGB component that datapoint represents. The ideal linear model ($y = x$) is also displayed for convenience as a solid line. The mean percent error on the test set for R, G, and B are 0.48%, 0.67%, and 3.31% respectively, showing that the DNN chosen predicts all three value with very good accuracy. The larger error for the blue is likely due to the relative lack of representation of low blue values in the dataset. The DNN was not given enough training data with blue values below 0.6 where we clearly see the most error in Fig. 5(b). We can see from the plot that numerical values are well predicted from the model.

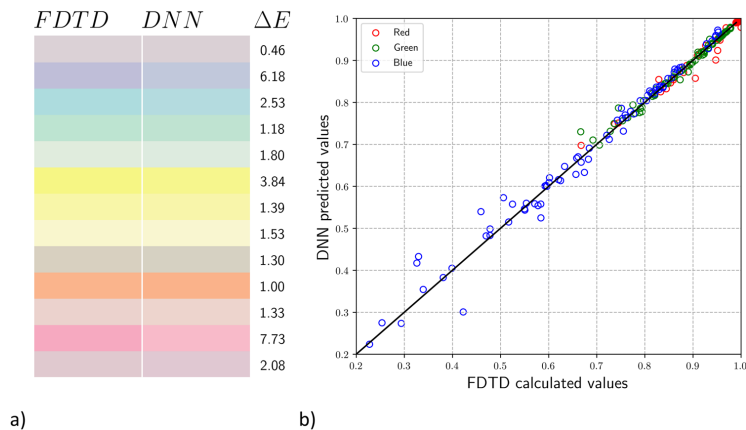


Figure 5. FDTD simulation colours vs. DNN prediction. (a) Comparison between FDTD simulation colours (left) and DNN predicted colours (right) on the test set with corresponding ΔE values. (b) Comparison between RGB values from the simulation results (x-axis) and the DNN predictions (y-axis) along with the ideal linear model ($y = x$). The colour of each circle represents whether the point represents an R (red), G (green), or B (blue) datapoint.

It is worth noting that this model can only be used for interpolation and very small extrapolations. Large extrapolations can lead to nonphysical results. Over the range spanned by our data, this model can predict the colours with high accuracy (less than 4% error) without having to run additional simulations.

Using laser machining data. In this section we repeat the above analysis for the laser parameter dataset. This dataset consisted originally of 10 laser parameters and 1841 datapoints. We eliminated 6 laser parameters due to lack of variation and effect on colour, which left 4 laser parameters under consideration: laser scanning speed, hatch spacing, number of laser pulses per burst, and laser fluence. Physically these parameters affect the total accumulated fluence on the metal surface. This was found empirically to be the key factor in colour production, so that different laser parameters with the same total accumulated fluence produced very similar colours¹¹. While eliminating laser parameters, we also removed the corresponding datapoints to ensure that the training process was not affected. The number of datapoints in our updated set is 1745.

We use the process outlined in the previous section to find the ideal DNN for this dataset. We choose our model to be a DNN with 3 HLs and 60 nodes per hidden layer.

We split our data randomly into two sets: 90% for training and 10% for testing. We again display a comparison between some of the DNN predicted colours and those measured in the lab in Fig. 6(a) along with the ΔE for the test set. The error is higher for this dataset in comparison to the simulation dataset. There is a much more complicated relationship between laser parameters and the RGB of the resulting colour. Despite this, the DNN has found enough relationships to be able to predict a colour within a given range of the measured colours, resulting in a mean ΔE of 5.27 for the test set, where 41 of the 175 test datapoints are greater than 7. We can see a few of these in Fig. 6(a), as the second, third, fifth, and last blocks from the top. The first and third, although having high ΔE values, are still subjectively well predicted. The last block however is the result of a large difference (greater than 0.1) in the R, G and B values. Large errors like these are rare but should be expected in a large test set such as this, where there will likely be points that are not well represented in the training data. This means that the DNN is not as well trained in these regions of parameter space. In Fig. 6(b) we display the comparison of RGB values for the entire test set. The mean percent error for R, G and B are 4.29%, 4.88%, and 5.47% respectively. We see that this DNN predicts each colour reasonably well. However we point out that care should be taken when applying DL to the prediction of subjective results – here, even a small error can change the colour perceptively.

Implicit parameters like temperature or humidity deviations in a laboratory environment may also affect the colour, in addition to the explicit parameters such as laser settings considered here. In commercial setting, the time of day may be relevant due to workplace traffic or machinery vibrations. The relationship between these implicit effects and the resulting colour is very complicated. Though not tracked in our work, these are all parameters that can be used as additional inputs into a deep learning model.

Inverse Design Prediction from Colours

Using simulation data. In this section we solve the inverse problem: we find the simulation geometries that are required to produce given colours, where the RGB values are now the inputs. Solving the inverse problem is not a trivial task as the solution may not be unique. Multiple geometries may produce the same colour so a DNN cannot be directly trained to predict geometric parameters from a given colour. Methods to get around the uniqueness problem have been suggested, including (1) training the inputs using back-propagation with the

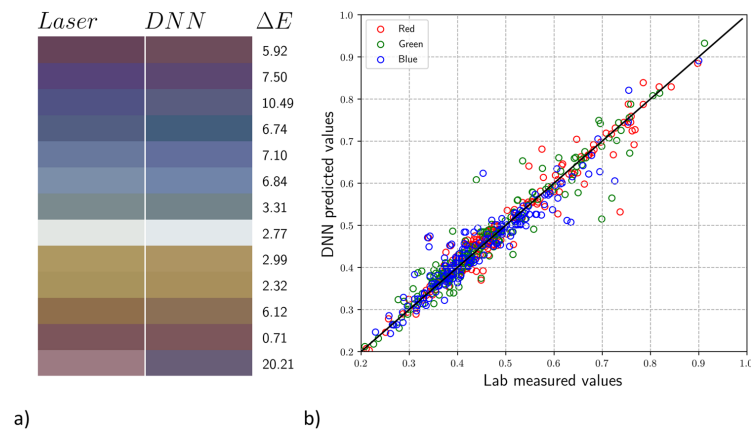


Figure 6. Laser parameter colours vs. DNN prediction. **(a)** Comparison between the measured colours (left) and DNN predicted colours (right) with corresponding ΔE values. **(b)** Comparison between RGB values from the experimental measurements (x -axis) and the DNN predictions (y -axis) along with the ideal linear model ($y=x$). The colour of each circle represents whether the point represents an R (red), G (green), or B (blue) datapoint.

weights of the DNN fixed⁷, (2) training the inverse DNN using the direct DNN^{8,10}, and (3) applying reinforcement learning⁹. Here we introduce a more straightforward iterative multivariable inverse design approach, which to our knowledge remains unreported.

Our approach involves training n neural networks; $\{DNN_1, DNN_2, \dots, DNN_n\}$, where n is the number of outputs (in this case our geometric parameters), given by $\{O_1, O_2, \dots, O_n\}$. DNN_1 is trained to calculate output O_1 from the desired inputs (in this case the RGB values) and the remainder of the outputs $\{O_2, \dots, O_n\}$. This means that the remainder of the outputs act as inputs for DNN_1 alongside the desired RGB values. DNN_2 is trained to calculate O_2 using the desired inputs and the remainder of the outputs $\{O_1, O_3, \dots, O_n\}$, and so on.

Once all n DNNs are trained, we can calculate the set of outputs in an iterative procedure. We first calculate O_n using the DNN_n with the desired inputs and the remaining outputs (which can be initialized as random numbers or the mean values of the dataset). Next we calculate O_{n-1} using DNN_{n-1} with the desired inputs and the remaining outputs including the updated O_n . This is repeated for all outputs. This process can be repeated until the values of the outputs have relaxed to a final value. This method is very simple in implementation since it requires no special treatment of the DNNs unlike inverse design methods suggested in previous studies^{7–10}.

In the case of our simulation geometry, $n=3$. DNN_1 is trained to predict the particle spacing from R, G, B, radius, and embedding. DNN_2 predicts the embedding from R, G, B, radius, and spacing. DNN_3 predicts the radius from R, G, B, embedding, and spacing. After training these three DNNs, we can then calculate the three geometric parameters using the iterative procedure outlined in Fig. 7(a).

First the radius and embedding are set as the mean radius and mean embedding of the dataset, providing a good starting place for the parameter search. We then calculate the spacing using our spacing DNN. The calculated spacing value and mean radius value are used to predict the embedding. This embedding and the spacing are then used to calculate the radius. We then repeat these steps until the geometric parameters relax to their final values. This generally takes 10–20 iterations, however, this may vary depending on the colours themselves. If the colours are not well represented in the training data, the number of iterations will increase. As sketched in Fig. 7(b), we use 5 HLs with 40 nodes per hidden layer for our backwards DNNs.

A sample colour palette of our test dataset is shown in Fig. 8(a) where on the left we display a set of desired colours, and on the right we display the colours generated by the output geometric parameters. We see an agreement between the input colour, and that generated by our iterative method. We also see ΔE values within the same range as the direct prediction with a mean of 2.83. The comparison between input and output RGB values are given in Fig. 8(b) where the percent error between the input and output colours are 0.83%, 1.34%, and 2.12% for R, G and B respectively. The slight increase in error between the direct and inverse prediction is expected due to error accumulation between the backward DNNs used to calculate the geometries and the forward DNN used for testing the geometries. The geometries output by the relaxation are not guaranteed to provide the closest possible colour to the input. They depend on the initial inputs (O_1 and O_2 in Fig. 7). Multiple initialized inputs can be used to find multiple possible geometries, and the one which has the least error between input and output colours should be used.

In traditional optimization methods such as gradient based methods and genetic algorithms, solutions are iteratively sought by trying to minimize an error function. Depending on the complexity of the problem, this can require hundreds or thousands of iterations where at least one simulation must be run for each iteration,

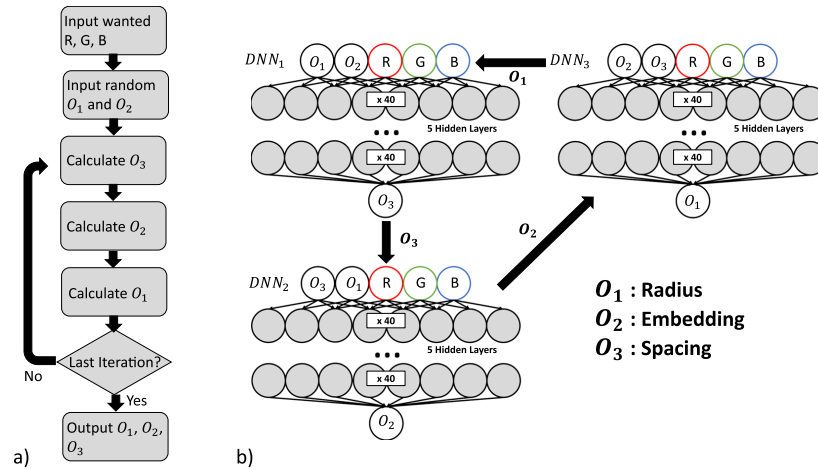


Figure 7. (a) Iterative multivariable inverse design algorithm for the prediction of geometric parameters from colours. We input the desired R, G, B and initial (random or mean) values for the radius and embedding. We then iterate through DNN calculations for spacing, embedding, radius until the last iteration. (b) DNN calculation iteration procedure.

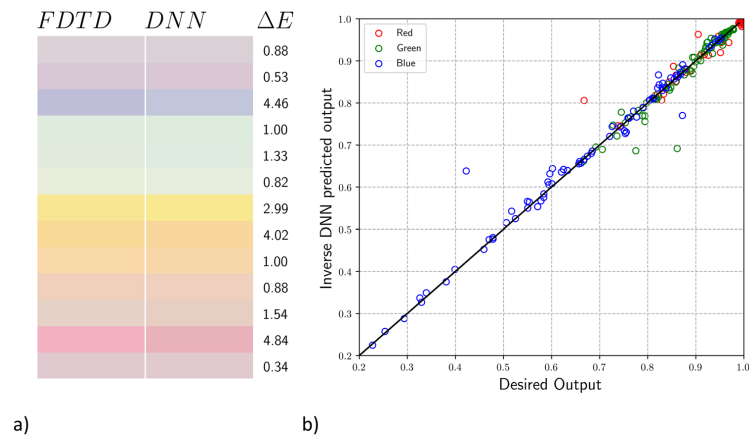


Figure 8. Geometric parameter inverse DNN prediction. (a) Comparison between the input test colours (left) and the colours produced by the output geometric parameters (right) with corresponding ΔE values. (b) Comparison between input RGB values (x-axis) and the values produced by the output geometric parameters (y-axis) along with the ideal linear model ($y = x$). The colour of each circle represents whether the point represents an R (red), G (green), or B (blue) datapoint.

and all this to obtain only one optimized solution. The multivariable inverse design method also requires a large number of simulations upfront for data acquisition, the number of which may be comparable to what is required for an optimization method. However, once these data are used to train the DNNs, it is then possible to obtain a large number of optimized solutions very rapidly. In this study, we seek to find many optimized solutions satisfying different constraints using an existing simulation dataset. Thus using DNNs – which take less than a second per prediction, versus hours on a super computer for a single simulation iteration – wins over traditional optimization.

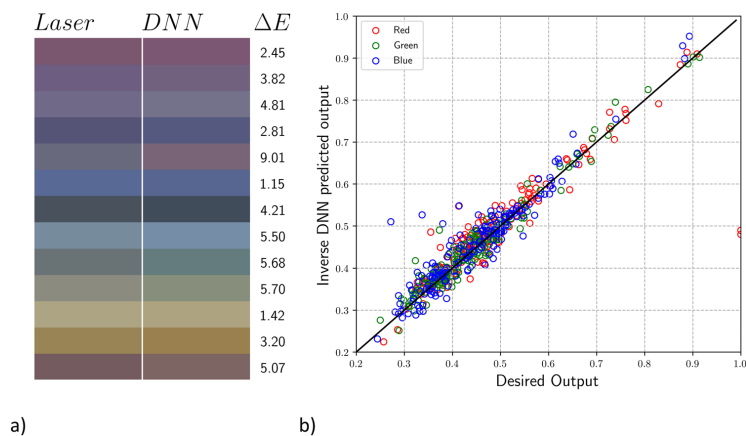


Figure 9. Laser parameter inverse DNN prediction. **(a)** Comparison between the input test colours (left) and the colours produced by the output laser parameters (right) with corresponding ΔE values. **(b)** Comparison between input RGB values (x -axis) and the values produced by the output laser parameters (y -axis) along with the ideal linear model ($y = x$). The colour of each circle represents whether the point represents an R (red), G (green), or B (blue) datapoint.

Using laser machining data. The laser parameter dataset was obtained from the Royal Canadian Mint, who uses this colouring technique to create artwork on silver¹¹. The colours for this artwork are chosen carefully using a palette of known experimentally-obtained colours. The laser is then programmed with the corresponding laser parameters to create this colour on the metal surface. This process is restricted to the list of known colours. However, by using DNNs to predict the laser parameters required to produce a given desired colour, this process can be simplified and laser parameters can be predicted for colours that are not within the known list.

We use the iterative multivariable inverse design procedure introduced in the previous section with $n = 4$ for the four laser parameters that we want to predict (the same ones as used for Fig. 6). We train our four backward DNNs with 90% of the data and test it with the other 10%. These DNNs have three HLs and 60 nodes per HL. We use the direct prediction DNN to calculate the colour from the predicted laser parameters. We show some of the colours output in Fig. 9(a) and display the numerical results in Fig. 9(b). The percent error between the input and output colours are 5.50%, 4.57%, and 6.09% for R, G, and B, respectively, and the mean of ΔE is 6.31, which remains less than our limit of 7. Numerically these errors are small and result in the linear pattern displayed in Fig. 9(b). However, these errors are large enough that some colours are switched in the process. For example, the fifth box from the top in the palette of Fig. 9(a) reveals a change from the desired grey to the output purple. Such changes are caused by differences in the RGB values of the order of 0.07. These differences should be expected due to the accumulation of error in experiment, the backward DNNs, and the forward DNN used to test the backward DNNs. Despite the error in some of the colours, the majority of them are accurately predicted.

