Anodic alumina as a scalable platform for structural coloration and optical rectification

By

Declan Oller

B.A., Clark University, 2011

M.S., Brown University, 2013

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Physics at Brown University

Providence, Rhode Island

May 2018

This dissertation by Declan Oller is accepted in its present form by the Department of Physics as satisfying the dissertation requirement for the degree of Doctor of Philosophy

Date	Jimmy Xu, advisor

Recommended to the Graduate Council

Date	James Valles Jr, Reader
Date	Vesna Mitrovic, Reader

Approved by the Graduate Council

Date	Andrew Campbell, Dean of the Graduate School

VITA

Declan Oller was born in New York City, New York, in 1989. He received a B.A. in Physics and Mathematics from Clark University in 2011, an M.S. from Brown University in 2013 and a Ph.D. from Brown University in 2018. He has taught lab courses at Brown.

ACKNOWLEDGMENTS

Many people have inspired, supported, and helped me during the writing of this dissertation. I am deeply indebted to Jimmy Xu for his patience, insight, and inspiration for research direction. James Valles and Vesna Mitrovic have provided me with indispensable suggestions and throughout the whole process have been sources of support and drive. Gustavo Fernandes spent countless hours with me solving problems both high and low. At Clark University, I thank Charles Agosta for introducing me to my first self-driven research.

While at Brown, I shared great times with a close group of friends. During the most difficult points, the experiences with them kept me going and happy. They are too many to name here, but they know who they are. Lastly, I would like to thank my mom, Deborah Maier. She inspired curiosity in me from a young age, and did everything she could to give me every opportunity. I am eternally grateful to her for this.

Contents

List of Figures

Introduction

Two projects are covered in this dissertation. Though they are in somewhat disparate areas of physics, they actually contain some overlap: both use Anodic Aluminum Oxide (AAO) as a platform, both are concerned with optical effects, and both have an emphasis on scalability in areas that have mostly used extremely non-scalable means to produce the desired effects.

In section 1, we present a broad overview of AAO, because it forms the basis for both projects. AAO is a very interesting material. It is the oxide of aluminum (Al), but what makes it interesting is that it easily forms an ordered pore structure; that is, when the oxide grows, long channels are formed from the surface of it down to the boundary it forms against the Al it grew from. AAO production itself is a whole field of research. In this section, we both describe basics of it, fabrication procedures used in this project, and also present optical measurements of industrial Al alloys. The motivation for this is that Al is already incredibly widely used in the world, and also widely used for research purposes – but the overlap and transfer from scientific research results using it have yet to be exploited. This is partly due to the disconnect between the nature of the Al used in research (very pure, expensive, and poor material properties) and the nature of Al alloys (dirtier, rougher, less controlled properties, etc). If Al alloys could be used as easily as pure Al, many brilliant yet currently less practical ideas could be implemented in the real world.

In section 2, we present a scalable, structural coloration platform leveraged to its full extent. Structural coloration is an exciting research field that, despite its age-old history and use, is seeing a resurgence, due partly to nanofabrication techniques and computational methods that were not available until recently. Structural coloration can be described as a method for producing colors that doesn't rely on short-range effects like dyes, but rather more directly physical

and long-range effects such as interference and scattering. Well known examples of structural coloration are the blue sky and wings of a butterfly. Because structural coloration is due to general physical principles rather than specific materials, it has potential that chemicals don't, such as more tunability, and even active tunability. Though, it should also be emphasized that structural coloration doesn't need to *replace* chemical coloration to be valuable – it can work in concert with it and simply needs to add. The structural coloration platform we investigate in this project is that of a Fabry-Perot (FP) cavity formed by an Al substrate, AAO grown on that Al, and an absorber layer. While many of the recent advances in coloration are very impressive, many of them rely on very expensive and non-scalable fabrication techniques like electron-beam lithography. Our structure is very scalable, and as can be seen, has been scaled to the macro (1cm x 1cm) scale already here. We use computational methods to investigate the properties, possibilities, and mechanisms behind this structure, such that it could become a fully-fledged, color-by-design platform, and fabricate and measure samples to verify and guide this. Notably, we investigate the broader picture of this structure: what are its *full* capabilities? What are the *best* colors it can theoretically and practically give rise to? What are the bottlenecks to this being used to great effect in the real world?

In section 3, we present the intersection of two fields that have had little interaction: resistive switching (RS) and Optical Rectification (OR). RS is, generally, a system that can have its resistance switched by various methods (often just application of the right bias). RS has a rich research history, tied to its prediction and demonstration as a platform for the "missing fourth circuit element", the memristor. However, the vast majority of research done with it has been in the realm of electronics, which we hope to change in this project. The other field, OR, is generally concerned with using a collector of light (often an

antenna, or just small plasmonic features, depending on the light frequency) as well as a rectifier (often a tunneling junction) to harvest or sense light. It holds advantages over current methods, most notably not being limited by certain speed/frequency bottlenecks that current methods are. Light harvesting is certainly a very lucrative reward, but the task at hand is proportionally massive, and despite decades of research, the satisfying results are slim. Importantly, there certainly *are* results for some frequency ranges like radio and microwaves, but due to physical constraints, higher frequencies are very difficult to rectify, and even more difficult to actually prove that's what has occurred. In this project, we have employed a common RS system (that produces a nanoscopic, tunable filament of metal) as a platform for OR. We first look at its electronic capabilities, which are interesting in their own right, and then present what may be the first OR signal produced without nanofabrication techniques.

1 Anodized aluminum overview

1.1 Basics, history, and issues

Many elements form an oxide compound through a reduction-oxidation reaction in which an atom of the element loses some number of electrons and the oxygen gains some. These ions then compose a new material, called the oxide of that element. For example, the chemical reaction for a typical metal such as aluminum (Al) looks like:

$$4\mathrm{Al} + 3\mathrm{O}_2 \Rightarrow 2\mathrm{Al}_2\mathrm{O}_3$$

The oxide produced is usually much more chemically stable than the element it is based on, and has radically different properties, owing to the fact that an oxide is typically a dielectric. Aluminum oxide (alumina) is a highly insulating

ceramic material. There are several common phases of alumina, most notably the alpha and gamma phases [11]. The alpha phase, also known as corundum, has a rhombohedral structure and is the most stable. When it is in a single crystal form, it is known as sapphire. Alumina can also be found in an amorphous state (α-Al_2O_3).

This oxidation reaction takes place in many ways with many different mechanisms. Many elements (most metals and semiconductors) form what is called a "native oxide" in which, when the element is exposed to air, the outside surface of the material is oxidized for some thickness [22]. It should be noted that oxidation is almost always a self limiting process, because the reaction requires both the unoxidized element and unreduced oxygen; once an oxide layer has already been formed to some thickness, further oxidation requires oxygen to diffuse through the formed oxide layer to reach the unoxidized element [10]. Therefore, the thicker the oxide layer, the slower the rate of oxidation. Certain factors can increase the oxidation rate, such as temperature and oxygen concentration [65].

However, there are other mechanisms of oxidation. One common one is known as anodization, in which the material to be oxidized is placed in an electrolyte bath, with a voltage bias established between the material to be oxidized and a counter electrode (usually an inert graphite electrode) [28], as in Fig. 1. A current is passed from one electrode to the other, and in the process, an oxidation reaction takes place at the interface between the material and its growing oxide. Due to the driving force of the applied bias, the oxide grows at a much faster rate and can grow much thicker than a typical oxide. Aluminum oxide grown this way is known as Anodic Aluminum Oxide (AAO). Although anodization can occur with many different materials [89, 7, 49], we choose here to focus on anodization done with Al.

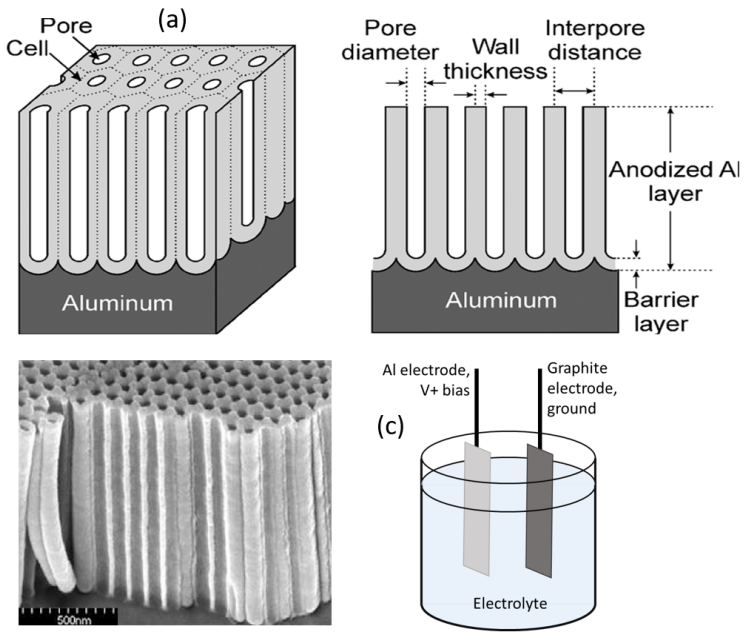

Figure 1: Overview of Al anodization. (a) Schematic of AAO and relevant dimensions. (b) SEM image of real, highly ordered AAO. (c) Schematic of anodization cell.

AAO can have one other very interesting property, as well: given the right anodization conditions, the oxide that is grown will form an ordered "pore" structure, with channels extending from the air/oxide interface down a certain depth. The point at which they stop is known as the "barrier layer", where there is a continuous oxide layer without any pores. The pores (looking from above the structure) naturally form a hexagonal lattice (see Fig. 1). Various parameters of this process (electrolyte temperature, compound, concentration, anodization voltage, aluminum purity, and others) can affect various properties of the AAO (growth rate, barrier layer thickness, pore diameter, pore period, pore ordering, etc) [8, 54], which will be covered later.

Alumina has many interesting properties and applications. While pure elemental Al is actually very soft, with a yield strength of 7–11 MPa [69], aluminum oxide (alumina) is very hard and durable [36, 16], making alumina a good protection against environmental and mechanical effects. Because it grows anywhere Al is revealed to the oxidizing environment, alumina will almost always be conformal and relatively uniform. Additionally, alumina is transparent, being optically non-lossy and having a refractive index of $\sim$1.6, similar to glass [13, 62], making it an ideal protective layer because it allows the Al it is protecting to remain reflective.

However, all of these properties are characteristic of alumina, which is aluminum oxide of any form; they are not unique to AAO. AAO itself, however, has several interesting properties that are unique to its naturally created pore structure. First, because the pores extend from the outer surface of the AAO to deep inside, there is a path for the electrolyte that is causing the anodization to get to the Al surface more easily, independent of the AAO depth. This breaks the usual self-limiting rate of the AAO and causes it to have a much faster growth rate (typically around 100nm/min) [101]. The pores also offer many

other opportunities. Most commonly, consumer Al products are protected and colored by anodizing them to create the porous AAO layer, then infiltrating the pores with a traditional dye, and the boiling the structure, causing the pores to seal up and trap the dye [28]. This is superior to simply painting the dyes on the surface, for which a scratch would easily scrape the dye off. AAO has also been used in a similar way as a matrix to suspend metal nanoparticles (similar to traditional stained glass) [100, 95]. Because it is porous, with the alumina parts (i.e., the non-pore parts) having a refractive index of 1.6, it can be easily wet etched, decreasing the alumina/air ratio of the AAO. Because the inter-pore distances and pore diameters are usually below the wavelengths of light, the AAO acts as an effective medium, meaning pore etching allows creation of AAO with a tunable effective medium between 1 and 1.6 (aside from it structurally degrading when it gets etched enough). This has been used for structural coloration purposes [102], as will be discussed later. Similarly, anodization parameters have been varied with time, creating a stacked structure of alternating layers of AAO with different properties. These have been used to create a Bragg Reflector [94]. Similarly, it has been shown that there exists an optical Fano resonance caused by the pore structure [76].

The pore structure has also been used for other purposes. For example, it can be used as various types of masks; in one case, an RIE etch mask, allowing large scale etching of features usually only possible with e-beam lithography (EBL) [55]. In another case, an evaporation mask for depositing a large two dimensional lattice of nanodots on a substrate [72]. Similarly, it has also been used as a template for creating nanowires, in which a material is deposited into the pores from an electrolyte solution, in a process similar to anodization. The deposited material fills the pores, creating nanowires in the negative space of the pores [71]. The AAO template can then be dissolved, leaving extremely

large numbers of nanowires behind. Lastly, AAO can be used to create a thin, standalone dielectric membrane, by dissolving the Al and leaving just a thin layer of AAO [85].

1.2 Physical mechanism of anodization

The physical mechanism behind AAO must be explained. Most basically, the reaction taking place is the one stated above, but that still leaves many questions open, most notably, once an oxide is formed, how does further oxide form? And, why do pores form at all? And, why, under the right circumstances, are these pores very well ordered? The answer to the first is that oxygen containing ions, O_2^- and OH^-, electromigrate through the barrier oxide at the bottom of the pore [45]. Simultaneously, Al^{3+} ions are being created at the metal/oxide interface. When the migrating oxide ions meet the Al^{3+} ions, alumina is formed. It should be noted that no oxide formation takes place an the oxide/electrolyte interface; any Al^{3+} ions that reach that interface are ejected into the solution.

The reason for pore growth itself is typically considered to be a result of several simultaneous mechanisms: the electric field distribution at the metal/oxide interface, heating from the formation reaction, and the mechanical stresses due to the increase in volume that occurs when Al is converted into alumina. It is integral to pore formation that the electrolyte can dissolve the oxide [47]; if an electrolyte which does not dissolve the oxide is used, an oxide coating will be formed (a "barrier type" oxide), but will not be porous and will have a constant thickness with time (dependent only on the anodization voltage). This is because, for a given voltage, it needs a certain magnitude of electric field across the oxide to drive oxide ions through it. When the oxide is thin enough that the E field is large enough, it forms until it isn't (because it decreases with the oxide thickness as 1/d).

The mechanism for pore formation can be explained as follows. Consider there already being a thin barrier oxide over the whole metal, which is submerged in the electrolyte which is at the anodization voltage with respect to the metal. Now, if the oxide happens to be a little thinner at some point, there will be a higher E field there, causing more current to flow. This increased current should cause more oxide to form, but also locally heats the electrolyte temperature, which increases the rate of it dissolving the formed oxide in this area. Thus, there is a positive feedback loop in which an area of thin oxide leads to higher current, which leads to higher temperature, which leads to more dissolution at that point, keeping the thin area thinner. This is essentially what the pore is: the "thin area" mechanism repeated for some depth. Meanwhile, provided that the reaction is not too violent, at the other areas, oxide will continue to be formed, but not dissolved.

It's very important to note that the barrier layer at the bottom of the pores is the same as what occurs for electrolytes that can't dissolve the oxide; however, in this case, the diminishing of the E field as the barrier layer grows is matched by the dissolving action of the electrolyte. Thus, there is a barrier layer of constant thickness at the bottom of the pore, that is being continuously dissolved by the electrolyte and formed anew at the metal/oxide interface.

The origin of the ordering/packing is a result of a "packing problem". It is assumed that the mechanism described above (a pore forming from an instability) is not a rare event, and happens all over the surface. Essentially, formation of a pore happens "where it can". That is, if there were somehow a large distance between two adjacent pores such that another pore could form in between them, the addition of this new pore would decrease the total resistance from the electrolyte to underlying Al and the instability effect discussed already would occur. However, this only explains why pores are dense, not why they

form a hexagonal grid, which is the *densest* packing. This is due to a minimization effect: the E field going from the end of the electrolyte column (at the very bottom of the pore) to the metal is roughly spherical. If two pores are nearby, their respective E fields combine, causing the reaction to occur at the metal/oxide interface even more quickly where they are closer, causing nearby pores to essentially migrate towards each other. Therefore, the mechanism that caused two pores to move more closely to each other will cause a third to meet at this intersection, which is the basis for a hexagonal packing. It should be noted that this doesn't immediately happen; it is an energy-minimization process that occurs over some period, which is the point of the first and second anodization steps.

1.3 Fabrication, parameters, and options

Almost every fabrication parameter of AAO anodization can be changed, creating AAO with different properties. Several of them are briefly covered here. Because each parameter can affect more than one property at once, if a specific set of properties is required, it may be necessary to tune several parameters in harmony such that the effect of one is balanced by the effect of another.

1.3.1 Electrolyte

The choice of the electrolyte bath is very instrumental to the properties of the AAO formed. The most common choices are oxalic acid [45, 47], sulfuric acid [45, 47], phosphoric acid [47], and chromic acid [47]. Other parameters being equal, they give different pore diameters and periods, though they often need different anodization voltages to work correctly.

It should also be noted that experiments have been done using mixtures of different electrolytes [107], resulting in AAO with properties corresponding

to values in between the values of AAO fabricated using either of the pure electrolytes, as intuitively expected.

1.3.2 Anodizing voltage

As discussed previously, the anodization voltage affects several properties simultaneously. Because it determines the equilibrium between the strength of the oxide formation and the dissolution rate of the oxide, it determines the thickness of any continuous section of oxide between two conductors in this system (i.e., either the electrolyte/oxide/metal sandwich at the bottom of the pore, or the electrolyte/oxide/electrolyte sandwich formed between pores, though it is slightly different for the latter). That is, while it also affects barrier thickness, it also affects pore spacing.

Because the constant barrier layer thickness is an equilibrium between being formed by the anodization voltage and being dissolved by the electrolyte, the barrier thickness is a function of both voltage and the electrolyte. However, this means that the barrier thickness per applied volt will be different for different electrolytes [40].

Recall that the mechanism by which a barrier layer is formed for "barrier type" electrolytes is such that oxide is formed by oxygen containing ions electromigrating through the already formed oxide until the oxide is thick enough that the electric field acting on those ions isn't high enough to drive them, and current stops. For porous type electrolytes, at equilibrium this is perfectly matched by the dissolving action of the electrolyte. However, this leads to an interesting effect. If, for a porous type anodization, the voltage is suddenly significantly dropped, the E field at the barrier layer is no longer high enough to drive electromigration through it. This wouldn't be an issue if this (lower) voltage were used since the beginning of formation, because then it would form a thinner barrier layer initially, but because it is being met with a thicker one,

no further anodization can occur. This fact was used by Hass [30] to measure the barrier layer thickness. For example, it was shown for 2% wt oxalic acid at 24°C that its barrier layer thickness is 1.2nm/V.

However, this is only true for very large changes. If the voltage is changed slowly, the barrier layer mechanism can respond as well. While by far the most common anodization process is to just use a single, constant voltage, much research has been done on changing the voltage over time, often alternating the voltage between two values [94]. This has the effect of changing the barrier layer thickness (and therefore the pore wall thickness as well) and can be used for interesting applications. In one instance, it was used to simply create a stack of alternating layers with different optical properties, creating a Bragg reflector with sharp resonances [94]. In another instance, researchers used the fact that it can cause pores to split at the interface between the two layers, giving rise to a 'Y' shape pore (which can then be used as a template for nanowires) [90].

1.3.3 Type of Al substrate

Several types of Al are used in this project for comparison and variability.

Pure Al strips The most commonly used form of Al in literature involving AAO fabrication is that of pure Al sheets, typically of 99.999% purity. These have several advantages: 1) Purity, 2) Ease of fabrication (typically ordered from a well known source, doesn't have to be evaporated) 3) thickness, so it can have EP or a first anodization done to it, and 4) uniformity (many factors go into film deposition, increasing difficulty of reproducibility). They tend to produce very high quality AAO.

However, there are several disadvantages. Pure Al is very soft, and the samples bend easily during processing. This makes optical measurements more difficult. Additionally, they are often more expensive than evaporating Al, mak-

ing mass production of them for research difficult.

Pure evaporated Al Another commonly used method is to create a thin Al film by e-beam evaporation. This produces a very high quality, smooth, mirror like finish that does not need to (and cannot) be EP'd. However, depending on the thickness, it can be difficult on the evaporation system (if evaporating $> 1^{-}$m), and is time consuming. Typically, a thin (5 nm) adhesion layer (Ti or Cr) is evaporated immediately prior (without breaking vacuum) to the Al.

Industrial Al

1.3.4 Motivation for use of industrial Al

So far our has been concerned with a very pure Al substrate, because it gives rise to high quality AAO and is simple to work with. However, pure Al has several downsides that prevent it from ever being used in practical contexts; most notably, it is actually very soft and malleable. In addition, high purity Al tends to be very expensive. Therefore, there is strong motivation to investigate the use of Al alloys in addition to pure Al in this project. Below we present relevant optical data of Al alloys which we will use the results of later.

1.3.5 Different properties of Al alloys

Although pure Al may be soft and not widely commercially used, only a few percent of the Al needs to be replaced by certain impurities in order to drastically change the alloy's properties. A look at the compositions of several of the most common Al alloys is shown in Fig. 2, where it can be seen that highest percent of Al that is replaced for any of the alloys listed is 7075 which still has 90.0 percent Al. In this section, we investigate the 5 most commonly used Al alloys: 2024, 3003, 5052, 6061, and 7075.

Alloy	Composition (percent by weight)						
	Al	Cu	Mn	Mg	Cr	Si	Zn
2024	93.5	4.4	0.6	1.5			
3003	98.6	0.12	1.2				
5052	97.2			2.5	0.25		
6061	97.9	0.28		1.0		0.6	
7075	90.0	1.6		2.5	0.23		5.6

Figure 2: Composition of the several Al alloys used in this project. These quantities represent the bulk of the elemental composition of each alloy, but even smaller percents of other elements may be present.

The question we wish to answer in this section is, how much do these impurities change the optical properties? And, are there other properties typically incidental to the manufacturing process of industrial Al that change the behavior as well?

Al indices The first property to check is the optical properties of the Al alloys as compared to pure Al. There are a few important details to mention here. The first is that metals are usually hard to get the optical properties of, due to their reflective/absorbing nature. Second, the values of the refractive indices that can be taken from the literature for pure bulk Al vary across different sources, often non-insignificantly [70]. This can be due to a multitude of reasons; 1) the aforementioned difficulty of measuring metals' optical properties, 2) how crystalline the Al is, 3) Al has a native oxide, so how recently the sample was made will determine the thickness of the native oxide and could affect the measured Al indices.

The other important facet to note is surface roughness. For evaporated Al on a polished Si substrate, surface roughness is very low and only really a result of random orientation of crystallites and hillocks on the Al surface [78]. However,

the Al alloy samples are pieces of metal meant for industry work and are thus relatively very rough, because a high polish isn't often needed. Additionally, due to the rolling process in which sheet Al is created, there is an inherent "grain" to the metal, analogous to a grain in wood. This causes it to naturally be somewhat anisotropic.

Keeping this in mind, we do several processes to prepare the surfaces of the Al alloy samples involving mechanical polishing (MP) and electropolishing (EP). Although it might seem natural to polish all samples to the highest polish possible, we actually process some of them to a lesser polish, to investigate the effects of a small amount of surface roughness. We must make a distinction when explaining polishing; here, when we say "polished", we really mean "smoothness", which really comes down to a matter of scale. For example, a sample could be flat on a large scale, but very rough looking in an SEM image; or alternatively, bent and warped when viewed with the naked eye, but very flat microscopically (or rough in both scales). This is why, to achieve the highest surface polish, we must do both MP and EP: MP removes large scale roughness, and oppositely, EP is an electrochemical process that attacks microscopic protrusions more violently because they concentrate electric field (but doesn't change the large scale structure).

Therefore, for most Al alloy samples (unless otherwise noted), we do a process of sequential mechanical polishing using an orbital sander with increasing sanding grits up to 2000 grit. It was found from SEM imaging that there was no noticeable difference for grits above 2000. In addition, an EP step is performed immediately after, at which point the surface roughness from the MP is low enough that the EP removes it.

AAO from Al alloys The refractive indices produced from Al alloys have some differences from AAO produced from pure Al, but surprisingly little. The

indices measured from VASE are shown in Fig. 3, where n is represented by the solid curves, and k by the dashed, for each Al alloy. Each AAO was measured at 3 or 4 different thicknesses; it is seen that there is often significant difference between them. This is concerning, because even though they are different thicknesses, the refractive indices should be independent of this. However, it is possible that, because for these small thicknesses, the barrier layer is still a significant percent of the AAO thickness, it is affecting the AAO indices disproportionately for the different thicknesses.

1.3.6 Lorentz-Drude Model fitted to Al alloys

The Lorentz-Drude model is a model for the optical constants of a material, most often the dielectric constants. It attempts to account for electrons that are unbound to any specific atom (the Drude part of it) as well as electrons that are more like bound oscillators to an atom (the Lorentz part). There are derivations for both parts that we briefly cover here. However, it should be made very clear that although the derivations for the two components give useful forms that are relatively physically reasonable, they are *not* necessarily what is physically occurring here; that is, they are a useful starting point, and in addition, the LD model is really a mathematical construct for fitting the optical constants that matches certain physical constraints. That said, it *can* and often does give physical insight (as will be explained) to what is occurring to give the results found, but it cannot be assumed that it is the full story. For example, the derivation for each part is entirely based on electrons, but phonon modes also give rise to optical effects, which affect the optical constants. Therefore, the LD model could be applied to fit to measured optical constants with phonon effects even though the derivation for the fitted term was for electrons.

The Drude model was first introduced in 1900 to explain the conduction mechanism for metals, which it matched well. Metals tend to have many free

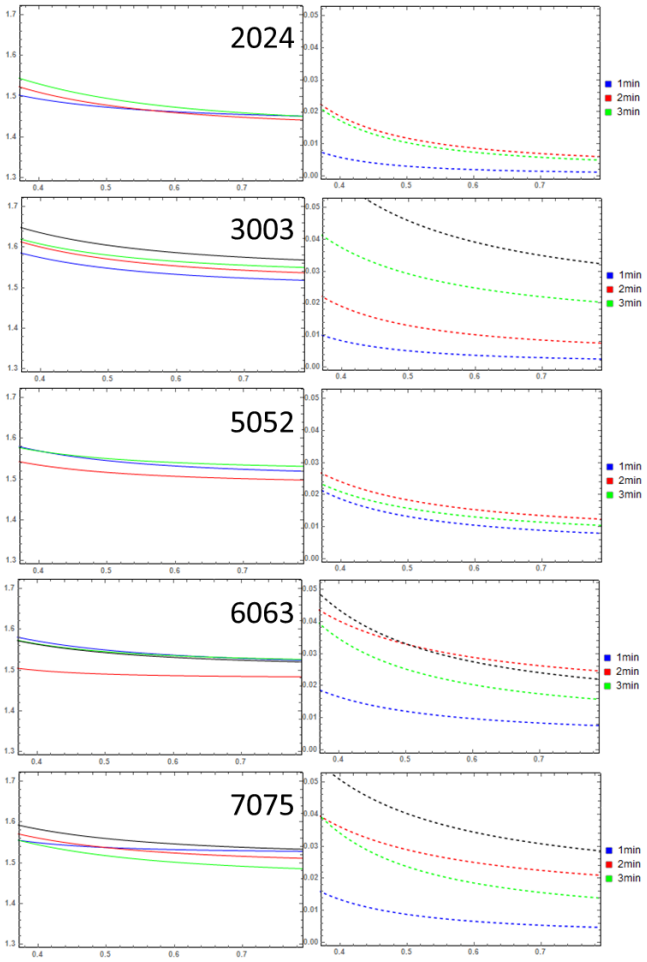

Figure 3: Refractive indices of AAO produced from different Al alloys. For each, n is shown on the left and k on the right. In each plot, the relevant index is measured for three different AAO thicknesses, to check for any variability.

carriers (electrons and holes, but we will just cover electrons for simplicity; the math should be nearly identical for holes), which is what the Drude model attempts to explain.

We look at the behavior of an electron of charge $-e$ and mass m_e under a constant (for now) electric field E. We are interested in the average behavior of all the free electrons. Naturally, a charge acted on by an electric field accelerates at a rate $a = -eE/m$. However, randomly and sporadically, the electron collides with atoms of the crystal lattice and has its momentum "reset". If an electron travels for a time t before it collides with an atom, in that time it will have picked up a total velocity at. Because the initial velocity it has after a collision is random and equally distributed in all directions on average, that initial velocity cancels out and doesn't contribute to an average velocity in an electric field. However, for the velocity picked up as a result of the E field, we can define a "relaxation time" τ that is the average collision time, which gives an average velocity for the electrons: $v_{avg} = -eE\tau/m$. Now, since the current at a point is just the total flow of charge, we have $j = -nev_{avg}$, where n is the electron density. Conventionally the conductivity is written $\sigma = ne^2\tau/m$, which gives $j = \sigma E$.

Now we use the momentum of an electron rather than its velocity and say that $j = nep/m$. Consider the average behavior of an electron with momentum at p at time t. To find the evolution of $p(t)$, we investigate the value of p a short time dt later, $p(t + dt)$. If τ is the time (on average) by which is must have a 100% chance of having a collision, then in time dt it has a dt/τ chance of having a collision. If it does, then the momentum is reset. However, if it doesn't, then in that time it picks up momentum $f(t)/m$ which is added to the momentum it already had at time t, where $f(t)$ is the force on it at time t. Since it either collides or doesn't, then the chance of it not colliding is $(1 - dt/\tau)$. Therefore,

we have

$$p(t+dt) = (1 - dt/\tau)(p(t) + f(t))$$

Which can easily be arranged to give

$$(p(t+dt) - p(t))/dt = \frac{dp}{dt} = -p(t)/\tau + f(t)$$

So we see that the effect of the collision is essentially to be a damping term, which intuitively makes sense – if there were no collisions, $\tau = \infty$, and dp/dt would be finite and p would increase without infinity. The detail of $1/\tau$ being the damping constant will be important later, which we will call $\gamma = 1/\tau$.

Now we consider an oscillating applied E field: $E(t) = E(\omega)e^{-i\omega t}$, plug this into the equation for the time evolution of the momentum we just derived, and assume a similar form for the momentum to try and solve it: $p(t) = p(\omega)e^{-i\omega t}$. Plugging this in and solving for $p(\omega)$, we get $p(\omega) = -eE/(\gamma - i\omega)$, which we can then plug into our momentum equation for current, $j = nep/m$, and relate that to the conductivity equation of current, $j = \sigma E$. Doing this gives

$$\sigma = \frac{\sigma_0}{1 - i\omega\tau}, \sigma_0 = \frac{ne^2\tau}{m}$$

Lastly, we can plug the current into Maxwell's equations to relate this form of the conductivity to the dielectric constant:

$$\nabla \times E = -\frac{1}{c}\frac{\partial H}{\partial t}, \nabla \times H = -\frac{4\pi}{c}j + \frac{1}{c}\frac{\partial E}{\partial t}$$

$$\Rightarrow \nabla \times (\nabla \times E) = \frac{i\omega}{c}(\nabla \times H) = \frac{i\omega}{c}(\frac{4\pi\sigma}{c}E - \frac{i\omega}{c}E)$$

$$\Rightarrow -\nabla^2 E = \frac{\omega^2}{c^2}(1 + \frac{4i\pi\sigma}{\omega})E$$

Since an electromagnetic wave usually has the form $-\nabla^2 E = \frac{\omega^2}{c^2}\epsilon(\omega)E$, this gives us the effective dielectric constant

$$\epsilon = 1 + \frac{4\pi i \sigma}{\omega}$$

There are a few things to note about this equation. First of all, we notice that it's monotonic: with this formulation, ϵ has a trend for increasing frequency, but there are no "special" or "anomalous" parts of the dielectric constant, aside from the plasma frequency, at which ϵ changes sign before and after it (and for our applications, we will be investigating effects far below the plasma frequency for all materials used). Second, it gives physical insight: if $P = \epsilon E$, then the real component of ϵ make P and E in phase and let a polarization dissipate energy, while the imaginary part puts them out of phase and means no energy is dissipated. This is backwards from what is expected from the behavior of the complex refractive index. The real and imaginary parts of ϵ are shown in Fig. 4 for arbitrarily chosen parameters to illustrate its generic behavior. Because it is standard to give the parameters in units made to match units of eV and the photon energy is proportional to the frequency we have used in the derivation, it is plotted with respect to photon energy. In addition, in this specific case we use the convention of making the imaginary parts of the dielectric constant negative.

The Drude model explains a large part of the optical properties of many materials, but still can't tell the whole story because it does not feature resonances, which are an important facet of optical response of real materials. For this, we must look at the Lorentz model.

The Lorentz model is meant to explain the behavior of bound electrons under an electric field. We take a simple model for a bound electron that has a restoring force linear with its displacement from its equilibrium position x (a reasonable

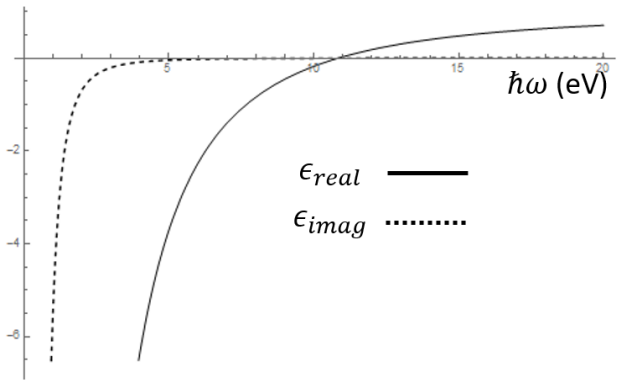

Figure 4: Example form of the real (solid curve) and imaginary (dashed curve) parts of ϵ due to the electrons under the Drude model. Note that the behavior for both parts is monotonic.

assumption if the potential it's subject to can be approximated as parabolic for small displacement), so $F_{restore} = -\omega_0^2 x$. In addition we assume there is a damping force which simply opposes the current velocity, so $F_{damp} = -\gamma \dot{x}$. Lastly, we consider a general driving E field force on the electron, $F_{drive} = eE$. Plugging these into Newton's 2nd equation, we get:

$$M\frac{d^2x}{dt^2} = F_{restore} + F_{damp} + F_{drive}$$

If the driving field is an oscillating electric field $E(t) = E_0 e^{-i\omega t}$ and we search for a similar solution for the displacement, $x(t) = x_0 e^{-i\omega t}$, we can get the frequency dependent amplitude term of the displacement $x_0(\omega)$. Similar to the Drude model, this can be combined with the dipole moment for this electron, $p = -ex$, and the definition of dielectric constant and E field to get:

$$x_0 = \frac{-e/m}{\omega_0^2 - \omega^2 - i\gamma\omega} E_0$$

$$p = -ex_0 = \epsilon_0 \chi_e E_0, \chi_e = \epsilon/\epsilon_0 - 1$$

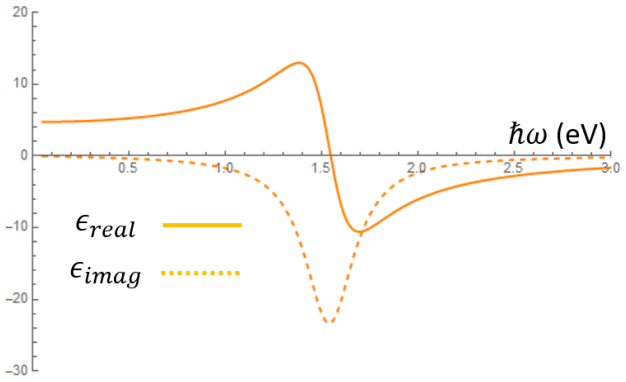

Figure 5: Example form of the real (solid) and imaginary (dashed) parts of ϵ for a Lorentz oscillator. Note that ϵ_{Im} has a peak where ϵ_{Re} goes to 0.

$$\frac{\epsilon(\omega)}{\epsilon_0} = 1 + \frac{Ne^2}{\epsilon_0 m} \frac{1}{\omega_0^2 - \omega^2 - i\gamma\omega}$$

Now we wish to generalize this to the whole material. We consider that different electrons (at different locations or energy levels) may have different resonant modes (that is, different ω_0 and γ_0 terms), so we actually cover a set of ω_j, γ_j's. Therefore, for each oscillator, we also define an "oscillator strength" f_j that is essentially the percent of electrons for which this resonant mode applies. Therefore, we use a summation over each of these oscillator modes, each weighted by their respective oscillator strength, and multiplied by the total electron volume density, to get the dielectric that results from all of them:

$$\frac{\epsilon(\omega)}{\epsilon_0} = 1 + \frac{Ne^2}{\epsilon_0 m} \sum_j \frac{f_j}{\omega_0^2 - \omega^2 - i\gamma\omega}$$

Now that we separately have the Drude and Lorentz parts, we can combine them to get the total ϵ, that accounts for both free (Drude) and bound (Lorentz) electrons:

$$\epsilon = \epsilon_{bound} + \epsilon_{free}$$

$$= 1 + \frac{4\pi i\sigma}{\omega} + \frac{Ne^2}{\epsilon_0 m}\sum_j \frac{f_j}{\omega_0^2 - \omega^2 - i\gamma\omega}$$

Now we can put the model to practical use. The procedure here is as follows. The optical properties of the material are measured, most commonly the dielectric constant. Then initial parameters are chosen for the model, which is calculated. The parameters are varied numerically, which varies the output of the model, which is compared to the measured results. This iterative process is repeated to a chosen level of accuracy.