Conclusion

We have applied deep learning to plasmonic metal colouration through (1) direct prediction of colours and (2) inverse prediction of parameters required for a desired colour, using both simulation and experimental data sets, with good prediction accuracy. The direct prediction of colours from new parameters means that we increase the palette of available colours without needing to run further simulations or experiments. The Royal Canadian Mint has started using laser colouring on silver¹¹ where the appropriate colours are chosen manually from the database. The inverse prediction can automate this process by finding the laser parameters (that are not necessarily in the database) that best match the desired colour. This method allows for the fixing of certain input parameters, like scanning speed or laser fluence. Fixing the scanning speed is of particular interest, as with higher scanning speeds, a surface can be coloured in less time. The direct prediction can be expanded for industrial applications where environments (temperature, humidity, vibrations, etc) are much less controlled. Environmental factors can be taken into account in the training of the deep neural networks and therefore the output colour of a laser machined surface can be predicted with higher accuracy. Finally, the relaxation technique we developed for the inverse prediction has numerous potential applications in nanophotonics¹⁴. Since it is much easier to implement than advanced methods such as reinforcement learning, and faster than traditional optimization techniques like gradient based methods, it is more accessible to those new to the field of deep learning.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding authors on reasonable request.

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nat.* **521**, 436–444, <https://doi.org/10.1038/nature14539> (2015).
2. Modarres, M. H. *et al.* Neural Network for Nanoscience Scanning Electron Microscope Image Recognition. *Sci. Reports* **7**, 13282, <https://doi.org/10.1038/s41598-017-13565-2> (2017).
3. Rivenson, Y. *et al.* Deep learning microscopy. *Opt.* **4**, 1437–1443, <https://doi.org/10.1364/OPTICA.4.001437> (2017).
4. Ma, W., Cheng, F. & Liu, Y. Deep-Learning-Enabled On-Demand Design of Chiral Metamaterials. *ACS Nano* **12**, 6326–6334, <https://doi.org/10.1021/acs.nano.8b03569> (2018).
5. Malkiel, I. *et al.* Plasmonic nanostructure design and characterization via Deep Learning. *Light. Sci. & Appl.* **7**, 60, <https://doi.org/10.1038/s41377-018-0060-7> (2018).
6. Pilozi, L., Farrelly, F. A., Marcucci, G. & Conti, C. Machine learning inverse problem for topological photonics. *Commun. Phys.* **1**, 57, <https://doi.org/10.1038/s42005-018-0058-8> (2018).
7. Peurifoy, J. *et al.* Nanophotonic particle simulation and inverse design using artificial neural networks. *Sci. Adv.* **4**, eaar4206, <https://doi.org/10.1126/sciadv.aar4206> (2018).
8. Liu, D., Tan, Y., Khoram, E. & Yu, Z. Training Deep Neural Networks for the Inverse Design of Nanophotonic Structures. *ACS Photonics* **5**, 1365–1369, <https://doi.org/10.1021/acsphotonics.7b01377> (2018).
9. Sajedian, I., Badloe, T. & Rho, J. Finding the best design parameters for optical nanostructures using reinforcement learning. *arXiv:1810.10964 [physics]* ArXiv: 1810.10964 (2018).
10. Liu, Z., Zhu, D., Rodrigues, S. P., Lee, K.-T. & Cai, W. Generative Model for the Inverse Design of Metasurfaces. *Nano Lett.* **18**, 6570–6576, <https://doi.org/10.1021/acs.nanolett.8b03171> (2018).
11. Guay, J.-M. *et al.* Laser-induced plasmonic colours on metals. *Nat. Commun.* **8**, 16095, <https://doi.org/10.1038/ncomms16095> (2017).
12. Guay, J.-M. *et al.* Topography Tuning for Plasmonic Color Enhancement via Picosecond Laser Bursts. *Adv. Opt. Mater.* **6**, 1800189, <https://doi.org/10.1002/adom.201800189> (2018).
13. Math|EasyRGB. Available at: <https://www.easyrgb.com/en/math.php>. (Accessed: 24th October 2018).
14. Molesky, S. *et al.* Inverse design in nanophotonics. *Nat. Photonics* **12**, 659, <https://doi.org/10.1038/s41566-018-0246-9> (2018).
15. Yee, K. Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media. *IEEE Transactions on Antennas Propag.* **14**, 302–307, <https://doi.org/10.1109/TAP.1966.1138693> (1966).
16. Calà Lesina, A., Vaccari, A., Berini, P. & Ramunno, L. On the convergence and accuracy of the FDTD method for nanoplasmonics. *Opt. Express* **23**, 10481–10497, <https://doi.org/10.1364/OE.23.010481> (2015).
17. Prokopydis, K. P. & Zografopoulos, D. C. A Unified FDTD/PML Scheme Based on Critical Points for Accurate Studies of Plasmonic Structures. *J. Light. Technol.* **31**, 2467–2476, <https://doi.org/10.1109/JLT.2013.2265166> (2013).
18. Vial, A., Laroche, T., Dridi, M. & Le Cunff, L. A new model of dispersion for metals leading to a more accurate modeling of plasmonic structures using the FDTD method. *Appl. Phys. A* **103**, 849–853, <https://doi.org/10.1007/s00339-010-6224-9> (2011).
19. BGQ - SciNetWiki. Available at: <https://wiki.scinet.utoronto.ca/wiki/index.php/BGQ>. (Accessed: 24th October 2018).
20. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics (Springer-Verlag, New York, 2013).
21. SOSCI GPU - SciNet Users Documentation. Available at: https://docs.scinet.utoronto.ca/index.php/SOSCI_GPU. (Accessed: 24th October 2018).
22. TensorFlow. Available at: <https://www.tensorflow.org/>. (Accessed: 24th October 2018).
23. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* ArXiv: 1412.6980 (2014).
24. Guay, J.-M. *et al.* Passivation of Plasmonic Colors on Bulk Silver by Atomic Layer Deposition of Aluminum Oxide. *Langmuir* **34**, 4998–5010, <https://doi.org/10.1021/acs.langmuir.8b00210> (2018).

Acknowledgements

SOSCI is funded by the Federal Economic Development Agency of Southern Ontario, the Province of Ontario, IBM Canada Ltd., Ontario Centres of Excellence, Mitacs and Ontario academic member institutions. The authors thank SOSCI for their computational resources and financial support. We acknowledge the computational resources and support from Scinet. We acknowledge financial support from the National Sciences and Engineering Research Council of Canada, and the Canada Research Chairs program. The authors also thank the Royal Canadian Mint for the use of their laser lab and the data acquired from it. We would like to thank Graham Killaire, Meagan Ginn, Guillaume Côté, and Martin Charron for their contributions in creating the colour palettes at the Royal Canadian Mint.

Author Contributions

A.C.L. proposed the application of machine learning for colour prediction and inverse design. J.B. conducted the deep learning study including data structuring, coding the deep neural networks, and developing the inverse design method under the supervision of A.C.L. and L.R. J.B. and A.C.L. created the figures. A.C.L. designed and conducted the FDTD simulations under the supervision of P.B. and L.R., J.-M.G. conducted the experiments and supplied the experimental data under the supervision of A.W. and P.B., J.B., A.C.L. and L.R. wrote the manuscript, and J.-M.G., A.W. and P.B. were involved in the editing process.

Additional Information

Competing Interests: The authors declare no competing interests.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

Chapter 8

Unfinished projects

The previous chapters address my contributions to published research. In this chapter, I will briefly discuss works that were left unfinished, but the ideas within are original. For that reason, these are important inclusions to this thesis. The first project concerns using deep neural networks to predict the multipole coefficients of plasmonic nanostructures, and the second project involves the inverse design of metalenses for solid-state high-harmonic generation. The first project was entirely self-conceived; however, I took a supervisory role for most of the work, where undergraduate honour students wrote much of the data engineering and neural network code. The second project is a collaborative effort with researchers from the National Research Council, whose expertise is in ultrafast optics and high-harmonic generation.

8.1 Deep learning for the prediction of multipoles

8.1.1 Introduction

Artificial deep neural networks (DNNs) are useful for modelling complex systems. Once trained, they can predict the outcome of complex experiments in less than a second. The obvious application to nanophotonics is replacing large-scale numerical calculations that can take hours to converge with compact, differentiable models that can be evaluated immediately. However, replacing a full-wave simulator with a DNN is seemingly impossible when accounting for some numerical approaches' flexibility. For this reason, DNNs are restricted to modelling specific situations; for example, optically characterizing relatively

simple, parameterized systems as in Chapter 7 and other works [118–120, 122, 165–168].

Nevertheless, more advanced network architectures allow for the relaxation of parameterization. Convolutional neural networks (CNNs) were popularized a decade ago with the success of *AlexNet* [169] and are now commonly used in most image-recognition applications. Although an image could be flattened and fed into a standard (dense) DNN, this is inefficient as the number of input nodes (equal to the number of pixels in the image) can be very large. This already is an unreasonably large network without considering the number of hidden layers required. Convolutional networks use a system of convolution and pooling layers to reduce the image data to a much smaller transient vector, which becomes the input to a dense DNN which performs the regression/classification. Convolutional networks have been applied to nanophotonics problems already to characterize the optical response from input images [123, 170, 171], or for reducing large array-based datasets [172–174].

The ability of a CNN to process image data allows us to generalize the inputs of a DNN from a vector to a matrix. The question behind this project is how to generalize the output for nanophotonic problems. How close can we get a DNN to learning Maxwell’s equations?

Artificial intelligence is not yet at a point where a single network can solve general electromagnetic problems, and as such, it is up to us to formulate a general problem. We turn to the multipole expansion, which provides a general solution to Maxwell’s equations in a source-free, homogeneous, isotropic medium [136]. More specifically, it can describe the electromagnetic field scattered from an arbitrary scatterer embedded in such an ambient medium [175, 176]. Each term of the expansion contains two unknown scalar quantities, and these scalars represent the entirety of the scatterer. If these scalars can be obtained for the first several orders of the expansion, the scattered field can be reconstructed. This project aims to use CNNs to predict the multipole coefficients from cross-sectional images of 3D plasmonic nanoparticles.

8.1.2 Methods

The radiated electromagnetic field from an arbitrary charge distribution localized in an otherwise source-free homogeneous medium can be described by the multipole expansion [136],

$$\mathbf{E}(\mathbf{r}) = Z \sum_{l,m} \left[\frac{i}{k} a_{l,m} \nabla \times (h_l^{(1)}(kr) \mathbf{X}_{l,m}(\theta, \phi)) + b_{l,m} h_l^{(1)}(kr) \mathbf{X}_{l,m}(\theta, \phi) \right], \quad (8.1)$$

$$\mathbf{H}(\mathbf{r}) = \sum_{l,m} \left[a_{l,m} h_l^{(1)}(kr) \mathbf{X}_{l,m}(\theta, \phi) - \frac{i}{k} b_{l,m} \nabla \times (h_l^{(1)}(kr) \mathbf{X}_{l,m}(\theta, \phi)) \right], \quad (8.2)$$

where $\mathbf{r}$ is assumed to be outside of the charge distribution (radiation zone), l and $|m| \leq l$ are the spherical harmonic degree and order respectively, $a_{l,m}/b_{l,m}$ are the electric/magnetic multipole coefficients for the (l, m) index set, k is the wavenumber, $h_l^{(1)}(kr)$ is the spherical Hankel function of the first kind, $\mathbf{X}_{l,m}(\theta, \phi)$ are the vector spherical harmonics, and Z is the impedance of the ambient medium.

If the electric field is known, the multipole coefficients can be obtained by integrating the radial component $\mathbf{r} \cdot \mathbf{E}$ and $\mathbf{r} \cdot \mathbf{H}$ of the scattered fields over a sphere of radius r_0 around the charge distribution via,

$$a_{l,m} = -\frac{k}{Z_0 h_l^{(1)}(kr) \sqrt{l(l+1)}} \int_{\text{sphere surf.}} Y_{l,m}^*(\theta, \phi) \mathbf{r} \cdot \mathbf{E} d\Omega \quad (8.3)$$

$$b_{l,m} = \frac{k}{h_l^{(1)}(kr) \sqrt{l(l+1)}} \int_{\text{sphere surf.}} Y_{l,m}^*(\theta, \phi) \mathbf{r} \cdot \mathbf{H} d\Omega \quad (8.4)$$

where $Y_{l,m}^*$ is the complex conjugate of the scalar spherical harmonic function, and $d\Omega$ is the differential solid angle. The total radiated field power can be calculated from the multipole coefficients via

$$P = \frac{Z_0}{2k^2} \sum_{l,m} [|a_{l,m}|^2 + |b_{l,m}|^2]. \quad (8.5)$$

Notice that to describe the radiation field fully, the problem boils down to finding the set of multipole coefficients $(a_{l,m}, b_{l,m})$. If we assume the charge distribution is a plasmonic nanostructure embedded in a dielectric medium, Eqs. 8.1 and 8.2 describe the scattered field from an irradiated plasmonic particle, and Eq. 8.5 is the scattered power. Therefore, the plasmonic scattering problem can be solved if one can accurately predict the $(a_{l,m},$

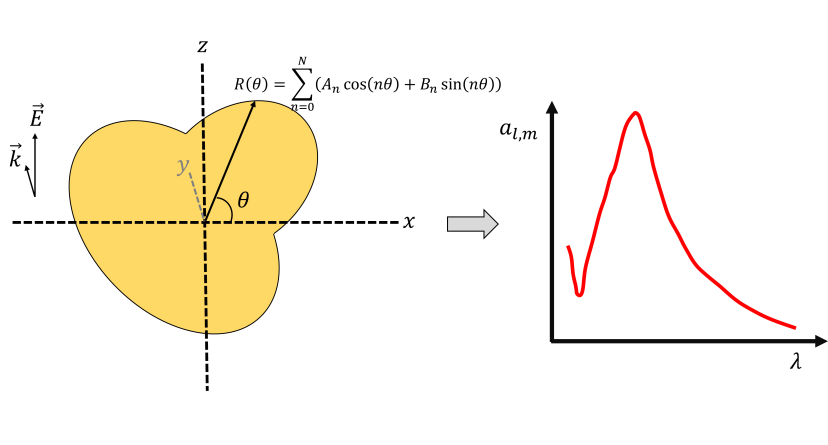


Figure 8.1: An example of a randomly generated nanoparticle from a Fourier series with randomly generated weights and truncation index. In the out-of-plane direction y , the particle has a uniform thickness of 40 nm. The incident electric field is polarized along z and travels along y . The output of the simulation is the multipole spectrum $a_{l,m}(\lambda)$ of the particle.

$b_{l,m}$) coefficients for the desired particle geometry.

With this in mind, we randomly generated a dataset of ~ 32000 gold nanoparticles by parameterizing their in-plane $x - z$ surface as a function describing the distance $R(\theta)$ from the origin to the surface of the particle as a function of angle as illustrated in Figure 8.1. The function R is chosen as a Fourier series so that $R(0) = R(2\pi)$ where the truncation index N and the weights A_n and B_n are randomly generated. In the out-of-plane (propagation) direction y , the particle is uniform with a constant thickness of 40 nm – thus enabling us to describe the geometry of each particle via a cross-sectional image.

Using our in-house GPU-parallelized FDTD program, simulations were run for each particle. A z - linearly polarized plane wave is propagated along the y axis to irradiate the particle. Electric and magnetic field data are collected at points forming a spherical shell around the particle for 100 wavelengths ranging from 300 – 800 nm. In post-processing, we extract $a_{l,m}$, $b_{l,m}$ by numerically integrating Eqs. 8.3 and 8.4. As plasmonic particles are dominated by their electric dipole ($a_{l=1}$) and quadrupole ($a_{l=2}$) responses, we thus truncate our analysis to these two terms.

Each data point is a 180×180 pixel image of the nanoparticle alongside a spectrum of electric multipole coefficients. With this data, a CNN is trained to predict the multipole coefficients from the nanoparticle image, where we make use of the Tensorflow Python library [131].