The most important parameters to figure out at the beginning are the plasma frequency (for the Drude part) and the number of oscillators and their energies (for the Lorentz part). One starting point can be physical measurements; fortunately, the plasma frequency is easily experimentally measured. The number of oscillators is trickier but can also rely on physical measurements from the literature; as mentioned before, while these LD model oscillators do not necessarily have to represent a real physical oscillating electron, they often actually do in the form of interband and other transitions that are well known for the material. These often give the starting points for oscillator parameters.

However, while there are various algorithms for fitting these parameters, it should be strongly noted that there is a large aspect of manual tuning that must be done. For example, while Rakic showed for a variety of materials that 5 oscillators gave very accurate results to data, often it is the case that including extra oscillators could give a slightly better numerical fit to the data. This is because one of the fitting parameters, aside from the oscillator frequency, is the oscillator strength. So, if there is a very small numerical wiggle, an oscillator could be added at that position with a tiny oscillator strength to match the data better, but would probably not be physically meaningful.

We begin by using the LD model parameters for pure Al fitted by Rakic as a starting point. It was found in [70] that four oscillators gave a very good fit

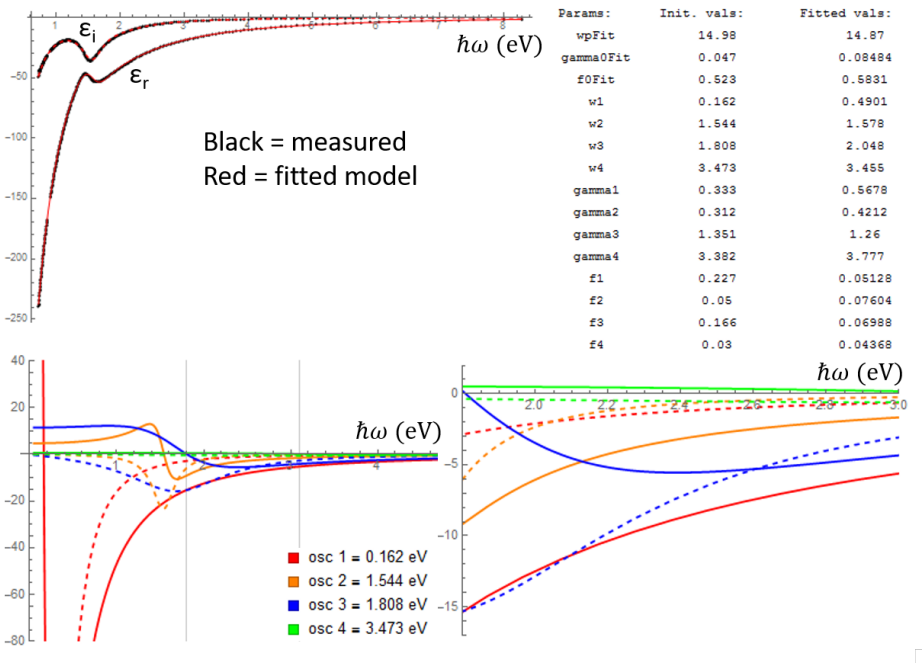

Figure 6: (a) Plot of fitted DL model to pure Al ϵ, real and imaginary parts (black curve = measured data from literature, red curve = DL model) (b) Four oscillators are used, with the fitted values shown in the table. All fitted values are in units of eV. (c) Real (solid curves) and imaginary (dashed curves) parts of each Lorentz oscillator shown separately for a large range. (d) The same as (c), but plotted for the relevant visible range. It can be seen that one oscillator has different curvature in this range because this range is on the other side of its resonant frequency with respect to the other oscillators.

throughout a large range (0.5 - 8eV), as shown in Fig. 6.

In addition, the 4 fitted oscillators are plotted separately, to illustrate the contribution from each. They are plotted over the entire fitted spectrum as well as just the visible spectrum, which is most relevant for our purposes. It can be seen that oscillators 1-3 are all relatively significant in the visible range, although none of their resonant frequencies are in this range. Oscillators 1 and 2 have the same curvature, while oscillator 3 changes the total curvature a little since its real part in the visible range is a different part of its oscillator's resonance.

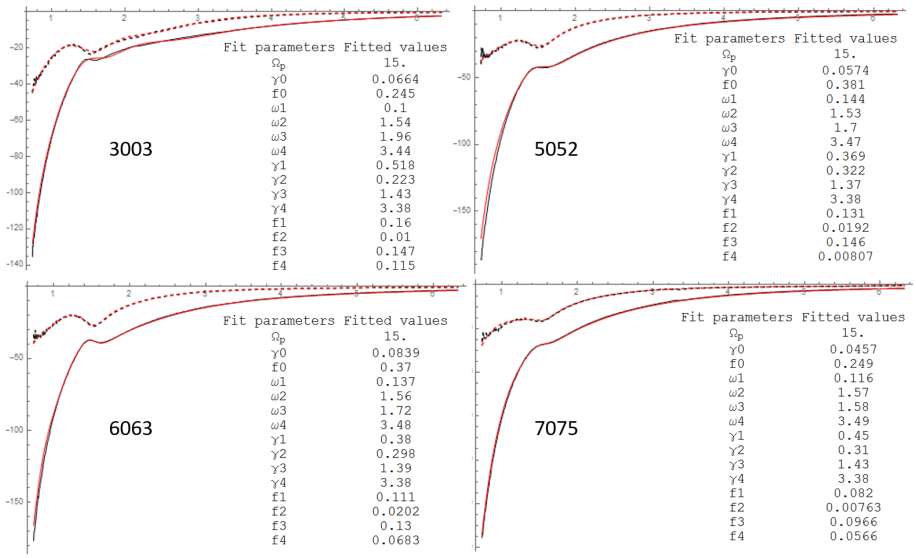

Figure 7: Measured and fitted dielectric constants, real (solid curve) and imaginary (dashed curve), for four Al alloys. For each, the measured VASE data is shown in black and the fitted model is shown in red for comparison.

Now that we are confident that we have a reasonable model for pure Al, we can extend this to industrial Al alloys. In Fig. 7, we present the measured optical constants for 5 alloys as well as their fitted LD models. Also included are their fitted parameters, all scaled to units of eV. Here, the black curves are the measured values (dielectric constants), and the red curves are the fitted LD model values. For both, the solid curves are the real part, and the dashed curves are the imaginary part, of the dielectric constant.

1.3.7 First vs second anodization; ordering

Anodization is a self ordering process, but it doesn't immediately form uniform pores with long range order. If anodization is done to an Al substrate with no surface preparation, alumina will be formed, but under most circumstances, it won't form an ordered array (see Fig. 9). It will still have porosity but it will be in the form of disordered pores, cracks, fissures, etc. however, as anodiza-

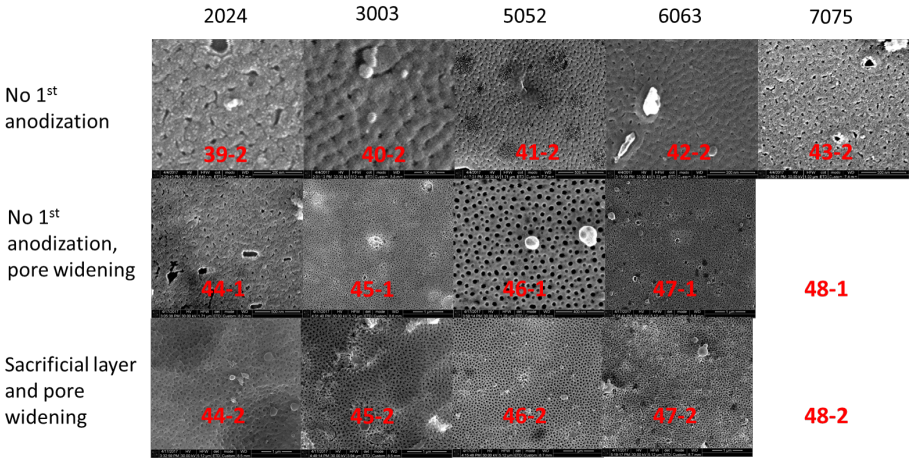

Figure 8: Overview of AAO from five Al alloys. Each column is a different Alloy, labeled above. Top row: one anodization step, therefore the most messy AAO, with no order expected. Middle row: the same as the top row, but with a pore widening step done, to illustrate the pore structure more clearly. Bottom row: two anodization steps, followed by a pore widening step.

tion continues, it will begin self-organizing at the Al-AAO interface into a more regular, periodic array. So, if anodization is done once, the pores closer to the air/AAO interface (the "oldest" ones) will be mostly disordered and messy, but become more ordered closer to the Al/AAO interface. Because, typically, completely ordered pores are desired, as well as the pores at the air/AAO interface, a process is done in which the first anodization layer is chemically etched off. This removes the AAO while minimally affecting the Al. This first AAO layer is called the "sacrificial layer" if it is removed. However, this Al surface has been changed by the first anodization and thus has a lattice of depressed/embossed features where ordered pores were being formed during the first step, as seen in the third step of Fig. 9. Now, if anodization is done on this Al, the pores will be ordered from the beginning, giving rise to a fully ordered pore structure.

A similar idea has also been used to create pores of different shapes. Because the properties of the AAO formed is so influenced by the initial underlying Al

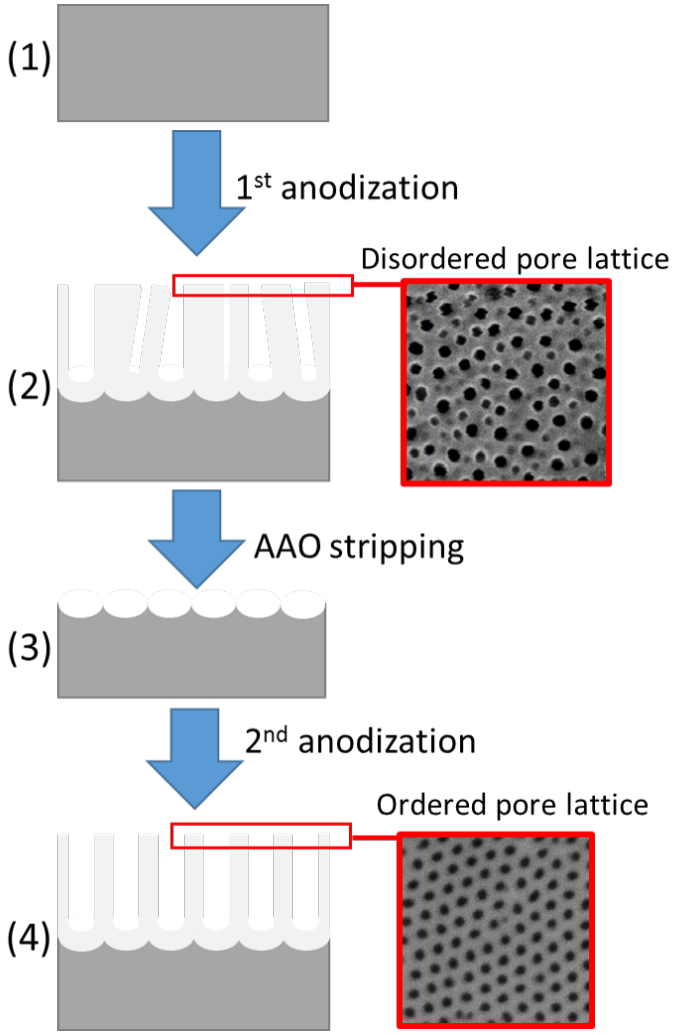

Figure 9: Illustration of AAO sacrificial layer and ordering process. (1) Flat piece of Al. (2) First anodization gives rise to disordered pores but an ordered lattice at the Al/AAO interface. (3) The AAO is stripped, leaving only an ordered, pointy lattice of Al. (4) Second anodization starts with this lattice and gives rise to ordered AAO. SEM images in (2) and (4) demonstrate disordered vs. ordered AAO.

surface, it has been shown that if the surface is nanoimprinted with a given pattern, pores can be formed in both different lattices, and different shapes [14].

1.4 Typical properties

1.4.1 Resistivity

The resistivity of alumina is highly dependent on its fabrication method, which can change its porosity, roughness, crystalline phase, crystallinity, and stoichiometry. However, it is typically found to be in the range of $10^{12} - 10^{14}$ ohm-cm [48].

AAO is naturally more complicated, with one large effect being ions and other molecules getting trapped in the fabricated alumina. Reports have found the resistivity of AAO to vary from $10^{9} - 10^{12}$ ohm-cm [29].

1.4.2 Optical properties

We must first discuss the properties of both Al and Al2O3/AAO for this project, because the optical properties we will be investigating involve them both.

Optical properties of Al Al is a highly reflective metal, due both to its refractive indices and its ability to be mechanically and electropolished to a high shine. Its refractive indices in the visible range are shown in Fig. 10. It has a nearly flat reflectivity of 90% across the visible range, leading to a typical shiny gray color, also shown in Fig. 10.

It has been demonstrated that it can be well modeled by the Drude-Lorentz model, with four optical oscillators, from 0.5-15eV. This is discussed later in the context of modeling industrial Al alloys. Al has a plasma frequency at 14.9 eV, where the reflectance drops to about 1% [70].

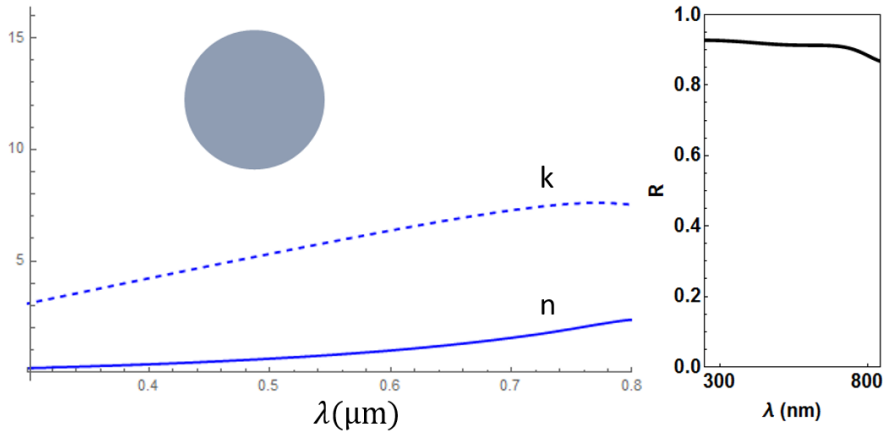

Figure 10: Basic optical properties of Al in the visible range. The indices are mostly flat with a large $k > n$, giving rise to a high, flat R across the visible. The inset color disk shows the calculated color of Al.

However, it should be pointed out (discussed later) that exact measurements of Al's refractive indices and other properties are widely varied, due to the inherent difficulty of measuring optical properties of bulk metal, the different preparation methods for Al, and the ever present oxide layer on Al.

Optical properties of AAO Aluminum oxide is ideally lossless in the visible range. It is an ideal dielectric in many ways, having a higher band gap and dielectric constant (9.9eV vs 9.0eV) and (10 vs 3.9) than SiO2 and can be prepared in many ways. As with Al itself, alumina can be produced in myriad ways, all having effects on the resulting optical properties [11]. Several common effects dependent on the fabrication process are the surface roughness, the void fraction (how "tightly packed" the alumina is vs how much of it is empty space), the crystal phase of the alumina, and the stoichiometry. The real part n of the refractive index in the visible range is usually relatively constant, decreasing slightly for increasing wavelength, and is usually centered from 1.5-1.7 depending

on fabrication parameters. The imaginary part k is usually very small, ranging from $10^{-4} - 10^{-2}$, also decreasing in the visible range. A thorough overview of different measured values for various preparations can be seen in [38].

The refractive indices of alumina can typically be well modeled by a Cauchy model with an Urbach absorption tail, in the visible-NIR range [11]. An Urbach absorption tail (or edge) is part of the absorption coefficient for long wavelengths that is typically due to messier, localized energy states in an amorphous or low crystalline material [93].

This covers the optical properties of plain alumina well enough, but AAO brings in several other facets. Most notable is the porosity factor, since it is inherent to AAO. A simple model for the resulting indices of AAO can be taken from a weighted average of the indices of solid alumina and air based on the filling fraction of the pores f:

$$n_{eff} = f n_{air} + (1 - f) n_{Al_2O_3},$$

$$f = \frac{8\pi r^2}{3\sqrt{2} p^2}$$

Where r is the pore radius and p is the pitch between the center of adjacent pores.

However, there are other details as well. For example, the porosity can lead to scattering, and as discussed already, the interface at the AAO/Al interface is necessarily textured, leading to more scattering. That said, the optical properties of AAO, unless specifically designed to be otherwise, are surprisingly similar to pure alumina. In [96], different fabrication parameters gave rise to n_{AAO} varying from 1.5-1.75 and $k_{AAO} \sim 0.003$ throughout the visible range, which is perfectly in line with the previously presented values for various fabrications of plain alumina. They found that both real and imaginary parts of the

refractive index are dependent on the anodization voltage and the anodization time. These dependencies are claimed to be caused by the variation of different molecular structures in the alumina making up the AAO; increasing the anodization voltage or time increases the content of the octohedral structure, which has higher refractive index [96].

1.4.3 Porosity

Etching Pore etching is typically performed with (70%) phosphoric acid. This has the benefit of etching the AAO but not the Al, and can easily be controlled by changing the concentration or temperature. An important consideration for AAO etching is that it behaves differently than typical wet etching of a uniform, non porous material. For a non-porous material, it etches uniformly from the top-down (the "top" being the material-etchant interface) and the etch rate is mostly uniform with time. However, with AAO, because of the pores, the etchant is attacking the AAO not only from the top, but also from inside the pores. This means that it's both less of a top-down etching, and also etches nonlinearly, because as more is etched, more surface area is exposed to etching. Furthermore, while a uniform film is uniformly etched from a full film to no film at all, because of the structure of the AAO and its pores, at some point the structure of the AAO gets degraded enough that it collapses and detaches from the surface, even if all the AAO is not actually dissolved. Thus, although the refractive index of the AAO can theoretically vary from 1.6 (full AAO) to 1.0 (vacuum), in practice it can't be reduced that much.

Ordering As explained earlier, ordering occurs naturally (if it is possible at all with the given substrate/other properties; see later discussion on industrial Al). However, different fabrication parameters will give rise to different levels of ordering [4]. Al purity is one important parameter [106], as well as temperature

and anodization currents [84].

Ordering may or may not be important for a given application, and if it is, the ordering range may be important. In this context, ordering "range" means the length scale at which the pores are ordered along the same lattice; typically, even for well ordered AAO, these lattices do not span the whole substrate and "domains" can be seen, conceptually similar to crystal domains.

The ordering can be measured by means of image analysis and a Fourier Transform (FT) [82]. Image analysis can easily identify pores and their positions from an SEM image of AAO, which can be easily analyzed with an FT to demonstrate the ordering. In Fig. 11, we have calculated the FFT of several SEM images of our AAO. The peaks in the FFT are a consequence of the regular spacing of the pores; by analyzing their respective distances and widths, we can quantitatively report their level of ordering, using the same technique as in [81].

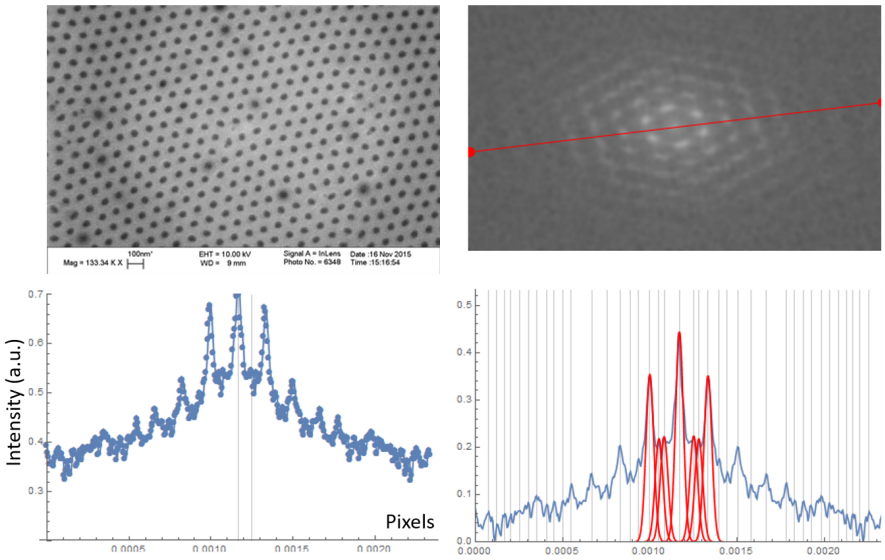

Figure 11: Fast Fourier Transform analysis of AAO. (a) SEM image of highly ordered AAO lattice. (b) FFT of SEM image in (a). A regular inverse lattice can be seen. (c) A plot of the image intensity as a function of position along the red line in (b). The position of the peaks allow us to figure out the average distance between nearest neighbor pores, second nearest neighbor pores, etc.

2 Structural coloration in AAO

AAO has many mechanisms and applications for producing visible color (although it should be noted that nearly all principles discussed here could be applied to parts of the electromagnetic spectrum other than the visible). First, by itself on the metal substrate from which it was grown, AAO has slight color. This is due to a simple thin film interference mechanism taking place in the AAO cavity between the Al and air. For a given AAO thickness, constructive and destructive interference can occur at different visible wavelengths, giving rise to reflection peaks and valleys. However, as will be discussed in the following, for simply AAO on Al, the peaks and valleys are not very deep and thus only faint colors are produced.

2.1 Origins of color perception

To understand how AAO can give rise to coloration, it is first necessary to explain how light is perceived by human vision and translated to color. Although the light spectrum producing principles discussed here are physical in nature, when we want to discuss coloration, it introduces an element of human biology that must be included and explained.

2.1.1 Human light sensitivity and XYZ tristimulus values

To begin, we recognize the concept of a light spectrum, denoted as $S(\lambda)$, where λ is the wavelength of the light, and $S(\lambda)$ is the intensity of the light at that wavelength (in units of J/m^2). We typically refer to the "visible spectrum" as the range of wavelengths from ∼380nm to ∼760nm, but it's important to note that there's nothing special or physically unique about this range; it is defined with respect to human vision because it is the only range of light that the human eye can sense (aside from unintended effects such as heating through microwave

absorption).

The way human vision senses light is constrained by biology and evolution and is not as straightforward as a manmade sensor might be. There are three types of light sensing cells [67] known as "cones", each of which responds to a wavelength range (see Fig. 12), sometimes called S, M, and L (for short, medium, long wavelengths). For convenience in the following, these cone sensitivities are often transformed to slightly different functions known as the CIE "color matching functions". To find out the human response to incident light, each color matching function is multiplied by the incident light spectrum and integrated over the visible range. These integrated quantities are called the tristimulus values:

These XYZ Tristimulus values allow quantification of human perception of a color due to incident light on our eyes. However, two matters complicate the matter: 1) the response of each cone is shaped roughly like a Gaussian function, so it responds to light within some range, but not evenly throughout that range, and 2) the sensitivity ranges for different types of cones overlap significantly. These two facts give rise to the complex nature of producing color from physical spectra.

2.1.2 Common and useful color spaces

These XYZ values form what can be called a "color space", that is, all possible values of them give the parameter space of colors that can be perceived by humans. The XYZ values are typically transformed to xyz values, which are essentially normalized XYZ values:

$$X + Y + Z = T$$

$$x = X/T, y = Y/T, z = Z/T$$

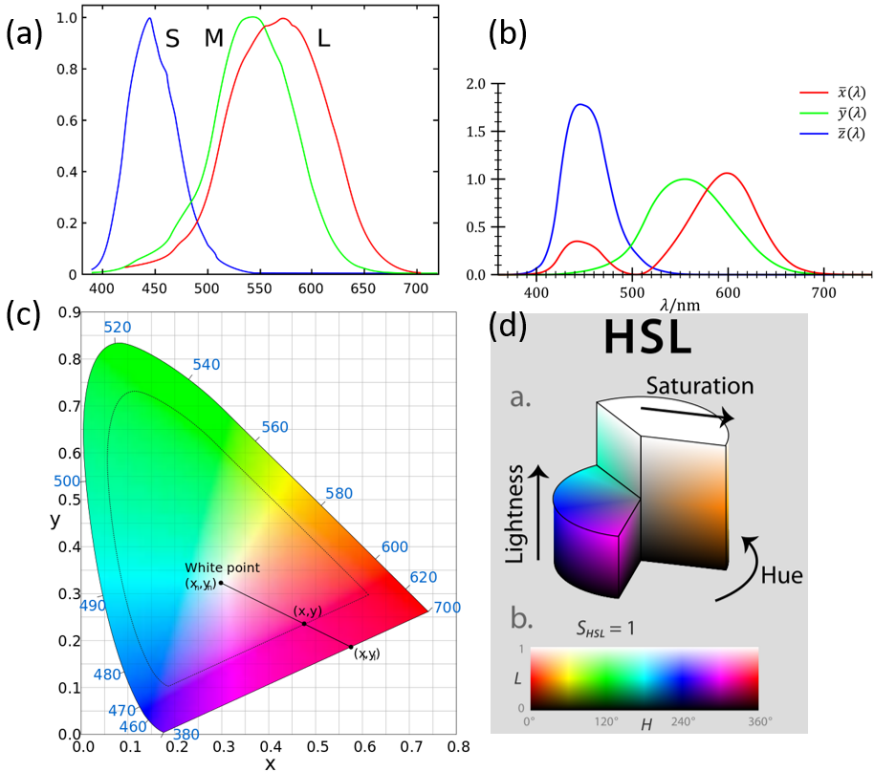

Figure 12: Human perception and color space information. (a) Sensitivity of human cones as a function of visible wavelength. (b) Color matching functions. Their positions can be seen to vaguely resemble the cone sensitivities in (a). (c) xy CIE color chart. The colors in this chart represent all the colors perceptible by humans at a given brightness level. The rounded border section of the chart represents the colors perceived for monochromatic light of the labeled wavelengths. (d) The HSL color space, which we will refer to periodically because it's more intuitive.

And x and y are taken to form the xy color space (Fig. 12). This is useful because it very cleanly illustrates all the hues that humans can perceive. There are two important things to note here though: 1) not all (x,y) values are in the color space; this is because the possible colors are essentially constrained by the color matching functions, and 2) if 3 values (X, Y, Z) were needed to describe all possible colors we can perceive, and the xy color space only has 2 values, what happened to the last value? The answer is that the xy color space doesn't include a concept of "brightness", really only concepts of "hue" and "saturation".

To continue, it's necessary to explain these concepts more quantitatively. The point in the middle of the horseshoe xy color space gamut is known as the "white point", which has a total absence of color. "Hue" is what is often meant by "color" colloquially; here we can take it to mean "the angle around the white point", measured from an arbitrarily chosen angle (in this project, we measure from a horizontal line extending to the right from the white point). The xy color space is intimately tied to human perception, since the color matching functions are essentially derived from our cone sensitivities. Thus, the curved part of the xy color space boundary corresponds to our perception of pure monochromatic light of that wavelength (see Fig. 12). Therefore, hue is closely related to (but not exactly the same as) the peak wavelength of the incident spectrum.

There is a complication to the concept of hue in the xy color space, however; for most colors in the space, they can be described by the monochromatic wavelength that the color lies on the line of between that wavelength and the white point. However, inspecting Fig. 12 shows that there are no equivalent monochromatic wavelengths for the straight segment connecting the 380nm and 700nm points of the xy space. This is because there are no equivalent monochromatic sources for the color range corresponding to magenta; it is created when

there are two peaks in the visible range, one in the blue range and one in the red range, and they're stimulating the S and L cones while minimally stimulating the M cone.

Also important is a concept of "saturation". This is typically thought of as the "purity" of a color. To be more quantitative, in the xy color space there is an analog known as the "excitation purity", which is calculated as follows:

$$EP = \frac{\sqrt{(x - x_n)^2 + (y - y_n)^2}}{\sqrt{(x_I - x_n)^2 + (y_I - y_n)^2}}$$

Where x_n, y_n are the coordinates of the white point, x, y is the point in question, and x_I, y_I is the point where the xy color space perimeter intersects a line segment from the white point to the point in question. More simply, it's the "percent from the white point to the point of monochromatic light of the same hue in the xy color space". This also has a close analogy with the incident spectrum, in which it is inversely related to the full width half maximum of a peak.

Lastly is the concept of brightness. This parameter is most intuitive, just being the amount of light in the spectrum. As mentioned before, this parameter is not included in the xy color space, because there are only two coordinates. To fully include this dimension of color, the xyY color space is often used. Recall that x, y, and z are simply the normalized versions of X, Y, and Z, which do have magnitude information. Therefore, to include the magnitude of the X, Y, Z point while still having x and y normalized so it can be plotted in a graph ranging from 0 to 1 in both dimensions, Y can be included. The xyY color space is now a 3D space that carries all color information. This has a straightforward connection to an incident spectrum, that is, it is the height of a peak. However, care must be taken because this is also connected to the saturation of a color; if a sharply peaked Gaussian curve is doubled in height only, it is therefore more

pure in color.

Other color spaces are commonly used as well, and have different strengths and weaknesses as compared to the xy or xyY color spaces. The xy(Y) space(s) are advantageous in that they connect closely to human perception of light. However, they suffer from not having a very intuitive coordinate system; that is, if you have a color with coordinates x,y, you must do calculations to figure out what the coordinates are for a color with the same saturation but different hue, or a color with the same hue but higher saturation. In contrast, a color space like HSB (hue, saturation, brightness) is less connected to human vision but a lot more intuitive to use; the hue varies from 0 to 1 and covers all hues, the saturation varies from 0 to 1 and covers everything from white to the purest color of that hue, and the brightness varies from 0 to 1 and covers everything from black to the brightest form of that color.

2.2 Structural coloration

As discussed previously, since light spectra are really just wavelength distributions of light intensity, there's really no concept of "color" without an observer translating the incident spectrum into the information known as color. So, color is really an observer phenomenon, and isn't concerned with the source of the incident spectrum. However, the color perceived is entirely dependent on the spectrum itself, and different spectra are produced via different means.

Light spectra come from a variety of sources. Some come from the actual source that created the light, such as when looking at the spectrum of an incandescent light bulb (created from blackbody radiation) or the spectrum of a laser (created from photons given off by atomic transitions between energy levels). However, many light spectra we see are the result of light coming from some source that created it, interacting with some material, and then transmitting,

reflecting, or scattering off that material into our eyes. Therefore, the spectrum that reaches our eyes is dependent on both the spectrum of the source (the "illuminant") and the result of what the material does to light incident on it. If incident light is reflecting off the material from the source and going to our eyes, we call it a "reflection spectrum", $R(\lambda)$, which is mostly what we'll be discussing in this case. Therefore, if the illuminant is $I(\lambda)$, we calculate the reflected light simply by $S(\lambda) = I(\lambda)R(\lambda)$.

A few assumptions should be noted here, though. Implicit in this equation is that the action on each wavelength is independent; that is, R is the same for a given wavelength whether other wavelengths are present or not. Similarly, it tells us that light isn't converted from one wavelength to another. Lastly, in any system we're looking at, energy will be conserved at every wavelength; that is, R is bounded between 0 and 1. If $R = 0$ for a given wavelength, that means that all of the energy of the illuminant at that wavelength is either absorbed by the reflecting structure, transmitted through the structure, or scattered away (which could arguably be considered reflection in a direction other than the viewer). Similarly, R will at most be 1, meaning that all the incident light at the wavelength is reflected to the viewer; it will never be more than 1 in our system because that would imply that the material is imparting energy.

There are many mechanisms by which light can be altered by the material. Most commonly used throughout history, both by manmade objects and nature, has been "chemical coloration", in which molecules have strong molecular resonances that can absorb energy at a given wavelength. These resonances are often sharp and thus the colors can be very bright and saturated. These are, for example, the colors created by dyes and pigments found in plants and animals. They also tend to be non-iridescent, meaning that the color looks the same no matter the angle it is being viewed at. As we will see later, this is important.

Chemical coloration has many advantages. In addition to its possibility of giving bright, saturated, non-iridescent colors, it is relatively easy to produce, cheap, relatively simple to understand, and colors can be mixed simply in intuitive ways to create new colors. However, they suffer from some disadvantages and other technologies are always worth pursuing. For example, many dyes degrade in heat and sunlight, or are made of environmentally unfriendly materials. They are typically "found" from already existing materials rather than designed in a continuously tunable way. They also usually require more volume, needing to be painted onto a surface, for example, which can make a big difference for applications requiring light weight, such as coloring an airplane chassis. Additionally, while we are mostly interested in the visible spectrum in this context, it would be desirable to have a method for designing spectra that could as easily be used for the IR or UV light ranges.

To complement, not necessarily displacing, the incumbent chemical coloration approach, one could explore the area known as "structural coloration" (or "physical coloration" sometimes). The idea is that, rather than altering and interacting with incident light by means of molecular resonances, other methods that can be more easily controlled and designed are used. Several example mechanisms include interference, scattering, diffraction, and non-chemical means of absorption. Interesting, all of these mechanisms have been used in nature as well, such as butterfly wings, peacock feathers, beetle shells, opals, and our familiar blue sky. A comprehensive review is given by *Sun et al.*[86], covering the different mechanisms giving rise to coloration in nature with examples of each.

Structural coloration can solve some or all of the problems associated to chemical coloration outlined above. It is well-established, yet also a quickly growing field of research. Human use of structural coloration is ancient as well, from tempering colors of metal oxides [21, 52] to use of nanoparticles in making

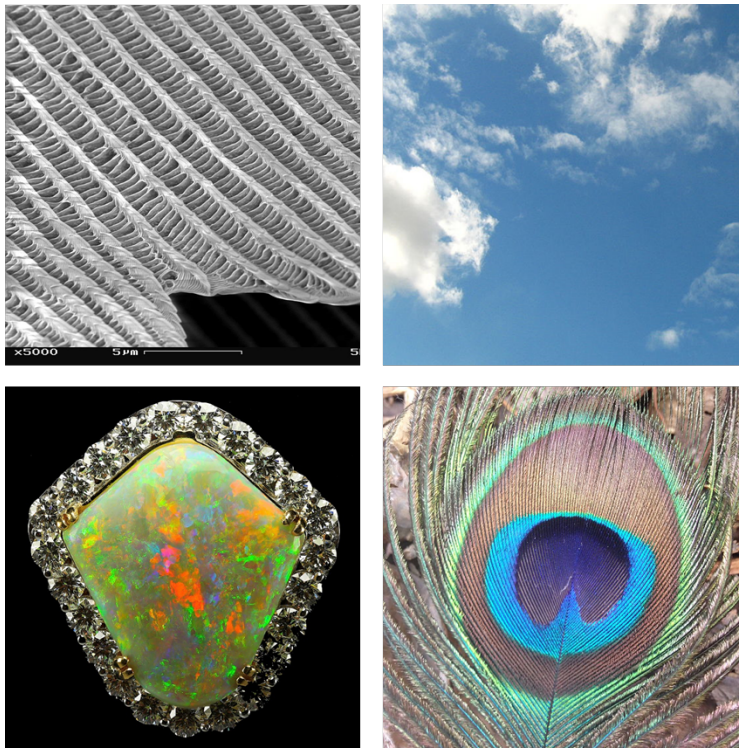

Figure 13: Examples of structural coloration. (a), (c), and (d) are ultimately interference mechanisms, (b) is due to Rayleigh scattering.

glass [18]. More recently, the advent of modern fabrication techniques has given rise to creative coloration methods. Notable recent examples are subwavelength plasmonic color filters [42], large-scale plasmonic pixel printing [43], actively controlled plasmonically enhanced pixels [25, 24], and a photonic crystal from anodized aluminum layers of alternating indices [94].

2.3 Fabry perot cavity structure

2.3.1 Overview

One very commonly employed mechanism of structural coloration is interference. Most generally, interference could be defined as when light is made to combine at a given spatial point, and in this combination either add together and be greater as a result (constructive interference), or cancel each other out and have less of a combined effect (destructive interference).

This can be employed in many platforms, but a very common one is in film interference, in which the light incident on a film reflects off the first surface it encounters, then penetrates into the film and reflects off subsequent interfaces. All the rays of light reflected back towards the viewer combine, and depending on how far they have traveled inside the film, and the nature of the reflections at the interfaces, all the rays of light will have different phase shifts associated with them. The electric field component of a plane wave ray of light in vacuum can be written mathematically as:

$$E(x,t) = E_0 e^{i(kx-\omega t)}$$

Where E_0 is the amplitude of the electric field, $k = \frac{\omega}{c} = \frac{2\pi n}{\lambda}$, n is the complex refractive index, x is its position in the x direction, t is time, and ω is the angular frequency. Because all the light in the systems we consider will be due to an incident ray, and all the light (independent of any medium it's in) will

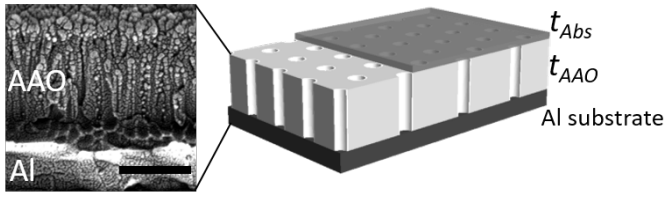

Figure 14: Schematic of the FP structure. Inset shows SEM image of typical AAO grown for this project. Scale bar is 500nm.

have the same angular frequency as the incident beam, we can effectively divide out the $e^{i\omega t}$ part of the E field term, and only consider its spatial dependence (keeping in mind that we can always multiply this term back in if we want to regain the time dependence).

At this point we have $E = E_0 e^{2\pi i n x/\lambda}$. From this it is evident that $E(x = 0) = E_0$, but $E(x = m\lambda/(2n)) = -E_0$ (where m is an odd integer). This means that if a ray of light travels a distance of an odd multiple of half the wavelength in the material and recombines with that original wave, the resulting wave will be 0 and they will have canceled out. Alternatively, for integers m, $E(x = m\lambda/n) = E_0$ again, and this wave will combine to form a resulting wave of value $2E_0$. So simply, we refer to the phase of the wave as everything in the exponent aside from the constant i (or, $2\pi n x/\lambda$), and understand that when the phase difference between two waves are an even multiple of π, they combine constructively, and when they're an odd multiple, they add destructively.

However, phase change due to path length difference is only one facet of interference. The other facet is phase change due to reflection. When a ray of light encounters an interface, due to the response of the different materials on each side of the interface, not only does the ray split into transmitted and reflected components, these components have a phase shift with respect to the incident wave, despite not having traveled farther.

It is worth mentioning a more general point about thin film reflection and

the FP cavity structure that will be discussed again momentarily. For all the light that enters a given structure, the light can only be some combination of transmitted (leaving through the structure in the direction it entered), reflected (leaving in the direction it initially came from), absorbed within the structure, or scattered (leaving the structure in directions other than the ones related to transmission and reflection). For the moment we also assume that there is no scattering, though we will amend this later.

If we assume that all the light is either transmitted, absorbed, or reflected, this is commonly written as $T + A + R = 1$, which gives insight to various platforms. For example, if we have a thin film of a lossless medium, meaning $A = 0$, then we know that for any interference that occurs, $T + R = 1$, and any dips in R must be paired with peaks in T, and vice versa. Similarly, for our FP structure, because we are assuming an optically thick Al substrate, $T = 0$ and we have $A + R = 1$. Because R is what we are interested in and measure or see, this tells us that any dips in R are cases where there is high absorption.

2.3.2 Comparison with single layer (just AAO) cavity

To illustrate the strong effects of the absorbing layer on top of the FP cavity structure, it is useful to compare its effects to the same structure but without the absorbing layer. As expected, the AAO layer still forms a cavity, but because there is much less opportunity for absorption (only in the Al), the dips in R are never as low, leading to low saturation because light of different wavelengths are all reflecting at similar amounts. This is illustrated with several AAO thicknesses in Fig. 15. It can be seen that there *are* dips in R, but they are very shallow. Also plotted is the energy loss in the Al (i.e, the integration of the magnitude of the E field in the Al), where the matching R peaks can be seen.

FP structures with the same AAO thicknesses, and a constant thickness

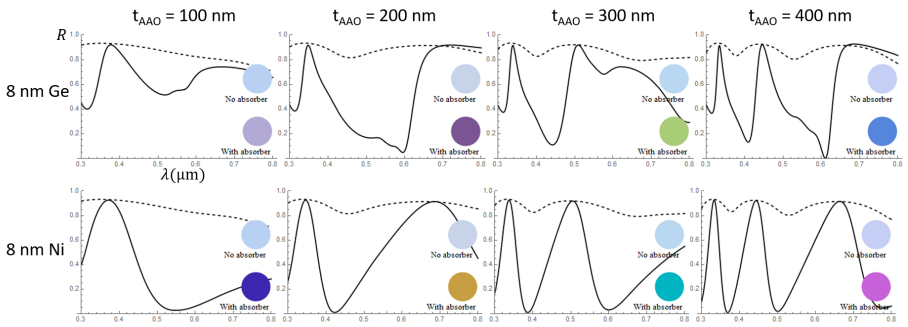

Figure 15: Illustration of effect of adding absorption layer. Top row is the effect of adding a 8nm Ge layer, bottom row is for 8nm Ni layer, each shown for 4 AAO thicknesses. In each plot, the solid curve is with the absorber layer, the dotted curve is without, and color disks are inset to show the color before and after adding the absorber layer.

absorber layer are also plotted for comparison. It can be seen that the position of the peaks and dips shift slightly (due to the extra path length of the light traveling through the absorber layer, as well as the phase shifts at the new interfaces). Additionally, the energy loss is plotted for the Al and the absorber layer. The difference is clear: while a small amount of energy is still lost in the Al, at the positions of the dips, the energy loss in the absorber is far greater.

2.3.3 Benefits and scalability

There are many benefits to structural coloration via an FP structure. One of the most important is about scalability of fabrication; many structural coloration methods rely on advanced fabrication techniques (for example, EBL) that will never be able to be mass produced. However, thin film deposition is already a well established process, widely used in producing anti-reflection coatings for eyeglasses, for example. Additionally, while DBR's can already produce very high saturation and brightness reflectance peaks, they typically require many layers to do so; this only requires a substrate (which would be there anyway), a dielectric that is already commonly grown on this material, and

an absorber layer. It is easy to imagine using a well characterized spray-on absorbing material that is known to dry at a given thickness.

2.4 Measurement methods

Several measurement techniques are used to characterize various parameters of experimental samples. Results from different methods are used to corroborate each other and verify the accuracy. The methods are detailed below.

2.4.1 Ellipsometry

Ellipsometry is a very common measurement method for getting optical information about a material. In its most basic form, it involves shining a beam of light of a known wavelength at a material and measuring the beam reflected off the sample. The beam is typically composed of light of two polarizations, and the sensor measures the ratio of the reflected beams of each polarization, as well as their phase shifts from the reflection. These can be used with the Fresnel equations to solve for the optical constants of the material. However (even if the Fresnel equations are being used internally), more commonly a model is chosen that the user expects accurately reflects the physical setup, and an iterative equation solver is used to fit the physical values (film thicknesses, surface roughnesses, and optical properties) that minimize the difference between the values returned by the model and the values measured by the machine.

There are different types of ellipsometers, but usually more measurements taken with slightly different measurement parameters gives more reliable information. For example, a very simple ellipsometer that operates at a single wavelength and incidence angle can be used to give a little information, but one that operates over a large spectrum and adjusts the incidence angle over some range can give better information. In addition, it often allows different

analysis methods to be used; for example, if the user wants to use an optical oscillator model to fit the optical constants, information from only one wavelength can hardly be used. In addition, the Kramers-Konig equations relate the real and imaginary parts of the optical constants of a material in a way that is separate from any sort of internal model, but require the values of those constants over a large range of wavelengths to reliably do so. Thus, one of the most comprehensive types of ellipsometry (used in this project) is Variable Angle Spectroscopic Ellipsometry, or VASE. The system used in this project is a J. A. Woollam M-2000 VASE that operates in a wavelength range of 190-1700nm and can measure at incidence angles from 45-90°. This machine provides very robust measurement of our system.

A few further details of these measurements should be mentioned. First, measuring bulk metals is typically very difficult to do, and less reliable than measuring the properties of films that are at least semi-transparent. This is because much of the information for semi-transparent films is gotten from interference occurring in the film. However, if the film is opaque, like in a metal, this information is impossible to get. A very thin film of metal that light may be able to transmit through can be deposited, but thin metal films often have different properties from their bulk forms. There are other reasons, that will be discussed later, for why measurement of metal optical constants is harder to do.

Another important point concerns the substrate the materials are measured from. For example, if we want to measure the optical constants of the material of our absorber layer, it will complicate matters significantly if we measured that layer by measuring the whole FP structure at once. Theoretically, it is possible to measure this structure, and if the properties of the other components of it (the Al substrate and AAO cavity) are already known, figure out the properties of the absorber layer, but in practice this will be much less reliable (in some cases, this

will be unavoidable; the AAO is necessarily grown on the Al so it will have to be measured on top of it). Instead, what is typically done is to deposit the film on the structure as it is desired, but simultaneously be depositing the same material on a standard, well characterized substrate such as polished Si, typically known as a "reference sample". This allows measurement of the material much more accurately. However, care should be taken to use a reference sample with a different enough refractive index. This is because, for the reflectance at an interface, R is proportional to the difference between the refractive indices of the different interfaces. Therefore, if they are too similar, it will be less of a "stable point" numerically when trying to figure out the location of the interface. To mitigate this, we use either SiO_2 ($n = 1.5$) or Si ($n = 3.5$) as the reference substrate.

2.4.2 Reflection measurements

Reflection measurements are another way of probing the optical properties of a structure. While VASE can potentially give much more detailed information, reflection measurements give less, but are more trustworthy and simple. In the simplest setup, a beam of monochromatic light is shone on the sample and the percent of the light that is reflected into a detector is measured. Naturally, this requires calibration. Typically, a highly reflective mirror is placed exactly where the sample is to be placed. The mirror's reflectivity is known beforehand from an external source, so it allows us to know what the total amount of incident light is, so the measurement of the sample itself will be an absolute measurement.

There are other details that can be used in reflection measurements, such as the polarization and incidence angle of the beam. However, commonly, when polarization is not being investigated, unpolarized light is used. Additionally, due to the difficulty of placing a detector and source in the same physical location (or using beam splitters), many reflection measurement setups don't actually

probe normal incidence, but slightly off-normal.

2.4.3 SEM

Scanning Electron Microscopy (SEM) is essential for investigating the surface of devices at a microscopic scale. For example, it is the simplest way to take a look at the pore structure of AAO. A FEI Helios NanoLab 600 DualBeam FIB/SEM was used extensively in this project, typically at an accelerating voltage of 10kV and a beam current of 1nA, to avoid charging the AAO.

2.4.4 Color balanced photography and LCD screen color compression

For a project about producing color, it's necessary to take color-faithful photos of samples produced. This is less simple than it sounds. As we've already discussed, perceived color is dependent on several factors: the illuminant, the sensor (our eyes, or the camera's sensor), and when we view a photo on a computer, the display on the screen. For example, a computer screen simply *can't* produce all the colors of the xy gamut. This is because an LCD computer screen's pixels produce colors by mixing light from three LCD colors (R, G, and B) in various combinations. Two colors in the xy color space can produce any color along the line connecting them by mixing them in some ratio; therefore, three colors can produce any color in the triangle enclosed by them. However, the xy space has a horseshoe shape, and thus can't be totally enclosed by any three points inside the shape. Typically, the actual color gamut available to computer LCD screens is far smaller. Therefore, what happens if the computer wants to display a color outside of its available gamut (or, what is it even doing when it's displaying the xy color space itself, if it technically can't show colors outside that gamut)?

Inevitably it is essentially "compressing" the colors. If we take the "real"

xy color space that was physically produced somehow, all the colors an LCD screen can display are in that RGB triangle. Therefore, when the xy color space is displayed on an LCD screen, it is actually composed of only the colors in that triangle. So that triangle is somehow mutated to fill out the horseshoe shape, so that the colors we see are changing from point to point inside of it. There is no procedure to correct for this but it still must be noted.

The colors recorded by a digital camera are another story. As discussed, the color we (or a camera) perceives is highly dependent on the spectral shape of the illuminant. Now, we could measure the spectrum of the illuminant (and we did), but what is a far more reliable and standard technique is to have a known reference sample next to the object you want to photograph. This is typically done with a "gray card", essentially a few cards of varying graytones. Because it is known how they should appear under various lights, by analyzing the color the gray card appears as, the illuminant can be figured out, and compensated for in the rest of the image.

2.5 Fabrication methods

2.5.1 Al/Ti on Si

The benefits and disadvantages of various types of Al substrate were previously discussed in the first section. Unless otherwise noted, evaporated Al on an Si wafer was the substrate used in this project.

The fabrication parameters are as follows. Polished Si wafer was cleaved into $\sim$ 1cm $\times$ 5cm strips, and then sonicated sequentially in acetone and then methanol. Finally, isopropyl alcohol was used to rinse the methanol and it was dried with a nitrogen jet. Samples were then loaded into the evaporator, which was pumped down to a pressure in the 10^{-6} mTorr range. At this point an adhesion layer of Ti was deposited by e-beam evaporation, typically about 5nm

at a rate of $\sim 1\text{Å}/s$ while the substrate holder was rotated. The Al layer, $\sim$ 600nm, was then deposited at a rate of $\sim 5\text{Å}/s$, also by e-beam evaporation. All evaporation was done at room temperature and the rate was monitored by a quartz crystal thickness monitor.

2.5.2 AAO

A standard oxalic acid anodization process was carried out to form the AAO. 0.3M oxalic acid was used as the electrolyte, cooled in a chiller to 10°C and magnetically stirred vigorously. The anodization voltage was 40V, with the positive bias applied to the sample and ground applied to the graphite counterelectrode. The anodization rate, as measured by VASE, was roughly $\sim$100nm/min.

Anodization was often performed in steps, to have sections on a given sample with different AAO thicknesses. To do this, nail polish was applied to the sample in sections and allowed to dry, before sequential anodization steps. The nail polish acted as a mask, preventing the areas it covered from being anodized.

Unless otherwise noted, no first anodization (sacrificial layer) or electropolishing was done; because of the necessarily large Al depth needed to create the sacrificial layer, and the violent nature of the electropolishing, these couldn't be done with the Al thickness limited by evaporated films. In addition, unless otherwise assumed, no pore widening was performed.

2.5.3 Absorber layer evaporation/masking

Finally, an absorber layer was evaporated on top of the AAO cavity. All materials used for the absorber layer (discussed later in more detail) were e-beam evaporated, typically at a rate of $\sim 1\text{Å}/s$. This allowed a slow enough rate for uniform deposition, but not slow enough that the uncertainty in thickness monitor could lead to large inaccuracies.

Similarly to the application of the nail polish, a shadow mask was used to physically cover parts of the sample before evaporation, allowing different regions to have different absorber thicknesses.

2.6 TMM analysis

Even if how the phase of the E field of light changes with position is known, and how it behaves (transmitting and reflecting) at a given interface, it is still not clear how to solve for the reflectance of a given system, such as our FP structure. For example, one can talk about a ray of light incident on the structure. A naive approach suggests that it encounters the first interface and splits into a reflected and transmitted term, and then the transmitted term propagates through the first medium until it hits the next interface, at which it splits again, and so on. Thus, the resulting reflected wave is the sum of an infinite number of waves that were split and reflected from the original wave.

The reflection can be calculated directly as described, choosing to cut off the summation of split waves at some point, knowing that the ones representing a wave that has reflected internally many times will be diminishing anyway. This is one approach, known as "partial wave summation", which can be done and is discussed later. However, a more elegant and more easily calculable solution is desired.

An alternate strategy can be realized if an insight is made about the previously described partial wave summation. If there is an incident ray, like a pulse where the light source is quickly turned on and off, then the subsequent reflections described are following it in time. However, in reality, if any light is incident on a structure, if it is continuously shining for any length of time (compared to the length of time the light spends in the structure), meaning that even though the reflections for a single pulse would occur after, because

the incident light is now continuous, all the reflections are occurring at the same time as the initial pulse. Thus, it is fair to say that the superposition of all of them are occurring in each medium of the structure, and the combined effect can be written as simply a single wave vector of forward and backwards going waves:

$$E(z) = E_{0+}e^{ikz} + E_{0-}e^{-ikz}$$

or, in vector form:

$$\begin{bmatrix} E_+(z) \\ E_-(z) \end{bmatrix}$$

So in each medium, the E field at every point z can be represented by this vector, composed of forward and backward traveling wave components. How can the values of these E fields in different media be calculated though?

The process used to do this is known as the Transfer Matrix Method (TMM). There are two aspects of this process: 1) Boundary matching, and 2) wave propagation. It should also be pointed out at this time that that normal incidence is a special case that simplifies to a one dimensional problem, but we wish to have a more general solution that can account for different incidence angles. However, non-normal incidence then introduces the complication of E field polarization. However, fortunately, it is possible to convert the non-normal incidence case to closely resemble the normal incidence case, with some minor substitutions. Therefore, we first present the normal case, and then later include the substitutions that allow for incidence angle variation. Here we use a formulation similar to that presented by *Electromagnetic Waves and Antennas* [79].

We first briefly discuss boundary matching. For a boundary between two media, Maxwell's equations dictate that there must be continuity of certain

components of the electromagnetic field. Namely, $E_\parallel$, $H_\parallel$, $B_\perp$, $D_\perp$, where the subscripts denote the parallel and perpendicular components of each term, respectively, $D = \epsilon E$, $B = \mu H$, and ϵ and μ are the dielectric constant and magnetic permeability, respectively. In the most general form, ϵ and μ are tensors, so D and H can have different directions than E and B (respectively), but for the media we're considering, they are scalars. Therefore, D is in the same direction as E and H is in the same direction as B. Furthermore, ϵ is typically very different between different materials, but $\mu = 1$ for most materials, which is assumed here. This causes some of the boundary matching equations to become redundant and simplifies matters.

The important result of this is that for the electromagnetic waves to obey these lower level continuity relations, the fields at the same position on the interface between two media but on either "side" of the interface (an apostrophe on a variable indicates it being on the other side of the interface as the variables without them), can be related by a matrix known as the "boundary matching matrix":

$$\begin{bmatrix} E_+ \\ E_- \end{bmatrix} = \frac{1}{\tau} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} E'_+ \\ E'_- \end{bmatrix}$$

where

$$\rho_T = \frac{n - n'}{n + n'}$$

and

$$\tau_T = \frac{2n}{n + n'}$$

So, this demonstrates how to calculate the fields on one side of an interface

if you have them on the other side already. The other aspect of the TMM is propagation. It is clear that, for an EM wave traveling only in the $+z$ direction, if the value of the E field at point z is $E(z) = E_0 e^{ikz}$, then the value at position $z + d$ (still within the same medium) is $E(z + d) = E_{0+} e^{ik(z+d)} = E_{0+} e^{ikz} e^{ikd} = E(z) e^{ikd}$. So there is a clear function for translating the value of the E field some distance within a medium. However, because our EM wave is generally the combination of the forward and backward propagating waves discussed above, the backward going wave must also be modified by this translation, but by a factor of e^{-ikd} instead. Therefore, we can write this in matrix form, as the propagation matrix:

$$\begin{bmatrix} E_{1+} \\ E_{1-} \end{bmatrix} = \begin{bmatrix} e^{ik_z d} & 0 \\ 0 & e^{-ik_z d} \end{bmatrix} \begin{bmatrix} E_{2+} \\ E_{2-} \end{bmatrix}$$

Lastly, since reflectance is the quantity we want, it must be defined. It is the ratio between the power of the incident and reflected EM waves:

$$R = \frac{P_{refl}}{P_{in}}$$

Where

$$P_{in} = \frac{\epsilon}{2} |E_+|^2$$

and

$$P_{refl} = \frac{\epsilon}{2} |E_-|^2$$

$$R = \frac{P_{refl}}{P_{in}} = |\frac{E_+}{E_-}|^2$$

Now we need to make some substitutions to turn this one dimensional solution into a more general angle-dependent version that can handle different polarizations. The key to this is recognizing that if the light ray is incident at some angle θ, its k vector (that was previously a scalar in the z direction, k_z) now has an x component as well (the coordinate system can be defined so that the k vector only has x and z components), but the behavior of the wave can still be calculated as a modified one dimensional problem in the z direction. That is, the incidence angle certainly matters, but it only changes the values that get used in the formulation already described for the one dimensional case.

The necessary substitutions are as follows. The E field being solved for now is the *transverse* E field:

$$E_T(z) = E_{T+}e^{ik_z z} + E_{T-}e^{-ik_z z}$$

with the transverse E field vector:

$$\begin{bmatrix} E_{T+} \\ E_{T-} \end{bmatrix}$$

$E_T(z)$ now propagates with the transverse wave vector:

$$k_z = kcos\theta = 2ncos\theta$$

and is boundary matched with the transverse reflection and transmission coefficients:

$$\rho_T = \frac{n_T - n'_T}{n_T + n'_T}, \tau_T = \frac{2n_T}{n_T + n'_T}$$

which now instead include the transverse refractive indices:

$$n_T = \begin{cases} \frac{n}{cos\theta}, \text{TM polarization} \\ \\ ncos\theta, \text{TE polarization} \end{cases}$$

The E field amplitudes are determined with simple geometry depending on the polarization. For the TM case, the full E field has transverse and non-transverse components, so $E_{T+} = E_+ cos(\theta)$. For the TE case, the full E field only has a transverse component at all, so $E_{T+} = E_+$. The expression for the reflectance is the same, because even though the incident and reflected power terms could pick up a factor of $cos\theta$, it will be in both terms and thus get canceled out when they divide.

2.6.1 Practical solution of TMM

The above explains the theory of the TMM well enough, but it's still a different matter to use it practically. At this point we have presented how to propagate the E field vector across a medium, or match it at the interface from one medium to the next, so the last step is to connect it to a "starting point", that is, the incident fields. Because many of the following results are calculated or verified using the TMM, we briefly discuss how to solve for the reflectance of a multilayer stack at an oblique angle here.

In the following, we assume the following: an incident angle θ_0, a list of the refractive indices for each medium $\{n_i\}$, and a list of the thicknesses of each medium $\{l_i\}$. We first note that with θ_0 and $\{n_i\}$, Snel's law gives a list $\{\theta_i\}$ of the propagation angles of the full E field in each medium:

$$n_0 sin\theta_0 = n_i sin\theta_i = ...$$

With these, the transverse "phase thickness" for each medium can be calcu-

lated:

$$\delta_i = \frac{2\pi}{\lambda} n_i l_i cos\theta_i$$

Now we use these and the previously defined transverse coefficients to write the relation that both propagates and boundary matches the E fields at one interface to the point in the next medium at the next interface:

$$\begin{bmatrix} E_{Ti+} \\ E_{Ti-} \end{bmatrix} = \frac{1}{\tau_{Ti}} \begin{bmatrix} 1 & \rho_{Ti} \\ \rho_{Ti} & 1 \end{bmatrix} \begin{bmatrix} e^{j\delta_i} & 0 \\ 0 & e^{-j\delta_i} \end{bmatrix} \begin{bmatrix} E'_{T,i+1,+} \\ E'_{T,i+1,-} \end{bmatrix}$$

This holds for every medium and its adjacent layers. Note that this is recursive, because the result from doing this operation from one medium to the next can then be plugged into the next medium after that:

$$\begin{bmatrix} E'_{T,i+1,+} \\ E'_{T,i+1,-} \end{bmatrix} = \frac{1}{\tau_{T,i+1}} \begin{bmatrix} 1 & \rho_{T,i+1} \\ \rho_{T,i+1} & 1 \end{bmatrix} \begin{bmatrix} e^{j\delta_{i+1}} & 0 \\ 0 & e^{-j\delta_{i+1}} \end{bmatrix} \begin{bmatrix} E'_{T,i+2,+} \\ E'_{T,i+2,-} \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} E_{Ti+} \\ E_{Ti-} \end{bmatrix} = \frac{1}{\tau_{Ti}} \begin{bmatrix} 1 & \rho_{Ti} \\ \rho_{Ti} & 1 \end{bmatrix} \begin{bmatrix} e^{j\delta_i} & 0 \\ 0 & e^{-j\delta_i} \end{bmatrix} \frac{1}{\tau_{T,i+1}}$$

$$\times \begin{bmatrix} 1 & \rho_{T,i+1} \\ \rho_{T,i+1} & 1 \end{bmatrix} \begin{bmatrix} e^{j\delta_{i+1}} & 0 \\ 0 & e^{-j\delta_{i+1}} \end{bmatrix} \begin{bmatrix} E'_{T,i+2,+} \\ E'_{T,i+2,-} \end{bmatrix}$$

Now the recursion must be ended to give a useful result. We use the fact that, in the very last medium (in order that the incident light reaches), it will have a forward traveling wave (transmission from the incident), but it can't have any backward traveling component. Therefore, the recursion is ended:

$$\begin{bmatrix} E_{T,M+1,+} \\ E_{T,M+1,-} \end{bmatrix} = \frac{1}{\tau_{T,M+1}} \begin{bmatrix} 1 & \rho_{T,M+1} \\ \rho_{T,M+1} & 1 \end{bmatrix} \begin{bmatrix} E'_{T,M+1,+} \\ 0 \end{bmatrix}$$

From the above, it can be seen that the operation of recursing (plugging in the formula for one medium into the previous one) just amounts to multiplication of 2×2 matrices and some constants. Therefore, the final form, including the recursion ending term, will have the form:

$$\begin{bmatrix} E_{T,0,+} \\ E_{T,0,-} \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} E'_{T,M+1,+} \\ 0 \end{bmatrix}$$

And, because only the ratio of the reflected E field to the incident one is needed:

$$\frac{E_{T,0,-}}{E_{T,0,+}} = \frac{m_{21}}{m_{11}} \Rightarrow R = |\frac{m_{21}}{m_{11}}|^2$$

2.7 Color calculation

2.7.1 Overview

We've now explained the basic concepts behind how light is reflected from an object, how that reflection spectrum is calculated, and how that reflected spectrum is translated into a color that has meaning to humans.

Recall that for an illuminant $I(\lambda)$ and reflection spectrum $R(\lambda)$, and the color matching functions $x(\lambda), y(\lambda), z(\lambda)$, the tristimulus values XYZ are calculated like so:

$$X = \int_{380nm}^{760nm} S(\lambda)\bar{x}(\lambda)d\lambda$$

$$Y = \int_{380nm}^{760nm} S(\lambda)\bar{y}(\lambda)d\lambda$$

$$Z = \int_{380nm}^{760nm} S(\lambda)\bar{z}(\lambda)d\lambda$$

$$x = \frac{X}{X+Y+Z}$$
$$y = \frac{Y}{X+Y+Z}$$
$$z = \frac{Z}{X+Y+Z}$$

We previously discussed how to calculate the reflectivity spectrum, so now we must discuss the illuminant and the color matching functions.

2.7.2 Illuminant

The illuminant is the power spectrum of the light incident on the reflecting object. Although this only changes the perception of the object, rather than its behavior, it can be very important. This is because the reflected light at a wavelength is proportional to the strength of the illuminant at that wavelength. Therefore, if the illuminant is not an even distribution across the visible spectrum, it will change the perception of the reflected spectrum in what could be considered inaccurate or deceptive ways. Most people are familiar with this, if they're ever seen a white object under a colored light; for example a white sheet normally looks white under typical white light because it is reflecting all wavelengths roughly equally and all those wavelengths are initially about the same strength, but under red light, it is still reflecting all wavelengths equally, but there is more red light to begin with, and thus it appears red. Another way of saying this is that the viewer only gets the product of the illuminant and reflecting object, so without either an expectation of what the color of the object should be under another illuminant, or another object as reference, it's impossible to tell what part of the color is due to the illuminant vs the object's

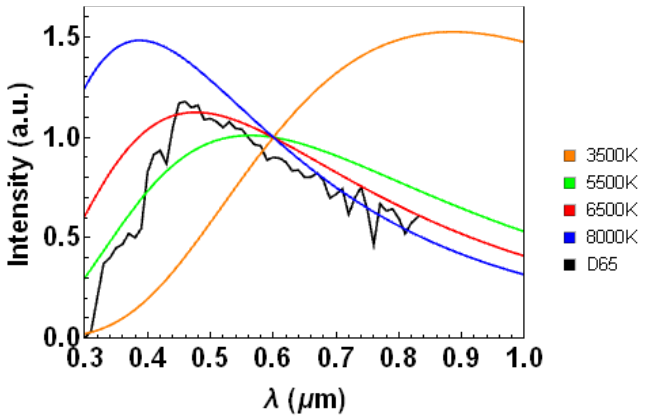

Figure 16: Blackbody radiation spectra of several temperatures compared to the D65 illuminant used in this paper. It can be seen that a 6500K blackbody source approximates D65 well.

reflectivity.

Completely evenly distributed illuminants are actually very rarely produced. For example, a common way of describing spectra is by specifying a temperature of a blackbody source. Blackbody radiation produces a smooth, peaked curve as a function of wavelength. For high enough temperatures ($\sim 5,000-7,000$K), a significant part of the spectrum is distributed across the visible range (this is what is occurring when a metal is heated "red hot"). In addition, the spectrum we receive from the sun at noon above the equator is relatively close to a blackbody radiation spectrum at 6,500K, known as a D65 illuminant. It should also be noted that the spectrum we receive from the sun changes throughout the day, season [27], with altitude [61], and other factors. Several blackbody radiation spectra and the D65 solar spectrum are plotted in Fig. 16 for comparison.

Because D65 is a standard, well known illuminant that can be closely reproduced anywhere, it is used in this paper.

2.7.3 Color matching functions

The color matching functions are a way to quantify how humans (taken as an average, because there are perception variations across populations) perceive incident spectra. Though this may seem like the most trivial part of color calculation, it is actually somewhat unintuitive. The goal is to be able to predict, given any possible spectrum across the visible range, what the perceived color will be. This is tricky,f because even if you measured responses to many different spectra, there are infinite possible visible spectra.

This procedure was actually originally done for the RGB color space, in 1931, before being transformed to the XYZ color space (A re-determination of the trichromatic coefficients of the spectral colours). The crux of the experiment is as follows. A human subject was made to look at a split screen. A given "test" color was projected on one half of the screen. On the other side of the screen, the resulting color mixture of three different primary sources were projected. However, the brightness of each of these primary sources could be changed by the subject. They were instructed to tune the brightnesses of each primary light source so that the colors on the two sides of the screen matched. In this way, the experimenters were able to find the composition breakdown of any test color in terms of the primary sources.

2.8 Gamut calculation

2.8.1 Overview and motivation of gamut calculation

We have now explained how to calculate the perceived color for a given FP structure with a chosen nAbs, t_{Abs}, and t_{AAO}, with a D65 illuminant. This is good, but only tells part of the story. If we wish to see the full potential of a structure, we would like to see the full "color gamut" it can produce. To this end, we must first define a color gamut; simply put, it is an area of possible

colors, in the most general sense. For simplicity, we could plot this gamut in the xy color space, where it would cover some area. For example, one could look at the gamut of the solar spectrum throughout the day, as if you looked directly into the sun. At a given time, the solar spectrum would be transformed (via the color matching functions) into a color which you could plot in the xy color space; as the spectrum changed, it would plot out a sequence of points. In this simple example, the gamut would simply be a line (see Fig. 17). However, if another dimension of control was added in addition to the solar spectrum, for example, a red lamp that could be dimmed continuously, then plotting out all the possible colors that result from the solar spectrum and the red lamp at various dimming settings both shining into the eyes would result in an area of points. This is illustrated in Fig. 17, where the line gamuts for several solar spectra at different dimming levels of the lamp are shown. Each one is a line, but it can easily be seen that if the dimming was swept continuously, it forms an area. So in this contrived example, the gamut is of the combined solar spectrum/red dimming lamp system.

In a similar way, we can speak of the color gamut of an FP structure. We could look at the effect of many parameters, but the two most obvious ones for this structure are t_{Abs} and t_{AAO}. It's worth first looking at a few limiting examples which will guide further analysis.

2.8.2 Calculation of gamut boundary from discrete points

So we have discussed how points are calculated and placed on the chromaticity plot. We have shown that by varying whatever parameters are under our control, we can produce many points that cover a large area. Now we wish to show that 1) with fine enough granularity in the the parameters, any value in a given area can be produced, and 2) we can plot a more easily visualized "boundary" of the gamut formed by the points.

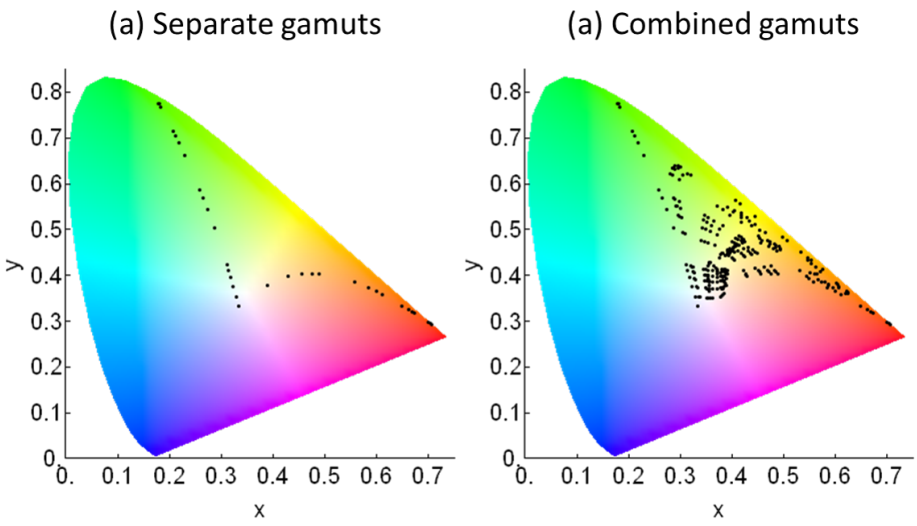

Figure 17: Illustration of the combination of two separate one dimensional gamuts (a) giving rise to a two dimensional gamut (b).

By the first goal, we mean the following. If a given number of points are plotted on the chromaticity plot, is it assured that a given point in between other points can be achieved for some combination of the parameters? It would be very difficult to definitively prove that, but it can be demonstrated that it seems likely that it can be reasonably achieved. The feasibility of it may be another matter, but that is discussed later.

The second goal is more of a visualization problem, but still interesting. It can be framed as follows. We can only plot a finite number of points on the chromaticity plot, but we want to plot an area that we can be reasonably confident represents the possible colors that can be achieved; not more, and not less. It is essentially searching for the "shape" of of a collection of points. One options would be to plot a large number of points, and then at each angle around the white point, take the outermost point, and then connect those points. Another option would be to use the mathematical concept of "hulls", which we have used in this project.