A full CNN is illustrated in Figure 8.2. The input image is fed to a convolution layer, where the input array is convoluted with a small filter array (typically a $\sim 3 \times 3$ matrix) called a kernel. The values of the convolution kernel arrays are optimized in the training process alongside the weights and biases of the dense neural network layers. The kernel is translated across the image pixel-by-pixel. At each pixel, the values of the surrounding sub-array are multiplied by the values of the kernel and the results are summed, becoming the corresponding pixel value of the output array. The kernel is then translated to the next pixel of the image and the process repeated. As such, the size of the output array is only reduced by a couple of pixels compared to the input image (depending on the boundary conditions of the convolution operation). Furthermore, in a single convolution layer, multiple kernels may be applied in parallel, generating multiple new arrays to be passed to the next layer. In Figure 8.2, we illustrate two convolution layers: the first uses three kernels and the second uses six. Following the convolution operation, an activation function (typically the rectified linear unit - ReLU) is applied for the nonlinearity.

The convolution operation introduces a lot of new data in the form of multiple output arrays. Following the convolution-activation layer comes a pooling (or subsampling) layer, where the convolved array is downscaled from an $N \times M$ array to $\sim N/c \times M/c$ by either taking the average over each $c \times c$ subarray (mean-pooling) or by taking the maximum value of each $c \times c$ subarray (max-pooling). In this case, the sub-arrays are fixed (not translated pixel-by-pixel) so that the numerical data is reduced.

The process outlined above is usually repeated several times. After applying a sequence of convolution-pooling layer pairs, the values of the resultant arrays are combined into a flattened array (vector) whose numerical values capture the necessary features to distinguish it from another nanoparticle. This *latent* vector is the input to a dense DNN (as covered in Chapter 2.4), which is trained to predict the spectrum of multipole coefficients.

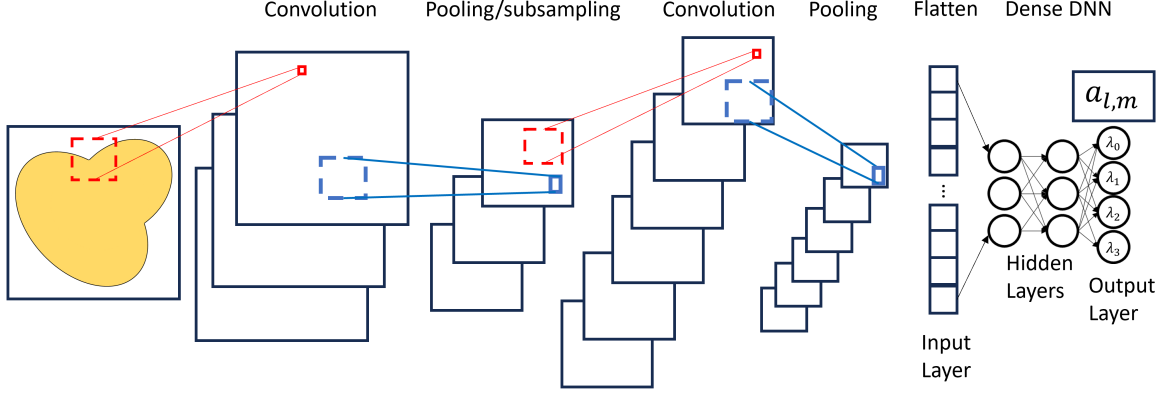


Figure 8.2: Diagram of a CNN, where an image is transformed by a sequence of convolutional and pooling layers to a flattened vector. This becomes the input of a dense DNN that predicts a vector of multipole coefficients for each wavelength. Red boxes represent convolution kernels, and blue boxes represent the pooling sub-arrays.

8.1.3 Results

Although many CNN architectures were experimented with, the architecture we found successful is illustrated in Figure 8.2. The image is fed into a convolution layer with a 3×3 filter, followed by a max-pooling layer with $c = 3$. This is repeated a second time. The output is then flattened and sent to a seven-hidden-layer dense DNN, each with 500 nodes. The output of the CNN is a vector of multipole values $a_{l_0, m_0}(\lambda)$ corresponding to the (l_0, m_0) multipole spectrum.

In Figure 8.3, we plot six selected data points from the test data set of particles unseen by the DNN during the training process. The line plot for each datapoint shows the FDTD-calculated (blue line) and DNN-predicted (red crosses) $a_{1,1}$ spectrum. The contour plot shows the corresponding nanoparticle cross-sections - the input to the CNN. Although the predictions are not perfect, we can see that the DNN captures key spectral features, such as the spectral dip evident in the last example. Some non-Fourier-based shapes were tested, including rectangular prisms, to validate generalizability. In Figure 8.4, we compare the DNN prediction (y-axis) of the entire test set (all wavelengths included) to the FDTD calculations (x-axis) via a scatter plot, where in the absence of error, the points would lie

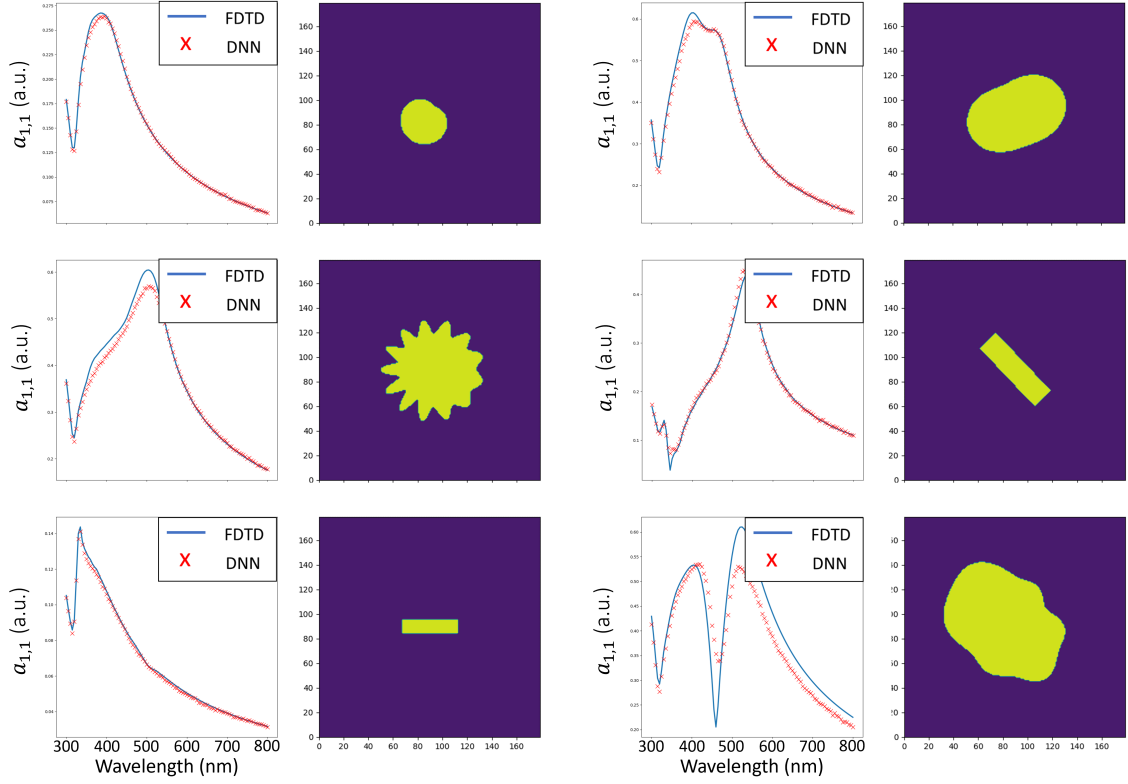


Figure 8.3: Selected data points from the test set, unseen by the CNN in training. The spectra show the magnitude of the first-order electric multipole (dipole) coefficient as calculated by FDTD (blue line) and predicted by the DNN (red x 's). The adjacent image shows the cross-section of the test particle.

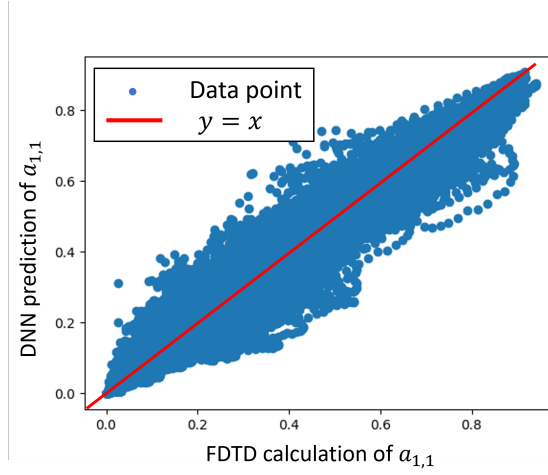


Figure 8.4: Comparison of the FDTD-calculated values of the first-order electric multipole (dipole) coefficient and the DNN-prediction for all data points in the test set (blue dots). A linear function ($y = x$) is included for convenience (red line).

on the $y = x$ line (red line).

8.1.4 Discussion

In the previous section, we demonstrated using CNNs to predict multipole scattering coefficients from plasmonic nanoparticles of arbitrary shape. The agreement between the network’s predictions and the FDTD calculations validates the idea behind this work. However, better agreement can be achieved with a more suitable network architecture, a task that would require more time. For example, only a little work was done in experimenting with the number of convolutional layers and the hyperparameters underlying these. That said, the project is being continued by a fellow graduate student in the Computational Nanophotonics research team.

The long-term goal of this project took inspiration from Ref. [121], where inverse design is accomplished by calculating the gradient of the DNN output with respect to the input via backpropagation through the pre-trained DNN. This idea can be extended to CNNs by their differentiability, and it may be possible to inverse-design nanoparticle geometries for a specified multipole spectrum. Furthermore, using the multipole coefficients, one may also use our CNN to predict the behaviour of plasmonic metasurfaces consisting of periodic

arrangements of these particles and the subsequent inverse problem.

8.2 Inverse design of a metalens for high harmonic generation

8.2.1 Introduction

High harmonic generation (HHG) is a process that occurs when a sufficiently intense laser induces the generation of odd harmonics from atoms in a medium [144] and is the basis for producing attosecond optical pulses. Originally discovered in rare gas atoms, HHG was recently measured from solids [177, 178], thus opening the door to tailored HHG via nanostructuring of the supporting medium. Like in nonlinear plasmonics, nanostructuring can enhance the HHG process and control the far-field radiation of the harmonics [179, 180].

For example, in Ref. [179], the diffraction pattern of HHG from nanostructured ZnO showed that HHG forms much stronger in electromagnetic hotspots and that the emission was spatially coherent. Furthermore, they showed that far-field control of HHG can be achieved via local amplitude masking by fabricating an HHG Fresnel zone plate (FZP) to focus the generated harmonics. As another example, in Ref. [180], a metasurface consisting of periodically arranged silicon nanorods (hosting an electric dipolar bright mode) coupled to silicon nanodisks (hosting a magnetic dipolar dark mode) is shown to enhance harmonic signals by orders of magnitude compared to the unpatterned silicon. It even revealed harmonic orders that were below the noise threshold for an unpatterned silicon surface.

It is thus not surprising that inverse design is being considered for the enhancement and control of HHG, and this is the subject of this section. Specifically, we wanted to design a metasurface to be etched onto an HHG-host medium to focus the resulting harmonics. In this work, we assume the harmonics are generated at the interface of bulk MgO and vacuum; however, no material models were used, and MgO/vacuum can be replaced with any other dielectric media for this work. In the following, we outline three ideas suggested/applied to this problem; automatic differentiation, parameterization/finite differences, and adjoint sensitivity analysis.

8.2.2 Methods

In order to design an HHG metalens, we first need to be able to simulate the generation and propagation of HHG fields. It is beyond this project's scope to derive/implement a self-consistent optical model for solid-state HHG. However, we use a simple model to convert the linear electric field to an approximate HHG field given by,

$$E_{z,HHG}(n\omega) \propto |E_z(\omega)|^2 \exp(in\angle E_z(\omega)) \quad (8.6)$$

where $E_z(\omega)$ is the linear electric field assuming the incident field is z -polarized at frequency ω , n is the harmonic order, and $\angle E_z$ is the phase of the linear field. This model is meant to act as an effective amplitude/phase mask that is applied to fields at the interface between the host medium and the ambient medium and provides us with a starting point from which to solve the problem as a whole. In fact, any model that can convert the linear electric field to the HHG field can be applied to the following methods.

Here we will outline the method used to calculate HHG emission from a metasurface. We use a two-step process that includes FDTD and the angular spectrum method (ASM). The ASM uses Fourier transforms to propagate electromagnetic fields and is very useful for propagating electromagnetic field planes to the far field [5]. The FDTD simulation calculates the linear electromagnetic fields from the metasurface under pulse irradiation. A schematic illustration of the simulation setup is shown in Figure 8.5a where the host medium is illustrated as grey, the ambient medium remains page colour, and the metasurface is shown to be concentric rings (although this is not necessary). As illustrated in Figure 8.5b, a Gaussian beam pump source is introduced on an $x - z$ plane inside the bulk of the host medium. Planar discrete Fourier transform (DFT) monitors are placed on the bottom and top of the etch as shown in Figure 8.5c and d, respectively. This field data will be used to calculate the HHG fields.

After the FDTD simulation, we enter the ASM step. First, Eq. 8.6 converts the monitor fields to the HHG fields. Note that this only applies to fields at the interface between the host and ambient medium. In Figure 8.5e, we show a cross-sectional summary of the simulation setup, where yellow x 's are used to show locations ignored in the propagation steps. These ignored fields are set to zero before propagation. The lower HHG-field plane is propagated a short distance to the top plane by applying a phase factor (green arrow in Figure 8.5e), and combined with the field on the top plane,

$$E_{z,DFT2}(x, y_2, z) = E_{z,DFT2}(x, y_2, z) + E_{z,DFT1}(x, y_1, z)e^{ik_y(y_2-y_1)}, \quad (8.7)$$

where y_1 and y_2 are the locations of planar DFT monitors 1 and 2, respectively, and $E_{z,DFT2}$ and $E_{z,DFT1}$ are the HHG fields at the monitors and are set to zero at the non-interface locations. The propagated field $E_{z,DFT1}$ fills the blank spots in the top plane (yellow x 's), combining with the HHG field at the top etch $E_{z,DFT2}$. This short propagation step assumes propagation is only along y and is valid if the propagation distance $y_2 - y_1$ is smaller than the wavelength - note that this may need to be re-evaluated in the case of HHG and the resulting small wavelengths.

Finally, the ASM propagates the second DFT plane to the focal plane in the far field. First $E_{z,DFT2}(x, y_2, z)$ is Fourier transformed in the x and z dimensions,

$$\hat{E}_{z,DFT2}(k_x, y_2, k_z) = \int dx dz E_{z,DFT2}(x, y_2, z) \exp\{(ik_x x + ik_z z)\}, \quad (8.8)$$

where k_x and k_z are the x and z spatial frequencies, respectively, of $E_{z,DFT2}(x, y_2, z)$. In the Fourier domain, we can propagate $\hat{E}_{z,DFT2}$ to the far field using a propagation step,

$$\hat{E}_{z,FF}(k_x, y_{FF}, k_z) = \hat{E}_{z,DFT2}(k_x, y_2, k_z) e^{ik_y(y_{FF}-y_2)}, \quad (8.9)$$

where y_{FF} is the location of interest in the far field, and where k_y is the y -component of the propagating wavevector. As we are using monochromatic electric fields, the magnitude of the wavevector $k = \omega/c$ is fixed, and k_y is dependent on k_x and k_z ,

$$k_y = \sqrt{k^2 - k_x^2 - k_z^2}. \quad (8.10)$$

In practice, the propagator $e^{ik_y(y_{FF}-y_2)}$ is an array of complex numbers. However, there exist pairs of k_x and k_z for which k_y will be imaginary, where $k_x^2 + k_z^2 > k^2$. If this inequality is met, $\hat{E}_{z,DFT2}(k_x, y_2, k_z)$ will evanescently decay to zero in the far field. Thus, only values of $\hat{E}_{z,DFT2}(k_x, y_2, k_z)$ for which $k_x^2 + k_z^2 \leq k^2$ need to be explicitly propagated. As an aside, this means that only small spatial frequencies make it to the far field. Thus, information regarding the smallest features (the highest spatial frequencies) is lost in propagation - this is the cause of the diffraction limit mentioned in the introduction of this thesis.

Finally, we must inverse Fourier transform this to the spatial domain to evaluate our objective,

$$E_{z,FF}(x, y_{FF}, z) = \int dk_x dk_z \hat{E}_{z,FF}(k_x, y_{FF}, k_z) \exp\{(-ik_x x - ik_z z)\}. \quad (8.11)$$

In practical application, the Fourier transforms are implemented using 2D fast Fourier transforms (FFTs) and their inverses (IFFTs) that are available in many scientific computing packages, namely auto differentiated libraries like *JAX* [132] and *PyTorch* [181], the latter of which was applied in this work. Therefore, after the first (approximate) propagation step, the far field $E_{z,FF}(x, y_{FF}, z)$ can be determined via,

$$E_{z,FF}(x, y_{FF}, z) = IFFT(FFT(E_{z,FF}(x, y_2, z))e^{ik_y(y_{FF}-y_2)}), \quad (8.12)$$

where, for completeness, it is worth noting that an FFT-shift and an inverse FFT-shift function must be applied to $\hat{E}_z(k_x, y, k_z)$ before and after (respectively) applying the propagator $e^{ik_y(y_{FF}-y_2)}$. These functions appropriately reorder the frequencies in the field arrays.

This workflow allows us to design a metaleins by coupling full-wave simulations with propagation methods. However, in our first approach, we make an approximation that allows us to avoid lengthy full-wave simulations, thus simplifying the extraction of gradients to guide optimization.

Automatic Differentiation

Initially, it was believed that such a metasurface would behave, essentially, as a phase mask due to the relative size disparity between the incident wavelength (in the near-infrared) and the height of the etch (on the order of 80 nm). The harmonics are formed at the interface between the host and ambient mediums; thus, the phase would be roughly constant across the unetched (Figure 8.5(c)) and etched (Figure 8.5(d)) interfaces. Under this assumption, first, the metasurface phase and amplitude profile would be implanted (multiplied) onto the incident Gaussian beam, and next, the HHG field is calculated via Eq. 8.6.

This assumption allows us to bypass the FDTD simulation. Only the ASM is required to propagate the HHG field to the focal plane in the far field. The ASM is computationally efficient, and automatic differentiation was used as the inverse design procedure. The objective function is set to be the far-field intensity at the focus $|E_{z,FF}(\mathbf{r}_{focus})|^2$, and the tuneable

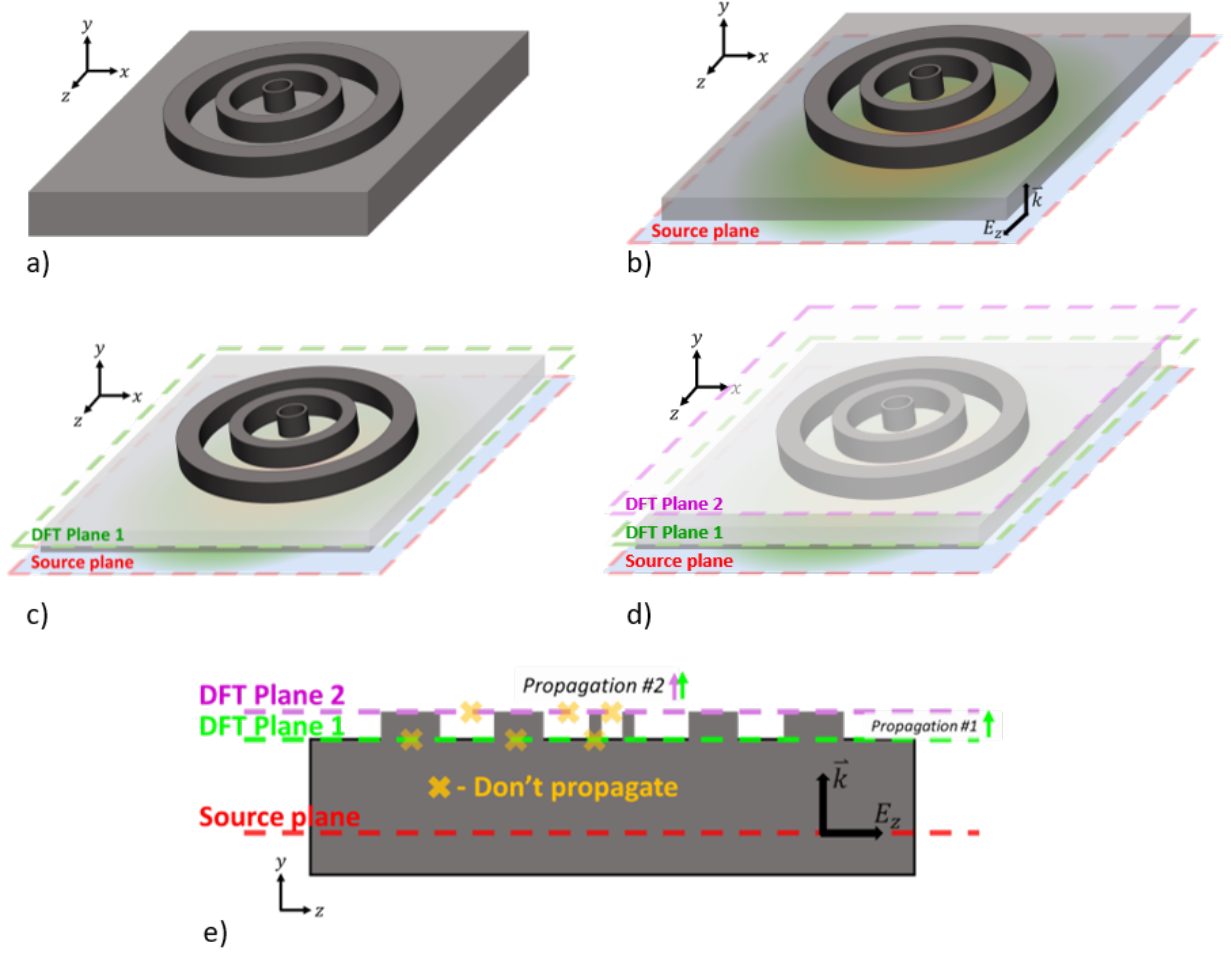


Figure 8.5: FDTD simulation setup for an HHG-focusing metalens used for inverse design; a) is a model metalens where the grey matter is MgO, and the ambient medium is air, b) a Gaussian source plane is placed inside the bulk MGO and light propagates along y and is polarized in the z direction, c) a discrete Fourier transform ($x - z$) plane monitor is placed at the bulk MgO-air interface, d) another plane monitor is placed at the top nanostructured MgO-air interface, e) a $y - z$ cross-section of the simulation setup where the green and purple arrows represent the first and second propagation steps respectively, and the yellow x 's represent non-interface regions where fields are set to zero before propagating.

parameters for which the gradient is required are the pixels of the phase and (or) amplitude profile of the metasurface. As in adjoint sensitivity analysis, the phase/amplitude profiles would be allowed to vary continuously during the optimization. Using filtering techniques outlined in Section 2.3.2, minimum feature sizes can be enforced, and the structure can be pushed toward a binary material distribution. This method was promising; however, it was later found that the phase plate approximation was not applicable, and the metasurface imparts more than a phase profile on the incident wave as depicted in results presented in Ref. [182]. As such, this method was abandoned; however, it remains feasible for designing phase/amplitude plates and inspired the adjoint sensitivity analysis workflow.

Parameterization/finite differences

Our second idea for designing this metasurface was to parameterize the metasurface and use finite differences to extract the gradient of the objective with respect to said parameters. The basis structures were concentric circular micro-rings as illustrated in Figure 8.5a, and the tuneable parameters are the radii and widths of each ring. The objective function remains the same as the previous method. The gradient requires several function evaluations (simulations); for a structure containing N rings, $2N + 1$ FDTD simulations (and subsequent propagation steps) are required to evaluate the gradient via finite differences. Each function evaluation is a 3D-FDTD simulation of a metasurface, a significant bottleneck for this method that hindered its exploration. It is worth noting that the benefit of such parameterized methods, despite their high-cost/low-return gradients, is that the feature size and binary material structure are implicitly obeyed, and no additional filtering is required.

Adjoint sensitivity analysis

As introduced in Chapter 2.3.1, the adjoint method can be used to calculate the gradient of an objective function with respect to the permittivity distribution, enabling the design of arbitrary geometries. Fortunately for our problem, the nonlinearity is applied externally to the full-wave simulation. Thus, the adjoint algorithm presented in Section 2.3.1 can be applied here, with some extra considerations.

Once again, we define our objective function to be $F = |E_{z,FF}(\mathbf{r}_{focus})|^2$ on the focal plane as evaluated using the two-step FDTD-ASM method. From Eq. 2.89, we know that the gradient is given by

$$\frac{dF}{d\varepsilon(\mathbf{r}_i)} = \frac{\partial F}{\partial E_z} \frac{\partial E_z}{\partial \varepsilon(\mathbf{r}_i)}, \quad (8.13)$$

where E_z is the electric field data collected by the DFT monitors in Figure 8.5. The function F is used to encapsulate the nonlinear conversion $E_{z,HHG}(\mathbf{r}) = N(E_z(\mathbf{r}))$ and propagation $E_{z,FF}(\mathbf{r}_{focus}) = P(E_{z,HHG}(\mathbf{r}))$ steps so that

$$F(E_z) = |E_{z,FF}(\mathbf{r}_{focus})|^2 = |P(N(E_z))|^2. \quad (8.14)$$

The gradient $\partial F / \partial E_z(\mathbf{r})$ can be evaluated using automatic differentiation. Our workflow has two such gradients, one for each DFT plane monitor. Each gradient is a 2D array of complex numbers that become the source amplitude of the adjoint FDTD simulation, which solves Eq. 2.98. Translating the complex fields into a real signal for FDTD is discussed in Chapter 2.3.2. In the adjoint simulation, the source planes are located at the DFT plane locations from Figure 8.5c and d. Additional 3D-DFT monitors are required in both the forward and adjoint simulations to capture the electric field in the design region for the objective function gradient with respect to the permittivity distribution (Eq. 2.99). This gradient guides the optimizer to maximize F , enabling the autonomous design of an HHG metalens.

This method was derived while writing this thesis, and the implementation and testing remain future work. Note that the above steps can be applied to the design of a linear metalens as well, where the nonlinear conversion is no longer required, and only a single DFT monitor is required above the metalens in the forward simulation (and thus only a single source plane in the adjoint simulation). Furthermore, a more proper treatment would replace the scalar ASM with a vectorial ASM or a near-to-far-field transform (see Chapter 8.5 of Ref. [129] for explicit instructions). The objective function would be the Poynting flux at the focal point. Vectorial methods can also be implemented using an automatic differentiation framework, and the above steps still apply.

8.2.3 Results

This section presents preliminary results for the automatic differentiation and finite-difference methods. As mentioned above, the adjoint method was developed while writing this thesis (after changing career paths), and its implementation still needs to be explored. Furthermore, a full exploration of the finite difference method has yet to be conducted for reasons discussed in the respective section.

Automatic Differentiation

For this inverse design method, the *PyTorch* [181] python library was used as it had the necessary tools for automatic differentiation of the ASM. The processes of multiplying the Gaussian beam profile by the amplitude and phase profiles of the metasurface, the HHG conversion (Eq. 8.6), and the ASM propagation are written in a Python function, and the gradient is calculated via backpropagation in the *PyTorch* backend. In this work, a custom gradient descent algorithm is used to update the design; however, a better choice would be an open-source module like the minimizer provided by the Scipy Optimize library [142].

In this section, we present two examples of amplitude plate design and phase plate design. For the amplitude plate designs, the final device is fully binarized and optimized to maximize the device’s performance exactly at the seventh harmonic $\lambda \sim 114$ nm of a $\lambda = 800$ nm incident beam - narrow-band device design. The phase plate designs were allowed to remain continuous (non-binary). Although impossible to fabricate, it allowed for interesting patterns. For the phase plates, the optimization is performed over a bandwidth around the seventh harmonic by summing the individual objective functions (the electric field intensity at each wavelength) and their respective gradients - broadband device design. The focal point for all examples remains constant at 60 microns from the center of metasurface. The metasurface was designed to be $30 \times 30 \mu m^2$ (despite the first few figures, which are cropped) and discretized using 2000×2000 pixels. A 90° rotational symmetry was also enforced to reduce the number of tuneable parameters. No feature-size filtering was applied in this section, however the amplitude masks used a sigmoid-function binary filter that was steepened approximately every 50 optimization iterations.

The first inverse-designed amplitude plate was initialized as a Fresnel zone plate (FZP) tuned to focus the seventh harmonic $\lambda \sim 114$ nm frequency component as illustrated in Figure 8.6a where yellow regions are fully transmitting, and purple regions have zero transmittance. The objective function is set to maximize the intensity of only the seventh harmonic frequency component at the focal point, and the amplitude profile is allowed to vary continuously between 0 and 1 during the optimization procedure. Every 50 (or so) iterations, an increase in the binarization is enforced to push the continuous amplitude profile toward a binary one, converging on Figure 8.6b. The resulting device resembles the effective medium approximation of a continuous (non-binary) FZP profile. In Figure 8.6c, the normalized intensity spectrum of the device in the focal plane is plotted for both the initial condition (yellow line) and the inverse-designed device (blue line). The final device has 3.7 times the intensity at the focal plane versus the FZP, but the focusing bandwidth

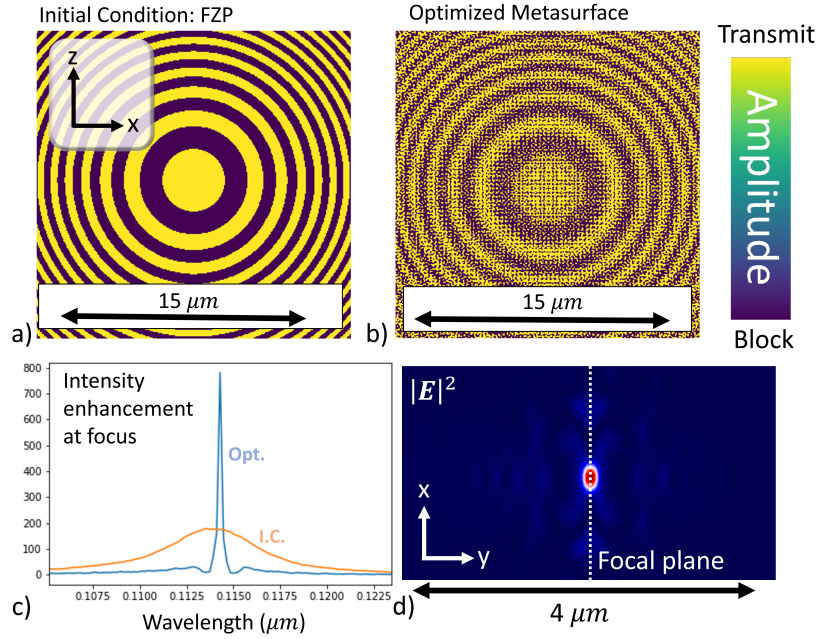


Figure 8.6: Inverse design of an amplitude plate to maximize focusing at the seventh harmonic $\lambda \sim 114$ nm via automatic differentiation of the ASM. a) Initial condition of the inverse design is chosen to be a Fresnel zone plate tuned for the seventh harmonic. Yellow regions represent unity transmittance, and purple regions represent zero transmittance. b) Final design optimized to maximize the seventh harmonic at the focal plane. c) Seventh harmonic intensity spectrum at the focal point, normalized to the incident intensity for the initial FZP (yellow line) and the inverse-designed device (blue line). d) Electric field intensity at the focal plane.

is greatly reduced. In Figure 8.6d, we plot the seventh harmonic field intensity around the focal plane, demonstrating a narrowed focal spot.