Hulls (in two dimensions, such as we're dealing with here) deal with an area needed to enclose a distribution of points in the 2D plane. However there is a little ambiguity here. For a distribution of points, there is a unique area that encloses all the points with maximum area as well as any lines connecting any two points. This is know as the "convex hull" and it is very simple to calculate (see Fig. 18). However, this often isn't useful; for example, consider a collection of tightly packed points that clearly resemble a shape like in Fig. 18. The convex hull for this collection would resemble just its outermost points and not represent the "shape" of the collection very well. Therefore, we often want to calculate a hull that "fits" to the "shape" of the collection more closely. This is called a "concave hull". However, while there is a unique convex hull, there is no unique concave hull; this is because, conceptually, the concave hull is wrapping to the points (such that it still contains all of them) with a certain factor of "tightness", very similar to shrink wrapping a physical object. Therefore, the tightness must be chosen carefully, though certain guidelines can be used. Different concave hulls for the same collection of points are shown in Fig. 19, created from different levels of tightness.

This tightness parameter is frequently called α. Because it's a relatively open problem, how to find the "correct" hull for a set of points (drawing attention to the fact that there's no objectively correct one), there are actually many proposed algorithms used for finding the result [6, 58], but that's beyond the scope of the project, being used merely as a tool.

2.9 Cartesian Chromaticity Plot

2.9.1 Overview and motivation

The xy color space is a great way to plot gamuts and coloration results for a number of reasons; it allows you to see how close to full saturation part of a

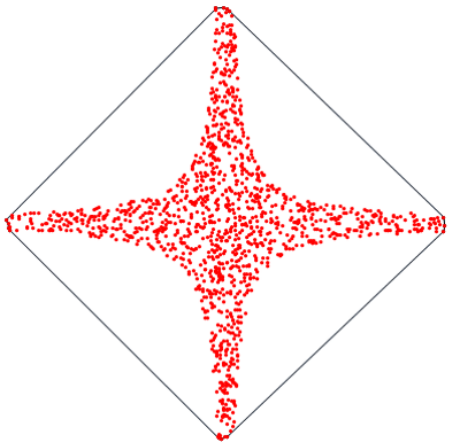

Figure 18: Example of a convex hull, determined only by its outermost points.

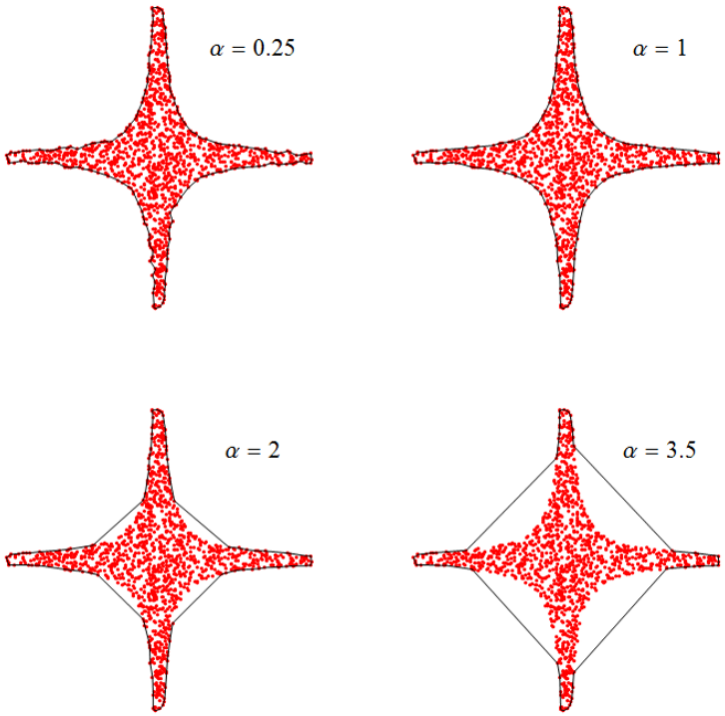

Figure 19: Several examples of concave hulls for different α parameters that determine the "tightness" of wrapping. It can be seen that $\alpha = 1$ provides the best result in this case.

gamut can reach, it allows you to easily see the area the gamut covers, and it has a close connection to human vision. However, it is somewhat removed from our actual structure, since it is really a more general tool that we're simply using here. Therefore, it's worthy to present the same data in a different form, which illuminates some interesting principles.

Some background is necessary here. So far, calculation of a gamut meant calculating for some range of t_{Abs} values and t_{AAO} values, the pairs $t_{Abs}, t_{AAO}, X, Y, Z$, and then calculating and plotting the corresponding points x,y on the xy color space. However, this is not helpful if you have a point t_{Abs}, t_{AAO}, x, y and you want to either predict where $t_{Abs} + \epsilon, t_{AAO}, x, y$ (for some small change ϵ) will be, or know what parameters are required to get a slightly different point, $x + \epsilon, y$. So, an alternate presentation is to plot all the same data points, but using axes of t_{AAO} and t_{Abs}, and at each point plotting the color itself. This has several benefits: 1) some emergent behavior is discovered, discussed later 2) while the xy plot doesn't show lightness information, each plotted point includes X,Y,Z, and therefore is actually the full color, and 3) it is actually more easy to choose a color, or similar colors to one color already in mind. We call this sort of plot a "Cartesian Chromaticity Plot" or CCP, because it plots chromaticities in a Cartesian plane. A typical one for a Fe absorber is shown in Fig. 20. As we will see, this visualization method is very useful.

2.9.2 Free spectral range, peak width, and color interference orders of CCP

Inspecting the CCP, several interesting points immediately emerge. Foremost, there are visible vertical "stripes". As expected, these are essentially interference orders, where the first one is due to one wavelength fitting in the cavity at that t_{AAO}, the 2nd one is due to two wavelengths fitting, etc. A few points of similar hue and their reflection spectra are plotted for different orders, illustrating this,

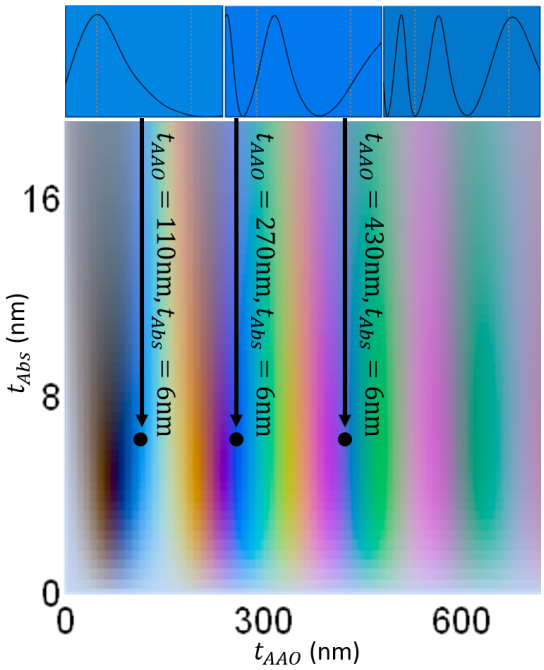

Figure 20: Example CCP with Fe used as the absorber layer. Three representative points are shown, with the same t_{Abs} but increasing t_{AAO}, to show that the same hue appears in different interference orders, with more peaks in the reflectance plot (top row).

in Fig. 20. However, a couple questions arise: 1) Why aren't all colors contained in a given order, if increasing t_{AAO} is essentially sweeping the resonance across the visible spectrum, and 2) Taking a closer look at the stripe in Fig. 20, the order of colors for increasing t_{AAO} is yellow, magenta, blue. However, we would intuitively expect the resonant wavelength to increase with increasing t_{AAO}, which would make it increase in the order blue, yellow, red. What's occurring here?

The explanation for both is due to the same important but not unexpected mechanism. As seen in Fig. 21, a zoomed in version of the first order stripe region, for increasing t_{AAO}, the resonant peak is certainly increasing in wavelength; nothing anomalous is occurring. However, a closer inspection of the reflection spectra, looking only at the range 400-700nm (the only range relevant to color), and a comparison to the plotted $x(\lambda), y(\lambda), z(\lambda)$ color matching functions, elucidates the answer. For a color that appears vivid, its peak wavelength activates the cone corresponding to that wavelength range, but just as importantly, the rest of its R spectrum *isn't* high, and therefore does not activate the other cones very much. On the other hand, looking at the R spectrum for the dull light green color in the middle of the selected points of Fig. 21 shows that, although it has a peak at a green wavelength, the whole spectrum is too broad and overlaps with every cone significantly.

So this answers the two questions posed above. To the first, the colors don't all occur because, even though spectra with peaks at the corresponding wavelengths occur, they don't match the human sensitivity cones well. From this, the second question is answered as well: what was intuitively assumed to be an interference order (a vertical stripe) is only based on our human color sensitivities. In reality, the interference "stripes" are in the "correct" order of increasing resonant wavelength, but not all the spectra give rise to color.

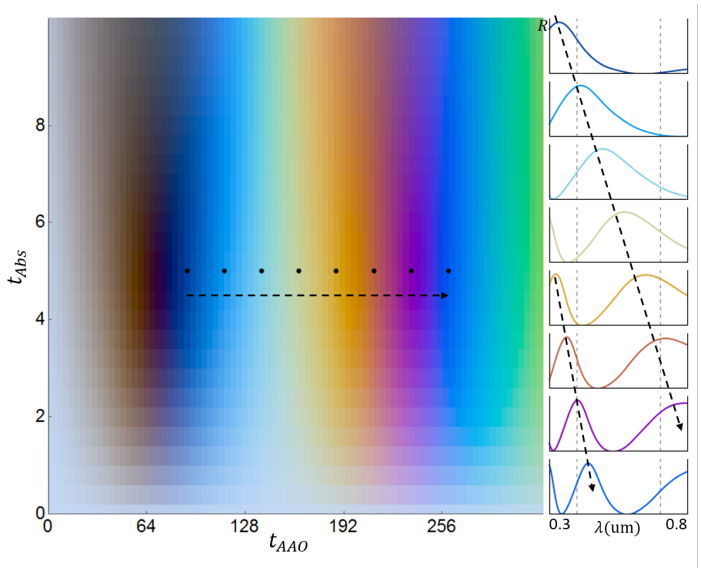

Figure 21: Reflection spectra for increasing t_{AAO} across an order in a CCP of Ni. For each R plot, the wavelength range is 300-800nm and the R range is 0-1. The peaks and valleys can be seen to be moving left to right as expected, but not all of them give rise to color, causing an illusion of colors corresponding to decreasing wavelengths for increasing t_{AAO}.

This is characteristic of problems involving structural coloration: because "color" is defined only in reference to coincidental human properties, the underlying physics must be understood and taken advantage of to mitigate this constraint. In this case, the underlying mechanism is due to the variation in the Free Spectral Range (FSR) and Full Width at Half Maximum of the R curves produced by this structure with respect to t_{AAO}. The FSR of an optical cavity can be simply explained as the wavelength spacing between two adjacent reflection peaks [31]:

$$\text{FSR} = \Delta\lambda = \frac{\lambda^2}{2nl}$$

where n is the refractive index of the cavity and l is the cavity width.

This can be illustrated by considering the CCP in Fig. 22 (and many others); green hues don't appear in the first stripe, but begin appearing for increasing t_{AAO}, before getting washed out for even larger t_{AAO}. This is due to both the FSR and the FWHM. It can be seen from the plotted R spectra that for all three examples, their peaks are at the same wavelength. The first spectrum, from t_{AAO}=150nm, has a large FSR, but its FWHM is large as well, activating all cones, as discussed above. For the next one, its FSR is large enough to keep the unwanted other peaks out of the visible range, yet the FWHM is small enough to only activate primarily the green cone. The last spectrum has nice peaks with a narrow FWHM, but at this point the FSR has decreased so that there are several peaks in the visible range, which activates several cones and muddies the color (discussed in more detail below). Green was chosen because it is in an especially difficult position to produce, having to overlap with the green cone spectrum in the middle of the visible range without overlapping with other ones too much. However, this principle applies to any desired color as well.

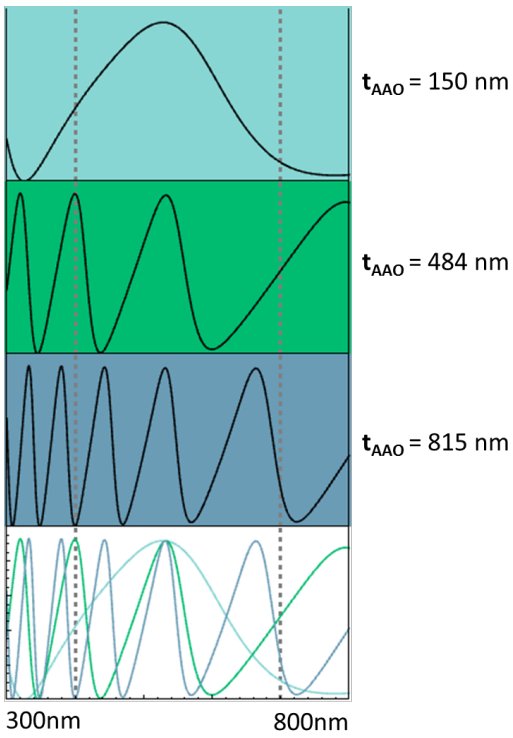

Figure 22: Separate reflectance spectra for three t_{AAO}, all with a t_{Abs}=7nm Co. The background color of each separate plot correspond to the color of each curve in the combined plot, which are the perceived colors produced by each spectra.

2.10 Color boundedness of structure

At this point we have introduced two of the main tools we will use to investigate and illustrate the structural coloration capabilities of the FP structure: the xy color gamut and the CCP. However, to be rigorous, one should ask: when we calculate the gamut as we have described, by numerically calculating a large number of discrete points and then finding their bounding contour on the xy CIE chart, how can we be sure that we have looked far enough? That is, how can we be sure that if we only searched up to some t_{AAO} and t_{Abs}, there aren't more vivid, "new" colors for structures with higher t_{AAO} and t_{Abs}? This section answers this question and gives more physical insight into this structure.

2.10.1 Limits of t_{AAO}

We wish to look at two limiting cases of t_{AAO}: $t_{AAO} \to 0$ and $t_{AAO} \to \infty$. For $t_{AAO} = 0$, it is simply a structure with an absorber on the Al substrate. Therefore, there is no dielectric cavity. However, this structure is not uninteresting; *Kats et al* [46] found that due to the complex phase shift at the dielectric/metal interface, interference effects can be seen in very sub-wavelength thickness dielectrics. However, this is a lower bound of our system and our TMM analysis will cover it as well.

The behavior of $t_{AAO} \to \infty$ is important to investigate, because if we wish to see the full gamut of possible colors that can be produced as a result of varying t_{AAO}, naively it could be assumed that we would need to vary t_{AAO} to infinity. However, this analysis shows that this is not the case.

To see why this is true, first recall what causes resonance for a given wavelength in a FP structure: resonance occurs when an integral number of wavelengths fits in the cavity. For some range (the range we'll be most concerned with), only one or two wavelengths in the visible range can fit an integral num-

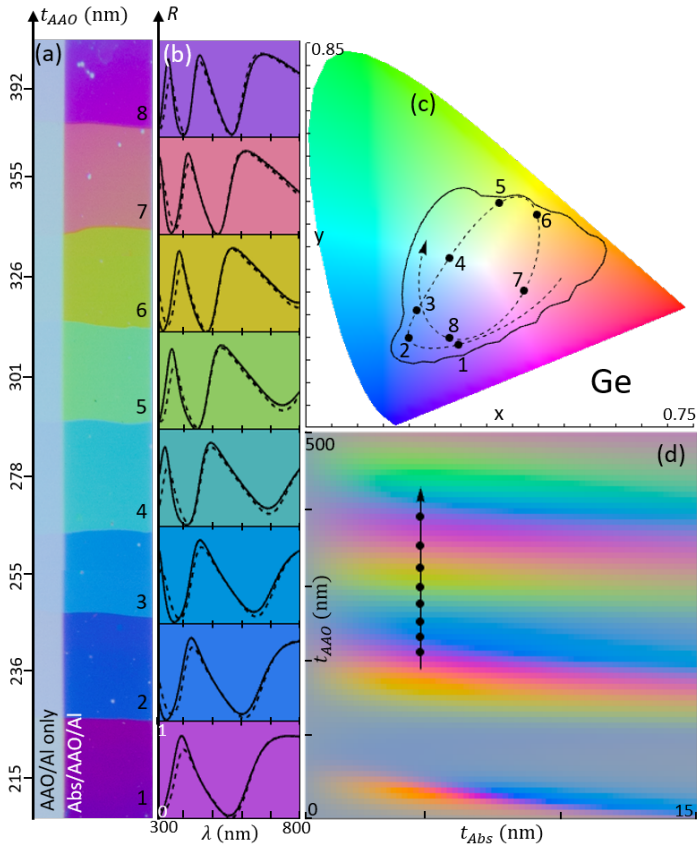

Figure 23: Fabricated sample with constant t_{Abs} and sections of increasing t_{AAO}. (a) Photo of physical sample. As labeled on the left, t_{AAO} is increasing from bottom to top. In the photo, the left side is just AAO, no absorber, and the right is AAO and absorber. (b) Measured (solid) and calculated (dashed) reflectivity spectra for the corresponding sections on the sample. The background color is the color calculated from the spectra. (c) xy plot with full gamut of Ge and corresponding points from (a) plotted along the path of increasing t_{AAO}. (d) CCP for Ge with corresponding points plotted.

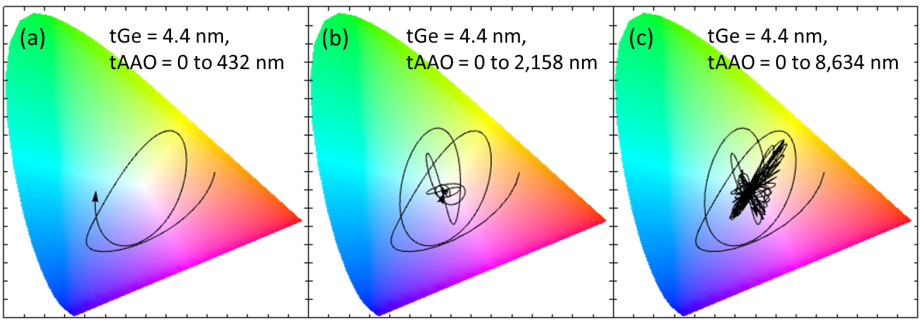

Figure 24: Demonstration of limiting effect of $t_{AAO} \to \infty$. Varying t_{AAO} up to ∼500 nm (a) produces new colors, but beyond that it just makes small oscillations around the white point.

ber of them into the cavity (for example, if t_{AAO} is 500nm, then . . .). However, if we look at a larger t_{AAO}, for example 3,000nm, then both 500nm and 600nm wavelengths would divide t_{AAO} evenly and have resonances. In fact, a t_{AAO} that is the product of any number of wavelengths will divide all of them and have many resonances all at once.

So what effect does this have? Effectively, having too many resonances at once is equivalent to having none; the reflection spectrum has too many peaks, activating all color matching functions at once, and the result is gray. This can be seen in Fig. 24, in which a path is shown for a constant t_{Abs} and increasing t_{AAO}; while t_{AAO} is increasing, for some range it still has some color, though they are becoming muddled by having an increased number of reflection peaks. This effect is manifested in its "spiraling to the gray point" behavior.

2.10.2 Limits of t_{Abs}

It is important to investigate the limits of this parameter as well, for the reasons stated above.

For $t_{Abs} = 0$, as we discussed previously, this is simply AAO on Al. This is still a cavity and will thus give rise to colors, but they tend to be faint because

AAO is non-lossy, and the Al substrate is mostly reflecting. Therefore, the only reflection peaks can come from loss in the Al substrate, which tends to be small. Examples of resonance for several AAO thicknesses and $t_{Abs} = 0$ are shown in Fig. 15.

For $t_{Abs} \to \infty$, the solution is also straightforward. Assuming the absorber layer is at all lossy (which is the case since we are only using lossy dielectrics and metals), the behavior of the cavity will be entirely dictated by the first interface reflection. Some of the light will be transmitted into the absorber layer and get absorbed as it propagates, while the rest of it will get reflected.

As t_{Abs} increases, some light will always still be able to get through it into the cavity, but investigating the most consequential path (i.e., the one contributing the most to any reflection) reveals that the amplitude quickly falls off: the light first enters the absorber, must propagate through it, losing amplitude, then either transmit into the AAO or reflect at the absorber/AAO boundary, and then propagate through the absorber thickness again, losing even more amplitude. Therefore, the range of thicknesses of t_{Abs} that can be used before the structure starts behaving as the bulk absorber is actually doubly limited because of this path an internal reflection must take, and in practice, for most absorber materials investigated, almost all color is gone for $t_{Abs} > 30nm$.

This is illustrated in Fig. 25, in which paths of increasing t_{Abs} is plotted for several t_{AAO} thicknesses. It is evident that the main range of the path is determined by t_{AAO}, and increasing t_{Abs} from 0 to some upper limit first increases the saturation (as the FP structure begins to occur), and then decreases it as the structure begins to be dominated by the bulk behavior of the absorber. It should also be noted that increasing t_{Abs} does change the hue, albeit not as drastically as the hue changes with t_{AAO}. This is actually an interesting detail which is later discussed.

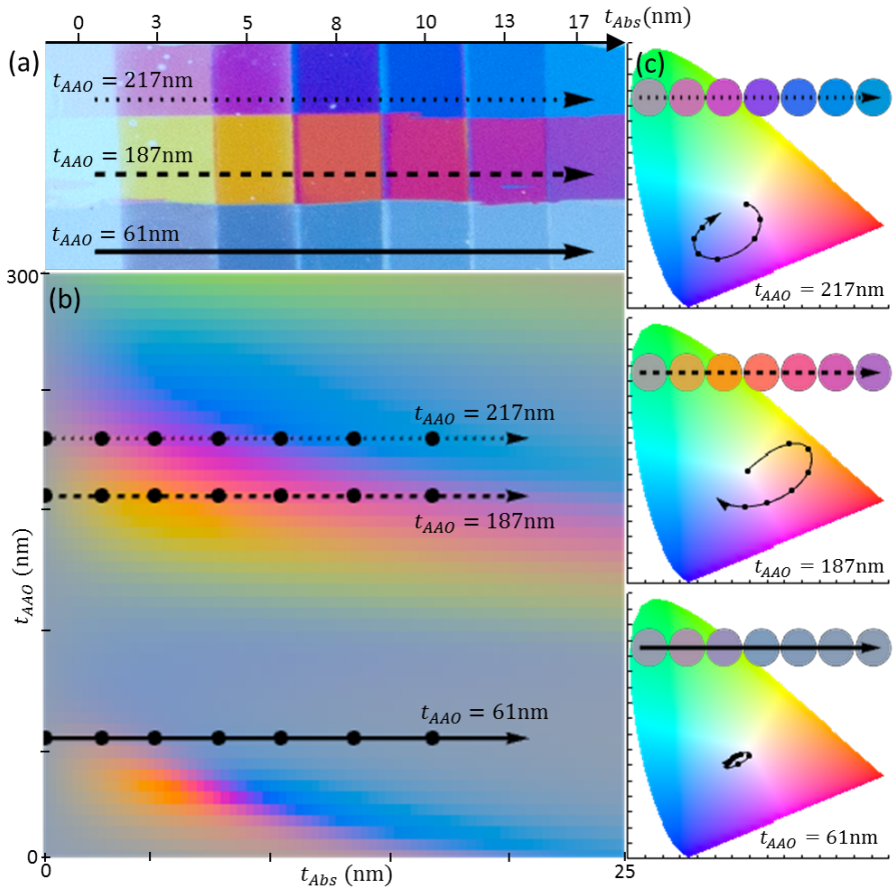

Figure 25: The effect of changing t_{Abs}. (a) A fabricated sample with each row a constant t_{AAO} thickness, and the absorber (Ge) thickness increasing from left to right. The tick marks at top indicate the absorber thickness for each column. (b) A CCP of the points from the sample in (a), showing them go through the color stripe order. (c) The paths on the xy chart for each row, with inset color disks.

2.11 Investigations of the FP structure

2.11.1 Gamut vs absorber layer

Scant attention in the literature so far has been paid to the role of the absorber layer itself. In Fig. 26, on the same macroscale sample, we anodized in steps and then evaporate a constant thickness of different absorber levels. The inset calculated color gamuts show their slight differences, and the measured points along the paths in them (corresponding to the actual points on the sample) verify the model and gamuts.

Below these, a plot of the excitation purity vs hue is plotted for each absorber, along its gamut, to more clearly illustrate the difference between absorber gamuts. It can be seen that the absorbers that can give rise to high EP for the red to green hue range (hue$\sim$0.05) have lower EP in the cyan to blue range (hue$\sim$0.7), and vice versa.

2.11.2 Gamut vs porosity

Because a controllable porosity is one of the main benefits of using AAO for the FP cavity structure, it is worthwhile to look at its effect, both on specific (t_{Abs}, t_{AAO}) points, and gamuts as a whole (i.e., the gamut for AAO that has had pore widening done to it to decrease its effective refractive index from 1.6). First, we simply look at the behavior of the color produced from the FP cavity structure for a few points as the pores are widened.

In the upper left of Fig. 27, three points are plotted on the xy chart for a FP cavity structure with different (t_{Abs}, t_{AAO}) parameters and a C absorber layer, and their paths are plotted for increasing refractive index. (decreasing porosity). The same is done for a Ni layer, in the bottom left of Fig. 27 (for different points). It can be seen that, as expected, because the optical path length is being changed, the resonant frequency changes over a wide range, similarly

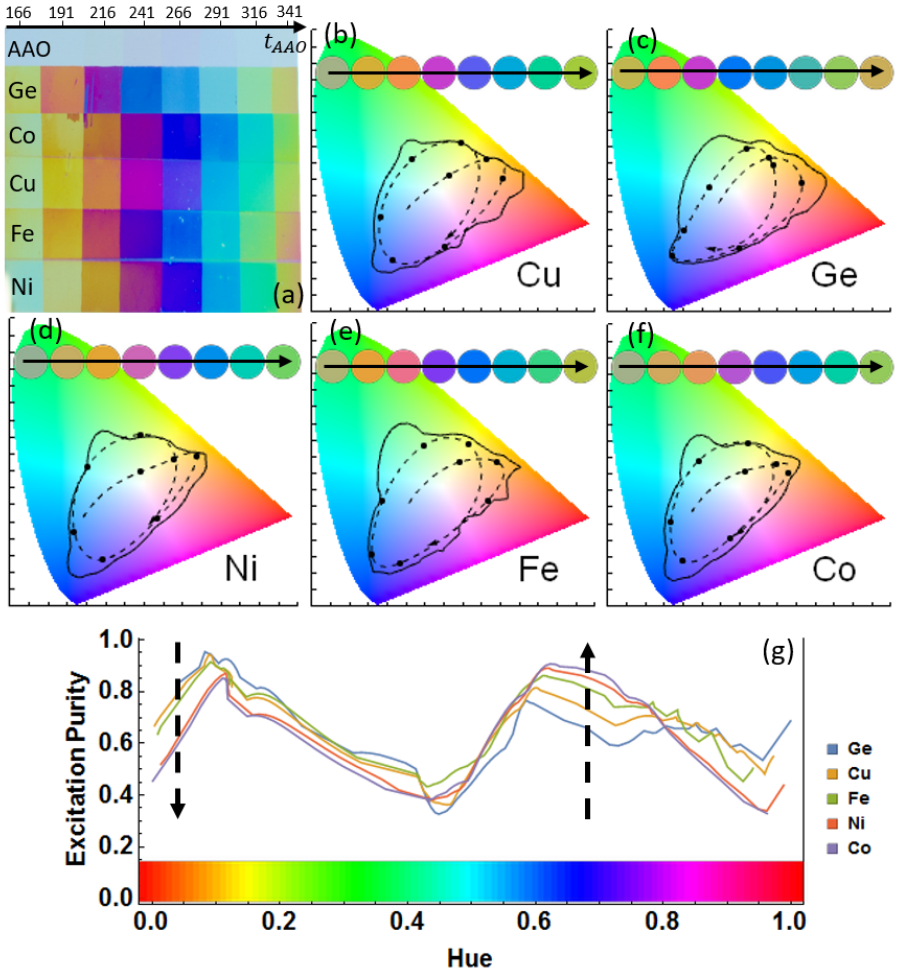

Figure 26: xy color gamuts for five different absorber layers. (a) A fabricated sample with five different absorber materials (in horizontal rows) as well as a row of only AAO at the top. For each absorber row, t_{AAO} is increasing from left to right, with the thicknesses as labeled above. (b-f) xy gamut for (Cu, Ge, Ni, Fe, Co) absorber layer with the points from (a) plotted along the path (dotted line) of constant t_{Abs} and increasing t_{AAO}.

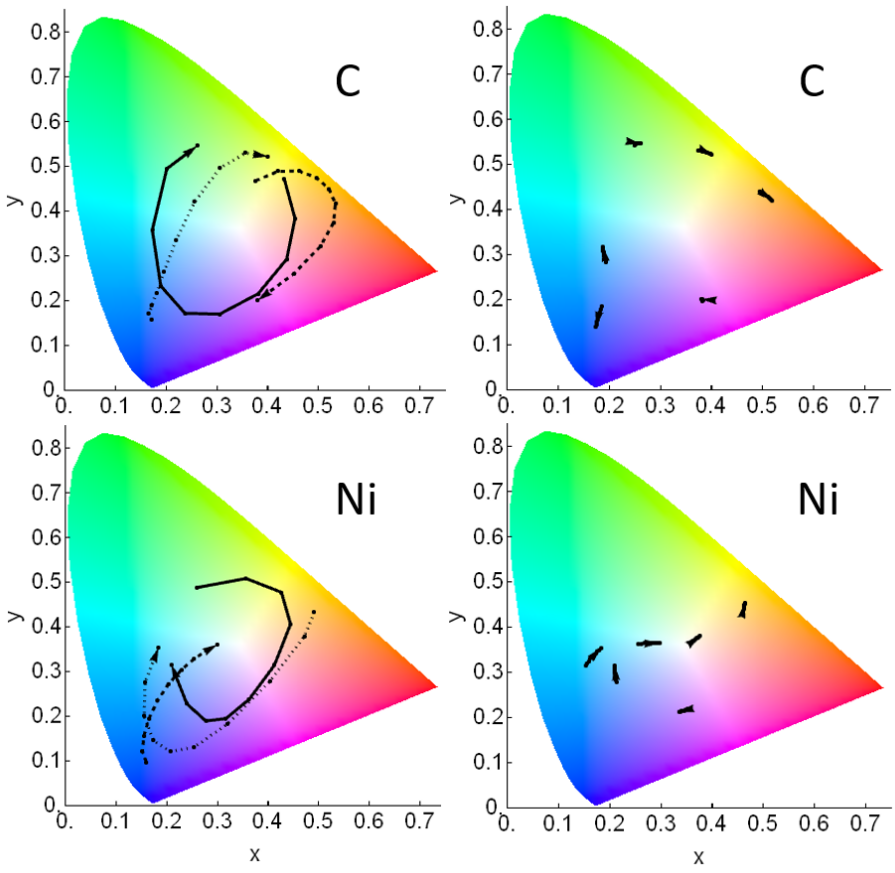

Figure 27: Example of colors changing for a few points of an FP cavity structure as the porosity is decreased. On the left are the paths for C and Ni absorber layers for simply increasing n_{AAO} and constant t_{AAO}. On the right, corresponding to the absorber materials on the left, are plotted points for increasing n_{AAO} with a corresponding decrease in t_{AAO} for each point to keep the optical path length constant.

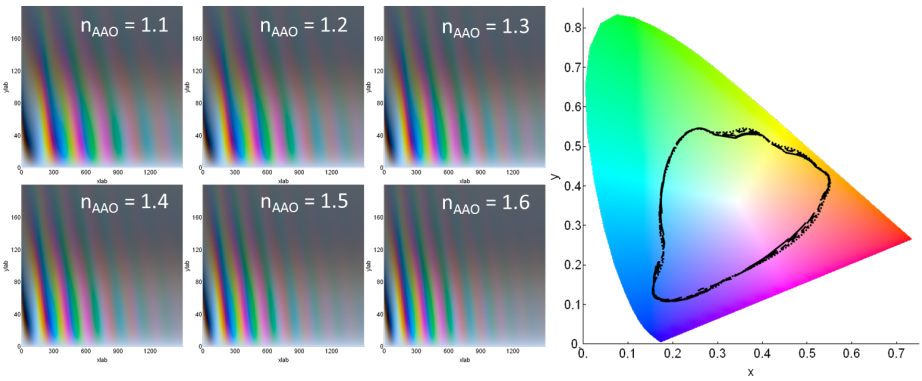

Figure 28: Gamut and CCPs for n_{AAO} 1.1-1.6 due to varying AAO porosity with a C absorber layer.

to simply changing t_{AAO}. However, on the right of Fig. 27, corresponding to the plots on the left, are points plotted for increasing n_{AAO}, but with a corresponding decrease in t_{AAO}, to keep the optical path length. The purpose of this is to see if the path length is truly all that effects the color. As illustrated, it pretty much is: because the propagation term, where most of the phase change comes from, is based on the product $n_{AAO}t_{AAO}$, keeping that constant keeps the phase change constant. However, a small amount of change can still be seen. This is due to the phase picked up at interfaces between the layers, which *only* depend on the refractive indices of each medium. However, these differences are overwhelmed by the propagation terms; the change from $n = 1.1$ to $n = 1.6$ is not very significant when considering that the reflection term is proportional to the difference between the complex indices of each medium, and AAO has a negligible imaginary term.

This demonstrates that coloration can be significantly changed by pore widening the AAO, but how does this change the gamut as a whole? This is illustrated in Fig. 28 and Fig. 29, where CCPs for a C absorber and Ni absorber (respectively) layer FP structure are shown for several AAO porosities. In addition, to the right for each figure, the gamuts are plotted on the same

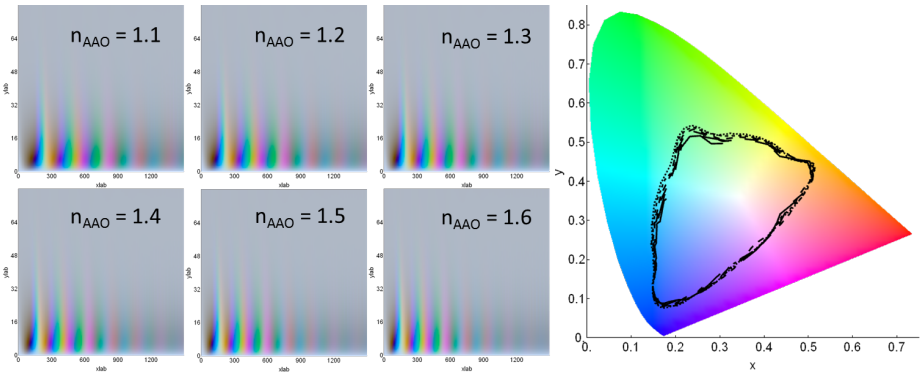

Figure 29: Gamut and CCPs for n_{AAO} 1.1-1.6 due to varying AAO porosity with a Ni absorber layer.

chart. Several points immediately emerge. First, looking closely at the CCPs, they appear mostly the same, but it's evident that the vertical stripes, the orders, are far more widely spaced out in t_{AAO} for the higher porosities. This makes sense: if n_{AAO} is lower, then the same difference in t_{AAO} will give rise to less accumulated phase from propagation. This could be useful, because precise control of the AAO thickness is experimentally difficult to achieve. Therefore, decreasing the sensitivity of the color to that parameter could make precision selection of colors easier to achieve. On the other hand, this would necessitate somewhat precise control of the etching rate, but this would likely be more easily controlled by using a very dilute solution of the etchant, making it take much longer (giving more "resolution" to the etching time).

The other noticeable effect is that the gamuts, overlaid on each other, are nearly exactly the same. This makes sense in light of the results on the right hand side of Fig. 27, in which it was shown that points of the same optical path length map to nearly the same colors. Therefore, every point in the gamut for porosity A can be produced by some point in gamut for porosity B, by choosing the one that has the same optical path length; the gamuts for different porosities have essentially a one to one mapping from one to the other.

While we have thus far demonstrated the results for AAO that can have its indices altered by etching (therefore, only decreasing n_{AAO} from its initial value), *Xue et al* [102] also demonstrated that the pores can be infiltrated with various materials. This means that it can actually increase the refractive index above that of the maximum achievable with AAO (1.6).