The second inverse-designed amplitude plate was initialized as a random grid of pixels as illustrated in Figure 8.7a. For this example, the objective function is set to maximize the intensity of only the seventh harmonic frequency component at the focal point and minimize the intensity at the fundamental wavelength. The optimized design is illustrated in Figure 8.7b. Despite the randomly generated initial condition, the resulting device resembles the FZP profile as with the previous example. In Figure 8.7c, the normalized

intensity spectrum of the device in the focal plane is plotted for both the initial FZP from the previous example (yellow line) and the inverse-designed device (blue line). The final device has approximately 2.8 times the intensity at the focal plane versus the FZP, with a greatly reduced focusing bandwidth. The intensity at the focus is decreased from the previous device, but this is made up for by the 10^{-6} factor reduction in intensity at the focus for the fundamental wavelength as shown in Figure 8.7d, which illustrates a log plot of the focal intensity spectrum around the fundamental wavelength. Theoretically, this reduction in the fundamental wavelength means the HHG emission could irradiate a sample without the overwhelming source beam.

The first inverse-designed phase plate was initialized as a Fresnel zone plate (FZP) tuned to focus the seventh harmonic $\lambda \sim 114$ nm frequency component as illustrated in Figure 8.8a where yellow regions induce a π phase shift on the incident beam, and purple impart zero phase shift. The objective is to maximize the intensity of ~ 20 frequency components within an 8 nm bandwidth of the seventh harmonic at the focal point. No binarization scheme is applied this time, and the final device is shown in Figure 8.8b. Unsurprisingly, the resulting device resembles a continuous (non-binary) FZP profile. In Figure 8.8c, we plot the normalized intensity spectrum of the phase plates in the focal plane for both the initial FZP (yellow line) and the inverse-designed device (blue line). At the center frequency, the final device has 1.7 times the intensity at the focal plane versus the FZP. Unlike the amplitude plates, the focusing is increased over the bandwidth of the harmonic, so a broader wavelength range is available for focusing. In Figure 8.8d, we plot the seventh harmonic field intensity around the focal plane, demonstrating a much broader focal spot compared to Figure 8.6d.

The second inverse-designed phase plate was initialized as a random grid of pixels as illustrated in Figure 8.9a. For this example, the objective function is set to maximize the intensity around a bandwidth of the seventh harmonic at the focal point and minimize the intensity for a bandwidth around the fundamental wavelength. The optimized design is illustrated in Figure 8.9b, where the device does not directly resemble an FZP despite the ring-based geometries that constitute it. In Figure 8.9c, the normalized intensity spectrum of the device in the focal plane is plotted for both the initial FZP from the previous example (yellow line) and the inverse-designed device (blue line). The intensity focused by the final device is reduced by a factor of 1.75 compared to the FZP. That said, the intensity at the focus for the fundamental wavelength is reduced by 10^{-3} as shown in Figure 8.9d, illustrating a log plot of the focal intensity spectrum around the fundamental wavelength.

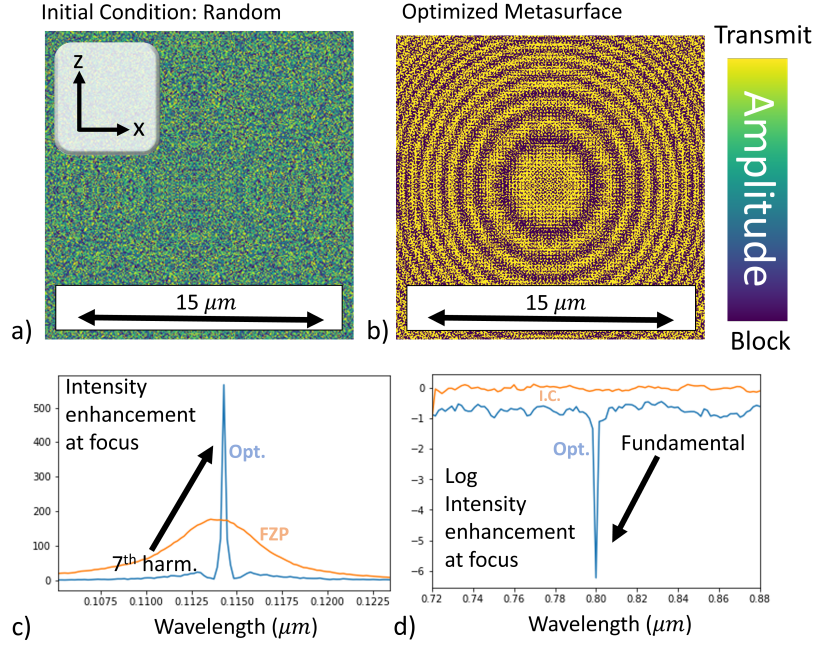


Figure 8.7: Inverse design of an amplitude plate to maximize focusing at the seventh harmonic $\lambda \sim 114\ \text{nm}$ and minimize the fundamental $\lambda = 800\ \text{nm}$ via automatic differentiation of the ASM. a) Initial condition of the inverse design is chosen to be a random grid of pixels ranging from fully transmitting (yellow) to total blocking (purple). b) Final design optimized to maximize the seventh harmonic and minimize the fundamental at the focal plane. c) Seventh harmonic intensity spectrum at the focal point, normalized to the incident intensity for the FZP (yellow line) and the inverse-designed device (blue line). d) Log plot of the fundamental intensity spectrum at the focal point, normalized to the incident intensity for the FZP (yellow line) and the inverse-designed device (blue line).

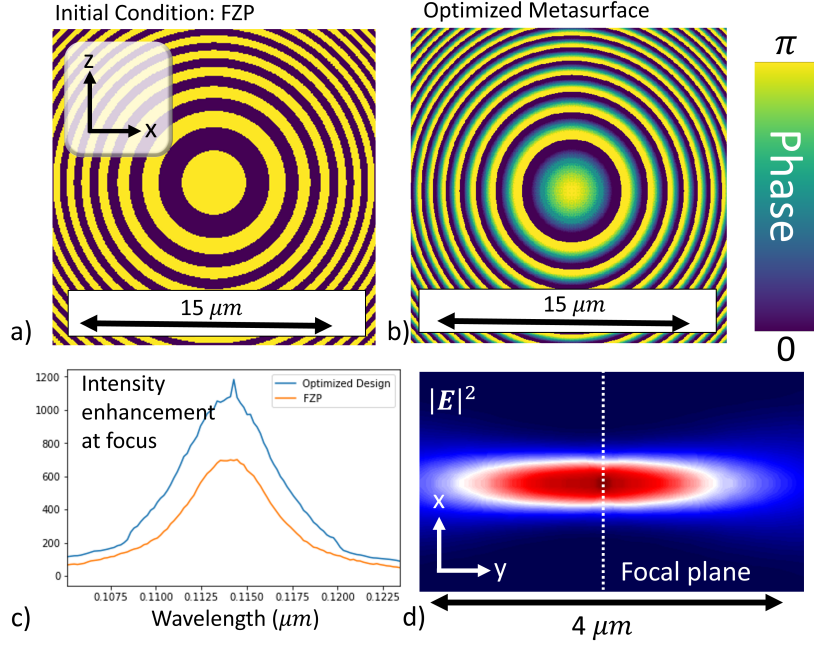


Figure 8.8: Broadband inverse design of a phase plate to maximize focusing around the seventh harmonic $\lambda \sim 114$ nm via automatic differentiation of the ASM. a) Initial condition of the inverse design is chosen to be a Fresnel zone plate tuned for the seventh harmonic. Yellow regions represent a π phase shift, and purple regions represent a zero phase shift. b) Final design optimized to maximize several wavelengths around the seventh harmonic at the focal plane. This device was not binarized. c) Seventh harmonic intensity spectrum at the focal point, normalized to the incident intensity for the initial FZP (yellow line) and the inverse-designed device (blue line). d) Electric field intensity at the focal plane, demonstrating a large focal spot.

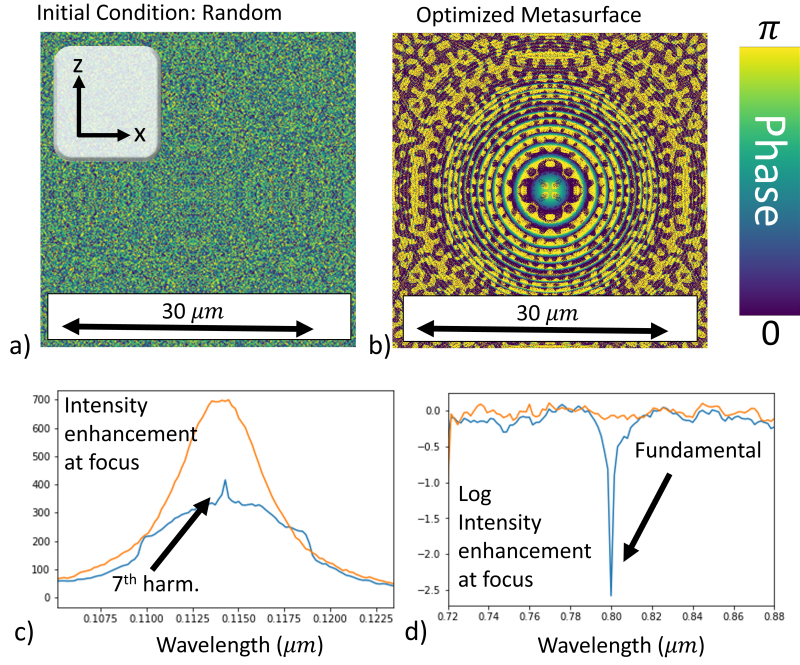


Figure 8.9: Broadband inverse design of a phase plate to maximize focusing around the seventh harmonic $\lambda \sim 114$ nm and minimize around the fundamental $\lambda = 800$ nm via automatic differentiation of the ASM. a) Initial condition of the inverse design is chosen to be a random grid of pixels ranging from π phase shift (yellow) to zero phase shift (purple). b) Final design optimized to maximize several wavelengths around the seventh harmonic and minimize the fundamental at the focal plane. c) Seventh harmonic intensity spectrum at the focal point, normalized to the incident intensity for the FZP (yellow line) and the inverse-designed device (blue line). d) Log plot of the fundamental intensity spectrum at the focal point, normalized to the incident intensity for the FZP (yellow line) and the inverse-designed device (blue line).

Parameterization/finite differences

In this section, we review what remains of the results of using finite difference gradients to design a metalens for solid-state HHG. This method was plagued by software bugs that caused our 3D-FDTD program to crash (seemingly) randomly during optimization. Despite my best efforts, the bug remained unsolved by the end of my research term. On top of the bug, the simulations required for these metalenses were massive and costly. The simulation size was $20 \times 20 \times 2 \mu m^3$, and the step sizes were $\Delta x, \Delta y, \Delta z = 40, 20, 40$ nm. The lateral step sizes were set to be larger than the propagation direction to allow for a larger design while not sacrificing the numerical dispersion in the propagation direction. The incident wavelength is $2 \mu m$, and the seventh harmonic (propagated in the ASM step) is thus ~ 286 nm. Since no automatic differentiation was required, the ASM was implemented using the Scipy python module [142] for this workflow.

Near the end of the term, one set of devices managed to optimize far enough to declare convergence. The devices were never characterized, so only images of the devices and the objective function evaluation were available at the time of writing this thesis. These are summarized in Figure 8.10, where the optimized devices are illustrated in Figure 8.10a,c,e, and the corresponding objective function evolution are adjacently plotted in Figure 8.10b,d,f. As evident from the figure, these devices have ten (a,b), six (c,d), and four (e,f) rings. A similar plot exists for eight rings, but it resembles the results of the six and ten-ring optimizations and converges on the same final objective.

Sadly these results are not very interesting upon closer inspection, as the final objectives in Figure 8.10b and d are roughly equivalent, meaning that the four outer rings had negligible effect. The result of losing two further rings has a small effect on the final objective in Figure 8.10f. Inspecting the final devices, we see that the inner four rings converge on the same location and thickness in all Figures 8.10a,c,e. It was found that closing the gaps between the rings in Figure 8.10e resulted in the same objective function - a flat surface was optimal for this set of initial conditions.

Writing this section months later, it is evident that a 2D exploration of this method should have been studied prior to trying 3D simulations. More than a single gradient-descent run is required for such a complex, nonlinear system like this. For example, at least in 2D, a global optimizer or a brute-force grid search should be applied to this problem for the global optimal ring positions rather than relying on initial conditions. That said, due to the size of the metalens, a 3D global design search would likely require a long compute

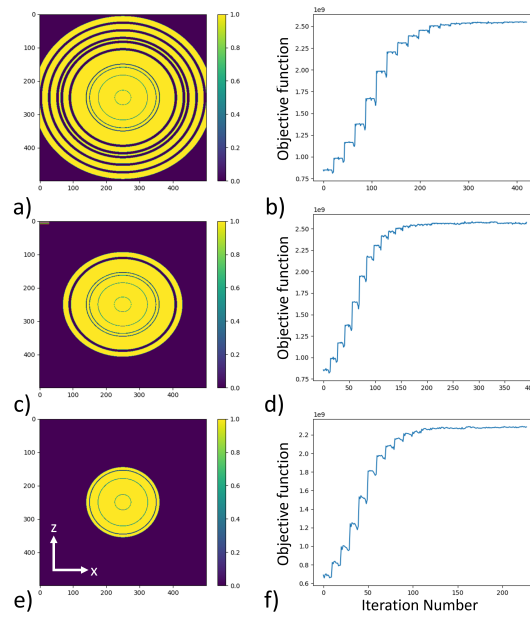


Figure 8.10: Inverse design of HHG metalens using finite difference gradients. a,c,e) Diagrams showing the optimal location and thickness of rings for 10, 6, and 4 rings, respectively. b,d,f) The evolution of the objective function during the gradient descent process for 10, 6, and 4 rings, respectively.

time (on the order of weeks/months) or a new computing system and software [134] [133]. Furthermore, implementing symmetric and antisymmetric boundary conditions would have reduced the simulation size by a factor of 4 and should be considered before trying a similar design scheme in 3D.

Although similar time constraints apply to the adjoint-based approach, the loss of parameterization and the ability to obtain the global gradient in two simulations would benefit it over the finite difference approach. However, this should be done in 2D first.

Chapter 9

Conclusion

In this thesis, we have delved into the exciting realm of computational nanophotonics, exploring various aspects of this rapidly evolving field. Through the use of advanced computational techniques, we have been able to investigate the behaviour of small plasmonic features and the nonlinear optical properties of plasmonic metasurfaces, the multi-physical and time-varying nature of epsilon-near-zero materials, the use of inverse design for optical pulse shaping and metasurfaces for high-harmonic generation, as well as harnessing the power of deep learning for predicting the optical response of plasmonic particles. The findings presented in this work have not only expanded our understanding of nanophotonics but also laid the groundwork for numerous future applications and advancements in this domain.

One of the key areas focused on was the computational modelling of small plasmonic features using nonlocal models in conjunction with the Finite-Difference Time-Domain (FDTD) method. By incorporating nonlocal effects into our classical plasmonic models, we are better able to capture patterns measured in experiments, such as the blue shifting and broadening of plasmonic resonances for increasingly smaller particles. Furthermore, nonlocality enables the smearing of the charge density near plasmonic interfaces, resulting in diminished stair-casing artifacts that plague plasmonic FDTD simulations. The application of nonlocal models is crucial in understanding the near-field enhancements and scattering phenomena in these nanostructures.

Furthermore, our investigation into nonlinear plasmonics using the hydrodynamic plasma model provided valuable insights into the nonlinear behaviour of plasmonic nanoprisms.

The hydrodynamic plasma model allowed us to examine the nonlinear scattering from triangular gold nanoprisms under intense optical excitation, validating a simpler phenomenological model in conjunction with experiments and nonlinear scattering theory. The qualitative and quantitative agreement between the hydrodynamic-FDTD simulations and experiments is remarkable, even replicating spectral features unseen from nonlinear scattering theory or the simple model, further cementing the hydrodynamic plasma model's validity in this domain. To add to this, an expression for the strength of the nonlinearity was derived using the underlying physics. The hydrodynamic description of free electron nonlinearities thus provides a compact yet accurate method for numerically simulating nonlinear plasmonic behaviour. Within our high-performance FDTD software, this model can be used for a myriad of numerical experiments inaccessible by commercial software and standard computing hardware.

Our ability to simulate complex nanophotonic systems in the time domain enabled us to explore the optical nonlinearity of degenerate semiconductors, namely indium tin oxide (ITO) in its epsilon-near-zero (ENZ) frequency range, where the real part of the permittivity is close to zero. ITO's permittivity experiences a strong, ultrafast perturbation when irradiated by intense ultrafast optical pulses, further enhanced in the ENZ bandwidth. Due to ITO's high doping concentration, this ultrafast nonlinearity is explained by applying a higher-order correction to the electronic dispersion relation to account for higher energy states and, thus, an energy-dependent plasma frequency. The changing plasma frequency explains the changing permittivity of this Drude-like material. Implementing the material model for ITO in our FDTD software enabled us to simulate experiments that were previously published in the literature with a high degree of agreement, including the time-domain response of ITO's nonlinearity and the quantitative strength of the nonlinearity, including the dependence on wavelength, incident pulse angle, and incident power. Further numerical experiments demonstrated time refraction in ITO films, where the frequency of light shifts due to the time-varying permittivity, and we predicted stronger time refraction in the reflected light rather than transmitted light (recent experiments only measured transmission results). ITO's ultrafast, strong nonlinearity has drawn interest from the wider photonics community for applications in active photonic devices where ultrafast modulation speeds are required, and numerical modelling is required for designing these devices. It is our hope that the model we present and validate in our work will be used by fellow researchers with a need for speed. Furthermore, this exploration opened us to the world of time-varying photonics and an interesting question that could be solved using inverse design.

Our exploration of inverse design techniques demonstrated the ability to tailor light-matter interactions for complex nanophotonic systems but in a unique way. This research into implementing inverse design algorithms intersected with our work in time-varying photonic materials (i.e. ITO) when a question was posed: Can optical pulses be designed to interact with time-varying materials? An answer to this question presented itself in the form of the time-domain adjoint inverse design method, which enables us to determine the sensitivity of electromagnetic quantities like the transmittance to the source current density in Maxwell's equations; we can calculate the derivative of the transmittance with respect to the source signal. Using gradient descent, we demonstrated the ability of inverse design to design light rather than matter in order to enhance light-matter interaction. We demonstrate our pulse shaping method for toy-model systems, including time-varying dielectrics and metals, and demonstrate optical amplification in some examples. Our math also shows the ability of light-matter co-optimization, where the optical pulse and the material distribution sensitivities can be extracted with the same simulations, which opens the door to some interesting future work.

Furthermore, using inverse design techniques, including automatic differentiation and adjoint sensitivity analysis, we were able to derive methods to design metasurfaces capable of focusing solid-state high-harmonic generated light, the ability of which will enable the future of ultrafast optics and attosecond physics. Although results are only available for the automatic differentiation method, the derived adjoint method opens the door to future work on this project. Furthermore, the ability to apply nonlinear transformations to linear fields and extract the gradients using automatic differentiation may also hold promise for other nonlinear photonics inverse design applications.

Perhaps one of the most intriguing aspects of the research presented here was the integration of deep learning in predicting the plasmonic response of particles. To start, we leveraged the power of artificial neural networks to predict the colours produced from laser-induced plasmonic metasurfaces, where the laser parameters and nanoparticle geometric parameters are the inputs, and the outputs are the RGB value of the colour. Furthermore, keeping up with the trends of the times, we developed a deep learning approach for the inverse problem - deriving laser parameters and nanoparticle geometric parameters for a desired colour. This work enabled a collaborative effort for the prediction of lifetime enhancement of plasmon-enhanced quantum emitters - a high-parameter system that is perfect for deep learning. Although eventually abandoned, this project enabled me to gain a much better understanding of light-matter interaction at the nanoscale.

As a follow-up project to the plasmonic colour predictions, we also determined that it was feasible to predict the radiative response of plasmonic particles using only their cross-sectional images via their multipole coefficients. After training on a large dataset for randomly shaped nanoparticles, we demonstrated that a convolutional neural network can accurately model the system and predict a spectrum of multipole coefficients for an input particle. Although further work needs to be done in optimizing the network architecture and predicting multiple multipole coefficients at once, the work presented thus far shows promise, and the project is being continued.

As we conclude this work, it is evident that computational nanophotonics is a vibrant and multifaceted field with tremendous potential for shaping the future of photonics technologies. The insights gained from our research contribute to the fundamental understanding of light-matter interactions at the nanoscale and lay the foundation for a plethora of applications, including but certainly not limited to near-field optics and microscopy, metasurfaces and metamaterials, nanophotonic sensors, phased arrays and beam steering, photonic modulators, ultrafast/attosecond technology, and photonic computing.

As we move forward, it is essential to continuously push the boundaries of computational nanophotonics using advancements in software and hardware technologies, such as advanced scientific computing programming libraries and powerful parallel processing units. Using software modules developed for artificial intelligence enables the automatic extraction of design sensitivities and the use of multiple processor types with minimal overhead. Similarly, the boom of artificial intelligence in the 21st century has inspired the development of increasingly powerful graphics processing units, which present an unseen level of computational parallizability and speed. The speed-up provided by these tools is two-fold; they enable an increase in computational speed and also an increase in development speed. Furthermore, these tools are becoming higher level; the barrier to entry is lower than ever for those who wish to solve differential equations with hardware acceleration or train a neural network. And as these tools become more user-friendly, inspiration abounds.

In conclusion, this exploration of computational nanophotonics has revealed a rich tapestry of possibilities, promising not only exciting advances in the field but also the potential to reshape the future of computational sciences more broadly.

References

- [1] James Clerk Maxwell. VIII. A dynamical theory of the electromagnetic field. *Philosophical Transactions of the Royal Society of London*, 155:459–512, January 1865. Publisher: Royal Society.
- [2] Bruce J. Hunt. Oliver Heaviside: A first-rate oddity. *Physics Today*, 65(11):48–54, November 2012.
- [3] Christopher Donaghy-Spargo. On Heaviside’s contributions to transmission line theory: waves, diffusion and energy flux. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2134):20170457, December 2018. Publisher: Royal Society.
- [4] Photonics, January 2023. Page Version ID: 1135229286.
- [5] Lukas Novotny and Bert Hecht. *Principles of Nano-Optics*. Cambridge University Press, Cambridge, 2012. OCLC: 939829538.
- [6] Gustav Mie. Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen. *Annalen der Physik*, 330(3):377–445, 1908. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.19083300302>.
- [7] Stefan A. Maier. *Plasmonics: fundamentals and applications*. Springer, New York, 2007.
- [8] Hitoshi Kuwata, Hiroharu Tamaru, Kunio Esumi, and Kenjiro Miyano. Resonant light scattering from metal nanoparticles: Practical analysis beyond Rayleigh approximation. *Applied Physics Letters*, 83(22):4625–4627, December 2003.
- [9] William L. Barnes. Particle plasmons: Why shape matters. *American Journal of Physics*, 84(8):593–601, August 2016. Publisher: American Association of Physics Teachers.

- [10] Gaëtan Lévêque and Olivier J. F. Martin. Tunable composite nanoparticle for plasmonics. *Optics Letters*, 31(18):2750–2752, September 2006. Publisher: Optica Publishing Group.
- [11] W. Dickson, G. A. Wurtz, P. Evans, D. O’Connor, R. Atkinson, R. Pollard, and A. V. Zayats. Dielectric-loaded plasmonic nanoantenna arrays: A metamaterial with tuneable optical properties. *Physical Review B*, 76(11):115411, September 2007. Publisher: American Physical Society.
- [12] Sergey V. Gaponenko and Dmitry V. Guzatov. Colloidal Plasmonics for Active Nanophotonics. *Proceedings of the IEEE*, pages 1–17, 2020. Conference Name: Proceedings of the IEEE.
- [13] Alexey V. Krasavin, Pavel Ginzburg, and Anatoly V. Zayats. Free-electron Optical Nonlinearities in Plasmonic Nanostructures: A Review of the Hydrodynamic Description. *Laser & Photonics Reviews*, 12(1):1700082, January 2018.
- [14] C. K. Chen, A. R. B. de Castro, and Y. R. Shen. Surface-Enhanced Second-Harmonic Generation. *Physical Review Letters*, 46(2):145–148, January 1981.
- [15] G. Vampa, B. G. Ghamsari, S. Siadat Mousavi, T. J. Hammond, A. Olivieri, E. Lisicka-Skretek, A. Yu Naumov, D. M. Villeneuve, A. Staudte, P. Berini, and P. B. Corkum. Plasmon-enhanced high-harmonic generation from silicon. *Nature Physics*, 13(7):659–662, July 2017. Number: 7 Publisher: Nature Publishing Group.
- [16] N C Panoiu, W E I Sha, D Y Lei, and G-C Li. Nonlinear optics in plasmonic nanostructures. *Journal of Optics*, 20(8):083001, August 2018.
- [17] Jan Renger, Romain Quidant, Niek van Hulst, and Lukas Novotny. Surface-Enhanced Nonlinear Four-Wave Mixing. *Physical Review Letters*, 104(4):046803, January 2010.
- [18] J. E. Sipe, V. C. Y. So, M. Fukui, and G. I. Stegeman. Analysis of second-harmonic generation at metal surfaces. *Physical Review B*, 21(10):4389–4402, May 1980.
- [19] M. Saad Bin-Alam, Joshua Baxter, Kashif M. Awan, Antti Kiviniemi, Yaryna Mamchur, Antonio Calà Lesina, Kosmas L. Tsakmakidis, Mikko J. Huttunen, Lora Ramunno, and Ksenia Dolgaleva. Hyperpolarizability of Plasmonic Meta-Atoms in Metasurfaces. *Nano Letters*, 21(1):51–59, January 2021. Publisher: American Chemical Society.

- [20] Brian K. Canfield, Hannu Husu, Janne Laukkanen, Benfeng Bai, Markku Kuittinen, Jari Turunen, and Martti Kauranen. Local Field Asymmetry Drives Second-Harmonic Generation in Noncentrosymmetric Nanodimers. *Nano Letters*, 7(5):1251–1255, May 2007.
- [21] Antonio Capretti, Gary F. Walsh, Salvatore Minissale, Jacob Trevino, Carlo Forestiere, Giovanni Miano, and Luca Dal Negro. Multipolar second harmonic generation from planar arrays of Au nanoparticles. *Optics Express*, 20(14):15797–15806, July 2012. Publisher: Optica Publishing Group.
- [22] Álvaro Rodríguez Echarri, Joel D. Cox, Fadil Iyikanat, and F. Javier García de Abajo. Nonlinear plasmonic response in atomically thin metal films. *Nanophotonics*, 10(16):4149–4159, November 2021. Publisher: De Gruyter.
- [23] Mikko J. Huttunen, Orad Reshef, Timo Stolt, Ksenia Dolgaleva, Robert W. Boyd, and Martti Kauranen. Efficient nonlinear metasurfaces by using multiresonant high-Q plasmonic arrays. *JOSA B*, 36(7):E30–E35, July 2019. Publisher: Optica Publishing Group.
- [24] Robert Czaplicki, Antti Kiviniemi, Mikko J. Huttunen, Xiaorun Zang, Timo Stolt, Ismo Vartiainen, Jérémy Butet, Markku Kuittinen, Olivier J. F. Martin, and Martti Kauranen. Less Is More: Enhancement of Second-Harmonic Generation from Metasurfaces by Reduced Nanoparticle Density. *Nano Letters*, 18(12):7709–7714, December 2018.
- [25] F. Bloch. Bremsvermögen von Atomen mit mehreren Elektronen. *Zeitschrift für Physik*, 81(5):363–376, May 1933.
- [26] R. H. Ritchie. Plasma Losses by Fast Electrons in Thin Films. *Physical Review*, 106(5):874–881, June 1957.
- [27] Ming Fang, Zhixiang Huang, Wei E. I. Sha, and Xianliang Wu. Maxwell–Hydrodynamic Model for Simulating Nonlinear Terahertz Generation From Plasmonic Metasurfaces. *IEEE Journal on Multiscale and Multiphysics Computational Techniques*, 2:194–201, 2017.
- [28] Andreas Hille, Matthias Moferdt, Christian Wolff, Christian Matyssek, Rogelio Rodríguez-Oliveros, Cristopher Prohm, Jens Niegemann, Stefan Grafström, Lukas M. Eng, and Kurt Busch. Second Harmonic Generation from Metal Nano-Particle Resonators: Numerical Analysis On the Basis of the Hydrodynamic Drude Model. *The Journal of Physical Chemistry C*, 120(2):1163–1169, January 2016.

- [29] Kirankumar R. Hiremath, Lin Zschiedrich, and Frank Schmidt. Numerical solution of nonlocal hydrodynamic Drude model for arbitrary shaped nano-plasmonic structures using Nédélec finite elements. *Journal of Computational Physics*, 231(17):5890–5896, July 2012.
- [30] Jinjie Liu, Moysey Brio, Yong Zeng, Armis R. Zakharian, Walter Hoyer, Stephan W. Koch, and Jerome V. Moloney. Generalization of the FDTD algorithm for simulations of hydrodynamic nonlinear Drude model. *Journal of Computational Physics*, 229(17):5921–5932, August 2010.
- [31] I. Tokatly and O. Pankratov. Hydrodynamic theory of an electron gas. *Physical Review B*, 60(23):15550–15553, December 1999.
- [32] Cristian Ciraci, Ekaterina Poutrina, Michael Scalora, and David R. Smith. Origin of second-harmonic generation enhancement in optical split-ring resonators. *Physical Review B*, 85(20):201403, May 2012.
- [33] Cristian Ciraci, Ekaterina Poutrina, Michael Scalora, and David R. Smith. Second-harmonic generation in metallic nanoparticles: Clarification of the role of the surface. *Physical Review B*, 86(11):115451, September 2012.
- [34] L. Rodríguez-Suné, M. Scalora, A. S. Johnson, C. Cojocaru, N. Akozbek, Z. J. Cop-pens, D. Perez-Salinas, S. Wall, and J. Trull. Study of second and third harmonic generation from an indium tin oxide nanolayer: Influence of nonlocal effects and hot electrons. *APL Photonics*, 5(1):010801, January 2020.
- [35] M. Scalora, M. A. Vincenti, D. de Ceglia, V. Roppo, M. Centini, N. Akozbek, and M. J. Bloemer. Second- and third-harmonic generation in metal-based structures. *Physical Review A*, 82(4):043828, October 2010.
- [36] Martti Kauranen and Anatoly V. Zayats. Nonlinear plasmonics. *Nature Photonics*, 6(11):737–748, November 2012.
- [37] Søren Raza, Sergey I Bozhevolnyi, Martijn Wubs, and N Asger Mortensen. Nonlocal optical response in metallic nanostructures. *Journal of Physics: Condensed Matter*, 27(18):183204, May 2015.
- [38] Andrew R. Melnyk and Michael J. Harrison. Resonant Excitation of Plasmons in Thin Films by Electromagnetic Waves. *Physical Review Letters*, 21(2):85–88, July 1968. Publisher: American Physical Society.