Infiltration with liquids and gases means that the color can be actively changed, but in reality the pores can be filled with anything. AAO is commonly used as a template for electrodeposited nanowires; this means that we can add a lossy component to the AAO in the form of metal nanowires.

2.12 Angular dependence study and iridescence suppression

2.12.1 Overview of iridescence

Iridescence is when appearance changes with viewing angle. Although there certainly are practical applications for iridescence, such as diffraction gratings, it is generally viewed as a disadvantage; certainly for consumer products, most people don't want an object that changes color depending on where the viewer is standing when they look at it. This is the major disadvantage of the FP structure. Taking a more extreme version of the FP structure (a multilayer film platform) results in a Distributed Bragg Reflector (DBR), which is a multilayer film structure made of many periods of alternating indices. The high number of layers give very sharp resonances and high purity colors, but also at the cost of high angular dependence. In this section, we discuss the cause of iridescence in these structures and ways of mitigating it.

The coloration strategies we've been investigating so far have been calculated for normal incidence, for simplicity. As we've already discussed, iridescence is an important factor for this structure, and we expect the color of a given

sample to change as we change the viewing angle. However, several questions arise: What's the nature of this angle dependence? Do some colors in the (normal incidence) gamut change less as the viewing angle increases? How can we mitigate this change, which is generally undesirable? How does it change for different polarizations? And, how does the gamut as a whole change?

2.12.2 General iridescence behavior of the FP cavity structure

First we look at the general behavior of the FP cavity structure as the viewing angle increases. Most vaguely, the color changes as a result of a non-normal viewing angle due to the accumulated phase changing, which can come from two sources: the different propagation length in the cavity, and the phase change occurring at the interface being different. Methods have been reported (discussed later) that cleverly design FP structures that make the propagation phase change cancel the interface phase change with increasing angle, to make a less iridescent structure. The exact mathematical cause can be seen in our derivation of the TMM, in which the propagation phase change δ_i is dependent on θ, as well as the transverse refractive index n_T, in addition to Snel's law.

Therefore, it's worth looking at a simple example of iridescence that occurs in our structure. However, this necessitates discussing polarization, which is a topic we haven't had to cover so far in our discussions, because at normal incidence, TE and TM polarizations are the same. To be clear, they are defined as follows. A "plane of incidence" is formed as the plane that contains the incident light ray and its corresponding reflected or transmitted ray. The polarization can be defined as the orientation of the incident E field with respect to this plane; two basic orientations are needed, and any other orientation can be written as a linear sum of these two basic ones. TE (transverse electric) polarization is when the E field is perpendicular to the plane of incidence; TM (transverse magnetic) is when the magnetic field is perpendicular to the plane of incidence.

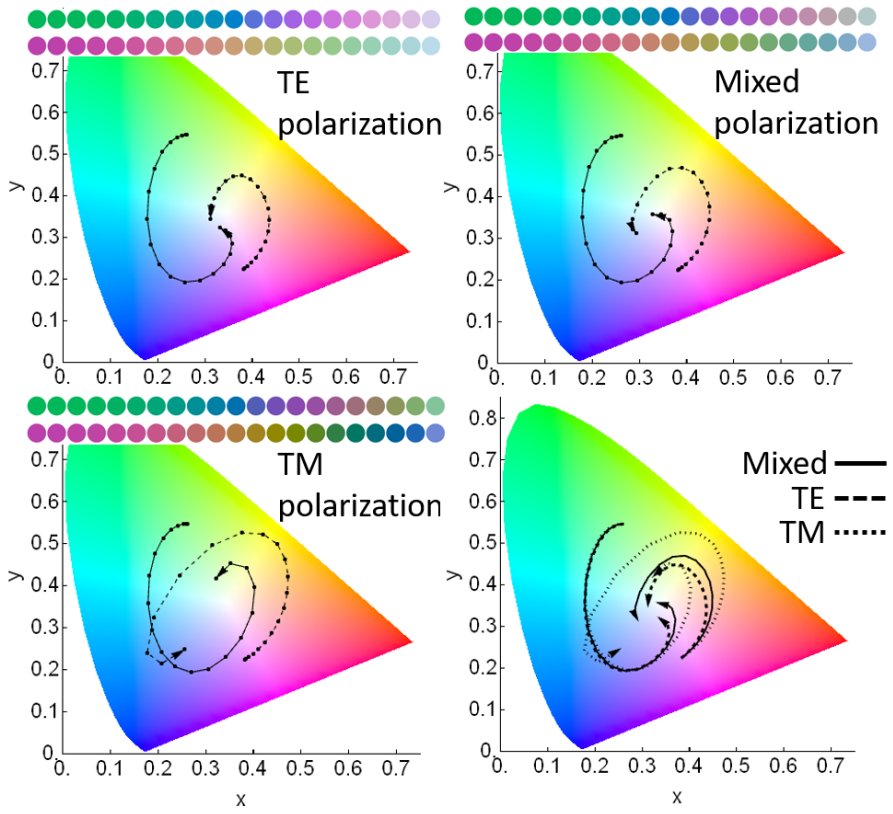

Figure 30: Effect of changing angle for two points of the FP structure with a C absorber layer, for the angular range 0-85°. In the upper left, half and half TE/TM polarization is shown. The others are as labeled, with the same points for each. For each, the colors inset are the colors produced by the structure as it goes through the angular range.

In Fig. 30, iridescence for two arbitrary pairs of points (($t_{Abs} = 35nm, t_{AAO} = 450nm$) and ($t_{Abs} = 55nm, t_{AAO} = 350nm$)) are shown for several variations of polarizations, as the viewing angle is increased from 0 to 85°. The absorber used in this theoretical structure is C. In the upper left, the iridescence of mixed polarization (half TE, half TM polarized light) is shown, as well as the colors (inset) produced by these two points for increasing angle (from left to right). The behavior of this mixed polarization light is not as physically clear, but it's important to illustrate because most light that we see is unpolarized, so this is a good approximation to the real behavior of the structure. It can be seen that the EP is relatively constant, but the hue is changing drastically with angle. Next, the TE and TM polarizations are shown separately for the same two points and angular range. The have the same rough trends (changing hue in the same way, but TM polarization appears to be much more sensitive to angle than TE. This intuitively makes sense: recall that one of the boundary conditions like going through an interface must obey is that the parallel component of E and the perpendicular component of D must be continuous. For TE polarization, both of these are constant with angle, but with TM polarization, as angle increases, the E field becomes less parallel (to the interface) and more perpendicular, so more change is expected. Lastly, all three are shown together. This is to illustrate that, all said and done, the two polarizations and their mixture aren't actually very different from each other. The paths illustrated in this figure go from 0-85°, which is a very large range, and they can be seen to really only diverge near the end for one point, and not significantly for the other. Therefore, unless otherwise specified in this paper, we will use the half and half mixture of polarizations for angular calculations, knowing that they represent the real behavior well and aren't a huge difference.

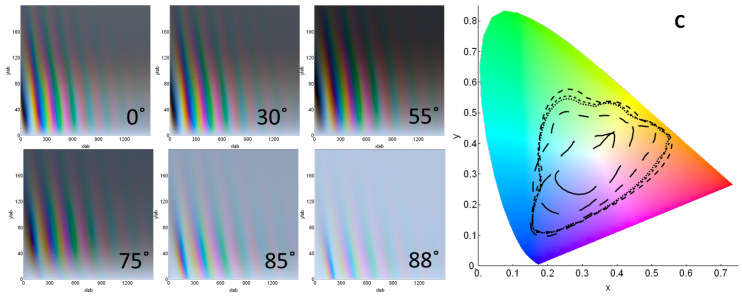

Figure 31: Gamut and CCPs for $\theta_{incidence}$ 0-88° with a C absorber layer.

2.12.3 Gamuts vs incidence angle

The last aspect of iridescence to look at for our typical FP structure is the full gamut as a function of viewing angle. These are plotted in Figures 31-35 for the absorber materials C, Ge, Ni, Fe, Cu, each for the angular range 0-88° (from the normal). In addition, for each material, their CCPs for each angle are shown. Interestingly, we can see that the gamuts are nearly independent of angle, until about 75° for most of them. Initially, this might appear strange, but it is actually analogous to the behavior previously presented for the gamuts produced by FP cavities with AAO of different porosities; recall that, for a FP structure with a given porosity (i.e., n_{AAO}) and t_{AAO}, a structure with a different porosity could have the same optical path length by having the right t_{AAO}. Similarly here, the incidence angle changes the phase thickness for each layer, but mostly the same optical path lengths (and therefore, colors) can be produced by selecting a t_{AAO} that accounts for this change. However, boundary matching is angle dependent and for the very extreme angles, the reflectance at the first (air/absorber) interface increases quickly, not giving much incident light a chance to get into the structure at all. Thus, all wavelengths are nearly completely reflected, which gives rise to the washed out gray color we see for the higher angles, as the gamuts start to shrink to the white point.

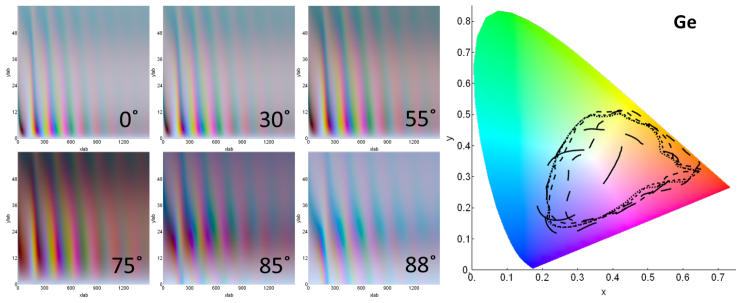

Figure 32: Gamut and CCPs for $\theta_{incidence}$ 0-88° with a Ge absorber layer.

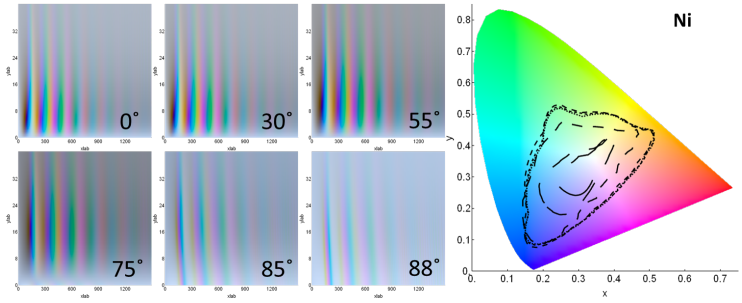

Figure 33: Gamut and CCPs for $\theta_{incidence}$ 0-88° with a Ni absorber layer.

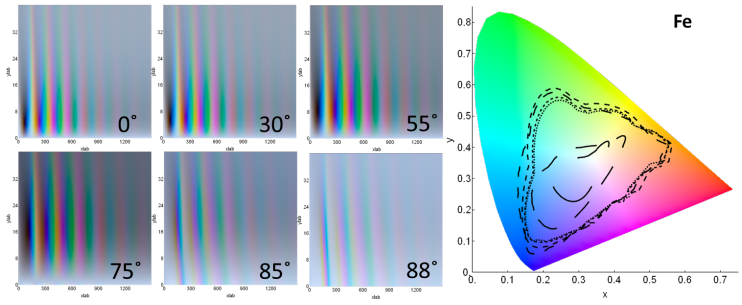

Figure 34: Gamut and CCPs for $\theta_{incidence}$ 0-88° with a Fe absorber layer.

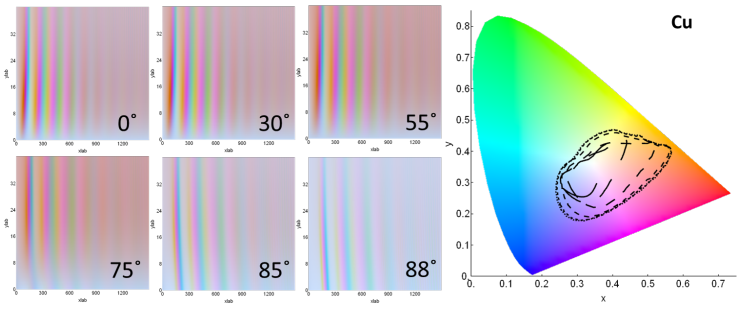

Figure 35: Gamut and CCPs for $\theta_{incidence}$ 0-88° with a Cu absorber layer.

2.12.4 Strategies for mitigating iridescence

High index cavity material A very clever method for mitigating iridescence in the FP cavity structure was reported on by several different groups in the past few years [103, 53]. The most important difference of their structure is that a larger (than AAO) index dielectric was used as the cavity material. This clever chosen combination of materials gives rise to a mechanism in which, as the viewing angle increases, the decrease in phase shift due to the changing propagation length is canceled out by a increase in phase shift from the reflections at the film interfaces, giving rise to a relatively constant (with angle) phase, maintaining the color.

Specific color stability One method for mitigating iridescence is as follows. Though it does provide a possible solution, its scope is certainly limited. The strategy here is to simply, for a given absorber material, produce its gamuts over a large angular range (as shown previously), but then intentionally search for points that minimally change their hue and EP. To do this, a vector is produced of the form $\{\{t_{AAO}, t_{Abs}\}, \{\{X,Y,Z\}_{\theta_1}, \{X,Y,Z\}_{\theta_2}, ...\}\}$ from the gamuts at the different angles combined, where $\{X,Y,Z\}_{\theta_i}$ are the X,Y,Z tristimulus values produced from viewing the FP cavity structure with $\{t_{AAO}, t_{Abs}\}$ at angle θ_i. From here, a tolerance for how much we let the XYZ values change between

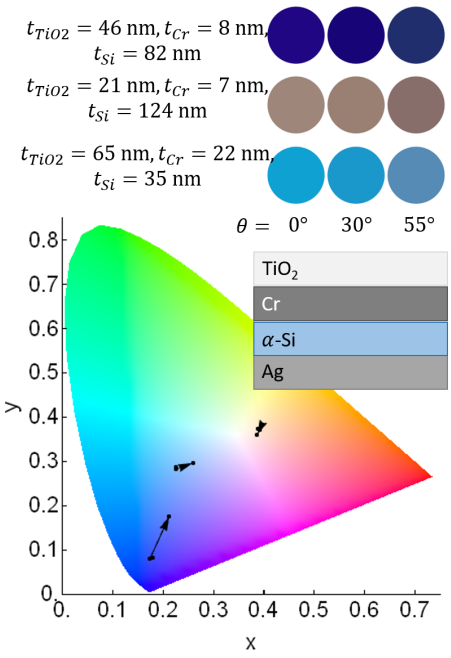

Figure 36: Three selected sets of parameters for the angle-robust FP cavity structure reported by [103]. The color changes a little, but the hue is nearly completely independent of angle.

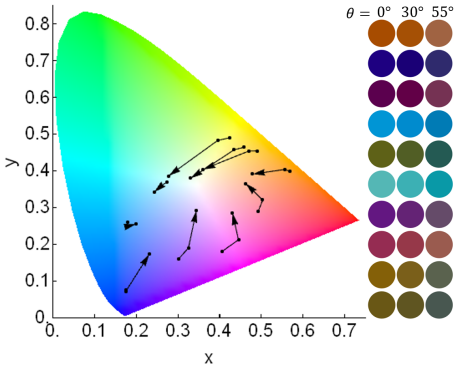

Figure 37: A series of other points for the same high index FP cavity structure as in Fig. 36. Note that even points that are moving a decent amount in the xy space change mostly in EP and minimally in hue.

angles is specified (normally by first converting the XYZ values to hue and EP), and only vectors that satisfy this tolerance are selected. The $\{t_{AAO}, t_{Abs}\}$ that have this property will be referred to as angle robust colors (ARCs).

The results of doing this procedure for several different absorber materials are shown in Fig. 38. For each absorber, several of the best ARCs (highest EP) are shown, as well as their reflectance spectra, to get a better sense of what's occurring physically to give rise to this. Once again, it can be seen to be due to a coincidental intersection between the FSR/FWHM and human cone spectra as discussed previously. For the ARCs, the spectra at different angles *are* changing, but in such a way that it doesn't change their convolution with the color matching functions very much.

2.13 Potential expansions of the FP structure and structural coloration and connections to resistive switching

At this point it is prudent to look forward at potential extensions of the formulation presented here. Since many of the structural coloration analysis techniques demonstrated here apply more generally than to just the simplest FP cavity structure, we can ask how the FP cavity structure could be expanded by the addition of the resistive switching (RS) platform to it, which is discussed and explored extensively in the rest of this document. Since the details of it are explained thoroughly in the next section, we briefly mention here that it is a system in which one material is actively made to move inside of another material. In addition, by fortunate coincidence, the RS structure investigated already has a very similar structure (metal/insulator/metal sandwich).

The most straightforward way RS could be combined with structural coloration in the FP platform would be simply taking advantage of the movement of

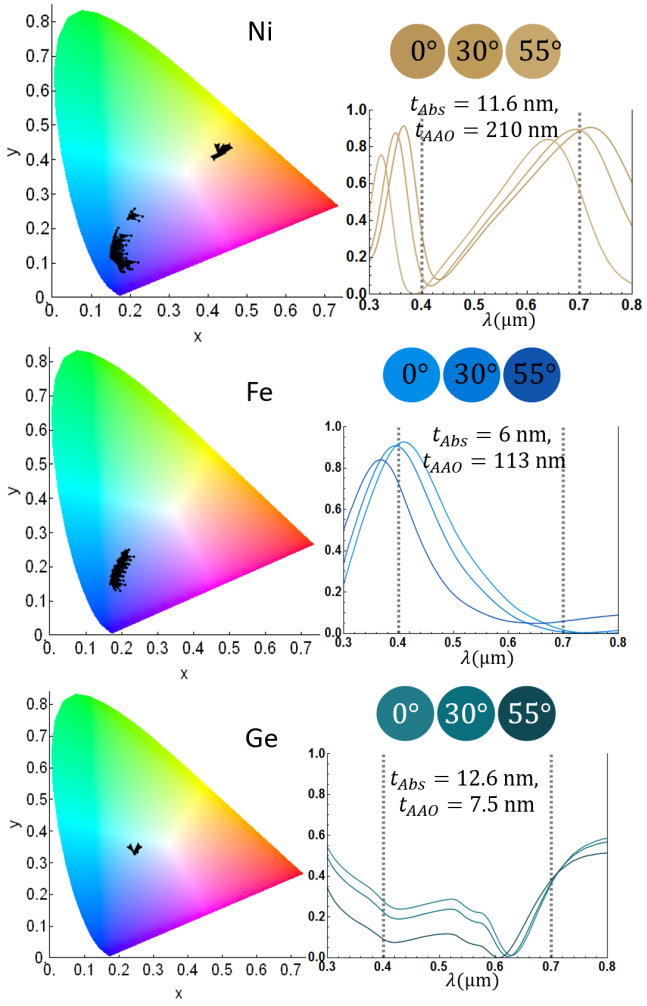

Figure 38: Specific angle robust points for three absorber materials. For these materials, the physical parameters inset in the R plots give rise to colors that happen to be mostly independent with viewing angle.

absorbing material. If RS moves an absorbing material (say, Ag) through a typically non-absorbing medium (AAO, for example), it would cause this medium to become a little absorbing, changing the FP cavity's optical response. This is essentially the mechanism used in [20], where the EM of Ag was used to block light.

However, this immediately runs into several objections. As discussed earlier, most RS is "filamentary", meaning that the conduction is dominated by a single (or very small number) of conducting filaments, and is mostly independent of the electrode area. Therefore, it can't cover a lot of area. This was not a problem for the research of [20] as they simply had to block a very thin waveguide with the filament, but a small number of filaments would have no macroscopically visible effect in this respect.

There are several paths that could be considered a solution to this problem. The first is that, in fact, there are reports of "homogeneous" RS (as opposed to filamentary RS), in which the EM of material is area-dependent [60, 39]. However, the materials used tend to be more exotic than simple metals and oxides, and the mechanisms proposed for its action are even less well agreed upon than for filamentary RS [73]. However, it still presents as a possibility.

Another mechanism for which RS could change the properties of a FP cavity could be by using filamentary RS, but in such a way that it effectively becomes homogeneous RS. Expounding, consider that the foremost cause of filamentary RS is that a positive feedback loop is created, making it impossible for any other filaments to form. On the other hand, we know that the main hypothesis for unipolar action is a "burning out" or "rupturing" of the filament [15] due to too much current, which is certainly a negative feedback effect. Depending on how the filament ruptures, it could cause that branch of the filament to not be able to continue growing. So if very precise characterization and control were used

for a system capable of unipolar RS, the repeated unipolar mechanism could give rise to bulk movement, though it would come with its own deficits, such as potential lack of repeatability. Similarly, since the proposed mechanism for the burning out is simply local Joule heating [104], that could be used as well to locally change the dielectric medium such that the EM material can no longer go through that area; one option for this would be changing the dielectric at that spot to crystalline, since many RS systems rely on an amorphous dielectric [98]. In fact, this very effect has been shown to occur in Si [44], with amorphous Si (a common RS dielectric for an Ag electrode) being more quickly turned into crystalline Si with the application of a modest E field and heating. Considering the local heating effect of unipolar resetting, this is even more plausible.

However, any of the above methods still have to grapple with another inherent problem. Because formation speed is exponentially dependent on the driving electric field magnitude [105], which is dependent on the applied bias and interelectrode distance, this typically dictates an interelectrode distance on the order of ∼100nm [98]. However, this will have the effect of constraining the sample geometry. The simplest method (assuming homogeneous RS has been achieved) would be simply to use a FP cavity, with the reflective substrate as one electrode and the absorber layer as the other. This has the fortunate coincidence that our FP cavities for the visible range already have widths of typically ∼100nm. On the other hand, making the FP cavity medium lossy may have minimal effect, in contrast to changing the absorber layer itself. An ideal geometry for the application of RS to the FP structure would be for the EM to occur *across* the top (normally absorbing) layer, such that it is by default a non-lossy dielectric, but turned into an absorber by the RS operation. However, this has the above limitation of interelectrode distance, unless nanofabrication were used to create a massive array of interdigitated electrodes.

Another mechanism by which the gains of RS research could be employed for structural coloration is by actively changing the carrier concentration of the absorber layer, thus changing its refractive index. The change in refractive indices with carrier concentration has been shown to occur for several semiconductor materials, such as InP, GaAs, and InGaAsP [9]. Several simultaneous mechanisms are hypothesized to be occur [9, 56], such as the Burnstein-Moss Effect [87] (in which the band gap absorption edge is increased as a result of carriers filling the conduction band), band gap shrinkage, and free carrier absorption. The same effect has been shown to occur in more realistically usable materials for this project, such as Ge and Si [80], though it should be noted that doped versions of these elements may need to be used to get their baseline carrier concentrations into the right range; Ge has an intrinsic carrier concentration in the $10^{13}/\text{cm}^3$ range, while the paper cited looks at the effect when the carrier concentration is doped into the $10^{18}/\text{cm}^3$ range. Then, assuming the refractive indices of a specifically chosen absorber layer (such as Ge, which this project has used extensively as the absorber layer) can be altered by changing the carrier concentration, this could be attained by actively forming and unforming the RS filament, either using it to directly electrically contact the absorber material and sink/source it with free carriers, or potentially act as a gate for it. Additionally, because our absorbers are typically of very small widths anyway ($\sim$10nm), they could already behave as a two dimensional electron gas, a quantum well in one dimension. If that were the case, they would be susceptible to the Quantum-Confined Stark Effect [57], which is used for optical modulators and has already been demonstrated in a Ge quantum well [50].

Lastly, we note one other potential connection of RS to the structural coloration FP platform. When the filament is nearly completely formed (i.e., an Ohmic connection), it is functionally a very sharp point. This gives rise to mas-

sive field enhancement for relatively meager applied biases, with enhancements of >1,000 already experimentally demonstrated [97]. So, the RS system gives us access to a tunable, easily modulated, extremely high E field source. At the same time, there are many reported examples of optically induced changes in refractive index. Ideally, we could create a system that would be inert to normal intensities of light (since, after all, the immediate goal is to use the structure to produce coloration), but would undergo some change with the application of the strong E field produced by a filament tip. In [5], a field enhanced tip was used to alter a photosensitive-organic containing film. A nonlinear polymer was shown in [77] to give a significant refractive index change with the application of 632nm incident light. Again, in [44], a strong E field and heating was shown to more easily transform amorphous Si into crystalline Si, which has slightly different optical properties. While several of the reported processes appear to be irreversible, it was reported in [23] that the same crystallization of amorphous semiconductors could be *reversed*, indicating a promising route.

3 Resistive switching and optical rectification in AAO

3.1 Resistive switching basics

Resistive switching (RS) is a field that has existed for a long time in various forms [35] and holds much promise for practical applications, but thus far has had only minimal use outside of the research setting. The most general definition for RS is a system in which the resistance between two points (usually electrodes on either side of a dielectric) can be changed by the application of an electric field or current. Typically, this is done with the goal of having well defined low resistance states (LRS) and high resistance states (HRS) that can be switched between quickly and repeatedly by application of a bias, and have the system stay in the state it was last set to without further applied bias. This is most frequently touted as a new path towards non-volatile resistive random access memory (RRAM) [1, 99, 41]. It has also been demonstrated to be a working example of a memristor, the fourth fundamental electronic component (after resistors, capacitors, and inductors) proposed by Leon Chua in 1971 [17]. A good overview of RS can be seen in [99].

So the motivation for such a platform is clear. Typical consumer dynamic RAM in a computer works by using a transistor to charge and discharge a capacitor that holds the memory state for that bit. It is successfully used on a large scale right now, but comes with downsides that will limit it in the long run. These include, but are not limited to, low endurance, low write speed, and high voltages needed to write memory states [99].

By what mechanism does RS usually work, though? There are a few different platforms, but by far the most common is a "sandwich" structure with one electrode, a dielectric, and another electrode. There are several variations to

this as well, but usually one of the electrodes (the "active electrode") is made of a material that is slightly soluble in the dielectric material (whether material A is soluble in material B just comes down to specific chemistry, essentially), while the other electrode isn't (the "inert electrode"). When a bias is applied across the electrodes, atoms from the soluble electrode ionize, migrate inside the dielectric towards the other electrode, and when they reach it, are deposited on the inert electrode and turned back into their neutral form. In doing this, material slowly bridges the gap between the electrodes, diminishing the distance. Because, for a given applied bias across the electrodes, the electric field is inversely proportional to the distance between the electrodes, the E field increases as material gets deposited, and the E field tends to get concentrated where material has already deposited, causing future material to deposit there as well. This causes it to have a "stringy" shape and it is commonly referred to as a "filament", which we do here. It should be clear from this that the formation is a positive feedback loop (barring other effects, which we will see).

It should be pointed out that the mechanism was intentionally left vague in order to be more general. For example, the simplest version of the migration mechanism (and the one used in this project) is shown in Fig. 39. A macroscopic schematic of the structure (discussed below) is shown in the upper left. In this scenario, the active electrode is a material like silver (Ag). During electromigration, a positive bias is applied to the Ag electrode with respect to the inert electrode, a neutral Ag atom gets ionized to Ag^+, and the electron from this Ag atom goes to the applied positive bias source. Now, charged and driven by the E field, the Ag^+ ion migrates, reaches the inert electrode, takes an electron from it, and is deposited as a neutral Ag atom (Fig. 39, upper right). However, there are other material platforms; we wanted to leave open the possibility of materials that are ionized with a negative bias, and there are

also more complicated ionic exchanges, most commonly with TiO_2 [51].

When the gap between the electrodes is similar to what it started (typically a few hundred nm), the current is very small, only due to the current from the migrating ions (electromigration), and perhaps some hopping between nearby ions. This is the HRS. When the distance is bridged more, if there is still a small gap between the filament and AE, direct tunneling can occur (Fig. 39, bottom left). Finally, when enough material is deposited, the filament forms a continuous, Ohmic junction from one electrode to the other. This is the LRS and the process for "setting" the system from the HRS to LRS has been described.

At this point we must explain the "resetting" action from the LRS to HRS. Unfortunately, it is more complicated than the setting action in that there are two main types, which we now explain.

3.1.1 Bipolar operation

Bipolar operation is actually the simpler of the two types of resetting. In bipolar operation, to reset a device in the LRS, if a positive bias was applied to the active electrode to set it (which we will be assuming is the case in this project), a negative bias must be applied to reset it (bipolar, due to the voltages of opposite sign). This makes intuitive sense; it's literally the same process at the set procedure (neutral Ag oxidized, Ag^+ ion travels, Ag^+ ion is reduced back to neutral Ag), but reversed. The "deposited" material ending up back on (or closer to) the active electrode.

However, the process can be difficult or impossible to occur depending on "how set" the filament was; if enough material of the active electrode was transported during a fast and violent set procedure and the formed filament bridges the electrodes strongly, when a reverse bias is applied, there will be very little E field contributing to electromigration. This could be called a "burned in" device. It should also be pointed out that it is unlikely that, if the device is reset to the

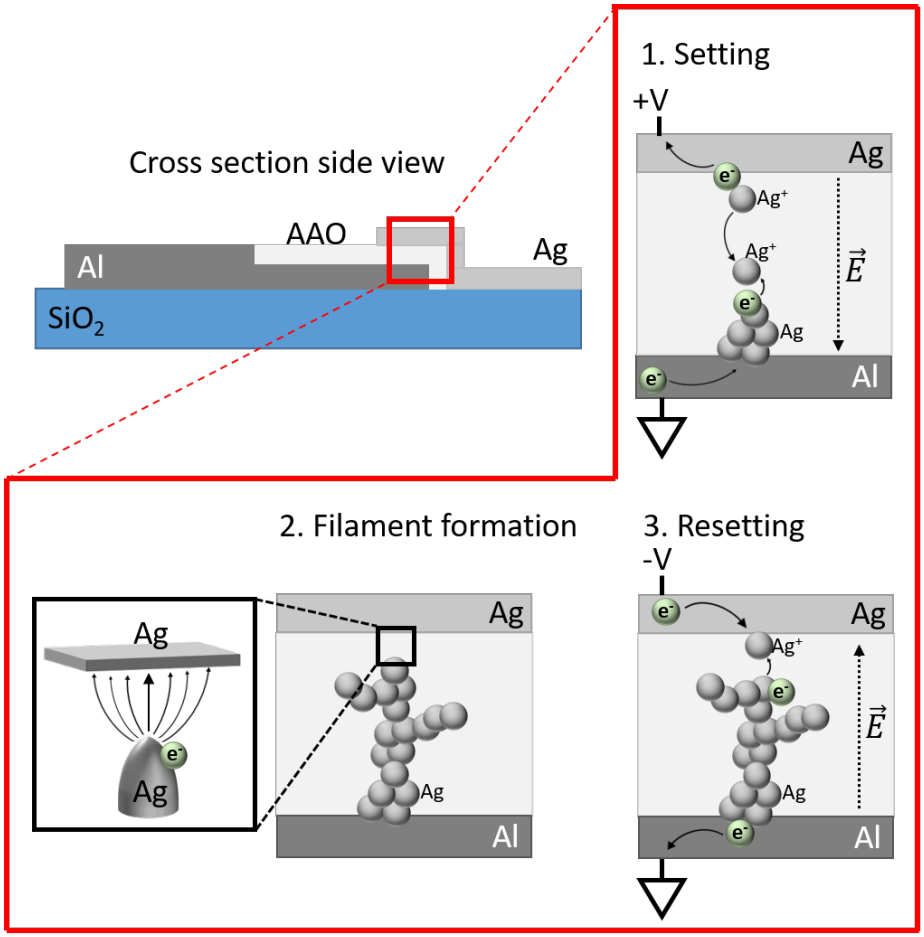

Figure 39: Process occurring at the nanoscale for OR. A side view of the macroscopic structure is shown above. Inset are three steps: 1) Setting the filament by application of a positive bias (with respect to the passive electrode), 2) Formation of a tunneling junction after some period, and 3) Resetting by application of a negative bias (bipolar operation).

HRS, it will go back to its initial form with the electrodes completely separate with only dielectric in between; it is much more likely that once enough material has electromigrated to break the filament connection and put the system in the HRS, further electromigration will be greatly diminished, leaving behind a disconnected but still mostly formed filament.

3.1.2 Unipolar operation

The other reset mechanism is known as unipolar, because voltages of the same sign are used to both set and reset the device. While bipolar operation is intuitive enough, unipolar must be explained and it should be pointed out that there is not total consensus on its physical mechanism [83] or that all data purporting to demonstrate it is truly showing this effect [99].

The typical behavior and use is as follows. To set it, the bias is increased up to some threshold, V_{set}, with a current compliance I_{set}. If an IV sweep is performed from 0 to V_{set} and back to 0, hysteresis can be seen because on the trip back, the device is in the LRS and the current is therefore higher. However, to reset it, a similar sweep is performed, but with a higher current compliance (and perhaps a higher bias V_{reset}). This higher compliance allows more current to flow, since it is already in the LRS, and the device can be seen to shortly go back to the HRS.

The mechanism for this is usually explained as the filament "burning out" due to too high a current density running through it. This may suffice as a handwavy solution, but this tends to raise more questions than it answers. For example, what does it mean to "burn out"? For a filament in the environment, the answer is simply that it gets oxidized into a different material. Similarly, for a incandescent light bulb with the filament in inert gas, it may not get oxidized, but it melts and the material is either exploded [51] or at least the retracts on itself in either direction due to cohesion, destroying the connection path.

However, in our structure, the filament material is ostensibly both not exposed to O_2 (by dint of both being under vacuum, and inside the dielectric), and seems unlikely to have the problem of the filament material going anywhere; it should still be confined to the dielectric.

3.1.3 Applications to optics

RS behavior has been known about for a long time and reported on thoroughly [99, 66, 68, 19], but has been reported on almost entirely in the context of electronic applications. However, it also offers interesting possibilities in the realm of optics. So far, only a few papers have used it for optical applications. The closest application to the one reported on in this work is [74], in which the behavior of the antenna formed by the filament was optically probed; however, they did not seem to be looking at optical rectification, and did not report on any current occurring as a result of incident light. In two cases, it was used essentially as a waveguide blocking system: In [20], researchers created a small waveguide that light could be coupled into and read out of. The waveguide was the dielectric medium of the RS structure in this case, and when the filament was formed, the waveguide was small enough that the filament blocked an appreciable percent of it and this reduction in the light getting through the waveguide could be measured. In another paper [37], a similar effect was demonstrated. Other tangentially related applications within optics are [59], in which an incident laser was used to alter the switching time of the RS system.

3.2 Optical rectification

3.2.1 Background and motivation

A promising use of RS that is more feasible involves its use as a platform for optical rectification (OR). While the previous optical applications of RS are

easily understood, OR requires some background. We can first explain the basics of OR in a high level way. More generally, rectification, for example in the typical electronic component sense, means only allowing current flow in one direction. Therefore, if the source is in alternating directions (for example, an AC signal in a circuit), a rectifier (such as a diode) converts this into a signal that is only in one direction (even if its magnitude in that direction is changing with time). This is the basic idea behind rectification.

The mechanism by which this happens with OR is different, but not entirely dissimilar to an electronic diode. This is partly because the device actually *is* a diode, but not a semiconductor diode that works via the well known mechanism involving a pn junction. Instead, an OR diode is typically one formed via some sort of tunneling junction. Additionally, the alternating potential that is to be rectified is not due to a potential applied elsewhere from the circuit (as in the typical diode component example above), it is due to the electric field from incident light, across the tunneling junction. So in either case, an alternating potential causes a net asymmetric flow of current in one direction, but the origins of that potential are different.

Additionally, some other key differences should be pointed out. Most traditional electronic circuits utilizing pn junction diodes have a bandwidth limit; that is, due to the mechanism by which they rectify alternating potential, they can only rectify light of some frequency up to a certain limit. This is due to the transit time of charges in the pn junction limiting their operation to the GHz range [75]; in general the idea governing frequency limits in this context is that, if the E field is switching directions faster than the carriers that give rise to current can respond, they will be acted upon in both directions of the E field, which will mostly cancel the effect (there's no hard cutoff in frequency but it will decrease sharply past a certain point).