- [39] I. Lindau and P. O. Nilsson. Experimental evidence for excitation of longitudinal plasmons by photons. *Physics Letters A*, 31(7):352–353, April 1970.
- [40] M. Anderegg, B. Feuerbacher, and B. Fitton. Optically Excited Longitudinal Plasmons in Potassium. *Physical Review Letters*, 27(23):1565–1568, December 1971. Publisher: American Physical Society.
- [41] U. Kreibig and C. v. Fragstein. The limitation of electron mean free path in small silver particles. *Zeitschrift fr Physik*, 224(4):307–323, August 1969.
- [42] H. C. van de Hulst. *Light scattering by small particles*. Dover Publications, New York, 1981.
- [43] R. Rupp. Optical Properties of a Plasma Sphere. *Physical Review Letters*, 31(24):1434–1437, December 1973.
- [44] R. Rupp. Optical properties of small metal spheres. *Physical Review B*, 11(8):2871–2876, April 1975.
- [45] Josef Tiggesbäumker, Lars Köller, Karl-Heinz Meiwes-Broer, and Ansgar Liebsch. Blue shift of the Mie plasma frequency in Ag clusters and particles. *Physical Review A*, 48(3):R1749–R1752, September 1993. Publisher: American Physical Society.
- [46] K.-P. Charlé, L. König, S. Nepijko, I. Rabin, and W. Schulze. The Surface Plasmon Resonance of Free and Embedded Ag-Clusters in the Size Range $1,5 \text{ nm} < D < 30 \text{ nm}$. *Crystal Research and Technology*, 33(7-8):1085–1096, 1998. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291521-4079%28199810%2933%3A7%3C1085%3A%3AAID-CRAT1085%3E3.0.CO%3B2-A>.
- [47] Søren Raza, Nicolas Stenger, Shima Kadkhodazadeh, Søren V. Fischer, Natalie Kostashe, Antti-Pekka Jauho, Andrew Burrows, Martijn Wubs, and N. Asger Mortensen. Blueshift of the surface plasmon resonance in silver nanoparticles studied with EELS. *Nanophotonics*, 2(2):131–138, 2013.
- [48] Jonathan A. Scholl, Ai Leen Koh, and Jennifer A. Dionne. Quantum plasmon resonances of individual metallic nanoparticles. *Nature*, 483(7390):421–427, March 2012.
- [49] J. J. Hopfield and D. G. Thomas. Theoretical and Experimental Effects of Spatial Dispersion on the Optical Properties of Crystals. *Physical Review*, 132(2):563–572, October 1963.

- [50] P. Halevi and R. Fuchs. Generalised additional boundary condition for non-local dielectrics. I. Reflectivity. *Journal of Physics C: Solid State Physics*, 17(21):3869–3888, July 1984.
- [51] R. Fuchs and R. G. Barrera. Dynamical response of a dipole near the surface of a nonlocal metal. *Physical Review B*, 24(6):2940–2950, September 1981.
- [52] W. H. Weber and G. W. Ford. Enhanced Raman Scattering by Adsorbates Including the Nonlocal Response of the Metal and the Excitation of Nonradiative Modes. *Physical Review Letters*, 44(26):1774–1777, June 1980.
- [53] Andrew R. Melnyk and Michael J. Harrison. Theory of Optical Excitation of Plasmons in Metals. *Physical Review B*, 2(4):835–850, August 1970.
- [54] Alan J. Bennett. Influence of the Electron Charge Distribution on Surface-Plasmon Dispersion. *Physical Review B*, 1(1):203–207, January 1970.
- [55] A. D. Boardman, B. V. Paranjape, and Y. O. Nakamura. Surface Plasmon-Polaritons in a Spatially Dispersive Inhomogeneous Medium. *physica status solidi (b)*, 75(1):347–359, May 1976.
- [56] Adolfo Eguiluz and J. J. Quinn. Hydrodynamic model for surface plasmons in metals and degenerate semiconductors. *Physical Review B*, 14(4):1347–1361, August 1976. Publisher: American Physical Society.
- [57] Jeffrey M. McMahon, Stephen K. Gray, and George C. Schatz. Nonlocal Optical Response of Metal Nanostructures with Arbitrary Shape. *Physical Review Letters*, 103(9):097403, August 2009.
- [58] Jeffrey M. McMahon, Stephen K. Gray, and George C. Schatz. Calculating non-local optical properties of structures with arbitrary shape. *Physical Review B*, 82(3):035423, July 2010.
- [59] Jeffrey Michael McMahon. *Topics in Theoretical and Computational Nanoscience: From Controlling Light at the Nanoscale to Calculating Quantum Effects with Classical Electrodynamics*. Springer Theses. Springer-Verlag, New York, 2011.
- [60] Søren Raza, Giuseppe Toscano, Antti-Pekka Jauho, Martijn Wubs, and N. Asger Mortensen. Unusual resonances in nanoplasmonic structures due to nonlocal response. *Physical Review B*, 84(12):121412, September 2011.

- [61] Aeneas Wiener, Antonio I. Fernández-Domínguez, Andrew P. Horsfield, John B. Pendry, and Stefan A. Maier. Nonlocal Effects in the Nanofocusing Performance of Plasmonic Tips. *Nano Letters*, 12(6):3308–3314, June 2012.
- [62] N. Schmitt, J. Viquerat, C. Scheid, S. Lanteri, M. Moefertdt, and K. Busch. A 3D discontinuous Galerkin Time-Domain method for nano plasmonics with a nonlocal dispersion model. In *2017 Progress In Electromagnetics Research Symposium - Spring (PIERS)*, pages 2792–2799, May 2017.
- [63] Xuezhi Zheng, Mario Kupresak, Raj Mittra, and Guy A. E. Vandenbosch. A Boundary Integral Equation Scheme for Simulating the Nonlocal Hydrodynamic Response of Metallic Antennas at Deep-Nanometer Scales. *IEEE Transactions on Antennas and Propagation*, 66(9):4759–4771, September 2018.
- [64] Søren Raza, Martijn Wubs, Sergey I. Bozhevolnyi, and N. Asger Mortensen. Nonlocal study of ultimate plasmon hybridization. *Optics Letters*, 40(5):839, March 2015.
- [65] C. Ciraci, R. T. Hill, J. J. Mock, Y. Urzhumov, A. I. Fernandez-Dominguez, S. A. Maier, J. B. Pendry, A. Chilkoti, and D. R. Smith. Probing the Ultimate Limits of Plasmonic Enhancement. *Science*, 337(6098):1072–1074, August 2012.
- [66] Jeffrey M. McMahon, Stephen K. Gray, and George C. Schatz. Optical Properties of Nanowire Dimers with a Spatially Nonlocal Dielectric Function. *Nano Letters*, 10(9):3473–3481, September 2010.
- [67] N. A. Mortensen, S. Raza, M. Wubs, T. Søndergaard, and S. I. Bozhevolnyi. A generalized non-local optical response theory for plasmonic nanostructures. *Nature Communications*, 5(1):3809, September 2014.
- [68] Christos Tserkezis, Johan R. Maack, Zhaowei Liu, Martijn Wubs, and N. Asger Mortensen. Robustness of the far-field response of nonlocal plasmonic ensembles. *Scientific Reports*, 6(1):28441, June 2016. Number: 1 Publisher: Nature Publishing Group.
- [69] C. Tserkezis, N. Asger Mortensen, and Martijn Wubs. How nonlocal damping reduces plasmon-enhanced fluorescence in ultranarrow gaps. *Physical Review B*, 96(8):085413, August 2017. Publisher: American Physical Society.
- [70] Harini Hapuarachchi and Jared H. Cole. Influence of a planar metal nanoparticle assembly on the optical response of a quantum emitter. *Physical Review Research*, 2(4):043092, October 2020. Publisher: American Physical Society.

- [71] Harini Hapuarachchi, Malin Premaratne, Qiaoliang Bao, Wenlong Cheng, Sarath D. Gunapala, and Govind P. Agrawal. Cavity QED analysis of an exciton-plasmon hybrid molecule via the generalized nonlocal optical response method. *Physical Review B*, 95(24):245419, June 2017.
- [72] Dasuni Lelwala Gamacharige, Sarath D. Gunapala, Mark I. Stockman, and Malin Premaratne. Significance of the nonlocal optical response of metal nanoparticles in describing the operation of plasmonic lasers. *Physical Review B*, 99(11):115405, March 2019. Publisher: American Physical Society.
- [73] Champi Abeywickrama, Malin Premaratne, and David L. Andrews. Analysis of Förster resonance energy transfer (FRET) in the vicinity of a charged metallic nanosphere via nonlocal optical response method. In *Nanophotonics VIII*, volume 11345, pages 75–84. SPIE, April 2020.
- [74] Sergejs Boroviks, Zhan-Hong Lin, Vladimir A. Zenin, Mario Ziegler, Andrea Dellith, P. a. D. Gonçalves, Christian Wolff, Sergey I. Bozhevolnyi, Jer-Shing Huang, and N. Asger Mortensen. Extremely confined gap plasmon modes: when nonlocality matters. *Nature Communications*, 13(1):3105, June 2022. Number: 1 Publisher: Nature Publishing Group.
- [75] Joshua Baxter, Antonio Calà Lesina, and Lora Ramunno. Parallel FDTD Modeling of Nonlocality in Plasmonics. *IEEE Transactions on Antennas and Propagation*, 69(7):3982–3994, July 2021. Conference Name: IEEE Transactions on Antennas and Propagation.
- [76] Benjamin Gallinet, Jérémy Butet, and Olivier J. F. Martin. Numerical methods for nanophotonics: standard problems and future challenges: Numerical methods for nanophotonics. *Laser & Photonics Reviews*, 9(6):577–603, November 2015.
- [77] Antonino Calà Lesina, Alessandro Vaccari, Pierre Berini, and Lora Ramunno. On the convergence and accuracy of the FDTD method for nanoplasmonics. *Optics Express*, 23(8):10481–10497, April 2015.
- [78] Eric G. Johnson, Alan D. Kathman, Diane H. Hochmuth, A. L. Cook, David Renick Brown, and William F. Delaney. Advantages of genetic algorithm optimization methods in diffractive optic design. In *Diffractive and Miniaturized Optics: A Critical Review*, volume 10271, pages 56–76. SPIE, December 1993.

- [79] Eric G. Johnson and Mustafa A. G. Abushagur. Microgenetic-algorithm optimization methods applied to dielectric gratings. *JOSA A*, 12(5):1152–1160, May 1995. Publisher: Optica Publishing Group.
- [80] Jared D. Stack and Michael R. Feldman. Recursive mean-squared-error algorithm for iterative discrete on-axis encoded holograms. *Applied Optics*, 31(23):4839–4846, August 1992. Publisher: Optica Publishing Group.
- [81] Ian M. Barton and Mohammad R. Taghizadeh. Application of optimization algorithms to the design of diffractive optical elements for custom laser resonators. *Optics Letters*, 23(3):198–200, February 1998. Publisher: Optica Publishing Group.
- [82] V. V Kotlyar, P. G Seraphimovich, and V. A Soifer. An iterative algorithm for designing diffractive optical elements with regularization. *Optics and Lasers in Engineering*, 29(4):261–268, May 1998.
- [83] M.M. Spuhler, B.J. Offrein, G.-L. Bona, R. Germann, I. Massarek, and D. Erni. A very short planar silica spot-size converter using a nonperiodic segmented waveguide. *Journal of Lightwave Technology*, 16(9):1680–1685, September 1998.
- [84] Steven J. Cox and David C. Dobson. Maximizing Band Gaps in Two-Dimensional Photonic Crystals. *SIAM Journal on Applied Mathematics*, 59(6):2108–2120, 1999.
- [85] Steven J. Cox and David C. Dobson. Band Structure Optimization of Two-Dimensional Photonic Crystals in H-Polarization. *Journal of Computational Physics*, 158(2):214–224, March 2000.
- [86] Thomas Felici and Heinz W Engl. On shape optimization of optical waveguides using inverse problem techniques. *Inverse Problems*, 17(4):1141–1162, August 2001.
- [87] Stefan Preble, Michal Lipson, and Hod Lipson. Two-dimensional photonic crystals designed by evolutionary algorithms. *Applied Physics Letters*, 86(6):061111, February 2005.
- [88] Jakob S. Jensen and Ole Sigmund. Systematic design of photonic crystal structures using topology optimization: Low-loss waveguide bends. *Applied Physics Letters*, 84(12):2022–2024, March 2004.
- [89] P. I. Borel, A. Harpøth, L. H. Frandsen, M. Kristensen, P. Shi, J. S. Jensen, and O. Sigmund. Topology optimization and fabrication of photonic crystal structures. *Optics Express*, 12(9):1996, May 2004.

- [90] Martin Burger, Stanley J Osher, and Eli Yablonovitch. Inverse Problem Techniques for the Design of Photonic Crystals. *IEICE TRANS. ELECTRON.*, page 8, 2004.
- [91] Martin Burger and Stanley J. Osher. A survey on level set methods for inverse problems and optimal design. *European Journal of Applied Mathematics*, 16(2):263–301, April 2005. Publisher: Cambridge University Press.
- [92] Owen Dennis Miller. Photonic Design: From Fundamental Solar Cell Physics to Computational Inverse Design. page 137.
- [93] Jakob S. Jensen and Ole Sigmund. Topology optimization of photonic crystal structures: a high-bandwidth low-loss T-junction waveguide. *Journal of the Optical Society of America B*, 22(6):1191, June 2005.
- [94] C. Y. Kao, S. Osher, and E. Yablonovitch. Maximizing band gaps in two-dimensional photonic crystals by using level set methods. *Applied Physics B*, 81(2-3):235–244, July 2005.
- [95] Yang Jiao, Shanhui Fan, and David A. B. Miller. Photonic crystal device sensitivity analysis with Wannier basis gradients. *Optics Letters*, 30(3):302, February 2005.
- [96] Yang Jiao, Shanhui Fan, and D.A.B. Miller. Systematic photonic crystal device design: global and local optimization and sensitivity analysis. *IEEE Journal of Quantum Electronics*, 42(3):266–279, March 2006. Conference Name: IEEE Journal of Quantum Electronics.
- [97] J.S. Jensen and O. Sigmund. Topology optimization for nano-photonics. *Laser & Photonics Reviews*, 5(2):308–321, March 2011.
- [98] Jesse Lu and Jelena Vučković. Nanophotonic computational design. *Optics Express*, 21(11):13351, June 2013.
- [99] Alexander Y. Piggott, Jesse Lu, Thomas M. Babinec, Konstantinos G. Lagoudakis, Jan Petykiewicz, and Jelena Vučković. Inverse design and implementation of a wavelength demultiplexing grating coupler. *Scientific Reports*, 4(1):7210, November 2014. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Applied mathematics;Optical materials and structures Subject_term_id: applied-mathematics;optical-materials-and-structures.

- [100] Alexander Y. Piggott, Jesse Lu, Konstantinos G. Lagoudakis, Jan Petykiewicz, Thomas M. Babinec, and Jelena Vučković. Inverse design and demonstration of a compact and broadband on-chip wavelength demultiplexer. *Nature Photonics*, 9(6):374–377, June 2015.
- [101] Logan Su, Alexander Y. Piggott, Neil V. Saprà, Jan Petykiewicz, and Jelena Vučković. Inverse Design and Demonstration of a Compact on-Chip Narrowband Three-Channel Wavelength Demultiplexer. *ACS Photonics*, 5(2):301–305, February 2018.
- [102] Christopher M. Lalau-Keraly, Samarth Bhargava, Owen D. Miller, and Eli Yablonovitch. Adjoint shape optimization applied to electromagnetic design. *Optics Express*, 21(18):21693, September 2013.
- [103] Photonic Inverse Design.
- [104] Bing Shen, Peng Wang, Randy Polson, and Rajesh Menon. An integrated-nanophotonics polarization beamsplitter with 2.4×2.4 m² footprint. *Nature Photonics*, 9(6):378–382, June 2015.
- [105] David Sell, Jianji Yang, Sage Doshay, and Jonathan A. Fan. Periodic Dielectric Metasurfaces with High-Efficiency, Multiwavelength Functionalities. *Advanced Optical Materials*, 5(23):1700645, December 2017.
- [106] David Sell, Jianji Yang, Sage Doshay, Rui Yang, and Jonathan A. Fan. Large-Angle, Multifunctional Metagratings Based on Freeform Multimode Geometries. *Nano Letters*, 17(6):3752–3757, June 2017.
- [107] Raphaël Pestourie, Carlos Pérez-Arancibia, Zin Lin, Wonseok Shin, Federico Capasso, and Steven G. Johnson. Inverse design of large-area metasurfaces. *Optics Express*, 26(26):33732–33747, December 2018. Publisher: Optica Publishing Group.
- [108] Adam S. Backer. Computational inverse design for cascaded systems of metasurface optics. *Optics Express*, 27(21):30308, October 2019.
- [109] Jinhie Skarda, Rahul Trivedi, Logan Su, Diego Ahmad-Stein, Hyounghan Kwon, Seunghoon Han, Shanhui Fan, and Jelena Vučković. Low-overhead distribution strategy for simulation and optimization of large-area metasurfaces. *npj Computational Materials*, 8(1):78, December 2022.
- [110] Machine learning, October 2022. Page Version ID: 1116598718.