Derivation of OR equations It is instructional to first look at the effects of rectification for a simple rectifying device with a nonlinear IV curve, such as a traditional pn junction diode, for a low frequency potential. We will consider an alternating bias on top of a DC bias as well, represented by $V(t) = V_{dc} + V_{ac}cos(\omega t)$. We can Taylor expand the IV curve around the DC bias point V_{dc}. If we do so, we get:

$$I(V) \approx I(V_{dc}) + \frac{1}{4}\frac{\partial^2 I}{\partial V^2}\bigg|_{V_{dc}} V_{ac}^2 + \frac{\partial I}{\partial V}\bigg|_{V_{dc}} V_{ac}cos(\omega t) - \frac{1}{4}\frac{\partial^2 I}{\partial V^2}\bigg|_{V_{dc}} V_{ac}^2 cos(2\omega t) + ...$$

Grouping the terms together, we see a couple important points. First, we see an addition (the second term) of DC current to the DC current we expected for an application of only V_{dc}, based on the IV curve. This is the rectified signal. Importantly, we can see that it's proportional to the value of d^2I/dV^2, the 2nd derivative of the IV curve with respect to V, evaluated at V_{dc}. In more intuitive terms, this is the nonlinearity at V_{dc}. This will be important later. Secondly, we see a term that has a frequency of 2ω. This means that our initial AC bias has given rise to a signal at twice its frequency! This has very useful applications, known as Second Harmonic Generation. This illustrates the behavior for a traditional nonlinear device.

Now we must present what happens for a similar situation, but with a tunneling junction rather than any nonlinear electronic device, and incident light rather than a simple applied alternating bias (on top of a DC offset bias). We will still consider applying the DC offset bias, since that's something we would be able to do with our device.

This was initially presented by Tien and Gordon [91] and the derivation is as follows. We consider a tunneling device composed of a metal, insulator (usually an oxide), metal sandwich. We assume that the energy band diagram for the

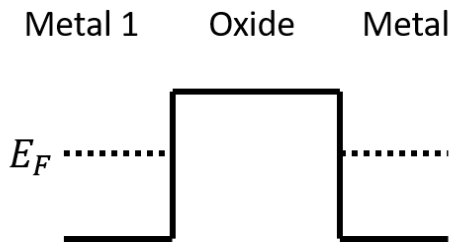

Figure 40: Example band diagram for MOM RS system.

device has already been figured out, by looking at the relevant materials' work functions, Fermi Levels, etc. A typical band diagram might look like what is shown in Fig. 40. For simplicity at this point, we consider a MOM structure that's symmetric, that is, at zero applied bias, the barrier is flat.

We now consider the case of an incident light beam on a tunneling junction, oriented and polarized such that the E field is perpendicular to the layers of the stack. We assume that the wave function in one of the metals is already solved and represented by a spatial part and a time dependent part:

$$\psi(x, y, z, t) = f(x, y, z)e^{-iEt/\hbar}$$

For the regular Hamilton H_0 with no incident light. If the E field couples to the electrons in the metal films, this will give rise to an alternating potential difference across it:

$$E(t) = E_0 cos(\omega t) \rightarrow V(t) = V_0 cos(\omega t)$$

This will create a new time dependent in the Hamiltonian, $H = H_0 + V_0 cos(\omega t)$. However, because the perturbing term is only time, not spatially, dependent, it will only affect the time dependence of the original wave function. The solution for this perturbed Hamilton can be written:

$$\psi(x,y,z,t) = f(x,y,z)e^{-iEt/\hbar}\sum_{n=-\infty}^{n=\infty} B_n e^{-in\omega t}$$

Plugging this and the perturbed Hamiltonian into the Schrödinger equation gives a recursion formula that leads to the value of these coefficients:

$$B_n = J_n(eV_0/\hbar\omega) = J_n(\alpha)$$

The solution for is the nth order Bessel function of the first kind. Now, plugging these back into the above equation of the modified wave function gives its solved version. We have introduced the variable $\alpha = eV_0/\hbar\omega$ which will be important later. The implications are immediate: the wave function is "split" from its original energy level E into a wave function with the same spatial dependence but a distribution of energies at $E, E \pm \hbar\omega, E \pm 2\hbar\omega$, etc, where $\hbar\omega$ is the incident photon energy. From this, we can look at how the density of states is modified as well:

$$\rho'(E) = \sum_{n=-\infty}^{n=\infty} \rho(E + n\hbar\omega)J_n^2(\alpha)$$

Now we can use the equation for tunneling current in a dielectric layer between two metal films and plug in this modified density of states:

$$I(V) \propto \int_{-\infty}^{\infty} (f(E-eV) - f(E)) \times \rho_L(E-eV)\rho_R(E)dE$$

$$\rightarrow I(V) \propto \sum_{n=-\infty}^{n=\infty} J_n^2(\alpha) \int_{-\infty}^{\infty} (f(E-eV) - f(E+n\hbar\omega)) \times \rho_L(E-eV)\rho_R(E+n\hbar\omega)dE$$

This is easily separated and reduced to an intuitive explanation. As discussed before, if we consider the effect of the incident photon E field on the junction to be splitting the wave function into weighted components separated by multiples of the photon energy, the resultant tunneling current equation is just the sum

of all those new energy levels tunneling separately as if the Fermi Level was at that energy for each one.

This explains the current expected but it is worth looking at some simplifying assumptions for specific cases. The parameter α that was introduced is important for further discussions. Taking a look at it, it is the potential difference imposed across the junction as a result of the incident light divided by the photon energy of that light. However, how is the potential difference determined? It is certainly dependent on the E field of the incident light, but is also largely dependent on the coupling of this E field to the two electrodes of the junction. This can certainly be smaller than the incident E field, but it can also be larger due to field enhancement at sharp tips. Therefore, it is not trivial to calculate V_0 and by extension α.

Two common simplifying assumptions are often made at this point [92]. The first is to expand the term for the current to first order in V_0 and solve for the difference in current between the tunneling current above and the DC current (without incident light):

$$\Delta I_{dc}(V) = \frac{1}{4} V_0^2 \frac{I_{dc}(V + \hbar\omega/e) - 2I_{dc}(V) + I_{dc}(V - \hbar\omega/e)}{(\hbar\omega/e)^2}$$

At this point it begins to look familiar. This is a finite difference calculation of the 2nd derivative of the DC current, where the step size is the photon energy! Note that what occurred here was to discard the terms besides $n = -1, 0, +1$, meaning that we aren't considering the case of higher energy levels. This is not an unfair assumption, because the amplitude of these terms drops off very quickly.

The second common assumption that is used to simplify the math is to say $eV_0 \gg \hbar\omega$, meaning that $\alpha \gg 1$, or in other words, there is strong coupling of the E field compared to the photon energy. For the case that Tien and Gordon

originally investigated, $\hbar\omega = 0.16$eV, so it is very easy for eV_0 to be larger than this. When this is applied to the above equation, we see that it is equivalent to taking $\hbar\omega/e \to 0$, which gives us the continuous second derivative of the DC tunneling current, as shown for the classical case previously! This is an incredible result. One implication is that a relatively robust way to check for the occurrence of OR is to compare the second derivative of the IV curve (either calculated, or more definitely, measured using a 2nd derivative measurement, as in [97] to the photocurrent from an incident laser.

However, this assumption is not necessarily valid. As discussed, $\hbar\omega$ for microwaves is very small so it is very possibly valid, but for the visible range, $\hbar\omega > 1.8$eV, which is a very large energy, relatively. There is a good chance the IV curve has curvature on this scale. In that case, the finite difference equation must be used. However, this runs into problems. First of all, because it requires data from $\pm\hbar\omega/e$, that is a minimum voltage range of 3.6eV for red light, to even get a single data point. Depending on the system, it may be difficult to measure that range without either burning out the sample, or running into other effects where tunneling is no longer the dominant current mechanism, which would render this analysis moot.

3.3 Our structure and fabrication

We investigated several platforms in this project, each with their own strengths and weaknesses. Each is based on a platform of the same instrumental materials; that is, Ag active electrode, AAO dielectric, and Al inert electrode. However, they differ in their geometry and preparation. We present the pitfalls involved in fabricating a structure that can work for our needs and the final structure that solves this problem.

3.3.1 Fabrication pitfalls

The planar or "sandwich" structure is by far the most common (note that although every type of RS device is a metal-dielectric-metal sandwich, this one is most straightforwardly so). It is simply a film of one electrode material, a film of the dielectric directly on top of that, and then a film of the other electrode on top of that.

This structure geometry is very simple to make; in fact, it can potentially be fabricated in a single evaporation step, although it is more desirable to split it up into several, as explained in the following. In Fig. 41 (a), we present what the device would look like if done with a single evaporation session (but multiple evaporations within this session). First we must note that, because we must electrically contact both electrodes, the films cannot just be evaporated over the whole substrate; if we did, the bottom electrode is completely covered by the subsequently deposited films and there would be no way of contacting it. So, a typical strategy is to first deposit the bottom electrode over a larger area, and then deposit the subsequent layers over a smaller area, so the bottom electrode is exposed (using a shadow mask or photolithography). This can be seen in Fig. 41 (b).

However, it is still a poor method to deposit the other films in the same session. There are two major pitfalls.

The first involves the contact of the top electrode. If the RS structure, from the bottom up, is electrode 1, dielectric, electrode 2, contacting electrode 1 is simple enough (just physically touch the bottom electrode where it is exposed), but contacting electrode 2 presents a problem. The dielectric is typically thin enough (100nm range) that contacting electrode 2 physically (with a flat alligator clip or even a micromanipulator probe) presents a strong likelihood that it would pierce the dielectric, causing an ohmic path between electrodes, as shown

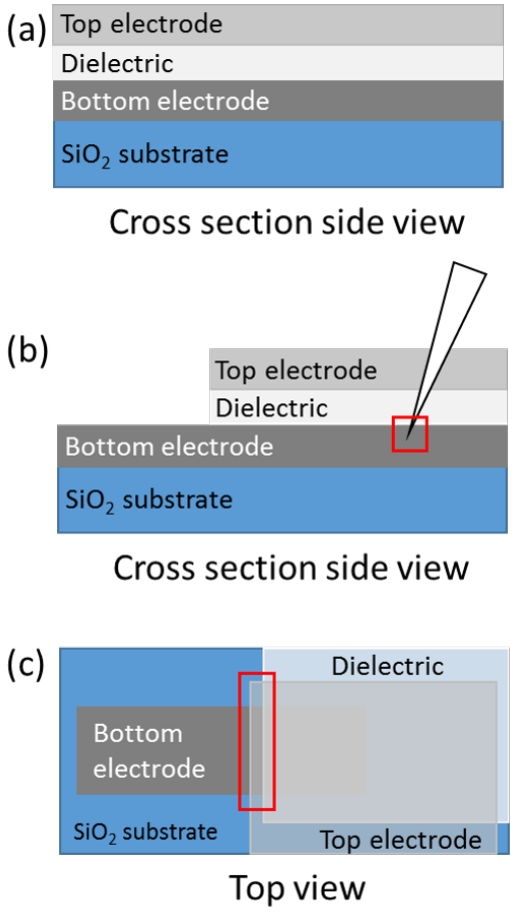

Figure 41: Potential fabrication pitfalls of the RS platform.

in Fig. 41 (b). This can be can be mitigated by a simple solution. Instead of the bottom electrode covering the whole substrate and the top electrode only being on top of the stack, the bottom electrode only covers a section of the substrate and the top electrode is evaporated over an area that goes down the side of the stack, always on top of the dielectric film, to the initial substrate. At that point it can be contacted directly without fear of "puncturing" through to anything. This is illustrated in Fig. 41 (c). Note however, that this requires also a change in the evaporation step of the dielectric as well, as it must now overlap and cover more area than the first electrode, so the second electrode doesn't contact the first electrode while going down the side.

The second possible pitfall must be avoided because we are investigating such low currents that any alternative path for current aside from electromigration will dominate and cause electromigration to either not be measurable in comparison, or not occur at all. So if the dielectric and top electrode films are evaporated in the same session (even if the first pitfall is avoided by recessing the bottom electrode), the dielectric and top electrodes will have the same boundary, defined by a shadow mask or other. However, because of fluctuations, wetting, migration during deposition, and the different positions of the evaporating sources, the boundaries for the different films won't be exactly the same. Thus, there is a non-negligible chance that the top electrode will go over the side of the dielectric, making contact with either the inert electrode or Au contact, as shown in Fig. 41 (c). This too can be mitigated by sequentially depositing the first electrode, the dielectric, and then the top electrode, moving the film boundaries by shadow masking appropriately.

3.3.2 Planar structure

To avoid all the pitfalls above, we created the structure with the fabrication procedure illustrated in Fig. 42. There are two details about the structure that

have nothing to do with the pitfalls above, however. First, it can be seen in Fig. 42 that the bottom electrode is made to be a narrow ($\sim$ 30 μm) strip. This was done because we want to optically probe the filamentation that's occurring. However, as mentioned previously, filamentation is typically a positive feedback loop, meaning that the filament usually occurs at a single point, as opposed to being a wide scale phenomenon (the behavior is mostly independent of the electrode areas). This means that if we need to know where the filament is to directly interact with it, it will be nearly impossible if the filament (covering an area of $(10\text{nm})^2$) is somewhere in the macroscopic area of the overlap of the RS film stack. Therefore, making the bottom electrode a thin strip (via photolithography) localizes the filament to a spot small enough for the laser spot to cover. Note that only one electrode needs to be made small, because filamentation can only occur somewhere in the region where there is a full RS stack of one electrode, the dielectric, and the other electrode.

The other difference is that of the formation of the dielectric layer. The most common choice in the literature for oxide-dielectric RS platforms is to use an evaporated, sputtered, or CVD film for the oxide layer. However, another method involves using anodization to create the dielectric layer. RS has already been demonstrated in AAO in a series of papers by *Takase et al* [26, 64, 88, 63] as well as some papers by Hickmott [34, 33, 32], although the ones by Hickmott are more suspect because he mentions that the mechanism seemed to not work unless the materials were evaporated in the presence of "rubber" (this suggests somehow organic impurities from the rubber are making their way into the AAO, which makes the relevancy of the results dubious with respect to the mechanism we're looking at).

For this method (Fig. 42), we first deposit a layer of Al, which will be the inert electrode. The thickness of the Al is enough that the desired amount of it

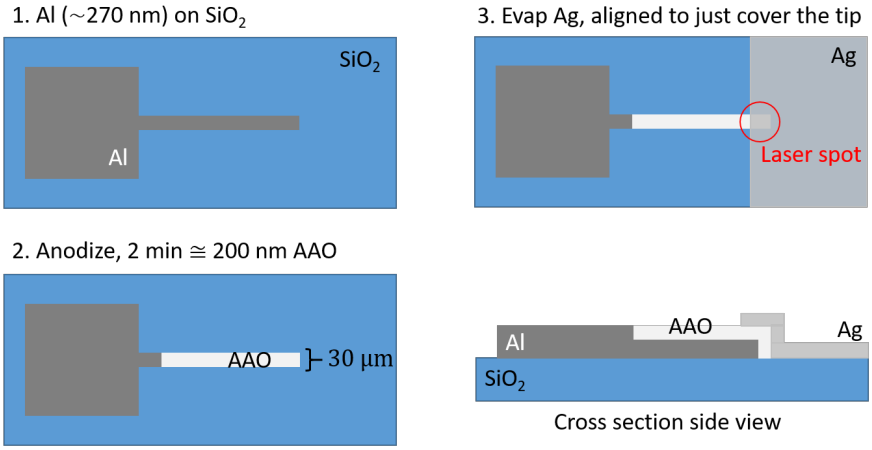

Figure 42: Planar structure for the RS platform. Step 2 can be done by either anodization or other oxide deposition methods.

(usually 100-200nm) can be anodized while keeping some thickness of Al intact underneath to act as the inert electrode (so typically an initial Al thickness of ~300nm). Then, the sample is held by a flat alligator clip attached to the large pad part of the Al and most of the thin strip of Al is submerged in the anodization bath. This causes the strip to be anodized. The small cross sectional area of the strip does not appear to give rise to any significant slowing or speeding up of the anodization rate.

3.4 Experimental setup

3.4.1 Overview, background, and requirements

The experimental setup for this project has several requirements. The most immediate facet is the electronic aspect (i.e., the setting and resetting effects), but the hope is to also witness OR, and thus we want the capabilities to measure it if it occurs. Therefore, there are several components. Additionally, RS is somewhat of a probabilistic process; while there will be a "typical" applied bias at which the device will set, over the course of many set/reset cycles, slightly

different applied biases over some range will also be able to set it. Much work has done on making the distribution of set/reset voltages less variable [12, 3, 2]. Thus, we can't simply do the same basic process every time. For example, if we wish to apply the smallest bias necessary to set the device, we couldn't apply the same bias every time: sometimes it would be too small and wouldn't set at all, and sometimes it would be bigger than necessary. Additionally, even when prepared with the same fabrication parameters, there is variation between devices, and when a device is first tested, its properties aren't known. However, scanning the whole parameter space manually would be tedious or impossible. Therefore, if we wish to investigate the properties related to OR and RS, we will need an adaptive system. To this, we employ carefully designed algorithms that take a basic process and repeat it with slightly changed variables each time, depending on the outcome of the previous run.

3.4.2 Physical setup

We have several requirements we wish to fulfill with our experimental setup. We want to 1) apply a bias to the device, 2) measure the current as a result of that applied bias, 3) shine a laser beam on the RS junction, 4) measure any current response to that applied laser beam, 5) apply a small AC bias in addition to the DC bias, 6) measure the 2nd harmonic of the AC bias, and 7) lower the temperature of the junction. The schematic of the entire setup is shown in Fig. 43.

Tasks 1 and 2 are both done by a pair of Keithley 236 Source Measure Units (SMU). They are used in voltage source/current measure mode. We should take a moment to define and discuss the concept of compliance. This is the maximum current that the machine will supply. To be clear, what the machine is actually doing when it limits the current, is monitoring the current it is supplying to the device at the given bias you have asked it to apply. When the current

it is supplying at that bias would be above the compliance level, it actually changes the bias it is applying to make sure only the compliance current can flow. Therefore, it is a little misleading when the machine hits the compliance current at some voltage level of the IV curve and reports that it's applying a higher voltage (while still hitting compliance). It is actually applying the bias it hit compliance at (or maybe even less, if the device has formed more since then, lowering its resistance). It also should be noted that the current compliance is an absolute quantity, so it is the maximum positive *or* negative current that is allowed to flow.

Two SMUs must be used, rather than one. The first SMU (the "bias SMU") is used to apply a DC bias to the sample. Since we also want to apply an AC bias, they must both be fed into a mixer that is applied to the sample. However, because the mixer is actively powered, the current that the bias SMU reports sourcing is not the current that is going through the sample. To know the current actually going through the sample, the current must be measured in series with the sample, directly before or after it. This is done with the second SMU (the "measure SMU").

To measure the optical response, a 250mW 632nm solid state laser is focused into a spot of diameter <1mm after passing through an optical chopper running at 2kHz. The mixed DC and AC bias is applied to the active electrode of the device, and the inert electrode then acts as the "signal test point", the spot in the circuit we wish to measure effects. For the optical response, this point is fed through a 100Ω series resistor and then connected to the input of a Stanford Research Systems SR830 lock-in amplifier (LIA), which acts as a virtual ground. The LIA is set to current input, AC input, (R, θ) mode, with the other parameters changing depending on the circumstance. Its reference frequency is supplied by the optical chopper.

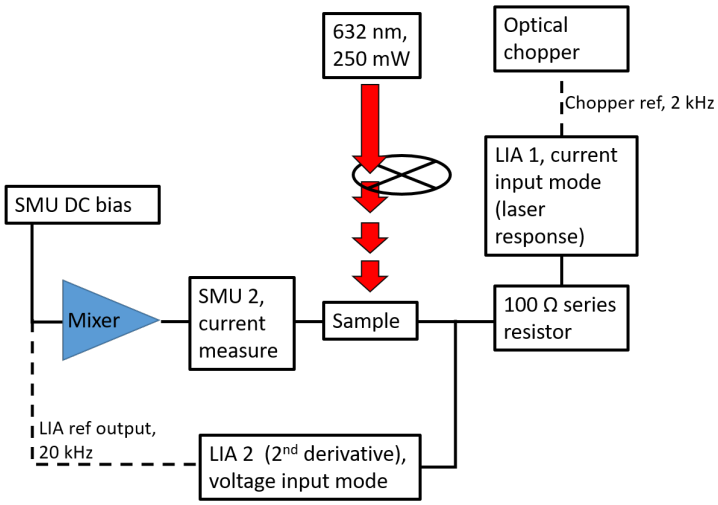

Figure 43: Schematic of experimental setup. Dashed lines indicate reference frequencies, solid indicate signal path.

To both apply and measure the small AC bias, a second LIA is used. The reference frequency is set internally to a 20kHz sine wave of amplitude 250mV, which is reduced to an amplitude of 10mV in the mixer. This is chosen to stay far away from the 2kHz chopper frequency as well as any of its harmonics. To measure the signal, the the signal test point is electrically connected to the input of this LIA, but in voltage mode, since now it's measuring the voltage across the test point and the virtual ground of the first LIA.

In addition, because electromigration speed appears to have an inversely exponential dependence on temperature, the setup must be able to alter that as well. While this was initially done using a cold finger cryostat with liquid helium cooling and a resistive heater to control the temperature, temperatures that cold weren't needed and the electrical noise from the cryo compressor and mechanical vibrations from the machine made it unworkable. Instead, the setup was modified to have the sample in physical contact with a Peltier cooler, the other side of which was water cooled by water circulating through a copper pipe.

Because the temperatures we used were typically below -10°C and we wanted to avoid condensation, this was performed under low vacuum.

3.4.3 Algorithms and software setup

To carefully probe the device characteristics of our devices, a large amount of parameter space must be covered. A few of the parameters that form this space are the applied bias, compliance current, sweep speed, integration time, temperature, among others. While it could be done manually, it is most efficiently done by an algorithm that can run many successive tests and adapt to the results of previous ones. The algorithms were programmed in LabVIEW, which supports a "finite state machine" architecture. Because this formulation is very useful and simplifies explanation of the algorithms, it is briefly explained here. A finite state machine (FSM) is a set of "states" a system can be in as well as a set of conditionals that can bring the system from one state to another. At all times the system is necessarily in one of those states. The arrows represent transitions from one state to another, and each has a condition, indicated by the text next to it. Typically there is a state the system starts in, and a final one that, when the system is in it, indicates that execution is finished. With this formulation, we can now more easily explain the algorithms used.

In this project, we used two main algorithms. They are similar to each other but ultimately probe different aspects of the resistive switching system. Because the data is produced from them, we briefly explain them here.

IV curve forming algorithm The following is a common procedure seen in the literature for RS devices. Very simply, a current compliance is set, and an IV curve is swept back and forth continuously; that is, while measuring the current, the voltage is swept from V_{lb} (can be negative or 0 depending on the application) to V_{ub}, and then immediately back down to V_{lb}, with no discontinuities in the

applied voltage.

This allows us to see several possible manifestations of interesting effects. First, it demonstrates (roughly) the smallest positive bias needed for some sort of formation to occur. This is visible in the form of sharply increasing current at some bias. However, depending on how sharp it increases, it could also be mistaken for another current mechanism. This leads to a second indicator that this test provides. If the forward IV sweep (from lower to higher bias) is different than the backward sweep (from higher to lower bias), that demonstrates hysteresis and a high likelihood that the device is changing as a result of the applied bias. A typical hysteresis IV curve that would indicate that RS is occurring might look like this:

This experiment is simple and effective, but has a few disadvantages. One in particular is that it is time dependent in an unideal way. Consider the following: if this test is performed and it's seen that current sharply increases at 3.0V, it's unlikely that it need *exactly* 3.0V. It's likely that 2.99V would cause formation, with possibly a longer wait. So the important point here is that the variable of formation time is important, and there is an inherent time aspect to IV curves that is often not considered or stated. The literature on this topic very rarely states the rate at which the IV curve was swept through. This itself doesn't diminish the experiment method, but speaks to an inescapable uncertainty with it.

However, this isn't as bad as it sounds. Even if there is a little uncertainty in the formation bias, for a well-behaving sample, it almost never forms for significantly smaller biases, and can't avoid forming for significantly larger biases. The point being here that the variability in forming time is largely dominated by the effect of its bias dependence, such with repeated runs the result becomes apparent.

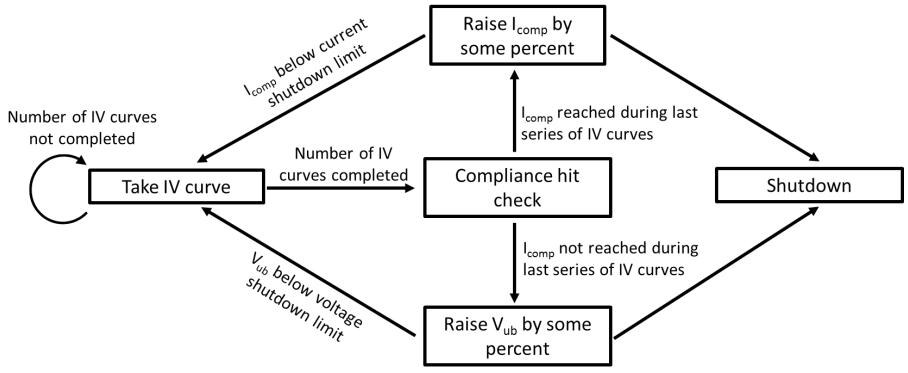

Figure 44: State machine for IV sweep forming algorithm.

Now we can explicitly look at the state machine in Fig. 44. For all these algorithms, there are a few inconsequential states that we omit (such as initialization procedures for the machinery), leaving just the important states. We must present a few important variables. The first is V_{ub}, which is the upper bound on the voltage that the IV curve is going to sweep to (and from, because for this algorithm we just sweep from -V_{ub} to +V_{ub}). This is how hard we are going to "drive" the formation and unformation for a given state. In addition, there is I_{comp}, which is the current compliance. Simply, the algorithm, for a given V_{ub}, I_{comp} state, does some number of IV curves. Then, if current compliance was hit for any of the IV curves, before the next run, I_{comp} is raised by some fixed percent. If compliance wasn't hit, instead, V_{ub} is raised. This allows searching of the parameter space in an efficient way. For example, without doing this operation of incrementally increasing V_{ub} and I_{comp}, we wouldn't know how high is too high an I_{comp}; if it's too high, it may "burn in" the device, making it permanently Ohmically conductive. Likewise, if we immediately set V_{ub} to too high a value and it hit compliance, we would not know where is the point at which it really starts introducing hysteresis.

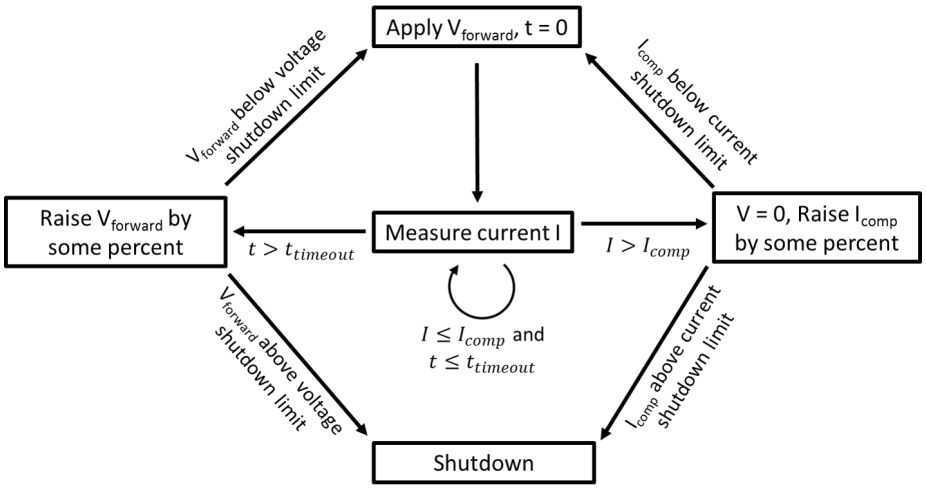

Figure 45: State machine for forming algorithm at constant bias.

Constant bias temperature switching algorithm In contrast to the IV curve forming algorithm, a simpler but perhaps more rigorous method can be used. The disadvantages of the IV curve method were explained previously, mostly concerning issues with time dependence. To solve this problem, we can use a method which explicitly includes the time dependence.

Here, in the simplest formulation, an initial constant bias is applied to the junction and the current is monitored as a function of time. If the current hits a set compliance, the applied bias is stopped and another action is done, depending on the subtype of this method, which will be covered below. If the compliance is not hit after a certain length of time, the constant bias is increased, and the process is repeated. A general state diagram for this algorithm is shown in Fig. 45.

This method has several disadvantages, most notably that only one bias is investigated at a time, and it typically requires longer experiment times for the same number of sets/reset cycles. However, its advantages are 1) it is rigorous, 2) the system can potentially be stopped at an intermediate forming

stage, and 3) other behavior can often be seen that might be obscured by an IV curve algorithm. With regards to 1), if it is said in a paper that "an initial voltage of 1V was applied for a timeout period of 120s and increased by 20% for every timeout", there is no missing information about the experiment there. Advantage 2) is important because one of the main goals of this project is to create a tunneling junction, which occurs in between the electromigration regime and the ohmic connected regime. If a good idea of the current magnitudes is had, it can easily be stopped when it reaches compliance. Lastly, it often more clearly displays the stochastic nature of formation; for example, sometimes for a given applied bias, the current will increase, implying some formation is occurring, and then suddenly decrease, implying that a filament is burning out.

We now discuss a few variants of this algorithm.

3.5 Experimental results

Here we present the results and discuss the electrical behavior of filamentation in alumina and AAO resistive switching devices.

3.5.1 IV sweeps algorithm results

The first experiment discussed here is the IV sweeps forming algorithm discussed above.

Hysteresis, On/Off Ratio, and normalized variance As discussed previously, this algorithm takes a series of IV sweeps forward and backward, to investigate data related to the hysteresis from RS. A typical IV sweep curve with hysteresis is shown in Fig. 46. The blue curve is the forward sweep, varying from -5V to 5V (increasing bias), and the red curve is the backward sweep, performed immediately after the forward sweep, varying from 5V to -5V (decreasing bias). There is clear evidence of hysteresis, demonstrated by the

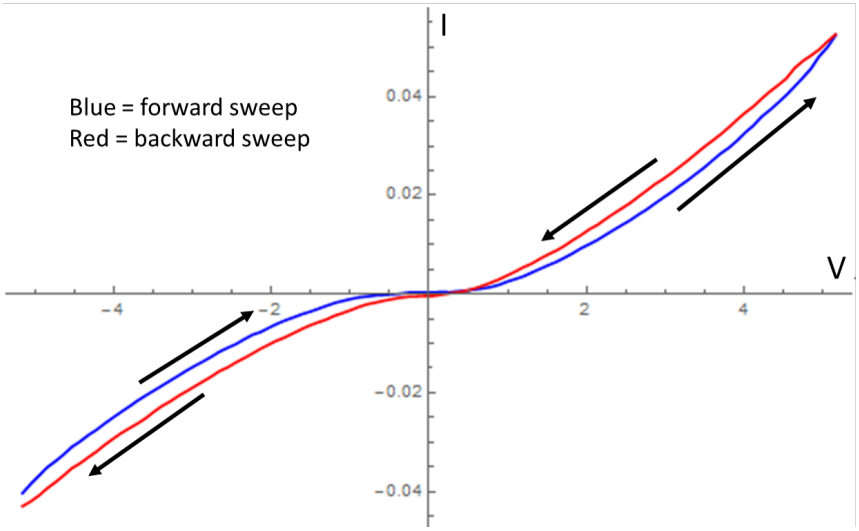

Figure 46: Example of forward and backward IV sweeps demonstrating hysteresis.

difference between the current of the forward and backward sweeps. However, it is not just the difference, but also the relative magnitudes of the currents. For positive biases, the device is being changed from a HRS to a LRS, and that is visible because points along the red curve are higher than the corresponding points along the blue curve with the same bias because the device has had more time for electromigration by the time that point is being swept.

This is for a single sweep pair. However, usually, the data taken for a single state (that is, a given V_{ub}, I_{comp}, and temperature) are Some number of IV sweep pairs taken in quick succession. This allows aggregate behavior to be seen, as well as being able to tell if any behavior that is witnessed are flukes or fluctuations, or an actual effect. A typical data set for a single state is shown in the top of Fig. 48. Ten IV sweep pairs are taken back to back with no delay in between. Each sweet direction is ∼100 data points, taking ∼10ms for each point, meaning that a given sweep direction takes ∼1s to perform.

One property of interest is the on/off ratio (OOR), simply defined as the

value of the LRS divided by the value of the HRS at each bias, as shown in Fig. 47. The middle of Fig. 47 shows how the OOR is calculated for a single forward/backward IV sweep pair, while the bottom shows the aggregate OOR's for the data at top. This is important for several reasons. The first is that, for practical use as a memory device, it represents the current difference between the "formed" and "unformed" states, which is what would have to be read to determine the memory state, and therefore a higher ratio would be desirable. The second reason, more important for our discussion, is that it demonstrates the V_{ub} or I_{comp} at which resistive switching starts occurring. Every IV sweep state covers at least the bias range of the previous states, but with either a higher V_{ub} or I_{comp}. Therefore, if there has been a low OOR for most states up to some point, and then either V_{ub} or I_{comp} is raised and the OOR increases dramatically, we know the bias and compliance current needed to give significant filamentation.

This is simple enough, but there are several important points to address. First, the OOR data is not meaningful for all of the bias range. Because the forward and backward sweeps both cross over zero current at some bias each, but those biases are usually not exactly the same, at the point where the off current (the denominator in the OOR) crosses zero, the on current will still be finite and the OOR will be extremely large. However, it will be nearly positive infinity on one side of the off current zero crossing, and negative infinity on the other side. We see this exactly in the bottom of Fig. 47, which displays the OORs for several IV sweeps on the same plot. The ratios tend towards positive and negative infinities for approaching the off current zero crossing from different directions.

Another caveat to discuss is the behavior of the OOR at V_{ub}. Because the sweeps are taken in quick succession, back and forth, even if electromigration

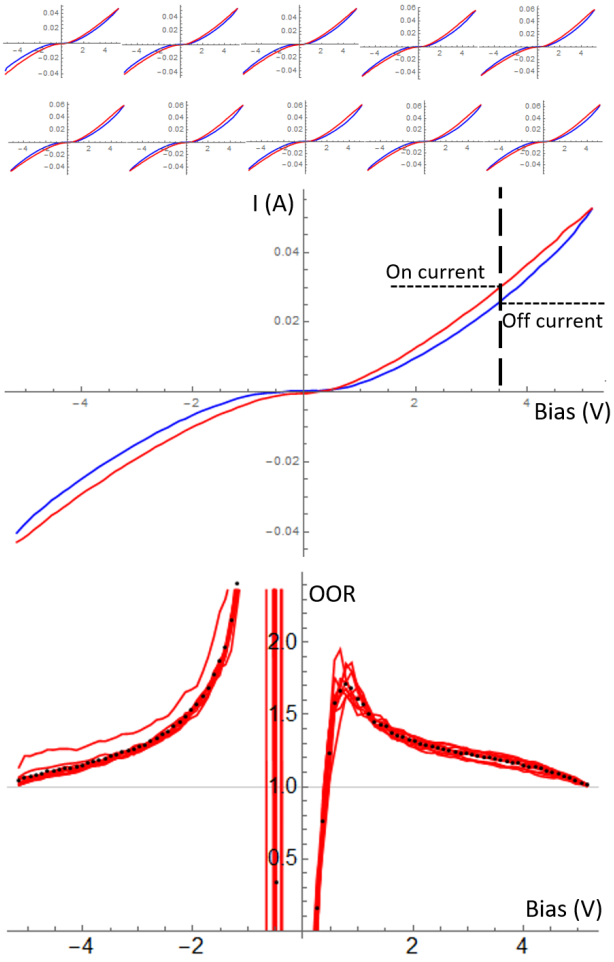

Figure 47: RS OOR behavior from a single (V_{ub}, I_{comp}) state. Top: ten forward/backward IV sweep pairs (blue = forward, red = backward). Middle: Illustration of OOR calculation for a single sweep pair. Bottom: calculated OOR for each sweep pair above.

has occurred and the on current is larger, at the exact point of V_{ub}, the currents must be the same, essentially due to continuity. This means that the OOR at V_{ub} will always approach 1, which can be seen in Fig. 47. It should be pointed out that this is mostly due to the backward sweep being taken immediately after the forward sweep; if it was held at V_{ub} for any length of time before decreasing the bias to do the backward sweep, as mentioned previously (one of the potential downsides of this experiment), electromigration would continue to occur and the current at that bias would continue to increase.

So, the typical behavior we see of the on-off ratio, for positive biases, is an increase from the previously discussed negative infinity, to a physically meaningful value, and then to a peak in the ratio, before it has to decrease again to tend towards the ratio of 1 we see at V_{ub}. Therefore, for all groups of sweeps in a given state, for the OOR, we are interested in the maximum OOR (within the physically meaningful bias range), the bias at which it occurs, and the V_{ub} of the state required to give rise to this max OOR. This is where a device could both be practically used, as well as a good bias point for testing the resistance state of the device. In Fig. 47, looking at the compilation of all the OOR curves, we can see that there is a max OOR of $\sim$1.8 at $V = 1$ Volt, from a V_{ub} of 5Volt.

Another parameter analyzed is the variance of the forward and backward sweeps, as shown in the middle of Fig. 48. If, for a given bias, we look at the values of all the forward sweeps, we can see that they cover some range, and likewise for the backward sweep. From each of these we can form the variance for the forward and backward sweeps at every bias. This is an important practical parameter, because if these devices were ever to be used practically, a low variance is necessary; it is essentially a measure of repeatability for these devices. However, a couple details should be mentioned: first, we must normalize this variance by the magnitude of the current at that bias, because otherwise the

raw variance will always be larger for larger currents, even if they have relatively less variance. Therefore, when we discuss the variance for a set of curves, we actually mean the variance divided by the mean of those curves, the normalized variance (NV).

Another facet that must be mentioned is that the NV data, too, is not meaningful for all of the bias range. This is due to the inherent noise level of the measurement and the normalization. For example, if our noise level (the measured fluctuations in the current even with a completely open circuit, due to radio noise/etc) is 10^{-11}A, then for very low biases in which the current of our device is below the noise level, the NV will be about equal to this noise level. However, because this noise is random, its average of several runs will tend towards zero. Therefore, the NV will be dividing a finite number by a very small one, and report an unphysically large NV.

Aggregate data trends The preceding has been analysis that is done to a single (V_{ub}, I_{comp}) state, but now we consider what can be figured out after the system has run through many states. As an example, consider the run shown in Fig. 49, where the IV sweeps are plotted for each state. Because the IV sweep algorithm increases either V_{ub} or I_{comp} after each state, and for each state we could calculate the OOR and NV, now we can analyze the OOR or NV as a function of each of those states. There are a few ways to do this, illustrated in Fig. 50. One way is to plot the relevant variable (OOR or NV) vs the V_{ub} for that state (Fig. 50 (a)); this is good because V_{ub} is the main parameter we control that influences the behavior of a state. However, recall that a state can have the same V_{ub} as its previous state if I_{comp} was reached and I_{comp} rather than V_{ub} was increased. This can make plots (for example, of the OOR vs V_{ub}) look confusing, because there will be several stacked points with different OORs but the same V_{ub}. An alternate way is to plot the points simply by the order of

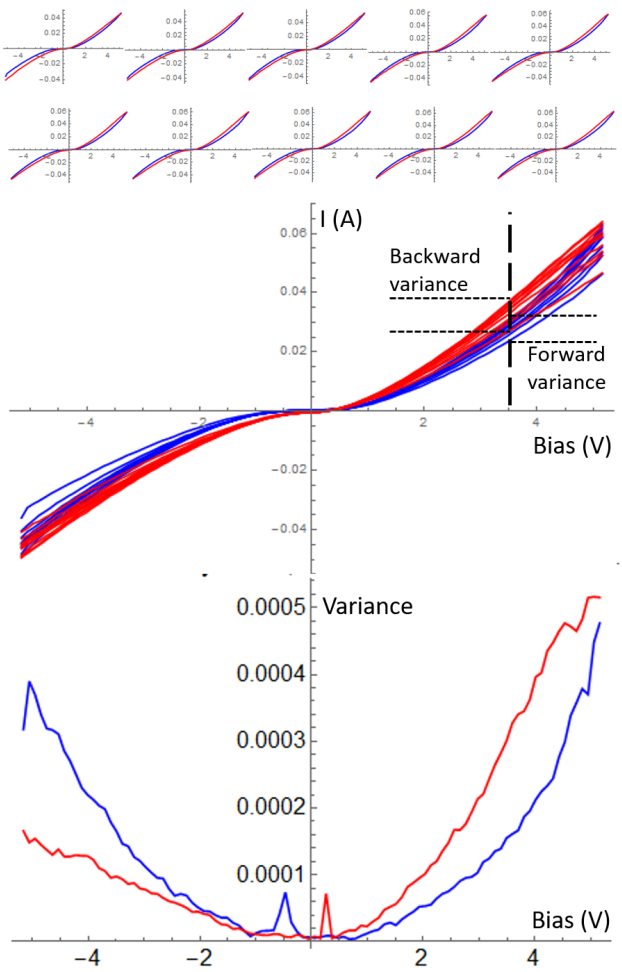

Figure 48: RS variance behavior from a single (V_{ub}, I_{comp}) state. Top: ten forward/backward IV sweep pairs (blue = forward, red = backward). Middle: Illustration of variance calculation for each sweep direction. Bottom: calculated variance for each sweep direction.

the states (Fig. 50 (b)), labeling them separately. This solves the problem by spreading out the bunches points that have the same V_{ub}. However, although both of these analyses are useful, they have another weakness. Often, as will be evident below, a sample can temporarily "burn out" during a run (similar to unipolar operation), and go back to a more HRS. This typically causes the sample to behave as it did for a lower V_{ub}, even though it currently has a higher V_{ub}, which doesn't represent trends that may be occurring well. A third analysis method, also used here, is to plot relevant variables against I_{max}, the maximum (of the average of the) forward sweep currents for a state (Fig. 50 (c)). This tends to show trends more clearly, because while V_{ub} and the state index are more like settings or labels for the state, I_{max} actually represents a measured quantity of that state.

Fig. 50 illustrates this. It represents the aggregate of the data in Fig. 49 and several effects can be seen from it. First, for this sample and run, looking at Fig. 50(a), OOR_{max} is increasing with I_{max}, which is expected, until V_{ub} =2.2V. At this point, it can be seen (compare with the corresponding IV in Fig. 49 as well) that the current is rising too sharply to maintain, and unipolar action takes place as the sample changes from a LRS to a HRS. Additionally, we can see that the NV measured at the the bias the OOR_{max} occurs at also has a trend of increasing with I_{max}, as seen in Fig. 50.

Another dataset, better behaved, can be seen in Fig. 51. Its aggregate behavior is also plotted in Fig. 52. It shows the same trends, that of OOR_{max} increasing with I_{max}, but without any unipolar behavior. Additionally, it also has the NV increasing with I_{max}. It appears that, in general, the NV is larger for the backward portion of an IV sweep pair than for the forward, at the same I_{max}. These data sets demonstrate an unfortunate balance that must be struck: OOR_{max} increases with I_{max} (desirable), but the NV also increases

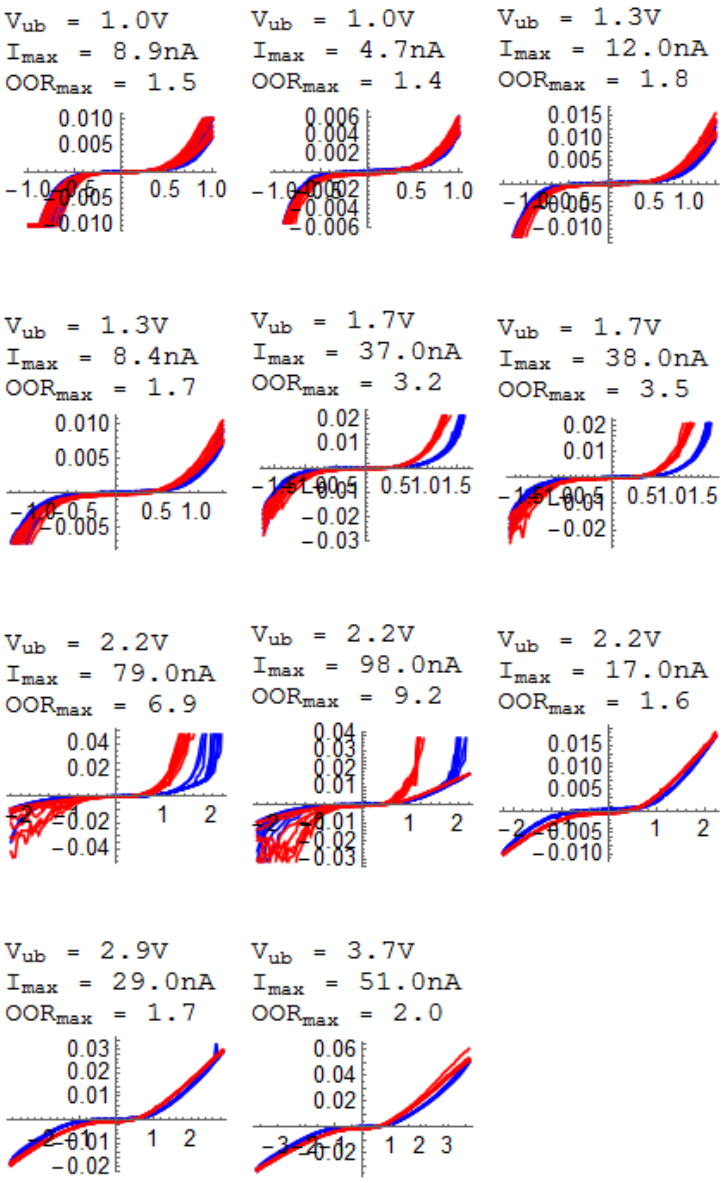

Figure 49: A series of states for a single run of the IV sweeps algorithm, progressing from left to right and top to bottom. The parameters for each state are displayed above the plot of the state sweeps.

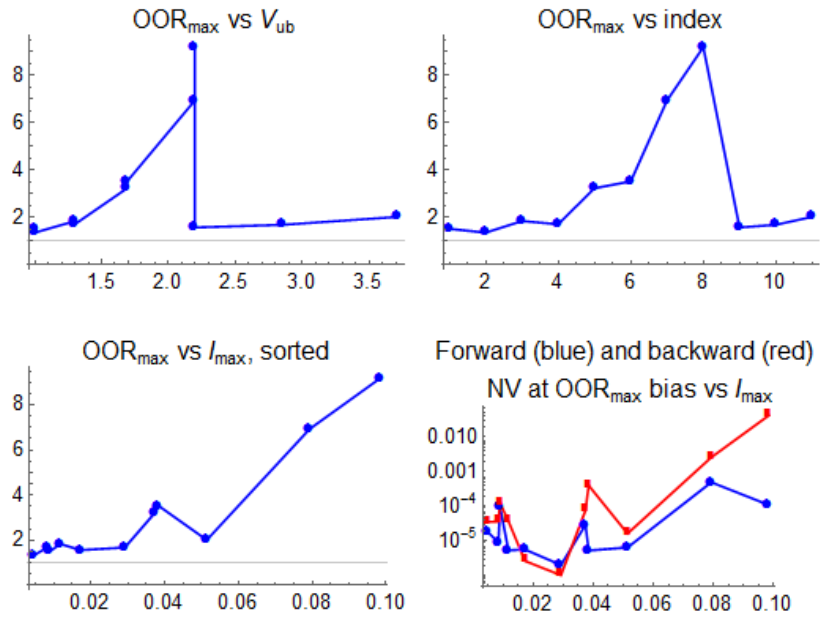

Figure 50: The aggregate data for Fig. 49.

with I_{max} (undesirable). Therefore, for both practical use of these devices and our own purposes, a decision must be made that determines where in the tradeoff between OOR_{max} and the NV the device will operate. Further, two other device characteristics should be taken into account. As seen in Fig. 50, unipolar switching occurred, which would want to be avoided (unless the device was specifically designed for that), so I_{comp} would determine this. A similar problem is that of "burn in"; too much current could also cause the device to become permanently Ohmic and impossible to reset.

Temperature dependence The data produced by the IV sweeps algorithm above illustrated interesting behavior of this RS system, but we can gain more information by also probing the temperature dependence of the same properties. To do this, we use the same algorithm, but with one extra facet. In addition to the sweeping, at every (V_{ub}, I_{comp}) state, the IV sweeps are repeated for both

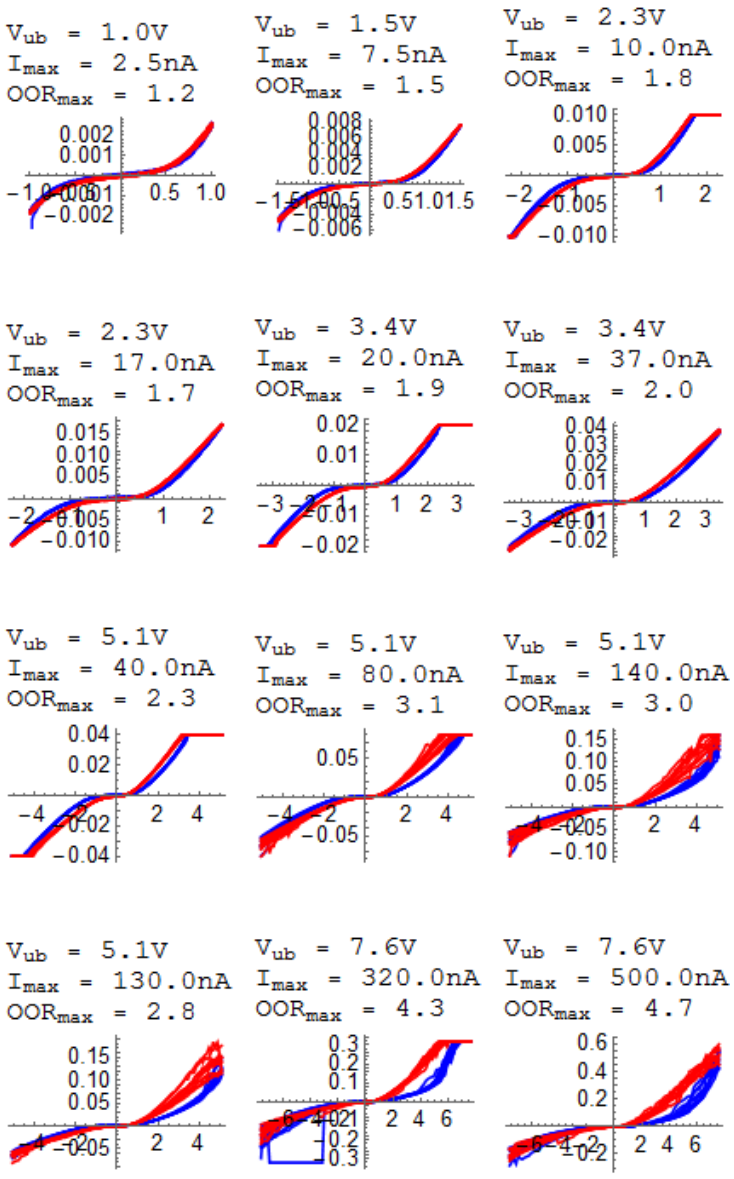

Figure 51: A series of states for a single run of the IV sweeps algorithm, progressing from left to right and top to bottom. The parameters for each state are displayed above the plot of the state sweeps.

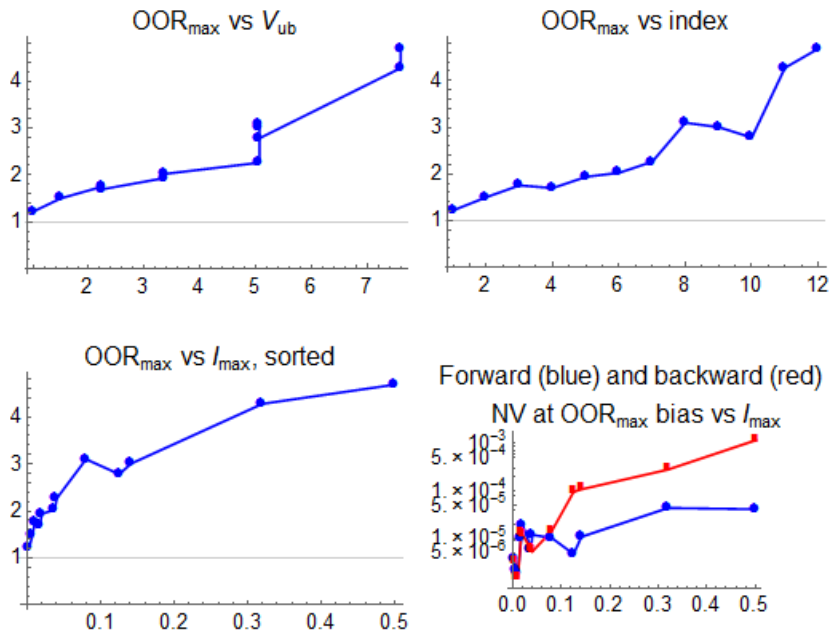

Figure 52: The aggregate data for Fig. 51. The aggregate behavior is shown to have ideal, monotonic behavior with I_{max}.

a hot temperature (T_h) and a cold temperature (T_c). The same procedure for raising either V_{ub} or I_{comp} is used, but I_{comp} is raised if it was reached during either the hot or cold state for the given properties (i.e., nothing is changed in between the hot and cold states).

This allows observation of several phenomena, which is presented below. The main cause of all of these temperature dependent phenomena will be that electromigration rates have a very nonlinear dependence on the temperature [105]; therefore, when the temperature is raised, we should see a larger difference between the LRS and HRS state currents at a given bias, as well as other effects. For these experiments, typically $T_c \sim$ -25°C, while $T_h \sim$50°C.

Therefore, we now discuss several trends that appear in the analysis of this data. The first and simplest to discuss is the maximum OOR for each state, as a function of that state's V_{ub}. Additionally, because each state is essentially repeated for both T_{high} and T_{low}, we can plot on the same plot the OOR vs I_{max} for both temperatures. Consider Fig. 53, showing a series of IV sweeps produced by this temperature changing algorithm. The temperature of the state is labeled above its IV plot with the other relevant variables. However, these states are only about half of the states for this run; there are too many for this run (and most runs) to present all of them (and they are usually repetitive), so from now on only the aggregate data will be presented unless there is specifically anomalous behavior.

As discussed before, in general, we expect a few main trends: 1) there to be a higher maximum OOR for higher V_{ub} or I_{max} (aside from erratic unipolar effects), 2) there to be less electromigration at T_c, meaning that for the same V_{ub}, the current is smaller, and 3) resulting from this, causing the OOR to be smaller at T_c. Looking at Fig. 54, the aggregate data of Fig. 53, this is exactly what we see. The maximum OORs for both temperatures are increasing with

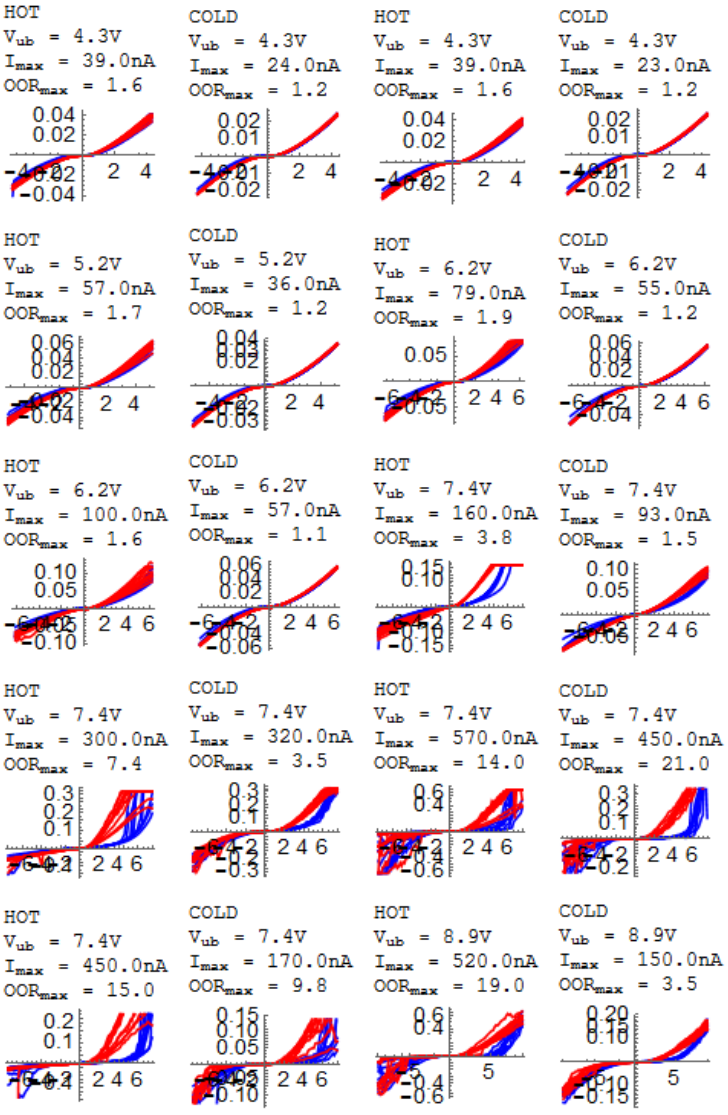

Figure 53: A series of states for a single run of the IV sweeps algorithm with changing temperature, progressing from left to right and top to bottom. The parameters for each state are displayed above the plot of the state sweeps. Note that every state is repeated for a hot and cold temperature, which are immediately adjacent. Not all states are shown for this run because there are too many.

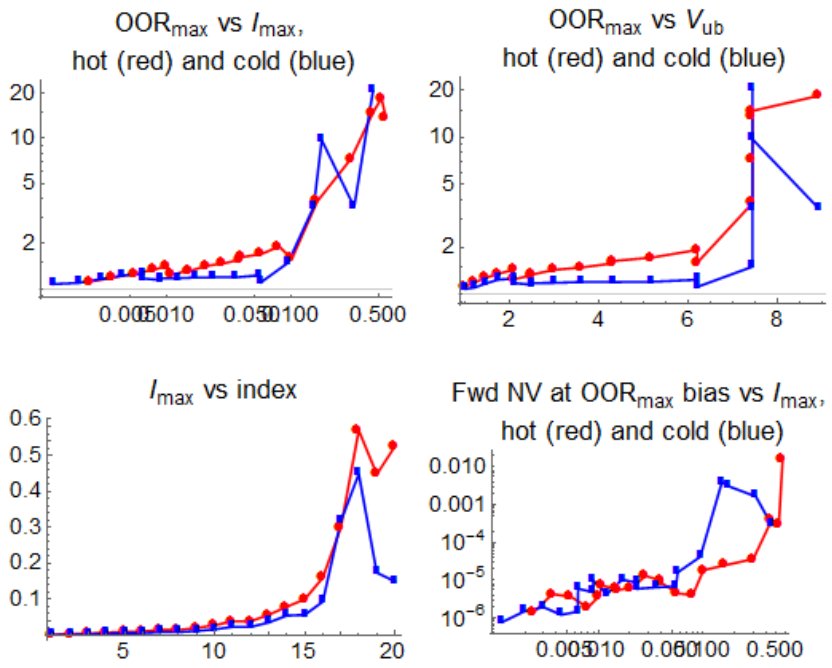

Figure 54: The aggregate data for Fig. 53, an evaporated alumina sample. The aggregate behavior is shown to have mostly monotonic behavior with I_{max} with noticeable trends in the hot vs cold sweeps.

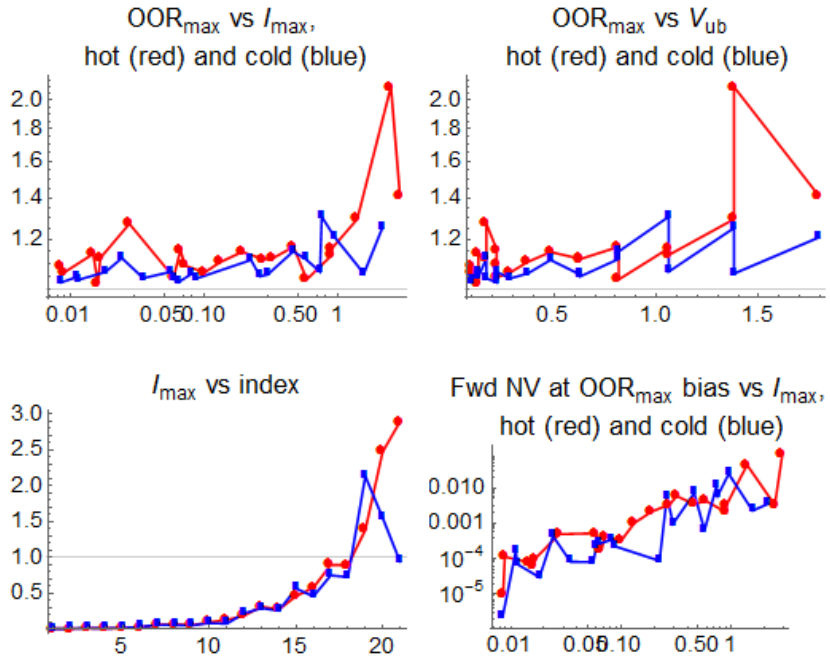

Figure 55: The aggregate data for another set of IV sweeps with changing temperature for a AAO dielectric sample.

V_{ub}, but the T_h one is usually larger. We can also see that the maximum OOR for both temperatures are roughly equal as a function of I_{max}, but this is still consistent with the first result because the hot sweeps are typically reaching a given I_{max} at a lower V_{ub}, as predicted. Lastly, we can see in Fig. 54 (d) that the NV has the trend of increasing with I_{max}, but there are no very obvious trends in the respective magnitudes of the NV for hot and cold temperatures. Lastly, there is likely a small amount of unipolar resetting occurring during the last couple states, when I_{max} decreases a little.

Another run showing similar behavior can be seen in Fig. 55, though the effects are less pronounced and it can be seen that the final V_{ub} is much smaller (1.8V). This could be due to the different sample preparations; the dielectric of the sample presented in Fig. 54 was formed by e-beam evaporation of 100nm

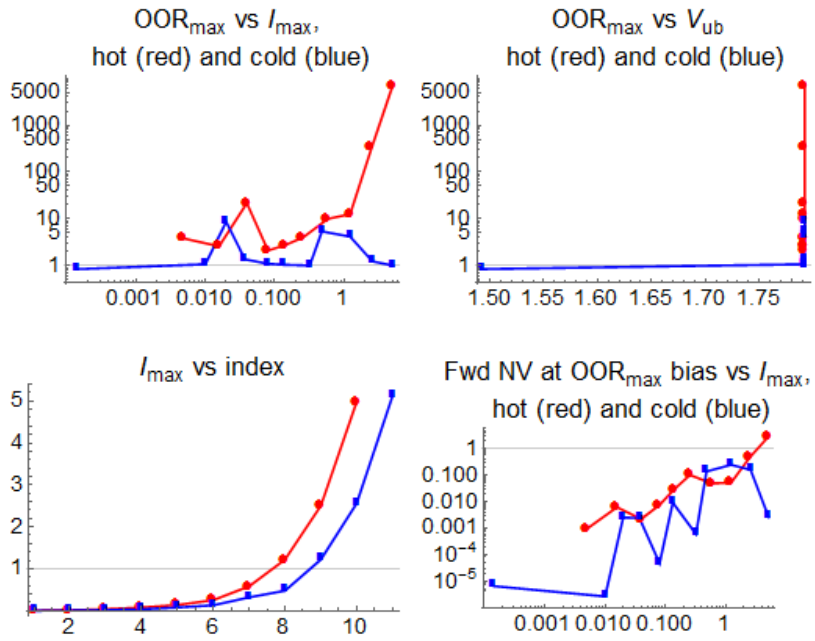

Figure 56: The aggregate data for another set of IV sweeps with changing temperature for another AAO dielectric sample.

of Al_2O_3 as a comparison to samples with an AAO dielectric layer, such as the sample presented in Fig. 55 whose dielectric layer was formed from 1 minute of anodization at 40V.

The aggregate behavior of another AAO dielectric sample is show in Fig. 56, though this sample was anodized for 2 minutes. We can see that similarly this sample has a low (compared to the evaporated alumina sample) final V_{ub} of ∼1.8V, but the aggregate data from this one is very different, with maximum OORs in the 10^3 range as well as very high NVs compared to other samples. To illustrate the cause, we present the last IV states of the run in Fig. 57. We can see a very large difference between the hot and cold states, with even subsequent cold states having a very small NV compared to the immediately previous hot state. Additionally, we can see the cause of the extraordinarily high maximum OORs for hot states: because unipolar switching occurs for some of the forward

sweeps, at certain biases those unipolar sweeps have very low current. The maximum OOR is calculated by taking the ratio of corresponding backward and forward sweeps as a function of bias, and then averaging all these OORs. This typically gives an accurate depiction of the OOR for a state even if the currents are changing between each sweep pair, but it has the weakness that a single sweep pair can mess up the whole OOR for that state if it's anomalous or bad. This could be mitigated by removing the sweeps pairs that show strong unipolar action from the OOR average for each state by filtering out OORs that are a certain standard deviation away from the median OOR. However, for this case the main point to elucidate is that AAO dielectric samples seem to have much higher temperature dependence for the EM, which we will use to our advantage.

Additionally, across the samples presented here and in general, AAO dielectric samples tend to need a lower V_{ub} to reach the same high I_{max}s than evaporated alumina dielectric samples. Similarly, they appear to have less dependence of V_{ub} on their anodization time, which is intuitively strange; shouldn't a thicker dielectric decrease the E field acting on EM particles and cause it to need a higher V_{ub}? There are a couple possible explanations for this. One is that, due to the structure of AAO, there is technically a constant alumina thickness: the barrier layer. While AAO grows at an approximate rate of 100nm/min, that is the growth rate of the pore thickness; the barrier layer is constant, independent of the pore thickness. That could explain both the smaller required V_{ub}, as well as the independence of it on the anodization time. Another explanation could be due to the topography of the Al underneath the barrier layer. The bottom of the pores form small points of Al where they intersect. This could give rise to field enhancement at those points, making it easier for filamentation at a given bias, compared to a flat Al electrode layer.

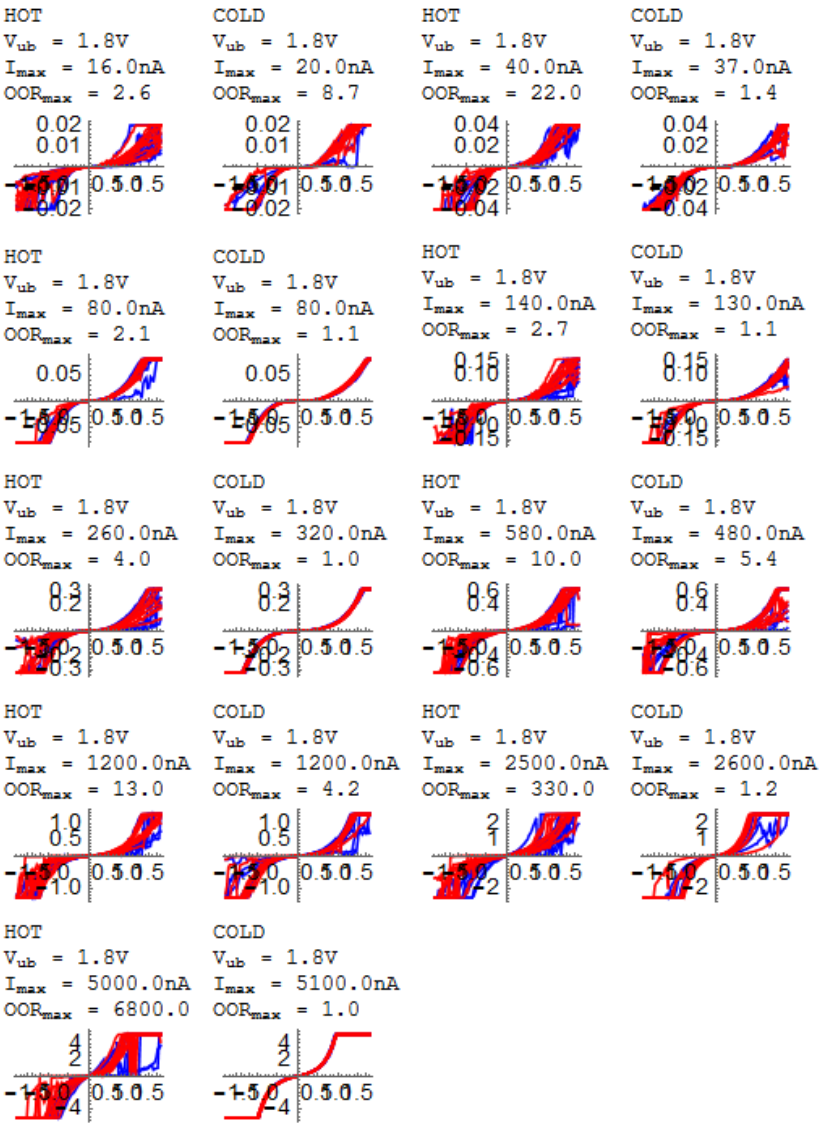

Figure 57: The IV sweeps data for Fig. 56, illustrating the cause of the physically insignificant giant OORs for some of the late states.

3.5.2 IVT algorithm results

The above results are good for probing setting and resetting behavior and trends. However, the original goal of this project was to demonstrate OR in a RS filamentation system. The data produced by the IV sweeps algorithm guides our search for this in an indirect way (e.g., by giving information about necessary setting biases/currents, temperature dependence, etc), but it is very hard to get reliable data that way for several reasons. The most salient reason is that the OR signal we are searching for is *very* small, if it's even there, and the LIA that is being used to search for it will respond to many other effects. Probably the largest obscuring effect can be any change in current; that is, the way the LIA works is by searching for signals at a known frequency, in our case 2kHz for the chopped laser signal. Ideally, a simple IV curve with no other effects present would have no components at this frequency and the LIA would have zero signal. However, whenever there is a change in current, the Fourier Transform of it includes many frequencies necessary to give rise to the transient change. Therefore, the LIA will respond to changes in current, and the fast IV sweeps performed by the IV sweep algorithm would naturally give rise to a misleading LIA signal. Indeed, LIA response to the laser was seen for much of the IV sweep algorithm data, but deemed unreliable for this reason.

A way around this and other problems can be solved by using our IVT algorithm, explained earlier. Again, there is a added detail, though. When the system is applying the constant bias and measuring the forming current, it is at an elevated temperature T_c. When the current hits I_{comp}, it first lowers the temperature to T_h. Similarly to before, usually $T_c \sim$ -25°C and $T_h \sim$50°C. The reason for doing this is that we want to separate the forming regime from the IV curve regime as much as possible; we want to cause formation and conduction of increasing currents (during the constant bias application) with as small a bias

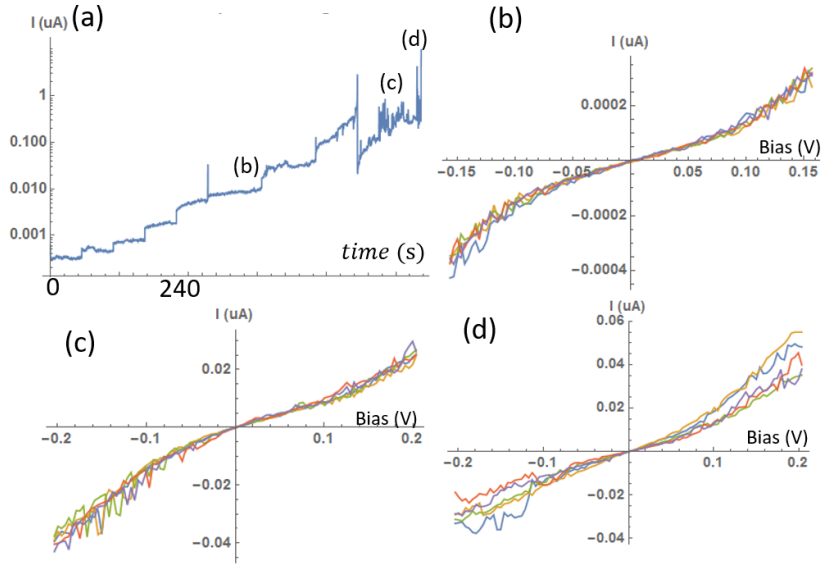

Figure 58: A typical IVT run and IV curves taken during the run at low temperature. (a) The IVT run. The horizontal axis is the index of the point taken, rather than time, because the stretches of constant bias are separated by long gaps when the IV curves are being taken. (b-d) IV curves taken at certain points in the IVT run, after a given compliance threshold has been hit. The positions of these IV curves are shown inset in the IVT progression. The magnitude of the IV curves is shown to be increasing, implying formation is occurring.

as possible. If we were to use a larger bias, it would likely drive EM current so quickly that the regime we want to probe for OR would be passed over immediately. So, using our information from the IV sweeps section, we raise the temperature during the constant bias formation process to make it easier to form at a smaller bias, and then drop the temperature to take the IV curves, as well as take the IV curves over a range that is only a fraction as high as the forming bias.