- [111] A. Patnaik, R.K. Mishra, G.K. Patra, and S.K. Dash. An artificial neural network model for effective dielectric constant of microstrip line. *IEEE Transactions on Antennas and Propagation*, 45(11):1697, November 1997.
- [112] Dragan A Cirovic. Feed-forward artificial neural networks: applications to spectroscopy. *TrAC Trends in Analytical Chemistry*, 16(3):148–155, March 1997.
- [113] A.H. Zaabab, Qi-Jun Zhang, and M. Nakhla. A neural network modeling approach to circuit optimization and statistical design. *IEEE Transactions on Microwave Theory and Techniques*, 43(6):1349–1358, June 1995.
- [114] P.M. Watson and K.C. Gupta. EM-ANN models for microstrip vias and interconnects in dataset circuits. *IEEE Transactions on Microwave Theory and Techniques*, 44(12):2495–2503, December 1996.
- [115] H.L. Southall, J.A. Simmers, and T.H. O'Donnell. Direction finding in phased arrays with a neural network beamformer. *IEEE Transactions on Antennas and Propagation*, 43(12):1369–1374, December 1995.
- [116] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554, July 2006.
- [117] Wei Ma, Zhaocheng Liu, Zhaxylyk A. Kudyshev, Alexandra Boltasseva, Wenshan Cai, and Yongmin Liu. Deep learning for the design of photonic structures. *Nature Photonics*, 15(2):77–90, February 2021.
- [118] Itzik Malkiel, Michael Mrejen, Achiya Nagler, Uri Arieli, Lior Wolf, and Haim Suchowski. Plasmonic nanostructure design and characterization via Deep Learning. *Light: Science & Applications*, 7(1):60, September 2018.
- [119] Dianjing Liu, Yixuan Tan, Erfan Khoram, and Zongfu Yu. Training Deep Neural Networks for the Inverse Design of Nanophotonic Structures. *ACS Photonics*, 5(4):1365–1369, April 2018.
- [120] Wei Ma, Feng Cheng, and Yongmin Liu. Deep-Learning-Enabled On-Demand Design of Chiral Metamaterials. *ACS Nano*, 12(6):6326–6334, June 2018.
- [121] John Peurifoy, Yichen Shen, Li Jing, Yi Yang, Fidel Cano-Renteria, Brendan G. DeLacy, John D. Joannopoulos, Max Tegmark, and Marin Soljačić. Nanophotonic particle simulation and inverse design using artificial neural networks. *Science Advances*, 4(6):eaar4206, June 2018.

- [122] Iman Sajedian, Trevon Badloe, and Junsuk Rho. Optimisation of colour generation from dielectric nanostructures using reinforcement learning. *Optics Express*, 27(4):5874–5883, February 2019.
- [123] Iman Sajedian, Jeonghyun Kim, and Junsuk Rho. Finding the optical properties of plasmonic structures by image processing using a combination of convolutional neural networks and recurrent neural networks. *Microsystems & Nanoengineering*, 5(1):27, December 2019.
- [124] J. Baxter, J. Desautels, and L. Ramunno. Deep Learning for the Prediction of Multi poles. In *2021 Photonics North (PN)*, pages 1–1, May 2021. ISSN: 2693-8316.
- [125] Kane Yee. Numerical solution of initial boundary value problems involving maxwell’s equations in isotropic media. *IEEE Transactions on Antennas and Propagation*, 14(3):302–307, May 1966.
- [126] MATRIX METHODS FOR SOLVING FIELD PROBLEMS. VOLUME 1. MATRIX TECHNIQUES AND APPLICATIONS. Technical report. Section: Technical Reports.
- [127] Allen Taflove. A Perspective on the 40-Year History of FDTD Computational Electrodynamics. 22(1):21, 2007.
- [128] Allen Taflove. Application of the Finite-Difference Time-Domain Method to Sinusoidal Steady-State Electromagnetic-Penetration Problems. *IEEE Transactions on Electromagnetic Compatibility*, EMC-22(3):191–202, August 1980. Conference Name: IEEE Transactions on Electromagnetic Compatibility.
- [129] Allen Taflove and Susan C. Hagness. *Computational electrodynamics: the finite-difference time-domain method*. Artech House antennas and propagation library. Artech House, Boston, 3rd ed edition, 2005.
- [130] M. Livesey, J. F. Stack, F. Costen, T. Nanri, N. Nakashima, and S. Fujino. Development of a CUDA Implementation of the 3D FDTD Method. *IEEE Antennas and Propagation Magazine*, 54(5):186–195, October 2012.
- [131] Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon

- Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2015.
- [132] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.
 - [133] Tidy3D: Lightning-fast Python-driven FDTD at Scale.
 - [134] Tyler W. Hughes, Momchil Minkov, Victor Liu, Zongfu Yu, and Shanhui Fan. A perspective on the pathway toward full wave simulation of large area metalenses. *Applied Physics Letters*, 119(15):150502, October 2021. Publisher: American Institute of Physics.
 - [135] Y. R. Shen. *The principles of nonlinear optics*. Wiley classics library. Wiley-Interscience, Hoboken, N.J, wiley classics library ed edition, 2003.
 - [136] John David Jackson. *Classical electrodynamics*. Wiley, New York, 3rd ed edition, 1999.
 - [137] P. B. Johnson and R. W. Christy. Optical Constants of the Noble Metals. *Physical Review B*, 6(12):4370–4379, December 1972. Publisher: American Physical Society.
 - [138] Mikhail N. Polyanskiy. Refractive index database.
 - [139] A. D. Boardman, editor. *Electromagnetic surface modes*. Wiley, Chichester ; New York, 1982.
 - [140] Joshua Baxter, Adriana Pérez-Casanova, Luis Cortes-Herrera, Antonio Calà Lesina, Israel De Leon, and Lora Ramunno. Understanding the nonlinear optical response of epsilon near zero materials in the time-domain. *arXiv:2110.14806 [physics]*, October 2021. arXiv: 2110.14806.
 - [141] The Feynman Lectures on Physics Vol. II Ch. 7: The Electric Field in Various Circumstances (Continued).
 - [142] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan

- Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.
- [143] David J. Griffiths. *Introduction to electrodynamics*. Pearson, Boston, fourth edition edition, 2013.
 - [144] Robert W. Boyd. *Nonlinear optics*. Academic Press, Amsterdam ; Boston, 3rd ed edition, 2008.
 - [145] J. Leng, J. Opsal, H. Chu, M. Senko, and D.E. Aspnes. Analytic representations of the dielectric functions of materials for device and structural modeling. *Thin Solid Films*, 313-314:132–136, February 1998.
 - [146] P. G. Etchegoin, E. C. Le Ru, and M. Meyer. An analytic model for the optical properties of gold. *The Journal of Chemical Physics*, 125(16):164705, October 2006.
 - [147] A. Vial and T. Laroche. Comparison of gold and silver dispersion laws suitable for FDTD simulations. *Applied Physics B*, 93(1):139–143, October 2008.
 - [148] A. Vial, T. Laroche, M. Dridi, and L. Le Cunff. A new model of dispersion for metals leading to a more accurate modeling of plasmonic structures using the FDTD method. *Applied Physics A*, 103(3):849–853, June 2011.
 - [149] K. P. Prokopidis and D. C. Zografopoulos. A Unified FDTD/PML Scheme Based on Critical Points for Accurate Studies of Plasmonic Structures. *Journal of Lightwave Technology*, 31(15):2467–2476, August 2013.
 - [150] Wolfram H. P. Pernice. Finite-Difference Time-Domain Methods and Material Models for the Simulation of Metallic and Plasmonic Structures. *Journal of Computational and Theoretical Nanoscience*, 7(1):1–14, January 2010.
 - [151] Benny Lautrup. *Physics of continuous matter: exotic and everyday phenomena in the macroscopic world*. Taylor & Francis, Boca Raton, 2nd ed edition, 2011.
 - [152] Zh. A. Moldabekov, M. Bonitz, and T. S. Ramazanov. Theoretical foundations of quantum hydrodynamics for plasmas. *Physics of Plasmas*, 25(3):031903, March 2018.

- [153] Giuseppe Toscano, Jakob Straubel, Alexander Kwiatkowski, Carsten Rockstuhl, Ferdinand Evers, Hongxing Xu, N. Asger Mortensen, and Martijn Wubs. Resonance shifts and spill-out effects in self-consistent hydrodynamic nanoplasmonics. *Nature Communications*, 6(1):7132, May 2015. Number: 1 Publisher: Nature Publishing Group.
- [154] Xiaoming Li, Hui Fang, Xiaoyu Weng, Lichao Zhang, Xiujie Dou, Aiping Yang, and Xiaocong Yuan. Electronic spill-out induced spectral broadening in quantum hydrodynamic nanoplasmonics. *Optics Express*, 23(23):29738, November 2015.
- [155] Alec M. Hammond, Ardavan Oskooi, Mo Chen, Zin Lin, Steven G. Johnson, and Stephen E. Ralph. High-performance hybrid time/frequency-domain topology optimization for large-scale photonics inverse design. *Optics Express*, 30(3):4467, January 2022.
- [156] B. S. Lazarov and O. Sigmund. Filters in topology optimization based on Helmholtz-type differential equations. *International Journal for Numerical Methods in Engineering*, 86(6):765–781, 2011. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/nme.3072>.
- [157] Boyan S. Lazarov and Fengwen Wang. Maximum length scale in density based topology optimization. *Computer Methods in Applied Mechanics and Engineering*, 318:826–844, May 2017.
- [158] Rasmus E. Christiansen, Rasmus E. Christiansen, Ole Sigmund, and Ole Sigmund. Inverse design in photonics by topology optimization: tutorial. *JOSA B*, 38(2):496–509, February 2021. Publisher: Optical Society of America.
- [159] Tyler W. Hughes, Momchil Minkov, Ian A. D. Williamson, and Shanhui Fan. Adjoint Method and Inverse Design for Nonlinear Nanophotonic Devices. *ACS Photonics*, 5(12):4781–4787, December 2018.
- [160] Steven G. Johnson. The NLopt nonlinear-optimization package.
- [161] Di Zhu, Michel Bosman, and Joel K. W. Yang. A circuit model for plasmonic resonators. *Optics Express*, 22(8):9809, April 2014.
- [162] Luis Enrique Cortes Herrera. Self-Consistent Numerical Modeling of the Large Hot-Electron-Induced Optical Nonlinearity of Indium Tin Oxide in its Near-Zero-Permittivity Regime. page 169.

- [163] Jean-Michel Guay, Antonino Calà Lesina, Guillaume Côté, Martin Charron, Daniel Poitras, Lora Ramunno, Pierre Berini, and Arnaud Weck. Laser-induced plasmonic colours on metals. *Nature Communications*, 8:16095, July 2017.
- [164] Jean-Michel Guay, Antonino Calà Lesina, Joshua Baxter, Graham Killaire, Lora Ramunno, Pierre Berini, and Arnaud Weck. Topography Tuning for Plasmonic Color Enhancement via Picosecond Laser Bursts. *Advanced Optical Materials*, 6(17):1800189, September 2018.
- [165] Xiaozhong Li, Jing Shu, Wenhua Gu, and Li Gao. Deep neural network for plasmonic sensor modeling. *Optical Materials Express*, 9(9):3857–3862, September 2019. Publisher: Optica Publishing Group.
- [166] Sunae So, Jungho Mun, and Junsuk Rho. Simultaneous Inverse Design of Materials and Structures via Deep Learning: Demonstration of Dipole Resonance Engineering Using Core–Shell Nanoparticles. *ACS Applied Materials & Interfaces*, 11(27):24264–24268, July 2019. Publisher: American Chemical Society.
- [167] Omid Hemmatyar, Sajjad Abdollahramezani, Yashar Kiarashinejad, Mohammadreza Zandehshahvar, and Ali Adibi. Full color generation with Fano-type resonant HfO₂ nanopillars designed by a deep-learning approach. *Nanoscale*, 11(44):21266–21274, November 2019. Publisher: The Royal Society of Chemistry.
- [168] Nathan Bryn Roberts and Mehdi Keshavarz Hedayati. A deep learning approach to the forward prediction and inverse design of plasmonic metasurface structural color. *Applied Physics Letters*, 119(6):061101, August 2021. Publisher: American Institute of Physics.
- [169] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.
- [170] Sunae So and Junsuk Rho. Designing nanophotonic structures using conditional deep convolutional generative adversarial networks. *Nanophotonics*, 8(7):1255–1261, July 2019. Publisher: De Gruyter.
- [171] Peter R. Wiecha and Otto L. Muskens. Deep Learning Meets Nanophotonics: A Generalized Accurate Predictor for Near Fields and Far Fields of Arbitrary 3D Nanostructures. *Nano Letters*, 20(1):329–338, January 2020.

- [172] Ronghui Lin, Yanfen Zhai, Chenxin Xiong, and Xiaohang Li. Inverse design of plasmonic metasurfaces by convolutional neural network. *Optics Letters*, 45(6):1362, March 2020.
- [173] Andrew Lininger, Michael Hinczewski, and Giuseppe Strangi. General Inverse Design of Layered Thin-Film Materials with Convolutional Neural Networks. *ACS Photonics*, 8(12):3641–3650, December 2021.
- [174] Takashi Asano and Susumu Noda. Optimization of photonic crystal nanocavities based on deep learning. *Optics Express*, 26(25):32704, December 2018.
- [175] Daria Smirnova and Yuri S. Kivshar. Multipolar nonlinear nanophotonics. *Optica*, 3(11):1241, November 2016.
- [176] Wei Liu and Yuri S. Kivshar. Multipolar interference effects in nanophotonics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 375(2090):20160317, March 2017. Publisher: Royal Society.
- [177] Shambhu Ghimire, Anthony D. DiChiara, Emily Sistrunk, Pierre Agostini, Louis F. DiMauro, and David A. Reis. Observation of high-order harmonic generation in a bulk crystal. *Nature Physics*, 7(2):138–141, February 2011.
- [178] A. Korobenko, S. Saha, A. T. K. Godfrey, M. Gertsvolf, A. Yu. Naumov, D. M. Villeneuve, A. Boltasseva, V. M. Shalaev, and P. B. Corkum. High-harmonic generation in metallic titanium nitride. *Nature Communications*, 12(1):4981, December 2021.
- [179] Murat Sivis, Marco Taucer, Giulio Vampa, Kyle Johnston, André Staudte, Andrei Yu. Naumov, D. M. Villeneuve, Claus Ropers, and P. B. Corkum. Tailored semiconductors for high-harmonic optoelectronics. *Science*, 357(6348):303–306, July 2017.
- [180] Hanzhe Liu, Cheng Guo, Giulio Vampa, Jingyuan Linda Zhang, Tomas Sarmiento, Meng Xiao, Philip H. Bucksbaum, Jelena Vučković, Shanhui Fan, and David A. Reis. Enhanced high-harmonic generation from an all-dielectric metasurface. *Nature Physics*, 14(10):1006–1010, October 2018.
- [181] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library.

In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

- [182] Aleksey Korobenko, Sabaa Rashid, Christian Heide, Andrei Yu Naumov, David A. Reis, Pierre Berini, Paul B. Corkum, and Giulio Vampa. *In-Situ* Nanoscale Focusing of Extreme Ultraviolet Solid-State High Harmonics. *Physical Review X*, 12(4):041036, December 2022.