An example of a typical IVT run is shown in Fig. 58. Fig. 58 (a) shows the "IVT progression", meaning the currents measured at constant bias, over time. Steps can be seen in this curve; those are regions of one specific bias. If a step is flat, that indicates that no forming EM was able to occur for that applied bias, so

it timed out and a higher bias was applied at the next step. This often leads to a higher baseline current, even if the current is not growing within that step. This is most likely due to small amounts of EM happening, but no serious filaments being formed (note the small units of current). However, at some point, a high enough bias is applied and filamentation occurs. For this sample, the bias had to be increased six times before the first compliance level of 10nA was hit. IV curves of interest taken at several points are shown in Fig. 58 (b-d), and labeled on the IVT plot. Two things should be noted here. The first is simply that the current magnitudes at the same biases for the different IV curves are drastically different ($\sim$100pA at 0.1V for Fig. 58 (b), $\sim$10nA at 0.1V for Fig. 58 (d)), which shows that this formation procedure is definitely causing the nature of the junction to change. Another point is that not just the magnitude of the IV curves, but their shape as well, are changing. This is another good indicator of a changing junction, though it can be difficult to identify the regime just by the shape of the IV curve, due to the complicated nature of tunneling junctions and EM. Lastly, the repeatability of the IV curves gives information regarding whether EM is occurring; as before, if they are different between successive runs, that strongly implies that the act of measuring the IV curve itself is affecting the junction. This could mean that the IV voltage bounds are too large, the temperature should be lowered, or it may just be very difficult to measure a junction that sensitive without affecting it. It can be seen that Fig. 58 (b-c) have very repeatable IV curves, but Fig. 58 (d) has more change between them, as well as different curvature from the IV curves of Fig. 58 (c); therefore, EM is occurring during the constant bias application between these two IV curve sets.

Additionally, the stochastic and erratic behavior of formation can be seen in Fig. 58(a). Looking at the current measured at around index 4500, we can see that a given current threshold was reached, then IV curves were taken, and

when the system went back to constant bias application, the current was actually *lower* than before. This could either be a sign of EM/resetting occurring during the IV curves, or unipolar operation. Unipolar operation is definitely a common occurrence regardless, as we've recorded many other IVT curves that have decreased in current without hitting the current threshold (meaning that resetting due to EM from the IV curves wasn't a possibility).

3.5.3 Evidence of OR

Measurement Lastly, we discuss data that may be indicative of OR occurring in our filamentary RS system. It is produced by the IVT algorithm above, also with the temperature change between formation/IV curve measurement. However, there are two important differences that had to be done in the measurement to achieve what we believe may be OR. The first is related to the number of data points, noise, and averaging. As mentioned previously, the effect is *very* small if it's there at all. Therefore, even a small noise level due to many other effects can obscure it. This was likely occurring before, when we were measuring several IV curves successively and quickly. Changing the measurement to use a 3s LIA integration time (for a 2kHz signal, so integrating over $3 \times 2,000$ cycles) as well as averaging each bias point over 200 points revealed behavior that was not visible before. Second, reducing the bias bounds of the IV curve allowed for higher resolution, and looking at a bias range that was more likely to give fruitful results. It was noticed previously that the LIA response signal often showed some matching to the 2nd derivative for low biases but diverged for higher biases. Ward et al reported the same in their seminal paper on OR in a nanogap [97]. This will be discussed further in the section about pitfalls of OR measurement.

2nd derivatives The primary way of showing that OR has been achieved is by showing that the LIA signal due to the laser is proportional to the 2nd derivative (D2) of the IV curve. There are several ways of doing this but they fall into two main categories, numerical and experimental. Numerical methods are ones where the D2 is not measured directly, but calculated from other data, usually the IV curve. A couple ways include interpolating the data and taking a finite difference D2, similar to the well known mathematical derivation of the derivative in calculus. Another way involves fitting an analytic function to the data, such as a high order polynomial or function that has a form similar to what could be expected for a tunneling IV curve, and then taking the analytic D2 of it. These certainly have their advantages: 1) One less measurement, 2) IV data tends to be much less noisy, 3) If an analytic model is fitted, it is essentially replacing the data by a clean function, so the D2 of it is also completely clean. However, they also have their disadvantages. Most notably, they are less "physical"; that is, it is *expected* that the D2 that the OR signal should match should be the same as the analytically calculated one from the IV curve, but it could potentially be different. Secondly, calculation of the D2 is never trivial. Shown in Fig. 59(a) is an experimental IV curve (blue) with its interpolation laid on top of it in red, which appear to be a very good match. However, shown in Fig. 59(b) is the LIA signal for that IV curve (blue) and the D2 of the interpolated function from Fig. 59(a) (red). This illustrates that, despite the interpolated function being nearly as perfect as can be, taking derivatives necessarily adds a ton of noise; recall that the D2 is essentially a measure of the *curvature* of a function, so any small nonlinearities get magnified. A similar problem presents for taking the D2 of a polynomial fit; any number of terms can be used to fit the original IV curve, achieving a very close fit, but more accuracy requires more terms, and the higher terms

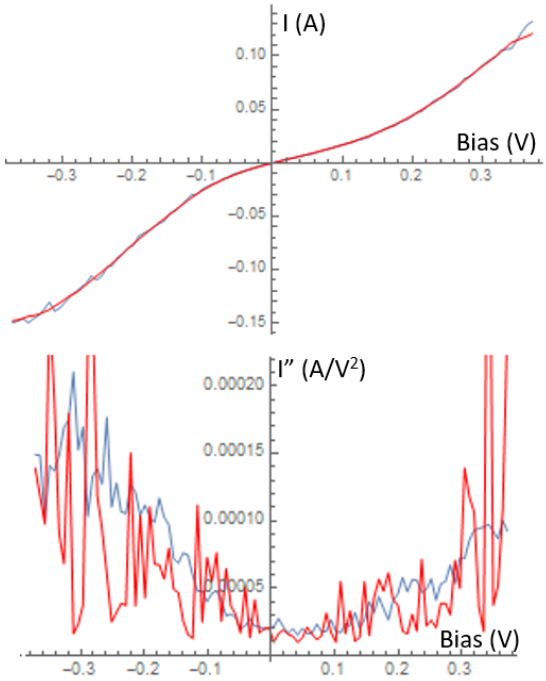

Figure 59: Top: A measured IV curve (blue) and a function interpolated to it (red). Bottom: A simultaneously measured LIA laser signal (blue) and the second derivative of the interpolated function from top. Note that it matches the shape roughly, but is too noisy to be meaningful.

have more zeros. Regardless, numerical derivatives were used in this project (as well as the following) because they're easy to calculate with data that is already being taken anyway, and can be a useful confirmation.

The better alternative to numerical methods, however, is experimentally. As seen in the Taylor expansion of the IV curve, if a small AC bias is added to the DC bias, the measured 2nd harmonic of the frequency of this AC bias will be proportional to the D2. This is a much more concrete way of measuring the D2, but is not perfect either. First of all, it is another measurement that must be done, that makes measurement of the LIA signal a little more difficult. For example, now a mixer must be used to apply bias to the sample, to apply both

the SMU DC bias and this AC bias. Additionally, the output signal from the sample must now be split into one path going to the LIA for the OR signal, and another LIA for the D2 measurement. Lastly, there is the question of what amplitude AC signal to use. It too is a small signal to be measured, so a larger input signal is easier to measure. However, the larger the signal, the more range this D2 method is "sampling"; that is, for an AC amplitude of V_{AC} at a DC bias V_{DC}, it is really sampling the curvature between $V_{DC} - V_{AC}$ and $V_{DC} + V_{AC}$. This isn't a big problem for small V_{AC} because the curvature shouldn't be changing too rapidly, but it still stands that the larger the amplitude, the less resolution achievable. For our measurements, we typically used an AC bias of frequency 20kHz and amplitude 10mV.

Experimental results Now we briefly present some experimental evidence for OR. This same signal has been produced for several samples now, repeatedly, but we only present results for one sample at several points here, because the others are similar enough that they don't reveal any different information.

Shown in Fig. 60 are three datasets taken during the IV curves run of the IVT algorithm, taken at three different formation points. For each, the IV curve is shown to the left, for comparison. On the right, the measured D2 curve is shown in blue, while the measured OR signal is shown in red. It can be seen that they match fairly well over the whole measured bias range. Importantly, as discussed in the next section, they are nonlinear as well, which provides stronger evidence for OR. Note the current magnitudes of the IV curves, 300-600nA, a lot for a typical tunneling current. However, this is in line with results others have reported, such as Ward et al.

Pitfalls of OR measurements Now we discuss the many pitfalls related to OR measurement to illustrate that, although the data appears to indicate that

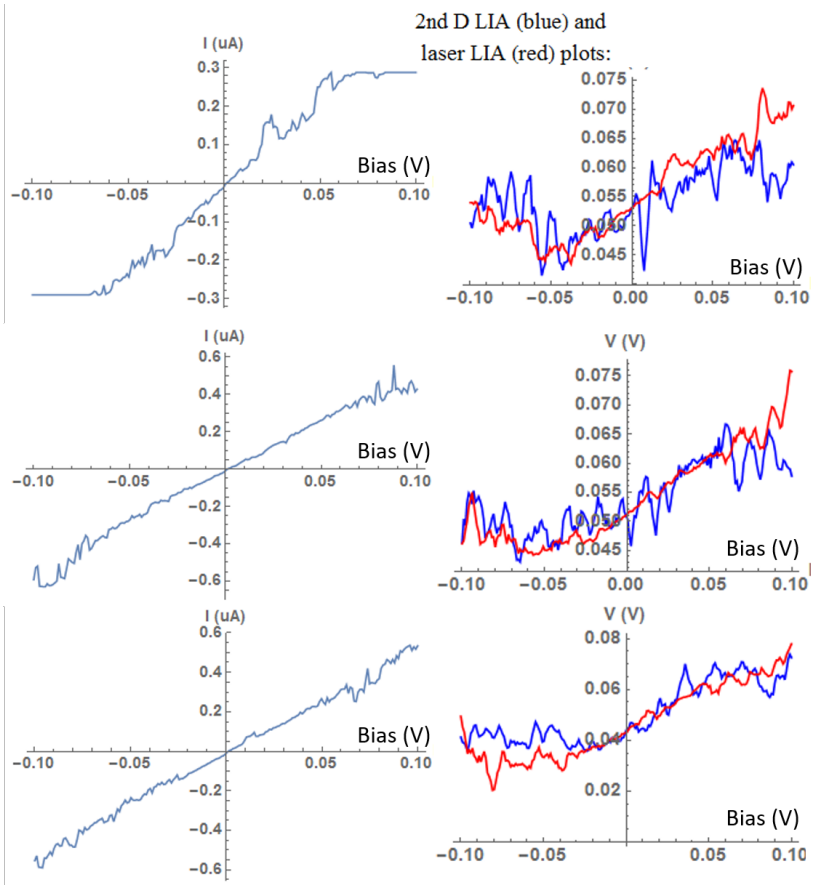

Figure 60: Possible OR signal for three IV curves, taken at different regions during the IVT progression. For each, the IV signal is on the left and the corresponding 2nd derivative measuring LIA (blue) and laser LIA (red) signals are on the right. It can be seen that the 2nd derivative and laser signals track each other fairly well throughout the bias range, but do not match the IV curves significantly.

OR is occurring, mean that we should be careful and search for more evidence, ideally of different types. We also bring it up here to highlight the reasons why our data plausibly avoids these.

The first pitfall is due to thermal effects. There are many types of *these* too, but they all originate from a focused, powerful laser hitting a part of the sample with very small thermal mass, quickly heating it up. They could be due to bolometer-type effects, where it is simply heating metal up, lowering its resistance, which will cause more current to flow at the same frequency as the laser. Alternately, it could be hitting the tunnel junction, causing the temperature to rise and the Fermi distribution to spread out, giving rise to a higher net current in one direction. Almost every paper on OR inevitably gives a handwavy justification for why their OR signal is not just thermal, that are usually plausible but not concretely convincing. For example, Ward et al provided the following justification: "We do not expect thermal voltages in our experiment because the device is a comparatively symmetric design, and with a centered laser spot there should be no temperature gradient between the electrodes, even in the presence of minor structural asymmetries. The macroscale unilluminated parts of the electrodes also act as thermal sinks." Most of this is either ignoring certain effects or is unrealistic. For example, a bolometer type effect would occur with asymmetry or not. Also, the other parts of the structure acting as heat sinks may be irrelevant, because the heating due to the laser pulse may occur and be measured more quickly than it is dissipated by the sinks (which, themselves, could also give rise to thermal effects).

One way to be more confident that thermal effects aren't occurring is based on the fact that most thermal effects will manifest, at least to first order, as proportional with the IV curve. Therefore, if a measured OR signal has the same exact shape as the IV curve, that doesn't mean it's only a thermal signal,

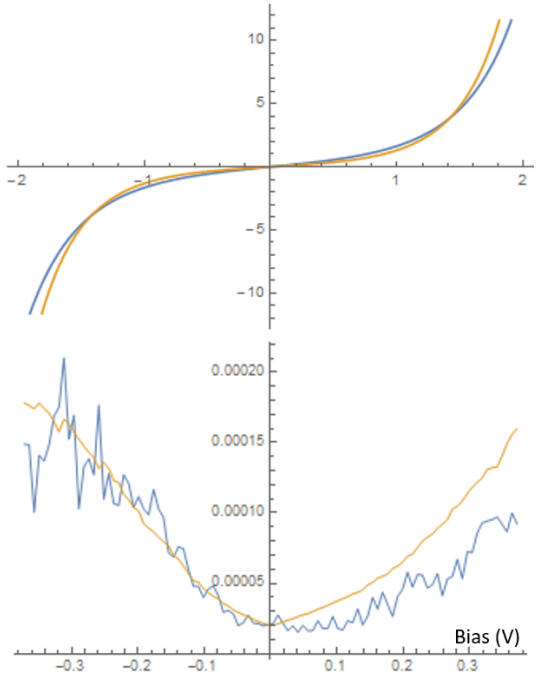

Figure 61: Top: an analytic nonlinear curve and its analytic 2nd derivative, to illustrate their mathematical similarity. Bottom: the laser LIA signal and absolute value of the IV curve from Fig. 59, demonstrating their similarity in the same way.

but it's possible it is. On the other hand, if the OR signal has a different enough shape, that is decent evidence that it is not entirely due to thermal effects. Our OR signals in Fig. 60 can be seen to have different shapes than their corresponding IV curves.

Related to this is the difficulty that, for the shape of IV curves seen in tunneling and RS, the D2 actually *does* often look like the original IV curve. This mean that even a measurement of the legitimate OR signal would give a signal that looks like a thermal signal. An example of this is illustrated in Fig. 61(a), where an analytic function $f(x) = xe^{0.5x^2}$ that resembles a typical IV curve we might see, is plotted (blue). Its analytically calculated D2,

$f''(x) = 3e^{0.5x^2} + x^3e^{0.5x^2}$, is also plotted (yellow), showing just how similar they are. With an added fuzz of noise on these signals, they could easily be mistaken to be the same curve. A practical example of this is shown in Fig. 61(b). The absolute value of the measured IV curve from Fig. 59(a) is plotted along with the measured OR signal from Fig. 59(b). They are a little different, but still very close.

References

[1] Hiroyuki Akinaga and Hisashi Shima. "Resistive random access memory (ReRAM) based on metal oxides". In: *Proceedings of the IEEE* 98.12 (2010), pp. 2237–2251.

[2] Stefano Ambrogio et al. "Statistical Fluctuations in HfO x Resistive-Switching Memory: Part I-Set/Reset Variability". In: *IEEE Transactions on Electron Devices* 61.8 (2014), pp. 2912–2919.

[3] S Ambrogio et al. "Understanding switching variability and random telegraph noise in resistive RAM". In: *Electron Devices Meeting (IEDM), 2013 IEEE International*. IEEE. 2013, pp. 31–5.

[4] Hidetaka Asoh et al. "Conditions for fabrication of ideally ordered anodic porous alumina using pretextured Al". In: *Journal of the Electrochemical Society* 148.4 (2001), B152–B156.

[5] Renaud Bachelot et al. "Apertureless near-field optical microscopy: A study of the local tip field enhancement using photosensitive azobenzene-containing films". In: *Journal of Applied Physics* 94.3 (2003), pp. 2060–2072.

[6] C Bradford Barber, David P Dobkin, and Hannu Huhdanpaa. "The quickhull algorithm for convex hulls". In: *ACM Transactions on Mathematical Software (TOMS)* 22.4 (1996), pp. 469–483.

[7] MIJ Beale et al. "An experimental and theoretical study of the formation and microstructure of porous silicon". In: *Journal of Crystal Growth* 73.3 (1985), pp. 622–636.

[8] A Belwalkar et al. "Effect of processing parameters on pore structure and thickness of anodic aluminum oxide (AAO) tubular membranes". In: *Journal of membrane science* 319.1 (2008), pp. 192–198.

[9] Brian R Bennett, Richard A Soref, and Jesus A Del Alamo. "Carrier-induced change in refractive index of InP, GaAs and InGaAsP". In: *IEEE Journal of Quantum Electronics* 26.1 (1990), pp. 113–122.

[10] Christian Bohling and Wolfgang Sigmund. "Self-Limitation of Native Oxides Explained". In: *Silicon* 8.3 (2016), pp. 339–343.

[11] Rémi Boidin et al. "Pulsed laser deposited alumina thin films". In: *Ceramics International* 42.1 (2016), pp. 1177–1182.

[12] An Chen and Ming-Ren Lin. "Variability of resistive switching memories and its impact on crossbar array performance". In: *Reliability Physics Symposium (IRPS), 2011 IEEE International*. IEEE. 2011, MY–7.

[13] Hao Ming Chen et al. "Controlling optical properties of aluminum oxide using electrochemical deposition". In: *Journal of The Electrochemical Society* 154.6 (2007), K11–K14.

[14] Sheng-Hui Chen et al. "Nanoimprinting pre-patterned effects on anodic aluminum oxide". In: *Japanese Journal of Applied Physics* 49.1R (2010), p. 015201.

[15] BJ Choi et al. “Resistive switching mechanism of TiO 2 thin films grown by atomic-layer deposition”. In: *Journal of Applied Physics* 98.3 (2005), p. 033715.

[16] TC Chou et al. “Microstructures and mechanical properties of thin films of aluminum oxide”. In: *Scripta metallurgica et materialia* 25.10 (1991), pp. 2203–2208.

[17] Leon Chua. “Memristor-the missing circuit element”. In: *IEEE Transactions on circuit theory* 18.5 (1971), pp. 507–519.

[18] Philippe Colomban, Aurelie Tournie, and Paola Ricciardi. “Raman spectroscopy of copper nanoparticle-containing glass matrices: ancient red stained-glass windows”. In: *Journal of Raman Spectroscopy* 40.12 (2009), pp. 1949–1955.

[19] G Dearnaley, AM Stoneham, and DV Morgan. “Electrical phenomena in amorphous oxide films”. In: *Reports on Progress in Physics* 33.3 (1970), p. 1129.

[20] Alexandros Emboras et al. “Nanoscale plasmonic memristor with optical readout functionality”. In: *Nano letters* 13.12 (2013), pp. 6151–6155.

[21] UR Evans. “Colors exhibited by thin films on metal”. In: *Journal of Colloid Science* 11.4 (1956), pp. 314–316.

[22] Francis P Fehlner. “Low temperature oxidation, the role of vitreous oxides”. In: (1986).

[23] J Feinleib et al. “Rapid reversible light-induced crystallization of amorphous semiconductors”. In: *Applied Physics Letters* 18.6 (1971), pp. 254–257.

[24] Daniel Franklin et al. “Actively addressed single pixel full-colour plasmonic display”. In: *Nature Communications* 8 (2017), p. 15209.

[25] Daniel Franklin et al. “Polarization-independent actively tunable colour generation on imprinted plasmonic surfaces”. In: *Nature communications* 6 (2015).

[26] Saeko Furuya et al. “Improvement of the Reproducibility of the Switching Voltage of Resistance Change Random Access Memory by Restricting Formation of Conductive Filaments”. In: *Japanese Journal of Applied Physics* 52.6S (2013), 06GF07.

[27] R Gottschalg, DG Infield, and MJ Kearney. “Experimental study of variations of the solar spectrum of relevance to thin film solar cells”. In: *Solar Energy materials and solar cells* 79.4 (2003), pp. 527–537.

[28] Charles A Grubbs. “Anodizing of aluminum”. In: *Metal Finishing* 100 (2002), pp. 463–478.

[29] Khaled Habib. “Measurement of surface resistivity/conductivity of anodized aluminium alloy by optical interferometry techniques”. In: *Proc. of SPIE Vol.* Vol. 7649. 2010, pp. 764922–1.

[30] Georg Hass. “On the preparation of hard oxide films with precisely controlled thickness on evaporated aluminum mirrors”. In: *JOSA* 39.7 (1949), pp. 532–540.

[31] Eugene Hecht. “Optics 4th edition”. In: *Optics, 4th Edition, Addison Wesley Longman Inc, 1998* (1998).

[32] TW Hickmott. “Electroforming and Ohmic contacts in Al-Al2O3-Ag diodes”. In: *Journal of Applied Physics* 111.6 (2012), p. 063708.

[33] TW Hickmott. “Electron Emission, Electroluminescence, and Voltage-Controlled Negative Resistance in Al–Al2O3–Au Diodes”. In: *Journal of Applied Physics* 36.6 (1965), pp. 1885–1896.

[34] TW Hickmott. "Low-frequency negative resistance in thin anodic oxide films". In: *Journal of Applied Physics* 33.9 (1962), pp. 2669–2682.

[35] Yooichi Hirose and Haruo Hirose. "Polarity-dependent memory switching and behavior of Ag dendrite in Ag-photodoped amorphous As2S3 films". In: *Journal of Applied Physics* 47.6 (1976), pp. 2767–2772.

[36] Bernard J Hockey. "Plastic deformation of aluminum oxide by indentation and abrasion". In: *Journal of the American Ceramic Society* 54.5 (1971), pp. 223–231.

[37] C Hoessbacher et al. "The plasmonic memristor: a latching optical switch". In: *Optica* 1.4 (2014), pp. 198–202.

[38] J Houska et al. "Overview of optical properties of Al 2 O 3 films prepared by various techniques". In: *Thin Solid Films* 520.16 (2012), pp. 5405–5408.

[39] Chi-Hsin Huang et al. "Manipulated transformation of filamentary and homogeneous resistive switching on ZnO thin film memristor with controllable multistate". In: *ACS applied materials & interfaces* 5.13 (2013), pp. 6017–6023.

[40] MS Hunter and P Fowle. "Determination of barrier layer thickness of anodic oxide coatings". In: *Journal of the Electrochemical Society* 101.9 (1954), pp. 481–485.

[41] Daniele Ielmini. "Resistive switching memories based on metal oxides: mechanisms, reliability and scaling". In: *Semiconductor Science and Technology* 31.6 (2016), p. 063002.

[42] Daisuke Inoue et al. "Polarization independent visible color filter comprising an aluminum film with surface-plasmon enhanced transmission

through a subwavelength array of holes". In: *Applied Physics Letters* 98.9 (2011), p. 093113.

[43] Timothy D James, Paul Mulvaney, and Ann Roberts. "The Plasmonic Pixel: large area, wide gamut color reproduction using aluminum nanostructures". In: *Nano letters* (2016).

[44] Jin Jang et al. "Electric-field-enhanced crystallization of amorphous silicon". In: *Nature* 395.6701 (1998), p. 481.

[45] O Jessensky, F Müller, and U Gösele. "Self-organized formation of hexagonal pore arrays in anodic alumina". In: *Applied physics letters* 72.10 (1998), pp. 1173–1175.

[46] Mikhail A Kats et al. "Nanometre optical coatings based on strong interference effects in highly absorbing media". In: *Nature materials* 12.1 (2013), pp. 20–24.

[47] F Keller, MS Hunter, and DL Robinson. "Structural features of oxide coatings on aluminum". In: *Journal of the Electrochemical Society* 100.9 (1953), pp. 411–419.

[48] James Kolodzey et al. "Electrical conduction and dielectric breakdown in aluminum oxide insulators on silicon". In: *IEEE Transactions on Electron Devices* 47.1 (2000), pp. 121–128.

[49] H Kroger, LN Smith, and DW Jillie. "Selective niobium anodization process for fabricating Josephson tunnel junctions". In: *Applied Physics Letters* 39.3 (1981), pp. 280–282.

[50] Yu-Hsuan Kuo et al. "Strong quantum-confined Stark effect in germanium quantum-well structures on silicon". In: *Nature* 437.7063 (2005), p. 1334.

[51] Deok-Hwang Kwon et al. “Atomic structure of conducting nanofilaments in TiO2 resistive switching memory”. In: *Nature nanotechnology* 5.2 (2010), pp. 148–153.

[52] CE Leberknight and Benjamin Lustman. “An optical investigation of oxide films on metals”. In: *JOSA* 29.2 (1939), pp. 59–66.

[53] Kyu-Tae Lee et al. “Strong Resonance Effect in a Lossy Medium-Based Optical Cavity for Angle Robust Spectrum Filters”. In: *Advanced Materials* 26.36 (2014), pp. 6324–6328.

[54] Feiyue Li, Lan Zhang, and Robert M Metzger. “On the growth of highly ordered pores in anodized aluminum oxide”. In: *Chemistry of materials* 10.9 (1998), pp. 2470–2480.

[55] Jianyu Liang et al. “A growth pathway for highly ordered quantum dot arrays”. In: *Applied physics letters* 85.24 (2004), pp. 5974–5976.

[56] JG Mendoza-Alvarez, FD Nunes, and NB Patel. “Refractive index dependence on free carriers for GaAs”. In: *Journal of Applied Physics* 51.8 (1980), pp. 4365–4367.

[57] David AB Miller et al. “Band-edge electroabsorption in quantum well structures: The quantum-confined Stark effect”. In: *Physical Review Letters* 53.22 (1984), p. 2173.

[58] Adriano Moreira and Maribel Yasmina Santos. “Concave hull: A k-nearest neighbours approach for the computation of the region occupied by a set of points”. In: (2007).

[59] NI Mou and Massood Tabib-Azar. “Photoreduction of Ag+ in Ag/Ag 2 S/Au memristor”. In: *Applied Surface Science* 340 (2015), pp. 138–142.

[60] Ruth Muenstermann et al. “Coexistence of Filamentary and Homogeneous Resistive Switching in Fe-Doped SrTiO3 Thin-Film Memristive Devices”. In: *Advanced materials* 22.43 (2010), pp. 4819–4822.

[61] David G Murcray et al. “Variation of the Infrared Solar Spectrum Between 700 cm- 1 and 2240 cm- 1 with Altitude”. In: *Applied optics* 8.12 (1969), pp. 2519–2536.

[62] Priyanka Nayar et al. “Structural, optical and mechanical properties of amorphous and crystalline alumina thin films”. In: *Thin Solid Films* 568 (2014), pp. 19–24.

[63] Shintaro Otsuka et al. “Additional Electrochemical Treatment Effects on the Switching Characteristics of Anodic Porous Alumina Resistive Switching Memory”. In: *Japanese Journal of Applied Physics* 51.6S (2012), 06FF11.

[64] Shintaro Otsuka et al. “Electric conduction mechanism of resistive switching memory fabricated with anodic aluminum oxide”. In: *ECS Transactions* 50.34 (2013), pp. 49–54.

[65] H Over and AP Seitsonen. “Oxidation of metal surfaces”. In: *Science* 297.5589 (2002), pp. 2003–2005.

[66] DP Oxley. “Electroforming, switching and memory effects in oxide thin films”. In: *Active and Passive Electronic Components* 3.4 (1977), pp. 217–224.

[67] Clyde W Oyster. *The human eye: structure and function*. Sinauer Associates, 1999.

[68] H Pagnia and N Sotnik. “Bistable switching in electroformed metal–insulator–metal devices”. In: *physica status solidi (a)* 108.1 (1988), pp. 11–65.

[69] Ian Polmear et al. *Light alloys: metallurgy of the light metals*. Butterworth-Heinemann, 2017.

[70] Aleksandar D Rakić. "Algorithm for the determination of intrinsic optical constants of metal films: application to aluminum". In: *Applied optics* 34.22 (1995), pp. 4755–4767.

[71] A Saedi and M Ghorbani. "Electrodeposition of Ni–Fe–Co alloy nanowire in modified AAO template". In: *Materials Chemistry and Physics* 91.2 (2005), pp. 417–423.

[72] Melissa S Sander and L-S Tan. "Nanoparticle arrays on surfaces fabricated using anodic alumina films as templates". In: *Advanced functional materials* 13.5 (2003), pp. 393–397.

[73] Akihito Sawa. "Resistive switching in transition metal oxides". In: *Materials today* 11.6 (2008), pp. 28–36.

[74] David T Schoen, Aaron L Holsteen, and Mark L Brongersma. "Probing the electrical switching of a memristive optical antenna by STEM EELS". In: *Nature communications* 7 (2016).

[75] Adel S Sedra and Kenneth Carless Smith. *Microelectronic circuits*. Vol. 1. New York: Oxford University Press, 1998.

[76] Guo Liang Shang et al. "Fano resonance in anodic aluminum oxide based photonic crystals". In: *Scientific reports* 4 (2014), p. 3601.

[77] Yongqiang Shi et al. "Large stable photoinduced refractive index change in a nonlinear optical polyester polymer with disperse red side groups". In: *Applied physics letters* 58.11 (1991), pp. 1131–1133.

[78] Tennyson Smith. "Effect of surface roughness on ellipsometry of aluminum". In: *Surface Science* 56 (1976), pp. 252–271.

[79] J Sophocles. "Orfanidis,'Electromagnetic waves and antennas". In: (2003).

[80] WG Spitzer and HY Fan. "Determination of optical constants and carrier effective mass of semiconductors". In: *Physical Review* 106.5 (1957), p. 882.

[81] Wojciech J Stkepniowski, Agata Nowak-Stkepniowski, and Zbigniew Bojar. "Quantitative arrangement analysis of anodic alumina formed by short anodizations in oxalic acid". In: *Materials Characterization* 78 (2013), pp. 79–86.

[82] Wojciech J Stkepniowski et al. "Fast Fourier transform based arrangement analysis of poorly organized alumina nanopores formed via self-organized anodization in chromic acid". In: *Materials Letters* 117 (2014), pp. 69–73.

[83] Dmitri B Strukov et al. "The missing memristor found". In: *nature* 453.7191 (2008), p. 80.

[84] GD Sulka and KG Parkoła. "Temperature influence on well-ordered nanopore structures grown by anodization of aluminium in sulphuric acid". In: *Electrochimica Acta* 52.5 (2007), pp. 1880–1888.

[85] Grzegorz D Sulka et al. "Through-hole membranes of nanoporous alumina formed by anodizing in oxalic acid and their applications in fabrication of nanowire arrays". In: *Electrochimica Acta* 55.14 (2010), pp. 4368–4376.

[86] Jiyu Sun, Bharat Bhushan, and Jin Tong. "Structural coloration in nature". In: *Rsc Advances* 3.35 (2013), pp. 14862–14889.

[87] SN Svitasheva and AM Gilinsky. "Influence of doping level on shift of the absorption edge of gallium nitride films (Burstein-Moss effect)". In: *Applied Surface Science* 281 (2013), pp. 109–112.

[88] Yusuke Tanimoto et al. “Effect of confining filaments on the current–voltage characteristics of resistive change memory by using anodic porous alumina”. In: *Japanese Journal of Applied Physics* 53.6S (2014), 06JF07.

[89] Tian Tian et al. “Study on titania nanotube arrays prepared by titanium anodization in NH4F/H2SO4 solution”. In: *Journal of materials science* 42.14 (2007), pp. 5539–5543.

[90] YT Tian et al. “Alumina nanowire arrays standing on a porous anodic alumina membrane”. In: *Nanotechnology* 15.1 (2003), p. 189.

[91] PK Tien and JP Gordon. “Multiphoton process observed in the interaction of microwave fields with the tunneling between superconductor films”. In: *Physical Review* 129.2 (1963), p. 647.

[92] J Tucker. “Quantum limited detection in tunnel junction mixers”. In: *IEEE Journal of Quantum Electronics* 15.11 (1979), pp. 1234–1258.

[93] Franz Urbach. “The long-wavelength edge of photographic sensitivity and of the electronic absorption of solids”. In: *Physical Review* 92.5 (1953), p. 1324.

[94] Biao Wang et al. “Preparation of photonic crystals made of air pores in anodic alumina”. In: *Nanotechnology* 18.36 (2007), p. 365601.

[95] Guoqiang Wang et al. “Synthesis and characterization of Ag nanoparticles assembled in ordered array pores of porous anodic alumina by chemical deposition”. In: *Materials Letters* 61.18 (2007), pp. 3795–3797.

[96] Jian Wang et al. “Optical constants of anodic aluminum oxide films formed in oxalic acid solution”. In: *Thin Solid Films* 516.21 (2008), pp. 7689–7694.

[97] Daniel R Ward et al. “Optical rectification and field enhancement in a plasmonic nanogap”. In: *Nature nanotechnology* 5.10 (2010), pp. 732–736.

[98] Rainer Waser and Masakazu Aono. “Nanoionics-based resistive switching memories”. In: *Nature materials* 6.11 (2007), pp. 833–840.

[99] Rainer Waser et al. “Redox-based resistive switching memories–nanoionic mechanisms, prospects, and challenges”. In: *Advanced materials* 21.25-26 (2009), pp. 2632–2663.

[100] Krongkamol Wong-Ek et al. “Silver nanoparticles deposited on anodic aluminum oxide template using magnetron sputtering for surface-enhanced Raman scattering substrate”. In: *Thin Solid Films* 518.23 (2010), pp. 7128–7132.

[101] Lee Woo et al. “Fast fabrication of long-range ordered porous alumina membranes by hard anodization”. In: *Nature materials* 5.9 (2006), p. 741.

[102] Jiancai Xue et al. “Scalable, full-colour and controllable chromotropic plasmonic printing”. In: *Nature communications* 6 (2015).

[103] Chenying Yang et al. “Compact multilayer film structure for angle insensitive color filtering”. In: *Scientific reports* 5 (2015).

[104] J Joshua Yang et al. “Memristive switching mechanism for metal/oxide/metal nanodevices”. In: *Nature nanotechnology* 3.7 (2008), pp. 429–433.

[105] Yuchao Yang et al. “Observation of conducting filament growth in nanoscale resistive memories”. In: *Nature communications* 3 (2012), p. 732.

[106] Leszek Zaraska et al. “Porous anodic alumina formed by anodization of aluminum alloy (AA1050) and high purity aluminum”. In: *Electrochimica Acta* 55.14 (2010), pp. 4377–4386.

[107] Li-Rong Zhao et al. “Anodic aluminum oxide films formed in mixed electrolytes of oxalic and sulfuric acid and their optical constants”. In: *Physica B: condensed matter* 405.1 (2010), pp. 456–460.